Springer Proceedings in Mathematics & Statistics

Josef Dick
Frances Y. Kuo
Gareth W. Peters
Ian H. Sloan *Editors*

Monte Carlo and Quasi-Monte Carlo Methods 2012

Springer

Springer Proceedings in Mathematics and Statistics

Volume 65

For further volumes:
http://www.springer.com/series/10533

Springer Proceedings in Mathematics and Statistics

This book series features volumes composed of selected contributions from workshops and conferences in all areas of current research in mathematics and statistics, including OR and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

Josef Dick • Frances Y. Kuo • Gareth W. Peters
Ian H. Sloan
Editors

Monte Carlo and Quasi-Monte Carlo Methods 2012

Springer

Editors
Josef Dick
Frances Y. Kuo
Gareth W. Peters
Ian H. Sloan
The University of New South Wales
School of Mathematics and Statistics
New South Wales
Sydney, Australia

ISSN 2194-1009 ISSN 2194-1017 (electronic)
ISBN 978-3-642-41094-9 ISBN 978-3-642-41095-6 (eBook)
DOI 10.1007/978-3-642-41095-6
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013956514

Mathematics Subject Classification (2010): Primary: 11K38, 11K45, 65-06, 65C05, 65C10, 65D30
Secondary: 11K38, 65D18, 65D32, 65R20, 91B25

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This volume represents the refereed proceedings of the Tenth International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing, which was held at the University of New South Wales, Sydney, Australia, from 13 to 17 February 2012. It contains a limited selection of articles based on presentations given at the conference. The conference program was arranged with the help of an international committee consisting of:

- William Chen, *Macquarie University, Australia*
- Ronald Cools, *KU Leuven, Belgium*
- Josef Dick, *University of New South Wales, Australia* (Conference organizer)
- Henri Faure, *CNRS Marseille, France*
- Alan Genz, *Washington State University, USA*
- Mike Giles, *University of Oxford, UK*
- Paul Glasserman, *Columbia University, USA*
- Michael Gnewuch, *University of Kaiserslautern, Germany*
- Stefan Heinrich, *University of Kaiserslautern, Germany*
- Fred J. Hickernell, *Illinois Institute of Technology, USA*
- Aicke Hinrichs, *University of Rostock, Germany*
- Stephen Joe, *University of Waikato, New Zealand*
- Aneta Karaivanova, *Bulgarian Academy of Science, Bulgaria*
- Alexander Keller, *NVIDIA, Germany*
- Dirk P. Kroese, *University of Queensland, Australia*
- Frances Y. Kuo, *University of New South Wales, Australia* (Conference organizer)
- Gerhard Larcher, *Johannes Kepler University Linz, Austria*
- Pierre L'Ecuyer, *Université de Montréal, Canada*
- Christiane Lemieux, *University of Waterloo, Canada*
- Peter Mathé, *Weierstrass Institute Berlin, Germany*
- Makoto Matsumoto, *Hiroshima University, Japan*
- Kerrie Mengersen, *Queensland University of Technology, Australia*
- Thomas Müller-Gronbach, *University of Passau, Germany*
- Harald Niederreiter, *RICAM Linz and University of Salzburg, Austria*
- Erich Novak, *University of Jena, Germany*
- Art B. Owen, *Stanford University, USA*

- Gareth W. Peters, *University of New South Wales, Australia, and University College London, UK* (Conference organizer)
- Friedrich Pillichshammer, *Johannes Kepler University Linz, Austria*
- Leszek Plaskota, *University of Warsaw, Poland*
- Eckhard Platen, *University of Technology Sydney, Australia*
- Klaus Ritter, *University of Kaiserslautern, Germany*
- Gareth Roberts, *University of Warwick, UK*
- Wolfgang Ch. Schmid, *University of Salzburg, Austria*
- Nikolai Simonov, *Russian Academy of Sciences, Russia*
- Ian H. Sloan, *University of New South Wales, Australia* (Conference organizer)
- Ilya M. Sobol', *Russian Academy of Sciences, Russia*
- Jerome Spanier, *Claremont, California, USA*
- Shu Tezuka, *Kyushu University, Japan*
- Xiaoqun Wang, *Tsinghua University, China*
- Grzegorz W. Wasilkowski, *University of Kentucky, USA*
- Henryk Woźniakowski, *Columbia University, USA, and University of Warsaw, Poland*

This conference continued the tradition of biennial MCQMC conferences initiated by Harald Niederreiter, held previously at:

- University of Nevada in Las Vegas, Nevada, USA, in June 1994
- University of Salzburg, Austria, in July 1996
- Claremont Colleges in Claremont, California, USA, in June 1998
- Hong Kong Baptist University in Hong Kong, China, in November 2000
- National University of Singapore, Republic of Singapore, in November 2002
- Palais des Congrès in Juan-les-Pins, France, in June 2004
- Ulm University, Germany, in July 2006
- Université de Montréal, Canada, in July 2008
- University of Warsaw, Poland, in August 2010

The next conference will be held at the KU Leuven, Belgium, in April 2014.

The proceedings of these previous conferences were all published by Springer-Verlag, under the following titles:

- *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing* (H. Niederreiter and P.J.-S. Shiue, eds.)
- *Monte Carlo and Quasi-Monte Carlo Methods 1996* (H. Niederreiter, P. Hellekalek, G. Larcher, and P. Zinterhof, eds.)
- *Monte Carlo and Quasi-Monte Carlo Methods 1998* (H. Niederreiter and J. Spanier, eds.)
- *Monte Carlo and Quasi-Monte Carlo Methods 2000* (K.-T. Fang, F.J. Hickernell, and H. Niederreiter, eds.)
- *Monte Carlo and Quasi-Monte Carlo Methods 2002* (H. Niederreiter, ed.)
- *Monte Carlo and Quasi-Monte Carlo Methods 2004* (H. Niederreiter and D. Talay, eds.)
- *Monte Carlo and Quasi-Monte Carlo Methods 2006* (A. Keller, S. Heinrich, and H. Niederreiter, eds.)

- *Monte Carlo and Quasi-Monte Carlo Methods 2008* (P. L'Ecuyer and A. Owen, eds.)
- *Monte Carlo and Quasi-Monte Carlo Methods 2010* (L. Plaskota and H. Woźniakowski, eds.)

The program of the conference was rich and varied with over 140 talks being presented. Highlights were the invited plenary talks given by Pierre Del Moral (INRIA and University of Bordeaux 1), Mike Giles (Oxford University), Fred J. Hickernell (Illinois Institute of Technology), Aicke Hinrichs (University of Jena), Michael Lacey (Georgia Institute of Technology), Kerrie Mengersen (Queensland University of Technology), Andreas Neuenkirch (University of Kaiserslautern), Art B. Owen (Stanford University), Leszek Plaskota (University of Warsaw), and Eckhard Platen (University of Technology Sydney), and the tutorials given by Art B. Owen (Stanford University), Pierre Del Moral (INRIA and University of Bordeaux 1), Josef Dick (University of New South Wales), and Alex Keller (NVIDIA).

The papers in this volume were carefully screened and cover both the theory and the applications of Monte Carlo and quasi-Monte Carlo methods. We thank the anonymous reviewers for their reports and many others who contributed enormously to the excellent quality of the conference presentations and to the high standards for publication in these proceedings by careful review of the abstracts and manuscripts that were submitted.

We gratefully acknowledge generous financial support of the conference by the School of Mathematics and Statistics of the University of New South Wales, the Australian Mathematical Society (AustMS), the Australian and New Zealand Industrial and Applied Mathematics (ANZIAM), the Australian Mathematical Sciences Institute (AMSI), the Commonwealth Scientific and Industrial Research Organisation (CSIRO), and the National Science Foundation (NSF).

Finally, we want to express our gratitude to Springer-Verlag for publishing this volume.

Sydney, Australia
September 2013

Josef Dick
Frances Y. Kuo
Gareth W. Peters
Ian H. Sloan

Contents

Part III Contributed Articles

Part I
Invited Articles

Computing Functionals of Square Root and Wishart Processes Under the Benchmark Approach via Exact Simulation

Jan Baldeaux and Eckhard Platen

Abstract The aim of the paper is to show how Wishart processes can be used flexibly in financial modeling. We explain how functionals, resulting from the benchmark approach to finance, can be accurately computed via exact simulation methods. We employ Lie symmetry methods to identify explicit transition densities and explicitly computable functionals. We illustrate the proposed methods via finance problems formulated under the benchmark approach. This approach allows us to exploit conveniently the analytical tractability of the considered diffusion processes.

1 Introduction

In mathematical finance, the pricing of financial derivatives can under suitable conditions be shown to amount to the computation of an expected value, see e.g. [50, 53]. We focus in this paper on the application of the benchmark approach, described e.g. in [53], where we show how Wishart processes can be flexibly used in financial modeling and derivative pricing. Depending on the financial derivative and the model under consideration, it might not be possible to compute the expected value explicitly, however, numerical methods have to be invoked. A candidate for the computation of such expectations is the Monte Carlo method, see e.g. [11, 29],

J. Baldeaux (✉)
Finance Discipline Group, University of Technology, PO Box 123, Broadway, Sydney, NSW, 2007, Australia
Current address: Quant Models & Development, Danske Bank, Denmark
e-mail: JanBaldeaux@gmail.com

E. Platen
Finance Discipline Group and School of Mathematical Sciences, University of Technology, PO Box 123, Broadway, Sydney, NSW, 2007, Australia
e-mail: Eckhard.Platen@uts.edu.au

J. Dick et al. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2012*,
Springer Proceedings in Mathematics & Statistics 65,
DOI 10.1007/978-3-642-41095-6_1, © Springer-Verlag Berlin Heidelberg 2013

and [40]. Applying the Monte Carlo method typically entails the sampling of the distribution of the relevant financial state variables, e.g. an equity index, a short rate, or a commodity price. It is then, of course, desirable to have at one's disposal a recipe for drawing samples from the relevant distributions. In case these distributions are known, one refers to exact simulation schemes, see e.g. [52], but also [7–9], and [16], for further references on exact simulation schemes. In particular, exact simulation is relevant for long term simulation. If exact simulation schemes are not applicable, discrete time approximations, as analyzed in [40] and [52] become relevant.

For modeling financial quantities of interest, it is important to know a priori if exact simulation schemes exist, so that financial derivatives can be priced accurately, even if expected values cannot be computed explicitly. In this paper, we discuss classes of square root and Wishart processes for which exact simulation is possible. For one-dimensional diffusions, Lie symmetry analysis, see [10], and [51] turns out to be a useful tool to identify tractable diffusion processes. Besides allowing one to discover transition densities, see [21], it also allows us to compute Laplace transforms of important multidimensional functionals, see e.g. [20]. In particular, squared Bessel processes fall into the class of diffusions that can be handled well via Lie symmetry methods.

The Wishart process [13], is the multidimensional extension of the squared Bessel process. It turns out, see [32] and [33], that Wishart processes are affine processes, i.e. their characteristic function is exponentially affine in the state variables. We point out that in [32], and [33] the concept of an affine process was generalized from real-valued processes to matrix-valued processes, where the latter category covers Wishart processes. Furthermore, the characteristic function can be computed explicitly, see [32], and [33]. Finally, we remark that in [1] an exact simulation scheme for Wishart processes was presented.

Modeling financial quantities, one aims for models which provide a reasonably accurate reflection of reality, whilst at the same time retaining analytical tractability. The benchmark approach, see [53], offers a unified rich modeling framework to derivative pricing, risk management, and portfolio optimization. It allows one to use a much wider range of empirically supported models than under the classical no-arbitrage approach. At the heart of the benchmark approach sits the growth optimal portfolio (GOP). It is the portfolio which maximizes expected log-utility from terminal wealth. In particular, the benchmark approach uses the GOP as numéraire and benchmark and the real world probability measure for taking expectations. The paper combines and reviews various recent results on Wishart processes, Lie symmetry group methods and the benchmark approach with focus on exact Monte carlo simulation for derivative pricing. We demonstrate using examples that the benchmark approach is easily applied for the mentioned class of processes for which exact simulation is possible.

The remaining structure of the paper is as follows: In Sect. 2 we introduce the benchmark approach using a particular model for illustration, the minimal market model (MMM), see [53]. Section 3 introduces Lie symmetry methods and discusses how they can be applied in the context of the benchmark approach. Section 4

presents Wishart processes and shows how they can be used to extend the MMM. Section 6 concludes the paper.

2 Benchmark Approach

We focus in our selection of stochastic processes and the choice of examples on their suitability under the benchmark approach. The GOP plays a pivotal role as benchmark and numéraire under the benchmark approach. It also enjoys a prominent position in the finance literature, see [39], but also [12, 41, 45–47], and [55]. The benchmark approach uses the GOP as the numéraire. Since the GOP is the numéraire portfolio, see [45], contingent claims are priced under the real world probability measure. This avoids the restrictive assumption on the existence of an equivalent risk-neutral probability measure. We remark, it is argued in [53] that the existence of such a measure may not be a realistic assumption. Finally, we emphasize that the benchmark approach can be seen as a generalization of risk-neutral pricing, as well as other pricing approaches, such as actuarial pricing, see [53].

To fix ideas in a simple manner, we model a well-diversified index, which we interpret as the GOP, using the stylized version of the MMM, see [53]. Though parsimonious, this model is able to capture important empirical characteristics of well-diversified indices. It has subsequently been extended in several ways, see e.g. [53], and also [4]. To be precise, consider a filtered probability space $(\Omega, \mathscr{A}, \underline{\mathscr{A}}, P)$, where the filtration $\underline{\mathscr{A}} = (\mathscr{A}_t)_{t\in[0,\infty)}$ is assumed to satisfy the usual conditions, which carries, for simplicity, one source of uncertainty, a standard Brownian motion $W = \{W(t),\, t \in [0,\infty)\}$. The deterministic savings account is modeled using the differential equation

$$dS_t^0 = r\, S_t^0\, dt\,,$$

for $t \in [0,\infty)$ with $S_0^0 = 1$, where r denotes the constant short rate. Next, we introduce the model for the well diversified index, the GOP $S_t^{\delta *}$, which is given by the expression

$$S_t^{\delta *} = S_t^0\, \bar{S}_t^{\delta *} = S_t^0\, Y_t\, \alpha_t^{\delta *}\,. \tag{1}$$

Here $Y_t = \frac{\bar{S}_t^{\delta *}}{\alpha_t^{\delta *}}$ is a square-root process of dimension four, satisfying the stochastic differential equation (SDE)

$$dY_t = (1 - \eta\, Y_t)\, dt + \sqrt{Y_t}\, dW(t)\,, \tag{2}$$

for $t \in [0,\infty)$ with initial value $Y_0 > 0$ and net growth rate $\eta > 0$. Here $W = \{W(t)\,,\, t \geq 0\}$ is a standard Brownian motion. The deterministic function of time $\alpha_t^{\delta *}$ is given by the exponential function

$$\alpha_t^{\delta *} = \alpha_0 \exp\{\eta t\}\,,$$

with scaling parameter $\alpha_0 > 0$. Furthermore, it can be shown by the Itô formula that $\alpha_t^{\delta*}$ is the drift at time t of the discounted GOP

$$\bar{S}_t^{\delta*} := \frac{S_t^{\delta*}}{S_t^0},$$

so that the parameters of the model are $S_0^{\delta*}$, α_0, η, and r. We note that one obtains for the GOP the SDE

$$dS_t^{\delta*} = S_t^{\delta*}\left(\left(r + \frac{1}{Y_t}\right)dt + \sqrt{\frac{1}{Y_t}}dW(t)\right). \tag{3}$$

This SDE models the well-observed leverage effect, since as the index $S_t^{\delta*}$ decreases, its volatility $\frac{1}{\sqrt{Y_t}} = \sqrt{\frac{\alpha_t^{\delta*}}{\bar{S}_t^{\delta*}}}$ increases and vice versa.

It is useful to define the transformed time $\varphi(t)$ as

$$\varphi(t) = \varphi(0) + \frac{1}{4}\int_0^t \alpha_s^{\delta*} ds.$$

Setting

$$X_{\varphi(t)} = \bar{S}_t^{\delta*},$$

we obtain the SDE

$$dX_{\varphi(t)} = 4d\varphi(t) + 2\sqrt{X_{\varphi(t)}}dW_{\varphi(t)}, \tag{4}$$

where

$$dW_{\varphi(t)} = \sqrt{\frac{\alpha_t^{\delta*}}{4}}dW(t),$$

for $t \in [0, \infty)$. This shows that $X = \{X_\varphi, \varphi \in [\varphi(0), \infty)\}$ is a time transformed squared Bessel process of dimension four and $W = \{W_\varphi, \varphi \in [\varphi(0), \infty)\}$ is a Wiener process in the transformed φ-time $\varphi(t) \in [\varphi(0), \infty)$, see [54]. The merit of the dynamics given by (4) is that transition densities of squared Bessel processes are well studied. In fact we derive them in Sect. 3 using Lie symmetry methods.

We remark that the MMM does not admit a risk-neutral probability measure because the Radon-Nikodym derivative $\Lambda_t = \frac{\bar{S}_0^{\delta*}}{\bar{S}_t^{\delta*}}$ of the putative risk-neutral measure, which is the inverse of a time transformed squared Bessel process of dimension four, is a strict local martingale and not a martingale, see [54]. On the other hand, $S^{\delta*}$, is the numéraire portfolio, and thus, when used as numéraire to denominate

any nonnegative portfolio, yields a supermartingale under the real-world probability measure P. This implies that the financial market under consideration is free of those arbitrage opportunities that are economically meaningful in the sense that they would allow to create strictly positive wealth out of zero initial wealth via a nonnegative portfolio, that is, under limited liability, see [44] and [53]. This also means that we can price contingent claims under P employing S^{δ_*} as the numéraire. This pricing concept is referred to as real-world pricing, which we now recall, see [53]: For a nonnegative contingent claim with payoff H at maturity T, where H is $\mathscr{A}_T$-measurable, and $E\left(\frac{H}{S_T^{\delta_*}}\right) < \infty$, we define the value process at time $t \in [0, T]$ by

$$V_t := S_t^{\delta_*} E\left(\frac{H}{S_T^{\delta_*}} \,\middle|\, \mathscr{A}_t\right). \tag{5}$$

Note that since $V_T = H$, the benchmarked price process $\frac{V_t}{S_t^{\delta_*}}$ is an $(\underline{\mathscr{A}}, P)$-martingale. Formula (5) represents the real-world pricing formula, which provides the minimal possible price and will be used in this paper to price derivatives. If the expectation in Eq. (5) cannot be computed explicitly, one can resort to Monte Carlo methods. In that case, it is particularly convenient, if the relevant financial quantities, such as $S_T^{\delta_*}$ can be simulated exactly. In the next section, we derive the transition density of S^{δ_*} via Lie symmetry methods, which then allows us to simulate $S_T^{\delta_*}$ exactly. Note, in Sect. 4 we generalize the MMM to a multidimensional setting and present a suitable exact simulation algorithm.

3 Lie Symmetry Methods

The aim of this section is to present Lie symmetry methods as an effective tool for designing tractable models in mathematical finance. Tractable models are, in particular, useful for the evaluation of derivatives and risk measures in mathematical finance. We point out that in the literature, Lie symmetry methods have been used to solve mathematical finance problems explicitly, see e.g. [19], and [37]. Within the current paper we want to demonstrate that they can also be used to design efficient Monte Carlo algorithms for complex multidimensional functionals.

The advantage of the use of Lie symmetry methods is that it is straightforward to check whether the method is applicable or not. If the method is applicable, then the relevant solution or its Laplace transform has usually already been obtained in the literature or can be systematically derived. We will demonstrate this in finance applications using the benchmark approach for pricing.

We now follow [20], and recall that if the solution of the Cauchy problem

$$u_t = bx^{\gamma} u_{xx} + f(x)u_x - g(x)u\,,\ x > 0\,,\ t \geq 0\,, \tag{6}$$

$$u(x, 0) = \varphi(x)\,,\ x \in \Omega = [0, \infty)\,, \tag{7}$$

is unique, then by using the Feynman-Kac formula it is given by the expectation

$$u(x,t) = E\left(\exp\left(-\int_0^t g(X_s)ds\right)\varphi(X_t)\right),$$

where $X_0 = x$, and the stochastic process $X = \{X_t\,,\ t \geq 0\}$ satisfies the SDE

$$dX_t = f(X_t)dt + \sqrt{2bX_t^{\gamma}}dW_t\,.$$

We now briefly indicate the intuition behind the application of Lie symmetry methods to problems from mathematical finance. In particular, the integral transform method developed in [43], and the types of results this approach can produce. Lie's method allows one to find vector fields

$$\mathbf{v} = \xi(x,t,u)\partial_x + \tau(x,t,u)\partial_t + \phi(x,t,u)\partial_u\,,$$

which generate one parameter Lie groups that preserve solutions of (6). It is standard to denote the action of $\mathbf{v}$ on solutions $u(x,t)$ of (6) by

$$\rho(\exp\epsilon\mathbf{v})u(x,t) = \sigma(x,t;\epsilon)u(a_1(x,t;\epsilon),a_2(x,t;\epsilon))\,, \tag{8}$$

for some functions σ, a_1, and a_2. Here ϵ is the parameter of the group, σ is referred to as the multiplier, and a_1 and a_2 are changes of variables of the symmetry, see [19] for more details. For the applications we have in mind, ϵ and σ are of crucial importance. The parameter ϵ will play the role of the transform parameter of the Fourier or Laplace transform and σ will usually be the Fourier or Laplace transform of the transition density. Following [19], we assume that (6) has a fundamental solution $p(t,x,y)$. For this paper, it suffices to recall that we can express a solution $u(x,t)$ of the PDE (6) subject to the boundary condition $u(x,0) = f(x)$ in the form

$$u(x,t) = \int_{\Omega} f(y)\,p(t,x,y)dy\,, \tag{9}$$

where $p(t,x,y)$ is a fundamental solution of (6). The key idea of the transform method is to connect (8) and (9). Now consider a stationary, i.e. a time-independent solution, say $u_0(x)$. Of course, (8) yields

$$\rho\left(\exp\epsilon\mathbf{v}\right)u_0(x) = \sigma\left(x,t;\epsilon\right)u_0\left(a_1(x,t;\varepsilon)\right)\,,$$

which also solves the initial value problem. We now set $t = 0$ and use (8) and (9) to obtain

$$\int_{\Omega}\sigma(y,0;\epsilon)u_0\left(a_1\left(y,0,;\epsilon\right)\right)p\left(t,x,y\right)dy = \sigma\left(x,t;\epsilon\right)u_0\left(a_1\left(x,t;\epsilon\right)\right)\,. \tag{10}$$

Since σ, u_0, and a_1 are known functions, we have a family of integral equations for $p(t, x, y)$. To illustrate this idea using an example, we consider the one-dimensional heat equation

$$u_t = \frac{1}{2} g^2 u_{xx} \,. \tag{11}$$

We will show that if $u(x, t)$ solves (11), then for ϵ sufficiently small, so does

$$\tilde{u}(z, t) = \exp\left\{ \frac{\epsilon t^2}{2g^2} - \frac{z\epsilon}{g^2} \right\} u\left(z - t\epsilon, t\right) \,.$$

Taking $u_0 = 1$, (10) gives

$$\int_{-\infty}^{\infty} \exp\left\{ -\frac{y\epsilon}{g^2} \right\} p(t, x, y) dy = \exp\left\{ \frac{t\epsilon^2}{2g^2} - \frac{x\epsilon}{g^2} \right\} \,.$$

Setting $a = -\frac{\epsilon}{g^2}$, we get

$$\int_{-\infty}^{\infty} \exp\{ay\} p(t, x, y) dy = \exp\left\{ \frac{a^2 g^2 t}{2} + ax \right\} \,. \tag{12}$$

We recognize that (12) is the moment generating function of the Gaussian distribution, so $p(t, x, y)$ is the Gaussian density with mean x and variance $g^2 t$. We alert the reader to the fact that ϵ plays the role of the transform parameter and σ corresponds to the moment generating function. Finally, we recall a remark from [17], namely the fact that Laplace and Fourier transforms can be readily obtained through Lie algebra computations, which suggests a deep relationship between Lie symmetry analysis and harmonic analysis. Lastly, we remark that in order to apply the approach, we require the PDE (6) to have nontrivial symmetries. The approach developed by Craddock and collaborators, see [17–20], and [21], provides us with the following: A statement confirming if nontrivial symmetries exist and an expression stemming from (10), which one only needs to invert to obtain $p(t, x, y)$. We first present theoretical results, and then apply these to the case of the MMM. Now we discuss the question whether the PDE (6) has nontrivial symmetries, see [20], Proposition 2.1.

Theorem 1. *If $\gamma \neq 2$, then the PDE*

$$u_t = b x^{\gamma} u_{xx} + f(x) u_x - g(x) u \,, \quad x \geq 0 \,, b > 0 \tag{13}$$

has a nontrivial Lie symmetry group if and only if $h(x) = x^{1-\gamma} f(x)$ satisfies one of the following families of drift equations

$$bxh' - bh + \frac{1}{2}h^2 + 2bx^{2-\gamma}g(x) = 2bAx^{2-\gamma} + B\,, \tag{14}$$

$$bxh' - bh + \frac{1}{2}h^2 + 2bx^{2-\gamma}g(x) = \frac{Ax^{4-2\gamma}}{2\,(2-\gamma)^2} + \frac{Bx^{2-\gamma}}{2-\gamma} + C\,, \tag{15}$$

$$bxh' - bh + \frac{1}{2}h^2 + 2bx^{2-\gamma}g(x) = \frac{Ax^{4-2\gamma}}{2\,(2-\gamma)^2} + \frac{Bx^{3-\frac{3}{2}\gamma}}{3-\frac{3}{2}\gamma} + \frac{Cx^{2-\gamma}}{2-\gamma} - \kappa\,, \tag{16}$$

with $\kappa = \frac{\gamma}{8}(\gamma-4)\,b^2$.

For the case $\gamma = 2$, a similar result was obtained in [20], Proposition 2.1. Regarding the first Ricatti equation (14), the following result was described in [20], Theorem 3.1.

Theorem 2. *Suppose* $\gamma \neq 2$ *and* $h(x) = x^{1-\gamma}f(x)$ *is a solution of the Ricatti equation*

$$bxh' - bh + \frac{1}{2}h^2 + 2bx^{2-\gamma}g(x) = 2bAx^{2-\gamma} + B\,.$$

Then the PDE (13) has a symmetry of the form

$$\overline{U}_\varepsilon(x,t) = \frac{1}{(1+4\varepsilon t)^{\frac{1-\gamma}{2-\gamma}}} \exp\left\{\frac{-4\varepsilon\left(x^{2-\gamma} + Ab\,(2-\gamma)^2\,t^2\right)}{b\,(2-\gamma)^2\,(1+4\varepsilon t)}\right\} \tag{17}$$

$$\times \exp\left\{\frac{1}{2b}\left(F\left(\frac{x}{(1+4\varepsilon t)^{\frac{2}{2-\gamma}}}\right) - F\,(x)\right)\right\} \tag{18}$$

$$\times\, u\left(\frac{x}{(1+4\varepsilon t)^{\frac{2}{2-\gamma}}}, \frac{t}{1+4\varepsilon t}\right)\,, \tag{19}$$

where $F'(x) = f(x)/x^\gamma$ *and* u *is a solution of the respective PDE. That is, for* ε *sufficiently small,* U_ε *is a solution of (13) whenever* u *is. If* $u(x,t) = u_0(x)$ *with* u_0 *an analytic, stationary solution, then there is a fundamental solution* $p(t,x,y)$ *of (13) such that*

$$\int_0^\infty \exp\{-\lambda y^{2-\gamma}\}u_0\,(y)\;p\,(t,x,y)\,dy = U_\lambda(x,t)\,.$$

Here $U_\lambda(x,t) = \overline{U}_{\frac{1}{4}b(2-\gamma)^2\lambda}$. *Further, if* $u_0 = 1$, *then* $\int_0^\infty p(t,x,y)dy = 1$.

For the remaining two Ricatti equations, (15) and (16), we refer the reader to Theorems 2.5 and 2.8 in [17].

We would now like to illustrate how the method can be used. Consider a squared Bessel process of dimension δ, where $\delta \geq 2$,

$$dX_t = \delta dt + 2\sqrt{X_t} dW_t \,,$$

where $X_0 = x > 0$. The drift $f(x) = \delta$ satisfies Eq. (14) with $A = 0$. Consequently, using Theorem 2 with $A = 0$ and $u(x,t) = 1$, we obtain

$$\overline{U}_\varepsilon(x,t) = \exp\left\{-\frac{4\epsilon x}{b\,(1+4\varepsilon t)}\right\}(1+4\varepsilon t)^{-\frac{\delta}{b}} \,,$$

where $b = 2$. Setting $\varepsilon = \frac{b\lambda}{4}$, we obtain the Laplace transform

$$\begin{aligned} U_\lambda(x,t) &= \int_0^\infty \exp\{-\lambda y\}\, p(t,x,y) dy \\ &= \exp\left\{-\frac{x\lambda}{1+2\lambda t}\right\}(1+2\lambda t)^{-\frac{\delta}{2}} \,, \end{aligned}$$

which is inverted to yield

$$p(t,x,y) = \frac{1}{2t}\left(\frac{x}{y}\right)^{\frac{\nu}{2}} I_\nu\left(\frac{\sqrt{xy}}{t}\right)\exp\left\{-\frac{(x+y)}{2t}\right\} \,, \tag{20}$$

where $\nu = \frac{\delta}{2} - 1$ denotes the index of the squared Bessel process. Here I_ν denotes the modified Bessel function of the first kind. Equation (20) shows the transition density of a squared Bessel process started at time 0 in x for being at time t in y. This result, together with the real world pricing formula (5), allows us to price a wide range of European style and path-dependent derivatives with payoffs of the type $H = f(S^*_{t_1}, S^*_{t_2}, \ldots, S^*_{t_d})$, where $d \geq 1$ and $t_1, t_2, \ldots, t_d$ are given deterministic times.

By exploiting the tractability of the underlying processes, Lie symmetry methods allow us to design efficient Monte Carlo algorithms, as the following example from [3] and [2] shows. We now consider the problem of pricing derivatives on realized variance. Here we define realized variance to be the quadratic variation of the log-index, and we formally compute the quadratic variation of the log-index in the form,

$$\left[\log(S^{\delta *}_\cdot)\right]_T = \int_0^T \frac{dt}{Y_t} \,.$$

Recall from Sect. 2 that $Y = \{Y_t \,,\, t \geq 0\}$ is a square-root process whose dynamics are given in Eq. (2). In particular, we focus on put options on volatility, where volatility is defined to be the square-root of realized variance. We remark that call options on volatility can be obtained via the put-call parity relation in Lemma 4.1 in [3]. The real-world pricing formula (5) yields the following price for put options on volatility

$$S_t^{\delta_*} E\left(\frac{(K-\sqrt{\frac{1}{T}\int_0^T \frac{ds}{Y_s}})^+}{S_T^{\delta_*}}\middle|\mathscr{A}_t\right). \tag{21}$$

For computing the expectation in (21) via Monte Carlo methods, one first needs to have access to the joint density of $(S_T^{\delta_*}, \int_0^T \frac{ds}{Y_s})$ and subsequently perform the Monte Carlo simulation. Before presenting the relevant result, we recall that $S_T^{\delta_*} = S_T^0 \alpha_T^{\delta_*} Y_T$, i.e. it suffices to have access to the joint distribution of $(Y_T, \int_0^T \frac{dt}{Y_t})$. We remark that if we have access to the Laplace transform of $(Y_T, \int_0^T \frac{dt}{Y_t})$, i.e.

$$E\left(\exp\left(-\lambda Y_T - \mu \int_0^T \frac{dt}{Y_t}\right)\right), \tag{22}$$

then we have, in principle, solved the problem. From the point of view of implementation though, inverting a two-dimensional Laplace transform numerically is expensive. The following result from [20], see Corollaries 5.8 and 5.9, goes further: In fact the fundamental solution corresponds to inverting the expression in (22) with respect to λ, which significantly reduces the computational complexity.

Lemma 1. *The joint Laplace transform of Y_T and $\int_0^T \frac{dt}{Y_t}$ is given by*

$$\begin{aligned} &E\left(\exp\left(-\lambda Y_T - \mu \int_0^T \frac{1}{Y_t} dt\right)\right) \\ &= \frac{\Gamma(3/2+\nu/2)}{\Gamma(\nu+1)} \beta x^{-1} \exp\left(\eta\left(T + x - \frac{x}{\tanh(\eta T/2)}\right)\right) \\ &\quad \frac{1}{\beta\alpha} \exp\left(\beta^2/(2\alpha)\right) M_{-k,\nu/2}\left(\frac{\beta^2}{\alpha}\right), \end{aligned}$$

where $\alpha = \eta\left(1+\coth(\frac{\eta t}{2})\right) + \lambda$, $\beta = \frac{\eta\sqrt{x}}{\sinh\left(\frac{\eta t}{2}\right)}$, $\nu = 2\sqrt{\frac{1}{4}+2\mu}$, and $M_{s,r}(z)$ denotes the Whittaker function of the first kind. In [20], the inverse with respect to λ was already performed explicitly and is given as

$$\begin{aligned} p(T,x,y) = &\frac{\eta}{\sinh(\eta T/2)} \left(\frac{y}{x}\right)^{1/2} \\ &\times \exp\left(\eta\left(T + x - y - \frac{x+y}{\tanh(\eta T/2)}\right)\right) I_\nu\left(\frac{2\eta\sqrt{xy}}{\sinh(\eta T/2)}\right). \end{aligned} \tag{23}$$

Consequently, to recover the joint density of $(Y_T, \int_0^T \frac{dt}{Y_t})$, one only needs to invert a one-dimensional Laplace transform. For further details, we refer the interested reader to [2]. By gaining access to the relevant joint densities, this example

demonstrates that Lie symmetry methods allow us to design efficient Monte Carlo algorithms for challenging problems in mathematical finance.

4 Wishart Processes

Very tractable and highly relevant to finance are models that generalize the previously mentioned MMM. Along these lines, in this section we discuss Wishart processes with a view towards exact simulation. As demonstrated in [13], Wishart processes turn out to be the multidimensional extensions of squared Bessel processes. However, they also turn out to be affine, see [32], and [33]. Prior to the latter two contributions, the literature was focused on affine processes taking values in the Euclidean space, see e.g. [28], and [27]. Subsequently, matrix-valued affine processes were studied, see e.g. [22], and [34]. Since [32], and [33], it has been more widely known that Wishart processes are analytically tractable, since their characteristic function is available in closed form; see also [30]. In this section, we exploit this fact when we discuss exact simulation of Wishart processes.

Firstly, we fix notation and present an existence result. Wishart processes are S_d^+ or $\overline{S_d^+}$ valued, i.e. they assume values in the set of positive definite or positive semidefinite matrices, respectively. This makes them natural candidates for the modeling of covariance matrices, as noted in [32]. Starting with [32] and [33], there is now a substantial body of literature applying Wishart processes to problems in finance, see [14, 15, 23–26], and [31]. In the current paper we study Wishart processes in a pure diffusion setting. For completeness, we mention that matrix valued processes incorporating jumps have been studied, see e.g. in [5], and [42]. These processes are all contained in the affine framework introduced in [22], where we direct the reader interested in affine matrix valued processes.

In the following, we introduce the Wishart process as described in the work of Grasselli and collaborators; see [25] and [34]. For $\boldsymbol{x} \in \overline{S_d^+}$, we introduce the $\overline{S_d^+}$ valued Wishart process $\boldsymbol{X}^{\boldsymbol{x}} = \boldsymbol{X} = \{\boldsymbol{X}_t\,,\, t \geq 0\}$, which satisfies the SDE

$$d\boldsymbol{X}_t = \left(\alpha \boldsymbol{a}^\top \boldsymbol{a} + \boldsymbol{b}\boldsymbol{X}_t + \boldsymbol{X}_t \boldsymbol{b}^\top\right) dt + \left(\sqrt{\boldsymbol{X}_t}\, d\boldsymbol{W}_t \boldsymbol{a} + \boldsymbol{a}^\top d\boldsymbol{W}_t^\top \sqrt{\boldsymbol{X}_t}\right), \quad (24)$$

where $\alpha \geq 0$, $\boldsymbol{b} \in \mathscr{M}_d$, $\boldsymbol{a} \in \mathscr{M}_d$. Here $\mathscr{M}_d$ denotes the set of $d \times d$ matrices taking values in $\Re$. An obvious question to ask is whether Eq. (24) admits a solution, and, furthermore, if such a solution is unique and strong. For results on weak solutions we refer the reader to [22], and for results on strong solutions to [48]. We now present a summary of results, which in this form also appeared in [1]; see Theorem 1 in [1].

Theorem 3. *Assume that $\boldsymbol{x} \in \overline{S_d^+}$, and $\alpha \geq d - 1$, then Eq. (24) admits a unique weak solution. If $\boldsymbol{x} \in S_d^+$ and $\alpha \geq d + 1$, then this solution is strong.*

In this paper, we are interested in exact simulation schemes to be used in Monte Carlo methods. Hence weak solutions suffice for our purposes and we assume that

$\alpha > d-1$, so that the weak solution is unique. As in [1], we use $WIS_d(\boldsymbol{x}, \alpha, \boldsymbol{b}, \boldsymbol{a})$ to denote a Wishart process and $WIS_d(\boldsymbol{x}, \alpha, \boldsymbol{b}, \boldsymbol{a}; t)$ for the value of the process at the time point t.

We begin with the study of some special cases, which includes an extension of the MMM to the multidimensional case. We use $\boldsymbol{B}_t$ to denote an $n \times d$ Brownian motion and set

$$\boldsymbol{X}_t = \boldsymbol{B}_t^\top \boldsymbol{B}_t \,. \tag{25}$$

Then it can be shown that $\boldsymbol{X} = \{\boldsymbol{X}_t \,,\, t \geq 0\}$ satisfies the SDE

$$d\boldsymbol{X}_t = n\boldsymbol{I}_d dt + \sqrt{\boldsymbol{X}_t} d\boldsymbol{W}_t + d\boldsymbol{W}_t^\top \sqrt{\boldsymbol{X}_t} \,,$$

where $\boldsymbol{W}_t$ is a $d \times d$ Brownian motion, and $\boldsymbol{I}_d$ denotes the $d \times d$ identity matrix. This corresponds to the case where we set

$$\boldsymbol{a} = \boldsymbol{I}_d \,,\, \boldsymbol{b} = \boldsymbol{0} \,,\, \alpha = n \,.$$

We now provide the analogous scalar result, showing that Wishart processes generalize squared Bessel processes: Let $\delta \in \mathcal{N}$, and set

$$x = \sum_{k=1}^{\delta} (w^k)^2 \,,$$

where $w^k \in \Re$, $k \in \{1, \ldots, \delta\}$. Now we set

$$X_t = \sum_{k=1}^{\delta} (W_t^k + w^k)^2 \,. \tag{26}$$

Then X can be shown to satisfy the SDE

$$dX_t = \delta dt + 2\sqrt{X_t} dB_t \,,$$

where $B = \{B_t \,,\, t \geq 0\}$ is a scalar Brownian motion. This shows that (25) is the generalization of (26). Furthermore, it is also clear how to simulate (25).

Next, we illustrate how Wishart processes can be used to extend the MMM from Sect. 2. We recall some results pertaining to matrix-valued random variables, see e.g. [35], and [49]. We introduce some auxiliary notation. We denote by $\mathcal{M}_{m,n}(\Re)$ the set of all $m \times n$ matrices with entries in $\Re$. Next, we present a one-to-one relationship between vectors and matrices.

Definition 1. Let $\boldsymbol{A} \in \mathcal{M}_{m,n}(\Re)$ with columns $\boldsymbol{a}_i \in \Re^m$, $i = 1, \ldots, n$, and define the function $vec : \mathcal{M}_{m,n}(\Re) \to \Re^{mn}$ via

$$vec(\boldsymbol{A}) = \begin{pmatrix} \boldsymbol{a}_1 \\ \vdots \\ \boldsymbol{a}_n \end{pmatrix}.$$

We can now define the matrix variate normal distribution.

Definition 2. A $p \times n$ random matrix is said to have a matrix variate normal distribution with mean $\boldsymbol{M} \in \mathscr{M}_{p,n}(\Re)$ and covariance $\boldsymbol{\Sigma} \otimes \boldsymbol{\Psi}$, where $\boldsymbol{\Sigma} \in \mathscr{S}_p^+$, $\boldsymbol{\Psi} \in \mathscr{S}_n^+$, if $vec(\boldsymbol{X}^\top) \sim \mathscr{N}_{pn}(vec(\boldsymbol{M}^\top), \boldsymbol{\Sigma} \otimes \boldsymbol{\Psi})$, where $\mathscr{N}_{pn}$ denotes the multivariate normal distribution on $\Re^{pn}$ with mean $vec(\boldsymbol{M}^\top)$ and covariance $\boldsymbol{\Sigma} \otimes \boldsymbol{\Psi}$. We will use the notation $\boldsymbol{X} \sim \mathscr{N}_{p,n}(\boldsymbol{M}, \boldsymbol{\Sigma} \otimes \boldsymbol{\Psi})$.

Next, we introduce the Wishart distribution, which we link in the subsequent theorem to the normal distribution.

Definition 3. A $p \times p$-random matrix $\boldsymbol{X}$ in $\mathscr{S}_p^+$ is said to have a noncentral Wishart distribution with parameters $p \in \mathscr{N}$, $n \geq p$, $\boldsymbol{\Sigma} \in \mathscr{S}_p^+$ and $\boldsymbol{\Theta} \in \mathscr{M}_p(\Re)$, if its probability density function is of the form

$$f_X(\boldsymbol{S}) = \left(2^{\frac{1}{2}np}\Gamma_p(\frac{n}{2})det(\boldsymbol{\Sigma})^{\frac{n}{2}}\right)^{-1} etr\left(-\frac{1}{2}(\boldsymbol{\Theta} + \boldsymbol{\Sigma}^{-1}\boldsymbol{S})\right)$$
$$\times det(\boldsymbol{S})^{\frac{1}{2}(n-p-1)} \; {}_0F_1\left(\frac{n}{2}; \frac{1}{4}\boldsymbol{\Theta}\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\right)$$

where $\boldsymbol{S} \in \mathscr{S}_p^+$, *etr* denotes the exponential of the trace, and ${}_0F_1$ is the matrix-valued hypergeometric function, see [35], and [49] for a definition. We write

$$\boldsymbol{X} \sim \mathscr{W}_p(n, \boldsymbol{\Sigma}, \boldsymbol{\Theta}).$$

Before stating the next result, recall that scalar non-central chi-squared random variables of integer degrees of freedom, can be constructed via sums of normal random variables; see e.g. [38]. The following result presents the matrix variate analogy.

Theorem 4. *Let* $\boldsymbol{X} \sim \mathscr{N}_{p,n}(\boldsymbol{M}, \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n)$, $n \in \{p, p+1, \dots\}$. *Then*

$$\boldsymbol{X}\boldsymbol{X}^\top \sim \mathscr{W}_p(n, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}^{-1}\boldsymbol{M}\boldsymbol{M}^\top).$$

Using Theorem 4 we immediately have access to the probability density functions of Wishart processes which are generated as products of matrix valued Brownian motions, see Eq. (25) for an example. This close link between the Wishart distribution and the Wishart process is employed in the next section.

5 Bivariate MMM

Theorem 4 is now employed to extend the MMM to a bivariate case. We consider exchange rate options, and follow the ideas from [36]. The GOP denominated in units of the domestic currency is denoted by S^a, and the GOP denominated in the foreign currency by S^b. An exchange rate at time t can be expressed in terms of a ratio of two GOP denominations. Then one would pay at time t, $\frac{S_t^a}{S_t^b}$ units of currency a to obtain one unit of the foreign currency b. As the domestic currency is indexed by a, the price of, say, a call option with maturity T on the exchange rate can be expressed via the real world pricing formula (5) as:

$$S_0^a E\left(\frac{\left(\frac{S_T^a}{S_T^b} - K\right)^+}{S_T^a}\right). \tag{27}$$

We now discuss a bivariate extension of the MMM from Sect. 2, which is still tractable, as we can employ the non-central Wishart distribution to compute (27). For $k \in \{a, b\}$, we set

$$S_t^k = S_t^{0,k} \bar{S}_t^k ,$$

where $S_t^{0,k} = \exp\{r_k t\}$, $S_0^{0,k} = 1$, so $S^{0,k}$ denotes the savings account in currency k, which for simplicity is assumed to be a deterministic exponential function of time. As for the stylized MMM, we model the discounted GOP, $\bar{S}_t^k$, denominated in units of the kth savings account, $S_t^{0,k}$, as a time-changed squared Bessel process of dimension four. We introduce the 2×4 matrix process $\boldsymbol{X} = \{\boldsymbol{X}_t , t \geq 0\}$ via

$$\boldsymbol{X}_t = \begin{bmatrix} \left(W^{1,1}_{\varphi^1(t)} + w^{1,1}\right) & \left(W^{2,1}_{\varphi^1(t)} + w^{2,1}\right) & \left(W^{3,1}_{\varphi^1(t)} + w^{3,1}\right) & \left(W^{4,1}_{\varphi^1(t)} + w^{4,1}\right) \\ \left(W^{1,2}_{\varphi^2(t)} + w^{1,2}\right) & \left(W^{2,2}_{\varphi^2(t)} + w^{2,2}\right) & \left(W^{3,2}_{\varphi^2(t)} + w^{3,2}\right) & \left(W^{4,2}_{\varphi^2(t)} + w^{4,2}\right) \end{bmatrix},$$

where $w^{1,1}, \ldots, w^{4,2}$ can be interpreted as initial parameters. The processes $W^{i,1}_{\varphi^1}$, $i = 1, \ldots, 4$, denote independent Brownian motions, subject to the deterministic time-change

$$\varphi^1(t) = \frac{\alpha_0^1}{4\eta^1}\left(\exp\{\eta^1 t\} - 1\right) = \frac{1}{4}\int_0^t \alpha_s^1 ds ,$$

c.f. Sect. 2. Similarly, also $W^{i,2}_{\varphi^2}$, $i = 1, \ldots, 4$, denote independent Brownian motions, subject to the deterministic time change

$$\varphi^2(t) = \frac{\alpha_0^2}{4\eta^2}\left(\exp\{\eta^2 t\} - 1\right) = \frac{1}{4}\int_0^t \alpha_s^2 ds\,.$$

Now, consider the process $\boldsymbol{Y} = \{\boldsymbol{Y}_t\,,\, t \geq 0\}$, which assumes values in S_2^+, and is given by

$$\boldsymbol{Y}_t := \boldsymbol{X}_t \boldsymbol{X}_t^\top\,,\, t \geq 0\,,$$

which yields

$$\boldsymbol{Y}_t = \begin{bmatrix} \sum_{i=1}^4 \left(W^{i,1}_{\varphi^1(t)} + w^{i,1}\right)^2 & \sum_{i=1}^4 \sum_{j=1}^2 \left(W^{i,j}_{\varphi^j(t)} + w^{i,j}\right) \\ \sum_{i=1}^4 \sum_{j=1}^2 \left(W^{i,j}_{\varphi^j(t)} + w^{i,j}\right) & \sum_{i=1}^4 \left(W^{i,2}_{\varphi^2(t)} + w^{i,2}\right)^2 \end{bmatrix}.$$

We set

$$\bar{S}_t^a = Y_t^{1,1} \quad \text{and} \quad \bar{S}_t^b = Y_t^{2,2}\,,$$

so we use the diagonal elements of $\boldsymbol{Y}_t$ to model the GOP in different currency denominations. Next, we introduce the following dependence structure: The Brownian motions $W^{i,1}$ and $W^{i,2}$, $i = 1, \ldots, 4$, covary as follows:

$$\langle W^{i,1}_{\varphi^1(\cdot)}, W^{i,2}_{\varphi^2(\cdot)} \rangle_t = \frac{\varrho}{4}\int_0^t \sqrt{\alpha_s^1 \alpha_0^2} ds, i = 1, \ldots, 4\,, \tag{28}$$

where $-1 < \varrho < 1$. The specification (28) allows us to employ the non-central Wishart distribution. We work through this example in detail, as it illustrates how to extend the stylized MMM to allow for a non-trivial dependence structure, but still exploit the tractability of the Wishart distribution. We recall that $vec(\boldsymbol{X}_T^\top)$ stacks the two columns of $\boldsymbol{X}_T^\top$, hence

$$vec(\boldsymbol{X}_T^\top) = \begin{bmatrix} \left(W^{1,1}_{\varphi^1(T)} + w^{1,1}\right) \\ \vdots \\ \left(W^{4,1}_{\varphi^1(T)} + w^{4,1}\right) \\ \left(W^{1,2}_{\varphi^2(T)} + w^{1,2}\right) \\ \vdots \\ \left(W^{4,2}_{\varphi^2(T)} + w^{4,2}\right) \end{bmatrix}.$$

It is easily seen that the mean matrix $\boldsymbol{M}$ of $vec(X_T^\top)$ satisfies

$$vec\left(\boldsymbol{M}^\top\right) = \begin{bmatrix} w^{1,1} \\ \vdots \\ w^{4,1} \\ w^{1,2} \\ \vdots \\ w^{4,2} \end{bmatrix} \tag{29}$$

and the covariance matrix of $vec(\boldsymbol{X}_T^\top)$ is given by

$$\boldsymbol{\Sigma} \otimes \boldsymbol{I}_4 = \begin{bmatrix} \Sigma^{1,1}\boldsymbol{I}_4 & \Sigma^{1,2}\boldsymbol{I}_4 \\ \Sigma^{2,1}\boldsymbol{I}_4 & \Sigma^{2,2}\boldsymbol{I}_4 \end{bmatrix}, \tag{30}$$

where $\boldsymbol{\Sigma}$ is a 2×2 matrix with $\Sigma^{1,1} = \varphi^1(T)$, $\Sigma^{2,2} = \varphi^2(T)$, and

$$\Sigma^{1,2} = \Sigma^{2,1} = \frac{\varrho}{4} \int_0^t \sqrt{\alpha_s^1 \alpha_s^2} ds\,.$$

We remark that assuming $-1 < \varrho < 1$ results in $\boldsymbol{\Sigma}$ being positive definite. It now immediately follows from Theorem 4 that

$$\boldsymbol{X}_T \boldsymbol{X}_T^\top \sim W_2\left(4, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}^{-1}\boldsymbol{M}\boldsymbol{M}^\top\right),$$

where $\boldsymbol{M}$ and $\boldsymbol{\Sigma}$ are given in Eqs. (29) and (30), respectively. Recall that we set

$$\begin{aligned} \boldsymbol{Y}_t &= \boldsymbol{X}_t \boldsymbol{X}_t^\top, \\ \bar{S}_t^a &= Y_t^{1,1}, \\ \bar{S}_t^b &= Y_t^{2,2}, \end{aligned}$$

hence we can compute (27) using

$$E\left(f(\boldsymbol{Y}_T)\right),$$

where $f : S_2^+ \to \Re$ is given by

$$f(y) = \frac{\left(\frac{\exp\{r_1 T\} y^{1,1}}{\exp\{r_2 T\} y^{2,2}} - K\right)^+}{\exp\{r_1 T\} y^{1,1}},$$

for $y \in S_2^+$, and $y^{i,i}$, $i = 1, 2$, are the diagonal elements of y, and the probability density function of $\boldsymbol{Y}_T$ is given in Definition 3.

We now discuss further exact simulation schemes for Wishart processes, where we rely on [1] and [6]. For integer valued parameters α in (24), we have the following exact simulation scheme, which generalizes a well-known result from the scalar case, linking Ornstein-Uhlenbeck and square-root processes. In particular, this lemma shows that, in principle, certain square-root processes can be simulated using Ornstein-Uhlenbeck processes.

Lemma 2. *Let $A > 0$, $Q > 0$, and define the SDEs*

$$dX_t^i = -AX_t^i dt + QdW_t^i ,$$

for $i = 1, \ldots, \beta$, where $\beta \in \mathcal{N}$, $W^1, W^2, \ldots, W^\beta$ are independent Brownian motions. Then

$$Z_t = \sum_{i=1}^{\beta} (X_t^i)^2$$

is a square-root process of dimension β, whose dynamics are characterised by an SDE

$$dZ_t = (\beta Q^2 - 2AZ_t)dt + 2Q\sqrt{Z_t}dB_t ,$$

where B is a resulting Brownian motion.

Proof. The proof follows immediately from the Itô-formula. □

This result is easily extended to the Wishart case, for integer valued α, see Sect. 1.2.2 in [6]. We define

$$\boldsymbol{V}_t = \sum_{k=1}^{\beta} \boldsymbol{X}_{k,t} \boldsymbol{X}_{k,t}^\top , \tag{31}$$

where

$$d\boldsymbol{X}_{k,t} = A\boldsymbol{X}_{k,t}dt + \boldsymbol{Q}^\top d\boldsymbol{W}_{k,t} , k = 1, \ldots, \beta , \tag{32}$$

where $A \in \mathscr{M}_d$, $\boldsymbol{X}_t \in \Re^d$, $\boldsymbol{Q} \in \mathscr{M}_d$, $\boldsymbol{W}_k \in \Re^d$, so that $\boldsymbol{V}_t \in \mathscr{M}_d$. The following lemma gives the dynamics of $\boldsymbol{V} = \{\boldsymbol{V}_t , t \geq 0\}$.

Lemma 3. *Assume that $\boldsymbol{V}_t$ is given by Eq. (31), where $\boldsymbol{X}_t$ satisfies Eq. (32). Then*

$$d\boldsymbol{V}_t = \left(\beta \boldsymbol{Q}^\top \boldsymbol{Q} + A\boldsymbol{V}_t + \boldsymbol{V}_t A^\top\right) dt + \sqrt{\boldsymbol{V}_t} d\boldsymbol{W}_t \boldsymbol{Q} + \boldsymbol{Q}^\top d\boldsymbol{W}_t^\top \sqrt{\boldsymbol{V}_t} ,$$

where $\boldsymbol{W} = \{\boldsymbol{W}_t , t \geq 0\}$ is a $d \times d$ matrix valued Brownian motion that is determined by

$$\sqrt{V_t}\, d\boldsymbol{W}_t = \sum_{k=1}^{\beta} \boldsymbol{X}_{k,t}\, d\boldsymbol{W}_{t,k}^{\top}.$$

Finally, we remind the reader that vector-valued Ornstein-Uhlenbeck processes can be simulated exactly, see e.g. Chap. 2 in [52].

For the general case, we refer the reader to [1]. In that paper, a remarkable splitting property of the infinitesimal generator of the Wishart process was employed to come up with an exact simulation scheme for Wishart processes without any restriction on the parameters. Furthermore, in [1] higher-order discretization schemes for Wishart processes and second-order schemes for general affine diffusions on positive semidefinite matrices were presented. These results emphasize that Wishart processes are suitable candidates for financial models from a computational perspective, since exact simulation schemes are readily available. They are also well suited from the perspective of the benchmark approach to finance, since they go in a natural way beyond the classical risk neutral modeling.

6 Conclusion

In this paper, we discussed, with a view towards financial modeling under the benchmark approach, classes of stochastic processes for which exact simulation schemes are available. In the one-dimensional case, our first theorem gives access to explicit transition densities via Lie symmetry group results. In the multidimensional case the probability law of Wishart processes is described explicitly. When considering applications in finance, one needs a framework that can accommodate these processes as asset prices, in particular, when they generate strict local martingales. We demonstrated that the benchmark approach is a suitable framework for these processes and allows us to systematically exploit the tractability of the models described.

References

1. Ahdida, A., Alfonsi, A.: Exact and higher order discretization schemes for Wishart processes and their affine extensions. Ann. Appl. Probab. **23**, 1025–1073 (2013)
2. Baldeaux, J., Chan, L., Platen, E.: Quasi-Monte Carlo methods for derivatives on realized variance of an index under the benchmark approach. ANZIAM J. **52**, 727–741 (2011)
3. Baldeaux, J., Chan, L., Platen, E.: Derivatives on realized variance and volatility of an index under the benchmark approach. University of Technology, Sydney (2012)
4. Baldeaux, J., Ignatieva, K., Platen, E.: A tractable model for indices approximating the growth optimal portfolio. Stud. Nonlinear Dyn. Control (2013, to appear)
5. Barndorff-Nielsen, O., Stelzer, R.: Positive-definite matrix processes of finite variation. Probab. Math. Statist. **27**, 3–43 (2007)

6. Benabid, A., Bensusan, H., El Karoui, N.: Wishart stochastic volatility: asymptotic smile and numerical framework. Working paper, Ecole Polytechnique, Paris (2010)
7. Beskos, A., Papaspiliopoulos, O., Roberts, G.: Retrospective exact simulation of diffusion sample paths with applications. Bernoulli **12**, 1077–1098 (2006)
8. Beskos, A., Papaspiliopoulos, O., Roberts, G.: A factorisation of diffusion measure and finite sample path constructions. Methodol. Comput. Appl. Probab. **10**, 85–104 (2008)
9. Beskos, A., Roberts, G.: Exact simulation of diffusions. Ann. Appl. Probab. **15**, 2422–2444 (2008)
10. Bluman, G., Kumai, S.: Symmetry and Differential Equations. Springer, New York (1989)
11. Boyle, P.P.: Options: a Monte Carlo approach. J. Financ. Econ. **4**, 323–338 (1977)
12. Breiman, L.: Investment policies for expanding business optimal in a long run sense. Naval Res. Logist. Q. **7**, 647–651 (1960)
13. Bru, M.F.: Wishart processes. J. Theoret. Probab. **4**, 725–743 (1991)
14. Burasci, B., Cieslak, A., Trojani, F.: Correlation risk and the term structure of interest rates. University of Lugano (2006)
15. Burasci, B., Porchia, P., Trojani, F.: Correlation risk and optimal portfolio choice. J. Finance **65**, 393–420 (2010)
16. Chen, N.: Exact simulation of stochastic differential equations. Chinese University of Hong Kong (working paper)
17. Craddock, M.: Fundamental solutions, transition densities and the integration of Lie symmetries. J. Differ. Equ. **246**, 2538–2560 (2009)
18. Craddock, M., Dooley, A.H.: On the equivalence of Lie symmetries and group representations. J. Differ. Equ. **249**, 621–653 (2010)
19. Craddock, M., Lennox, K.: Lie group symmetries as integral transforms of fundamental solutions. J. Differ. Equ. **232**, 652–674 (2007)
20. Craddock, M., Lennox, K.: The calculation of expectations for classes of diffusion processes by Lie symmetry methods. Ann. Appl. Probab. **19**, 127–157 (2009)
21. Craddock, M., Platen, E.: Symmetry group methods for fundamental solutions. J. Differ. Equ. **207**, 285–302 (2004)
22. Cuchiero, C., Filipović, D., Mayerhofer, E., Teichmann, J.: Affine processes on positive semidefinite matrices. Ann. Appl. Probab. **21**, 397–463 (2011)
23. Da Fonseca, J., Grasselli, M., Ielpo, F.: Estimationg the Wishart affine stochastic correlation model using the empirical characteristic function. Working paper, University Padova (2008)
24. Da Fonseca, J., Grasselli, M., Ielpo, F.: Hedging (co)variance risk with variance swaps. Int. J. Theor. Appl. Finance **14**, 899, (2011)
25. Da Fonseca, J., Grasselli, M., Tebaldi, C.: Option pricing when correlations are stochastic: an analytical framework. Review of Derivatives Research **10**, 151–180 (2007)
26. Da Fonseca, J., Grasselli, M., Tebaldi, C.: A multifactor volatility Heston model. Quant. Finance **8**, 591–604 (2008)
27. Duffie, D., Filipović, D., Schachermayer, W.: Affine processes and applicatiosn in finance. Ann. Appl. Probab. **13**, 984–1053 (2003)
28. Duffie, D., Pan, J., Singleton, K.: Transform analysis and asset pricing for affine jump-diffusion. Econometrica **68**, 1343–1376 (2000)
29. Glasserman, P.: Monte Carlo Methods in Financial Engineering. Springer, New York (2004)
30. Gnoatto, A., Grasselli, M.: The explicit Laplace transform for the Wishart process. University Padova (2011, submitted)
31. Gouriéroux, C., Montfort, A., Sufana, R.: International money and stock market contingent claims. Working paper, CREST (2007)
32. Gourieroux, C., Sufana, R.: Wishart quadratic term structure models, CREF 03-10, HEC Montreal (2003)
33. Gourieroux, C., Sufana, R.: Derivative pricing with multivariate stochastic volatility: application to credit risk. Working paper, CREST (2004)
34. Grasselli, M., Tebaldi, C.: Solvable affine term structure models. Math. Finance **18**, 135–153 (2008)

35. Gupta, A.K., Nagar, D.K.: Matrix Valued Stochastic Processes. Chapman & Hall/CRC. Boca Raton, FL (2000)
36. Heath, D., Platen, E.: Currency derivatives under a minimal market model with random scaling. Int. J. Theor. Appl. Finance **8**, 1157–1177 (2005)
37. Itkin, A.: New solvable stochastic volatility models for pricing volatility derivatives. Review of Derivatives Research **16**, 111–134 (2013)
38. Johnson, N.L., Kotz, S., Balakrishnan, N.: Continuous Univariate Distributions. Wiley Series in Probability and Mathematical Statistics, vol. 2, 2nd edn. Wiley, New York (1995)
39. Kelly, J.R.: A new interpretation of information rate. Bell Syst. Tech. J. **35**, 917–926 (1956)
40. Kloeden, P., Platen, E.: Numerical Solution of Stochastic Differential Equations, 3rd edn. Springer, Berlin (1999)
41. Latané, H.: Criteria for choice among risky ventures. J. Political Economy **38**, 145–155 (1959)
42. Leippold, M., Trojani, F.: Asset pricing with matrix affine jump diffusions. Working paper, University of Lugano (2008)
43. Lennox, K.: Lie symmetry methods for multidimensional linear parabolic pdes and diffusions. PhD thesis, University of Technology, Sydney (2011)
44. Loewenstein, M., Willard, G.A.: Local martingales, arbitrage, and viability: free snacks and cheap thrills. Econ. Theory **16**, 135–161 (2000)
45. Long, J.B.: The numeraire portfolio. J. Financ. Econ. **26**, 29–69 (1990)
46. MacLean, L.C., Thorp, E., Ziemba, W.: The Kelly Capital Growth Investment Criterion. World Scientific, Singapore/Hackensack (2011)
47. Markowitz, H.: Investment for the long run: new evidence for an old rule. J. Finance **XXXI**, 1273–1286 (1976)
48. Mayerhofer, E., Pfaffel, O., Stelzer, R.: On strong solutions for positive definite jump diffusions. Technical report, University of Munich (2011)
49. Muirhead, R.J.: Aspects of Multivariate Statistical Theory. Wiley, New York (1982)
50. Musiela, M., Rutkowski, M.: Martingale Methods in Financial Modelling, 2nd edn. Springer, Berlin/New York (2005)
51. Olver, P.J.: Applications of Lie Groups to Differential Equations. Graduate Texts in Mathematics. Springer, New York (1993)
52. Platen, E., Bruti-Liberati, N.: Numerical Solution of Stochastic Differntial Equations with Jumps in Finance. Springer, Berlin/Heidelberg (2010)
53. Platen, E., Heath, D.: A Benchmark Approach to Quantitative Finance, 2nd edn. Springer, Berlin (2010)
54. Revuz, D., Yor, M.: Continuous Martingales and Brownian Motion, 3rd edn. Springer, Berlin/New York (1999)
55. Thorp, E.O.: A favourable strategy for twenty-one. Proc. Nat. Acad. Sci. **47**, 110–112 (1961)

The Supremum Norm of the Discrepancy Function: Recent Results and Connections

Dmitriy Bilyk and Michael Lacey

Abstract A great challenge in the analysis of the discrepancy function D_N is to obtain universal lower bounds on the L^∞ norm of D_N in dimensions $d \geq 3$. It follows from the L^2 bound of Klaus Roth that $\|D_N\|_\infty \geq \|D_N\|_2 \gtrsim (\log N)^{(d-1)/2}$. It is conjectured that the L^∞ bound is significantly larger, but the only definitive result is that of Wolfgang Schmidt in dimension $d = 2$. Partial improvements of the Roth exponent $(d-1)/2$ in higher dimensions have been established by the authors and Armen Vagharshakyan. We survey these results, the underlying methods, and some of their connections to other subjects in probability, approximation theory, and analysis.

1 Introduction

We survey recent results on the sup-norm of the discrepancy function. For integers $d \geq 2$, and $N \geq 1$, let $\mathscr{P}_N \subset [0,1]^d$ be a finite point set with cardinality $\sharp \mathscr{P}_N = N$. Define the associated discrepancy function by

$$D_N(x) = \sharp(\mathscr{P}_N \cap [0,x)) - N|[0,x)|, \tag{1}$$

where $x = (x_1, \dots, x_d)$ and $[0,x) = \prod_{j=1}^d [0,x_j)$ is a rectangle with antipodal corners at 0 and x, and $|\cdot|$ stands for the d-dimensional Lebesgue measure. The dependence upon the selection of points $\mathscr{P}_N$ is suppressed, as we are mostly

D. Bilyk (✉)
School of Mathematics, University of Minnesota, Minneapolis, MN 55455, USA

Mathematics, University of South Carolina, Columbia, SC, USA
e-mail: dbilyk@math.umn.edu

M. Lacey
School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332, USA
e-mail: lacey@math.gatech.edu

J. Dick et al. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2012*,
Springer Proceedings in Mathematics & Statistics 65,
DOI 10.1007/978-3-642-41095-6_2, © Springer-Verlag Berlin Heidelberg 2013

interested in bounds that are universal in $\mathscr{P}_N$. The discrepancy function D_N measures equidistribution of $\mathscr{P}_N$: a set of points is *well-distributed* if this function is small in some appropriate function space.

It is a basic fact of the theory of irregularities of distribution that relevant norms of this function in dimensions 2 and higher must tend to infinity as N grows. The classic results are due to Roth [21] in the case of the L^2 norm and Schmidt [23] for L^p, $1 < p < 2$.

Theorem 1. *For $1 < p < \infty$ and any collection of points $\mathscr{P}_N \subset [0,1]^d$, we have*

$$\|D_N\|_p \gtrsim (\log N)^{(d-1)/2}. \tag{2}$$

Moreover, we have the endpoint estimate

$$\|D_N\|_{L(\log L)^{(d-1)/2}} \gtrsim (\log N)^{(d-1)/2}. \tag{3}$$

The symbol "$\gtrsim$" in this paper stands for "greater than a constant multiple of", and the implied constant may depend on the dimension, the function space, but *not* on the configuration $\mathscr{P}_N$ or the number of points N. The Orlicz space notation, such as $L(\log L)^\beta$, is explained in the next section, see (10).

We should mention that there exist sets $\mathscr{P}_N$ that meet the L^p bounds (2) in all dimensions. This remarkable fact is established by beautiful and quite non-trivial constructions of the point distributions $\mathscr{P}_N$. We refer to the reader to one of the very good references [2,11,12] on the subject for more information on this important complement to the subject of this note.

While the previous theorem is quite adequate for L^p, $1 < p < \infty$, the endpoint cases of L^∞ and L^1 are not amenable to the same techniques. Indeed, the extremal L^∞ bound should be larger than the average L^2 norm. In dimension $d = 2$ the endpoint estimates are known – it is the theorem of Schmidt [22].

Theorem 2. *The following estimate is valid for all collections $\mathscr{P}_N \subset [0,1]^2$:*

$$\|D_N\|_\infty \gtrsim \log N. \tag{4}$$

This is larger than Roth's L^2 bound by $\sqrt{\log N}$. The difference between the two estimates points to the fact that for extremal choices of sets $\mathscr{P}_N$, the L^∞ norm of D_N is obtained on a set so small it cannot be seen on the scale of L^p spaces. We will return to this point below.

In dimensions 3 and higher partial results began with a breakthrough work of J. Beck [1] in dimension $d = 3$. The following result is due to Bilyk and Lacey [5] in dimension $d = 3$, and Bilyk, Lacey, Vagharshakyan [8] in dimensions $d \geq 4$.

Theorem 3. *In dimensions $d \geq 3$ there exists $\eta = \eta(d) \geq c/d^2$ for which the following estimate holds for all collections $\mathscr{P}_N \subset [0,1]^d$:*

$$\|D_N\|_\infty \gtrsim (\log N)^{\frac{d-1}{2}+\eta}. \tag{5}$$

This is larger than Roth's bound by $(\log N)^{\eta}$. Beck's original result in dimension $d = 3$ had a much smaller doubly logarithmic term $(\log\log N)^{\frac{1}{8}-\varepsilon}$ in place of $(\log N)^{\eta}$. The proof strategy begins with the fundamental orthogonal function method of Roth and Schmidt, which we turn to in the next section. In Sect. 3 we turn to a closely related combinatorial inequality for "hyperbolic" sums of multiparameter Haar functions. It serves as the core question which has related the progress on lower bounds for the discrepancy function to questions in probability and approximation theory. Based upon this inequality, it is natural to conjecture that the optimal form of the L^{∞} estimate is

Conjecture 1. In dimensions $d \geq 3$ there holds $\|D_N\|_{\infty} \gtrsim (\log N)^{d/2}$.

We should mention that at the present time there is no consensus among the experts about the sharp form of the conjecture (in fact, a great number of specialist believes that $\|D_N\|_{\infty} \gtrsim (\log N)^{d-1}$ is the optimal bound, which is supported by the best known examples). However, in this paper we shall advocate our belief in Conjecture 1 by comparing it to other sharp conjectures in various fields of mathematics. In particular, the sharpness of Conjecture 2 in Sect. 3 suggests that the estimate above is the best that could be obtained by the orthogonal function techniques.

The reader can consult the papers [5, 8], as well as the surveys of the first author [3, 4] for more detailed information.

2 The Orthogonal Function Method

All progress on these universal lower bounds has been based upon the orthogonal function method, initiated by Roth, with the modifications of Schmidt, as presented here. Denote the family of all dyadic intervals $I \subset [0,1]$ by $\mathscr{D}$. Each dyadic interval I is the union of two dyadic intervals I_- and I_+, each of exactly half the length of I, representing the left and right halves of I respectively. Define the Haar function associated to I by $h_I = -\chi_{I_-} + \chi_{I_+}$. Here and throughout we will use the L^{∞} (rather than L^2) normalization of the Haar functions.

In dimension d, the d-fold product $\mathscr{D}^d$ is the collection of dyadic intervals in $[0,1]^d$. Given $R = R_1 \times \cdots \times R_d \in \mathscr{D}^d$, the Haar function associated with R is the tensor product

$$h_R(x_1, \ldots, x_d) = \prod_{j=1}^{d} h_{R_j}(x_j)\,.$$

These functions are pairwise orthogonal as $R \in \mathscr{D}^d$ varies.

For a d-dimensional vector $r = (r_1, \ldots, r_d)$ with non-negative integer coordinates let $\mathscr{D}_r$ be the set of those $R \in \mathscr{D}^d$ that for each coordinate $1 \leq j \leq d$, we have $|R_j| = 2^{-r_j}$. These rectangles partition $[0,1]^d$. We call f_r an r-function

(a generalized Rademacher function) if for some choice of signs $\{\varepsilon_R : R \in \mathscr{D}_r\}$, we have

$$f_r(x) = \sum_{R\in\mathscr{D}_r} \varepsilon_R h_R(x)\,.$$

The following is the crucial lemma of the method. Given an integer N, we set $n = \lceil 1+\log_2 N\rceil$, where $\lceil x\rceil$ denotes the smallest integer greater than or equal to x.

Lemma 1. *In all dimensions $d \geq 2$ there is a constant $c_d > 0$ such that for each r with $|r| := \sum_{j=1}^d r_j = n$, there is an r-function f_r with $\langle D_N, f_r\rangle \geq c_d$. Moreover, for all r-functions there holds $|\langle D_N, f_r\rangle| \lesssim N2^{-|r|}$.*

The proof of the lemma is straightforward, see e.g. [3, 21, 23]. With this lemma at hand, the proof of Roth's Theorem in L^2 is as follows. Note that the requirement that $|r| = n$ says that the coordinates of r must partition n into d parts. It follows that the number of ways to select the coordinates of r is bounded above and below by a multiple of n^{d-1}, agreeing with the simple logic that there are $d-1$ "free" parameters: d dimensions minus the restriction $|r| = n$. Set $F_d = \sum_{r\,:\,|r|=n} f_r$. Orthogonality implies that $\|F_d\|_2 \lesssim n^{(d-1)/2}$. Hence, by Cauchy–Schwarz

$$n^{d-1} \lesssim \sum_{r\,:\,|r|=n} \langle D_N, f_r\rangle = \langle D_N, F_d\rangle \tag{6}$$

$$\leq \|D_N\|_2 \cdot \|F_d\|_2 \leq \|D_N\|_2 \cdot n^{(d-1)/2}\,. \tag{7}$$

The universal lower bound $n^{(d-1)/2} \lesssim \|D_N\|_2$ follows.

Deeper properties of the discrepancy function may be deduced from finer properties of r-functions. A key property is the classical Littlewood–Paley inequality for Haar functions:

Theorem 4. *For $p \geq 2$, we have the inequality*

$$\Big\|\sum_{I\in\mathscr{D}} \alpha_I h_I\Big\|_p \leq C\sqrt{p}\,\Big\|\Big[\sum_{I\in\mathscr{D}} |\alpha_I|^2\chi_I\Big]^{1/2}\Big\|_p\,, \tag{8}$$

where C is an absolute constant, and the coefficients α_I take values in a Hilbert space **H**.

The right-hand side is the Littlewood–Paley (martingale) square function of the left hand side. This inequality can be viewed as an extension of orthogonality and Parseval's identity to values of p other than 2, and it is often useful to keep track of the growth of L^p norms. The fact that one can allow Hilbert space value coefficients permits repeated application of the inequality. The role of the Hilbert space valued coefficients is the focus of [14], which includes more information about multiparameter harmonic analysis, relevant to this subject.

Consider the dual function in (7), $F_d = \sum_{r\,:\,|r|=n} f_r$. As discussed earlier, the index set $\{r : |r| = n\}$ has $d-1$ free parameters. The function F_d is a Haar series

in the first variable, so the inequality (8) applies. On the right-hand side, the square function can be viewed as an ℓ^2-valued Haar series in the second variable, hence (8) applies again, see [3, 8] for details. Continuing this $d-1$ times, one arrives at

$$\|F_d\|_p \lesssim p^{(d-1)/2} n^{(d-1)/2}, \qquad 2 \le p < \infty. \tag{9}$$

Repeating (7) verbatim (with Hölder replacing Cauchy–Schwarz), one obtains $n^{(d-1)/2} \lesssim \|D_N\|_q$ for $1 < q < 2$.

If one is interested in endpoint estimates, it is useful to rephrase the inequalities for F_d above in the language of Orlicz spaces. For a convex increasing function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ with $\psi(0) = 0$, the Orlicz space L^ψ is defined as the space of measurable functions $f : [0,1]^d \to \mathbb{R}$ for which

$$\|f\|_{L^\psi} = \inf\left\{ K > 0 : \int_{[0,1]^d} \psi\big(|f(x)|/K\big)\, dx \le 1 \right\}. \tag{10}$$

In particular, for $\psi(t) = t^p$ one obtains the standard L^p spaces, while $\exp(L^\alpha)$ and $L(\log L)^\beta$ denote Orlicz spaces generated by functions equal to e^{t^α} and $t \log^\beta t$ respectively, when t is large enough. These spaces serve as refinements of the endpoints of the L^p scale, as for each $1 < p < \infty$, $\alpha, \beta > 0$ we have the embeddings $L^\infty \subset \exp(L^\alpha) \subset L^p$ and $L^p \subset L(\log L)^\beta \subset L^1$.

The polynomial growth in the L^p norms of F_d (9) translates into exponential integrability estimates, namely $\|F_d\|_{\exp(L^{2/(d-1)})} \lesssim n^{(d-1)/2}$, since

$$\|f\|_{\exp(L^\alpha)} \simeq \sup_{p>1} p^{-1/\alpha} \|f\|_p, \qquad \alpha > 0.$$

The dual space to $\exp(L^{2/(d-1)})$ is $L(\log L)^{(d-1)/2}$, hence we see that

$$n^{(d-1)/2} \lesssim \|D_N\|_{L(\log L)^{(d-1)/2}}.$$

A well-known result of Halász [15] is a '$\sqrt{\log L}$' improvement of this estimate in dimension $d = 2$. Indeed, we have the following theorem valid for all dimensions, see [18].

Theorem 5. *For dimensions $d \ge 2$, there holds $\|D_N\|_{L(\log L)^{(d-2)/2}} \gtrsim (\log N)^{(d-1)/2}$.*

Notice that for $d = 2$ one recovers Halász's L^1 bound

$$\sqrt{\log N} \lesssim \|D_N\|_1. \tag{11}$$

In dimension $d = 2$, the argument of Halász can be rephrased into the estimate

$$\sqrt{n} \lesssim \langle D_N, \sin\big(\tfrac{c}{\sqrt{n}} F_2\big)\rangle, \qquad 0 < c < 1 \text{ sufficiently small.} \tag{12}$$

This immediately shows that $\|D_N\|_1 \gtrsim \sqrt{n}$ in dimension $d = 2$. There is a relevant endpoint estimate of the Littlewood–Paley inequalities, namely the Chang–Wilson–Wolff inequality [10]. Employing extensions of this inequality and the estimate above, one can give a proof of Theorem 5 in dimensions $d \geq 3$.

It is a well-known conjecture that in all dimensions $d \geq 3$ one has the estimate

$$\|D_N\|_1 \gtrsim (\log N)^{(d-1)/2} \tag{13}$$

on the L^1 norm of the discrepancy function. Any improvement of Theorem 5 would yield progress on this conjecture.

3 The Small Ball Inequality

Lower bounds on the discrepancy function are related through proof techniques to subjects in different areas of mathematics. They include, in particular, the so-called *small deviation inequalities* for the Brownian sheet in probability theory, complexity bounds for certain Sobolev spaces in approximation theory, and a combinatorial inequality involving multivariate Haar functions in the unit cube. We refer the reader to the references [3–5, 8] for more information, and emphasize that the questions in probability and approximation theory are parts of very broad areas of investigation with additional points of contact with discrepancy theory and many variations of the underlying themes.

According to the idea introduced in the previous section, the behavior of D_N is essentially defined by its projection onto the span of $\{h_R : R \in \mathscr{D}^d, |R| = 2^{-n}\}$. It is therefore reasonable to model the discrepancy bounds by estimates of the linear combinations of such Haar functions (we call such sums "hyperbolic"). The problem of obtaining lower bounds for sums of Haar functions supported by rectangles of fixed volume – known as the *Small Ball inequality* – arises naturally in the aforementioned problems in probability and approximation theory. While in the latter fields versions of this inequality have important formal implications, its connection to discrepancy estimates is still only intuitive and is not fully understood. However, most known proof methods are easily transferred from one problem to another. The conjectured form of the inequality is the following.

Conjecture 2 (The Small Ball Conjecture). For dimension $d \geq 3$ we have the inequality

$$2^{-n} \sum_{|R|=2^{-n}} |\alpha_R| \lesssim n^{(d-2)/2} \Big\| \sum_{R\in\mathscr{D}^d : |R|=2^{-n}} \alpha_R h_R \Big\|_\infty \tag{14}$$

valid for all real-valued coefficients α_R.

The subject of the conjecture is the exact exponent of n the right-hand side. This conjecture is better, by one square root of n, than a trivial estimate available from

the Cauchy–Schwartz inequality. Indeed, with $n^{(d-2)/2}$ replaced by $n^{(d-1)/2}$ it holds for the L^2 norm:

$$\Big\| \sum_{R\in\mathscr{D}^d:\,|R|=2^{-n}} \alpha_R h_R \Big\|_2 = \Big(\sum_{|R|=2^{-n}} |\alpha_R|^2 2^{-n} \Big)^{\frac{1}{2}} \tag{15}$$

$$\gtrsim \frac{\sum_{|R|=2^{-n}} |\alpha_R| 2^{-n/2}}{\big(n^{d-1}2^n\big)^{\frac{1}{2}}} = n^{-\frac{d-1}{2}} \cdot 2^{-n} \sum_{|R|=2^{-n}} |\alpha_R|,$$

where we have used the fact that the total number of rectangles $R \in \mathscr{D}^d$ is $\approx n^{d-1}2^n$. This computation is similar in spirit to (7) establishing Roth's L^2 discrepancy bound. Generally, the Small Ball Conjecture bears a strong resemblance to Conjecture 1 about the discrepancy function. Indeed, in both cases one gains a square root of the logarithm over the L^2 bound.

One can consider a restricted version of inequality (14), which appears to contain virtually all the complexity of the general inequality and is sufficient for applications:

$$\Big\| \sum_{|R|=2^{-n}} \varepsilon_R h_R \Big\|_\infty \gtrsim n^{d/2}\,, \qquad \varepsilon_R \in \{-1, 0, 1\}\,, \tag{16}$$

subject to the requirement that $\sum_{|R|=2^{-n}} |\varepsilon_R| \geq c 2^n n^{d-1}$ for a fixed small constant $c > 0$, in other words, at least a fixed proportion of the coefficients ε_R are non-zero. The relation to the discrepancy estimates becomes even more apparent for this form of the inequality. For instance, the trivial bound (15) becomes

$$\Big\| \sum_{|R|=2^{-n}} \varepsilon_R h_R \Big\|_\infty \geq \Big\| \sum_{|R|=2^{-n}} \varepsilon_R h_R \Big\|_2 \gtrsim n^{(d-1)/2}\,. \tag{17}$$

Compare this to Roth's bound (2), and compare (16) to Conjecture 1. The similarities between the discrepancy estimates and the Small Ball inequality are summarized in Table 1.

A more restrictive version of inequality (14) with $\varepsilon_R = \pm 1$ (the *signed small ball inequality*) does allow for some proof simplifications, but has no direct consequences. The papers [8, 9] study this restricted inequality, using only the fundamental inequality – Lemma 2 of Sect. 6. This case will likely continue to be a proving ground for new techniques in this problem.

Conjecture 2 is sharp: for independent random selection of coefficients (either random signs or Gaussians), the supremum is at most $C n^{d/2}$,

$$\mathbb{E} \Big\| \sum_{|R|=2^{-n}} \alpha_R h_R \Big\|_\infty \simeq n^{d/2}\,.$$

Table 1 Discrepancy estimates and the signed Small Ball inequality.

Discrepancy estimates	**Small Ball inequality (signed)**
Dimension $d = 2$	
$\|D_N\|_\infty \gtrsim \log N$ (Schmidt, '72; Halász, '81)	$\left\| \sum_{\|R\|=2^{-n}} \varepsilon_R h_R \right\|_\infty \gtrsim n$ (Talagrand, '94; Temlyakov, '95)
Higher dimensions, L^2 bounds	
$\|D_N\|_2 \gtrsim (\log N)^{(d-1)/2}$	$\left\| \sum_{\|R\|=2^{-n}} \varepsilon_R h_R \right\|_2 \gtrsim n^{(d-1)/2}$
Higher dimensions, conjecture	
$\|D_N\|_\infty \gtrsim (\log N)^{d/2}$	$\left\| \sum_{\|R\|=2^{-n}} \varepsilon_R h_R \right\|_\infty \gtrsim n^{d/2}$
Higher dimensions, known results	
$\|D_N\|_\infty \gtrsim (\log N)^{\frac{d-1}{2}+\eta}$	$\left\| \sum_{\|R\|=2^{-n}} \varepsilon_R h_R \right\|_\infty \gtrsim n^{\frac{d-1}{2}+\eta}$

Unfortunately, random selection of coefficients does not seem to be a guide to the sums that are hardest to analyze. The sharpness of the Small Ball Conjecture justifies our belief in the optimality of Conjecture 1 in discrepancy theory.

4 Connections to Probability and Approximation Theory

We briefly touch upon the connections of the Small Ball inequality (14) to problems in other fields. A very detailed account of these relations is contained in [4].

4.1 Approximation Theory: Metric Entropy of Classes with Dominating Mixed Smoothness

Let MW^p be the image of the unit ball $L^p([0,1]^d)$ under the integration operator $\big(\mathscr{T} f\big)(x) = \int_0^{x_1} \dots \int_0^{x_d} f(y)dy$, i.e. in some sense MW^p is the set of functions on $[0,1]^d$ whose mixed derivative $\frac{\partial^d f}{\partial x_1 \partial x_2 \dots \partial x_d}$ has L^p norm bounded by one. This set is compact in the L^∞ metric and its compactness may be measured by

covering numbers. Let $N(\varepsilon, p, d)$ be the cardinality of the smallest ε-net of MW^p in the L^∞ norm. The exact asymptotics of these numbers as $\varepsilon \downarrow 0$ is a subject of conjecture.

Conjecture 3. For $d \geq 2$, we have $\log N(\varepsilon, 2, d) \simeq \varepsilon^{-1}(\log 1/\varepsilon)^{d-1/2}$, as $\varepsilon \downarrow 0$.

The case $d = 2$ is settled [26], and the upper bound is known in all dimensions [13]. Inequalities similar to the Small Ball Conjecture (14) lead to lower bounds on the covering numbers.

4.2 Probability: The Small Ball Problem for the Brownian Sheet

Consider the Brownian sheet B_d, i.e. a centered multiparameter Gaussian process characterized by the covariance relation $\mathbb{E}B_d(s) \cdot B_d(t) = \prod_{j=1}^d \min\{s_j, t_j\}$. The problem deals with the precise behavior of $\mathbb{P}(\|B\|_{C([0,1]^d)} < \varepsilon)$, the *small deviation (or small ball) probabilities* of B_d.

There is an exciting formal equivalence established by Kuelbs and Li [16, 17] between the small ball probabilities and the metric entropy of the unit ball of the reproducing kernel Hilbert space, which in the case of the Brownian sheet is WM^2. This yields an equivalent conjecture:

Conjecture 4. In dimensions $d \geq 2$, for the Brownian sheet B we have

$$-\log \mathbb{P}(\|B\|_{C([0,1]^d)} < \varepsilon) \simeq \varepsilon^{-2}(\log 1/\varepsilon)^{2d-1}, \quad \varepsilon \downarrow 0.$$

The upper bounds are known for $d \geq 2$ [13], while the lower bound for $d = 2$ has been obtained by Talagrand [26] using (14). It is worth mentioning that Conjecture 4 explains the nomenclature *small ball inequality*.

4.3 Summary of the Connections

The connections between the Small Ball Conjecture and these problems is illustrated in Fig. 1. Solid arrows represent known formal implications, while a dashed line denotes an informal heuristic relation. Hopefully, other lines, as well as other nodes, will be added to this diagram in the future. In particular, we expect that the theory of empirical processes may connect the discrepancy bounds to the small deviation probabilities.

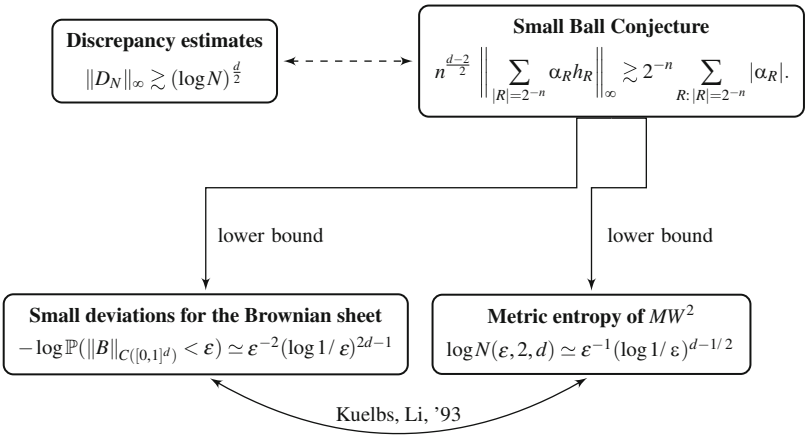


Fig. 1 Connections between the Small Ball Conjecture and other problems.

5 Riesz Product Techniques

The only case in which the Small Ball inequality (14) is known in full generality is dimension $d = 2$, which was proved by M. Talagrand [26].

Theorem 6. *In dimension $d = 2$, there holds for all n,*

$$2^{-n} \sum_{|R|=2^{-n}} |\alpha_R| \lesssim \Big\| \sum_{|R|\geq 2^{-n}} \alpha_R h_R \Big\|_\infty .$$

Soon after M. Talagrand proved Conjecture 2 in dimension $d = 2$, V. Temlyakov [27] has given an alternative elegant proof of this inequality, which strongly resonated with the argument of Halász [15] for (4). We shall present this technically simpler proof and then explain the adjustments needed to obtain the discrepancy bound.

All the endpoint estimates in dimension $d = 2$ are based upon a very special property of the two-dimensional Haar functions and the associated r-functions, **product rule**: if $R,\ R' \in \mathscr{D}^2$ are not disjoint, $R \neq R'$, and $|R| = |R'|$, then

$$h_R \cdot h_{R'} = \pm h_{R\cap R'}, \tag{18}$$

i.e. the product of two Haar functions is again Haar, or equivalently, if $|r| = |s| = n$, then the product $f_r \cdot f_s = f_t$ is also an r function, where $t = (\min\{r_1, s_1\}, \min\{r_2, s_2\})$. In higher dimensions two different boxes of the same volume may coincide in one of the coordinates, in which case $h_{R_k} \cdot h_{R'_k} = h^2_{R_k} = \mathbf{1}_{R_k}$. This loss of orthogonality leads to major complications in dimensions three and above.

Proof. For each $j = 0, \dots, n$ consider the r-functions $f_{(j,n-j)} = \sum_{\substack{|R|=2^{-n},\\ |R_1|=2^{-j}}} \operatorname{sgn}(\alpha_R) h_R$. In dimension $d = 2$ the summation conditions uniquely define the shape of a dyadic rectangle. The product rule drives this argument. We construct the following Riesz product

$$\Psi := \prod_{j=1}^{n} \left(1 + f_{(j,n-j)}\right) = 1 + \sum_{R \in \mathscr{D}^d : |R| = 2^{-n}} \operatorname{sgn}(\alpha_R) h_R \; + \; \Psi_{>n}, \tag{19}$$

where, by the product rule, $\Psi_{>n}$ is a linear combination of Haar functions supported by rectangles of area less than 2^{-n}, and make three simple observations

1. $\Psi \geq 0$, since each factor is either 0 or 2.
2. Next, $\int \Psi(x)dx = 1$. Indeed, expand the product in (19) – the initial term is 1, while all the higher-order terms are Haar functions with mean zero.
3. Therefore Ψ has L^1 norm 1: $\|\Psi\|_1 = 1$.

By the same token, using orthogonality,

$$\left\| \sum_{|R|=2^{-n}} \alpha_R h_R \right\|_\infty \geq \left\langle \sum_{|R|=2^{-n}} \alpha_R h_R, \Psi \right\rangle = \; 2^{-n} \cdot \sum_{|R|=2^{-n}} |\alpha_R|, \tag{20}$$

since $\langle h_R, h_R \rangle = 2^{-n}$. □

Rather than proving Schmidt's discrepancy lower bound, we shall explain how the above argument could be adapted to obtain Halász's proof of (4). These are the necessary changes:

- *Building blocks:* Instead of the r-functions $f_{(j,n-j)} = \sum \operatorname{sgn}(\alpha_R) h_R$ used above, we take the r-functions provided by Lemma 1 with the property that $\langle D_N, f_r \rangle \gtrsim 1$.
- *Riesz product:* The test function $\Psi := \prod_{j=1}^{n} \left(1 + f_{(j,n-j)}\right)$ should be replaced by a slightly more complicated $\Phi = \prod_{j=1}^{n} \left(1 + \gamma f_{(j,n-j)}\right) - 1$, where $\gamma > 0$ is a small constant.

These adjustments play the following roles: -1 in the end forces the "zero-order" term $\int D_N(x)dx$ to disappear, while a suitable choice of the small constant γ takes care of the "higher-order" terms and ensures that their contribution is small. Otherwise, the proof of (4) is verbatim the same as the proof of the two-dimensional Small Ball Conjecture; the details can be found in [3, 8, 15, 19] etc. The Small Ball Conjecture may therefore be viewed as a linear term in the discrepancy estimates. These same comments apply to the proof of the L^1 estimate (12) of Halász.

Table 2 Discrepancy function and lacunary Fourier series.

Discrepancy function $D_N(x) = \#\{\mathscr{P}_N \cap [0,x)\} - Nx_1x_2$	**Lacunary Fourier series** $f(x) \sim \sum_{k=1}^{\infty} c_k \sin n_k x,\ \frac{n_{k+1}}{n_k} > \lambda > 1$
$\|D_N\|_2 \gtrsim \sqrt{\log N}$ (Roth, '54)	$\|f\|_2 \equiv \sqrt{\sum |c_k|^2}$
$\|D_N\|_\infty \gtrsim \log N$ (Schmidt, '72; Halász, '81) Riesz product: $\prod(1 + cf_k)$	$\|f\|_\infty \gtrsim \sum |c_k|$ (Sidon, '27) Riesz product: $\prod \big(1 + \cos(n_k x + \phi_k)\big)$
$\|D_N\|_1 \gtrsim \sqrt{\log N}$ (Halász, '81) Riesz product: $\prod \left(1 + i \cdot \frac{c}{\sqrt{\log N}} f_k\right)$	$\|f\|_1 \gtrsim \|f\|_2$ (Sidon, '30) Riesz product: $\prod \left(1 + i \cdot \frac{\lvert c_k\rvert}{\|f\|_2} \cos(n_k x + \theta_k)\right)$

The power of the Riesz product approach in discrepancy problems and the Small Ball Conjecture can be intuitively justified. The maximal values of the discrepancy function (as well as of hyperbolic Haar sums) are achieved on a very sparse, fractal set. Riesz products are known to capture such sets extremely well. In fact, $\Psi = 2^{n+1}\mathbf{1}_E$, where E is the set on which all the functions f_k are positive, i.e. Ψ defines a uniform measure on the set where the L^∞ norm is achieved. In particular, E is essentially the low-discrepancy van der Corput set [3] if all $\varepsilon_R = 1$ (in this case, $f_{(k,n-k)}$ are Rademacher functions).

Historically, Riesz products were designed to work with lacunary Fourier series, see e.g. [20,24,25,28], that is, Fourier series with harmonics supported on lacunary sequences $\{n_k\}$ with $n_{k+1}/n_k > \lambda > 1$, e.g., $n_k = 2^k$. The terms of such series behave like independent random variables, which resembles our situation, since the functions $f_{(j,n-j)}$ are actually independent. The failure of the product rule explains the loss of independence in higher dimensions (see [7] for this approach towards the conjecture). The strong similarity of the two-dimensional Small Ball inequality and Sidon's theorem on lacunary Fourier series [24]

$$\left\| \sum_{|R|=2^{-n}} \alpha_R h_R \right\|_\infty \gtrsim 2^{-n} \sum_{R:\,|R|=2^{-n}} |\alpha_R| \qquad \text{vs.} \qquad \left\| \sum_k c_k \sin n_k x \right\|_\infty \gtrsim \sum_k |c_k| \tag{21}$$

may be explained heuristically: the condition $|R| = 2^{-n}$ effectively leaves only one free parameter, and the supports of Haar functions are dyadic – thus we

obtain a one-parameter system with lacunary frequencies. The similarities between discrepancy estimates, lacunary Fourier series, and the corresponding Riesz product techniques are shown in Table 2.

6 Recent Results

An improvement of the Small Ball inequality in higher dimensions has been obtained by Bilyk, Lacey, and Vagharshakyan [5, 8].

Theorem 7. *For all dimensions $d \geq 3$, there is an $\eta = \eta(d) > c/d^2$ so that for all integers n there holds*

$$2^{-n} \sum_{|R|=2^{-n}} |\alpha_R| \lesssim n^{\frac{d-1}{2}-\eta} \Big\| \sum_{|R|\geq 2^{-n}} \alpha_R h_R \Big\|_\infty .$$

We shall briefly explain some ideas and complications that arise in the higher-dimensional case.

All simple approaches to these questions are blocked by the dramatic failure of the product rule in dimensions $d \geq 3$. This failure, as well as potential remedies, was first addressed in the breakthrough paper of József Beck [1]. Recall that the product rule breaks when some sides of the dyadic rectangles coincide. There is a whole range of inequalities which partially compensate for the absence of the product rule and the presence of coincidences. The simplest of these inequalities is the so-called *Beck gain*.

Lemma 2 (Beck gain). *In dimensions $d \geq 3$ there holds*

$$\Big\| \sum_{\substack{r \neq s \,:\, |r|=|s|=n \\ r_1=s_1}} f_r \cdot f_s \Big\|_p \lesssim p^{d-1} n^{\frac{2d-3}{2}}, \qquad 1 < p < \infty . \tag{22}$$

The meaning of this bound can be made clear by simple parameter counting. The summation conditions $|r| = |s| = n$ and $r_1 = s_1$ "freeze" three parameters. Thus the pair of vectors r and s has $2d-3$ free parameters, and the estimate says that they behave in an orthogonal fashion, nearly as if we had just applied the Littlewood-Paley inequality $2d - 3$ times. The actual proof is more complicated, of course, since the variables in the sum are not free as they are in (9). The paper of Beck [1] contains a weaker version of the lemma above in the case of $d = 3$, $p = 2$. The L^p version is far more useful: the case $d = 3$ is in [5], and an induction on dimension argument [8] proves the general case.

To apply the Riesz product techniques one has to be able to deal with longer, more complicated patterns of coincidences. This would require inequalities of the type

$$\left\| \sum f_{r_1} \cdots f_{r_k} \right\|_p \lesssim p^{\alpha M} n^{\frac{M}{2}}, \tag{23}$$

where the summation is extended over all k-tuples of d-dimensional integer vectors $r_1, \dots, r_k$ with a specified configuration of coincidences and M is the number of free parameters imposed by this configuration, i.e. the free parameters should still behave orthogonally even for longer coincidences. If $k = 2$, this is just (22); in [8] a partial result in this direction is obtained for $k > 2$.

While the breakdown of the product rule is a feature of the method, there are intrinsic issues that demonstrate that the higher-dimensional inequality is much more delicate and difficult than the case $d = 2$. There is no simple closed form for the dual function in this situation. Indeed, assume that all $|\alpha_R| = 1$. One then wants to show that the sum $\sum_{R\,:\,|R|=2^{-n}} \alpha_R h_R(x) \gtrsim n^{d/2}$ for some values of x. But every x is contained in many more, namely $cn^{d-1} \gg n^{d/2}$, rectangles of volume 2^{-n}. That is, one has to identify a collection of points which capture only a very slight disbalance between the number of positive and negative summands. There doesn't seem to be any canonical way to select such a set of points in the higher-dimensional setting, let alone construct a function similar to the Riesz product (19), which would be close to uniform measure on such a set, see [7].

6.1 Other Endpoint Estimates

The Small Ball Conjecture provides supporting evidence for Conjecture 1 on the behavior of the L^∞ norm of the discrepancy function in dimensions $d \geq 3$, $\|D_N\| \gtrsim (\log N)^{d/2}$. On the other hand, the best known examples of point sets $\mathscr{P}_N$ satisfy $\|D_N\|_\infty \lesssim (\log N)^{d-1}$. However, the techniques of the orthogonal function method cannot prove anything better than the Small Ball inequality.

As we have pointed out repeatedly, the set on which D_N achieves its L^∞ norm is a small set. Exactly how small has been quantified in the two-dimensional setting by Bilyk, Lacey, Parissis, Vagharshakyan [6].

Theorem 8. *In dimension $d = 2$, for any integer N*

(a) For any point set $\mathscr{P}_N$ with $\#\mathscr{P}_N = N$, and $2 < q < \infty$, we have

$$\|D_N\|_{\exp(L^q)} \gtrsim (\log N)^{1-1/q}; \tag{24}$$

(b) There exists a set $\mathscr{P}_N$ (a shifted van der Corput set) such that for $2 \leq q < \infty$,

$$\|D_N\|_{\exp(L^q)} \lesssim (\log N)^{1-1/q}.$$

This theorem is an interpolation between Roth's and Schmidt's bounds in dimension two: when $q = 2$ (the subgaussian case) the estimates resembles the

L^2 behavior, $\sqrt{\log N}$, while as q approaches infinity, the bounds become close to the L^∞ estimate, $\log N$.

The crucial index $q = 2$ is the exact limit of Roth's Theorem: $\|D_N\|_{\exp(L^2)} \gtrsim \sqrt{\log N}$ by Roth's theorem, and there is an example of $\mathscr{P}_N$ for which the reverse inequality holds. It is very tempting to speculate that the Orlicz space $\exp(L^2)$ of subgaussian functions is the sharp space in all dimensions.

Conjecture 5. For all dimensions d

$$\inf_{\mathscr{P}_N} \|D_N\|_{\exp(L^2)} \lesssim (\log N)^{(d-1)/2}.$$

This would imply that in the extremal case the set $\{x : D_N(x) \geq (\log N)^{d/2}\}$ would have measure at most N^{-c}, for some positive c. We are of course very far from verifying such conjectures, though they can be helpful in devising potential proof strategies for the main goal – Conjecture 1.

Acknowledgements This research is supported in part by NSF grants DMS 1101519, 1260516 (Dmitriy Bilyk), DMS 0968499, and a grant from the Simons Foundation #229596 (Michael Lacey).

References

1. Beck, J.: A two-dimensional van Aardenne-Ehrenfest theorem in irregularities of distribution. Compos. Math. **72**, 269–339 (1989)
2. Beck, J., Chen, W.W.L.: Irregularities of Distribution. Cambridge University Press, Cambridge (1987)
3. Bilyk, D.: On Roth's orthogonal function method in discrepancy theory. Unif. Distrib. Theory **6**, 143–184 (2011)
4. Bilyk, D.: Roth's orthogonal function method in discrepancy theory and some new connections. In: Chen, W., Srivastav, A., Travaglini, G. (eds.) Panorama of Discrepancy Theory. Springer (2013–14, to appear)
5. Bilyk, D., Lacey, M.T.: On the small ball inequality in three dimensions. Duke Math. J. **143**, 81–115 (2008)
6. Bilyk, D., Lacey, M.T., Parissis, I., Vagharshakyan, A.: Exponential squared integrability of the discrepancy function in two dimensions. Mathematika **55**, 1–27 (2009)
7. Bilyk, D., Lacey, M.T., Parissis, I., Vagharshakyan, A.: A three-dimensional signed small ball inequality. In: Dependence in Probability, Analysis and Number Theory, pp. 73–87. Kendrick, Heber City (2010)
8. Bilyk, D., Lacey, M.T., Vagharshakyan, A.: On the small ball inequality in all dimensions. J. Funct. Anal. **254**, 2470–2502 (2008)
9. Bilyk, D., Lacey, M.T., Vagharshakyan, A.: On the signed small ball inequality. Online J. Anal. Comb. **3** (2008)
10. Chang, S.-Y.A., Wilson, J.M., Wolff, T.H.: Some weighted norm inequalities concerning the Schrödinger operators. Comment. Math. Helv. **60**, 217–246 (1985)
11. Dick, J., Pillichshammer, F.: Digital Nets and Sequences. Cambridge University Press, Cambridge (2010)
12. Drmota, M., Tichy, R.: Sequences, Discrepancies and Applications. Springer, Berlin (1997)

13. Dunker, T., Kühn, T., Lifshits, M., Linde, W.: Metric entropy of the integration operator and small ball probabilities for the brownian sheet. C. R. Acad. Sci. Paris Sér. I Math. **326**, 347–352 (1998)
14. Fefferman, R., Pipher, J.: Multiparameter operators and sharp weighted inequalities. Amer. J. Math. **119**, 337–369 (1997)
15. Halász, G.: On Roth's method in the theory of irregularities of point distributions. In: Recent Progress in Analytic Number Theory, vol. 2, pp. 79–94. Academic, London (1981)
16. Kuelbs, J., Li, W.V.: Metric entropy and the small ball problem for Gaussian measures. C. R. Acad. Sci. Paris Sér. I Math. **315**, 845–850 (1992)
17. Kuelbs, J., Li, W.V.: Metric entropy and the small ball problem for Gaussian measures. J. Funct. Anal. **116**, 133–157 (1993)
18. Lacey, M.: On the discrepancy function in arbitrary dimension, close to L^1. Anal. Math. **34**, 119–136 (2008)
19. Matoušek, J.: Geometric Discrepancy. Springer, Berlin (2010)
20. Riesz, F.: Über die Fourierkoeffizienten einer stetigen Funktion von beschränkter Schwankung. Math. Z. **2**, 312–315 (1918)
21. Roth, K.F.: On irregularities of distribution. Mathematika **1**, 73–79 (1954)
22. Schmidt, W.M.: Irregularities of distribution, VII. Acta Arith. **21**, 45–50 (1972)
23. Schmidt, W.M.: Irregularities of distribution, X. In: Zassenhaus, H. (ed.) Number Theory and Algebra, pp. 311–329. Academic, New York (1977)
24. Sidon, S.: Verallgemeinerung eines Satzes über die absolute Konvergenz von Fourierreihen mit Lücken. Math. Ann. **97**, 675–676 (1927)
25. Sidon, S.: Ein Satz über trigonometrische Polynome mit Lücken und seine Anwendung in der Theorie der Fourier-Reihen. J. Reine Angew. Math. **163**, 251–252 (1930)
26. Talagrand, M.: The small ball problem for the Brownian sheet. Ann. Probab. **22**, 1331–1354 (1994)
27. Temlyakov, V.N.: An inequality for trigonometric polynomials and its application for estimating the entropy numbers. J. Complexity **11**, 293–307 (1995)
28. Zygmund, A.: Trigonometric Series, vols. I, II. Cambridge University Press, Cambridge (2002)

An Introduction to Stochastic Particle Integration Methods: With Applications to Risk and Insurance

Pierre Del Moral, Gareth W. Peters, and Christelle Vergé

Abstract This article presents a guided introduction to a general class of interacting particle methods and explains throughout how such methods may be adapted to solve general classes of inference problems encountered in actuarial science and risk management. Along the way, the resulting specialized Monte Carlo solutions are discussed in the context of how they complemented alternative approaches adopted in risk management, including closed form bounds and asymptotic results for functionals of tails of risk processes.

The development of the article starts from the premise that whilst interacting particle methods are increasingly used to sample from complex and high-dimensional distributions, they have yet to be generally adopted in inferential problems in risk and insurance. Therefore, we introduce a range of methods which can all be interpreted in the general framework of interacting particle methods, which goes well beyond the standard particle filtering framework and Sequential Monte Carlo frameworks. For the applications we consider in risk and insurance we focus on particular classes of interacting particle genetic type algorithms. These stochastic particle integration techniques can be interpreted as a universal acceptance-rejection sequential particle sampler equipped with adaptive and interacting recycling mechanisms. We detail how one may reinterpret these stochastic particle integration

P. Del Moral (✉)
INRIA Bordeaux-Sud Ouest, and Bordeaux Mathematical Institute,
Ecole Polytechnique (CMAP), Palaiseau, Paris, France
e-mail: Pierre.Del_Moral@inria.fr

G.W. Peters
School of Mathematics and Statistics, University of New South Wales, Sydney, NSW, Australia
CMIS – Commonwealth Scientific and Industrial Research Organization, Sydney, Australia
Current address: Department of Statistical Science, University College London, London, UK
e-mail: gareth.peters@ucl.ac.uk

C. Vergé
ONERA – The French Aerospace Lab, F-91761, Palaiseau, Paris, France
e-mail: christelle.verge@onera.fr

J. Dick et al. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2012*,
Springer Proceedings in Mathematics & Statistics 65,
DOI 10.1007/978-3-642-41095-6_3, © Springer-Verlag Berlin Heidelberg 2013

techniques under a Feynman-Kac particle integration framework. In the process, we illustrate how such frameworks act as natural mathematical extensions of the traditional change of probability measures, common in designing importance samplers for risk managements applications.

1 Introduction to Stochastic Particle Integration

The intention of this paper is to introduce a class of stochastic particle based integration techniques to a broad community, with a focus on risk and insurance practitioners. We will demonstrate that a range of problems in risk and insurance can directly benefit from the development of such methods. A key motivation for this endeavour is the fact that stochastic particle integration models have been extensively used in engineering, statistics and physics under sometimes different names, such as: particle filters, bootstrap or genetic filters, population Monte Carlo methods, sequential Monte Carlo models, genetic search models, branching and multi-level splitting particle rare event simulations, condensation models, go-with-the winner, spawning models, walkers, population reconfigurations, pruning-enrichment strategies, quantum and diffusion Monte Carlo, rejuvenation models, and many others. However, they have not yet been routinely applied to develop solutions in important financial domains such as those we discuss in this tutorial type overview.

We begin with an introduction to the fundamental background for interacting particle systems by highlighting key papers in their development in a range of different scientific disciplines, before introducing aspects of these stochastic methods to risk and insurance. It is the intention of this article to explain the key papers and ideas in a general Feynman-Kac interacting particle framework which is much more encompassing than the special subset of the well known particle filter based algorithms. We proceed through a selection of key features of the development of interacting particle systems, focusing on a sub-class of such methods of relevance to the application domain explored in this manuscript, risk and insurance.

The origins of stochastic particle simulation certainly starts with the seminal paper of N. Metropolis and S. Ulam [52]. As explained by these two physicists in the introduction of their pioneering article, the Monte Carlo method is, "essentially, a statistical approach to the study of differential equations, or more generally, of integro-differential equations that occur in various branches of the natural sciences". The links between genetic type particle Monte Carlo models and quadratic type parabolic integro-differential equations has been developed in the beginning of 2000' in the series of articles on continuous time models [24, 26].

The earlier works on heuristic type genetic particle schemes seem to have started in Los Alamos National Labs with works of M.N. Rosenbluth and A.W. Rosenbluth [68], and earlier works by H. Kahn and T. Harris [44]. We also quote the work on artificial life of Nils Aall Barricelli [5,6]. In all of these works, the genetic Monte Carlo scheme is always presented as a natural heuristic resampling type algorithm

to generate random population models, to sample molecular conformations, or to estimate high energy particle distributions, without a single convergence estimate to ensure the performance, nor the robustness of the Monte Carlo sampler.

The mathematical foundations, and the performance analysis of all of these discrete generation particle models is a rather recent development. The first rigorous study in this field seems to be the article [11] published in 1996 on the applications of particle methods to non-linear estimation problems. This article provides the first proof of the unbiased property of particle likelihood approximation models (Lemma 3 page 12); and adaptive resampling criteria w.r.t. the weight dispersions (see Remark 1 on page 4). We also quote the first articles presenting heuristic type particle filters [41, 45], and a series of earlier research reports [27, 29–31].

For an in-depth description of the origins of particle methods and their applications we refer to the following studies [18, 28]. These articles also contain new stochastic models and methods including look-ahead type strategies (Sect. 4.2.2), reducing the variance using conditional explorations w.r.t. the observation sequences (Example 3 page 40), local errors transport models (see the proof of Theorem 1 on page 11) and mean field models w.r.t. the occupation measures of random trees (Sect. 3.2).

A more detailed review of particle models in discrete and continuous time can be found in [19, 23]. In the research monograph [19], the reader will find a detailed discussion on particle models and methods including acceptance-rejection with recycling particle strategies, interacting Kalman filters a.k.a. Rao-Blackwellized particle filters (Sects. 2.6 and 12.6.7), look-ahead type strategies (Sect. 12.6.6), genealogical tree models and branching strategies (Sect. 11), and interacting Metropolis-Hasting models (Chap. 5).

The practitioner will find in the research books [19, 20, 22, 23] a source of useful convergence estimates as well as a detailed list of concrete examples of particle approximations for real models, including restricted Markov chain simulations, random motions in absorbing media, spectral analysis of Schrodinger operators and Feynman-Kac semi-groups, rare event analysis, sensitivity measure approximations, financial pricing numerical methods, parameter estimation in HMM models, island particle models, interacting MCMC models, statistical machine learning, Bayesian inference, Dirichlet boundary problems, non-linear filtering problems, interacting Kalman-Bucy filters, directed polymer simulations, stochastic optimization, and interacting Metropolis type algorithms. For further discussion on the origins and the applications of these stochastic models, we refer the reader to the following texts [3, 12, 13, 21, 33, 34, 46, 53, 57, 60], and the references therein.

Despite this, particle methods are yet to be routinely or widely introduced to the areas of risk and insurance modelling. The initial examples that have been developed are detailed in [61], where a special sub-class of such methods for an important set of risk management problems was explained. It is therefore the intention of this paper to highlight aspects of this class of problems and the stochastic particle solutions that will aid further development of these approaches in risk modelling.

2 Motivation for Stochastic Particle Solutions: *Examining How Such Methods May Complement Risk Asymptotics*

In the following subsections we provide motivation and context that will explain how and why risk management and actuarial sciences can benefit from the development of interacting particle system solutions. In particular, we will focus on a few key estimation problems that form an important subset of generic problems faced by practitioners in these domains. This will involve consideration of single risk loss processes described under a Loss Distributional Approach (hereafter LDA) framework, see discussion in [54, 62, 69] and the books [47] and [70] for the background on such modelling approaches in risk. For basic discussions on how such problems relate to a large class of non-life insurance problems see examples in [60].

2.1 The Loss Distributional Approach and Risk Management: A Tale of Light to Heavy Tails

In this section we first motivate and introduce the context of LDA modelling in risk and insurance. Then we present three key challenges associated with working with such LDA models that are commonly encountered by risk and insurance practitioners, thereby explaining some important inference challenges faced by such practitioners. Next, we provide a brief specifically selected survey of closed form analytic results known in the actuarial and risk literature for sub-classes of such LDA models as the Single Loss Approximations (hereafter SLA). We first detail the closed form solution for the light tailed severity distribution case. Then we explain how such results that are applicable in the light tailed case cannot be obtained in such a form in the heavy tailed sub-exponential risk process settings. Consequently, we briefly present the results recently developed in actuarial literature for the heavy tailed case corresponding to the first order and second order asymptotic approximations, see comprehensive discussions in a general context in [1, 16], and the books [4] and [14].

The fact that SLA approximations are inherently asymptotic in nature, and may be inaccurate outside of the neighbourhood of infinity, typically means that in practice risk managers must resort to numerical procedures to estimate risk measures and capital, see discussions in [66]. It is in these cases we will explain and motivate the utility of interacting particle based solutions.

Consider the widely utilised insurance model known as a single risk LDA model. This represents the standard under the Basel II/III capital accords [7] and involves an annual loss in a risk cell (business line/event type) modelled as a compound distributed random variable,

$$Z_t^{(j)} = \sum_{s=1}^{N_t^{(j)}} X_s^{(j)}(t)\,, \tag{1}$$

for $t = 1, 2, \ldots, T$ discrete time (in annual units) and index j identifies the risk cell. Furthermore, the annual number of losses is denoted by $N_t^{(j)}$ which is a random variable distributed according to a frequency counting distribution $P^{(j)}(\cdot)$, typically Poisson, Binomial or Negative Binomial. The severities (losses) in year t are represented by random variables $X_s^{(j)}(t)$, $s \geq 1$, distributed according to a severity distribution $F^{(j)}(\cdot)$ and there are $N_t^{(j)}$ of them in year t.

In general, we will suppress the risk cell index j and time index t unless explicitly utilised. Therefore, we denote by $F(x)$ a distribution with positive support for the severity model characterizing the loss distribution for each random variable X_s for $s \in \{1, 2, \ldots, N\}$. We denote the annual loss (aggregated loss) by Z with annual loss distribution $G = F_Z$ and the partial sum of n random losses by S_n with distribution $F_{S_n} = F^{n*}$ where F^{n*} denotes the n-fold convolution of the severity distribution for the independent losses. The densities, when they exist, for the severity distribution and annual loss distributions will be denoted by $f(x)$ and $f_Z(x)$ respectively.

We assume that all losses are i.i.d. with $X_s^{(j)}(t) \sim F(x)$ where $F(x)$ is continuous with no atoms in the support $[0, \infty)$. As a consequence, linear combinations (aggregation) of losses in a given year satisfy

$$S_n(t) = \sum_{s=1}^{n} X_s^{(j)}(t) \sim F_{S_n}(x)$$

and have the following analytic representation:

$$F_{S_n}(x) = (F \star F \star \cdots F)(x) = \int_{[0,\infty)} F^{(n-1)\star}(x - y) dF(x).$$

In [38] it is shown that if $F(x)$ has no atoms in $[0, \infty)$ then the n-fold convolution of such severity distributions will also admit no atoms in the support $[0, \infty)$. The implications of this for interacting particle based numerical procedures is that it ensures numerical techniques are well defined for such models. In particular the ratios of densities on the support $[0, \infty)$ are well defined. This is important as it is often required for interacting particle methods. In addition we note that continuity and boundedness of a severity distribution $F(x)$ is preserved under n-fold convolution. Hence, if $F(x)$ admits a density $\frac{d}{dx}F(x)$ then so does the distribution of the partial sum F_{S_n}, for any $n \in \{1, 2, \ldots\}$ and compound process (random sum) F_Z.

In practice the choice of severity distribution $F(x)$ should be considered carefully for each individual risk process. As discussed in [66] it is common to consider sub-exponential severity distributions that we denote by membership $(F(x) \in \mathscr{F})$. The sub-exponential family of distributions $\mathscr{F}$ defines a class of heavy tailed severity models that satisfy the limits

$$\lim_{x\to\infty} \frac{1 - F^{n\star}(x)}{1 - F(x)} = n, \tag{2}$$

if and only if,

$$\lim_{x\to\infty} \frac{1 - F^{2\star}(x)}{1 - F(x)} = 2. \tag{3}$$

Alternatively, one may characterize the family of distributions $F \in \mathscr{F}$ by those that satisfy asymptotically the tail ratio

$$\lim_{x\to\infty} \frac{\overline{F}(x - y)}{\overline{F}(x)} = 1, \ \forall y \in [0, \infty). \tag{4}$$

Severity models $F \in \mathscr{F}$ are of interest in high consequence loss modelling since they include models with *infinite mean loss* and *infinite variance*. In addition, the class $\mathscr{F}$ includes all severity models in which the tail distribution under the log transformed r.v., $\overline{F}(\log(x))$, is a slowly varying function of x at infinity.

To further understand LDA modelling with sub-exponential severity models it will be beneficial to recall the notion of asymptotic equivalence in which a probability distribution function $F(x)$ is *asymptotically equivalent* to another probability distribution function $G(x)$, denoted by $F(x) \sim G(x)$ as $x \to \infty$ if it holds that, $\forall \epsilon > 0, \exists x_0$ such that $\forall x > x_0$ the following is true

$$\left| \frac{F(x)}{G(x)} - 1 \right| < \epsilon. \tag{5}$$

Furthermore, we say that a probability distribution function is *max-sum-equivalent*, denoted by $F \sim_M G$, when the convolution of the tail distribution of two random variables is distributed according to the sum of the two tail distributions asymptotically,

$$1 - (F \star G)(x) = (\overline{F \star G})(x) \sim \overline{F}(x) + \overline{G}(x), \quad x \to \infty.$$

Then for the class of heavy tailed sub-exponential LDA models we have that a probability distribution function F will belong to the sub-exponential class $\mathscr{F}$ if $F \sim_M F$, i.e. it is max-sum-equivalent with itself and that the class $\mathscr{F}$ is closed under convolutions. The implications of this for LDA models is clear when one observes that sub-exponential LDA models are compound process random sums comprised of an infinite mixture of convolved distributions,

$$G(x) = \sum_{n=0}^{\infty} p_n F^{n\star}(x), \tag{6}$$

for a suitable series $\{p_n\}$, (*e.g. convergent sequence satisfying Kolmogorov three series theorem*). Using the property of max-sum equivalence one can show the practically relevant asymptotic equivalence between the severity distribution F and the annual loss distribution G in which selecting $F \in \mathscr{F}$ results in $G \in \mathscr{F}$ and

$$\lim_{x\to\infty} \frac{\overline{G}(x)}{\overline{F}(x)} = \lambda.$$

This asymptotic equivalence relationship between the severity distribution F and the annual loss distribution G, present for sub-exponential LDA models, greatly benefits the formulation of asymptotic approximations of tail functionals used in the estimation of bank capital.

Based on these aforementioned properties we can obtain asymptotic approximations to the annual loss distribution tails which typically fall under one of the following classifications:

- "First-Order" and "Second-Order" Single Loss Approximations: recently discussed in [10, 16, 17] and references therein.
- "Higher-Order" Single Loss Approximations: see discussions in [8] and recent summaries in [1] and references therein.
- Extreme Value Theory (EVT) Single Loss Approximations (Penultimate Approximations): the EVT based asymptotic estimators for linear normalized and power normalized extreme value domains of attraction were recently discussed in [17].
- Doubly Infinitely Divisible Tail Asymptotics given α-stable severity models discussed in [58, 64].

We now briefly detail the first and second order asymptotics that are known in the risk literature for light and heavy tailed severity distributions in LDA models. Then we explain how stochastic particle methods can be utilised to complement such closed form expressions in practical banking models and scenarios.

2.1.1 A Light Tale of Light Tails

Here we recall some asymptotic results known for light tailed models as these will inform the results obtained in the heavy tailed expansions. A useful result in the light tailed case was provided by Embrechts and Puccetti [37] where they consider frequency distributions $p_n = \mathbb{P}\mathrm{r}(N = n)$ satisfying

$$p_n \sim w^n n^\gamma C(n), \quad \text{as} \quad n \to \infty,$$

for some $w \in (0, 1)$, $\gamma \in \mathbb{R}$ and a function $C(n)$ slowly varying at ∞. Then, if there exists $\kappa > 0$, such that the Laplace transform of the severity

$$L_X(s) = \mathscr{L}[F(x)] = \int_0^\infty \exp(-sx) dF(x), \ \forall s \in \mathbb{R},$$

matches the radius of convergence of the generating function of the frequency distribution,

$$w^{-1} = L_X(-\kappa),$$

with $-L'_X(-\kappa) < \infty$, one can state the following asymptotic equivalence for the compound process tail distribution,

$$\overline{F}_{Z_N}(x) \sim \frac{x^\gamma \exp(-\kappa x) C(x)}{\kappa \left(-wL'_X(-\kappa)\right)^{\gamma+1}}, \quad x \to \infty.$$

This light tailed asymptotic result demonstrates that the behaviour of the compound loss distribution tail is determined by either the frequency or the severity depending on which has the heavier tail. In addition, it is clear that the Poisson distribution tail is too light for this result to be valid, since the radius of convergence of generating function is infinite. There are therefore alternative expansions developed for compound Poisson risk processes such as the saddle point approximation.

So how do light tailed results motivate the context we are considering in sub-exponential LDA models?

In the sub-exponential heavy tailed setting the Laplace transform does not exist and hence these results do not apply. This is unfortunate, since rates of convergence and approximation accuracy are studied for such results. There are many important examples of LDA models for which these light tailed results do not apply, these include severity distributions with power law tail decay (*Pareto, Burr, log gamma, Cauchy, α-Stable, tempered stable and t-distribution*). In the sub-exponential model setting it is often possible to develop alternative asymptotic results, however asymptotic convergence rates are typically not available. This is one area where particle integration methods can also be informative and complementary in the study of such LDA closed form asymptotics.

2.1.2 A Heavier Tale of Heavy Tails

In this subsection we briefly detail the asymptotic fist and second order tail results for the LDA models when sub-exponential severity distributions are considered. The sub-exponential LDA first order tail asymptotics involve obtaining a closed form approximate expression for $\overline{F_Z}(x)$, see details in [9, 17]. To proceed, consider the annual loss distribution $G(z) = F_Z(z)$ under LDA formulation with the severity distribution satisfying $F \in \mathscr{F}$,

$$G(z) = F_Z(z) = \sum_{n=0}^{\infty} Pr[Z \leq z | N = n] \, Pr[N = n] = \sum_{n=0}^{\infty} p_n F^{(n)\star}(z). \quad (7)$$

Furthermore, to ensure convergence when evaluating Eq. (7) it is required that one assumes that for some $\epsilon > 0$, the following condition is satisfied

$$\sum_{n=0}^{\infty} (1+\epsilon)^n \, p_n < \infty.$$

Then the right tail of the annual loss distribution $F_Z(z)$ for the annual loss random variable Z, is approximated according to a SLA given by,

$$\overline{F_Z}(x) = \mathbb{E}[N] \, \overline{F}(x) \, (1 + o(1)) \quad \text{as} \quad x \to \infty.$$

To understand this basic result of the first order tail asymptotic $\overline{F_Z}(x)$ consider the following two steps:

1. Obtain an upper bound on the asymptotic ratio of $\overline{F}_{S_n}(x)$ and severity $\overline{F}(x)$ for all $n \in \mathbb{J}$. Typically one can apply Kesten's Bound which states that for sub-exponential severity distributions F there exists a constant $K = K(\epsilon) < \infty$ for $\epsilon > 0$ s.t. $\forall n \geq 2$ the following bound holds [15],

$$\frac{\overline{F^{*n}}(x)}{\overline{F}(x)} \leq K(1+\epsilon)^n, \; x \geq 0.$$

2. Then utilise Kesten's bound to motivate the application of dominated convergence theorem to interchange the order of summation and limit and recall the characterization of heavy tailed sub-exponential severity models to obtain $\overline{F_Z}(x) \sim \mathbb{E}[N]\overline{F}(x)$ since,

$$\lim_{x\to\infty} \frac{\overline{F_Z}(x)}{\overline{F}(x)} = \lim_{x\to\infty} \sum_{n=1}^{\infty} p_n \frac{\overline{F^{*n}}(x)}{\overline{F}(x)} = \sum_{n=1}^{\infty} n p_n = \mathbb{E}[N].$$

As discussed in [16], and the papers therein, the second order asymptotic results can be developed in a wide class of risk models by considering the following further assumptions.

Assumption 1. *F is zero at the origin ($x = 0$) and satisfies that both the tail distribution $\overline{F}$ and density f are sub-exponential.*

Assumption 2. *The frequency distribution $N \sim F_N(n)$ is such that its probability generating function given by*

$$p_N(v) = \mathbb{E}\left[v^N\right] = \sum_{n=0}^{\infty} \mathbb{P}\mathrm{r}(N = n)v^n,$$

is analytic at $v = 1$.

Examples of severity models widely used in risk and insurance settings that satisfy such assumptions include: Log-Normal, Weibull (heavy tailed), Benktander Type I and Type II, Inverse Gaussian, α-Stable, Halphen Family and certain members of Generalized Hypergeometric family.

Given a distribution satisfying Assumptions 1 and 2, then two second order results are obtained. One for finite mean loss models and the other for infinite mean loss models. If the loss r.v. has finite mean ($\mathbb{E}[X] < \infty$) then the following result can be derived, see [55] and [73] for details,

$$\lim_{x\to\infty} \frac{\overline{F}_Z(x) - \mathbb{E}[N]\overline{F}(x)}{f(x)} = \mathbb{E}[X]\mathbb{E}[(N-1)N]. \tag{8}$$

Alternatively, if the loss r.v. has an infinite mean but the severity density satisfies the regular variation condition $f \in RV_{-1/\beta-1}$ for $1 \leq \beta < \infty$ then,

$$\lim_{x\to\infty} \frac{\overline{F}_Z(x) - \mathbb{E}[N]\overline{F}(x)}{f(x)\int_0^x \overline{F}(s)ds} = c_\beta \mathbb{E}[(N-1)N],$$

with $c_1 = 1$ and $c_\beta = (1-\beta)\frac{\Gamma^2(1-1/\beta)}{2\Gamma(1-2/\beta)}$ for $\beta \in (1,\infty)$.

2.2 *Inferential Challenges for Risk and Insurance: Asymptotics and the Role for Stochastic Particle Integration*

The asymptotic approximation methods just surveyed were developed in the actuarial literature to tackle the serious statistical and computational challenges posed by estimation of tail quantiles and expectations for heavy tailed LDA models. The continued interest in such asymptotic results primarily stems from the fact that such closed form expressions bypass the significant computational challenges involved in estimation of risk measures for such heavy tailed annual loss distributions under traditional integration methods, Fourier methods, recursions (Panjer) or basic Monte Carlo approaches. However, they do have associated issues, see discussions in [42].

The properties of such asymptotic single loss approximation estimates are still an active subject of study with regard to explicit approximation errors, asymptotic rates of convergence and sensitivity to parameter estimation. To understand these features for loss approximations as well as to provide an alternative estimation approach for tail functionals we propose the application of interacting particle methods.

As summarized in [66] these single loss approximations can be utilised to form estimation of risk and capital approximations by obtaining an expression for the LDA model quantile function. For example, based on second order asymptotic results in the heavy tailed LDA models, one can show that if the severity distribution F satisfies Assumptions 1 and 2 with a finite mean, and the hazard rate $h(x) = \frac{f(x)}{1-F(x)}$ is of regular variation $h \in RV_{-\beta}$ for $\beta \geq 0$, then as $\alpha \to 1$ one has for the inverse of the annual loss distribution the result (see [1]),

$$F_Z^{-1}(\alpha) = F^{-1}\left(1 - \frac{1-\alpha}{\mathbb{E}[N]}\left\{1 + \tilde{c}_\beta g_1\left(F^{-1}(\tilde{\alpha})\right) + o\left(g_1\left(F^{-1}(\tilde{\alpha})\right)\right)\right\}^{-1}\right) \tag{9}$$

where $\tilde{\alpha} = 1 - (1-\alpha)/\mathbb{E}[N]$ and

$$g_1(x) = \begin{cases} \frac{f(x)}{1-F(x)}, & \text{if } \mathbb{E}[X] < \infty, \\ \frac{\int_0^x \overline{F}(s)ds\, f(x)}{1-F(x)}, & \text{if } \mathbb{E}[X] = \infty.; \end{cases}$$

$$\tilde{c}_\beta = \begin{cases} \frac{\mathbb{E}[X]\mathbb{E}[(N-1)N]}{\mathbb{E}[N]}, & \text{if } \mathbb{E}[N] < \infty, \\ \frac{c_\beta \mathbb{E}[(N-1)N]}{\mathbb{E}[N]}, & \text{if } \mathbb{E}[N] = \infty. \end{cases}$$

Using this result it is then possible to consider asymptotic approximations of key risk management quantities known as risk measures which are used in the allocation of capital and reserving in all financial institutions and stipulated as standards under regulatory accords in both Basel II/III and Solvency II.

For example, one may now utilise this approximation to the annual loss quantile function to obtain estimates of common risk measures and regulatory capital, see [2] and [50]. Examples of such risk measures include the *Value-at-Risk (VaR)* which is defined for a level $\alpha \in (0,1)$ and corresponds to the quantile of the annual loss distribution,

$$\begin{aligned} \text{VaR}_Z(\alpha) &= F_Z^{\leftarrow}(\alpha) = \inf\{z \in \mathbb{R} : F_Z(z) \geq \alpha\} \\ &\approx F_Z^{\leftarrow}\left(1 - \frac{1-\alpha}{\mathbb{E}[N]}[1 + o(1)]\right) \sim F^{\leftarrow}\left(1 - \frac{1-\alpha}{\mathbb{E}[N]}\right), \end{aligned} \tag{10}$$

where $F^{\leftarrow}(\cdot)$ is the generalized inverse, see [36]. A second alternative, which includes the Expected Shortfall as a special case, is the *Spectral Risk Measure (SRM)*, which for a weight function $\phi : [0,1] \mapsto \mathbb{R}$ is given by

$$\begin{aligned} \text{SRM}_Z(\phi) &= \int_0^1 \phi(s)\text{VaR}_Z(s)ds \\ &\approx \mathscr{K}(\alpha, \phi_1) F^{\leftarrow}\left(1 - \frac{1-\alpha}{\mathbb{E}[N]}\right) \sim \mathscr{K}(\alpha, \phi_1) VaR_Z(\alpha), \end{aligned} \tag{11}$$

with $\forall t \in (1,\infty)$ a function $\phi_1(1-1/t) \leq Kt^{-1/\beta+1-\epsilon}$ for some $K > 0$ and $\epsilon > 0$ where

$$\mathscr{K}(\alpha,\phi_1) = \int_1^\infty s^{1/\beta-2}\phi_1(1-1/s)ds.$$

2.2.1 The Role for Stochastic Particle Methods

Though the asymptotic results presented are elegant for some LDA models and efficient to evaluate in closed form, they do warrant careful consideration in their application, see discussions in [37]. In practice it may often be the case that one requires calculation of VaR, ES and Spectral Risk Measures at levels which do not satisfy such asymptotic properties, rendering such approximations inaccurate. In addition, though not yet a regulatory requirement, it is always good practice to consider the uncertainty associated with the estimation of the tail functionals and quantiles. This can be achieved via statistical confidence intervals, however these are non-trivial to obtain under such asymptotic expansion results. Thirdly, as discussed in [1] and [4], the asymptotic rates of convergence of such approximations are still only known in a little-oh Landau sense and therefore do not inform or guide the applicability of such results. Finally, there is a significant interest in diversification benefits that may be gained through the modelling of tail dependence features in the multi-variate risk process setting. Extending these asymptotic results to multiple risk processes coupled with a copula structure makes the derivation of asymptotic approximations highly challenging, see [43].

It is for these reasons that we argue stochastic particle based numerical solutions for the estimation of risk measures and tail functionals in LDA structures can be of direct utility to complement such asymptotic results. *However, as all practitioners will know, the naive implementation of standard Monte Carlo and stochastic integration approaches to such problems will produce often poor results even for a considerable computational budget, see discussions in [48].* We therefore require specific interacting particle methods to provide accurate and computationally efficient solutions.

3 Selected Topics in Stochastic Integration Methods

In this section we will introduce practitioners to a variety of stochastic integration methods, presenting them formally from a mathematical perspective and making clear the properties of such methods. Note, in this section the notation adopted is utilised to reflect that which is utilised in the statistics and probability literature where much of the formal study of these methods has taken place.

3.1 Standard Monte Carlo Techniques for Risk and Insurance

Here we consider a conceptually simple problem involving a d-dimensional random variable denoted by X and some measurable subset denoted by $A \subset \mathbb{R}^d$. Now suppose we want to compute the quantity $\mathbb{P}(X \in A) := \mathbb{P}_X(A)$. For example in an LDA model in risk one may naturally consider defining A according to the interval for the annual loss Z given by $A = \left[F_Z^{\leftarrow}(\alpha), \infty\right)$ for some quantile level $\alpha \in [0, 1]$ which is typically very close to one. Then we wish to evaluate the probability that the annual loss for a risk process falls in this interval. If α is close to one then such a probability calculation poses a challenging computational task involving rare-event simulation.

The simplest and least computationally efficient approach to such a computation would involve a basic Monte Carlo simulation. To understand this approach we further assume that it is straightforward to generate a sequence $(X^i)_{1 \le i \le N}$ of independent copies of the random variable X. In this situation, the traditional Monte Carlo approximation of the distribution $\mathbb{P}_X$ is given by the empirical measures

$$\mathbb{P}_X^N = \frac{1}{N} \sum_{1 \le i \le N} \delta_{X^i} \longrightarrow_{N \uparrow \infty} \mathbb{P}_X.$$

Now we define a generic, bounded, measurable test function φ on $\mathbb{R}^d$ that will be used through the remaining sections. Then we can say, more precisely, that the convergence can be understood as the weak convergence of empirical measures, in the sense that the sequence of random variables

$$\mathbb{P}_X^N(\varphi) := \int \varphi(x)\, \mathbb{P}_X^N(dx) = \frac{1}{N} \sum_{1 \le i \le N} \varphi(X^i)$$

converges almost surely, to the limiting integrals

$$\mathbb{P}_X(\varphi) = \int \varphi(x)\, \mathbb{P}_X(dx) = \mathbb{E}(\varphi(X)).$$

Using indicator functions of cells in $\mathbb{R}^d$, the shape of the measure $\mathbb{P}_X$ can be obtained by plotting the histograms of the samples X^i in each dimension. By the strong law of large numbers, the above convergence is also met for integrable functions w.r.t. the measure $\mathbb{P}_X$.

For indicator functions $\varphi = 1_A$, sometimes we make a slight abuse of notation and we set $\mathbb{P}_X^N(A)$ and $\mathbb{P}_X(A)$ instead of $\mathbb{P}_X^N(1_A)$ and $\mathbb{P}_X(1_A)$. From the above discussion, we already have that

$$\mathbb{P}_X^N(A) := \frac{1}{N} \sum_{1 \le i \le N} 1_A(X^i) \longrightarrow_{N \uparrow \infty} \mathbb{P}_X(A) = \mathbb{E}(1_A(X)).$$

The following properties are readily checked

$$\mathbb{E}(\mathbb{P}_X^N(A)) = \mathbb{P}_X(A) \quad \text{and} \quad \mathrm{Var}\left(\mathbb{P}_X^N(A)\right) = \frac{1}{N}\,\mathbb{P}_X(A)\,\left(1-\mathbb{P}_X(A)\right).$$

In addition, an N-approximation of the conditional distribution of X w.r.t. the event $\{X \in A\}$ is given by

$$\frac{1}{\mathbb{P}_X^N(A)}\,1_A(x)\,\mathbb{P}_X^N(dx) \longrightarrow_{N\uparrow\infty} \frac{1}{\mathbb{P}_X(A)}\,1_A(x)\,\mathbb{P}_X(dx) = \mathbb{P}\left(X \in dx \mid X \in A\right). \tag{12}$$

The l.h.s. terms in the above display are well defined as soon as $\mathbb{P}_X^N(A) > 0$. For rare event probabilities $\mathbb{P}_X(A)$, say of order 10^{-6}, the practical implementation of this Monte Carlo algorithm meets the difficulty that we need too many samples to estimate $\mathbb{P}_X(A)$ using the proportion of success of such an event occurring only once in millions of attempts. It is therefore in general not recommended to consider such basic Monte Carlo techniques when studying or estimating the asymptotic risk measures discussed in this paper.

We illustrate this basic Monte Carlo on a standard model in risk and insurance based on the Poisson-Log Normal LDA model of a single risk process. This example, though simple, is both widely utilised in practice and also illustrative of the complementary role of the asymptotic approximations and the role Monte Carlo plays, since this specific model admits a closed form expression for the survival quantile of the annual loss under the first order asymptotic.

Example 1 (Single Risk LDA Poisson-Log-Normal Family). Consider the heavy tailed severity model, selected to model the sequence of i.i.d. losses in each year t, denoted $\{X_i(t)\}_{i=1:N_t}$, and chosen to be a Log-Normal distribution $X_i \sim LN(\mu,\sigma)$ where the two parameters in this model correspond to parametrizing the shape of the distribution for the severity σ and the log-scale of the distribution μ. The survival and quantile functions of the severity are given by

$$\begin{aligned}
f(x;\mu,\sigma) &= \frac{1}{x\sqrt{2\pi\sigma^2}}\,e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}},\ x>0;\ \mu\in\mathbb{R}\ \sigma>0,\\
\overline{F}(x;\mu,\sigma) &= 1-F(x) = \int_x^\infty \frac{1}{\sqrt{2\pi\sigma u}}\exp\left(-\frac{1}{2\sigma^2}\left(\log(u)-\mu^2\right)\right)du\\
&= \frac{1}{2}+\frac{1}{2}\,\mathrm{erf}\Big[\frac{\ln x-\mu}{\sqrt{2\sigma^2}}\Big],\ x>0;\ \mu\in\mathbb{R}\ \sigma>0,\\
Q(p) &= \exp\left(\mu+\sigma\Phi^{-1}(p)\right),\ 0<p<1.
\end{aligned}$$

Therefore the closed form SLA for the VaR risk measure at level α would be presented in this case under a first order approximation for the annual loss $Z = \sum_{n=1}^N X_i$ according to Eq. (13)

$$\mathrm{VaR}_\alpha[Z] = \exp\left[\mu - \sigma\Phi^{-1}\left(\frac{1-\alpha}{\lambda}\right)\right]. \tag{13}$$

We illustrate the basic Monte Carlo solution for the VaR for a range of quantile levels of the annual loss distribution, we display these along with the measured confidence intervals in the point estimators after a long run of 5,000,000 samples of annual years so that the Monte Carlo accuracy was sufficient. We compare these to the first order SLA asymptotic result on the quantile levels $\alpha \in \{0.70, 0.75, 0.80, 0.85, 0.9, 0.95, 0.99, 0.995, 0.9995\}$, where the 99.5 and 99.95 % quantile levels do in fact correspond to regulatory standards of reporting in Basel II/III (Fig. 1).

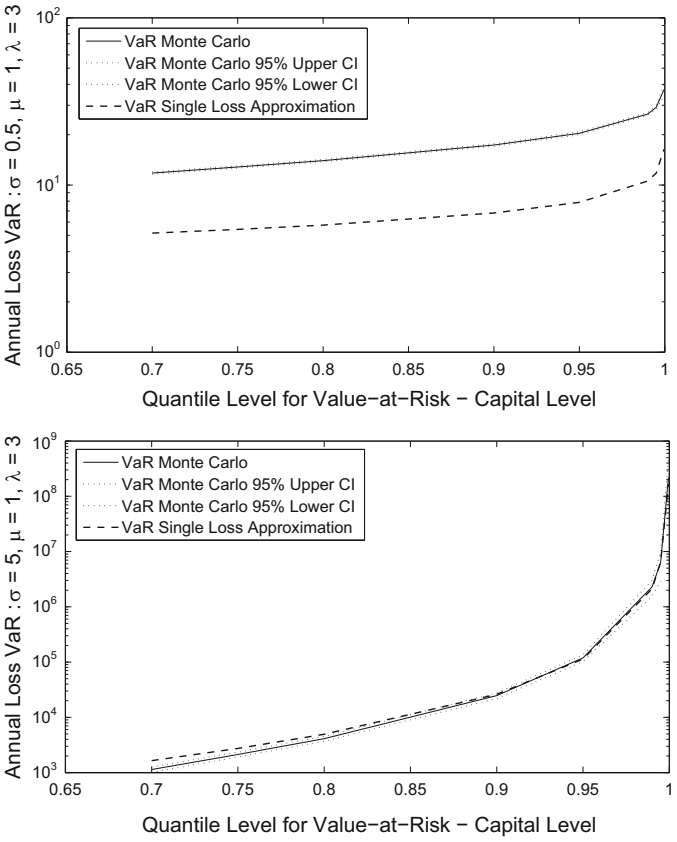


Fig. 1 Annual loss VaR capital estimate versus quantile level for Poisson-Log normal LDA risk process. *Top plot:* severity distribution $\mu = 1, \sigma = 0.5, \lambda = 3$. *Bottom plot:* severity distribution $\mu = 1, \sigma = 5, \lambda = 3$.

This example provides a clear motivation for consideration of particle methods, especially in cases where the SLA results are not accurate. One can see that even

in this relatively simple example, depending on the values of the parameters in the LDA risk model, the asymptotic VaR approximation may or may not be accurate at quantile levels of interest to risk management. Therefore, even small amounts of parameter uncertainty in the LDA model estimation may manifest in significantly different accuracies in the SLA capital estimates. Further examples for higher order asymptotics and other models are provided in [66].

Since the rate of convergence of SLA's is still an active topic of research for such approximations. This simple study illustrates the fact that in practice the only way to ensure accuracy of such methods, for a given set of estimated or specified parameters, is to complement these approximations with a numerical solution such as a Monte Carlo or more efficient interacting particle solution.

3.2 Importance Sampling Techniques for Risk and Insurance

One could argue that the second most widely utilised class of stochastic integration methods considered in risk and insurance settings would be Importance Sampling, see for example [39] and in insurance settings [61]. Here we consider the understanding of these classes of methods in the context of risk and insurance estimation of tail functions. This involves considering undertaking sampling using another random variable for which the occurrence probability of the desired event $\mathbb{P}(Y \in A) := \mathbb{P}_Y(A)$ is closer to 1. This well known importance sampling strategy often gives efficient results for judicious choices of twisted measures $\mathbb{P}_Y$. Nevertheless, in some practical situations, it is impossible to find a judicious $\mathbb{P}_Y$ that achieves a given efficiency. Furthermore, this importance sampling technique is intrusive, in the sense that it requires the user to change the reference statistical or physical model into a twisted sampling rule.

To be more precise, sampling N independent copies $(Y^i)_{1\leq i\leq N}$ with the same dominating probability measure $\mathbb{P}_Y \gg \mathbb{P}_X$, the traditional Monte Carlo approximation is now given by

$$\mathbb{P}_Y^N\left(1_A \frac{d\mathbb{P}_X}{d\mathbb{P}_Y}\right) := \frac{1}{N}\sum_{1\leq i\leq N} 1_A(X^i)\,\frac{d\mathbb{P}_X}{d\mathbb{P}_Y}(Y^i) \longrightarrow_{N\uparrow\infty} \mathbb{P}_Y\left(1_A \frac{d\mathbb{P}_X}{d\mathbb{P}_Y}\right) = \mathbb{P}_X(A).$$

The following properties are readily checked

$$\mathbb{E}\left(\mathbb{P}_Y^N\left(1_A \frac{d\mathbb{P}_X}{d\mathbb{P}_Y}\right)\right) = \mathbb{P}_X(A),$$

and

$$\mathrm{Var}\left(\mathbb{P}_Y^N\left(1_A \frac{d\mathbb{P}_X}{d\mathbb{P}_Y}\right)\right) = \frac{1}{N}\left(\mathbb{P}_X\left(1_A \frac{d\mathbb{P}_X}{d\mathbb{P}_Y}\right) - \mathbb{P}_X(A)^2\right).$$

It is easily checked that

$$\mathbb{P}_Y(dx) = \frac{1}{\mathbb{P}_X(A)}\, 1_A(x)\, \mathbb{P}_X(dx) \Rightarrow \mathrm{Var}\left(\mathbb{P}_Y^N\left(1_A\, \frac{d\mathbb{P}_X}{d\mathbb{P}_Y}\right)\right) = 0.$$

In other words, the optimal twisted measure $\mathbb{P}_Y$ is given by the unknown conditional distribution of X w.r.t. the event $\{X \in A\}$. In practice, we try to find a judicious choice of twisted measure that is easy to sample, with a probability distribution that resembles as closely as possible the desired conditional distribution.

Another approach is to use the occupation measure of a judiciously chosen Markov Chain Monte Carlo (hereafter MCMC) sampler with prescribed target measure

$$\eta(dx) := \mathbb{P}\left(X \in dx \mid X \in A\right).$$

Of course, the first candidate is to take a sequence of independent copies of random variables with common distribution η. Several exact sampling techniques can be used, including the inversion of the repartition function, change of variables principles, the coupling from the past, and acceptance-rejection techniques. A random sample X_i with distribution $\mathbb{P}_X$ is accepted whenever it enters in the desired subset A. In this interpretation, we need to sample N independent copies of X to obtain $\overline{N} := N \times \mathbb{P}_X^N(A)$ independent samples with common law η. However, for probabilities $\mathbb{P}_X(A)$ of order 10^{-6}, this method requires millions of samples, so we consider more computationally efficient solutions.

3.3 Markov Chain Monte Carlo for Risk and Insurance

MCMC samplers have been used in insurance applications in non-life reserving models for example in Chain Ladder models [63, 67, 72] and Paid Incurred Claims models [51] and [59], in Operational Risk models in [62, 65] and in credit risk modelling for example in [49]. Hence, we now present the fundamental mathematical description of the underlying Monte Carlo algorithm that is developed for all the risk and insurance applications discussed in these references.

MCMC algorithms are based on sampling a Markov chain with invariant measure η. In this context, the limiting measure η is often called the target measure. It is not difficult to construct these random processes. For instance, let us assume that the law of X is reversible w.r.t. some Markov transition $K(x, dy)$. In this case, starting from the set A, we sample a sequence of random states using the Markov proposal K, rejecting sequentially all the states falling outside the set A. The algorithm is well defined as soon as $K(x, A) = K(1_A)(x) > 0$, and the resulting Markov chain X_n coincides with the Metropolis-Hasting algorithm with probability transition given by the following formulae

$$M(x, dy) := K(x, dy)\, 1_A(y) + \left(1 - \int K(x, dz)\, 1_A(z)\right) \delta_x(dy).$$

It is not difficult to check that η is an invariant measure of the chain with transition M, that is we have that

$$(\eta M)(dy) := \int \eta(dx)\, M(x, dy) = \eta(dy).$$

Note, the exact acceptance-rejection method discussed above corresponds to the special case

$$K(x, dy) = \mathbb{P}(X \in dy).$$

In more general situations, the proposal transition $K(x, dy)$ amounts of moving randomly around the starting point x. The individual (sometimes also called the walker) makes a number of tentative steps until it succeeds to enter into the desired set A. In general, the random state at that (random) hitting time of A is not distributed according to η. Roughly speaking, when the proposal transition K is based on local moves, the individual tends to hit the set A near the boundary of A. To be more precise, starting from an initial state $X_0 = x \in \mathbb{R}^d - A$, the hitting time

$$T := \inf\{n \geq 0 \,:\, X_n \in A\},$$

is a geometric random variable with distribution

$$\mathbb{P}(T = n \mid X_0 = x) = (1 - K(x, A))^{n-1}\ K(x, A),$$

and we have

$$\mathbb{E}(\varphi(X_T) \mid X_0 = x) = K_A(\varphi)(x) := K(\varphi 1_A)(x)/K(1_A)(x).$$

When the chain enters in A, it remains for all times confined to the set A. In addition, under some weak regularity conditions on the Markov transition K, the target measure η is approximated by the occupation measures of the states; that is, we have the following asymptotic convergence result

$$\frac{1}{n+1} \sum_{0 \leq p \leq n} \delta_{X_p} \longrightarrow_{n \uparrow \infty} \eta \quad \text{and} \quad \mathbb{P}(X_n \in dy \mid X_0 = x) := M^n(x, dy) \longrightarrow_{n \uparrow \infty} \eta(dy). \tag{14}$$

In the above display, $M^n(x, dy)$ stands for the n compositions of the integral operator M defined by the induction formulae

$$M^n(x, dy) = \int M^{n-1}(x, dz) M(z, dy) = \int M(x, dz) M^{n-1}(z, dy),$$

with the convention $M^0(x, dy) = \delta_x(dy)$, for $n = 0$. It is of course out of the scope of this article to prove the ergodic theorem stated in the l.h.s. of (14).

3.4 Sequential Monte Carlo for Risk and Insurance

Application of Sequential Monte Carlo (hereafter SMC) methods in risk and insurance modelling is still relatively underdeveloped, hence the motivation for this article. In the context of risk modelling see the example in [60] and the references therein for more discussion. We start this section with a motivating class of algorithms targeting rare-event simulation via the restriction of a target measure to a contracting, increasingly rare set, such as a tail event.

SMC methods are acceptance-rejection techniques equipped with a recycling mechanism that allows a gradual sampling of a population of individuals w.r.t. a sequence of probabilities with increasing complexity, see a tutorial in [35]. We illustrate this methodology in the situation discussed above. Let us choose a decreasing sequence of subsets $(A_p)_{0\leq p\leq n}$ joining $A_0 = \mathbb{R}^d$ to the desired lower subset $A_n = A$:

$$A_0 = \mathbb{R}^d \supset A_1 \supset A_2 \supset \ldots \supset A_{n-1} \supset A_n = A.$$

Now, let's try to sample sequentially random copies of the random variable X w.r.t the conditioning events $\{X \in A_p\}$, with $p \leq n$. To get one step further, we let η_p be the sequence of measures

$$\eta_p(dy) := \mathbb{P}\left(X \in dx \mid X \in A_p\right) \quad \text{with} \quad p \leq n.$$

By construction, $\left(\eta_p\right)_{0\leq p\leq n}$ is a decreasing sequence of measures w.r.t. the absolutely continuous partial order relation $\mu \ll \nu$ between probability measures[1]; that is, we have that

$$\eta_n \ll \eta_{n-1} \ll \ldots \ll \eta_2 \ll \eta_1 \ll \eta_0 = \text{Law}(X).$$

Example 2 (Single Risk LDA Doubly-Infinitely Divisible Poisson-α-Stable Family). Consider a single risk LDA model, then such a sequence of measures may correspond to construction of a sequence for the annual loss distribution. As an example, consider the sequence given by

$$\eta_n(dz) := F_Z(dz|Z \in A_n), \tag{15}$$

where $A_n = [VaR_Z(\alpha_n), \infty)$ is one set, corresponding to the n-th element in the strictly increasing sequence $(\alpha_p)_{0\geq p\geq n}$ as $\alpha_p \uparrow 1$, which results in a contracting sequence of subsets $A_0 = [0, \infty) \supset A_1 = [VaR_Z(\alpha_1), \infty) \supset \cdots \supset A_n = [VaR_Z(\alpha_n), \infty)$. Given samples from this measure it is then simple to see that one could estimate quantities such as $\overline{F}_Z(VaR_Z(\alpha_n))$ which would be the normalizing constant of this probability distribution when restricted to the set A_n. As an explicit

[1] We recall that $\mu \ll \nu$ as soon as $\nu(A) = 0 \Rightarrow \mu(A) = 0$, for all measurable subset $A \subset \mathbb{R}^d$.

example we consider the α-Stable severity model in a Poisson LDA framework with strictly positive support. Consider the α-Stable severity model with parameters $\alpha \in [0,2]$, $\beta \in [-1,1]$, $\gamma > 0$ and $\delta \in \mathbb{R}$ for the i.i.d. α-Stable distributed random loses with common α. Then w.l.o.g. the density function of an *alpha*-Stable severity distribution (standardized such that $\gamma = 1$ and $\delta = 0$) can be evaluated point-wise according to the series expansions [74, Eq. 2.4.6, page 89]

$$f_X(x;\alpha,\beta,1,0;S(0)) = \begin{cases} \frac{1}{\pi}\sum_{n=1}^{\infty}(-1)^{n-1}\frac{\Gamma(\frac{n}{\alpha}+1)}{\Gamma(n+1)}\sin(n\pi\rho)x^{n-1}, & \text{if } \alpha > 1, \beta \in [-1,1], x \in \mathbb{R}, \\ \frac{1}{\pi}\sum_{n=1}^{\infty}(-1)^{n-1}n b_n x^{n-1}, & \text{if } \alpha = 1, \beta \in (0,1], x \in \mathbb{R}, \\ \frac{1}{\pi}\frac{\Gamma(n\alpha+1)}{\Gamma(n+1)}\sin(n\pi\rho\alpha)x^{-n\alpha-1}, & \text{if } \alpha < 1, \beta \in [-1,1], x \in \mathbb{R}^+, \end{cases} \quad (16)$$

where the coefficients b_n are given by

$$b_n = \frac{1}{\Gamma(n+1)}\int_0^{\infty}\exp(-\beta u \ln u)\, u^{n-1}\sin\left[(1+\beta)u\frac{\pi}{2}\right]du. \quad (17)$$

The resulting LDA model annual loss distribution F_Z is given by

$$Z = \sum_{i=1}^{N} X_i \sim F_{Z_N} = \sum_{n=1}^{\infty}\exp(-\lambda)\frac{\lambda^n}{n!}S_\alpha\left(z;\tilde{\beta}_n,\tilde{\gamma}_n,\tilde{\delta}_n;S(0)\right), \quad (18)$$

where the parameters of each mixture component are analytically known as expressions of the base severity model according to

$$\tilde{\gamma}^\alpha = \sum_{i=1}^{N}|\gamma_i|^\alpha,\ \tilde{\beta} = \frac{\sum_{i=1}^{N}\beta_i|\gamma_i|^\alpha}{\sum_{i=1}^{N}|\gamma_i|^\alpha},$$
$$\tilde{\delta} = \begin{cases} \sum_{i=1}^{N}\delta_i + \tan\frac{\pi\alpha}{2}\left(\tilde{\beta}\tilde{\gamma} - \sum_{i=1}^{N}\beta_j\gamma_j\right) & \text{if } \alpha \neq 1 \\ \sum_{i=1}^{N}\delta_i + \frac{2}{\pi}\left(\tilde{\beta}\tilde{\gamma}\log\tilde{\gamma} - \sum_{i=1}^{N}\beta_j\gamma_j\log|\gamma_i|\right) & \text{if } \alpha = 1. \end{cases} \quad (19)$$

Hence, one observes that as a result of closure under convolution of the α-stable severity model, the resulting distribution for the annual loss can be presented exactly as a mixture representation, see discussions in [58]. Now, consider the Levy sub-family of models in which we consider $X \sim S(0.5,1,\gamma,\delta;S(0))$ with positive real support $x \in [\delta,\infty]$. The density and distribution functions are analytic and given respectively, for $\delta < x < \infty$, by

$$f_X(x) = \sqrt{\frac{\gamma}{2\pi}}\frac{1}{(x-\delta)^{3/2}}\exp\left(-\frac{\gamma}{2(x-\delta)}\right),\quad F_X(x) = \text{erfc}\left(\sqrt{\frac{\gamma}{2(x-\delta)}}\right),$$

where $\text{erfc}(x) = 1 - \text{erf}(x) = 1 - \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$. Under this severity model, the distribution of the annual loss process Z, is represented by a compound process

model with LDA structure in which the frequency is $N \sim Po(\lambda)$ and the severity is $X_i \sim S(0.5, 1, \gamma, \delta; S(0))$. The exact density of the annual loss process can then be expressed analytically as a mixture density comprised of α-stable components with Poisson mixing weights for $N_t^{(j)} > 0$ given by,

$$f_Z(z) = \sum_{n=1}^{\infty} \exp(-\lambda)\frac{\lambda^n}{n!}\left[\sqrt{\frac{\tilde{\gamma}_n}{2\pi}}\frac{1}{\left(z-\tilde{\delta}_n\right)^{3/2}}\exp\left(-\frac{\tilde{\gamma}_n}{2\left(z-\tilde{\delta}_n\right)}\right)\right]\mathbb{I}\left[\tilde{\delta}_n < z < \infty\right] \tag{20}$$

with $\widetilde{\beta}_n = 1$ and

$$\widetilde{\gamma}_n^{0.5} = \sum_{i=1}^{n}|\gamma_i|^{0.5} = n|\gamma|^{0.5},\ \ \widetilde{\delta}_n = \sum_{i=1}^{n}\delta_i + \tan\frac{\pi}{4}\left(\tilde{\gamma}_n - \sum_{j=1}^{n}\gamma_j\right) = n\delta + \tan\frac{\pi}{4}\left(n^2|\gamma| - n\gamma\right),$$

and $f_Z(0) = \Pr(N_t^{(j)} = 0) = \exp(-\lambda)$ for $N = 0$. The exact form of the annual loss cumulative distribution function is also expressible in closed-form,

$$F_Z(z) = \sum_{n=1}^{\infty}\exp(-\lambda)\frac{\lambda^n}{n!}\text{erfc}\left(\sqrt{\frac{\tilde{\gamma}_n}{2\left(z-\tilde{\delta}_n\right)}}\right)\mathbb{I}\left[\tilde{\delta}_n < z < \infty\right] + \exp(-\lambda)\mathbb{I}[z=0]. \tag{21}$$

Now, given these expressions, we note that the simple existence of a closed form expression for the compound process distribution does not make it simple to sample the distribution under the restriction to some set A_n, as may be required for certain tail functional estimates. Therefore, we are still required to consider a sequence of measures, which in this particular example may be defined by the restriction of the annual loss to a decreasing tail set $A_n = [VaR_Z(\alpha_n), \infty)$ as $\alpha_n \uparrow 1$. Hence, the resulting sequence of target measures for the annual loss distribution is known explicitly in a functional closed form according to

$$\begin{aligned}\eta_k(dz) &:= F_Z(dz|Z \in A_k) = \overline{F}_Z(A_k) = 1 - F_Z(A_k)\\ &= 1 - \sum_{n=1}^{\infty}\exp(-\lambda)\frac{\lambda^n}{n!}\text{erfc}\left(\sqrt{\frac{\tilde{\gamma}_n}{2\left(dz-\tilde{\delta}_n\right)}}\right)\mathbb{I}\left[\left\{\tilde{\delta}_n < dz < \infty\right\}\cap\{dz \in A_k\}\right]\\ &\quad - \exp(-\lambda)\mathbb{I}[dz=0].\end{aligned}$$

This example is just one of many illustrations one could construct in order to demonstrate the possible sequences of distributions of relevance in risk modelling that will naturally fit into such Monte Carlo frameworks.

3.4.1 A Markov Chain Monte Carlo Model

We assume that we have a dedicated MCMC style probability transition M_p with invariant measure $\eta_p = \eta_p M_p$, for any $p \leq n$. We start drawing a sequence of random states $(X_p)_{0\leq p\leq n_1}$ with transitions M_1, and initial condition η_0. For a sufficiently large time horizon n_1, both the occupation measure $\frac{1}{n_1}\sum_{1\leq p\leq n_1}\delta_{X_p}$ and the law of the terminal state $\text{Law}(X_{n_1}) = \eta_0 M_1^{n_1} := \pi_1$ will approximate the target measure η_1. We also notice that the chain $(X_p)_{p_1\leq p\leq n_1}$ is confined to the set A_1 as soon as one of the random states $X_{p_1} \in A_1$ hits the set A_1 for some $p_1 \leq n_1$.

In the second step, starting from X_{n_1} we run a sequence of random states $(X_{n_1+p})_{0\leq p\leq n_2}$ with transitions M_2 (and initial condition π_1). For a sufficiently large time horizon n_2, both the occupation measure $\frac{1}{n_2}\sum_{1\leq p\leq n_1}\delta_{X_{n_1+p}}$ and the law of the terminal state $\text{Law}(X_{n_1+n_2}) = \pi_1 M_2^{n_2}$ will now approximate the target measure η_2. As before, the chain $(X_{n_1+p})_{p_2\leq p\leq n_2}$ is confined to the set A_2 as soon as one of the random states $X_{n_1+p_2} \in A_2$ hits the set A_2 for some $p_2 \leq n_2$,

$$\eta_0 \xrightarrow{M_1^{n_1}} \eta_0 M_1^{n_1} := \pi_1 \xrightarrow{M_2^{n_2}} \pi_1 M_2^{n_2} = \pi_2 \xrightarrow{M_3^{n_3}} \pi_2 M_3^{n_3} = \pi_3 \ldots . \quad (22)$$

3.4.2 An Acceptance-Rejection Markov Model

Our next objective is to better understand the evolution of the flow of measures η_p, from the origin $p = 0$ up to the final time $p = n$. Firstly, it is readily checked that

$$\mathbb{P}\left(X \in dx \mid X \in A_{p+1}\right) = \frac{1}{\mathbb{P}\left(X \in A_{p+1} \mid X \in A_p\right)}\, 1_{A_{p+1}}(x)\, \mathbb{P}\left(X \in dx \mid X \in A_p\right)$$

and

$$\mathbb{P}\left(X \in A_{p+1} \mid X \in A_p\right) = \int 1_{A_{p+1}}(x)\, \mathbb{P}\left(X \in dx \mid X \in A_p\right).$$

Therefore, if we specifically set $G_p(x) = 1_{A_{p+1}}(x)$, then we have that

$$\eta_{p+1} = \Psi_{G_p}(\eta_p)$$

with the Boltzmann-Gibbs Ψ_{G_p} transformation defined by:

$$\eta_p(dx) \longrightarrow \Psi_{G_p}(\eta_p)(dx) := \frac{1}{\eta_p(G_p)}\, G_p(x)\, \eta_p(dx).$$

The next formula provides an interpretation of the Boltzmann-Gibbs transformation in terms of a non-linear Markov transport equation

$$\Psi_{G_p}(\eta_p)(dy) = \left(\eta_p S_{p,\eta}\right)(dy) := \int \eta_p(dx) S_{p,\eta_p}(x, dy)$$

with the Markov transition S_{p,η_p} defined below

$$S_{p,\eta_p}(x,dy) = G_p(x)\ \delta_x(dy) + \big(1 - G_p(x)\big)\ \Psi_{G_p}(\eta_p)(dy).$$

In summary, we have shown that $(\eta_p)_{0\leq p\leq n}$ satisfies the following evolution equation

$$\eta_0 \xrightarrow{S_{0,\eta_0}} \eta_1 \xrightarrow{S_{1,\eta_1}} \eta_2 \xrightarrow{S_{2,\eta_2}} \eta_3 \xrightarrow{S_{3,\eta_3}} \eta_4 \ldots$$

In other words, $\eta_p = \text{Law}(X_p^\star)$ can be interpreted as the law of the random state of a Markov chain $X_p^\star$ with transitions S_{p,η_p}; that is, we have that

$$X_0^\star \xrightarrow{S_{0,\eta_0}} X_1^\star \xrightarrow{S_{1,\eta_1}} X_2^\star \xrightarrow{S_{2,\eta_2}} X_3^\star \xrightarrow{S_{3,\eta_3}} \ldots.$$

The Markov chain $X_p^\star$ can be interpreted as the optimal sequential acceptance-rejection scheme along the non-increasing sequence of subsets A_p, in the sense that

$$\left\{ \begin{array}{c} X_p^\star \in A_{p+1} \Rightarrow X_{p+1}^\star = X_p^\star, \\ X_p^\star \in A_p - A_{p+1} \Rightarrow X_{p+1}^\star = X_{p+1}^{\star\star}, \end{array} \right.$$

where $X_{p+1}^{\star\star}$ stands for an independent random sample with distribution $\eta_{p+1} = \Psi_{G_p}(\eta_p)$. When the sample $X_p^\star$ is not in the desired subset A_{p+1}, it jumps instantly to a new state $X_{p+1}^{\star\star}$ randomly chosen with the desired distribution $\eta_{p+1} = \Psi_{G_p}(\eta_p)$. Next we provide a brief discussion on the optimality property of this Markov chain model. We recall that

$$\begin{aligned} &\|\eta_{p+1} - \eta_p\|_{tv} = \sup\big\{[\eta_{p+1} - \eta_p](f)\ :\ \text{osc}(f) \leq 1\big\} \\ &= \inf\big\{\mathbb{P}\left(X_p \neq X_{p+1}\right)\ :\ (X_p, X_{p+1})\ \text{s.t.}\ \text{Law}(X_p) = \eta_p\ \text{ and }\ \text{Law}(X_{p+1}) = \eta_{p+1}\big\}. \end{aligned}$$

In the above display $\text{osc}(\varphi) = \sup_{x,y}(|\varphi(x) - \varphi(y)|)$ stands for the oscillation of a given function φ on $\mathbb{R}^d$.

Proposition 1. *The chain $X_p^\star$ with Markov transitions S_{p,η_p} realizes the optimal coupling between the sequence of distributions η_p, in the sense that*

$$\|\eta_{p+1} - \eta_p\|_{tv} = \|\eta_p S_{p,\eta_p} - \eta_p\|_{tv} = \mathbb{P}\left(X_p^\star \neq X_{p+1}^\star\right). \tag{23}$$

A proof of this assertion is provided in the appendix.

3.4.3 Feynman-Kac Distributions

As the reader may have noticed, the MCMC model and the acceptance-rejection Markov chain models discussed may have very poor stability properties, in the

sense that the distributions of the random states may strongly depend on the initial distribution η_0. For instance, we notice that η_p coincides with the restriction of η_0 to the subset A_p; more formally, we have that

$$\eta_p(dx) = \Psi_{G_{p-1}}(\eta_0) = \frac{1}{\eta_0(A_p)}\, 1_{A_p}(x)\, \eta_0(dx).$$

The sequential Monte Carlo methodology is based on combining the MCMC methodology presented in (22) with the sequential acceptance-rejection technique discussed above. To describe with some precision this method, we let M_p be an MCMC transition with invariant measure $\eta_p = \eta_p M_p$. In this case, we have the evolution equation

$$\eta_{p+1} = \eta_{p+1} M_{p+1} = \Psi_{G_p}(\eta_p) M_{p+1} := \Phi_{p+1}(\eta_p).$$

Notice that Φ_{p+1} maps the set of probability measures η s.t. $\eta(G_p) > 0$ into the set of probability measures, and it is the composition of an updating transformation Ψ_{G_p} and a Markov transport equation w.r.t. M_{p+1}; that is, we have that

$$\eta_p \xrightarrow{\Psi_{G_p}} \hat{\eta}_p := \Psi_{G_p}(\eta_p) \xrightarrow{M_{p+1}} \hat{\eta}_p M_{p+1} = \Phi_{p+1}(\eta_p).$$

The solution of this equation is given by the Feynman-Kac measures defined for any measurable function φ on $\mathbb{R}^d$ by the following formulae

$$\eta_p(\varphi) = \gamma_p(\varphi)/\gamma_p(1) \quad \text{with} \quad \gamma_p(\varphi) = \mathbb{E}\left(\varphi(X_p) \prod_{0\le q<p} G_q(X_q)\right). \tag{24}$$

To prove this claim, we use the Markov property to check that

$$\gamma_{p+1}(\varphi) = \mathbb{E}\left(M_{p+1}(\varphi)(X_p)\, G_p(X_p) \prod_{0\le q<p} G_q(X_q)\right) = \gamma_p(G_p M_{p+1}(\varphi)).$$

This clearly implies that

$$\eta_{p+1}(\varphi) = \frac{\gamma_p(G_p M_{p+1}(\varphi))/\gamma_p(1)}{\gamma_p(G_p)/\gamma_p(1)} = \frac{\eta_p(G_p M_{p+1}(\varphi))}{\eta_p(G_p)} = \Psi_{G_p}(\eta_p) M_{p+1}(\varphi).$$

We already mention that the unnormalized measures γ_n can be expressed in terms of the flow of measures $(\eta_p)_{0\le p\le n}$ with the following multiplicative formulae

$$\gamma_p(\varphi) = \eta_p(\varphi) \times \prod_{0\le q<p} \eta_q(G_q). \tag{25}$$

This result is a direct consequence of the following observation

$$\gamma_p(1) = \mathbb{E}\left(G_{p-1}(X_{p-1}) \prod_{0\le q<p-1} G_q(X_q) \right) = \gamma_{p-1}(G_{p-1}) = \eta_{p-1}(G_{p-1})\ \gamma_{p-1}(1).$$

It is readily checked that the measures η_n are the n-th time marginals of the Feynman-Kac measures on the path space defined by the following formulae

$$d\mathbb{Q}_n := \frac{1}{\mathscr{Z}_n} \left\{ \prod_{0\le p<n} G_p(X_p) \right\} d\mathbb{P}_n \tag{26}$$

with some normalizing constants $\mathscr{Z}_n = \gamma_n(1)$ and the reference measures

$$\mathbb{P}_n = \text{Law}(X_{0:n}) \quad \text{with} \quad X_{0:n} := (X_0, \ldots, X_n).$$

This class of path space measures goes beyond the MCMC model discussed above. These measures represent the distribution of the trajectories of a reference Markov process, weighted by a collection of potential functions. *These functional models are natural mathematical extensions of the traditional change of probability measures, commonly used in importance sampling.*

From a purely probabilistic viewpoint, these measures can be interpreted as the conditional distribution of a given Markov chain w.r.t. to a sequence of events. For instance, if we take $G_n = 1_{A_n}$ indicator potential functions of some measurable subsets $A_n \in E_n$, then it can be readily checked that

$$\mathbb{Q}_n = \text{Law}(X_{0:n} \mid \forall 0 \le p < n \ \ X_p \in A_p) \ \text{ and } \ \mathscr{Z}_n = \mathbb{P}\left(\forall 0 \le p < n \ \ X_p \in A_p\right).$$

For a thorough discussion on the application domains of these Feynman-Kac models, we refer the reader to the books [13, 19, 22, 34].

Example 3 (Multiple LDA Risk Conditional Tail Expectations). Consider the class of problems in risk management involving the evaluation of a coherent capital allocation. We consider the $X \in E$ to be a random vector $X = \left[Z^{(1)}, Z^{(1)}, \ldots, Z^{(d)}\right]$ for d LDA structured risk processes, with the space on which this random vector is defined given by $E = [0, \infty)^d$. In this case we can consider the multi-variate distribution for the d risk processes, for which we can consider dependence if required, according to

$$\begin{aligned}\mathbb{Q}_n &:= \text{Law}(X_{0:n} \mid \forall 0 \le p < n \ \ X_p \in A_p) \\ &= F(Z_0^{(1)}, \ldots, Z_0^{(d)}, Z_1^{(1)}, \ldots, Z_n^{(d)} \mid \forall 0 \le p < n \ \ X_p \in A_p).\end{aligned}$$

If one considers the event $X_p \in A_p$ as corresponding to the sequence of multivariate loss draws that produce the rare-event that the total loss $Z_T = \sum_{i=1}^d Z^{(i)}$ gives $Z_T \in (\text{VaR}_{Z_T}(\alpha) - \epsilon_n, \text{VaR}_{Z_T}(\alpha) + \epsilon_n)$ for some $\epsilon_n \downarrow 0$, then one has a

mechanism for calculating conditional tail expectations, relevant to assessing multivariate risk measures, tail dependence and capital allocation problems.

3.5 Non-linear McKean Markov Chains

The central idea behind Feynman-Kac particle samplers is to observe that *any* evolution equation of probability measures

$$\eta_n = \Phi_n(\eta_{n-1})$$

on some measurable state spaces E_n can be interpreted as the law

$$\eta_n = \text{Law}\left(\overline{X}_n\right)$$

of a Markov chain $\overline{X}_n$ with initial distribution η_0 and Markov transitions

$$\mathbb{P}\left(\overline{X}_n \in dx_n \mid \overline{X}_{n-1} = x_{n-1}\right) = K_{n,\eta_{n-1}}(x_{n-1}, dx_n).$$

The Markov transitions $K_{n,\eta_{n-1}}$ are chosen so that

$$\forall n \geq 1, \qquad \eta_{n-1}K_{n,\eta_{n-1}} = \Phi_n(\eta_{n-1}).$$

The Markov chain $\overline{X}_n$ incorporate free evolution moves according to M_n, with sequential updates of the measures η_n, so that the law of the random states $\overline{X}_n$ coincide with the desired distributions η_n, at every time step. This chain can be interpreted as a perfect sequential sampler of the sequence of measures η_n.

The choice of the transitions K_{n+1,η_n} is not unique. For instance, for the Feynman-Kac models on $E_n = \mathbb{R}^d$ discussed above, if we take

$$K_{n+1,\eta_n}(x, dy) := \left[S_{n,\eta_n}M_{n+1}\right](x, dy) \quad \text{or} \quad K_{n+1,\eta_n}(x, dy) := \Phi_{n+1}(\eta_n)(dy)$$

we readily check that

$$\eta_n K_{n+1,\eta_n} = \Phi_{n+1}(\eta_n) = \Psi_{G_n}(\eta_n)M_{n+1} = \eta_n S_{n,\eta_n}M_{n+1}.$$

We also mention that the law of the random trajectories $(\overline{X}_0, \ldots, \overline{X}_n)$ are given by the so-called McKean measures

$$\overline{\mathbb{P}}_n(dx_{0:n}) = \eta_0(dx_0)\, K_{1,\eta_0}(x_0, dx_1) \ldots K_{n,\eta_{n-1}}(x_{n-1}, dx_n),$$

where $dx_{0:n} = d(x_0, \ldots, x_n)$ stands for an infinitesimal neighbourhood of the trajectory $x_{0:n} := (x_0, \ldots, x_n)$.

We further assume that the Markov transitions $M_n(x_{n-1}, dx_n)$ are absolutely continuous with respect to some reference measure ν_n and we set

$$Q_n(x_{n-1}, dx_n) := G_{n-1}(x_{n-1}) M_n(x_{n-1}, dx_n) = H_n(x_{n-1}, x_n)\ \nu_n(dx_n).$$

In this situation, we have the following time reversal formulae

$$\mathbb{Q}_n(dx_{0:n}) = \eta_n(dx_n)\ \mathbb{M}_{n,\eta_{n-1}}(x_n, dx_{n-1}) \ldots \mathbb{M}_{1,\eta_0}(x_1, dx_0), \tag{27}$$

with the Markov transitions

$$\mathbb{M}_{n,\eta_{n-1}}(x_n, dx_{n-1}) := \frac{\eta_{n-1}(dx_{n-1})\ H_n(x_{n-1}, x_n)}{\eta_{n-1}\left(H_n(\bullet, x_n)\right)}.$$

We prove this backward formula using the fact that

$$\eta_n(dx_n) = \Psi_{G_{n-1}}(\eta_{n-1}) M_n(dx_n) = \frac{\eta_{n-1}\left(H_n(\bullet, x_n)\right)}{\eta_{n-1}(G_{n-1})}\ \nu_n(dx_n),$$

from which we find that

$$\eta_n(dx_n)\ \mathbb{M}_{n,\eta_{n-1}}(x_n, dx_{n-1}) = \frac{1}{\eta_{n-1}(G_{n-1})}\ \eta_{n-1}(dx_{n-1})\ Q_n(x_{n-1}, dx_n).$$

Iterating this process, we prove (27).

3.5.1 Mean Field Particle Simulation

This section is concerned with particle approximations of the Feynman-Kac model (24) and (26). We also present a series of exponential concentration inequalities that allows one to estimate the deviation of the particle estimates around their limiting values.

In the remainder of this section φ_n stands for some function s.t. $\|\varphi_n\| \leq 1$, and (c_1, c_2) represent two constants related to the bias and the variance of the particle approximation scheme, and c stands for some universal constant. The values of these constants may vary from line to line but they don't depend on the time horizon. Furthermore, we assume that the Feynman-Kac model satisfies some strong stability properties. For a more detailed description of the stability properties, and the description of the quantities (c, c_1, c_2) in terms of the Feynman-Kac model (24), we refer the reader to the books [19, 22].

We approximate the transitions

$$\overline{X}_n \rightsquigarrow \overline{X}_{n+1} \sim K_{n+1,\eta_n}(\overline{X}_n, dx_{n+1}),$$

by running a Markov chain $\xi_n = (\xi_n^1, \ldots, \xi_n^N) \in E_n^N$ that approximate the distribution η_n when $N \uparrow \infty$

$$\frac{1}{N} \sum_{1 \leq i \leq N} \delta_{\xi_n^i} := \eta_n^N \longrightarrow_{N \uparrow \infty} \eta_n.$$

A natural choice of particle transitions is to take at every time step a sequence of conditionally independent particles

$$\xi_n^i \rightsquigarrow \xi_{n+1}^i \sim K_{n+1,\eta_n^N}(\xi_n^i, dx_{n+1}).$$

For the Feynman-Kac models discussed above, we can chose the transitions $K_{n+1,\eta_n} = S_{n,\eta_n} M_{n+1}$. In this context, the evolution of the particle algorithm is decomposed into two steps:

$$\begin{bmatrix} \xi_n^1 \\ \vdots \\ \xi_n^i \\ \vdots \\ \xi_n^N \end{bmatrix} \xrightarrow{S_{G_n,\eta_n^N}} \begin{bmatrix} \hat{\xi}_n^1 & \xrightarrow{M_{n+1}} & \xi_{n+1}^1 \\ \vdots & & \vdots \\ \hat{\xi}_n^i & \longrightarrow & \xi_{n+1}^i \\ \vdots & & \vdots \\ \hat{\xi}_n^N & \longrightarrow & \xi_{n+1}^N \end{bmatrix}.$$

During the first step, every particle ξ_n^i evolves to a new particle $\hat{\xi}_n^i$ randomly chosen with the distribution

$$S_{\eta_n^N}(\xi_n^i, dx) := G_n(\xi_n^i)\, \delta_{\xi_n^i}(dx) + \left(1 - G_n(\xi_n^i)\right)\, \Psi_{G_n}(\eta_n^N)(dx),$$

with the updated measures

$$\Psi_{G_n}(\eta_n^N) = \sum_{j=1}^N \frac{G_n(\xi_n^j)}{\sum_{k=1}^N G_n(\xi_n^k)} \delta_{\xi_n^j} \longrightarrow_{N \uparrow \infty} \Psi_{G_n}(\eta_n) = \eta_{n+1}.$$

This transition can be interpreted as an acceptance-rejection scheme with a recycling mechanism. In the second step, the selected particles $\hat{\xi}_n^i$ evolve randomly according to the Markov transitions M_{n+1}. In other words, for any $1 \leq i \leq N$, we sample a random state ξ_{n+1}^i with distribution $M_{n+1}(\hat{\xi}_n^i, dx)$.

3.6 A Sequential Monte Carlo Formulation

Most of the SMC technology developed for Bayesian inference, is based on finding judicious sequential importance sampling representations of a given sequence of target measures, on some general state space models defined on E_n. More precisely, let us suppose that we are given a sequence of target measures of the following form

$$\mathbb{Q}_n(dx_{0:n}) \propto \mathbb{Q}_{n-1}(dx_{0:n-1}) \times Q_n(x_{n-1}, dx_n), \tag{28}$$

for some bounded positive integral operators $Q_n(x_{n-1}, dx_n)$ from $\mathscr{B}_b(E_n)$ into $\mathscr{B}_b(E_{n-1})$. By construction, we observe that these target measures can alternatively be defined by

$$\mathbb{Q}_n(dx_{0:n}) := \frac{1}{\mathscr{Z}_n}\, \eta_0(dx_0) Q_1(x_0, dx_1) \ldots Q_n(x_{n-1}, dx_n),$$

for some normalizing constant $\mathscr{Z}_n$. Given a sequence of importance sampling transition M_{n+1} s.t.

$$Q_{n+1}(x_n, \bullet) \ll M_{n+1}(x_n, \bullet),$$

for any $x_n \in E_n$ we denote by W_n the sequential importance weights

$$\begin{aligned} W_n(x_n, x_{n+1}) &\propto \frac{\text{Target at time (n+1)}}{\text{Target at time (n)} \times \text{ Twisted transition}} \\ &\propto \frac{\mathbb{Q}_{n+1}(dx_{0:n+1}))}{\mathbb{Q}_n(dx_{0:n}) \times M_{n+1}(x_n, dx_{n+1})} := \frac{dQ_{n+1}(x_n, \bullet)}{dM_{n+1}(x_n, \bullet)}(x_{n+1}). \end{aligned} \tag{29}$$

The corresponding change of measure has the following form

$$\mathbb{Q}_n(dx_{0:n}) = \frac{1}{\mathscr{Z}_n} \left\{ \prod_{0 \le p < n} W_p(x_p, x_{p+1}) \right\} \mathbb{P}_n(dx_{0:n}). \tag{30}$$

We consider the Markov chain on the transition space defined by

$$\mathbf{X_n} := (X_n, X_{n+1}) \in \mathbf{E_n} = (E_n \times E_{n+1}).$$

In this notation, for any bounded measurable function φ_n on the product state space $(E_0 \times \ldots \times E_n)$, we have the following importance sampling formulae

$$\mathbb{E}\left(\varphi_n(X_{0:n}) \prod_{0 \le p < n} W_p(X_p, X_{p+1})\right) = \mathbb{E}\left(\boldsymbol{\varphi_n(X_{0:n})} \prod_{0 \le p < n} \mathbf{G_p(X_p)}\right)$$

with the functions

$$\boldsymbol{\varphi_n(X_{0:n})} = \varphi_n(X_{0:n}), \quad \text{and the potential functions} \quad \mathbf{G_p} := W_p\,.$$

This implies that

$$\mathbb{Q}_n(\varphi_n) = \mathbf{Q_n}(\mathbf{'_n}), \tag{31}$$

with the Feynman-Kac measure $\mathbf{Q_n}$ associated with the Markov chain $\mathbf{X_n}$ on the transition space $\mathbf{E_n} = (E_n \times E_{n+1})$ and the potential functions $\mathbf{G_n}$. In this formulation, sequential Monte Carlo samplers coincide with the mean field particle interpretations discussed in Sect. 3.5.1.

3.6.1 Some Non-asymptotic Estimates: Finite Sample Accuracy for Particle Integration

The exponential concentration inequalities developed below are satisfied under some regularity conditions on the Feynman-Kac parameters (G_n, M_n), on some general state space models defined on E_n. It is clearly out of the scope to present here all the details of the proof of these inequalities. As shown in Sect. 3.6, the importance sampling Feynman-Kac representation of a given sequence of target measures is far from unique. Roughly speaking, the twisted transitions M_n and the corresponding potential weight functions G_n have to be chosen so that the non-linear semi-group associated with evolution equation

$$\eta_n = \Psi_{G_{n-1}}(\eta_{n-1})M_n,$$

of the n-th time marginals η_n of the Feynman-Kac target measures $\mathbb{Q}_n$ are sufficiently stable. One way to satisfy this stability property is to choose sufficiently mixing twisted transitions, with bounded relative oscillations of the weight functions. For a more thorough discussion on these stability conditions, we refer the reader to [19, 20, 22].

We note that the exponential concentration inequalities presented below are also valid for non necessarily stable Feynman-Kac semi-groups. Nevertheless, in this degenerate situation the constants c and (c_1, c_2) depend on the time parameter. Using the concentration analysis of mean field particle models developed in [32], the following exponential estimate was proved in [22]. For any $x \geq 0$, $n \geq 0$, and any population size $N \geq 1$, the probability of the event

$$\left[\eta_n^N - \eta_n\right](\varphi) \leq \frac{c_1}{N}\,\left(1 + x + \sqrt{x}\right) + \frac{c_2}{\sqrt{N}}\,\sqrt{x},$$

is greater than $1 - e^{-x}$.

3.6.2 Non-asymptotic Estimates for Risk Measure Estimation via Interacting Particle Systems

In addition, for any $x = (x_i)_{1\leq i\leq d}$ and any $(-\infty, x] = \prod_{i=1}^{d}(-\infty, x_i]$ cells in $E_n = \mathbb{R}^d$, we let

$$F_n(x) = \eta_n\left(1_{(-\infty,x]}\right) \quad \text{and} \quad F_n^N(x) = \eta_n^N\left(1_{(-\infty,x]}\right).$$

For any $y \geq 0$, $n \geq 0$, and any population size $N \geq 1$, the probability of the following event

$$\sqrt{N}\ \left\|F_n^N - F_n\right\| \leq c\ \sqrt{d\ (y+1)},$$

is greater than $1 - e^{-y}$. This concentration inequality ensures that the particle repartition function F_n^N converges to F_n, almost surely for the uniform norm. For $d = 1$, we let $F_n^{\leftarrow}$ be the generalized inverse on $[0, 1]$ of the function F_n; that is, we have that

$$F_n^{\leftarrow}(\alpha) := \inf\{x \in \mathbb{R}\ :\ F_n(x) \geq \alpha\}.$$

We let $F_n^{\leftarrow}(\alpha) = q_{n,\alpha}$ be the quantile, of order α, and we denote by ζ_n^i the order particle statistic associated with the particle system ξ_n^i at time n; that is, we have that

$$\zeta_n^1 := \xi_n^{\sigma(1)} \leq \zeta_n^2 := \xi_n^{\sigma(2)} \leq \ldots \leq \zeta_n^N := \xi_n^{\sigma(N)},$$

for some random permutation σ. We also denote by $q_{n,\alpha}^N := \zeta_n^{1+\lfloor N\alpha\rfloor}$ the α-particle quantile. By construction, we have that

$$\begin{aligned}\left|F_n\left(q_{n,\alpha}^N\right) - F_n(q_{n,\alpha})\right| &\leq \left|F_n\left(q_{n,\alpha}^N\right) - F_n^N(q_{n,\alpha}^N)\right| + \left|F_n^N(q_{n,\alpha}^N) - \alpha\right| \\ &\leq \left\|F_n^N - F_n\right\| + \left(\frac{1+\lfloor N\alpha\rfloor}{N} - \alpha\right) \leq \left\|F_n^N - F_n\right\| + 1/N.\end{aligned}$$

This clearly implies that $q_{n,\alpha}^N$ converges almost surely to $q_{n,\alpha}$, as N tends to ∞. In addition, for any $y \geq 0$, $n \geq 0$, and any population size $N \geq 1$, the probability of the following event

$$\sqrt{N}\ \left|F_n\left(q_{n,\alpha}^N\right) - \alpha\right| \leq c\ \sqrt{d\ (y+1)} + \frac{1}{\sqrt{N}},$$

is greater than $1 - e^{-y}$.

If we interpret the mutation-selection particle algorithm as a birth and death branching process, then we can trace back in time the whole ancestral line $\gamma_n^i = (\xi_{p,n}^i)_{0\leq p\leq n}$ of the individual ξ_n^i at the n-th generation

$$\xi_{0,n}^i \longleftarrow \xi_{1,n}^i \longleftarrow \ldots \longleftarrow \xi_{n-1,n}^i \longleftarrow \xi_{n,n}^i = \xi_n^i.$$

The random state $\xi_{p,n}^i$ represents the ancestor of the individual ξ_n^i at the level p, with $0 \leq p \leq n$, and $1 \leq i \leq N$. It is more or less well known that γ_n coincides with the particle approximation of the Feynman-Kac model defined in (24) by replacing X_n by the historical process $(X_p)_{0\leq p\leq n}$. This interpretation provides an alternative particle approximation scheme of the measures (26), that is we have that

$$\eta_n^N = \frac{1}{N} \sum_{1\leq i\leq N} \delta_{(\xi_{0,n}^i, \xi_{1,n}^i, \ldots, \xi_{n,n}^i)} \cdot \longrightarrow_{N\uparrow\infty} \mathbb{Q}_n.$$

More precisely, we proved in [22] the following exponential concentration estimate. For any test function $\boldsymbol{\varphi_n}$ on path space s.t. $\|\boldsymbol{\varphi_n}\| \leq 1$, for any $y \geq 0$, $n \geq 0$, and any $N \geq 1$, the probability of the event

$$\left[\eta_n^N - \mathbb{Q}_n\right](\varphi) \leq c_1 \frac{n+1}{N} \left(1 + x + \sqrt{x}\right) + c_2 \sqrt{\frac{(n+1)}{N}} \sqrt{x},$$

is greater than $1 - e^{-x}$.

Further details on these genealogical tree models can be found in [19, 22, 25]. Mimicking formulae (25) and (27), we define *an unbiased* particle estimate γ_n^N of the unnormalized measures γ_n and a particle backward measures $\mathbb{Q}_n^N$ by setting

$$\gamma_n^N(\varphi) = \eta_n^N(\varphi) \times \prod_{0\leq q<n} \eta_q^N(G_q),$$

and

$$\mathbb{Q}_n^N(d(x_0, \ldots, x_n)) = \eta_n^N(dx_n)\, \mathbb{M}_{n,\eta_{n-1}^N}(x_n, dx_{n-1}) \ldots \mathbb{M}_{1,\eta_0^N}(x_1, dx_0).$$

We end this section with a couple of exponential concentration estimates proved in [22]. For any $x \geq 0$, $n \geq 0$, $N \geq 1$, and any $\epsilon \in \{+1, -1\}$, the probability of the event

$$\frac{\epsilon}{n} \log \frac{\gamma_n^N(1)}{\gamma_n(1)} \leq \frac{c_1}{N} \left(1 + x + \sqrt{x}\right) + \frac{c_2}{\sqrt{N}} \sqrt{x},$$

is greater than $1 - e^{-x}$. In addition, for any normalized additive functional

$$\boldsymbol{\varphi_n}(x_0, \ldots, x_n) = \frac{1}{n+1} \sum_{0\leq p\leq n} \varphi_p(x_p)$$

with $\|\varphi_p\| \leq 1$, for $x \geq 0$, $n \geq 0$, and any population size $N \geq 1$, the probability of the event

$$\left[\mathbb{Q}_n^N - \mathbb{Q}_n\right](\overline{\varphi}_n) \leq c_1 \frac{1}{N}(1 + (x + \sqrt{x})) + c_2 \sqrt{\frac{x}{N(n+1)}},$$

is greater than $1 - e^{-x}$.

4 Illustration of Interacting Particle Solutions for Risk and Insurance Capital Estimation

In this section we detail a special subset of algorithms, from within the stochastic particle integration methods, that were specifically developed to solve problems for risk and insurance in [61]. The class of recursive solutions developed is applicable to a wide range of insurance and risk settings. We provide a novel result in this illustration which extends the framework originally presented in [61] through consideration of a higher-order Panjer recursion whilst avoiding the need to perform discretisation of the severity distribution. We shall present a generic version of this approach which adopts an interacting particle solution. In addition, we illustrate how this method may be used in inference for tail quantiles of compound processes to complement the results considered for the SLA approximations.

4.1 Recursions for Loss Distributions: Panjer and Beyond

We extend the framework proposed in [61] for developing a recursive numerical solution to estimation of such risk measures through estimation of the density of the compound process. In particular, we briefly summarize an approach to transform the standard actuarial solution known as the Panjer recursion [56] to a sequence of expectations. We note that recursions for the evaluation of single risk process distributions, under discretisation, are ubiquitous in risk and insurance modelling, see discussions in [71]. We consider an advanced development that avoids the need to discretise the severity distribution via development of a stochastic particle integration based solution.

Consider the actuarial recursions for evaluating $\overline{F}_Z(x)$ based around the Panjer class of frequency distribution relationships defined by

$$p_n = \left(a + \frac{b}{n}\right) p_{n-1}, \tag{32}$$

with members Poisson ($a = 0, b = l, p_0 = e^{-\lambda}$), Binomial ($a = \frac{-q}{(1-q)}, b = \frac{(m+1)q}{(1-q)}, p_0 = (1-q)m$) and Negative Binomial ($a = \frac{b}{1+b}, b = \frac{(r-1)b}{1+b}, p_0 = (1 + b) - r$). In addition, we consider the higher order Panjer recursion for an extended class of frequency distributions given by the generalized Poisson distribution (GPD). The GPD model is defined via the probability mass function

$$\mathbb{P}\mathrm{r}\,(N=n)=p_n(\lambda,\theta)=\begin{cases}\lambda(\lambda+n\theta)^{n-1}\,, & \forall n=0,1,2,\ldots\\ 0, & \text{if } n>m,\ \text{when } \theta<0,\end{cases}$$

with $\lambda>0$ and $\max(-1,\lambda/m)\leq\theta<1$ and $m\geq 4$ is the largest positive integer s.t. $\lambda+\theta m>0$ when θ is negative, where the GPD is Poisson for $\theta=0$; over-dispersed $\theta>0$ and under-dispersed $\theta<0$.

One can then derive closed form recursions for the annual loss LDA compound process distribution given by

$$f_Y\,(x)=p_1 f_X\,(x)+\int_0^x\left(a+\frac{by}{x}\right)f_X\,(y)\,f_Y\,(x-y)\,dy, \tag{33}$$

or the generalized higher order Panjer recursion [40],

$$f_Y\,(x)=p_1(\lambda,\theta)f_X\,(x)+\frac{\lambda}{\lambda+\theta}\int_0^x\left(\theta+\lambda\frac{y}{x}\right)f_X\,(y)\,f_Y\,(x-y)\,dy. \tag{34}$$

To understand how these recursions are obtained, consider the convolution identity for an i.i.d. partial sum $S_{n+1}=X_1+\ldots+X_{n+1}$ with density

$$f^{*(n+1)}(x)=\int_0^x f(\tau)f^{*n}(x-\tau)d\tau,\quad \forall n=1,2,3,\ldots. \tag{35}$$

Substitute the conditional of X_1 when $S_{n+1}=x$,

$$f_{X_1}\,(\tau|X_1+\cdots+X_{n+1}=x)=\frac{f(\tau)f^{*n}(x-\tau)}{f^{*(n+1)}(x)}, \tag{36}$$

into the average given $S_{n+1}=x$ to get

$$\mathbb{E}\,[X_1|X_1+\cdots+X_{n+1}=x]=\int_0^x\tau\frac{f_{X_1}(\tau)f_{X_1}^{*n}(x-\tau)}{f_{X_1}^{*(n+1)}(x)}d\tau. \tag{37}$$

Then observe that with i.i.d. losses one also gets

$$\begin{aligned}\mathbb{E}\,[X_1|X_1+\cdots+X_{n+1}=x]&=\frac{1}{n+1}\sum_{i=1}^{n+1}\mathbb{E}\,[X_i|X_1+\cdots+X_{n+1}=x]\\ &=\frac{1}{n+1}\mathbb{E}\,[X_1+\cdots+X_{n+1}|X_1+\cdots+X_{n+1}=x]=\frac{x}{n+1}.\end{aligned} \tag{38}$$

Equating these conditional mean expressions and rearranging gives

$$\frac{1}{n+1} f_{X_1}^{*(n+1)}(x) = \frac{1}{x} \int_0^x \tau f_{X_1}(\tau) f_{X_1}^{*n}(x-\tau) d\tau. \tag{39}$$

Now utilise the Panjer class of frequency distributions satisfying for some a and b,

$$\mathbb{P}\mathrm{r}\,(N = n) = p_n = \left(a + \frac{b}{n}\right) p_{n-1}. \tag{40}$$

Upon substitution and some elementary algebra one obtains the Panjer recursion.

4.2 Stochastic Particle Methods as Solutions to Panjer Recursions

Avoiding the distributional discretisation of the severity model in applying the Panjer recursion reduces the computational cost when considering heavy-tailed severity models. It also reduces the discretisation error. It was noted in [61] that the Panjer recursions could be re-expressed as linear Volterra integral equations of the second kind via the mapping

$$\begin{aligned} &x_1 = x - y,\ g(x) = p_1 F(x),\ f\,(x_1) = f_Z(x_1), \text{ and} \\ &k\,(x, x_1) = \left(a + b\frac{x - x_1}{x}\right) F\,(x - x_1)\,. \end{aligned} \tag{41}$$

where the kernel $k : E \times E \mapsto \mathbb{R}$ and the function $g : E \mapsto \mathbb{R}$ are known whilst the function $f : E \mapsto \mathbb{R}$ is unknown. Furthermore, if one defines $k^0(x, y) \triangleq 1$, $k^1(x, y) \triangleq k(x, y)$ and

$$k^n(x, y) \triangleq \int k(x, y) k^{n-1}(z, y) dz$$

and these kernels satisfy that

$$\sum_{n=0}^{\infty} \int_E |k^n\,(x_0, x_n)\, g\,(x_n)|\, dx_n < \infty,$$

then one can identify the resolvent kernel and Neumann series through iterative expansion of the recursion to obtain for a sequence of domains $E_{1:n}$

$$f\,(x_0) = g(x_0) + \sum_{n=0}^{\infty} \int_0^{x_0} \cdots \int_0^{x_{n-1}} g\,(x_n) \prod_{l=1}^{n} k\,(x_{l-1}, x_l)\, dx_{1:n}.$$

Under this formulation it was shown in [61] how to address two problems: *estimation of the annual loss density over a set A and estimation of the annual loss density pointwise*. These are both directly relevant to obtaining estimates of the risk measures specified for capital estimation.

To achieve this we convert the Neumann series above into a sequence of expectations with respect to an importance sampling distribution. This is performed by making the following associations

$$f_0(x_0) = g(x_0), \text{ and } f_n(x_{0:n}) = g(x_n) \prod_{l=1}^{n} k(x_{l-1}, x_l)$$

$$\therefore f(x_0) = f_0(x_0) + \sum_{n=1}^{\infty} \int_0^{x_0} \cdots \int_0^{x_{n-1}} f_n(x_{0:n})\, dx_{0:n}.$$

Now we may develop this reformulated problem as an expectation with respect to a sequence of distributions $\{\pi(n, x_{1:n})\}_{n\geq 0}$:

$$f(x) = \frac{f_0(x)}{\pi(0)}\pi(0) + \sum_{n=1}^{\infty} \int_{A_{1:n}(x)} \frac{f_n(x, x_{1:n})}{\pi(n, x_{1:n})} \pi(n, x_{1:n})\, dx_{1:n}$$
$$= \mathbb{E}_{\pi(n, x_{1:n})}\left[\frac{f_n(x, x_{1:n})}{\pi(n, x_{1:n})}\right],$$

with the sets $A_{1:n}(x_0) = \{(x_1, \ldots, x_n) : x_0 > x_1 > \cdots > x_n\}$ playing an analogous role to the sequence of level sets described previously.

We note that there are now two path-space based particle solutions available, those that consider estimating $f(x)$ point-wise via an importance sampling solution on the path-space defined by

$$\cup_{n=0}^{\infty} \{n\} \times A_{1:n}(x).$$

The other alternative involves characterizing $f(x)$ over some interval by obtaining samples from its restriction to that interval $[x_a, x_b]$, via importance sampling on a slightly larger space

$$\bigcup_{n=0}^{\infty} \{n\} \times A_{1:n}([x_a, x_b]).$$

In [61] a path space based Sequential Importance Sampling (SIS) approximation to this sequence of expectations is obtained. This involves considering a Markov chain with initial distribution $\mu(x) > 0$ on E and transition kernel $M(x, y) > 0$ if $k(x, y) \neq 0$ and M has absorbing state $d \notin E$ such that $M(x, d) = P_d$ for any $x \in E$. Under this framework the interacting particle solution to the Panjer recursion is summarized in Algorithm 1. This is directly applicable to the higher order Panjer recursions discussed above.

Algorithm 1 (Path Space Stochastic Particle Methods for Panjer Recursions).

1. *Generate N independent Markov chain paths $\left\{X^{(i)}_{0:n^{(i)}+1}\right\}_{i=1}^{N}$ until absorption $X^{(i)}_{n^{(i)}+1} = d$.*
2. *Evaluate the importance weights for each particle on the path space by,*

$$W\left(X^{(i)}_{0:n^{(i)}}\right) = \begin{cases} \frac{1}{\mu\left(X^{(i)}_0\right)}\left(\prod_{n=1}^{n^{(i)}} \frac{k\left(X^{(i)}_{n-1}, X^{(i)}_n\right)}{M\left(X^{(i)}_{n-1}, X^{(i)}_n\right)}\right)\frac{g\left(X^{(i)}_{n^{(i)}}\right)}{P_d}, & \text{if } n^{(i)} \geq 1, \\ \frac{g\left(X^{(i)}_0\right)}{\mu\left(X^{(i)}_0\right)P_d}, & \text{if } n^{(i)} = 0. \end{cases} \tag{42}$$

4.3 Stochastic Particle Solutions to Risk Measure Estimation

If we consider $\mu\left(X^{(i)}_0\right) = \delta\left(X^{(i)}_0\right)$, $\forall i \in \{1, \ldots, N\}$, the empirical measure at *a point* x_0 is given by

$$\hat{f}_Z(x_0) = \frac{1}{N}\sum_{i=1}^{N} W\left(x_0, X^{(i)}_{1:n^{(i)}}\right),$$

or over an interval by

$$\hat{f}_Z(x_0) = \frac{1}{N}\sum_{i=1}^{N} W_1\left(X^{(i)}_{0:n^{(i)}}\right)\delta\left(x_0 - X^{(i)}_0\right).$$

These estimators can be used to construct unbiased Monte Carlo approximations of the expectation of $f_Z(z)$ for any set A given by $\mathbb{E}\left[\int_A \hat{f}(x_0)dx_0\right] = \int_A f(x_0)dx_0$.

Having obtained this particle based approximation, this weighted Dirac measure can then be utilised to estimate any of the required risk measures such as VaR and SRM for any desired level α. This can be performed in two ways, depending on whether the particle solution is obtained for the evaluation of the recursions pointwise over a fixed grid or alternatively over an interval, which could be increasing in size. In the case of an interval, or contracting set, one considers perhaps a set of interest to be $A = [0, x_{\max}]$ such that $x_{max} >> F^{\leftarrow}\left(1 - \frac{1-\alpha}{\mathbb{E}[N]}\right)$, and then utilises this to construct an unbiased particle approximation of the distribution of the annual loss up to any level $\alpha \in (0, 1)$. This could be obtained from growing a set $A_1 = [0, x_1] \subset A_2 \subset \cdots \subset A = [0, x_{\max}]$ recursively, as discussed in previous sections.

If the partition is made pointwise over a linear or a non-linear spacing $[0, z] = \bigcup_{m=1}^{M} [(m-1)\triangle, m\triangle)$ and the distribution evaluated pointwise, this leads to an estimation of

$$\hat{F}_Z(z) = \sum_{m=0}^{M} \triangle f\,(m\triangle) \approx \frac{1}{N} \sum_{m=0}^{M} \sum_{i=1}^{N} \triangle W\left(m\triangle, X_{1:n^{(i,m)}}^{(i,m)}\right). \tag{43}$$

Alternatively, if the estimation is performed over an interval $A(x_{\max}) = [0, x_{\max}]$, then for any $z < x_{max}$ one may use the construction of the resulting empirical measure to obtain,

$$\hat{F}_Z\,(z) = \frac{1}{N} \sum_{i=1}^{N} W\left(X_{0:n^{(i)}}^{(i)}\right) \mathbb{I}\left(X_{0:n^{(i)}}^{(i)} \in [0, z]\right) \to_{N\uparrow\infty} \int_0^z f_Z(z)dz. \tag{44}$$

Practical advice: consider a range for the support $[0, x_{\max}]$ such that $x_{\max} >> F^{\leftarrow}\left(1 - \frac{1-\alpha}{\mathbb{E}[N]}\right)$.

From these unbiased particle approximations of the annual loss density and distribution we can reconstruct the (inverse cdf) quantile function of the annual loss LDA model. This can either be based on a random set of particle locations or on a discrete deterministic grid as follows:

Deterministic Grid Solution: Given partition $[0, x_{\max}] = \bigcup_{m=1}^{M} [(m-1)\triangle, m\triangle)$ for some step $\triangle$ s.t.

$$\hat{Q}(p) = \inf\left\{ x \in \{0, \triangle, \ldots, M\triangle\} :\ p \le \frac{1}{N} \sum_{m=0}^{M} \sum_{i=1}^{N} \triangle W\left(x, X_{1:n^{(i,m)}}^{(i,m)}\right)\right\}. \tag{45}$$

Interval Solution: Construct the empirical measure over $A(\infty) = [0, \infty)$ s.t.

$$\hat{Q}(p) = \inf\left\{ x \in \left\{X_{(0)}^{(i)}\right\}_{i=1:N} :\ p \le \frac{1}{N} \sum_{i=1}^{N} W\left(X_{(0):n^{(i)}}^{(i)}\right) \mathbb{I}\left(X_{(0):n^{(i)}}^{(i)} \in [0, x]\right)\right\} \tag{46}$$

$X_{(0)}^{(i)}$ represents the order statistics for the particles.

Given the quantile function estimate we get the risk measure estimates for any $\alpha \in (0, 1)$ by:

Value-at-Risk (VaR): directly obtained using the estimated quantile function!

Spectral Risk (SRM): the SRM for a weight function $\phi : [0, 1] \mapsto \mathbb{R}$ is given by

$$\widehat{\mathrm{SRM}}_Z(\phi) = \frac{1}{N} \sum_{i=1}^{N} X_{(0):n^{(i)}}^{(i)} \phi\,(p_i)\,\triangle_{p_i}$$

with $p_i = \sum_{j=1:i} W\left(X^{(i)}_{(0):n^{(i)}}\right)$.

For additional discussions and detailed examples of this numerical approach to risk estimation, we refer the reader to the examples found in [61] and [70].

Example 4 (Poisson-Log Normal LDA Model (continued)). Consider the Poisson-Log Normal compound process detailed in Example 1. We demonstrate results for the standard Monte Carlo approach and compare results to the path-space particle solution discussed above. The Monte Carlo solution involved $N = 50$ million samples, hence it is considered effectively exact since the resulting Monte Carlo error was insignificant. In addition, a grid based solution was adopted for the particle solution with $N = 50$k per grid point giving a total of $NT = 500$k, with a grid width = 1. The estimated quantiles (rounded to integer values) are provided in the following table for two sets of parameter settings of $\lambda = 2$, $\mu = 2$ and $\sigma = 0.5$ and $\lambda = 2$, $\mu = 2$ and $\sigma = 1$. The particle solution presents a 95 % confidence interval and the single loss approximation simply reports the asymptotic approximation no explicit error can be calculated for the point estimated quantile (as discussed above).

Table 1 Standard Monte Carlo solution (exact) versus particle solution and first order single loss approximations.

Quantile level	Standard Monte Carlo		Particle solution (Algorithm 1)		Single loss approximation	
(%)	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 0.5$	$\sigma = 1$
50	14	16	15 [14,16]	16 [13,17]	10	14
80	27	39	25 [26,28]	41 [39,43]	14	26
90	35	57	33 [31,35]	55 [52,59]	16	38
95	42	77	40 [38,43]	74 [70,79]	19	52
99	57	129	55 [54,56]	123 [119,127]	26	97
99.5	77	234	73 [68,79]	227 [218,240]	38	198
99.95	83	276	79 [73,91]	270 [261,282]	42	240

The results that we present in Table 1 are obtained on a linearly spaced grid. However, this can be changed to either include a non-linear spacing, placing more points around the mode and less points in the tails, or as we detailed, straight out evaluation on an interval, avoiding the discretisation of the grid. For the sake of comparison between the standard Monte Carlo and the importance sampling estimates, we histogram the standard Monte Carlo procedure samples using unit length bins. We can see two things from Table 1, firstly as expected the particle based solution performs accurately under any parameter settings for a modest computational budget. When compared to the Single Loss Approximation, we see that there is two clear advantages in having a complementary particle solution, since we obtain measures of uncertainty in the quantile point estimates, trivially. Secondly, we demonstrate that the Single Loss Approximations may not be as accurate as required for even these simple models at quantiles that may be of

interest to assessment and are required for reporting of capital figures under financial regulation standards.

Appendix: Optimal Coupling Updated Models

This section is mainly concerned with the proof of the coupling formula (23). By construction, we clearly have that

$$\mathbb{P}\left(X_{p+1}^{\star} \neq X_{p}^{\star}\right) = \eta_p(A_p - A_{p+1}) = \eta_p(1 - G_p) = 1 - \eta_p(G_p).$$

On the other hand, we have

$$\begin{aligned}\eta_{p+1}(\varphi) - \eta_p(\varphi) &= \eta_p\left(S_{p,\eta_p}(\varphi) - \varphi\right) \\ &= \eta_p\left(\left[1 - G_p\right]\left[\varphi - \Psi_{G_p}(\eta_p)(\varphi)\right]\right).\end{aligned}$$

Choosing $\varphi = 1 - G_p$, so that

$$\Psi_{G_p}(\eta_p)(\varphi) = 1 - \Psi_{G_p}(\eta_p)(G_p) = 0$$

and

$$\eta_p\left(\left[1 - G_p\right]\left[\varphi - \Psi_{G_p}(\eta_p)(\varphi)\right]\right) = \eta_p\left(\left[1 - G_p\right]^2\right) = 1 - \eta_p\left(G_p\right).$$

This ends the proof of the optimal coupling formulae (23). Next, we observe that

$$1 - \eta_p\left(G_p\right) = 1 - \eta_0(A_{p+1})/\eta_0(A_p) \quad (\text{with} \quad \eta_0 = \text{Law}(X))$$

from which we conclude that

$$\eta_0\left(A_p\right) \geq \eta_0\left(A_{p+1}\right) \geq (1 - \epsilon)\,\eta_0\left(A_p\right) \Longrightarrow \mathbb{P}\left(X_{p+1}^{\star} = X_{p}^{\star}\right) \geq 1 - \epsilon. \quad (47)$$

References

1. Albrecher, H., Hipp, C., Kortschak, D.: Higher-order expansions for compound distributions and ruin probabilities with subexponential claims. Scand. Actuar. J. **2010**, 105–135 (2010)
2. Artzner, P., Delbaen, F., Eber, J., Heath, D.: Coherent measures of risk. Math. Finance **9**, 203–228 (1999)
3. Arulampalam, M., Maskell, S., Gordon, N., Clapp, T.: A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. IEEE Trans. Signal Process. **50**, 174–188 (2002)

4. Barbe, P., McCormick, W.: Asymptotic Expansions for Infinite Weighted Convolutions of Heavy Tail Distributions and Applications. American Mathematical Society, Providence (2009)
5. Barricelli, N.: Esempi numerici di processi di evoluzione. Methodos **6**, 45–68 (1954)
6. Barricelli, N.: Symbiogenetic evolution processes realized by artificial methods. Methodos **9**, 143–182 (1957)
7. BASEL, I., Bank for International Settlements, Basel Committee on Banking Supervision: Risk Management Principles for Electronic Banking (2001)
8. Bingham, N., Goldie, C., Teugels, J.: Regular Variation, vol. 27. Cambridge University Press, Cambridge (1989)
9. Bocker, K., Kluppelberg, C.: Operational var: a closed-form approximation. Risk-London-Risk Magazine Limited **18**, 90 (2005)
10. Böcker, K., Klüppelberg, C.: First order approximations to operational risk: dependence and consequences. In: Gregoriou, G.N. (ed.) Operational Risk Towards Basel III, Best Practices and Issues in Modeling, Management and Regulation, pp. 219–245. Wiley, Hoboken (2009)
11. Cappé, O., Godsill, S., Moulines, E.: Non linear filtering: interacting particle solution. Markov Process. Related Fields **2**, 555–580 (1996)
12. Cappé, O., Godsill, S., Moulines, E.: An overview of existing methods and recent advances in sequential Monte Carlo. Proc. IEEE **95**, 899–924 (2007)
13. Cappé, O., Moulines, E., Rydén, T.: Inference in Hidden Markov Models. Springer Science and Business Media, New York (2005)
14. Cruz, M., Peters, G., Shevchenko, P.: Handbook on Operational Risk. Wiley, New York (2013)
15. Daley, D., Omey, E., Vesilo, R.: The tail behaviour of a random sum of subexponential random variables and vectors. Extremes **10**, 21–39 (2007)
16. Degen, M.: The calculation of minimum regulatory capital using single-loss approximations. J. Oper. Risk **5**, 1–15 (2010)
17. Degen, M., Embrechts, P.: Scaling of high-quantile estimators. J. Appl. Probab. **48**, 968–983 (2011)
18. Del Moral, P.: Measure-valued processes and interacting particle systems. Application to nonlinear filtering problems. Ann. Appl. Probab. **8**, 438–495 (1998)
19. Del Moral, P.: Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications. Probability and Applications. Springer, New York (2004)
20. Del Moral, P.: Mean Field Simulation for Monte Carlo Integration, 600p. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Boca Raton (2013)
21. Del Moral, P., Doucet, A., Jasra, A.: Sequential Monte Carlo samplers. J. R. Stat. Soc. Ser. B Stat. Methodol. **68**, 411–436 (2006)
22. Del Moral, P., Hu, P., Wu, L.: On the concentration properties of interacting particle processes (2011). Arxiv preprint arXiv:1107.1948
23. Del Moral, P., Miclo, L.: Branching and interacting particle systems approximations of Feynman-Kac formulae with applications to non-linear filtering. In: Azema, J., Emery, M., Ledoux, M., Yor, M. (eds.) Seminaire de Probabilites XXXIV. Lecture Notes in Mathematics, vol. 1729, pp. 1–145. Springer, Berlin/Heidelberg (2000)
24. Del Moral, P., Miclo, L.: A Moran particle system approximation of Feynman-Kac formulae. Stochastic Process. Appl. **86**, 193–216 (2000)
25. Del Moral, P., Miclo, L.: Genealogies and increasing propagation of chaos for Feynman-Kac and genetic models. Ann. Appl. Probab. **11**, 1166–1198 (2001)
26. Del Moral, P., Miclo, L.: Particle approximations of lyapunov exponents connected to Schrödinger operators and Feynman-Kac semigroups. ESAIM Probab. Stat. **7**, 171–208 (2003)
27. Del Moral, P., Noyer, J.-C., Rigal, G., Salut, G.: Particle filters in radar signal processing: detection, estimation and air targets recognition. Research report no. 92495, LAAS-CNRS, Toulouse (1992)
28. Del Moral, P., Patras, F., Rubenthaler, S.: A mean field theory of nonlinear filtering. In: Crisan, D., Rozovskiĭ, B. (eds.) The Oxford Handbook of Nonlinear Filtering, pp. 705–740. Oxford University Press, Oxford (2011)

29. Del Moral, P., Rigal, G., Salut, G.: Nonlinear and non Gaussian particle filters applied to inertial platform repositioning. Research report no. 92207, STCAN/DIGILOG-LAAS/CNRS convention STCAN no. A.91.77.013, LAAS-CNRS, Toulouse, pp. 1–94 (1991)
30. Del Moral, P., Rigal, G., Salut, G.: Estimation and nonlinear optimal control: particle resolution in filtering and estimation: experimental results. Convention DRET no. 89.34.553.00.470.75.01, research report no.2, pp. 1–54 (1992)
31. Del Moral, P., Rigal, G., Salut, G.: Estimation and nonlinear optimal control: particle resolution in filtering and estimation: theoretical results. Convention DRET no. 89.34.553.00.470.75.01, research report no.3, pp. 1–123 (1992)
32. Del Moral, P., Rio, E.: Concentration inequalities for mean field particle models. Ann. Appl. Probab. **21**, 1017–1052 (2011)
33. Del Moral, P., Salut, G.: Maslov optimization theory: optimality versus randomness. Idempotency Anal. Appl., Math. Appl. **401**, 243–302 (1997)
34. Doucet, A., De-Freitas, N., Gordon, N.: Sequential Monte Carlo Methods in Practice. Springer, New York (2001)
35. Doucet, A., Johansen, A.: A tutorial on particle filtering and smoothing: fifteen years later. In: Crisan, D., Rozovsky, B. (eds.) Handbook of Nonlinear Filtering, pp. 656–704. Cambridge University Press, Cambridge (2009)
36. Embrechts, P., Hofert, M.: A note on generalized inverses. Math. Methods Oper. Res., 1–10 (2011).
37. Embrechts, P., Puccetti, G., Rüschendorf, L.: Model uncertainty and VaR aggregation. J. Bank. Finance **37**, 2750–2764 (2013). Embrechts, P., Maejima, M., Teugels, J.: Asymptotic behaviour of compound distributions. Astin Bull. **15**, 45–48 (1985)
38. Feller, W.: An introduction to Probability Theory, vol. II. Wiley, New York (1966)
39. Glasserman, P.: Monte Carlo Methods in Financial Engineering, vol. 53. Springer, New York (2003)
40. Goovaerts, M., Kaas, R.: Evaluating compound generalized Poisson distributions recursively. Astin Bull. **21**, 193–198 (1991)
41. Gordon, N., Salmond, D., Smith, A.: Novel approach to nonlinear/non-Gaussian Bayesian state estimation. IEE Proceedings F Radar and Signal Processing **140**, 107–113 (1993)
42. Hess, C.: Can the single-loss approximation method compete with the standard Monte Carlo simulation technique? J. Oper. Risk **6**, 31–43 (2011)
43. Hua, L., Joe, H.: Tail comonotonicity: properties, constructions, and asymptotic additivity of risk measures. Insurance Math. Econom. **51**, 492–503 (2012)
44. Kahn, H., Harris, T.: Estimation of particle transmission by random sampling. National Bureau of Standards Applied Mathematics Series **12**, 27–30 (1951)
45. Kitagawa, G.: Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. J. Comput. Graph. Statist. **5**, 1–25 (1996)
46. Kitagawa, G., Gersch, W.: Smoothness Priors Analysis of Time Series, vol. 116. Springer, New York (1996)
47. Klugman, S., Panjer, H., Willmot, G., Venter, G.: Loss Models: From Data to Decisions, vol. 2. Wiley, New York (1998)
48. Luo, X., Shevchenko, P.: Computing tails of compound distributions using direct numerical integration (2009). arXiv preprint, arXiv:0904.0830
49. Luo, X., Shevchenko, P.: Lgd credit risk model: estimation of capital with parameter uncertainty using mcmc (2010). Arxiv preprint, arXiv:1011.2827
50. McNeil, A., Frey, R., Embrechts, P.: Quantitative Risk Management: Concepts, Techniques and Tools. Princeton University Press, Princeton (2005)
51. Merz, M., Wüthrich, M.: Paid–incurred chain claims reserving method. Insurance Math. Econom. **46**, 568–579 (2010)
52. Metropolis, N., Ulam, S.: The Monte Carlo method. J. Amer Statist. Assoc. **44**, 335–341 (1949)
53. Najim, K., Ikonen, E., Del Moral, P.: Open-loop regulation and tracking control based on a genealogical decision tree. Neural Comput. Appl. **15**, 339–349 (2006)

54. Nešlehová, J., Embrechts, P., Chavez-Demoulin, V.: Infinite mean models and the LDA for operational risk. J. Oper. Risk **1**, 3–25 (2006)
55. Omey, E., Willekens, E.: Second order behaviour of the tail of a subordinated probability distribution. Stochastic Process. Appl. **21**, 339–353 (1986)
56. Panjer, H.: Recursive evaluation of a family of compound distributions. Astin Bull. **12**, 22–26 (1981)
57. Peters, G.: Topics in sequential Monte Carlo samplers. M.Sc., Department of Engineering, University of Cambridge (2005)
58. Peters, G., Byrnes, A., Shevchenko, P.: Impact of insurance for operational risk: is it worthwhile to insure or be insured for severe losses? Insurance Math. Econom. **48**, 287–303 (2011)
59. Peters, G., Dong, A., Kohn, R.: A copula based Bayesian approach for paid-incurred claims models for non-life insurance reserving (2012). arXiv preprint, arXiv:1210.3849
60. Peters, G., Fan, Y., Sisson, S.: On sequential Monte Carlo, partial rejection control and approximate Bayesian computation. Stat. Comput. **22**, 1209–1222 (2012)
61. Peters, G., Johansen, A., Doucet, A.: Simulation of the annual loss distribution in operational risk via Panjer recursions and volterra integral equations for value at risk and expected shortfall estimation. J. Oper. Risk **2**, 29–58 (2007)
62. Peters, G., Shevchenko, P., Wüthrich, M.: Dynamic operational risk: modeling dependence and combining different sources of information. J. Oper. Risk **4**, 69–104 (2009)
63. Peters, G., Shevchenko, P., Wüthrich, M.: Model uncertainty in claims reserving within Tweedie's compound Poisson models. Astin Bull. **39**, 1–33 (2009)
64. Peters, G., Shevchenko, P., Young, M., Yip, W.: Analytic loss distributional approach models for operational risk from the-stable doubly stochastic compound processes and implications for capital allocation. Insurance Math. Econom. **49**, 565 (2011)
65. Peters, G., Sisson, S.: Bayesian inference, Monte Carlo sampling and operational risk. J. Oper. Risk **1**, 27–50 (2006)
66. Peters, G.W., Targino, R.S., Shevchenko, P.V.: Understanding operational risk capital approximations: first and second orders (2013). arXiv preprint, arXiv:1303.2910
67. Peters, G., Wüthrich, M., Shevchenko, P.: Chain ladder method: Bayesian bootstrap versus classical bootstrap. Insurance Math. Econom. **47**, 36–51 (2010)
68. Rosenbluth, M., Rosenbluth, A.: Monte-Carlo calculations of the average extension of macromolecular chains. J. Chem. Phys. **23**, 356–359 (1955)
69. Shevchenko, P.: Implementing loss distribution approach for operational risk. Appl. Stoch. Models Bus. Ind. **26**, 277–307 (2009)
70. Shevchenko, P.: Modelling Operational Risk Using Bayesian Inference. Springer, Berlin (2011)
71. Sundt, B., Vernic, R.: Recursions for Convolutions and Compound Distributions with Insurance Applications. Springer, Berlin/Heidelberg (2009)
72. Verrall, R.J.: A Bayesian generalized linear model for the Bornhuetter-Ferguson method of claims reserving. North Am. Actuarial J. **8**, 67–89 (2004)
73. Willekens, E.: Asymptotic approximations of compound distributions and some applications. Bulletin de la Société Mathématique de Belgique Série. B **41**, 55–61 (1989)
74. Zolotarev, V.: Univariate Stable Distributions. Nauka, Moscow (1983)

Multilevel Monte Carlo Methods

Michael B. Giles

Abstract The author's presentation of multilevel Monte Carlo path simulation at the MCQMC 2006 conference stimulated a lot of research into multilevel Monte Carlo methods. This paper reviews the progress since then, emphasising the simplicity, flexibility and generality of the multilevel Monte Carlo approach. It also offers a few original ideas and suggests areas for future research.

1 Introduction

1.1 Control Variates and Two-Level MLMC

One of the classic approaches to Monte Carlo variance reduction is through the use of a control variate. Suppose we wish to estimate $\mathbb{E}[f]$, and there is a control variate g which is well correlated to f and has a known expectation $\mathbb{E}[g]$. In that case, we can use the following unbiased estimator for $\mathbb{E}[f]$:

$$N^{-1}\sum_{n=1}^{N}\left\{f^{(n)}-\lambda\left(g^{(n)}-\mathbb{E}[g]\right)\right\}.$$

The optimal value for λ is $\rho\sqrt{\mathbb{V}[f]\,/\,\mathbb{V}[g]}$, where ρ is the correlation between f and g, and the variance of the control variate estimator is reduced by factor $1-\rho^2$ compared to the standard estimator.

A two-level version of MLMC (multilevel Monte Carlo) is very similar. If we want to estimate $\mathbb{E}[P_1]$ but it is much cheaper to simulate $P_0 \approx P_1$, then since

M.B. Giles (✉)
Mathematical Institute, University of Oxford, Oxford, UK
e-mail: mike.giles@maths.ox.ac.uk

J. Dick et al. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2012*,
Springer Proceedings in Mathematics & Statistics 65,
DOI 10.1007/978-3-642-41095-6_4, © Springer-Verlag Berlin Heidelberg 2013

$$\mathbb{E}[P_1] = \mathbb{E}[P_0] + \mathbb{E}[P_1 - P_0]$$

we can use the unbiased two-level estimator

$$N_0^{-1} \sum_{n=1}^{N_0} P_0^{(n)} \;+\; N_1^{-1} \sum_{n=1}^{N_1} \left(P_1^{(n)} - P_0^{(n)} \right).$$

Here $P_1^{(n)}$–$P_0^{(n)}$ represents the difference between P_1 and P_0 for the same underlying stochastic sample, so that $P_1^{(n)} - P_0^{(n)}$ is small and has a small variance; the precise construction depends on the application and various examples will be shown later. The two key differences from the control variate approach are that the value of $\mathbb{E}[P_0]$ is not known, so has to be estimated, and we use $\lambda = 1$.

If we define C_0 and C_1 to be the cost of computing a single sample of P_0 and $P_1 - P_0$, respectively, then the total cost is $N_0\, C_0 + N_1\, C_1$, and if V_0 and V_1 are the variance of P_0 and $P_1 - P_0$, then the overall variance is $N_0^{-1} V_0 + N_1^{-1} V_1$, assuming that $\displaystyle\sum_{n=1}^{N_0} P_0^{(n)}$ and $\displaystyle\sum_{n=1}^{N_1} \left(P_1^{(n)} - P_0^{(n)} \right)$ use independent samples.

Hence, treating the integers N_0, N_1 as real variables and performing a constrained minimisation using a Lagrange multiplier, the variance is minimised for a fixed cost by choosing $N_1 / N_0 = \sqrt{V_1 / C_1} \,/ \sqrt{V_0 / C_0}$.

1.2 Multilevel Monte Carlo

The full multilevel generalisation is quite natural: given a sequence $P_0, P_1, \ldots,$ which approximates P_L with increasing accuracy, but also increasing cost, we have the simple identity

$$\mathbb{E}[P_L] = \mathbb{E}[P_0] + \sum_{\ell=1}^{L} \mathbb{E}[P_\ell - P_{\ell-1}],$$

and therefore we can use the following unbiased estimator for $\mathbb{E}[P_L]$,

$$N_0^{-1} \sum_{n=1}^{N_0} P_0^{(0,n)} \;+\; \sum_{\ell=1}^{L} \left\{ N_\ell^{-1} \sum_{n=1}^{N_\ell} \left(P_\ell^{(\ell,n)} - P_{\ell-1}^{(\ell,n)} \right) \right\}$$

with the inclusion of the level ℓ in the superscript (ℓ, n) indicating that the samples used at each level of correction are independent.

If we define C_0, V_0 to be the cost and variance of one sample of P_0, and C_ℓ, V_ℓ to be the cost and variance of one sample of $P_\ell - P_{\ell-1}$, then the overall cost and variance of the multilevel estimator is $\sum_{\ell=0}^{L} N_\ell\, C_\ell$ and $\sum_{\ell=0}^{L} N_\ell^{-1}\, V_\ell$, respectively.

For a fixed cost, the variance is minimised by choosing $N_\ell = \lambda \sqrt{V_\ell / C_\ell}$ for some value of the Lagrange multiplier λ. In particular, to achieve an overall variance of ε^2 requires that $\lambda = \varepsilon^{-2} \sum_{\ell=0}^{L} \sqrt{V_\ell\, C_\ell}$. The total computational cost is then

$$C = \varepsilon^{-2} \left(\sum_{\ell=0}^{L} \sqrt{V_\ell\, C_\ell} \right)^2 . \tag{1}$$

It is important to note whether the product $V_\ell\, C_\ell$ increases or decreases with ℓ, i.e. whether or not the cost increases with level faster than the variance decreases. If it increases with level, so that the dominant contribution to the cost comes from $V_L\, C_L$ then we have $C \approx \varepsilon^{-2} V_L\, C_L$, whereas if it decreases and the dominant contribution comes from $V_0\, C_0$ then $C \approx \varepsilon^{-2} V_0\, C_0$. This contrasts to the standard MC cost of approximately $\varepsilon^{-2} V_0\, C_L$, assuming that the cost of computing P_L is similar to the cost of computing $P_L - P_{L-1}$, and that $\mathbb{V}[P_L] \approx \mathbb{V}[P_0]$. This shows that in the first case the MLMC cost is reduced by factor V_L / V_0, corresponding to the ratio of the variances $\mathbb{V}[P_L - P_{L-1}]$ and $\mathbb{V}[P_L]$, whereas in the second case it is reduced by factor C_0 / C_L, the ratio of the costs of computing P_0 and $P_L - P_{L-1}$. If the product $V_\ell\, C_\ell$ does not vary with level, then the total cost is $\varepsilon^{-2} L^2\, V_0\, C_0 = \varepsilon^{-2} L^2\, V_L\, C_L$.

1.3 Earlier Related Work

Prior to the author's first publications [20, 21] on MLMC for Brownian path simulations, Heinrich developed a multilevel Monte Carlo method for parametric integration, the evaluation of functionals arising from the solution of integral equations, and weakly singular integral operators [33–37]. Parametric integration concerns the estimation of $\mathbb{E}[f(x, \lambda)]$ where x is a finite-dimensional random variable and λ is a parameter. In the simplest case in which λ is a real variable in the range $[0, 1]$, having estimated the value of $\mathbb{E}[f(x, 0)]$ and $\mathbb{E}[f(x, 1)]$, one can use $\frac{1}{2}(f(x, 0) + f(x, 1))$ as a control variate when estimating the value of $\mathbb{E}[f(x, \frac{1}{2})]$. This approach can then be applied recursively for other intermediate values of λ, yielding large savings if $f(x, \lambda)$ is sufficiently smooth with respect to λ. Although this does not quite fit into the general MLMC form given in the previous section, the recursive control variate approach is very similar and the complexity analysis is also very similar to the analysis to be presented in the next section.

Although not so clearly related, there are papers by Brandt et al. [9, 10] which combine Monte Carlo techniques with multigrid ideas in determining thermodynamic limits in statistical physics applications. It is the multigrid ideas of Brandt and

others for the iterative solution of systems of equations which were the inspiration for the author in developing the MLMC method for SDE path simulation.

In 2005, Kebaier [41] developed a two-level approach for path simulation which is very similar to the author's approach presented in the next section. The only differences are the use of only two levels, and the use of a general multiplicative factor as in the standard control variate approach. A similar multilevel approach was under development at the same time by Speight, but was not published until later [49,50].

2 MLMC Theorem

In the Introduction, we considered the case of a general multilevel method in which the output P_L on the finest level corresponds to the quantity of interest. However, in many infinite-dimensional applications, such as in SDEs and SPDEs, the output P_ℓ on level ℓ is an approximation to a random variable P. In this case, the mean square error (MSE) has the usual decomposition into the total variance of the multilevel estimator, plus the square of the bias $(\mathbb{E}[P_L - P])^2$. To achieve an MSE which is less than ε^2, it is sufficient to ensure that each of these terms is less than $\frac{1}{2}\varepsilon^2$. This leads to the following theorem:

Theorem 1. *Let P denote a random variable, and let P_ℓ denote the corresponding level ℓ numerical approximation.*

If there exist independent estimators Y_ℓ based on N_ℓ Monte Carlo samples, and positive constants $\alpha, \beta, \gamma, c_1, c_2, c_3$ such that $\alpha \geq \frac{1}{2}\min(\beta, \gamma)$ and

$$\begin{aligned}
&(i) \quad |\mathbb{E}[P_\ell - P]| \leq c_1\, 2^{-\alpha \ell} \\
&(ii) \quad \mathbb{E}[Y_\ell] = \begin{cases} \mathbb{E}[P_0], & \ell = 0 \\ \mathbb{E}[P_\ell - P_{\ell-1}], & \ell > 0 \end{cases} \\
&(iii) \quad \mathbb{V}[Y_\ell] \leq c_2\, N_\ell^{-1}\, 2^{-\beta \ell} \\
&(iv) \quad \mathbb{E}[C_\ell] \leq c_3\, N_\ell\, 2^{\gamma \ell}, \text{ where } C_\ell \text{ is the computational complexity of } Y_\ell
\end{aligned}$$

then there exists a positive constant c_4 such that for any $\varepsilon < e^{-1}$ there are values L and N_ℓ for which the multilevel estimator

$$Y = \sum_{\ell=0}^{L} Y_\ell,$$

has a mean-square-error with bound

$$MSE \equiv \mathbb{E}\left[(Y - \mathbb{E}[P])^2\right] < \varepsilon^2$$

with a computational complexity C with bound

$$\mathbb{E}[C] \leq \begin{cases} c_4\,\varepsilon^{-2}, & \beta > \gamma, \\ c_4\,\varepsilon^{-2}(\log\varepsilon)^2, & \beta = \gamma, \\ c_4\,\varepsilon^{-2-(\gamma-\beta)/\alpha}, & \beta < \gamma. \end{cases}$$

The statement of the theorem is a slight generalisation of the original theorem in [21]. It corresponds to the theorem and proof in [15], except for the minor change to expected costs to allow for applications such as jump-diffusion modelling in which the simulation cost of individual samples is itself random.

The theorem is based on the idea of a geometric progression in the levels of approximation, leading to the exponential decay in the weak error in condition *(i)*, and the variance in condition *(iii)*, as well as the exponential increase in the expected cost in condition *(iv)*. This geometric progression was based on experience with multigrid methods in the iterative solution of large systems of linear equations, but it is worth noting that it is not necessarily the optimal choice in all circumstances.

The result of the theorem merits some discussion. In the case $\beta > \gamma$, the dominant computational cost is on the coarsest levels where $C_\ell = O(1)$ and $O(\varepsilon^{-2})$ samples are required to achieve the desired accuracy. This is the standard result for a Monte Carlo approach using i.i.d. samples; to do better would require an alternative approach such as the use of Latin hypercube sampling or quasi-Monte Carlo methods. In the case $\beta < \gamma$, the dominant computational cost is on the finest levels. Because of condition *(i)*, $2^{-\alpha L} = O(\varepsilon)$, and hence $C_L = O(\varepsilon^{-\gamma/\alpha})$. If $\beta = 2\alpha$, which is usually the largest possible value for a given α, for reasons explained below, then the total cost is $O(C_L)$ corresponding to $O(1)$ samples on the finest level, again the best that can be achieved. The dividing case $\beta = \gamma$ is the one for which both the computational effort, and the contributions to the overall variance, are spread approximately evenly across all of the levels; the $(\log\varepsilon)^2$ term corresponds to the L^2 factor in the corresponding discussion in Sect. 1.2.

The natural choice for the multilevel estimator is

$$Y_\ell = N_\ell^{-1} \sum_i P_\ell(\omega_i) - P_{\ell-1}(\omega_i), \tag{2}$$

where $P_\ell(\omega_i)$ is the approximation to $P(\omega_i)$ on level ℓ, and $P_{\ell-1}(\omega_i)$ is the corresponding approximation on level $\ell-1$ for the same underlying stochastic sample ω_i. Note that $\mathbb{V}[P_\ell - P_{\ell-1}]$ is usually similar in magnitude to $\mathbb{E}[(P_\ell - P_{\ell-1})^2]$ which is greater than $(\mathbb{E}[P_\ell - P_{\ell-1}])^2$; this implies that $\beta \leq 2\alpha$ and hence the condition in the theorem that $\alpha \geq \frac{1}{2}\min(\beta,\gamma)$ is satisfied.

However, the multilevel theorem allows for the use of other estimators, provided they satisfy the restriction of condition *(ii)* which ensures that $\mathbb{E}[Y] = \mathbb{E}[P_L]$. Two examples of this will be given later in the paper. In the first, slightly different numerical approximations are used for the coarse and fine paths in SDE simulations, giving

$$Y_\ell = N_\ell^{-1} \sum_i P_\ell^f(\omega_i) - P_{\ell-1}^c(\omega_i).$$

Provided $\mathbb{E}[P_\ell^f] = \mathbb{E}[P_\ell^c]$ so that the expectation on level ℓ is the same for the two approximations, then condition *(ii)* is satisfied and no additional bias (other than the bias due to the approximation on the finest level) is introduced into the multilevel estimator. The second example defines an antithetic ω_i^a with the same distribution as ω_i, and then uses the multilevel estimator

$$Y_\ell = N_\ell^{-1} \sum_i \tfrac{1}{2}\left(P_\ell(\omega_i) + P_\ell(\omega_i^a)\right) - P_{\ell-1}(\omega_i).$$

Since $\mathbb{E}[P_\ell(\omega_i^a)] = \mathbb{E}[P_\ell(\omega_i)]$, then again condition *(ii)* is satisfied. In each case, the objective in constructing a more complex estimator is to achieve a greatly reduced variance $\mathbb{V}[Y_\ell]$ so that fewer samples are required.

3 SDEs

3.1 Euler Discretisation

The original multilevel path simulation paper [21] treated SDEs using the simple Euler-Maruyama discretisation together with the natural multilevel estimator (2).

Provided the SDE satisfies the usual conditions (see Theorem 10.2.2 in [42]), the strong error for the Euler discretisation with timestep h is $O(h^{1/2})$, and therefore for Lipschitz payoff functions P (such as European, Asian and lookback options in finance) the variance $V_\ell \equiv \mathbb{V}[P_\ell - P_{\ell-1}]$ is $O(h_\ell)$. If $h_\ell = 4^{-\ell} h_0$, as in [21], then this gives $\alpha = 2$, $\beta = 4$ and $\gamma = 2$. Alternatively, if $h_\ell = 2^{-\ell} h_0$, then $\alpha = 1$, $\beta = 2$ and $\gamma = 1$. In either case, Theorem 1 gives the complexity to achieve a root-mean-square error of ε to be $O(\varepsilon^{-2}(\log \varepsilon)^2)$, which is near-optimal as Müller-Gronbach and Ritter have proved an $O(\varepsilon^{-2})$ lower bound for the complexity [46].

For other payoff functions the complexity is higher. $V_\ell \approx O(h^{1/2})$ for the digital option which is a discontinuous function of the SDE solution at the final time, and the barrier option which depends discontinuously on the minimum or maximum value over the full time interval. Loosely speaking, this is because there is an $O(h^{1/2})$ probability of the coarse and fine paths being on opposite sides of the discontinuity, and in such cases there is an $O(1)$ difference in the payoff. Currently, there is no known "fix" for this for the Euler-Maruyama discretisation; we will return to this issue for the Milstein discretisation when there are ways of improving the situation.

Table 1 summarises the observed variance convergence rate in numerical experiments for the different options, and the theoretical results which have been obtained; the digital option analysis is due to Avikainen [4] while the others are due to Giles,

Higham and Mao [24]. Although the analysis in some of these cases is for one-dimensional SDEs, it also applies to multi-dimensional SDEs [22].

Table 1 Observed and theoretical convergence rates for the multilevel correction variance for scalar SDEs, using the Euler-Maruyama and Milstein discretisations. δ is any strictly positive constant.

	Euler-Maruyama		Milstein	
Option	Numerics	Analysis	Numerics	Analysis
Lipschitz	$O(h)$	$O(h)$	$O(h^2)$	$O(h^2)$
Asian	$O(h)$	$O(h)$	$O(h^2)$	$O(h^2)$
Lookback	$O(h)$	$O(h)$	$O(h^2)$	$o(h^{2-\delta})$
Barrier	$O(h^{1/2})$	$o(h^{1/2-\delta})$	$O(h^{3/2})$	$o(h^{3/2-\delta})$
Digital	$O(h^{1/2})$	$O(h^{1/2}\log h)$	$O(h^{3/2})$	$o(h^{3/2-\delta})$

3.2 Milstein Discretisation

For Lipschitz payoffs, the variance V_ℓ for the natural multilevel estimator converges at twice the order of the strong convergence of the numerical approximation of the SDE. This immediately suggests that it would be better to replace the Euler-Maruyama discretisation by the Milstein discretisation [20] since it gives first order strong convergence under certain conditions (see Theorem 10.3.5 in [42]).

This immediately gives an improved variance for European and Asian options, as shown in Table 1, but to get the improved variance for lookback, barrier and digital options requires the construction of estimators which are slightly different on the coarse and fine path simulations, but which respect the condition that $\mathbb{E}[P_\ell^f] = \mathbb{E}[P_\ell^c]$.

The construction for the digital option will be discussed next, but for the lookback and barrier options, the key is the definition of a Brownian Bridge interpolant based on the approximation that the drift and volatility do not vary within the timestep. For each coarse timestep, the mid-point of the interpolant can be sampled using knowledge of the fine path Brownian increments, and then classical results can be used for the distribution of the minimum or maximum within each fine timestep for both the fine and coarse path approximations [29]. The full details are given in [20], and Table 1 summarises the convergence behaviour observed numerically, and the supporting numerical analysis by Giles, Debrabant and Rößler [23].

The outcome is that for the case in which the number of timesteps doubles at each level, so $h_\ell = 2^{-\ell} h_0$, then $\gamma = 1$ and either $\beta = 2$ (European, Asian and lookback) or $\beta = 1.5$ (barrier and digital). Hence, we are in the regime where $\beta > \gamma$ and the overall complexity is $O(\varepsilon^{-2})$. Furthermore, the dominant computational cost is on the coarsest levels of simulation.

Since the coarsest levels are low-dimensional, they are well suited to the use of quasi-Monte Carlo methods which are particularly effective in lower dimensions

because of the existence of $O((\log N)^d/N)$ error bounds, where d is the dimension and N is the number of QMC points. The bounds are for the numerical integration of certain function classes on the unit hypercube, and are a consequence of the Koksma-Hlawka inequality together with bounds on the star-discrepancy of certain sequences of QMC points.

This has been investigated by Giles and Waterhouse [28] using a rank-1 lattice rule to generate the quasi-random numbers, randomisation with 32 independent offsets to obtain confidence intervals, and a standard Brownian Bridge construction of the increments of the driving Brownian process. The numerical results show that MLMC on its own was better than QMC on its own, but the combination of the two was even better. The QMC treatment greatly reduced the variance per sample for the coarsest levels, resulting in significantly reduced costs overall. In the simplest case of a Lipschitz European payoff, the computational complexity was reduced from $O(\varepsilon^{-2})$ to approximately $O(\varepsilon^{-1.5})$.

3.2.1 Digital Options

As discussed earlier, discontinuous payoffs pose a challenge to the multilevel Monte Carlo approach, because small differences in the coarse and fine path simulations can lead to an $O(1)$ difference in the payoff function. This leads to a slower decay in the variance V_ℓ, and because the fourth moment is also much larger it leads to more samples being required to obtain an accurate estimate for V_ℓ, which is needed to determine the optimal number of samples N_ℓ.

This is a generic problem. Although we will discuss it here in the specific context of a Brownian SDE and an option which is a discontinuous function of the underlying at the final time, the methods which are discussed are equally applicable in a range of other cases. Indeed, some of these techniques have been first explored in the context of pathwise sensitivity analysis [12] or jump-diffusion modelling [52].

Conditional Expectation

The conditional expectation approach builds on a well-established technique for payoff smoothing which is used for pathwise sensitivity analysis (see, for example, pp. 399–400 in [29]).

We start by considering the fine path simulation, and make a slight change by using the Euler-Maruyama discretisation for the final timestep, instead of the Milstein discretisation. Conditional on the numerical approximation of the value S_{T-h} one timestep before the end (which in turn depends on all of the Brownian increments up to that time) the numerical approximation for the final value S_T now has a Gaussian distribution, and for a simple digital option the conditional expectation is known analytically.

The same treatment is used for the coarse path, except that in the final timestep, we re-use the known value of the Brownian increment for the second last fine

timestep, which corresponds to the first half of the final coarse timestep. This results in the conditional distribution for the coarse path underlying at maturity matching that of the fine path to within $O(h)$, for both the mean and the standard deviation [23]. Consequently, the difference in payoff between the coarse and fine paths near the payoff discontinuity is $O(h^{1/2})$, and so the variance is approximately $O(h^{3/2})$.

Splitting

The conditional expectation technique works well in 1D where there is a known analytic value for the conditional expectation, but in multiple dimensions it may not be known. In this case, one can use the technique of "splitting" [3]. Here the conditional expectation is replaced by a numerical estimate, averaging over a number of sub-samples. i.e. for each set of Brownian increments up to one fine timestep before the end, one uses a number of samples of the final Brownian increment to produce an average payoff. If the number of sub-samples is chosen appropriately, the variance is the same, to leading order, without any increase in the computational cost, again to leading order. Because of its simplicity and generality, this is now my preferred approach. Furthermore, one can revert to using the Milstein approximation for the final timestep.

Change of Measure

The change of measure approach is another approximation to the conditional expectation. The fine and coarse path conditional distributions at maturity are two very similar Gaussian distributions. Instead of following the splitting approach of taking corresponding samples from these two distributions, we can instead take a sample from a third Gaussian distribution (with a mean and variance perhaps equal to the average of the other two). This leads to the introduction of a Radon-Nikodym derivative for each path, and the difference in the payoffs from the two paths is then due to the difference in their Radon-Nikodym derivatives.

In the specific context of digital options, this is a more complicated method to implement, and the resulting variance is no better. However, in other contexts a similar approach can be very effective.

3.2.2 Multi-dimensional SDEs

The discussion so far has been for scalar SDEs, but the computational benefits of Monte Carlo methods arise in higher dimensions. For multi-dimensional SDEs satisfying the usual commutativity condition (see, for example, p. 353 in [29]) the Milstein discretisation requires only Brownian increments for its implementation, and most of the analysis above carries over very naturally.

The only difficulties are in lookback and barrier options where the classical results for the distribution of the minimum or maximum of a one-dimensional Brownian motion, do not extend to the joint distribution of the minima or maxima of two correlated Brownian motions. An alternative approach may be to sub-sample from the Brownian Bridge interpolant for those timesteps which are most likely to give the global minimum or maximum. This may need to be combined with splitting for the barrier option to avoid the $O(1)$ difference in payoffs. An alternative might be to use adaptive time-stepping [40].

For multi-dimensional SDEs which do not satisfy the commutativity condition the Milstein discretisation requires the simulation of Lévy areas. This is unavoidable to achieve first order strong convergence; the classical result of Clark and Cameron says that $O(h^{1/2})$ strong convergence is the best that can be achieved in general using just Brownian increments [14].

However, Giles and Lukasz have developed an antithetic treatment which achieves a very low variance despite the $O(h^{1/2})$ strong convergence [26]. The estimator which is used is

$$Y_\ell = N_\ell^{-1} \sum_i \tfrac{1}{2} \left(P_\ell(\omega_i) + P_\ell(\omega_i^a) \right) - P_{\ell-1}(\omega_i).$$

Here ω_i represents the driving Brownian path, and ω_i^a is an antithetic counterpart defined by a time-reversal of the Brownian path within each coarse timestep. This results in the Brownian increments for the antithetic fine path being swapped relative to the original path. Lengthy analysis proves that the average of the fine and antithetic paths is within $O(h)$ of the coarse path, and hence the multilevel variance is $O(h^2)$ for smooth payoffs, and $O(h^{3/2})$ for the standard European call option.

This treatment has been extended to handle lookback and barrier options [27]. This combines sub-sampling of the Brownian path to approximate the Lévy areas with sufficient accuracy to achieve $O(h^{3/4})$ strong convergence, with an antithetic treatment at the finest level of resolution to ensure that the average of the fine paths is within $O(h)$ of the coarse path.

3.3 Lévy Processes

3.3.1 Jump-Diffusion Processes

With finite activity jump-diffusion processes, such as in the Merton model [44], it is natural to simulate each individual jump using a jump-adapted discretisation [47].

If the jump rate is constant, then the jumps on the coarse and fine paths will occur at the same time, and the extension of the multilevel method is straightforward [52].

If the jump rate is path-dependent then the situation is trickier. If there is a known upper bound to the jump rate, then one can use Glasserman and Merener's

"thinning" approach [31] in which a set of candidate jump times is simulated based on the constant upper bound, and then a subset of these are selected to be real jumps. The problem with the multilevel extension of this is that some candidate jumps will be selected for the coarse path but not for the fine path, or vice versa, leading to an $O(1)$ difference in the paths and hence the payoffs. Xia overcomes this by using a change of measure to select the jump times consistently for both paths, with a Radon-Nikodym derivative being introduced in the process [52].

3.3.2 More General Processes

With infinite activity Lévy processes it is impossible to simulate each jump. One approach is to simulate the large jumps and either neglect the small jumps or approximate their effect by adding a Brownian diffusion term [17,18,43]. Following this approach, the cutoff δ_ℓ for the jumps which are simulated varies with level, and $\delta_\ell \to 0$ as $\ell \to \infty$ to ensure that the bias converges to zero. In the multilevel treatment, when simulating $P_\ell - P_{\ell-1}$ the jumps fall into three categories. The ones which are larger than $\delta_{\ell-1}$ get simulated in both the fine and coarse paths. The ones which are smaller than δ_ℓ are either neglected for both paths, or approximated by the same Brownian increment. The difficulty is in the intermediate range $[\delta_\ell, \delta_{\ell-1}]$ in which the jumps are simulated for the fine path, but neglected or approximated for the coarse path. This is what leads to the difference in path simulations, and hence to a non-zero value for $P_\ell - P_{\ell-1}$.

Alternatively, for many SDEs driven by a Lévy process it is possible to directly simulate the increments of the Lévy process over a set of uniform timesteps [16,48], in exactly the same way as one simulates Brownian increments. For other Lévy processes, it may be possible in the future to simulate the increments by constructing approximations to the inverse of the cumulative distribution function. Where this is possible, it may be the best approach to achieve a close coupling between the coarse and fine path simulations, and hence a low variance V_ℓ, since the increments of the driving Lévy process for the coarse path can be obtained trivially by summing the increments for the fine path.

4 SPDEs

After developing the MLMC method for SDE simulations, it was immediately clear that it was equally applicable to SPDEs, and indeed the computational savings would be greater because the cost of a single sample increases more rapidly with grid resolution for SPDEs with higher space-time dimension.

In 2006, the author discussed this with Thomas Hou in the specific context of elliptic SPDEs with random coefficients, and Hou's postdoc then performed the first unpublished MLMC computations for SPDEs. The first published work was by

a student of Klaus Ritter in her Diploma thesis [32]; her application was to parabolic SPDEs. Since this early work, there has been a variety of papers on elliptic [6, 13, 15, 51], parabolic [5, 25] and hyperbolic [45] SPDEs.

In almost all of this work, the construction of the multilevel estimator is quite natural, using a geometric sequence of grids and the usual estimators for $P_\ell - P_{\ell-1}$. It is the numerical analysis of the variance of the multilevel estimator which is often very challenging.

4.1 Elliptic SPDE

The largest amount of research on multilevel for SPDEs has been for elliptic PDEs with random coefficients. The PDE typically has the form

$$-\nabla \cdot (k(\mathbf{x},\omega)\nabla p(\mathbf{x},\omega)) = 0, \qquad \mathbf{x} \in D.$$

with Dirichlet or Neumann boundary conditions on the boundary ∂D. For subsurface flow problems, such as the modelling of groundwater flow in nuclear waste repositories, the diffusivity (or permeability) k is often modelled as a lognormal random field, i.e. $\log k$ is a Gaussian field with a uniform mean (which we will take to be zero for simplicity) and a covariance function of the general form $R(\mathbf{x},\mathbf{y}) = r(\mathbf{x}-\mathbf{y})$. Samples of $\log k$ are provided by a Karhunen-Loève expansion:

$$\log k(\mathbf{x},\omega) = \sum_{n=0}^{\infty} \sqrt{\theta_n}\, \xi_n(\omega)\, f_n(\mathbf{x}),$$

where θ_n are the eigenvalues of $R(\mathbf{x},\mathbf{y})$ in decreasing order, f_n are the corresponding eigenfunctions, and ξ_n are independent unit Normal random variables. However, it is more efficient to generate them using a circulant embedding technique which enables the use of FFTs [19].

The multilevel treatment is straightforward. The spatial grid resolution is doubled on each level. Using the Karhunen-Loève generation, the expansion is truncated after K_ℓ terms, with K_ℓ increasing with level [51]; in unpublished work, a similar approach has also been used with the circulant embedding generation.

In both cases, $\log k$ is generated using a row-vector of independent unit Normal random variables ξ. The variables for the fine level can be partitioned into those for the coarse level $\xi_{\ell-1}$, plus some additional variables z_ℓ, giving $\xi_\ell = (\xi_{\ell-1}, z_\ell)$. It is possible to develop an antithetic treatment similar to that used for SDEs by defining $\xi_\ell^a = (\xi_{\ell-1}, -z_\ell)$. This gives a second $\log k_\ell^a$ field on the fine grid, and then the multilevel estimator can be based on the average of the two outputs obtained on the fine grid, minus the output obtained on the coarse grid using $\log k_{\ell-1}$. Unfortunately, numerical experiments indicate it gives little benefit; it is mentioned here as another

illustration of an antithetic estimator, and as a warning that it does not always yields significant benefits.

The numerical analysis of the multilevel approach for these elliptic SPDE applications is challenging because the diffusivity is unbounded, but Charrier, Scheichl and Teckentrup [13] have successfully analysed it for certain output functionals, and Teckentrup et al. have further developed the analysis for other output functionals and more general log-normal diffusivity fields [51].

4.2 Parabolic SPDE

Giles and Reisinger [25] consider an unusual SPDE from credit default modelling,

$$\mathrm{d}p = -\mu \frac{\partial p}{\partial x}\,\mathrm{d}t + \frac{1}{2}\frac{\partial^2 p}{\partial x^2}\,\mathrm{d}t - \sqrt{\rho}\frac{\partial p}{\partial x}\,\mathrm{d}M_t, \quad x > 0$$

subject to boundary condition $p(0,t) = 0$. Here $p(x,t)$ represents the probability density function for firms being a distance x from default at time t. The diffusive term is due to idiosyncratic factors affecting individual firms, while the stochastic term due to the scalar Brownian motion M_t corresponds to the systemic movement due to random market effects affecting all firms. The payoff corresponds to different tranches of a credit derivative which depends on the integral $\int_0^\infty p(x,t)\,\mathrm{d}x$ at a set of discrete times.

A Milstein time discretisation with timestep k, and a central space discretisation of the spatial derivatives with uniform spacing h gives the numerical approximation

$$\begin{aligned} p_j^{n+1} = p_j^n &- \frac{\mu\, k + \sqrt{\rho\, k}\, Z_n}{2h}\left(p_{j+1}^n - p_{j-1}^n\right) \\ &+ \frac{(1-\rho)\, k + \rho\, k\, Z_n^2}{2h^2}\left(p_{j+1}^n - 2p_j^n + p_{j-1}^n\right) \end{aligned}$$

where $p_j^n \approx p(j\,h, n\,k)$, and the Z_n are standard Normal random variables so that $\sqrt{h}\,Z_n$ corresponds to an increment of the driving scalar Brownian motion.

The multilevel implementation is very straightforward, with $k_\ell = k_{\ell-1}/2$ and $h_\ell = h_{\ell-1}/4$ due to numerical stability considerations which are analysed in the paper. As with SDEs, the coupling between the coarse and fine samples comes from summing the fine path Brownian increments in pairs to give the increments for the coarse path. The computational cost increases by factor 8 on each level, and numerical experiments indicate that the variance decreases by factor 8, so the overall computational complexity to achieve an $O(\varepsilon)$ RMS error is again $O(\varepsilon^{-2}(\log\varepsilon)^2)$.

5 Continuous-Time Markov Chain Simulation

Anderson and Higham have recently developed a very interesting new application of multilevel to continuous-time Markov Chain simulation [2]. Although they present their work in the context of stochastic chemical reactions, when species concentrations are extremely low and so stochastic effects become significant, they point out that the method has wide applicability in other areas.

In the simplest case of a single chemical reaction, the "tau-leaping" method (which is essentially the Euler-Maruyama method, approximating the reaction rate as being constant throughout the timestep) gives the discrete equation

$$\mathbf{x}_{n+1} = \mathbf{x}_n + P(h\ \lambda(\mathbf{x}_n)),$$

where h is the timestep, $\lambda(\mathbf{x}_n)$ is the reaction rate (or propensity function), and $P(t)$ represents a unit-rate Poisson random variable over time interval t.

If this equation defines the fine path in the multilevel simulation, then the coarse path, with double the timestep, is given by

$$\mathbf{x}^c_{n+2} = \mathbf{x}^c_n + P(2h\ \lambda(\mathbf{x}^c_n))$$

for even timesteps n. The question then is how to couple the coarse and fine path simulations.

The key observation by Anderson and Higham [2] is that for any $t_1, t_2 > 0$, the sum of two independent Poisson variates $P(t_1)$, $P(t_2)$ is equivalent in distribution to $P(t_1+t_2)$. Based on this, the first step is to express the coarse path Poisson variate as the sum of two Poisson variates, $P(h\ \lambda(\mathbf{x}^c_n))$ corresponding to the first and second fine path timesteps. For the first of the two fine timesteps, the coarse and fine path Poisson variates are coupled by defining two Poisson variates based on the minimum of the two reactions rates, and the absolute difference,

$$P_1 = P\left(h \min(\lambda(\mathbf{x}_n), \lambda(\mathbf{x}^c_n))\right), \quad P_2 = P\left(h\,|\lambda(\mathbf{x}_n) - \lambda(\mathbf{x}^c_n)|\right),$$

and then using P_1 as the Poisson variate for the path with the smaller rate, and $P_1 + P_2$ for the path with the larger rate. This elegant approach naturally gives a small difference in the Poisson variates when the difference in rates is small, and leads to a very effective multilevel algorithm.

In their paper [2], Anderson and Higham treat more general systems with multiple reactions, and include an additional coupling at the finest level to an SSA (Stochastic Simulation Algorithm) computation, so that their overall multilevel estimator is unbiased, unlike the estimators discussed earlier for SDEs. Finally, they give a complete numerical analysis of the variance of their multilevel algorithm.

Because stochastic chemical simulations typically involve 1000's of reactions, the multilevel method is particularly effective in this context, providing computational savings in excess of a factor of 100 [2].

6 Wasserstein Metric

In the multilevel treatment of SDEs, the Brownian or Lévy increments for the coarse path are obtained by summing the increments for the fine path. Similarly, in the Markov Chain treatment, the Poisson variate for the coarse timestep is defined as the sum of two Poisson variates for fine timesteps.

This sub-division of coarse path random variable into the sum of two fine path random variables should work in many settings. The harder step in more general applications is likely to be the second step in the Markov Chain treatment, tightly coupling the increments used for the fine and coarse paths over the same fine timestep.

The general statement of this problem is the following: given two very similar scalar probability distributions, we want to obtain samples Z_f, Z_c from each in a way which minimises $\mathbb{E}[\,|Z_f - Z_c|^p]$. This corresponds precisely to the Wasserstein metric which defines the "distance" between two probability distributions as

$$\left(\inf_{\gamma} \int \left\| Z_f - Z_c \right\|^p \, d\gamma(Z_f, Z_c)\right)^{1/p},$$

where the minimum is over all joint distributions with the correct marginals. In 1D, the Wasserstein metric is equal to

$$\left(\int_0^1 \left|\Phi_f^{-1}(u) - \Phi_c^{-1}(u)\right|^p \, \mathrm{d}u\right)^{1/p},$$

where Φ_f and Φ_c are the cumulative probability distributions for Z_f and Z_c [8], and this minimum is achieved by choosing $Z_f = \Phi_f^{-1}(U)$, $Z_c = \Phi_c^{-1}(U)$, for the same uniform $[0, 1]$ random variable U. This suggests this may be a good general technique for future multilevel applications, provided one is able to invert the relevant cumulative distributions, possibly through generating appropriate spline approximations.

7 Other Uses of Multilevel

7.1 Nested Simulation

The pricing of American options is one of the big challenges for Monte Carlo methods in computational finance, and Belomestny and Schoenmakers have recently written a very interesting paper on the use of multilevel Monte Carlo for this purpose [7]. Their method is based on Anderson and Broadie's dual simulation method [1]

in which a key component at each timestep in the simulation is to estimate a conditional expectation using a number of sub-paths.

In their multilevel treatment, Belomestny and Schoenmakers use the same uniform timestep on all levels of the simulation. The quantity which changes between different levels of simulation is the number of sub-samples used to estimate the conditional expectation. To couple the coarse and fine levels, the fine level uses N_ℓ sub-samples, and the coarse level uses $N_{\ell-1} = N_\ell/2$ of them.

Related unpublished research by N. Chen for a similar multilevel treatment of nested simulation found that the multilevel correction variance is reduced if the payoff on the coarse level is replaced by an average of the payoffs obtained using the first $N_\ell/2$ and the second $N_\ell/2$ samples. This is similar in some ways to the antithetic approach described earlier.

In future research, Belomestny and Schoenmakers intend to also change the number of timesteps on each level, to increase the overall computational benefits of the multilevel approach.

7.2 Truncated Series Expansions

Building on earlier work by Broadie and Kaya [11], Glasserman and Kim have recently developed an efficient method [30] of exactly simulating the Heston stochastic volatility model [38]. The key to their algorithm is a method of representing the integrated volatility over a time interval $[0, T]$, conditional on the initial and final values, v_0 and v_T as

$$\left(\int_0^T V_s \, ds \;\middle|\; V_0 = v_0, V_T = v_T \right) \overset{d}{=} \sum_{n=1}^{\infty} x_n + \sum_{n=1}^{\infty} y_n + \sum_{n=1}^{\infty} z_n$$

where x_n, y_n, z_n are independent random variables.

In practice, they truncate the series expansions at a level which ensures the desired accuracy, but a more severe truncation would lead to a tradeoff between accuracy and computational cost. This makes the algorithm a candidate for a multilevel treatment in which the level ℓ computation performs the truncation at N_ℓ, so the level ℓ computation would use

$$\sum_{n=1}^{N_\ell} x_n + \sum_{n=1}^{N_\ell} y_n + \sum_{n=1}^{N_\ell} z_n$$

while the level $\ell-1$ computation would truncate the summations at $N_{\ell-1}$, but would use the same random variables x_n, y_n, z_n for $1 \le n \le N_{\ell-1}$.

This kind of multilevel treatment has not been tested experimentally, but it seems that it might yield some computational savings even though Glasserman and Kim typically only need to retain ten terms in their summations through the use of a

carefully constructed estimator for the truncated remainder. The savings may be larger in other circumstances which require more terms to be retained for the desired accuracy.

7.3 Mixed Precision Arithmetic

The final example of the use of multilevel is unusual, because it concerns the computer implementation of Monte Carlo algorithms. In the latest CPUs from Intel and AMD, each core has a vector unit which can perform eight single precision or four double precision operations with one instruction. Also, double precision data takes twice as much time to transfer as single precision data. Hence, single precision computations can be twice as fast as double precision on CPUs, and the difference can be even greater on GPUs. This raises the question of whether single precision arithmetic is sufficient for Monte Carlo simulation.

My view is that it usually is since the finite precision rounding errors are smaller than the other sources of error: statistical error due to Monte Carlo sampling; bias due to SDE discretisation; model uncertainty. However, there can be significant errors when averaging unless one uses binary tree summation [39] to perform the summation, and in addition computing sensitivities by perturbing input parameters (so-called "bumping") can greatly amplify the rounding errors.

The best solution is perhaps to use double precision for the final averaging, and pathwise sensitivity analysis or the likelihood ratio method for computing sensitivities, but if there remains a need for the path simulation to be performed in double precision then one could use the two-level MLMC approach in which level 0 corresponds to single precision and level 1 corresponds to double precision, with the same random numbers being used for both.

7.4 Multiple Outputs

In all of the discussion so far, we have been concerned with a single expectation arising from a stochastic simulation. However, there are often times when one wishes to estimate the expected value of multiple outputs.

Extending the analysis in Sect. 1.2, when using multilevel to estimate M different expectations, using N_l samples on each level, the goal is to achieve an acceptably small variance for each output

$$\sum_{\ell=0}^{L} N_\ell^{-1} V_{\ell,m} \;\leq\; \varepsilon_m^2, \qquad m = 1, \ldots, M,$$

with the desired accuracy ε_m being allowed to vary from one output to another, and to do so with the minimum computational cost which is given as usual as

$$\sum_{\ell=0}^{L} N_\ell \, C_\ell,$$

assuming that the cost of computing the output functions is negligible compared to the cost of obtaining the stochastic sample (e.g. through an SDE path simulation).

This leads naturally to a constrained optimisation problem with a separate Lagrange multiplier for each output. However, a much simpler idea, due to Tigran Nagapetyan, which in practice is almost always equivalent, is to define

$$V_\ell = \max_m \frac{V_{\ell,m}}{\varepsilon_m^2}$$

and make the variance constraint $\sum_{\ell=0}^{L} N_\ell^{-1} \, V_\ell \;\leq\; 1$.

This is sufficient to ensure that all of the individual constraints are satisfied, and we can then use the standard approach with a single Lagrange multiplier. This multi-output approach is currently being investigated by Nagapetyan, Ritter and the author for the approximation of cumulative distribution functions and probability density functions arising from stochastic simulations.

8 Conclusions

In the past 6 years, considerable progress has been achieved with the multilevel Monte Carlo method for a wide range of applications. This review has attempted to emphasise the conceptual simplicity of the multilevel approach; in essence it is simply a recursive control variate strategy, using cheap approximations to some random output quantity as a control variate for more accurate but more costly approximations.

In practice, the challenge is to develop a tight coupling between successive approximation levels, to minimise the variance of the difference in the output obtained from each level. In the context of SDE and SPDE simulations, strong convergence properties are often relied on to obtain a small variance between coarse and fine simulations. In the specific context of a digital option associated with a Brownian SDE, three treatments were described to effectively smooth the output: a analytic conditional expectation, a "splitting" approximation, and a change of measure. Similar treatments have been found to be helpful in other contexts.

Overall, multilevel methods are being used for an increasingly wide range of applications. The biggest savings are in situations in which the coarsest approximation is very much cheaper than the finest. So far, this includes multi-dimensional

SPDEs, and chemical stochastic simulations with 1000's of timesteps. In SDE simulations which perhaps only require 32 timesteps for the desired level of accuracy, the potential savings are naturally quite limited.

Although this is primarily a survey article, a few new ideas have been introduced:

- Equation (1) giving the total computational cost required for a general unbiased multilevel estimator is new, as is the discussion which follows it, although the underlying analysis is not;
- Based on the 1D Wasserstein metric, it seems that inverting the relevant cumulative distributions may be a good way to couple fine and coarse level simulations in multilevel implementations;
- The multilevel approach could be used in applications which involve the truncation of series expansions;
- A two-level method combining single and double precision computations might provide useful savings, due to the lower cost of single precision arithmetic;
- A multilevel approach for situations with multiple expectations to be estimated.

Looking to the future, exciting areas for further research include:

- More use of multilevel for nested simulations;
- Further investigation of multilevel quasi-Monte Carlo methods;
- Continued research on numerical analysis, especially for SPDEs;
- Development of multilevel estimators for new applications.

For further information on multilevel Monte Carlo methods, see the webpage
http://people.maths.ox.ac.uk/gilesm/mlmc_community.html
which lists the research groups working in the area, and their main publications.

References

1. Andersen, L., Broadie, M.: A primal-dual simulation algorithm for pricing multi-dimensional American options. Management Science, **50**, 1222–1234 (2004)
2. Anderson, D., Higham, D.J.: Multi-level Monte Carlo for continuous time Markov chains with applications in biochemical kinetics. SIAM Multiscale Model. Simul. **10**, 146–179 (2012)
3. Asmussen, A., Glynn, P.: Stochastic Simulation. Springer, New York (2007)
4. Avikainen, R.: On irregular functionals of SDEs and the Euler scheme. Finance Stoch. **13**, 381–401 (2009)
5. Barth, A., Lang, A.: Multilevel Monte Carlo method with applications to stochastic partial differential equations. Int. J. Comput. Math. **89**, 2479–2498 (2012)
6. Barth, A., Schwab, C., Zollinger, N.: Multi-level Monte Carlo finite element method for elliptic PDEs with stochastic coefficients. Numer. Math. **119**, 123–161 (2011)
7. Belomestny, D., Schoenmakers, J.: Multilevel dual approach for pricing American style derivatives. Finance Stoch. **17**, 717–742 (2013)
8. Bickel, P.J., Freedman, D.A.: Some asymptotic theory for the bootstrap. Ann. Statist. **9**, 1196–1217 (1981)

9. Brandt, A., Galun, M., Ron, D.: Optimal multigrid algorithms for calculating thermodynamic limits. J. Stat. Phys. **74**, 313–348 (1994)
10. Brandt, A., Ilyin, V.: Multilevel Monte Carlo methods for studying large scale phenomena in fluids. J. Mol. Liquids **105**, 245–248 (2003)
11. Broadie, M., Kaya, O.: Exact simulation of stochastic volatility and other affine jump diffusion processes. Oper. Res. **54**, 217–231 (2006)
12. Burgos, S., Giles, M.B.: Computing Greeks using multilevel path simulation. In: Plaskota, L., Woźniakowski, H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2010, pp. 281–296. Springer, Berlin/Heidelberg (2012)
13. Charrier, J., Scheichl, R., Teckentrup, A.: Finite element error analysis of elliptic PDEs with random coefficients and its application to multilevel Monte Carlo methods. SIAM J. Numer. Anal. **51**, 322–352 (2013)
14. Clark, J.M.C., Cameron, R.J.: The maximum rate of convergence of discrete approximations for stochastic differential equations. In: Grigelionis, B. (ed.) Stochastic Differential Equations. Lecture Notes in Control and Information Sciences, vol. 25. Springer, New York (1980)
15. Cliffe, K.A., Giles, M.B., Scheichl, R., Teckentrup, A.: Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients. Comput. Vis. Sci. **14**, 3–15 (2011)
16. Cont, R., Tankov, P.: Financial Modelling with Jump Processes. CRC, Boca Raton (2004)
17. Dereich, S.: Multilevel Monte Carlo algorithms for Lévy-driven SDEs with Gaussian correction. Ann. Appl. Probab. **21**, 283–311 (2011)
18. Dereich, S., Heidenreich, F.: A multilevel Monte Carlo algorithm for Lévy-driven stochastic differential equations. Stochastic Process. Appl. **121**, 1565–1587 (2011)
19. Dietrich, C.R., Newsam, G.H.: Fast and exact simulation of stationary Gaussian processes through circulant embedding of the covariance matrix. SIAM J. Sci. Comput. **18**, 1088–1107 (1997)
20. Giles, M.B.: Improved multilevel Monte Carlo convergence using the Milstein scheme. In: Keller, A., Heinrich, S., Niederreiter, H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2006, pp. 343–358. Springer, New York (2008)
21. Giles, M.B.: Multilevel Monte Carlo path simulation. Oper. Res. **56**, 607–617 (2008)
22. Giles, M.B.: Multilevel Monte Carlo for basket options. In: Rossetti, M.D., Hill, R.R., Johansson, B., Dunkin, A., Ingalls, R.G. (eds.) Proceedings of the 2009 Winter Simulation Conference, Austin, pp. 1283–1290. IEEE (2009)
23. Giles, M.B., Debrabant, K., Rößler, A.: Numerical analysis of multilevel Monte Carlo path simulation using the Milstein discretisation. ArXiv preprint: 1302.4676 (2013)
24. Giles, M.B., Higham, D.J., Mao, X.: Analysing multilevel Monte Carlo for options with non-globally Lipschitz payoff. Finance Stoch. **13**, 403–413 (2009)
25. Giles, M.B., Reisinger, C.: Stochastic finite differences and multilevel Monte Carlo for a class of SPDEs in finance. SIAM J. Financial Math. **3**, 572–592 (2012)
26. Giles, M.B., Szpruch, L.: Antithetic multilevel Monte Carlo estimation for multi-dimensional SDEs without Lévy area simulation. Ann. Appl. Probab. (2013, to appear)
27. Giles, M.B., Szpruch, L.: Antithetic multilevel Monte Carlo estimation for multidimensional SDEs. In: Dick, J., Kuo, F.Y., Peters, G.W., Sloan, I.H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2012, this volume 367–384. Springer, Berlin/Heidelberg (2013)
28. Giles, M.B., Waterhouse, B.J.: Multilevel quasi-Monte Carlo path simulation. In: Advanced Financial Modelling, Radon Series on Computational and Applied Mathematics, pp. 165–181. De Gruyter, Berlin (2009)
29. Glasserman, P.: Monte Carlo Methods in Financial Engineering. Springer, New York (2004)
30. Glasserman, P., Kim, K.-K.: Gamma expansion of the Heston stochastic volatility model. Finance Stoch. **15**, 267–296 (2011)
31. Glasserman, P., Merener, N.: Convergence of a discretization scheme for jump-diffusion processes with state-dependent intensities. Proc. R. Soc. Lond. A **460**, 111–127 (2004)
32. Graubner, S.: Multi-level Monte Carlo Methoden für Stochastische Partielle Differentialgleichungen. Diplomarbeit, TU Darmstadt (2008)

33. Heinrich, S.: Monte Carlo complexity of global solution of integral equations. J. Complexity **14**, 151–175 (1998)
34. Heinrich, S.: The multilevel method of dependent tests. In: Balakrishnan, N., Melas, V.B., Ermakov, S. (eds.) Advances in Stochastic Simulation Methods, pp. 47–61. Springer, New York (2000)
35. Heinrich, S.: Multilevel Monte Carlo Methods. Lecture Notes in Computer Science, vol. 2179, pp. 58–67. Springer, New York (2001)
36. Heinrich, S.: Monte Carlo approximation of weakly singular integral operators. J. Complexity **22**, 192–219 (2006)
37. Heinrich, S., Sindambiwe, E.: Monte Carlo complexity of parametric integration. J. Complexity **15**, 317–341 (1999)
38. Heston, S.I.: A closed-form solution for options with stochastic volatility with applications to bond and currency options. Rev. Financ. Stud. **6**, 327–343 (1993)
39. Higham, N.J.: The accuracy of floating point summation. SIAM J. Sci. Comput. **14**, 783–799 (1993)
40. Hoel, H., von Schwerin, E., Szepessy, A., Tempone, R.: Adaptive multilevel Monte Carlo simulation. In: Engquist, B., Runborg, O., Tsai, Y.-H.R. (eds.) Numerical Analysis of Multiscale Computations. Lecture Notes in Computational Science and Engineering, vol. 82, pp. 217–234. Springer, New York (2012)
41. Kebaier, A.: Statistical Romberg extrapolation: a new variance reduction method and applications to options pricing. Ann. Appl. Probab. **14**, 2681–2705 (2005)
42. Kloeden, P.E., Platen, E.: Numerical Solution of Stochastic Differential Equations. Springer, Berlin (1992)
43. Marxen, H.: The multilevel Monte Carlo method used on a Lévy driven SDE. Monte Carlo Methods Appl. **16**, 167–190 (2010)
44. Merton, R.C.: Option pricing when underlying stock returns are discontinuous. J. Finance **3**, 125–144 (1976)
45. Mishra, S., Schwab, C., Sukys, J.: Multi-level Monte Carlo finite volume methods for nonlinear systems of conservation laws in multi-dimensions. J. Comput. Phys. **231**, 3365–3388 (2012)
46. Müller-Gronbach, T., Ritter, K.: Variable subspace sampling and multi-level algorithms. In: L'Ecuyer, P., Owen, A. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2008, pp. 131–156. Springer, Berlin/Heidelberg (2010)
47. Platen, E., Bruti-Liberati, N.: Numerical Solution of Stochastic Differential Equations with Jumps in Finance. Springer, Berlin/Heidelberg (2010)
48. Schoutens, W.: Lévy Processes in Finance: Pricing Financial Derivatives. Wiley, Chichester/ New York (2003)
49. Speight, A.L.: A multilevel approach to control variates. J. Comput. Finance **12**, 1–25 (2009)
50. Speight, A.L.: Multigrid techniques in economics. Oper. Res. **58**, 1057–1078 (2010)
51. Teckentrup, A., Scheichl, R., Giles, M.B., Ullmann, E.: Further analysis of multilevel Monte Carlo methods for elliptic PDEs with random coefficients. Numer. Math. **125**, 569–600 (2013)
52. Xia, Y., Giles, M.B.: Multilevel path simulation for jump-diffusion SDEs. In: Plaskota, L., Woźniakowski, H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2010, pp. 695–708. Springer, Berlin/Heidelberg (2012)

Guaranteed Conservative Fixed Width Confidence Intervals via Monte Carlo Sampling

Fred J. Hickernell, Lan Jiang, Yuewei Liu, and Art B. Owen

Abstract Monte Carlo methods are used to approximate the means, μ, of random variables Y, whose distributions are not known explicitly. The key idea is that the average of a random sample, $Y_1, \ldots, Y_n$, tends to μ as n tends to infinity. This article explores how one can reliably construct a confidence interval for μ with a prescribed half-width (or error tolerance) ε. Our proposed two-stage algorithm assumes that the *kurtosis* of Y does not exceed some user-specified bound. An initial independent and identically distributed (IID) sample is used to confidently estimate the variance of Y. A Berry-Esseen inequality then makes it possible to determine the size of the IID sample required to construct the desired confidence interval for μ. We discuss the important case where $Y = f(\boldsymbol{X})$ and $\boldsymbol{X}$ is a random d-vector with probability density function ρ. In this case μ can be interpreted as the integral $\int_{\mathbb{R}^d} f(\boldsymbol{x})\rho(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$, and the Monte Carlo method becomes a method for multidimensional cubature.

1 Introduction

Monte Carlo algorithms provide a flexible way to approximate $\mu = \mathbb{E}(Y)$ when one can generate samples of the random variable Y. For example, Y might be the discounted payoff of some financial derivative, which depends on the future

F.J. Hickernell (✉) · L. Jiang
Department of Applied Mathematics, Illinois Institute of Technology, Chicago, IL 60616, USA
e-mail: hickernell@iit.edu; ljiang14@hawk.iit.edu

Y. Liu
School of Mathematics and Statistics, Lanzhou University, Lanzhou City, Gansu, China 730000
e-mail: lyw@lzu.edu.cn

A.B. Owen
Department of Statistics, Stanford University, Stanford, CA 94305, USA
e-mail: owen@stanford.edu

J. Dick et al. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2012*,
Springer Proceedings in Mathematics & Statistics 65,
DOI 10.1007/978-3-642-41095-6_5, © Springer-Verlag Berlin Heidelberg 2013

performance of assets that are described by a stochastic model. Then μ is the fair option price. The goal is to obtain a *confidence interval*

$$\Pr[|\mu - \hat{\mu}| \le \varepsilon] \ge 1 - \alpha, \tag{1}$$

where

- μ is approximated by the sample average of n independent and identically distributed (IID) samples of Y,

$$\hat{\mu} = \hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i, \tag{2}$$

- ε is the half-width of the confidence interval, which also serves as an *error tolerance*, and
- α is the level of *uncertainty*, e.g., 1 % or 0.1 %, which is fixed in advance.

Often the sample size, n, is fixed in advance, and the central limit theorem (CLT) provides an approximate value for ε in terms of n and

$$\sigma^2 = \mathrm{Var}(Y) = \mathbb{E}[(Y - \mu)^2], \tag{3}$$

which itself may be approximated by the sample variance. The goal here is somewhat different. We want to fix ε in advance and then determine how large the sample size must be to obtain a fixed width confidence interval of the form (1). Moreover, we want to make sure that our confidence interval is correct, not just approximately correct, or correct in the limit of vanishing ε. In this paper we present Algorithm 1 for obtaining such a fixed width confidence interval for the mean of a real random variable when one is performing Monte Carlo sampling.

Before presenting the method, we outline the reasons that existing fixed width confidence intervals are not suitable. In summary, there are two drawbacks of existing procedures. Much existing theory is *asymptotic*, i.e., the proposed procedure attains the desired coverage level in the limit as $\varepsilon \to 0$ but does not provide coverage guarantees for fixed $\varepsilon > 0$. We want such fixed ε guarantees. A second drawback is that the theory may make distributional assumptions that are too strong. In Monte Carlo applications one typically does not have much information about the underlying distribution. The form of the distribution for Y is generally not known, $\mathrm{Var}(Y)$ is generally not known, and Y is not necessarily bounded. We are aiming to derive fixed width confidence intervals that do not require such assumptions.

The width (equivalently length) of a confidence interval tends to become smaller as the number n of sampled function values increases. In special circumstances, we can choose n to get a confidence interval of at most the desired length and at least the desired coverage level, $1 - \alpha$. For instance, if the variance, $\sigma^2 = \mathrm{Var}(Y)$, is known then an approach based on Chebychev's inequality is available, though the actual coverage will usually be much higher than the nominal level, meaning that much

narrower intervals would have sufficed. Known variance in addition to a Gaussian distribution for Y supports a fixed width confidence interval construction that is not too conservative. The CLT provides a confidence interval that is asymptotically correct, but our aim is for something that is definitely correct for finite sample sizes. Finally, conservative fixed width confidence intervals for means can be constructed for bounded random variables, by appealing to exponential inequalities such as Hoeffding's or Chernoff's inequality. Unfortunately, Y is often unbounded, e.g., in the case where it represents the payoff of a call option.

If the relevant variance or bound is unknown, then approaches based on sequential statistics [24] may be available. In sequential methods one keeps increasing n until the interval is narrow enough. Sequential confidence intervals require us to take account of the stopping rule when computing the confidence level. Unfortunately, all existing sequential methods are lacking in some aspects.

Serfling and Wackerly [21] consider sequential confidence intervals for the mean (alternatively for the median) in parametric distributions, symmetric about their center point. The symmetry condition is not suitable for general purpose Monte Carlo applications.

Chow and Robbins [2] develop a sequential sampling fixed width confidence interval procedure for the mean, but its guarantees are only asymptotic (as $\varepsilon \to 0$). Mukhopadhyay and Datta [14] give a procedure similar to Chow and Robbins', and it has similar drawbacks.

Bayesian methods can support a fixed width interval containing μ with $1 - \alpha$ posterior probability, and Bayesian methods famously do not require one to account for stopping rules. They do however require strong distributional assumptions.

There is no assumption-free way to obtain exact confidence intervals for a mean, as has been known since Bahadur and Savage [1]. Some kind of assumption is needed to rule out settings where the desired quantity is the mean of a heavy tailed random variable in which rarely seen large values dominate the mean and spoil the estimate of the variance. The assumption we use is an upper bound on the modified kurtosis (normalized fourth moment) of the random variable Y:

$$\tilde{\kappa} = \frac{\mathbb{E}[(Y - \mu)^4]}{\sigma^4} \le \tilde{\kappa}_{\max}. \tag{4}$$

(The quantity $\tilde{\kappa} - 3$ is commonly called the kurtosis.) Under such an assumption we present a two-stage algorithm: the first stage generates a conservative upper bound on the variance, and the second stage uses this variance bound and a Berry-Esseen Theorem, which can be thought of as a non-asymptotic CLT, to determine how large n must be for the sample mean to satisfy confidence interval (1). Theorem 5 demonstrates the validity of the fixed width confidence interval, and Theorem 6 demonstrates that the cost of this algorithm is reasonable. These are our main new theoretical results.

Our procedure is a two-stage procedure rather than a fully sequential one. In this it is similar to the method of Stein [26, 27], except that the latter requires normally distributed data.

One might question whether assumption (4), which involves fourth moments of Y, is more reasonable than an assumption involving only the second moment of Y. For example, using Chebychev's inequality with the assumption

$$\sigma^2 \le \sigma_{\max}^2 \tag{5}$$

also yields a fixed width confidence interval of the form (1). We would argue that (4) is indeed more reasonable. First, if Y satisfies (4), then so does cY for any nonzero c, however, the analog does not hold for (5). In fact, if σ is nonzero, then (5) must be violated by cY for c sufficiently large. Second, making $\tilde{\kappa}_{\max}$ a factor of 10 or 100 larger than $\tilde{\kappa}$ does not significantly affect the total cost (number of samples required) of our two-stage Monte Carlo Algorithm 1 for a large range of values of σ/ε. However, the cost of our Monte Carlo algorithm, and indeed any Monte Carlo algorithm based on IID sampling is proportional to σ^2, so overestimating σ^2 by a factor of 10 or 100 or more to be safe increases the cost of the algorithm by that factor.

An important special case of computing $\mu = \mathbb{E}(Y)$ arises in the situation where $Y = f(\boldsymbol{X})$ for some function $f : \mathbb{R}^d \to \mathbb{R}$ and some random vector $\boldsymbol{X}$ with probability density function $\rho : \mathbb{R}^d \to [0, \infty)$. One may then interpret the mean of Y as the multidimensional integral

$$\mu = \mu(f) = \mathbb{E}(Y) = \int_{\mathbb{R}^d} f(\boldsymbol{x})\rho(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}. \tag{6}$$

Note that unlike the typical probability and statistics setting, where f denotes a probability density function, in this paper f denotes an integrand, and ρ denotes the probability density function. Given the problem of evaluating $\mu = \int_{\mathbb{R}^d} g(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$, one must choose a probability density function ρ for which one can easily generate random vectors $\boldsymbol{X}$, and then set $f = g/\rho$. The quantities σ^2 and $\tilde{\kappa}$ defined above can be written in terms of weighted $\mathcal{L}_p$-norms of f:

$$\|f\|_p := \left\{ \int_{\mathbb{R}^d} |f(\boldsymbol{x})|^p \, \rho(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \right\}^{1/p}, \qquad \sigma^2 = \|f - \mu\|_2^2, \qquad \tilde{\kappa} = \frac{\|f - \mu\|_4^4}{\|f - \mu\|_2^4}. \tag{7}$$

For a given g, the choice of ρ is not unique, and making an optimal choice belongs to the realm of *importance sampling*. The assumption of bounded kurtosis, (4), required by Algorithm 1, corresponds to an assumption that the integrand f lies in the *cone* of functions

$$\mathcal{C}_{\tilde{\kappa}_{\max}} = \{f \in \mathcal{L}_4 : \|f - \mu(f)\|_4 \le \tilde{\kappa}_{\max}^{1/4} \|f - \mu(f)\|_2\}. \tag{8}$$

This is in contrast to a *ball* of functions, which would be the case if one was satisfying a bounded variance condition, (5).

From the perspective of numerical analysis, if ρ has independent marginals, one may apply a product form of a univariate quadrature rule to evaluate μ. However, this consumes a geometrically increasing number of samples as d increases, and moreover, such methods often require rather strict smoothness assumptions on f.

If f satisfies moderate smoothness conditions, then (randomized) quasi-Monte Carlo methods, or low discrepancy sampling methods for evaluating μ are more efficient than simple Monte Carlo [3, 9, 16, 25]. Unfortunately, practical error estimation remains a challenge for quasi-Monte Carlo methods. Heuristic methods have been proposed, but they lack theoretical justification. One such heuristic is used with reasonable success in the numerical examples of Sect. 4. Independent randomizations of quasi-Monte Carlo rules of fixed sample size can be used to estimate their errors, but they do not yet lead to guaranteed, fixed width confidence intervals.

Computational mathematicians have also addressed the problem of constructing automatic algorithms, i.e., given an error tolerance of ε, one computes an approximation, $\hat{\mu}$, based on n evaluations of the integrand f, such that $|\mu - \hat{\mu}| \leq \varepsilon$. For example, MATLAB [28], a popular numerical package, contains `quad`, an adaptive Simpson's rule for univariate quadrature routine developed by Gander and Gautschi [4]. Although `quad` and other automatic rules generally work well in practice, they do not have any rigorous guarantees that the error tolerance is met, and it is relatively simple to construct functions that fool them. This is discussed in Sect. 4. Since a random algorithm, like Monte Carlo, gives a random answer, any statements about satisfying an error criterion must be probabilistic. This leads us back to the problem of finding a fixed width confidence interval, (1).

An outline of this paper follows. Section 2 defines key terminology and provides certain inequalities used to construct our fixed width confidence intervals. The new two-stage Algorithm 1 is described in Sect. 3, where rigorous guarantees of its success and its cost are provided. Section 4 illustrates the challenges of computing μ to a guaranteed precision through several numerical examples. This paper ends with a discussion of our results and further work to be done.

2 Background Probability and Statistics

In our Monte Carlo applications, a quantity of interest is written as an expectation: $\mu = \mathbb{E}(Y)$, where Y is a real valued random variable. As mentioned above, very often $Y = f(\boldsymbol{X})$ where $\boldsymbol{X} \in \mathbb{R}^d$ is a random vector with probability density function ρ. In other settings the random quantity $\boldsymbol{X}$ might have a discrete distribution or be infinite dimensional (e.g., a Gaussian process) or both. For Monte Carlo estimation, we can work with the distribution of Y alone. The Monte Carlo estimate of μ is the sample mean, as given in (2), where the Y_i are IID random variables with the same distribution as Y.

2.1 Moments

Our methods require conditions on the first four moments of Y as described here. The variance of Y, as defined in (3), is denoted by σ^2, and its non-negative square root, σ, is the standard deviation of Y. Some of our expressions assume without stating it that $\sigma > 0$, and all will require $\sigma < \infty$. The skewness of Y is $\gamma = \mathbb{E}[(Y-\mu)^3]/\sigma^3$, and the kurtosis of Y is $\kappa = \tilde{\kappa} - 3 = \mathbb{E}[(Y-\mu)^4]/\sigma^4 - 3$ (see (4)). The mysterious 3 in κ is there to make it zero for Gaussian random variables. Also, $\mu, \sigma^2, \gamma, \kappa$ are related to the first four cumulants [12, Chap. 2] of the distribution of Y, meaning that

$$\log(\mathbb{E}[\exp(tY)]) = \mu t + \frac{\sigma^2 t^2}{2} + \frac{\gamma \sigma^3 t^3}{3!} + \frac{\kappa \sigma^4 t^4}{4!} + o(t^4).$$

Our main results require a known upper bound for κ, which then implies that σ and γ are finite.

2.2 CLT Intervals

A random variable Z has the standard normal distribution, denoted by $\mathcal{N}(0,1)$, if

$$\Pr(Z \le z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp(-t^2/2)\,\mathrm{d}t =: \Phi(z).$$

Under the central limit theorem, the distribution of $\sqrt{n}(\hat{\mu}_n - \mu)/\sigma$ approaches $\mathcal{N}(0,1)$ as $n \to \infty$, where $\hat{\mu}_n$ denotes the sample mean of n IID samples. As a result

$$\Pr\bigl(\hat{\mu}_n - 2.58\sigma/\sqrt{n} \le \mu \le \hat{\mu}_n + 2.58\sigma/\sqrt{n}\bigr) \to 0.99 \tag{9}$$

as $n \to \infty$. We write the interval in (9) as $\hat{\mu}_n \pm 2.58\sigma/\sqrt{n}$. Equation (9) cannot be used when σ^2 is unknown, but the usual estimate

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \hat{\mu}_n)^2 \tag{10}$$

may be substituted, yielding the interval $\hat{\mu}_n \pm 2.58 s_n/\sqrt{n}$ which also satisfies the limit in (9) by Slutsky's theorem [8]. For an arbitrary confidence level $1-\alpha \in (0,1)$, we replace the constant 2.58 by $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$. The width of this interval is $2z_{\alpha/2}s_n/\sqrt{n}$, and when μ is in the interval then the absolute error $|\mu - \hat{\mu}_n| \le \varepsilon := z_{\alpha/2}s_n/\sqrt{n}$.

The coverage level of the CLT interval is only asymptotic. In more detail, Hall [6, p. 948] shows that

$$\Pr\bigl(|\mu - \hat{\mu}_n| \le 2.58s/\sqrt{n}\bigr) = 0.99 + \frac{1}{n}(A + B\gamma^2 + C\kappa) + O\Bigl(\frac{1}{n^2}\Bigr) \qquad (11)$$

for constants A, B, and C that depend on the desired coverage level (here 99 %). Hall's theorem requires only that the random variable Y has sufficiently many finite moments and is not supported solely on a lattice (such as the integers). It is interesting to note that the $O(1/n)$ coverage error in (11) is better than the $O(1/\sqrt{n})$ root mean squared error for the estimate $\hat{\mu}_n$ itself.

2.3 Standard Probability Inequalities

Here we present some well known inequalities that we will use. First, Chebychev's inequality ensures that a random variable (such as $\hat{\mu}_n$) is seldom too far from its mean.

Theorem 1 (Chebychev's Inequality). *[10, 6.1c, p. 52] Let Z be a random variable with mean μ and variance $\sigma^2 \ge 0$. Then for all $\varepsilon > 0$,*

$$\Pr[|Z - \mu| \ge \varepsilon] \le \frac{\sigma^2}{\varepsilon^2}.$$

In some settings we need a one sided inequality like Chebychev's. We will use this one due to Cantelli.

Theorem 2 (Cantelli's Inequality). *[10, 6.1e, p. 53] Let Z be any random variable with mean μ and finite variance σ^2. For any $a \ge 0$, it follows that:*

$$\Pr[Z - \mu \ge a] \le \frac{\sigma^2}{a^2 + \sigma^2}.$$

Berry-Esseen type theorems govern the rate at which a CLT takes hold. We will use the following theorem which combines recent work on both uniform and non-uniform (x-dependent right hand side) versions.

Theorem 3 (Berry-Esseen Inequality). *Let $Y_1, \ldots, Y_n$ be IID random variables with mean μ, variance $\sigma^2 > 0$, and third centered moment $M_3 = E\,|Y_i - \mu|^3 / \sigma^3 < \infty$. Let $\hat{\mu}_n = (Y_1 + \cdots + Y_n)/n$ denote the sample mean. Then*

$$\left|\Pr\left[\frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}} < x\right] - \Phi(x)\right|$$

$$\le \Delta_n(x, M_3) := \frac{1}{\sqrt{n}} \min\left(A_1(M_3 + A_2), \frac{A_3 M_3}{1 + |x|^3}\right) \qquad \forall x \in \mathbb{R},$$

where $A_1 = 0.3328$ *and* $A_2 = 0.429$ *[23], and* $A_3 = 18.1139$ *[15].*

The constants in the Berry-Esseen Inequality above have been an area of active research. We would not be surprised if there are further improvements in the near future.

Our method requires probabilistic bounds on the sample variance, s_n^2. For that, we will use some moments of the variance estimate.

Theorem 4. *[13, Eq. (7.16), p. 265] Let* $Y_1, \ldots, Y_n$ *be IID random variables with variance* σ^2 *and modified kurtosis* $\tilde{\kappa}$ *defined in* (4)*. Let* s_n^2 *be the sample variance as defined in* (10)*. Then the sample variance is unbiased,* $\mathbb{E}(s_n^2) = \sigma^2$*, and its variance is*

$$\mathrm{Var}(s_n^2) = \frac{\sigma^4}{n}\left(\tilde{\kappa} - \frac{n-3}{n-1}\right).$$

3 Two-Stage Confidence Interval

Our two-stage procedure works as follows. In the first stage, we take a sample of independent values $Y_1, \ldots, Y_{n_\sigma}$ from the distribution of Y. From this sample we compute the sample variance, $s_{n_\sigma}^2$, according to (10) and estimate the variance of Y_i by $\hat{\sigma}^2 = \mathfrak{C}^2 \hat{s}_{n_\sigma}^2$, where $\mathfrak{C}^2 > 1$ is a "variance inflation factor" that will reduce the probability that we have underestimated $\sigma^2 = \mathrm{Var}(Y)$. For the second stage, we use the estimate $\hat{\sigma}^2$ as if it were the true variance of Y_i and use Berry-Esseen theorem to obtain a suitable sample size, n_μ, for computing the sample average, $\hat{\mu}$, that satisfies the fixed width confidence interval (1).

The next two subsections give details of these two steps that will let us bound their error probabilities. Then we give a theorem on the method as a whole.

3.1 Conservative Variance Estimates

We need to ensure that our first stage estimate of the variance σ^2 is not too small. The following result bounds the probability of such an underestimate.

Lemma 1. *Let* $Y_1, \ldots, Y_n$ *be IID random variables with variance* $\sigma^2 > 0$ *and kurtosis* κ*. Let* s_n^2 *be the sample variance defined at* (10)*, and let* $\tilde{\kappa} = \kappa + 3$*. Then*

$$\Pr\left[s_n^2 < \sigma^2 \left\{1 + \sqrt{\left(\tilde{\kappa} - \frac{n-3}{n-1}\right)\left(\frac{1-\alpha}{\alpha n}\right)}\right\}\right] \ge 1 - \alpha, \tag{12a}$$

$$\Pr\left[s_n^2 > \sigma^2\left\{1 - \sqrt{\left(\tilde{\kappa} - \frac{n-3}{n-1}\right)\left(\frac{1-\alpha}{\alpha n}\right)}\right\}\right] \geq 1 - \alpha. \tag{12b}$$

Proof. Applying Theorem 4 and choosing

$$a = \sqrt{\mathrm{Var}(s_n^2)\frac{1-\alpha}{\alpha}} = \sigma^2\sqrt{\left(\tilde{\kappa} - \frac{n-3}{n-1}\right)\left(\frac{1-\alpha}{\alpha n}\right)} > 0,$$

it follows from Cantelli's inequality (Theorem 2) that

$$\begin{aligned}\Pr\left[s_n^2 - \sigma^2 \geq \sigma^2\sqrt{\left(\tilde{\kappa} - \frac{n-3}{n-1}\right)\left(\frac{1-\alpha}{\alpha n}\right)}\right] &= \Pr\left[s_n^2 - \sigma^2 \geq a\right] \\ &\leq \frac{\mathrm{Var}(s_n^2)}{a^2 + \mathrm{Var}(s_n^2)} = \frac{\mathrm{Var}(s_n^2)}{\mathrm{Var}(s_n^2)\frac{1-\alpha}{\alpha} + \mathrm{Var}(s_n^2)} = \frac{1}{\left(\frac{1-\alpha}{\alpha}\right) + 1} = \alpha.\end{aligned}$$

Then (12a) follows directly. By a similar argument, applying Cantelli's inequality to the expression $\Pr\left[-s_n^2 + \sigma^2 \geq a\right]$ implies (12b). □

Using Lemma 1 we can bound the probability that $\hat{\sigma}^2 = \mathfrak{C}^2 s_{n_\sigma}^2$ overestimates σ^2. Equation (12a) implies that

$$\Pr\left[\frac{s_{n_\sigma}^2}{1 - \sqrt{\left(\tilde{\kappa} - \frac{n_\sigma - 3}{n_\sigma - 1}\right)\left(\frac{1-\alpha}{\alpha n_\sigma}\right)}} > \sigma^2\right] \geq 1 - \alpha.$$

Thus, it makes sense for us to require the modified kurtosis, $\tilde{\kappa}$, to be small enough, relative to n_σ, α, and $\mathfrak{C}$, in order to ensure that $\Pr(\hat{\sigma}^2 > \sigma^2) \geq 1 - \alpha$. Specifically, we require

$$\frac{1}{1 - \sqrt{\left(\tilde{\kappa} - \frac{n_\sigma - 3}{n_\sigma - 1}\right)\left(\frac{1-\alpha}{\alpha n_\sigma}\right)}} \leq \mathfrak{C}^2,$$

or equivalently,

$$\tilde{\kappa} \leq \frac{n_\sigma - 3}{n_\sigma - 1} + \left(\frac{\alpha n_\sigma}{1-\alpha}\right)\left(1 - \frac{1}{\mathfrak{C}^2}\right)^2 =: \tilde{\kappa}_{\max}(\alpha, n_\sigma, \mathfrak{C}). \tag{13}$$

This condition is the explicit version of (4) mentioned in the introduction.

3.2 Conservative Interval Widths

Here we consider how to choose the sample size n_μ to get the desired coverage level from an interval with half-length at most ε. We suppose here that σ is known. In practice we will use a conservative (biased high) estimate for σ.

First, if the CLT held exactly and not just asymptotically, then we could use a CLT sample size of

$$N_{\text{CLT}}(\varepsilon, \sigma, \alpha) = \left\lceil \left(\frac{z_{\alpha/2}\sigma}{\varepsilon} \right)^2 \right\rceil$$

independent values of Y_i in an interval like the one in (9).

Given knowledge of σ, but no assurance of a Gaussian distribution for $\hat{\mu}_n$, we could instead select a sample size based on Chebychev's inequality (Theorem 1). Taking

$$N_{\text{Cheb}}(\varepsilon, \sigma, \alpha) = \left\lceil \frac{\sigma^2}{\alpha\varepsilon^2} \right\rceil \tag{14}$$

IID observations of Y gives the confidence interval (1). Naturally $N_{\text{Cheb}} \geq N_{\text{CLT}}$.

Finally, we could use the non-uniform Berry-Esseen inequality from Theorem 3. This inequality requires a finite scaled third moment $M_3 = E\left|Y_i - \mu\right|^3 / \sigma^3$. If $\hat{\mu}_n$ denotes a sample mean of n IID random instances of Y, then the non-uniform Berry-Esseen inequality implies that

$$\begin{aligned}
\Pr\left[|\mu - \hat{\mu}_n| \leq \varepsilon\right] &= \Pr\left[\frac{\hat{\mu}_n - \mu}{\sigma/\sqrt{n}} \leq \frac{\sqrt{n}\varepsilon}{\sigma}\right] - \Pr\left[\frac{\hat{\mu}_n - \mu}{\sigma/\sqrt{n}} < -\frac{\sqrt{n}\varepsilon}{\sigma}\right] \\
&\geq \left[\Phi(\sqrt{n}\varepsilon/\sigma) - \Delta_n(\sqrt{n}\varepsilon/\sigma, M_3)\right] \\
&\qquad - \left[\Phi(-\sqrt{n}\varepsilon/\sigma) + \Delta_n(-\sqrt{n}\varepsilon/\sigma, M_3)\right] \\
&= 1 - 2[\Phi(-\sqrt{n}\varepsilon/\sigma) + \Delta_n(\sqrt{n}\varepsilon/\sigma, M_3)],
\end{aligned} \tag{15}$$

since $\Delta_n(-x, M_3) = \Delta_n(x, M_3)$. The probability of making an error no greater than ε is bounded below by $1 - \alpha$, i.e., the fixed width confidence interval (1) holds with $\hat{\mu} = \hat{\mu}_n$, provided $n \geq N_{\text{BE}}(\varepsilon, \sigma, \alpha, M_3)$, where the Berry-Esseen sample size is

$$N_{\text{BE}}(\varepsilon, \sigma, \alpha, M_3) := \min\left\{ n \in \mathbb{N} : \Phi\left(-\sqrt{n}\varepsilon/\sigma\right) + \Delta_n(\sqrt{n}\varepsilon/\sigma, M_3) \leq \frac{\alpha}{2} \right\}. \tag{16}$$

To compute $N_{\text{BE}}(\varepsilon, \sigma, \alpha, M_3)$, we need to know M_3. In practice, substituting an upper bound on M_3 yields an upper bound on the necessary sample size.

Note that if the Δ_n term in (16) were absent, N_{BE} would correspond to the CLT sample size N_{CLT}, and in general $N_{\text{BE}} > N_{\text{CLT}}$. It is possible that in some situations $N_{\text{BE}} > N_{\text{Cheb}}$ might hold, and in such cases we could use N_{Cheb} instead of N_{BE}.

3.3 Algorithm and Proof of Its Success

In detail, the two-stage algorithm works as described below.

Algorithm 1 (Two Stage). *The user specifies four quantities:*

- *An initial sample size for variance estimation,* $n_\sigma \in \{2, 3, \ldots\}$,
- *A variance inflation factor* $\mathfrak{C}^2 \in (1, \infty)$,
- *An uncertainty* $\alpha \in (0, 1)$, *and,*
- *An error tolerance or confidence interval half-width,* $\varepsilon > 0$.

At the first stage of the algorithm, $Y_1, \ldots, Y_{n_\sigma}$ *are sampled independently from the same distribution as* Y. *Then the conservative variance estimate,* $\hat{\sigma}^2 = \mathfrak{C}^2 s^2_{n_\sigma}$, *is computed in terms of the sample variance,* $s^2_{n_\sigma}$, *defined by* (10).

To prepare for the second stage of the algorithm we compute $\tilde{\alpha} = 1 - \sqrt{1-\alpha}$ *and then* $\tilde{\kappa}_{\max} = \tilde{\kappa}_{\max}(\tilde{\alpha}, n_\sigma, \mathfrak{C})$ *using Eq.* (13). *The sample size for the second stage is*

$$n_\mu = N_\mu(\varepsilon, \hat{\sigma}, \tilde{\alpha}, \tilde{\kappa}^{3/4}_{\max}), \tag{17}$$

where

$$N_\mu(\varepsilon, \sigma, \alpha, M) := \max\bigl(1, \min\bigl(N_{Cheb}(\varepsilon, \sigma, \alpha), N_{BE}(\varepsilon, \sigma, \alpha, M)\bigr)\bigr). \tag{18}$$

Recall that N_{Cheb} *is defined in* (14) *and* N_{BE} *is defined in* (16).

After this preparation, the second stage is to sample $Y_{n_\sigma+1}, \ldots, Y_{n_\sigma+n_\mu}$ *independently from the distribution of* Y, *and independently of* $Y_1, \ldots, Y_{n_\sigma}$. *The algorithm then returns the sample mean,*

$$\hat{\mu} = \frac{1}{n_\mu} \sum_{i=n_\sigma+1}^{n_\sigma+n_\mu} Y_i. \tag{19}$$

The success of this algorithm is guaranteed in the following theorem. The main assumption needed is an upper bound on the kurtosis.

Theorem 5. *Let* Y *be a random variable with mean* μ, *and either zero variance or positive variance with modified kurtosis* $\tilde{\kappa} \le \tilde{\kappa}_{\max}(\tilde{\alpha}, n_\sigma, \mathfrak{C})$. *It follows that Algorithm 1 above yields an estimate* $\hat{\mu}$ *given by* (19) *which satisfies the fixed width confidence interval condition*

$$\Pr(|\hat{\mu} - \mu| \le \varepsilon) \ge 1 - \alpha.$$

Proof. If $\sigma^2 = 0$, then $s^2_{n_\sigma} = 0$, $n_\mu = 1$ and $\hat{\mu} = \mu$ with probability one. Now consider the case of positive variance. The first stage yields a variance estimate satisfying $\Pr(\hat{\sigma}^2 > \sigma^2) \ge 1 - \tilde{\alpha}$ by the argument preceding the kurtosis bound in (13) applied with uncertainty $\tilde{\alpha}$. The second stage yields $\Pr(|\hat{\mu} - \mu| \le \varepsilon) \ge 1 - \tilde{\alpha}$ by the Berry-Esseen result (15), so long as $\hat{\sigma} \ge \sigma$ and $M_3 \le \tilde{\kappa}_{\max}(\tilde{\alpha}, n_\sigma, \mathfrak{C})^{3/4}$.

The second condition holds because $M_3 \le \tilde{\kappa}^{3/4}$ by Jensen's Inequality [10, 8.4.b]. Thus, in the two-stage algorithm we have

$$\begin{aligned}\Pr(|\hat{\mu}-\mu| \le \varepsilon) &= \mathbb{E}\big[\Pr(|\hat{\mu}-\mu| \le \varepsilon \mid \hat{\sigma})\big] \\ &\ge \mathbb{E}\left[(1-\tilde{\alpha})1_{\sigma \le \hat{\sigma}}\right] \\ &\ge (1-\tilde{\alpha})(1-\tilde{\alpha}) = 1-\alpha. \qquad \square\end{aligned}$$

Remark 1. As pointed out earlier, the guarantees in this theorem require that the modified kurtosis of Y not exceed the specified upper bound $\tilde{\kappa}_{\max}$. As it is presented, Algorithm 1 takes as inputs, n_σ, $\mathfrak{C}$, and α, and uses these to compute $\tilde{\kappa}_{\max}$ according to (13). The reason for doing so is that one might have a better intuition for n_σ, $\mathfrak{C}$, and α. Alternatively, one may specify n_σ and $\tilde{\kappa}_{\max}$ and use (13) to compute $\mathfrak{C}$, or specify $\mathfrak{C}$ and $\tilde{\kappa}_{\max}$ and use (13) to compute n_σ. The issue of how one should choose n_σ, $\mathfrak{C}$, and $\tilde{\kappa}_{\max}$ in practice is discussed further in Sect. 5.

Remark 2. In this algorithm it is possible to choose n_μ much smaller than n_σ if the sample variance is small. As a practical matter we suggest that if one is willing to invest n_σ samples to estimate the variance then one should be willing to invest at least that many additional samples to estimate the mean. Therefore, in the numerical examples of Sect. 4 we use

$$N_\mu(\varepsilon,\sigma,\alpha,M) := \max\big(n_\sigma, \min\big(N_{\text{Cheb}}(\varepsilon,\sigma,\alpha), N_{\text{BE}}(\varepsilon,\sigma,\alpha,M)\big)\big) \tag{20}$$

instead of (18) to determine the sample size for the sample mean. Because the variance is typically harder to estimate accurately than the mean, one may wonder whether n_σ should be chosen greater than n_μ. However, for Monte Carlo simulation we only need the variance to one or two digits accuracy, whereas we typically want to know the mean to a much higher accuracy. By the error bound following from Chebychev's inequality (Theorem 1), the definition of N_μ in (20) means that the fixed width confidence interval constructed by Algorithm 1 also holds for any random variables, Y, with small variance, namely, $\sigma^2 \le \varepsilon^2 \alpha n_\sigma$, even if its kurtosis is arbitrarily large.

As mentioned in the introduction, one frequently encountered case occurs when Y is a d-variate function of a random vector $\boldsymbol{X}$. Then μ corresponds to the multivariate integral in (6) and Theorem 5 may be interpreted as below:

Corollary 1. *Suppose that $\rho : \mathbb{R}^d \to \mathbb{R}$ is a probability density function, the integrand $f : \mathbb{R}^d \to \mathbb{R}$ has finite $\mathcal{L}_4$ norm as defined in (7), and furthermore f lies in the cone $\mathcal{C}_{\tilde{\kappa}_{\max}}$ defined in (8), where $\tilde{\kappa}_{\max} = \tilde{\kappa}_{\max}(\tilde{\alpha}, n_\sigma, \mathfrak{C})$. It follows that Algorithm 1 yields an estimate, $\hat{\mu}$, of the multidimensional integral μ defined in (6), which satisfies the fixed width confidence interval condition*

$$\Pr(|\hat{\mu}-\mu| \le \varepsilon) \ge 1-\alpha.$$

3.4 Cost of the Algorithm

The number of function values required by the two-stage Algorithm 1 is $n_\sigma + n_\mu$, the sum of the initial sample size used to estimate the variance of Y and the sample size used to estimate the mean of Y. Although n_σ is deterministic, n_μ is a random variable, and so the cost of this algorithm might be best defined probabilistically. Moreover, the only random quantity in the formula for n_μ in (17) is $\hat{\sigma}^2$, the upper bound on variance. Clearly this depends on the unknown population variance, σ^2, and we expect $\hat{\sigma}^2$ not to overestimate σ^2 by much. Thus, the algorithm cost is defined below in terms of σ^2 and the error tolerance (interval half-width) ε. An upper bound on the cost is then derived in Theorem 6.

Let A be any random algorithm that takes as its input, a method for generating random samples, $Y_1, Y_2, \ldots$ with common distribution function F having variance σ^2 and modified kurtosis $\tilde{\kappa}$. Additional algorithm inputs are an error tolerance, ε, an uncertainty, α, and a maximum modified kurtosis, $\tilde{\kappa}_{\max}$. The algorithm then computes $\hat{\mu} = A(F, \varepsilon, \alpha, \tilde{\kappa}_{\max})$, an approximation to $\mu = \mathbb{E}(Y)$, based on a total of $N_{\text{tot}}(\varepsilon, \alpha, \tilde{\kappa}_{\max}, F)$ samples. The probabilistic cost of the algorithm, with uncertainty β, for integrands of variance no greater than $\sigma^2_{\max}$ and modified kurtosis no greater than $\tilde{\kappa}_{\max}$ is defined as

$$N_{\text{tot}}(\varepsilon, \alpha, \beta, \tilde{\kappa}_{\max}, \sigma_{\max}) := \sup_{\substack{\tilde{\kappa} \le \tilde{\kappa}_{\max} \\ \sigma \le \sigma_{\max}}} \min \{N : \Pr[N_{\text{tot}}(\varepsilon, \alpha, \tilde{\kappa}_{\max}, F) \le N] \ge 1 - \beta\}.$$

Note that $\tilde{\kappa}_{\max}$ is an input to the algorithm, but $\sigma_{\max}$ is not. The cost of an arbitrary algorithm, A may also depend on other parameters, such as n_σ and $\mathfrak{C}$ in our Algorithm 1, which are related to $\tilde{\kappa}_{\max}$. However, this dependence is not shown explicitly to keep the notation simple.

The cost of the particular two-stage Monte Carlo algorithm defined in Algorithm 1 is

$$\sup_{\substack{\tilde{\kappa} \le \tilde{\kappa}_{\max} \\ \sigma \le \sigma_{\max}}} \min \left\{N : \Pr(n_\sigma + N_\mu(\varepsilon, \hat{\sigma}, \tilde{\alpha}, \tilde{\kappa}^{3/4}_{\max}) \le N) \ge 1 - \beta\right\}.$$

Since n_σ is fixed, bounding this cost depends on bounding $N_\mu(\varepsilon, \hat{\sigma}, \tilde{\alpha}, \tilde{\kappa}^{3/4}_{\max})$, which depends on $\hat{\sigma}$ as given by Algorithm 1. Moreover, $\hat{\sigma}$ can be bounded above using (12a) in Lemma 1. For $\tilde{\kappa} \le \tilde{\kappa}_{\max}$,

$$\begin{aligned} 1 - \beta &\le \Pr\left[s^2_{n_\sigma} < \sigma^2 \left\{1 + \sqrt{\left(\tilde{\kappa} - \frac{n_\sigma - 3}{n_\sigma - 1}\right)\left(\frac{1-\beta}{\beta n_\sigma}\right)}\right\}\right] \\ &\le \Pr\left[\hat{\sigma}^2 = \mathfrak{C}^2 s^2_{n_\sigma} < \mathfrak{C}^2\sigma^2 \left\{1 + \sqrt{\left(\tilde{\kappa}_{\max}(n_\sigma, \tilde{\alpha}, \mathfrak{C}) - \frac{n_\sigma - 3}{n_\sigma - 1}\right)\left(\frac{1-\beta}{\beta n_\sigma}\right)}\right\}\right] \end{aligned}$$

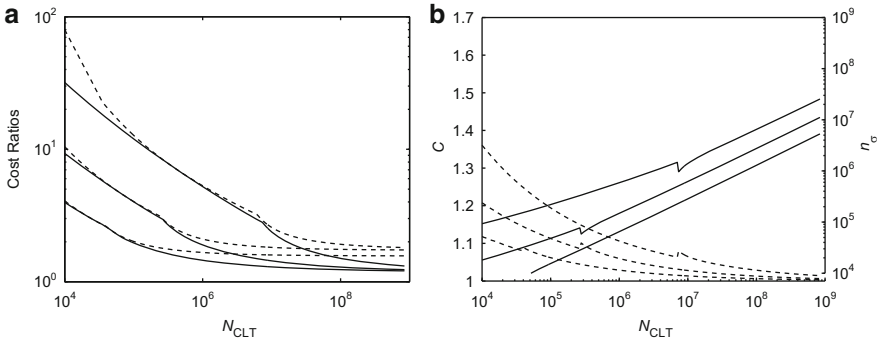


Fig. 1 (**a**) The cost ratios of $N_{\text{up}}(\varepsilon, 0.01, 0.01, \tilde{\kappa}_{\max}, \sigma)/N_{\text{CLT}}(\varepsilon, \sigma, 0.01)$ for $\tilde{\kappa}_{\max} = 2, 10$, and 100, with $n_\sigma = 4{,}000\,\tilde{\kappa}_{\max}$ (*dashed*) and n_σ optimized (*solid*); (**b**) the optimal values of n_σ (*solid*) and $\mathfrak{C}$ (*dashed*).

$$= \Pr\left[\hat{\sigma}^2 < \sigma^2 v^2(\tilde{\alpha}, \beta, \mathfrak{C})\right],$$

where

$$v^2(\tilde{\alpha}, \beta, \mathfrak{C}) := \mathfrak{C}^2 + \left(\mathfrak{C}^2 - 1\right)\sqrt{\frac{\tilde{\alpha}(1-\beta)}{(1-\tilde{\alpha})\beta}} > 1.$$

Noting that $N_\mu(\varepsilon, \cdot, \tilde{\alpha}, \tilde{\kappa}_{\max}^{3/4})$ is a non-decreasing function allows one to derive the following upper bound on the cost of the adaptive Monte Carlo algorithm.

Theorem 6. *The two-stage Monte Carlo algorithm for fixed width confidence intervals based on IID sampling described in Algorithm 1 has a probabilistic cost bounded above by*

$$N_{\text{tot}}(\varepsilon, \alpha, \beta, \tilde{\kappa}_{\max}, \sigma_{\max})$$
$$\le N_{\text{up}}(\varepsilon, \alpha, \beta, \tilde{\kappa}_{\max}, \sigma_{\max}) := n_\sigma + N_\mu(\varepsilon, \sigma_{\max} v(\tilde{\alpha}, \beta, \mathfrak{C}), \tilde{\alpha}, \tilde{\kappa}_{\max}^{3/4}).$$

Note that the Chebychev sample size, N_{Cheb}, defined in (14), the Berry-Esseen sample size, N_{BE}, defined in (16), and thus N_μ all depend on σ and ε through their ratio, σ/ε. Thus, ignoring the initial sample used to estimate the variance, $N_{\text{tot}}(\varepsilon, \alpha, \beta, \tilde{\kappa}_{\max}, \sigma_{\max})$ is roughly proportional to $\sigma_{\max}^2/\varepsilon^2$, even though $\sigma_{\max}$ is not a parameter of the algorithm. Algorithm 1 *adaptively* determines the sample size, and thus the cost, to fit the unknown variance of Y. Random variables, Y, with small variances will require a lower cost to estimate μ with a given error tolerance than random variables with large variances.

Figure 1a shows the ratio of the upper bound of the cost, $N_{\text{up}}(\varepsilon, 0.01, 0.01, \tilde{\kappa}_{\max}, \sigma)$, to the ideal CLT cost, $N_{\text{CLT}}(\varepsilon, \sigma, 0.01) = \lceil (2.58\sigma/\varepsilon)^2 \rceil$, for a range of σ/ε ratios and for $\tilde{\kappa}_{\max} = 2, 10$, and 100. In these graphs the formula defining N_{up}

in Theorem 6 uses the alternative and somewhat costlier formula for N_μ in (20). The dashed curves in Fig. 1a show these cost ratios with $n_\sigma = 4{,}000\,\tilde{\kappa}_{\max}$, which corresponds to $\mathfrak{C} \approx 1.1$. The solid curves denote the case where n_σ and $\mathfrak{C}$ vary with σ/ε to minimize N_{up}. Figure 1b displays the optimal values of n_σ (solid) and $\mathfrak{C}$ (dashed). In both figures, higher curves correspond to higher values of $\tilde{\kappa}_{\max}$.

Here, N_{CLT} denotes the ideal cost if one knew the variance of Y a priori and knew that the distribution of the sample mean was close to Gaussian. The cost ratio is the penalty for having a guaranteed fixed width confidence interval in the absence of this knowledge about the distribution of Y. For smaller values of N_{CLT}, equivalently smaller σ/ε, this cost ratio can be rather large. However the absolute effect of this large penalty is mitigated by the fact that the total number of samples needed is not much. For larger N_{CLT}, equivalently larger σ/ε, the cost ratio approaches somewhat less than 1.4 in the case of optimal n_σ and $\mathfrak{C}$, and somewhat less than 2 for $n_\sigma = 1{,}000\,\tilde{\kappa}_{\max}$.

The discontinuous derivatives in the curves in Fig. 1 arise from the minimum and maximum values arising in formulas (16) and (20) for N_{BE} and N_μ, respectively. Taking the upper dashed curve in Fig. 1a as an example, for N_{CLT} less than about 3.5×10^4, $N_\mu = n_\sigma$. For N_{CLT} from about 3.5×10^4 to about 6×10^6, N_μ corresponds to the second term in the minimum in the Berry-Esseen inequality, (16), i.e., the non-uniform term. For N_{CLT} greater than 6×10^6, N_μ corresponds to the first term in the minimum in the Berry-Esseen inequality, (16), i.e., the uniform term.

The ideal case of optimizing n_σ and $\mathfrak{C}$ with respect to σ/ε is impractical, since σ is not known in advance. Our suggestion is to choose $\mathfrak{C}$ around 1.1, and then choose n_σ as large as needed to ensure that $\tilde{\kappa}_{\max}$ is as large as desired. For example with $\mathfrak{C} = 1.1$ and $\tilde{\kappa}_{\max} = 2, 10$, and 100 we get $n_\sigma = 6{,}593$, 59,311, and 652,417 respectively.

4 Numerical Examples

4.1 Univariate Fooling Functions for Deterministic Algorithms

Several commonly used software packages have automatic algorithms for integrating functions of a single variable. These include

- `quad` in MATLAB [28], adaptive Simpson's rule based on `adaptsim` by Gander and Gautschi [4],
- `quadgk` in MATLAB [28], adaptive Gauss-Kronrod quadrature based on `quadva` by Shampine [22], and
- The `chebfun` [5] toolbox for MATLAB [28], which approximates integrals by integrating interpolatory Chebychev polynomial approximations to the integrands.

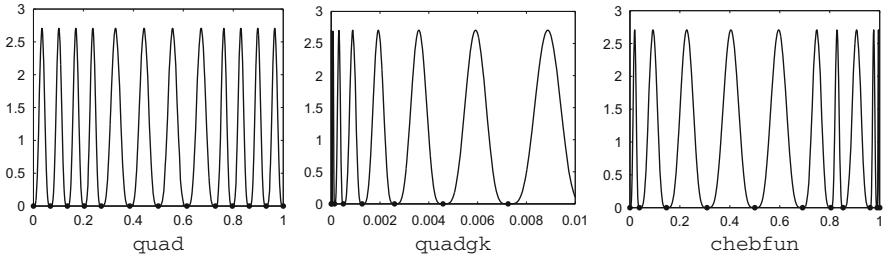


Fig. 2 Plots of fooling functions, f, with $\mu = \int_0^1 f(x)\,\mathrm{d}x = 1$, but for which the corresponding algorithms return values of $\hat{\mu} = 0$.

For these three automatic algorithms one can easily probe where they sample the integrand, feed the algorithms zero values, and then construct fooling functions for which the automatic algorithms will return a zero value for the integral. Figure 2 displays these fooling functions for the problem $\mu = \int_0^1 f(x)\,\mathrm{d}x$ for these three algorithms. Each of these algorithms is asked to provide an answer with an absolute error no greater than 10^{-14}, but in fact the absolute error is 1 for these fooling functions. The algorithms `quad` and `chebfun` sample only about a dozen points before concluding that the function is zero, whereas the algorithm `quadgk` samples a much larger number of points (only those between 0 and 0.01 are shown in the plot).

4.2 Integrating a Single Hump

Accuracy and timing results have been recorded for the integration problem $\mu = \int_{[0,1]^d} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ for a single hump test integrand

$$f(\boldsymbol{x}) = a_0 + b_0 \prod_{j=1}^{d} \left[1 + b_j \exp\left(-\frac{(x_j - h_j)^2}{c_j^2} \right) \right]. \tag{21}$$

Here $\boldsymbol{x}$ is a d dimensional vector, and $a_0, b_0, \ldots, b_d, c_1, \ldots, c_d, h_1, \ldots, h_d$ are parameters. Figures 3 and 4 show the results of different algorithms being used to integrate 500 different instances of f. For each instance of f, the parameters are chosen as follows:

- $b_1, \ldots, b_d \in [0.1, 10]$ with $\log(b_j)$ being i.i.d. uniform,
- $c_1, \ldots, c_d \in [10^{-6}, 1]$ with $\log(c_j)$ being i.i.d. uniform,
- $h_1, \ldots, h_d \in [0, 1]$ with h_j being i.i.d. uniform,
- b_0 chosen in terms of the $b_1, \ldots, b_d, c_1, \ldots, c_d, h_1, \ldots, h_d$ to make $\sigma^2 = \|f - \mu\|_2^2 \in [10^{-2}, 10^2]$, with $\log(\sigma)$ being i.i.d. uniform for each instance, and
- a_0 chosen in terms of the $b_0, \ldots, b_d, c_1, \ldots, c_d, h_1, \ldots, h_d$ to make $\mu = 1$.

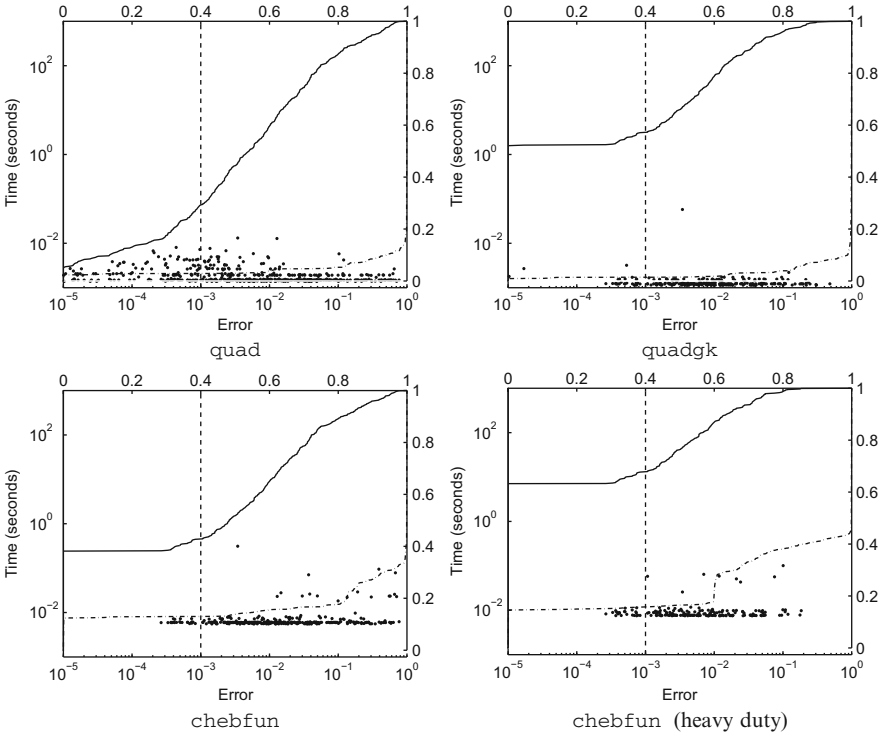


Fig. 3 Execution times and errors for test function (21) for $d = 1$ and error tolerance $\varepsilon = 10^{-3}$, and a variety of parameters giving a range of σ and $\tilde{\kappa}$. Those points to the left/right of the *dashed vertical line* represent successes/failures of the automatic algorithms. The *solid line* shows that cumulative distribution of actual errors, and the *dot-dashed line* shows the cumulative distribution of execution times.

These ranges of parameters are chosen so that the algorithms being tested fail to meet the error tolerance a significant number of times.

These 500 random constructions of f with $d = 1$ are integrated using `quad`, `quadgk`, `chebfun`, Algorithm 1, and an automatic quasi-Monte Carlo algorithm that uses scrambled Sobol' sampling [3, 7, 11, 17–19]. For the Sobol' sampling algorithm the error is estimated by an inflation factor of 1.1 times the sample standard deviation of 8 internal replicates of one scrambled Sobol' sequence [20]. The sample size is increased until this error estimate decreases to no more than the tolerance. We have not yet found simple conditions on integrands for which this procedure is guaranteed to produce an estimate satisfying the error tolerance, and so we do not discuss it in detail. We are however, intrigued by the fact that it does seem to perform rather well in practice.

For all but `chebfun`, the specified absolute error tolerance is $\epsilon = 0.001$. The algorithm `chebfun` attempts to do all calculations to near machine precision. The observed error and execution times are plotted in Figs. 3 and 4. Whereas

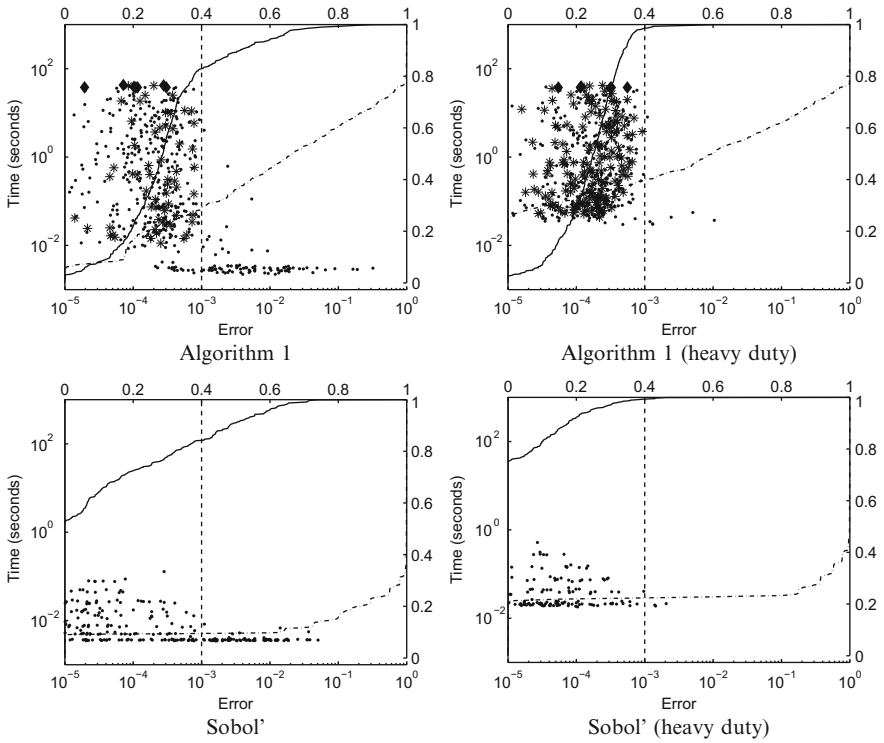


Fig. 4 Execution times and errors for test function (21) for $d = 1$ and error tolerance $\varepsilon = 10^{-3}$, and a variety of parameters giving a range of σ and $\tilde{\kappa}$. Those points to the left/right of the *dashed vertical line* represent successes/failures of the automatic algorithms. The *solid line* shows that cumulative distribution of actual errors, and the *dot-dashed line* shows the cumulative distribution of execution times. For Algorithm 1 the points labeled * are those for which the Corollary 1 guarantees the error tolerance.

`chebfun` uses a minimum of $2^3 + 1 = 9$ function values, the figure labeled "`chebfun` (heavy duty)" displays the results of requiring `chebfun` to use at least $2^8+1 = 257$ function values. Algorithm 1 takes $\alpha = 0.01$, and $\mathfrak{C} = 1.1$. For the plot on the left, $n_\sigma = 2^{13} = 8{,}192$, which corresponds to $\tilde{\kappa}_{\max} = 2.24$. For the heavy duty plot on the right, $n_\sigma = 2^{18} = 262{,}144$, which corresponds to $\tilde{\kappa}_{\max} = 40.1$. The same initial sample sizes are used for the Sobol' sampling algorithm.

Figure 3 shows that `quad` and `quadgk` are quite fast, nearly always providing an answer in less than 0.01 s. Unfortunately, they successfully meet the error tolerance only about 30 % of the time for `quad` and 50–60 % of the time for `quadgk`. The difficult cases are those where c_1 is quite small, and these algorithms miss the sharp peak. The performance of `chebfun` is similar to that of `quad` and `quadgk`. The heavy duty version of `chebfun` fares somewhat better. For both of the `chebfun` plots there are a significant proportion of the data that do not appear because their errors are smaller than 10^{-5}.

In the plots for Algorithm 1 in Fig. 4 the alternative and somewhat costlier formula for N_μ in (20) is employed. An asterisk is used to label those points satisfying $\tilde{\kappa} \le \tilde{\kappa}_{\max}$, where $\tilde{\kappa}$ is defined in (7). All such points fall within the prescribed error tolerance, which is even better than the guaranteed confidence of 99 %. For Algorithm 1 (heavy duty) $\tilde{\kappa}_{\max}$ is larger, so there are more points for which the guarantee holds. Those points labeled with a dot, are those for which $\tilde{\kappa} > \tilde{\kappa}_{\max}$, and so no guarantee holds. The points labeled with a diamond are those for which Algorithm 1 attempts to exceed the cost budget that we set, i.e., it wants to choose n_μ such that $n_\sigma + n_\mu > N_{\max} := 10^9$. In these cases n_μ is chosen as $\lfloor 10^9 - n_\sigma \rfloor$, which often is still large enough to get an answer that satisfies the error tolerance. Algorithm 1 performs somewhat more robustly than `quad`, `quadgk`, and `chebfun`, because it requires only a low degree of smoothness and takes a fairly large minimum sample. Algorithm 1 is generally much slower than the other algorithms because it does not assume any smoothness of the integrand. The more important point is that Algorithm 1 has a guarantee, whereas to our knowledge, the other routines do not.

From Fig. 4, the Sobol' sampling algorithm is more reliable and takes less time than Algorithm 1. This is due primarily to the fact that in dimension one, Sobol' sampling is equivalent to stratified sampling, where the points are more evenly spread than IID sampling.

Figure 5 repeats the simulation shown in Fig. 4 for the same test function (21), but now with $d = 2, \ldots, 8$ chosen randomly and uniformly. For this case the univariate integration algorithms are inapplicable, but the multidimensional routines can be used. There are more cases where the Algorithm 1 tries to exceed the maximum sample size allowed, i.e., $(n_\sigma + n_\mu)d > N_{\max} := 10^9$, but the behavior seen for $d = 1$ still generally applies.

4.3 Asian Geometric Mean Call Option Pricing

The next example involves pricing an Asian geometric mean call option. Suppose that the price of a stock S at time t follows a geometric Brownian motion with constant interest rate, r, and constant volatility, v. One may express the stock price in terms of the initial condition, $S(0)$, as

$$S(t) = S(0)\exp[(r - v^2/2)t + vB(t)], \qquad t \ge 0,$$

where B is a standard Brownian motion. The discounted payoff of the Asian geometric mean call option with an expiry of T years, a strike price of K, and assuming a discretization at d times is

$$Y = \max\left([\sqrt{S(0)}S(T/d)S(2T/d)\cdots S(T(d-1)/d)\sqrt{S(T)}]^{1/d} - K, 0\right)\mathrm{e}^{-rT}. \tag{22}$$

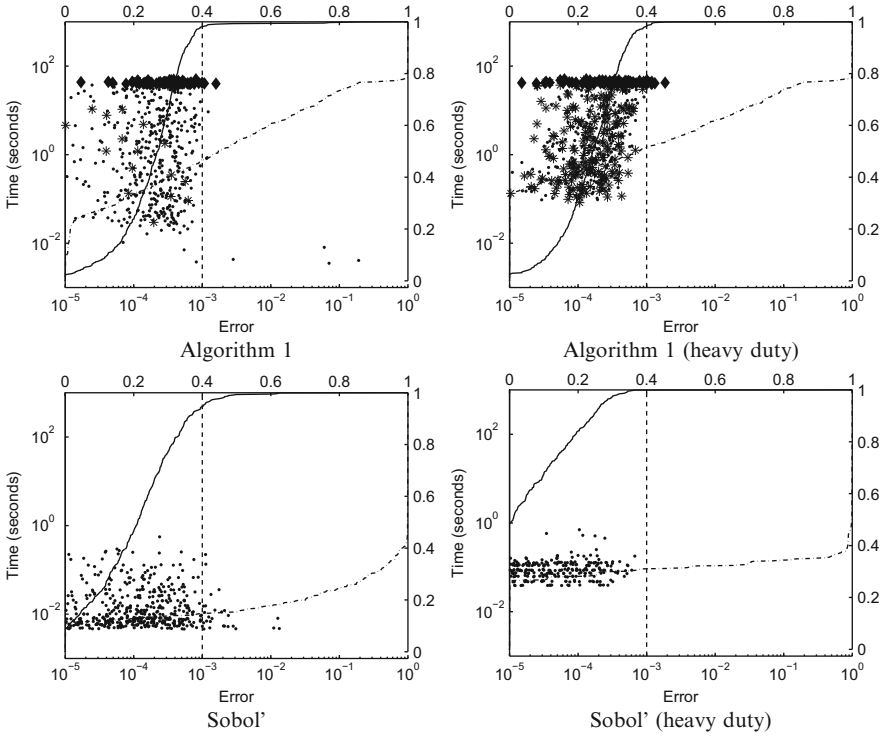


Fig. 5 Execution times and errors for test function (21) for $d = 2, \ldots, 8$ and $\varepsilon = 10^{-3}$, with the rest of the parameters as in Fig. 4.

The fair price of this option is $\mu = \mathbb{E}(Y)$. One of our chief reasons for choosing this option for numerical experiments is that its price can be computed analytically, while the numerical computation is non-trivial.

In our numerical experiments, the values of the Brownian motion at different times required for evaluating the stock price, $B(T/d), B(2T/d), \ldots, B(T)$, are computed via a Brownian bridge construction. This means that for one instance of the Brownian motion we first compute $B(T)$, then $B(T/2)$, etc., using independent Gaussian random variables $X_1, \ldots, X_d$, suitably scaled. The Brownian bridge accounts for more of the low frequency motion of the stock price by the X_j with smaller j, which allows the Sobol' sampling algorithm to do a better job.

The option price, $\mu = \mathbb{E}(Y)$, is approximated by Algorithm 1 and the Sobol' sampling algorithm using an error tolerance of $\varepsilon = 0.05$, and compared to the analytic value of μ. The result of 500 replications is given in Fig. 6. Some of the parameters are set to be fixed values, namely,

$$S(0) = K = 100, \qquad T = 1, \qquad r = 0.03.$$

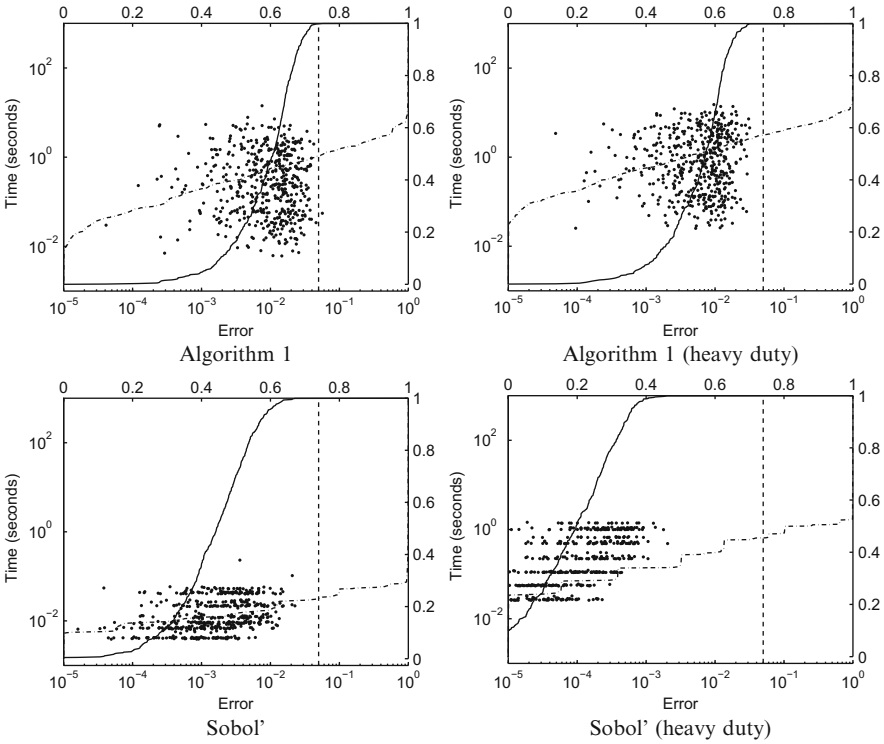


Fig. 6 Execution times and errors for the Asian geometric mean call option for $d = 1, 2, 4, 8, 16, 32$ and $\varepsilon = 0.05$.

The volatility, v, is drawn uniformly between 0.1 and 0.7. The number of time steps, d, is chosen to be uniform over $\{1, 2, 4, 8, 16, 32\}$. The true value of μ for these parameters is between about 2.8 and 14.

For this example the true kurtosis of Y is unknown. Both Algorithm 1 and the Sobol' sampling algorithm compute the option price to the desired error tolerance with high reliability. For the IID sampling Algorithm 1 and the ordinary Sobol' sampling algorithm it can be seen that some of the errors are barely under the error tolerance, meaning that the sample size is not chosen too conservatively. For the heavy duty Sobol' algorithm, the high initial sample size seems to lead to smaller than expected errors and larger than necessary computation times.

5 Discussion

Practitioners often construct CLT-based confidence intervals with the true variance estimated by the sample variance, perhaps multiplied by some inflation factor. Often, this approach works, but it has no guarantee of success. The two-stage

algorithm presented here is similar to the approach just described, but it carries guarantees. These are derived by employing Cantelli's inequality to ensure a reliable variance upper bound, and by employing a Berry-Esseen inequality to ensure a large enough sample for the sample mean.

In certain cases our procedure multiplies the computational cost by a large factor such as 2 or 10 or even 100 compared to what one might spend based on the CLT with a known value of σ (see Fig. 1). While this seems inefficient, one should remember that the total elapsed time may still be well below several seconds. Furthermore, one typically does not know σ in advance, and our adaptive algorithm estimates σ and then an appropriate sample size n_μ from the data. Our algorithmic cost will be low when the unknown σ is small and large when σ is large.

Like any algorithm with guarantees, our algorithm does need to make assumptions about the random variable Y. We assume a known bound on the kurtosis of Y, either specified directly or implied by the user's choice of the sample size for estimating the variance, n_σ, and the variance inflation factor, $\mathfrak{C}^2$. This is a philosophical choice. We prefer not to construct an algorithm that assumes a bound on the variance of Y, because such an algorithm would not be guaranteed for cY with $|c|$ large enough. If our algorithm works for Y, it will also work for cY, no matter how large $|c|$ is.

In practice the user may not know a priori if $\tilde{\kappa} \le \tilde{\kappa}_{\max}$ since it is even more difficult to estimate $\tilde{\kappa}$ from a sample than it is to estimate σ^2. Thus, the choice of $\tilde{\kappa}_{\max}$ relies on the user's best judgement. Here are a few thoughts that might help. One might try a sample of typical problems for which one knows the answers and use these problems to suggest an appropriate $\tilde{\kappa}_{\max}$. Alternatively, one may think of $\tilde{\kappa}_{\max}$ not as a parameter to be prescribed, but as a reflection of the robustness of one's Monte Carlo algorithm having chosen α, n_σ and $\mathfrak{C}$. The discussion at the end of Sect. 3.4 provides guidance on how to choose n_σ and $\mathfrak{C}$ to achieve a given $\tilde{\kappa}_{\max}$ in a manner that minimizes total computational cost. Briefly, one should not skimp on n_σ, but choose n_σ to be several thousand times $\tilde{\kappa}_{\max}$ and employ a $\mathfrak{C}$ that is relatively close to unity. Another way to look at the Theorem 5 is that, like a pathologist, it tells you what went wrong if the two-stage adaptive algorithm fails: the kurtosis of the random variable must have been too large. In any case, as one can see in Fig. 1, in the limit of vanishing ε/σ, i.e., $N_{\text{CLT}} \to \infty$, the choice of $\tilde{\kappa}_{\max}$ makes a negligible contribution to the total cost of the algorithm. The main determinant of computational cost is ε/σ.

Bahadur and Savage [1] prove in Corollary 2 that it is *impossible* to construct exact confidence intervals for the mean of random variable whose distribution lies in a set satisfying a few assumptions. One of these assumptions is that the set of distributions is convex. This assumption is violated by our assumption of bounded kurtosis in Theorem 5. Thus, we are able to construct guaranteed confidence intervals.

Our algorithm is adaptive because n_μ is determined from the sample variance. Information-based complexity theory tells us that adaptive information does not help for the integration problem for symmetric, convex sets of integrands, f, in the worst case and probabilistic settings [29, Chap. 4, Theorem 5.2.1; Chap. 8, Corollary

5.3.1]. Here, in Corollary 1 the cone, $\mathcal{C}_{\tilde{\kappa}_{\max}}$, although symmetric, is not a convex set, so it is possible for adaption to help.

There are a couple of areas that suggest themselves for further investigation. One is relative error, i.e., a fixed width confidence interval of the form

$$\Pr[|\mu - \hat{\mu}| \le \varepsilon\,|\mu|] \ge 1 - \alpha.$$

Here the challenge is that the right hand side of the first inequality includes the unknown mean.

Another area for further work is to provide guarantees for automatic quasi-Monte Carlo algorithms. Here the challenge is finding reliable formulas for error estimation. Typical error bounds involve a semi-norm of the integrand that is harder to compute than the original integral. For randomized quasi-Monte Carlo an estimate of the variance of the sample mean using n samples does not tell you much about the variance of the sample mean using a different number of samples.

Acknowledgements The first and second authors were partially supported by the National Science Foundation under DMS-0923111 and DMS-1115392. The fourth author was partially supported by the National Science Foundation under DMS-0906056.

The authors gratefully acknowledge discussions with Erich Novak and Henryk Woźniakowski, and the comments of the referees. The plots of the univariate fooling functions were prepared with the help of Nicholas Clancy and Caleb Hamilton. The first and fourth authors would like to express their thanks to the local organizers of the Tenth International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing for hosting a wonderful conference.

References

1. Bahadur, R.R., Savage, L.J.: The nonexistence of certain statistical procedures in nonparametric problems. Ann. Math. Stat. **27**, 1115–1122 (1956)
2. Chow, Y.S., Robbins, H.: On the asymptotic theory of fixed-width sequential confidence intervals for the mean. Ann. Math. Stat. **36**, 457–462 (1965)
3. Dick, J., Pillichshammer, F.: Digital nets and sequences: discrepancy theory and quasi-Monte Carlo integration. Cambridge University Press, Cambridge (2010)
4. Gander, W., Gautschi, W.: Adaptive quadrature—revisited. BIT **40**, 84–101 (2000)
5. Hale, N., Trefethen, L.N., Driscoll, T.A.: Chebfun version 4 (2012)
6. Hall, P.: Theoretical comparisons of bootstrap confidence intervals. Ann. Statist. **16**, 927–953 (1988)
7. Hong, H.S., Hickernell, F.J.: Algorithm 823: implementing scrambled digital nets. ACM Trans. Math. Software **29**, 95–109 (2003)
8. Lehmann, E.L., Romano, J.P.: Testing Statistical Hypotheses, 3rd edn. Springer, New York (2005)
9. Lemieux, C.: Monte Carlo and Quasi-Monte Carlo Sampling. Springer Science+Business Media, Inc., New York (2009)
10. Lin, Z., Bai, Z.: Probability Inequalities. Science Press/Springer, Beijing/Berlin (2010)
11. Matoušek, J.: On the L_2-discrepancy for anchored boxes. J. Complexity **14**, 527–556 (1998)
12. McCullagh, P.: Tensor Methods in Statistics. Chapman and Hall, London (1987)
13. Miller, R.: Beyond ANOVA, Basics of Applied Statistics. Wiley, New York (1986)

14. Mukhopadhyay, N., Datta, S.: On sequential fixed-width confidence intervals for the mean and second-order expansions of the associated coverage probabilities. Ann. Inst. Stat. Math. **48**, 497–507 (1996)
15. Nefedova, Yu.S., Shevtsova, I.G.: On non-uniform convergence rate estimates in the central limit theorem. Theory Probab. Appl. **57**, 62–97 (2012)
16. Niederreiter, H.: Random Number Generation and Quasi-Monte Carlo Methods. CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia (1992)
17. Owen, A.B.: Randomly permuted (t, m, s)-nets and (t, s)-sequences. In: Niederreiter, H., Shiue, P.J.-S. (eds.) Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing. Lecture Notes in Statistics, vol. 106, pp. 299–317. Springer, New York (1995)
18. Owen, A.B.: Monte Carlo variance of scrambled net quadrature. SIAM J. Numer. Anal. **34**, 1884–1910 (1997)
19. Owen, A.B.: Scrambled net variance for integrals of smooth functions. Ann. Statist. **25**, 1541–1562 (1997)
20. Owen, A.B.: On the Warnock-Halton quasi-standard error. Monte Carlo Methods Appl. **12**, 47–54 (2006)
21. Serfling, R.J., Wackerly, D.D.: Asymptotic theory of sequential fixed-width confidence procedures. J. Amer. Statist. Assoc. **71**, 949–955 (1976)
22. Shampine, L.F.: Vectorized adaptive quadrature in Matlab. J. Comput. Appl. Math. **211**, 131–140 (2008)
23. Shevtsova, I.: On the absolute constants in the Berry–Esseen type inequalities for identically distributed summands. arXiv:1111.6554v1 [math.PR] (2011)
24. Siegmund, D.: Sequential analysis: tests and confidence intervals. Springer, New York (1985)
25. Sloan, I.H., Joe, S.: Lattice methods for multiple integration. Oxford University Press, Oxford (1994)
26. Stein, C.: A two sample test for a linear hypothesis whose power is independent of the variance. Ann. Math. Stat. **16**, 243–258 (1945)
27. Stein, C.: Some problems in sequential estimation. Econometrica **17**, 77–78 (1949)
28. The MathWorks, Inc.: MATLAB 7.12, Natick (2012)
29. Traub, J.F., Wasilkowski, G.W., Woźniakowski, H.: Information-based complexity. Academic, Boston (1988)

Discrepancy, Integration and Tractability

Aicke Hinrichs

Abstract The discrepancy function of a point distribution measures the deviation from the uniform distribution. Different versions of the discrepancy function capture this deviation with respect to different geometric objects. Via Koksma-Hlawka inequalities the norm of the discrepancy function in a function space is intimately connected to the worst case integration error of the quasi-Monte Carlo integration rule determined by the point set for functions from the unit ball of a related function space. So the a priori very geometric concept of the discrepancy function is a crucial tool for studying numerical integration.

In this survey article we want to discuss aspects of this interplay between discrepancy, integration and tractability questions. The main focus is on the exposition of some more recent results as well as on identifying open problems whose solution might advance our knowledge about this interplay of discrepancy, integration and randomization.

Via the Koksma-Hlawka connection, the construction of point sets with small discrepancy automatically yields good quasi-Monte Carlo rules. Here we discuss how the explicit point sets constructed by Chen and Skriganov as low discrepancy sets in L_p for $1 < p < \infty$ provide also good quasi-Monte Carlo rules in Besov spaces of dominating mixed smoothness.

Lower bounds for norms of the discrepancy function show the limits of this approach using function values and deterministic algorithms for the computation of integrals. Randomized methods may perform better, especially if the dimension of the problem is high. In this context we treat recent results on the power of importance sampling.

The study of average discrepancies is of interest to gain insight into the behavior of typical point sets with respect to discrepancy and integration errors. Very general

A. Hinrichs (✉)
Institut für Mathematik, Universität Rostock, Ulmenstraße 69, Haus 3, D-18051 Rostock, Germany
e-mail: aicke.hinrichs@uni-rostock.de

J. Dick et al. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2012*,
Springer Proceedings in Mathematics & Statistics 65,
DOI 10.1007/978-3-642-41095-6_6, © Springer-Verlag Berlin Heidelberg 2013

notions of the discrepancy function are related to empirical processes, average discrepancies then are expectations of certain norms of such empirical processes. We explain this connection and discuss some recent results on the limit behavior of average discrepancies as the number of points goes to infinity.

1 Introduction

For a point set $\mathscr{P}$ of $n \geq 1$ points in the d-dimensional unit cube $[0,1)^d$ the *discrepancy function* $D_{\mathscr{P}}$ is defined as

$$D_{\mathscr{P}}(x) := \frac{1}{n} \sum_{z \in \mathscr{P}} \mathbf{1}_{B_x}(z) - \operatorname{vol}(B_x). \tag{1}$$

Here $\operatorname{vol}(B_x) = x_1 \cdots x_d$ denotes the volume of the rectangular box $B_x = [0, x_1) \times \cdots \times [0, x_d)$ for $x = (x_1, \cdots, x_d) \in [0,1)^d$ and $\mathbf{1}_{B_x}$ is the characteristic function of the box B_x. Then the sum in the discrepancy function counts the number of points of $\mathscr{P}$ contained in B_x and the discrepancy function measures the deviation of this number from the fair number of points $n \operatorname{vol}(B_x)$ which would be achieved by a perfect (but impossible) uniform distribution of the points of $\mathscr{P}$.

There are two major goals in the study of the distribution of point sets $\mathscr{P}$ via the discrepancy function. The first goal is to construct point sets as uniformly distributed as possible. The deviation from a uniform distribution is then measured by some norm of the discrepancy function $D_{\mathscr{P}}$. Constructions of such point sets abound in the literature, see e.g. [15, 45, 63]. Later on we mention some constructions which are relevant for our purposes.

The second goal is to study the limits of uniformity a finite point set of fixed size n can achieve. This is done by proving lower bounds for norms of the discrepancy function which any point set of size n in the unit cube $[0,1)^d$ has to satisfy. This line of research started in 1945 with the paper [2] of van Aardenne-Ehrenfest. Since then, the search for lower bounds for the discrepancy function has continued unabated. Many of the proofs of lower bounds are inspired by the ingenious idea of Roth in 1954 [54] using orthogonal functions to pick off bits of discrepancy and add them up by orthogonality. A recent comprehensive survey on the use of the orthogonal function method is given in Bilyk's paper [5].

The study of the discrepancy function is an interesting problem on its own. What makes it even more significant is its relation to numerical integration. This brings us to the second subject of this article. Via the Hlawka-Zaremba identity and the Koksma-Hlawka inequality, the norm of the discrepancy function in some function space is also the worst case error of the quasi-Monte Carlo-rule

$$Q_n(f) = \frac{1}{n} \sum_{z \in \mathscr{P}} f(z)$$

for the integration of functions f over the unit cube $[0,1)^d$ from the unit ball of some related function space. Then limits to uniformity provide also limits to the achievable error, and point sets with low discrepancy provide good quasi-Monte Carlo rules. Section 2 deals with bounds for the norm of the discrepancy function and integration errors in function spaces.

Sometimes, randomization can be used to break barriers which are inherent in deterministic approaches like quasi-Monte Carlo rules. In this article we want to study two aspects of randomization connected to discrepancy and integration. In Sect. 3 we consider the average behavior of the discrepancy function with respect to random point sets in the unit cube $[0,1)^d$, where points are chosen independent and uniformly distributed. It is rather clear that the uniformity of the distribution of an average point set is much worse then that of the best point set of the same cardinality. Nevertheless, it is an interesting object to study. Moreover, and perhaps quite surprisingly, the average point sets are rather good for high-dimensional problems and have small L_∞-norm of the discrepancy function, which we as usual call the star discrepancy. This was demonstrated in [28].

Nevertheless, finding explicit point sets which have star discrepancy as good as a typical point set is a difficult problem. In Sect. 4 we consider an approach to use structured point sets in moderate dimension which have the best known star discrepancy of explicitly given sets. This approach is new and, as it stands, is rather a proof of concept. Nevertheless, we expect that in this direction much more can be done.

In Sect. 5 we return to the question of the complexity of integration. We review recent results showing that tractability can be achieved via randomized algorithms for integration in Hilbert spaces of functions. Importance sampling can be shown to be optimal in this setting.

The purpose of this survey article is by no means a comprehensive treatment of all aspects of discrepancy theory and numerical integration. Rather, we want to concentrate on the explanation of some recent results on different aspects of discrepancy theory. The choice of these aspects is entirely due to the authors preferences. Nevertheless, we hope that this paints an interesting picture and inspires some future work. For the latter reason, we explicitly state a number of open problems some of which are well-known and probably difficult, others are just on the edge of our current knowledge.

Finally, for a comprehensive introduction into the subject, we have to refer to the monographs [12, 43, 46–48, 50, 66].

2 Discrepancy and Integration in Function Spaces

We start by recalling the duality between the norm of the discrepancy function of a point set $\mathscr{P} = \{t^1, \dots, t^n\} \subset [0,1)^d$ and the integration error of the quasi-Monte Carlo-rule

$$Q_n(f) = \frac{1}{n} \sum_{i=1}^{n} f(t^i)$$

for a smooth function $f : [0,1]^d \to \mathbb{R}$. For simplicity, we assume that f is identical 0 on the boundary of the cube $[0,1]^d$. Then, with

$$\mathrm{INT}_d(f) = \int_{[0,1]^d} f(x)\,\mathrm{d}x,$$

the Hlawka-Zaremba identity from [35, 70] tells us that the integration error can be computed as

$$\mathrm{INT}_d(f) - Q_n(f) = (-1)^{d+1} \int_{[0,1]^d} D_{\mathscr{P}}(x) \frac{\partial^d f(x)}{\partial x_1 \dots \partial x_d}\,\mathrm{d}x.$$

So, a duality pairing leads to the estimate

$$\left|\mathrm{INT}_d(f) - Q_n(f)\right| \le \left\| D_{\mathscr{P}} \mid X' \right\| \left\| \partial_{\mathrm{mix}} f \mid X \right\|$$

where X is some suitable normed function space with dual space X'. Here $\partial_{\mathrm{mix}} f$ abbreviates the mixed derivative $\frac{\partial^d f(x)}{\partial x_1 \dots \partial x_d}$ and $\|g \mid X\|$ is the norm of g in the normed space X. Moreover, up to some minor technicalities, $\left\| D_{\mathscr{P}} \mid X' \right\|$ is the worst case error

$$\sup \left|\mathrm{INT}_d(f) - Q_n(f)\right|$$

where the supremum is taken over all such functions f with mixed derivative in the unit ball of X. Such inequalities go under the name Koksma-Hlawka inequality and go back to [36, 39].

For $1 \le p \le \infty$, the usual Lebesgue space of p-integrable functions on the cube $[0,1]^d$ is denoted by L_p. Then it is well known that for $1 < p < \infty$ there exists a constant $c_1(p) > 0$ such that, for any $n \ge 1$, the discrepancy function of any point set $\mathscr{P}$ in $[0,1)^d$ with n points satisfies

$$\|D_{\mathscr{P}} \mid L_p\| \ge c_1(p)\, n^{-1} \left(\log n\right)^{(d-1)/2}. \tag{2}$$

This lower bound was proved by Roth in [54] for $p = 2$ (and, therefore, for $p > 2$) and extended to $1 < p < 2$ by Schmidt in [59]. Constructions of point sets $\mathscr{P}$ in $[0,1)^d$ with n points with

$$\|D_{\mathscr{P}} \mid L_p\| \le c_2(p)\, n^{-1} \left(\log n\right)^{(d-1)/2}$$

where given in [11, 25] for $p = 2$ and $d = 2$, in [56] for $p = 2$ and $d = 3$, in [57] for $p = 2$ and $d \ge 4$, in [55] for $2 < p < \infty$ and $d = 2$ and in [7] for $2 < p < \infty$

and $d \geq 2$. Since the constructions for $p = 2$ are equally suitable for $1 < p < 2$, this solved the problem of the asymptotic behavior of the discrepancy function in L_p for $1 < p < \infty$. All the above constructions for $d > 2$ were probabilistic, so they just gave existence results. Explicit constructions for $p = 2$ where found in [8] and for $p > 2$ in [60].

The boundary cases $p = 1$ and $p = \infty$ turned out to be more difficult. In dimension $d = 2$ the construction of van der Corput in [9, 10] and the lower bound of Schmidt in [58] show that there exist constants $c_1, c_2 > 0$ such that the right asymptotics is

$$\inf_{\#\mathscr{P}=n} \left\| D_{\mathscr{P}} | L_\infty([0,1)^2) \right\| \asymp n^{-1} \log n.$$

In dimension $d > 2$ we only have an upper bound of order $n^{-1} (\log n)^{d-1}$ due to Halton [24] using Hammersley points [26] and a slight improvement over the lower bound of Roth of order $n^{-1} (\log n)^{(d-1)/2+\eta_d}$ for some small constant η_d proved only recently by Bilyk, Lacey and Vagharshakyan in [6]. So we still have what is sometimes called the *great open problem* of discrepancy theory.

Problem 1. What is the right asymptotics in n of

$$\inf_{\#\mathscr{P}=n} \left\| D_{\mathscr{P}} | L_\infty([0,1)^d) \right\|$$

for fixed $d \geq 3$?

For $p = 1$, we only have a lower bound of order $n^{-1} (\log n)^{1/2}$ due to Halász [23] and the upper bound $n^{-1} (\log n)^{(d-1)/2}$ which already follows from the L_2-constructions above. This is fine for $d = 2$, but for $d \geq 3$ it again leaves an important open problem.

Problem 2. What is the right asymptotics in n of

$$\inf_{\#\mathscr{P}=n} \left\| D_{\mathscr{P}} | L_1([0,1)^d) \right\|$$

for fixed $d \geq 3$?

Apart from the L_p-norms, until recently there was little done for other norms. This changed when Lacey and collaborators started to make use of the full power of Littlewood-Paley Theory to obtain good lower bounds for the discrepancy function in Hardy spaces, in Orlicz spaces of type $L(\log L)^\alpha$ and $\exp(L^\alpha)$. This story is nicely explained in Bilyk's survey [5].

In [68] and, in particular, in the book [67] Triebel promoted the study of the discrepancy function in other function spaces such as suitable Sobolev, Besov or Triebel-Lizorkin spaces to gain more insight into its behavior and into applications to numerical integration. In this section, we want to discuss the recently obtained results for discrepancy and corresponding integration errors in these spaces.

The first main tool to prove such results are characterizations of Besov spaces of dominating mixed smoothness via the coefficients of the expansion of the functions in dyadic or b-adic Haar bases. Then the lower bounds can be obtained via Roth's original idea of using the easy to compute coefficients for Haar functions whose support does not contain a point of the point set. Also the best known constants in the lower bound (2) for $p = 2$ obtained in [32] rely on this approach. For the upper bounds, suitable point sets like Hammersley type point sets and Chen-Skriganov point sets can be used. With the help of known embeddings between function spaces of different type, such results can then be transferred from Besov spaces to Triebel-Lizorkin spaces and Sobolev spaces of dominating mixed smoothness.

2.1 Function Spaces of Dominating Mixed Smoothness and Haar Functions

In this section we introduce the spaces $S^r_{pq}B([0,1)^d)$ and give a characterization in terms of Haar expansions. Spaces of dominating mixed smoothness have a long history and a huge literature. Since we are going to recall only the relevant parts from [67] for our purposes we refer to the references therein for pointers to surveys and results about these spaces.

We let $\mathscr{S}(\mathbb{R}^d)$ stand for the Schwartz space and $\mathscr{S}'(\mathbb{R}^d)$ for the space of tempered distributions on $\mathbb{R}^d$. For $f \in \mathscr{S}'(\mathbb{R}^d)$, the Fourier transform and its inverse are denoted by $\hat{f}$ and $\check{f}$, respectively. Let $\phi_0 \in \mathscr{S}(\mathbb{R})$ satisfy $\phi_0(t) = 1$ for $|t| \le 1$ and $\phi_0(t) = 0$ for $t > 3/2$. Define

$$\phi_\ell(t) = \phi_0\big(2^{-\ell}t\big) - \phi_0\big(2^{-\ell+1}t\big) \quad \text{for } t \in \mathbb{R},\ \ell \in \mathbb{N}$$

and

$$\phi_k(x) = \phi_{k_1}(x_1)\dots\phi_{k_d}(x_d) \quad \text{for } k = (k_1,\dots,k_d) \in \mathbb{N}_0^d,\ x = (x_1,\dots,x_d) \in \mathbb{R}^d.$$

The functions ϕ_k form a (dyadic) resolution of unity. The functions $(\phi_k \hat{f})\check{}$ are entire analytic functions for any $f \in \mathscr{S}'(\mathbb{R}^d)$.

For $0 < p,q \le \infty$ and $r \in \mathbb{R}$, the Besov space $S^r_{pq}B(\mathbb{R}^d)$ of dominating mixed smoothness can now be defined as the collection of all $f \in \mathscr{S}'(\mathbb{R}^d)$ for which the quasi-norm

$$\big\| f \mid S^r_{pq}B(\mathbb{R}^d)\big\| = \left(\sum_{k\in\mathbb{N}_0^d} 2^{r(k_1+\dots+k_d)q}\, \big\|(\phi_k \hat{f})\check{}\,\big|\, L_p(\mathbb{R}^d)\big\|^q\right)^{1/q}$$

is finite, with the usual modification if $q = \infty$. Let $\mathscr{D}([0,1)^d)$ stand for the collection of all complex-valued infinitely differentiable functions on $\mathbb{R}^d$ with

compact support in the interior of $[0,1)^d$ and let $\mathscr{D}'([0,1)^d)$ be its dual space of all distributions in $[0,1)^d$. Finally, the Besov spaces $S^r_{pq}B([0,1)^d)$ of dominating mixed smoothness on the domain $[0,1]^d$ is the collection of all distributions $f \in \mathscr{D}'([0,1]^d)$ which are restrictions of some $g \in S^r_{pq}B(\mathbb{R}^d)$, with quasi-norm given by

$$\left\| f \mid S^r_{pq}B([0,1)^d) \right\| = \inf \left\{ \left\| f \mid S^r_{pq}B(\mathbb{R}^d) \right\| \;\middle|\; g|_{[0,1)^d} = f \right\}.$$

The spaces $S^r_{pq}B(\mathbb{R}^d)$ and $S^r_{pq}B([0,1)^d)$ are quasi-Banach spaces.

Observe that in [67] function spaces are defined on domains which are open subsets of $\mathbb{R}^d$. Since discrepancy is formally better dealt with on $[0,1)^d$, we abused notation a little.

This extrinsic definition of the Besov spaces of dominating mixed smoothness is not of much use to us for the computation of the norm of the discrepancy function. To facilitate explicit computations we need the following intrinsic characterization in terms of Haar expansions. For this purpose, we first introduce the Haar system on the interval $[0,1)$ and the tensor Haar system on $[0,1)^d$.

A dyadic interval of length 2^{-j}, $j \in \mathbb{N}_0$, in $[0,1)$ is an interval of the form $I = I_{j,m} := \left[2^{-j}m, 2^{-j}(m+1)\right)$ for $m = 0,1,\ldots,2^j-1$. The left and right half of $I = I_{j,m}$ are the dyadic intervals $I^+ = I^+_{j,m} = I_{j+1,2m}$ and $I^- = I^-_{j,m} = I_{j+1,2m+1}$, respectively. The Haar function $h_I = h_{j,m}$ with support I is the function on $[0,1)$ which is $+1$ on the left half of I, -1 on the right half of I and 0 outside of I. The L_∞-normalized Haar system consists of all Haar functions $h_{j,m}$ with $j \in \mathbb{N}_0$ and $m = 0,1,\ldots,2^j-1$ together with the indicator function $h_{-1,0}$ of $[0,1)$. Normalized in $L_2([0,1))$ we obtain the orthonormal Haar basis of $L_2([0,1))$.

Let $\mathbb{N}_{-1} = \{-1,0,1,2,\ldots\}$ and define $\mathbb{D}_j = \{0,1,\ldots,2^j-1\}$ for $j \in \mathbb{N}_0$ and $\mathbb{D}_{-1} = \{0\}$ for $j = -1$. For $j = (j_1,\ldots,j_d) \in \mathbb{N}^d_{-1}$ and $m = (m_1,\ldots,m_d) \in \mathbb{D}_j := \mathbb{D}_{j_1} \times \cdots \times \mathbb{D}_{j_d}$, the Haar function $h_{j,m}$ is given as the tensor product $h_{j,m}(x) = h_{j_1,m_1}(x_1) \ldots h_{j_d,m_d}(x_d)$ for $x = (x_1,\ldots,x_d) \in [0,1)^d$. We will also call the rectangles $I_{j,m} = I_{j_1,m_1} \times \cdots \times I_{j_d,m_d}$ dyadic boxes. The L_∞-normalized tensor Haar system consists of all Haar functions $h_{j,m}$ with $j \in \mathbb{N}^d_{-1}$ and $m \in \mathbb{D}_j$. Normalized in $L_2([0,1)^d)$ we obtain the orthonormal Haar basis of $L_2([0,1)^d)$.

The announced intrinsic characterization of the Besov spaces $S^r_{pq}B([0,1)^d)$ of dominating mixed smoothness can now be formulated as follows. For $d = 2$, this is Theorem 2.41 in [67]. The characterization can be extended to dimension $d > 2$ more or less straightforward.

Theorem 1. *Let $0 < p \le \infty, 0 < q \le \infty$, ($1 < q \le \infty$ if $p = \infty$) and $1/p - 1 < r < \min(1/p, 1)$. Let $f \in \mathscr{D}'([0,1)^d)$. Then $f \in S^r_{pq}B([0,1)^d)$ if and only if it can be represented as*

$$f = \sum_{j \in \mathbb{N}^d_{-1}} \sum_{m \in \mathbb{D}_j} \mu_{j,m} 2^{\max\{0,j_1\}+\cdots+\max\{0,j_d\}} h_{j,m}$$

for some sequence $(\mu_{j,m})$ *with*

$$\left(\sum_{j \in \mathbb{N}_{-1}^d} 2^{(j_1+\cdots+j_d)(r-1/p+1)q} \Big(\sum_{m \in \mathbb{D}_j} |\mu_{j,m}|^p \Big)^{q/p} \right)^{1/q} < \infty, \tag{3}$$

where the convergence is unconditional in $\mathscr{D}'([0,1)^d)$ *and in any* $S^{\varrho}_{pq}B([0,1)^d)$ *with* $\varrho < r$. *This representation of* f *is unique with the Haar coefficients*

$$\mu_{j,m} = \mu_{j,m}(f) = \int_{[0,1)^d} f(x) h_{j,m}(x)\,\mathrm{d}x. \tag{4}$$

Moreover, the expression (3) *is an equivalent quasinorm on* $S^r_{pq}B([0,1)^d)$.

This theorem shows in particular that $S^0_{2,2}B([0,1)^d) = L_2([0,1)^d)$ which reflects the fact that, after proper normalization, the system of Haar functions is an orthonormal basis of $L_2([0,1)^d)$.

For the application to the point sets of Chen-Skriganov type which are nets in base b where b can be bigger than 2 a similar characterization of the Besov spaces of dominating mixed smoothness with expansions into Haar functions in base b is needed. We do not formally state the definitions and the characterization. This was carried out by Markhasin in [40–42] to which we refer the interested reader.

In the next section we also want to state results for Triebel-Lizorkin and Sobolev spaces of dominating mixed smoothness. The Triebel-Lizorkin space with dominating mixed smoothness $S^r_{pq}F(\mathbb{R}^d)$ consists of all $f \in \mathscr{S}'(\mathbb{R}^d)$ with finite quasi-norm

$$\left\| f \,|S^r_{pq}F(\mathbb{R}^d)\right\| = \left\| \left(\sum_{k \in \mathbb{N}_0^d} 2^{r|k|q} \left| (\phi_k \hat{f})\check{}(\cdot) \right|^q \right)^{\frac{1}{q}} |L_p(\mathbb{R}^d) \right\|$$

with the usual modification if $q = \infty$. Here $|k| = k_1 + \cdots + k_d$ is the ℓ_1-norm of k. The Triebel-Lizorkin space with dominating mixed smoothness $S^r_{pq}F([0,1)^d)$ consists of all $f \in \mathscr{D}'([0,1)^d)$ with finite quasi-norm

$$\left\| f \,|S^r_{pq}F([0,1)^d)\right\| = \inf \left\{ \left\| g|S^r_{pq}F(\mathbb{R}^d)\right\| : \, g \in S^r_{pq}F(\mathbb{R}^d),\ g|_{[0,1)^d} = f \right\}.$$

The Sobolev spaces of dominating mixed smoothness can then be obtained as special cases of Triebel-Lizorkin spaces

$$S^r_p H([0,1)^d) = S^r_{p\,2}F([0,1)^d).$$

For this and much more information we refer to [40, 42, 67].

2.2 Discrepancy of Hammersley and Chen-Skriganov Point Sets

We now want to survey the known results on norms of the discrepancy function in Besov, Triebel-Lizorkin and Sobolev spaces of dominating mixed smoothness. First bounds for the norms of the discrepancy function in $S^r_{pq}B([0,1)^d)$-spaces and the $S^r_{pq}F([0,1)^d)$-spaces have been established by Triebel in [67]. There were gaps between the exponents of the lower and the upper bounds which have been closed for certain parameter values and for $d = 2$ in [31]. It turned out that the lower bounds obtained by Triebel were the right ones. The upper bounds where established using Hammersley type point sets

$$\left\{\left(\frac{t_m}{2} + \frac{t_{m-1}}{2^2} + \ldots + \frac{t_1}{2^m}, \frac{s_1}{2} + \frac{s_2}{2^2} + \ldots + \frac{s_m}{2^m}\right) \mid t_1, \ldots, t_m \in \{0,1\}\right\}$$

for some $m \in \mathbb{N}$. Here s_i can be chosen for each i independently as $s_i = t_i$ or $s_i = 1 - t_i$, so the set contains $n = 2^m$ points. For these sets, the Haar coefficients of the discrepancy function could be computed quite explicitly. Then the characterization of the Besov spaces of dominating mixed smoothness given in Theorem 1 gives optimal estimates of the norm.

This approach was generalized to Hammersley type point sets in arbitrary base b by Markhasin in [41] using the corresponding b-adic characterization of the Besov spaces of dominating mixed smoothness. It was already conjectured in [31] that the point sets which Chen and Skriganov used to give explicit examples of sets with optimal L_2- and L_p-discrepancy in [8, 60] also give the optimal discrepancy in $S^r_{pq}B([0,1)^d)$. This was finally proved by Markhasin in [40, 42]. Via embedding theorems between different function spaces of dominating mixed smoothness this also leads to matching upper and lower bounds for the discrepancy in $S^r_{pq}F([0,1)^d)$ and $S^r_pH([0,1)^d)$. The cases with matching upper and lower bounds are summarized in the next theorem.

Theorem 2. *(i) Let $1 \le p,q \le \infty$ and $q < \infty$ if $p = 1$ and $q > 1$ if $p = \infty$. Let $0 < r < \frac{1}{p}$. Then there exist constants $c_1, C_1 > 0$ such that for any integer $n \ge 2$ we have*

$$c_1\, n^{r-1}\, (\log n)^{\frac{d-1}{q}} \le \inf_{\#\mathscr{P}=n} \left\| D_{\mathscr{P}} | S^r_{pq}B([0,1)^d) \right\| \le C_1\, n^{r-1}\, (\log n)^{\frac{d-1}{q}}.$$

(ii) Let $1 \le p,q < \infty$. Let $0 < r < \frac{1}{\max(p,q)}$. Then there exist constants $c_2, C_2 > 0$ such that for any integer $n \ge 2$ we have

$$c_2\, n^{r-1}\, (\log n)^{\frac{d-1}{q}} \le \inf_{\#\mathscr{P}=n} \left\| D_{\mathscr{P}} | S^r_{pq}F([0,1)^d) \right\| \le C_2\, n^{r-1}\, (\log n)^{\frac{d-1}{q}}.$$

(iii) Let $1 \le p < \infty$ Let $0 \le r < \frac{1}{\max(p,2)}$. Then there exist constants $c_3, C_3 > 0$ such that for any integer $n \ge 2$ we have

$$c_3\, n^{r-1}\, (\log n)^{\frac{d-1}{2}} \le \inf_{\#\mathscr{P}=n} \left\| D_{\mathscr{P}} | S_p^r H([0,1)^d) \right\| \le C_3\, n^{r-1}\, (\log n)^{\frac{d-1}{2}}.$$

The constants are independent of n. They depend on d, p, q, r though. The method of proof for the upper bound is the computation of the b-adic Haar coefficients of the discrepancy function of the Chen-Skriganov point sets together with the b-adic characterization of Besov spaces of dominating mixed smoothness via Haar expansions. For some upper and lower bounds for other parameter values, where they are not yet matching, we refer the reader to [40, 42, 67].

We finally comment on the restrictions in the parameter values. The restriction to $r < \frac{1}{p}$ is necessary since the discrepancy has to be in the corresponding function spaces. The restriction $r \ge 0$ in the upper bounds is necessary for the considered point sets of Hammersley type or Chen-Skriganov type. For $r < 0$ they do not even yield the right power n^{r-1} in the main term. So there is an interesting transition taking place at $r = 0$, for $r < 0$ points from a hyperbolic cross become better than nets. But for $r < 0$ there are still gaps between lower and upper bounds, so we formulate this as an open problem.

Problem 3. What is the asymptotic behavior of

$$\inf_{\#\mathscr{P}=n} \left\| D_{\mathscr{P}} | S_{pq}^r B([0,1)^d) \right\|$$

for $r < 0$?

A solution to this problem could again be transferred with the known embeddings of function spaces to the Triebel-Lizorkin and Sobolev spaces. The restriction $r < \frac{1}{\max(p,q)}$ in the case of Triebel-Lizorkin and Sobolev spaces, which is an additional restriction for $p < q$, is due to the limitations of the embedding method. This leads to the next interesting open problem.

Problem 4. What is the asymptotic behavior of

$$\inf_{\#\mathscr{P}=n} \left\| D_{\mathscr{P}} | S_{pq}^r F([0,1)^d) \right\|$$

for $\frac{1}{\max(p,q)} < r < \frac{1}{p}$ in the case $1 \le p < q$?

2.3 *Integration Errors in Besov Spaces of Dominating Mixed Smoothness*

Let n be a positive integer and $M([0,1)^d)$ be some Banach space of functions on $[0,1)^d$. Let $M_0^1([0,1)^d)$ be the subset of the unit ball of $M([0,1)^d)$ with the property that the extensions of all elements of $M_0^1([0,1)^d)$ vanish whenever one of

the coordinates of the argument is 1. The error of quadrature formulas in $M([0,1)^d)$ with n points is

$$\mathrm{err}_n(M) = \inf_{\{x_1,\dots,x_n\}\subset[0,1)^d} \sup_{f\in M_0^1([0,1)^d)} \left| \int_{[0,1)^d} f(x)\,\mathrm{d}x - \frac{1}{n}\sum_{i=1}^{n} f(x_i) \right|.$$

Now the already mentioned Koksma-Hlawka duality can be formulated as follows, see [67].

Let $1 \le p, q \le \infty$ and $\frac{1}{p} < r < 1$. Let

$$\frac{1}{p} + \frac{1}{p'} = \frac{1}{q} + \frac{1}{q'} = 1.$$

Then there exist constants $c_1, c_2 > 0$ such that, for any integer $n \ge 2$, we have

(i)

$$c_1 \inf_{\#\mathscr{P}=n} \left\| D_{\mathscr{P}} | S^{1-r}_{p'q'} B([0,1)^d) \right\| \le \mathrm{err}_n(S^r_{pq} B) \le c_2 \inf_{\#\mathscr{P}=n} \left\| D_{\mathscr{P}} | S^{1-r}_{p'q'} B([0,1)^d) \right\|,$$

(ii)

$$c_1 \inf_{\#\mathscr{P}=n} \left\| D_{\mathscr{P}} | S^{1-r}_{p'q'} F([0,1)^d) \right\| \le \mathrm{err}_n(S^r_{pq} F) \le c_2 \inf_{\#\mathscr{P}=n} \left\| D_{\mathscr{P}} | S^{1-r}_{p'q'} F([0,1)^d) \right\|.$$

With this duality, Theorem 2 can be translated into the next result:

Theorem 3. *(i) Let $1 \le p, q \le \infty$ and $q < \infty$ if $p = 1$ and $q > 1$ if $p = \infty$. Let $\frac{1}{p} < r < 1$. Then there exist constants $c_1, C_1 > 0$ such that, for any integer $n \ge 2$, we have*

$$c_1 \frac{(\log n)^{\frac{(q-1)(d-1)}{q}}}{n^r} \le \mathrm{err}_n(S^r_{pq} B) \le C_1 \frac{(\log n)^{\frac{(q-1)(d-1)}{q}}}{n^r},$$

(ii) Let $1 \le p, q < \infty$. Let $\frac{1}{\min(p,q)} < r < 1$. Then there exist constants $c_2, C_2 > 0$ such that, for any integer $n \ge 2$, we have

$$c_2 \frac{(\log n)^{\frac{(q-1)(d-1)}{q}}}{n^r} \le \mathrm{err}_n(S^r_{pq} F) \le C_2 \frac{(\log n)^{\frac{(q-1)(d-1)}{q}}}{n^r},$$

(iii) Let $1 \le p < \infty$ Let $\frac{1}{\min(p,2)} < r \le 1$. Then there exist constants $c_3, C_3 > 0$ such that, for any integer $n \ge 2$, we have

$$c_3 \frac{(\log n)^{\frac{(q-1)(d-1)}{q}}}{n^r} \le \mathrm{err}_n(S^r_{p} H) \le C_3 \frac{(\log n)^{\frac{(q-1)(d-1)}{q}}}{n^r}.$$

Again, for some upper and lower bounds for other parameter values, where they are not yet matching, we refer the reader to [40, 42, 67].

3 Average Discrepancy and Brownian Bridges

In this section we focus on the behavior of the discrepancy of a *typical* point set in $[0,1]^d$. Here typical means that we take a *random* point set $\{t^1, \dots, t^n\} \subset [0,1)^d$, where the t^i are independent random points uniformly distributed in $[0,1)^d$. While it is rather clear that such a typical point set has much worse discrepancy than the best possible point set, this approach was used in [28] to obtain the best known upper bounds for the star discrepancy of point sets of moderate size n in large dimension d. This motivated further studies of the average discrepancy. In this section we want to present recent results on the average discrepancy based on the papers [34, 64]. We start with explaining how the average discrepancy can be considered as an empirical process.

3.1 Average Discrepancy as an Empirical Process

For a fixed integer n, let $X, X_1, \dots, X_n$ be independent and identically distributed random variables defined on the same probability space with values in some measurable space M. Assume that we are given a sufficiently small class $\mathscr{F}$ of measurable real functions on M. The *empirical process* indexed by $\mathscr{F}$ is given by

$$\alpha_n(f) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big(f(X_i) - \mathbb{E}\, f(X)\big)$$

for $f \in \mathscr{F}$.

To show the relation of the empirical process and the discrepancy function, we want to use a rather general notion of discrepancy. To this end, let (Ω_d, μ_d) be a probability space. For each fixed $x \in \Omega_d$ we consider one Lebesgue-measurable subset $B(x) \subset [0,1)^d$. Furthermore, we need that the mapping $(t,x) \mapsto \mathbf{1}_{B(x)}(t)$ is also measurable. The discrepancy function of the point set $\mathscr{P} = \{t^1, \dots, t^n\} \subset [0,1)^d$ at $x \in \Omega_d$ is now given as

$$D_{\mathscr{P}}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{B(x)}(t^i) - \lambda^d(B(x)),$$

where λ^d is the d-dimensional Lebesgue measure. The classical discrepancy function of lower left corners is obtained if $\Omega_d = [0,1]^d$ and $B(x) = [0,x)$ for $x \in [0,1]^d$.

Now let $X_i = t^i$ be independent random points uniformly distributed in $[0,1)^d$ and let $\mathscr{F}$ be the class of functions $\mathbf{1}_{B(x)}$ with $x \in \Omega_d$. Then we see that

$$\alpha_n\big(\mathbf{1}_{B(x)}\big) = \sqrt{n} D_{\mathscr{P}}(x)$$

for $x \in \Omega_d$, so $\sqrt{n} D_{\mathscr{P}}$ is an empirical process indexed by $x \in \Omega_d$ (more exactly, by $\mathbf{1}_{B(x)}$).

Although we do not pursue this path in this paper, we want to comment on the approximation of empirical processes by Gaussian processes, in particular by Brownian bridges. The hope is that results on such approximations can be used to study the behavior of the discrepancy of typical sets. Although the results on the average discrepancies discussed in the subsequent subsections are proved directly, we expect that the connection to empirical processes can be used in further studies.

Let us return to the setting of general empirical processes as above. The *Brownian bridge* process G indexed by $\mathscr{F}$ is the mean zero Gaussian process with the same covariance function as the empirical process α_n which is given as

$$\langle f, g\rangle = \operatorname{cov}(G(f), G(g)) = \mathbb{E}\, f(X)g(X) - \mathbb{E}\, f(X)\, \mathbb{E}\, g(X)$$

for $f, g \in \mathscr{F}$. Under certain conditions for the class $\mathscr{F}$, it can then be shown that there exist versions of $X_1, \ldots, X_n$ and G such that

$$\sup_{f \in \mathscr{F}} |\alpha_n(f) - G(f)|$$

is very small with high probability, i.e. with high probability the corresponding paths of the empirical process α_n and of the Brownian bridge G stay close together. For concrete versions of these approximation results, we refer to [3].

The connection of the general type of discrepancy function as above to empirical processes and Brownian bridges brings up the following general question.

Problem 5. For which norms on Ω_d do we have the limit relation

$$\lim_{n\to\infty} \sqrt{n}\, \mathbb{E}\, \|D_{\mathscr{P}}\| = \mathbb{E}\, \|G\|,$$

where G is the corresponding Brownian bridge to the discrepancy considered as an empirical process? If this limit relation holds, give quantitative estimates for the speed of convergence.

The special case of the star discrepancy might be the most interesting one. For this discrepancy, the corresponding Brownian bridge is the standard multivariate Brownian bridge process on $[0,1]^d$ and the norm is the sup-norm.

While it is open if the answer to the problem above is positive for the L_p-norms of the discrepancy function of lower left corners, for the p-th powers the corresponding limit relations can be shown. This is the content of the next subsection.

3.2 Average L_p-Discrepancy

Heinrich, Novak, Wasilkowski and Woźniakowski show in [28] some results for the inverse of the star-discrepancy

$$n^*_\infty(\varepsilon, d) = \min\{n : \mathrm{disc}^*_\infty(n, d) \le \varepsilon\},$$

with the minimal L_p-star discrepancy defined by

$$\mathrm{disc}^*_p(n, d) = \inf_{\#\mathscr{P}=n, \mathscr{P}\subset[0,1)^d} \left\| D_{\mathscr{P}} | L_p \right\|.$$

They show that for the inverse of the star discrepancy the upper bound

$$n^*_\infty(\varepsilon, d) \le C d \varepsilon^{-2} \tag{5}$$

holds, where the constant C is not known. The proof of this upper bound uses the empirical process approach described above together with results on the expectation of the supremum of empirical processes. For details, we refer to [28]. Because of the unknown constant C, this term can not be computed for explicit values of ε and d. Thus, the authors introduce two other bounds for $n^*_\infty(\varepsilon, d)$ with known constants, namely

$$n^*_\infty(\varepsilon, d) \le C_k d^2 \varepsilon^{-2-1/k} \quad \text{for } k = 1, 2, \ldots$$

and

$$n^*_\infty(\varepsilon, d) = O(d \varepsilon^{-2}(\log d + \log \varepsilon^{-1})).$$

To prove the first one, the authors use a technique which is based on the analysis of the average L_p-star discrepancy, defined as

$$\mathrm{av}^*_p(n, d) = \left(\mathbb{E} \left\| D_{\mathscr{P}} | L_p \right\|^p\right)^{1/p} = \left(\int_{[0,1]^{nd}} \left\| D_{\{t^1,\ldots,t^n\}} | L_p \right\|^p \mathrm{d}t \right)^{1/p}, \tag{6}$$

for independent and uniformly distributed points $t^1, \ldots, t^n \in [0, 1)^d$. For even p they compute an explicit expression for the average L_p-star discrepancy

$$\mathrm{av}^*_p(n, d)^p = \sum_{r=p/2}^{p-1} C(r, p, d) n^{-r}, \tag{7}$$

with known constants $C(r, p, d)$, which depend on Stirling numbers of the first and second kind. Because the explicit expression for $\mathrm{av}^*_p(n, d)$ is a sum of alternating terms, it is hard to handle. Thus, the authors show the upper bound

$$\mathrm{av}^*_p(n, d) \le 3^{2/3} 2^{5/2+d/p} p (p+2)^{-d/p} n^{-1/2},$$

with p again even. To improve this bound, Hinrichs and Novak [33] used symmetrization. This technique yields an expression with only positive summands for the average L_p-star discrepancy and leads to

$$\begin{aligned} \mathrm{av}_p^*(n,d) &\le 2^{1/2+d/p} p^{1/2} (p+2)^{-d/p} n^{-1/2}, \text{ for } p \ge 2d, \\ \mathrm{av}_p^*(n,d) &\le 2^{3/2-d/p} n^{-1/2}, \qquad\qquad \text{for } p < 2d. \end{aligned}$$

This idea of symmetrization was applied by Gnewuch [18]. He computed bounds for the average L_p-extreme discrepancy $\mathrm{av}_p(n,d)$. To get this type of discrepancy axis-parallel boxes in $[-1,1]^d$ instead of boxes in $[0,1]^d$ anchored in the origin are studied. Gnewuch used symmetrization and rather simple combinatorial arguments, to get the bounds

$$\begin{aligned} \mathrm{av}_p(n,d) &\le 2^{1/2+3d/p} p^{1/2} (p+2)^{-d/p} (p+4)^{-d/p} n^{-1/2}, \text{ for } p \ge 4d, \\ \mathrm{av}_p(n,d) &\le 2^{5/4} 3^{1/4-d} n^{-1/2}, \qquad\qquad \text{for } p < 4d. \end{aligned}$$

Bounds for general $p \in [2,\infty)$ can be obtained by using Hölder's inequality (see e.g. Gnewuch [18]).

Recently, Aistleitner proved (5) with the constant $C = 100$ in [1]. Furthermore, there exists also a lower bound for the inverse of the star-discrepancy

$$n_\infty^*(\varepsilon,d) \ge \tilde{C}\,\frac{d}{\varepsilon}, \text{ with } 0 < \varepsilon < \varepsilon_0$$

which was proven by Hinrichs in [29].

The first one who conceived limit relations as discussed in the introduction to this section was Steinerberger in [64]. Although his proof contained a gap, this could be closed in [34]. We present the results for different discrepancies as discussed in [34, 64]. To this end, we define the L_p-B-discrepancy as

$$\mathrm{disc}_p^B(t^1,\dots,t^n) = \left(\int_{\Omega_d} \left| \lambda^d(B(x)) - \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{B(x)}(t^i) \right|^p \mathrm{d}\mu_d(x) \right)^{1/p}. \tag{8}$$

This definition is similar to the L_p-B-discrepancy defined by Novak and Woźniakowski in [48]. While they use densities, we use measures. If the measure μ_d is absolutely continuous with respect to the Lebesgue measure, we obtain the definition of Novak and Woźniakowski via the Radon-Nikodym theorem. Furthermore, we define the average L_p-B-discrepancy by

$$\mathrm{av}_p^B(n,d) = \left(\int_{[0,1]^{nd}} \mathrm{disc}_p^B(t^1,\dots,t^n)^p \mathrm{d}t \right)^{1/p}. \tag{9}$$

So this is nothing but the p-th root of the expectation of the p-th power of the L_p-norm of the discrepancy function considered above. We also mention that the L_2-B-discrepancy was generalized to a weighted geometric L_2-discrepancy by Gnewuch in [19].

The general limit result as proved by H. Weyhausen and the author in [34] is as follows.

Theorem 4. *Let $p > 0, d \in \mathbb{N}$, let further (Ω_d, μ_d) be a probability space and $\{B(x) : x \in \Omega_d\} \subset 2^{[0,1]^d}$ the allowed sets. Then*

$$\lim_{n\to\infty} n^{p/2} \mathrm{av}_p^B(n,d)^p = \frac{2^{p/2}}{\pi^{1/2}} \Gamma\left(\frac{1+p}{2}\right) \int_{\Omega_d} \left[\lambda^d(B(x))\big(1-\lambda^d(B(x))\big)\right]^{p/2} \mathrm{d}\mu_d(x). \tag{10}$$

Proof. Switching the order of integration we get

$$\begin{aligned} \mathrm{av}_p^B(n,d)^p &= \int_{[0,1]^{nd}} \int_{\Omega_d} \left| \lambda^d(B(x)) - \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{B(x)}(t^i) \right|^p \mathrm{d}\mu_d(x)\mathrm{d}t \\ &= \int_{\Omega_d} \int_{[0,1]^{nd}} \left| \lambda^d(B(x)) - \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{B(x)}(t^i) \right|^p \mathrm{d}t\mathrm{d}\mu_d(x). \end{aligned}$$

Now, we take a closer look at the inner integral.

Therefore, we interpret for fixed $x \in \Omega_d$ the characteristic functions $\mathbf{1}_{B(x)}(t^i)$ as Bernoulli random variables $X_i : [0,1]^{nd} \to \{0,1\}$ with probability $\lambda = \lambda^d(B(x))$, where we first assume $\lambda \neq 0, 1$. Their expected value is $\mathbb{E}(X_i) = \lambda$ and their variance is $\sigma^2(X_i) = \lambda(1-\lambda)$. Hence, the sum $\sum_{i=1}^n X_i$ is binomial distributed with expected value $\mathbb{E}\left(\sum_{i=1}^n X_i\right) = n\lambda$ and variance $\sigma^2\left(\sum_{i=1}^n X_i\right) = n\lambda(1-\lambda)$. The central limit theorem now gives for fixed $x \in \Omega_d$ with $\lambda^d(B(x)) \neq 0, 1$

$$X_{n,\lambda^d(B(\mathbf{x}))} = \left(\lambda - \frac{1}{n} \sum_{i=1}^n X_i\right) \sqrt{n} \xrightarrow{\mathscr{D}} f(\lambda) Y, \tag{11}$$

with $Y \sim \mathcal{N}(0,1)$ and $f(\lambda) = \sqrt{\lambda(1-\lambda)}$. The notation $X_n \xrightarrow{\mathscr{D}} X$ means that the random variables X_n converge in distribution to the random variable X. Observe, that (11) holds obviously for $\lambda = 1$ and $\lambda = 0$ too.

This is only a pointwise convergence for fixed x. Because there is no uniform convergence given, it is not enough to integrate over $x \in \Omega_d$ to get the result.

Instead, we will use the following approach. Let Λ be a random variable on the probability space (Ω_d, μ_d), given by

$$\Lambda(x) = \lambda^d(B(x))$$

and independent of Y. Now $X_{n,\Lambda}$ is a random variable obtained by first choosing λ according to the distribution of Λ and then using $X_{n,\lambda^d(B(x))}$. Then

$$n^{p/2}\mathrm{av}_p^B(n,d)^p = \mathbb{E}\,|X_{n,\Lambda}|^p\,.$$

We will show the equation

$$\lim_{n\to\infty} \mathbb{E}\,|X_{n,\Lambda}|^p = \mathbb{E}\,|f(\Lambda)Y|^p = \mathbb{E} f(\Lambda)^p\ \mathbb{E}\,|Y|^p\,. \tag{12}$$

This finally yields the result

$$\begin{aligned}\lim_{n\to\infty} n^{p/2}\mathrm{av}_p^B(n,d)^p &= \mathbb{E}|Y|^p \int_{\Omega_d} \Big[\lambda^d(B(x))\big(1-\lambda^d(B(x))\big)\Big]^{p/2}\mathrm{d}\mu_d(x)\\ &= \frac{2^{p/2}}{\pi^{1/2}}\Gamma\left(\frac{1+p}{2}\right)\int_{\Omega_d}\Big[\lambda^d(B(x))\big(1-\lambda^d(B(x))\big)\Big]^{p/2}\mathrm{d}\mu_d(x).\end{aligned}$$

It is enough to show

$$X_{n,\Lambda} \xrightarrow{\mathscr{D}} f(\Lambda)Y, \tag{13}$$

because $\mathrm{av}_p^B(n,d)^p$ is of order $n^{-p/2}$ for even p, which is shown in [34, Lemma 1]. Hence, we have for every even p

$$\sup_{n\in\mathbb{N}} \mathbb{E}\left(|X_{n,\Lambda}|^p\right) = \sup_{n\in\mathbb{N}} n^{p/2}\ \mathrm{av}_p^B(n,d)^p \le \sup_{n\in\mathbb{N}} n^{p/2}\ n^{-p/2}c(p) < \infty,$$

which yields (12).

Instead of (13), we will show for the characteristic functions, that

$$\lim_{n\to\infty} \varphi_{X_{n,\Lambda}} = \varphi_{f(\Lambda)Y} \tag{14}$$

holds pointwise. These functions are given by

$$\begin{aligned}\varphi_{X_{n,\Lambda}}(s) = \mathbb{E}e^{isX_{n,\Lambda}} &= \int_\Omega e^{isX_{n,\Lambda}}\mathrm{d}\mathbb{P} = \int_{\Omega_d} \mathbb{E}_t e^{isX_{n,\lambda^d(B(x))}}\mathrm{d}\mu_d(x) \ \text{ and}\\ \varphi_{f(\Lambda)Y}(s) = \mathbb{E}e^{isf(\Lambda)Y} &= \int_\Omega e^{isf(\Lambda)Y}\,\mathrm{d}\mathbb{P} = \int_{\Omega_d} \mathbb{E}_t e^{isf(\lambda^d(B(x)))Y}\mathrm{d}\mu_d(x).\end{aligned}$$

Now we have to show for fixed $s\in\mathbb{R}$, that

$$\lim_{n\to\infty}\int_{\Omega_d} \mathbb{E}_t e^{isX_{n,\lambda^d(B(x))}}\mathrm{d}\mu_d(x) = \int_{\Omega_d} \mathbb{E}_t e^{isf(\lambda^d(B(x)))Y}\mathrm{d}\mu_d(x). \tag{15}$$

The dominated convergence theorem gives us (15): the absolute value of the integrand on the left hand side is dominated by the function $g \in L_1(\Omega_d,\mu_d)$,

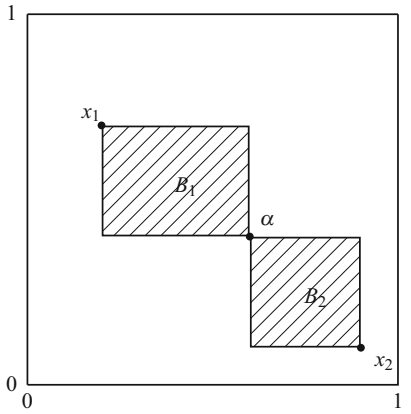


Fig. 1 L_p-discrepancy anchored in α. Boxes B_1, B_2 for points $x_1, x_2 \in \Omega_2 = [0,1]^2$.

defined by $g(x) = 1$ for $x \in \Omega_d$. Furthermore, we have for fixed x and therefore fixed $\lambda = \lambda^d(B(x))$ the equation

$$\lim_{n\to\infty} \mathbb{E}_t e^{isX_{n,\lambda^d(B(x))}} = \mathbb{E}_t e^{isf(\lambda)Y}$$

because of (11) and the fact, that the exponential function is bounded and continuous.

This yields (14) and the Lévy-Cramér continuity theorem ([4], Theorem 26.3) gives (13) and finishes the proof. □

If we choose $\Omega_d = [0,1]^d$, $\mu_d = \lambda^d$ and $B(x) = [0,x)$ we obtain the average L_p-star discrepancy.

Now we use Theorem 4 for different types of discrepancies. For estimates of the obtained limits we refer to [34, 64].

Example 1 (L_p-discrepancy anchored in α). To get $\mathrm{av}_p^{*,\alpha}(n,d)$, the average L_p-discrepancy anchored in α, we choose

$$\Omega_d = [0,1]^d \text{ and } \mu_d = \lambda^d.$$

The boxes $B(x)$ for fixed $x \in \Omega_d$ are defined as

$$B(x) = \bigtimes_{i=1}^{d} \Big[\min\{x_i, \alpha_i\}, \max\{x_i, \alpha_i\} \Big).$$

Figure 1 illustrates the Boxes B for different x.

These boxes have the Lebesgue measure

$$\lambda^d(B(x)) = \prod_{i=1}^{d} |x_i - \alpha_i|.$$

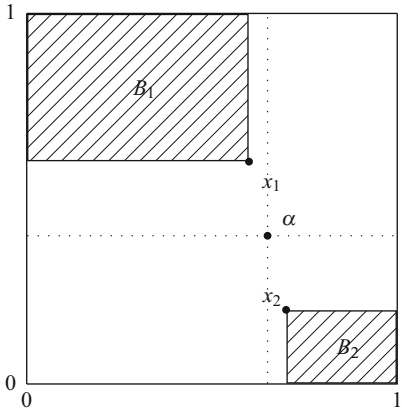


Fig. 2 Quadrant L_p-discrepancy in α. Boxes B_1, B_2 for points $x_1, x_2 \in \Omega_2 = [0,1]^2$.

Theorem 4 gives

$$\lim_{n\to\infty} n^{p/2}\mathrm{av}_p^{*,\alpha}(n,d)^p$$
$$= \frac{2^{p/2}}{\pi^{1/2}}\Gamma\left(\frac{1+p}{2}\right)\int_{[0,1]^d}\left[\prod_{i=1}^d|x_i-\alpha_i|\left(1-\prod_{i=1}^d|x_i-\alpha_i|\right)\right]^{p/2}\mathrm{d}x.$$

If we choose $\alpha = 0$ we get the boxes $[0,x)$, thus the average L_p-star discrepancy $\mathrm{av}_p^*(n,d)$. Theorem 4 gives

$$\lim_{n\to\infty} n^{p/2}\mathrm{av}_p^*(n,d)^p = \frac{2^{p/2}}{\pi^{1/2}}\Gamma\left(\frac{1+p}{2}\right)\int_{[0,1]^d}\left[\prod_{i=1}^d x_i\left(1-\prod_{i=1}^d x_i\right)\right]^{p/2}\mathrm{d}x.$$

Example 2 (Quadrant L_p-discrepancy in α). To get $\mathrm{av}_p^\alpha(n,d)$, the average quadrant L_p-discrepancy in α, we choose

$$\Omega_d = [0,1]^d \text{ and } \mu_d = \lambda^d.$$

The boxes $B(x)$ for fixed $x \in \Omega_d$ are defined as

$$B(x) = \bigtimes_{i=1}^d\Big[\mathbf{1}_{[\alpha_i,1]}(x_i)\cdot x_i, \mathbf{1}_{[\alpha_i,1]}(x_i) + \mathbf{1}_{[0,\alpha_i)}(x_i)\cdot x_i\Big).$$

Figure 2 illustrates the Boxes B for different $\mathbf{x}$.

These boxes have the Lebesgue measure

$$\lambda^d(B(x)) = \prod_{i=1}^d\left(\mathbf{1}_{[\alpha_i,1]}(x_i)(1-x_i) + \mathbf{1}_{[0,\alpha_i)}(x_i)x_i\right). \tag{16}$$

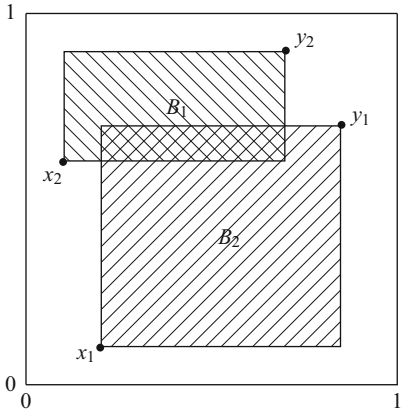


Fig. 3 Extreme L_p-discrepancy. Boxes B_1, B_2 for points $(x_1, y_1), (x_2, y_2) \in \Omega_2 \subset [0,1]^2 \times [0,1]^2$.

Theorem 4 gives

$$\lim_{n\to\infty} n^{p/2} \mathrm{av}_p^{\alpha}(n,d)^p = \frac{2^{p/2}}{\pi^{1/2}} \Gamma\left(\frac{1+p}{2}\right) \int_{[0,1]^d} \left[\lambda^d(B(x))\big(1-\lambda^d(B(x))\big)\right]^{p/2} \mathrm{d}x.$$

with $\lambda^d(B(x))$ given by (16).

Example 3 (Extreme L_p-discrepancy). To get the average extreme L_p-discrepancy $\mathrm{av}_p(n,d)$ on $[0,1]^d$ we choose

$$\Omega_d = \left\{x = (x^1, x^2) \in [0,1]^d \times [0,1]^d : x^1 \le x^2\right\} \subset [0,1]^{2d}.$$

The boxes $B(x)$ for fixed $x = (x^1, x^2) \in \Omega_d$ are defined as

$$B(x) = [x^1, x^2).$$

Figure 3 illustrates the Boxes B for different x. The measure μ_d is a normalized Lebesgue measure $c\lambda^{2d}$. To get the normalization factor c, we have to compute $\lambda^{2d}(\Omega_d)$. This yields

$$\begin{aligned}\lambda^{2d}(\Omega_d) &= \int_{[0,1]^d} \int_{[x^1,1]} 1 \mathrm{d}x^2 \mathrm{d}x^1 = \prod_{i=1}^{d} \left(\int_0^1 \int_{x_i^1}^1 1 \mathrm{d}x_i^2 \mathrm{d}x_i^1 \right) \\ &= \prod_{i=1}^{d} \left(\int_0^1 (1 - x_i^1) \mathrm{d}x_i^1 \right) = \left(\frac{1}{2}\right)^d.\end{aligned}$$

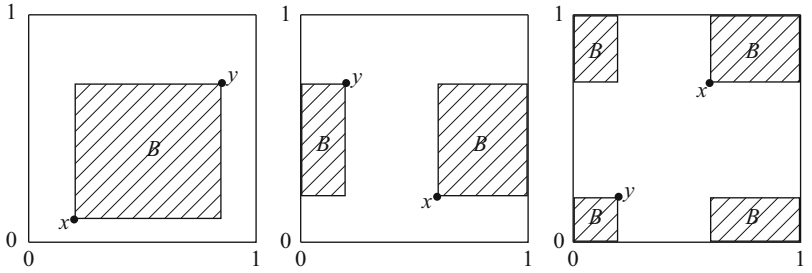


Fig. 4 Periodic L_p-discrepancy. Boxes B for points $(x, y) \in \Omega_2 = [0,1]^2 \times [0,1]^2$.

Hence, we get the measure

$$\mu_d = 2^d \lambda^{2d}.$$

The boxes $B(x)$ for fixed $x = (x^1, x^2) \in \Omega_d$ have the Lebesgue measure

$$\lambda^d(B(x)) = \prod_{i=1}^{d} \left(x_i^2 - x_i^1\right).$$

Theorem 4 yields

$$\lim_{n\to\infty} n^{p/2} \mathrm{av}_p(n,d)^p$$
$$= \frac{2^{p/2}}{\pi^{1/2}} \Gamma\left(\frac{1+p}{2}\right) \int_{[0,1]^d} \int_{[y,1]} \left[\prod_{i=1}^{d} \left(x_i^2 - x_i^1\right) \left(1 - \prod_{i=1}^{d} \left(x_i^2 - x_i^1\right)\right)\right]^{p/2} 2^d \, \mathrm{d}x^1 \mathrm{d}x^2.$$

Example 4 (Periodic L_p-discrepancy). To get the average periodic L_p-discrepancy $\mathrm{av}_p^\circ(n,d)$ we choose

$$\Omega_d = [0,1]^d \times [0,1]^d \text{ and } \mu_d = \lambda^{2d}.$$

We define the Boxes $B(x)$ for fixed $x = (x^1, x^2) \in \Omega_d$ as

$$B(x) = \bigtimes_{i=1}^{d} \left[x_i^1, \mathbf{1}_{\{x_i^1 > x_i^2\}} + \mathbf{1}_{\{x_i^1 \le x_i^2\}} x_i^2\right) \cup \left[0, \mathbf{1}_{\{x_i^1 > x_i^2\}} x_i^2\right).$$

Figure 4 illustrates the Boxes B for different x. These boxes have the Lebesgue measure

$$\lambda^d(B(x)) = \prod_{i=1}^{d} \left(\mathbf{1}_{[0,x_i^1)}(x_i^2)(1 + x_i^2 - x_i^1) + \mathbf{1}_{[x_i^1,1)}(x_i^2)(x_i^2 - x_i^1)\right). \tag{17}$$

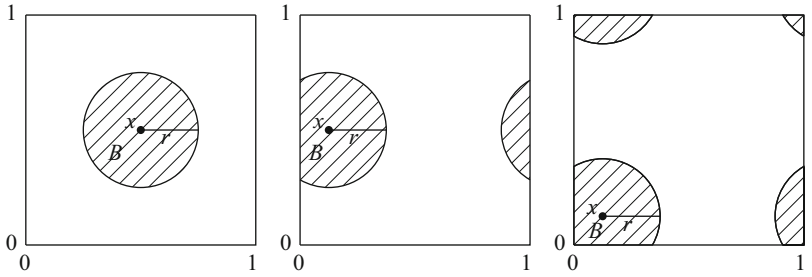


Fig. 5 Periodic ball L_p-discrepancy. Boxes B for vectors $(x, r) \in \Omega_2 = [0,1]^2 \times [0, 1/2]$.

Theorem 4 gives

$$\lim_{n\to\infty} n^{p/2} \mathrm{av}_p^\circ(n,d)^p = \frac{2^{p/2}}{\pi^{1/2}} \Gamma\left(\frac{1+p}{2}\right) \int_{[0,1]^d} \int_{[0,1]^d} \left[\lambda^d(B(x))\big(1-\lambda^d(B(x))\big)\right]^{p/2} \mathrm{d}x^1 \mathrm{d}x^2,$$

with $\lambda^d(B(x))$ given by (17).

Example 5 (Periodic ball L_p-discrepancy). To define the average periodic ball L_p-discrepancy $\mathrm{av}_p^\bullet(n,d)$ let $0 \le r_1 < r_2 \le \frac{1}{2}$ and e_j the jth canonical unit vector in dimension d. We choose

$$\Omega_d = [0,1]^d \times [r_1, r_2].$$

The boxes $B(y)$ for fixed $y = (x, r) \in \Omega_d$ are defined as

$$B(x,r) = \bigcup_{J \subset [d]} \left(B_r\left(x + \sum_{j\in J} e_j\right) \cap [0,1]^d \right),$$

where $B_r(x)$ is the open ball with center x and radius r. In the case $p = 2, d = 2$ this type of discrepancy was investigated by Gräf, Potts, and Steidel [22]. Figure 5 illustrates the Boxes B for different x and fixed $r = 1/4$. The measure μ_d is a normalized Lebesgue measur $c\lambda^{d+1}$. To get the normalization factor c, we have to compute $\lambda^{d+1}(\Omega_d)$. This yields

$$\lambda^{d+1}(\Omega_d) = \int_{[0,1]^d} \int_{r_1}^{r_2} 1 \mathrm{d}r \mathrm{d}x = r_2 - r_1.$$

Hence, we get the measure

$$\mu_d = \frac{1}{r_2 - r_1} \lambda^{d+1}.$$

The boxes $B(x,r)$ for fixed $(x,r) \in \Omega_d$ have the Lebesgue measure

$$\lambda^d(B(x,r)) = r^d \frac{\pi^{d/2}}{\Gamma(d/2+1)}.$$

Theorem 4 gives

$$\lim_{n\to\infty} n^{p/2} \mathrm{av}_p^\bullet(n,d)^p$$
$$= \frac{2^{p/2}}{\pi^{1/2}} \Gamma\left(\frac{1+p}{2}\right) \int_{[0,1]^d} \int_{r_1}^{r_2} \left[r^d \frac{\pi^{d/2}}{\Gamma(d/2+1)} \left(1 - r^d \frac{\pi^{d/2}}{\Gamma(d/2+1)}\right)\right]^{p/2} \mathrm{d}r\mathrm{d}x.$$

4 Star Discrepancy of Structured Sets

Since the star discrepancy is such a useful measure for the uniformity of a point distribution, it is desirable to compute it for a given point set. Whereas the L_2-norm of the discrepancy function can be computed efficiently via the Warnock formula or with an algorithm developed by Heinrich and Frank, see [16,27], the computation of the star discrepancy is NP-hard [20] and $W[1]$-hard with respect to the dimension parameter d [17].

But for very structured sets the computation of the star discrepancy is much easier. So, if all n points of the point set are on the diagonal of the unit cube $\{(x,\ldots,x) : x \in [0,1]\}$, the star-discrepancy is easily computed within time $O(n)$ (provided that the points are already sorted with respect to x). Indeed, if the coordinates of the diagonal points are $0 \le x_1 \le \cdots \le x_n \le 1$, the star discrepancy of the set

$$\mathscr{P} = \{(x_i,\ldots,x_i) : i = 1,\ldots,n\}$$

is

$$\mathrm{disc}^*(\mathscr{P}) = \|D_{\mathscr{P}}|L_\infty\| = \max_{1\le i\le n} \max \left\{ x_i - \frac{i}{n}, \frac{i-1}{n} - x_i^d \right\}.$$

Of course, these point sets can not have very small discrepancy. In fact, the best possible star discrepancy of such sets quickly approaches $1/2$ as d goes to infinity.

Nevertheless, this example illustrates the point that for certain structured sets the computation or efficient estimation of the star discrepancy might be much easier than for generic point sets. This will be used in this section to construct point sets in moderate dimensions with the best known star discrepancy of explicit point sets.

4.1 Discrepancy with Respect to a Lower Dimensional Subspace

In this section we first analyze the effect on the discrepancy which stems from considering only a lower dimensional subspace. We do not do this in full generality but rather in certain cases which are relevant for our later results and constructions. Obviously, there is much to be explored here.

So let us fix the following setting. Let $k > 1$ and e be fixed natural numbers, typically k will be $2, 3, 4$, and let the dimension be $d = ke$. Let M^e be the subset

$$M^e = \{x = (x_1, \dots, x_d) \in [0,1]^d : x_{k(m-1)+1} = x_{k(m-1)+2} = \cdots = x_{km}, \\ m = 1, 2, \dots, e\}.$$

This is an e-dimensional box inside $[0,1]^d$. Let μ be a probability measure on M^e. We define the discrepancy function of μ as

$$\operatorname{disc}(x, \mu) = \operatorname{vol}(B_x) - \mu(B_x \cap M^e) \quad \text{for} \quad x \in [0,1]^d,$$

where as before $B_x = [0, x_1) \times \cdots \times [0, x_d)$ is the anchored rectangular box with upper right corner x. The star discrepancy of μ then is

$$\operatorname{disc}^*(\mu) = \sup_{x \in [0,1]^d} |\operatorname{disc}(x, \mu)|.$$

Let us parametrize the points in M^e by $z = (z_1, \dots, z_e) \in [0,1]^e$ in the natural way by setting

$$z_m = x_{k(m-1)+1} = x_{k(m-1)+2} = \cdots = x_{km} \quad \text{for} \quad x = (x_1, \dots, x_d) \in M^e \text{ and} \\ m = 1, 2, \dots, e. \tag{18}$$

We now describe a special probability measure μ_0 by its cumulative distribution function Φ with respect to this parametrization. This cumulative distribution function is given by

$$\Phi(z) = \mu_0(B_z) = \frac{\operatorname{vol}(B_z) + \operatorname{vol}(B_z)^k}{2} \quad \text{for} \quad z \in [0,1]^d.$$

The probability measure μ_0 minimizes the star discrepancy among all probability measures on M^e.

Proposition 1. *The star discrepancy of the measure μ_0 described above is*

$$\operatorname{disc}^*(\mu_0) = \frac{1}{2} k^{1/(1-k)} (1 - k^{-1}).$$

Proof. For $x \in [0,1]^d$, the intersection of B_x with M^e is the box of points parametrized by B_z with

$$z_m = \min\{x_{k(m-1)+1} = x_{k(m-1)+2} = \cdots = x_{km}\} \quad \text{for} \quad m = 1, 2, \ldots, e. \tag{19}$$

Hence

$$\operatorname{disc}(x, \mu_0) = \prod_{i=1}^{d} x_i - \frac{\prod_{j=1}^{e} z_j + \prod_{j=1}^{e} z_j^k}{2}.$$

It follows that

$$\operatorname{disc}(x, \mu_0) \le \prod_{j=1}^{e} z_j - \frac{\prod_{j=1}^{e} z_j + \prod_{j=1}^{e} z_j^k}{2} = \frac{\prod_{j=1}^{e} z_j - \prod_{j=1}^{e} z_j^k}{2}$$

and

$$\operatorname{disc}(x, \mu_0) \ge \prod_{j=1}^{e} z_j^k - \frac{\prod_{j=1}^{e} z_j + \prod_{j=1}^{e} z_j^k}{2} = -\frac{\prod_{j=1}^{e} z_j - \prod_{j=1}^{e} z_j^k}{2}.$$

Moreover, for fixed $z \in [0,1]^e$ these bounds are attained for a certain $x \in [0,1]^d$. We conclude that

$$\operatorname{disc}^*(\mu_0) = \sup_{t \in [0,1]} \frac{t - t^k}{2} = \frac{1}{2} k^{1/(1-k)} (1 - k^{-1}).$$

This completes the proof. □

Remark 1. Observe that we obtain for the star discrepancy of μ_0 for $k = 2, 3, 4$ the values $0.125, 0.19245\ldots, 0.23623\ldots$, respectively. Moreover, for larger k the star discrepancy of μ_0 quickly approaches $\frac{1}{2}$. Hence this type of approximation is only suitable for star discrepancies which are not to small. More ingenious methods are necessary for smaller values. Since already values for the star-discrepancy around $\frac{1}{4}$ (see Open Problem 42 in [48]) are interesting we concentrate here on this simpler approach.

4.2 Approximation of μ_0 by Point Sets

We now approximate the measure μ_0 in the classical discrepancy fashion with an average of point masses. Together with the results from the preceding subsection we obtain point sets in M^e whose star discrepancy as set in $[0,1]^d$ can be easily estimated.

We use the following construction which starts with a point set $\mathscr{P}_n$ in $[0,1]^e$ with cardinality n. For each $z = (z_1, \dots, z_e) \in \mathscr{P}_n$ let $\tilde{z} = (z_1^k, \dots, z_e^k) \in [0,1]^e$ and let $\tilde{\mathscr{P}}_n = \{z : \tilde{z} \in \mathscr{P}_n\}$. Let $\mathscr{P}_{2n}$ be the union of $\mathscr{P}_n$ and $\tilde{\mathscr{P}}_n$. To be more precise, if a point occurs in both $\mathscr{P}_n$ and $\tilde{\mathscr{P}}_n$ we have to take it twice.

We again identify M^e with $[0,1]^e$ via the parametrization (18) and obtain a point set with cardinality $2n$ in $M^e \subset [0,1]^d$. The main result is the following theorem.

Theorem 5. *The star discrepancy of the set $\mathscr{P}_{2n}$ in $[0,1]^d$ can be estimated as*

$$\operatorname{disc}^*(\mathscr{P}_{2n}) \le \operatorname{disc}^*(\mu_0) + \operatorname{disc}^*(\mathscr{P}_n) = \frac{1}{2} k^{1/(1-k)}(1 - k^{-1}) + \operatorname{disc}^*(\mathscr{P}_n).$$

Proof. For $z \in [0,1]^e$ we define the relative discrepancy function of a finite set $\mathscr{P} \subset [0,1]^e$ with respect to μ_0 as

$$\operatorname{disc}(z, \mu_0, \mathscr{P}) = \mu_0(B_z) - \frac{\#B_z \cap \mathscr{P}}{\#\mathscr{P}}.$$

Obviously, for $x \in [0,1]^d$

$$\operatorname{disc}(x, \mathscr{P}_{2n}) \le \operatorname{disc}(x, \mu_0) + \operatorname{disc}(z, \mu_0, \mathscr{P}_{2n})$$

where z is again given by (19). Now Proposition 1 implies that to prove the theorem it is enough to show

$$|\operatorname{disc}(z, \mu_0, \mathscr{P}_{2n})| \le \operatorname{disc}_e^*(\mathscr{P}_n). \tag{20}$$

To this end we calculate

$$\begin{aligned}
\operatorname{disc}(z, \mu_0, \mathscr{P}_{2n}) &= \mu_0(B_z) - \frac{\#B_z \cap \mathscr{P}_{2n}}{2n} \\
&= \frac{\operatorname{vol}(B_z) + \operatorname{vol}(B_z)^k}{2} - \frac{\#B_z \cap \mathscr{P}_n + \#B_z \cap \tilde{\mathscr{P}}_n}{2n} \\
&= \frac{1}{2}\left(\operatorname{vol}(B_z) - \frac{\#B_z \cap \mathscr{P}_n}{n} + \operatorname{vol}(B_z)^k - \frac{\#B_z \cap \tilde{\mathscr{P}}_n}{n}\right) \\
&= \frac{1}{2}\left(\operatorname{vol}(B_z) - \frac{\#B_z \cap \mathscr{P}_n}{n} + \operatorname{vol}(B_{\tilde{z}}) - \frac{\#B_{\tilde{z}} \cap \mathscr{P}_n}{n}\right) \\
&= \frac{1}{2}\big(\operatorname{disc}(z, \mathscr{P}_n) + \operatorname{disc}(\tilde{z}, \mathscr{P}_n)\big)
\end{aligned}$$

Now (20) follows and the proof is completed. □

Example 6. As an example we treat the Open Problem 42 from [48] which asks to find explicitly n points in $[0,1]^d$ with star discrepancy bounded by $1/4$ and

(a) $n \leq 1{,}528$ and $d = 15$
(b) $n \leq 3{,}187$ and $d = 30$
(c) $n \leq 5{,}517$ and $d = 50$.

The existence of such sets is ensured by the non-constructive upper bounds in [14]. Now Theorem 5 provides such points if we find point sets with

(a) $n \leq 764$ and disc ≤ 0.125, $d = 8$ or disc ≤ 0.0575, $d = 5$ or disc ≤ 0.0136, $d = 4$
(b) $n \leq 1{,}593$ and disc ≤ 0.125, $d = 15$ or disc ≤ 0.0575, $d = 10$ or disc ≤ 0.0136, $d = 8$
(c) $n \leq 2{,}758$ and disc ≤ 0.125, $d = 25$ or disc ≤ 0.0575, $d = 17$ or disc ≤ 0.0136, $d = 13$.

Actually, there exist 128 Sobol points for $d = 8$ with discrepancy 0.1202, see [21]. This already solves the problem instance (a). It seems that there even exists a point set in dimension $d = 8$ with 97 points and star discrepancy 0.1214 (De Rainville, F.-M., Winzen, C.: Private communication).

We close this section with two general problems.

Problem 6. Can structured sets be used to find point sets of moderate size with small star discrepancy in arbitrary dimension?

Problem 7. Can structure be used to considerably speed up the algorithms for computing the star discrepancy?

5 Importance Sampling and Tractability of High Dimensional Integration

In this section we return to algorithms for the integration problem with a focus on *randomized algorithms* and *tractability*. The spaces of functions under consideration are reproducing kernel Hilbert spaces. In the first part we describe the approach via importance sampling which was developed in [30] and comment on its optimality which was demonstrated in [49]. The abstract approach via change of density results from Banach space theory does not yield explicit sampling densities. For one important example, which is the kernel

$$K_d(x,y) = \prod_{j=1}^{d} \big(1 + \min(x_j, y_j)\big),$$

we give an explicit sampling density. This result is new.

5.1 Optimality of Importance Sampling

For a probability density function ρ on a Borel measurable set $D \subseteq \mathbb{R}^d$ we consider the integration problem

$$\mathrm{INT}_d(f) = \int_D f(x)\rho(x)\,\mathrm{d}x, \tag{21}$$

where the functions $f : D_d \to \mathbb{R}$ belong to some Hilbert space H of functions. We consider randomized algorithms using n function evaluations of f to approximate the integral $\mathrm{INT}_d(f)$. In order to have function values well defined we assume that H is a reproducing kernel Hilbert space with kernel $K : D \times D \to \mathbb{R}$.

For the integration problem (21) to be well defined it is necessary that

$$\left| \int_D f(x)\rho(x)\,\mathrm{d}x \right| < \infty \quad \text{for } f \in H \tag{22}$$

which implies that also

$$\int_D |f(x)|\rho(x)\,\mathrm{d}x < \infty \quad \text{for } f \in H,$$

i.e. H is a subset of $L_1(\rho)$.

If the kernel is positive, this is equivalent to the requirement that the initial error and the norm of the functional INT_d

$$C^{\mathrm{init}} = \left(\int_D \int_D K(x,y)\rho(x)\rho(y)\,\mathrm{d}x\mathrm{d}y \right)^{1/2} < \infty, \tag{23}$$

is finite.

In general, (22) or does not imply that $H \subset L_2(\rho)$. Hence standard Monte-Carlo approximation of the integral does not necessarily have finite error, and it is not clear whether randomized algorithms with error of order $n^{-1/2}$ exist. The main result of this paper is that importance sampling is a possible remedy. We will need that the embedding operator

$$J_H : H \to L_1(\rho) \tag{24}$$

is not only well-defined but also bounded which means that

$$\int_D |f(x)|\rho(x)\,\mathrm{d}x \le C^{\mathrm{norm}} \|f\|_H \quad \text{for } f \in H$$

where $C^{\mathrm{norm}} = \|J_H\|$ is the operator norm of the embedding operator (24).

The boundedness of J_H is a consequence of the Closed Graph Theorem as follows. Assume that a sequence $(f_n) \subset H$ converges to f in H and to g in $L_1(\rho)$. Since H is a reproducing kernel Hilbert space, f is also the pointwise limit of the sequence (f_n). Moreover, convergence in $L_1(\rho)$ implies convergence in measure with respect to the measure $\rho\,\mathrm{d}x$. Convergence of (f_n) to g in measure now implies that a subsequence of (f_n) converges to g almost everywhere with respect to $\rho\,\mathrm{d}x$. Now f and g are equal almost everywhere with respect to $\rho\,\mathrm{d}x$, so they are equal in $L_1(\rho)$, the graph of J_H is indeed closed.

Importance sampling with another probability density function ω on D means that we write the integral (21) as

$$\mathrm{INT}_d(f) = \int_D \frac{f(x)\rho(x)}{\omega(x)}\,\omega(x)\,\mathrm{d}x,$$

choose n random sample points $x_1, x_2, \ldots, x_n$ according to the probability density ω and use the Monte-Carlo algorithm

$$Q_n(f) = \frac{1}{n}\sum_{i=1}^{n} \frac{f(x_i)\rho(x_i)}{\omega(x_i)}. \tag{25}$$

In the case $\omega = \rho$ we obtain the standard Monte-Carlo algorithm.

The worst case error of the randomized algorithm (25) is e_n given by the formula

$$\begin{aligned} e_n^2 &= \sup_{\|f\|_H \le 1} \mathbb{E}\,|\mathrm{INT}_d(f) - Q_n(f)|^2 \\ &= \frac{1}{n} \sup_{\|f\|_H \le 1} \left(\int_D \frac{f(x)^2\rho(x)^2}{\omega(x)}\,\mathrm{d}x - \mathrm{INT}_d(f)^2 \right), \end{aligned} \tag{26}$$

where the expectation is with respect to the random choice of the sample points according to the probability density ω. Now, independent of the concrete integral $\mathrm{INT}_d(f)$ in question, we have the estimate

$$e_n \le n^{-1/2}\, C(\omega),$$

where $C(\omega)$ is given by

$$C(\omega) = \left(\sup_{\|f\|_H \le 1} \int_D \frac{f(x)^2\rho(x)^2}{\omega(x)}\,\mathrm{d}x \right)^{1/2}.$$

Let

$$C^{\mathrm{imps}} = \inf_{\omega} C(\omega)$$

where the infimum is taken over all probability densities on D. Hence $C^{\text{imps}} < \infty$ is a sufficient condition for importance sampling to have a worst case error of order $n^{-1/2}$.

In the paper [52] the authors use the inequality $|f(t)| \le \sqrt{K(t,t)}$ for $\|f\| \le 1$ to conclude that

$$C(\omega) \le \left(\int_D \frac{K(x,x)\rho^2(x)}{\omega(x)} \, \mathrm{d}x \right)^{1/2}. \tag{27}$$

For the standard Monte-Carlo case $\omega = \rho$ one obtains the standard diagonal kernel condition

$$C^{\text{std}} := \left(\int_D K(x,x)\rho(x) \, \mathrm{d}x \right)^{1/2} < \infty$$

as a sufficient condition for a worst case error of standard Monte-Carlo of order $n^{-1/2}$.

Of course one can define ω^* to minimize the right side of inequality (27)

$$\omega^*(x) = \frac{\sqrt{K(x,x)}\rho(x)}{C} \tag{28}$$

and it is clear that $C(\omega^*) < \infty$ if the diagonal kernel condition

$$C^{\text{sqrt}} = \int_D \sqrt{K(x,x)}\rho(x) \, \mathrm{d}x < \infty.$$

is satisfied.

The chain of inequalities

$$C^{\text{init}} \le C^{\text{imps}} \le C^{\text{sqrt}} \le C^{\text{std}}$$

is obvious. In [52], examples are analyzed for which C^{std} is infinite but C^{sqrt} is finite. Here we will go beyond this condition and analyze cases where C^{sqrt} is infinite but we still have finite C^{imps}.

We will proceed as follows. In Sect. 5.1.1 we describe the necessary tools from Banach space theory. In particular we need a certain amount of knowledge about p-summing operators, the Little Grothendieck Theorem and its application to change of density results. Since our paper seems to be the first application of this part of Banach space theory to complexity studies of integration algorithms we go into some detail here. In particular, we present the key steps of the proof of the required change of density result in order to illuminate how the density ω is constructed.

Section 5.1.2 the contains the main result and its proof which shows that importance sampling works with worst case error of the order $n^{-1/2}$. In Sect. 5.1.3 we study tractability of multivariate integration problems. In Sect. 5.1.4 we present examples which show that our approach gives new algorithms and new information about the integration problem in the randomized setting.

5.1.1 Change of Density

We know that $C^{\text{init}} < \infty$ means that H is a subspace of $L_1(\rho)$ and that $J_H : H \to L_1(\rho)$ is a bounded operator. We would like to change the density so that H is actually a subspace of L_2. This suggests to try the change of density approach which is used in Banach space theory to study the structure of (mainly finite dimensional) subspaces of L_p.

Our result will be that, after an appropriate change of density, H is not only a subspace of L_2 but the corresponding embedding operator is also bounded.

We need the concept of p-summing operators. We recall here the definition and the Pietsch Domination Theorem, which is basic in theory and application of p-summing operators. For further information we refer the reader to the books [13,51, 65] and, for a gentle introduction, to [37].

Let $1 \le p < \infty$. A bounded linear operator T from a Banach space X to a Banach space Y is called p-summing if there exists a constant $c \ge 0$ such that

$$\sum_{i=1}^{n} \|Tx_i\|^p \le c^p \sup_{\|a\|_{X'} \le 1} \sum_{i=1}^{n} |a(x_i)|^p$$

for every n and every family $x_1, x_2, \ldots, x_n \in X$. Here X' is the dual Banach space of linear and bounded functionals on X. In other words, a p-summing operator maps weakly p-summable sequences in X to strongly p-summable sequences in Y. The smallest possible constant c is the p-summing norm of T and is denoted by $\pi_p(T)$.

The Pietsch Domination Theorem is the fundamental characterization of p-summing operators. For the formulation given here, see e.g. [51, 17.3.2 and 17.3.3].

Theorem 6. *The operator $T : X \to Y$ is p-summing if and only if there exists a constant $c \ge 0$ and a regular Borel probability measure ν on the weak-$*$-compact closed unit ball $B_{X'}$ of X' such that*

$$\|Tx\|^p \le c^p \int_{B_{X'}} |a(x)|^p \, \mathrm{d}\nu(a)$$

for all $x \in X$.

If $X = C(M)$ is the space of continuous functions on a compact Hausdorff space M then ν can be chosen as a regular Borel probability measure on M such that

$$\|Tf\|^p \le c^p \int_M |f(t)|^p \, \mathrm{d}\nu(t)$$

for all $f \in C(M)$. Moreover, $\pi_p(T)$ is in both cases the smallest possible constant c.

The proof of this theorem requires a Hahn-Banach type argument. Hence, in general, the measure ν is not obtained in a constructive way. However, in many

cases such a measure can be given explicitly. The measure ν is often called a Pietsch measure for the operator T.

We now provide the change of density result that we need. In the form stated here it is taken from W. B. Johnson and G. Schechtman [38, Proposition 1, Chap. 19]. If X is actually a subspace of $L_1(\Omega,\mu)$ with the norm inherited from $L_1(\Omega,\mu)$ then this result is due to H. Rosenthal [53, Theorem 1, implication $2 \Rightarrow 3$]. The proof in that paper can be literally carried over to give the result as stated here. We also refer to B. Maurey [44].

Theorem 7. *Let X be a Banach space of functions in $L_1(\Omega,\mu)$ where μ is a probability measure. Let $J : X \to L_1(\mu)$ be the embedding and let $C = \|J\|$ be its operator norm which is assumed to be finite. Additionally, assume that X has full support, i.e. that there does not exist a measurable subset of Ω with positive measure such that all $f \in X$ equal 0 almost everywhere on this subset. If the dual operator $J' : L_\infty(\Omega,\mu) \to X'$ is q-summing for some $1 \le q < \infty$ then there exists a measurable function $g > 0$ on Ω such that $\int_\Omega g \,\mathrm{d}\mu = 1$ and such that the isometry*

$$M : L_1(\Omega,\mu) \to L_1(\Omega, g\,\mathrm{d}\mu) \ \text{ given by } Mf = fg^{-1}$$

maps X to a space $\tilde{X} = M(X)$ which is contained in $L_p(\Omega, g\,\mathrm{d}\mu)$, where p is the dual index of q defined as $1/p + 1/q = 1$. Moreover, if we equip $\tilde{X}$ with the norm from X, i.e. if we set

$$\|Mf|\tilde{X}\| = \|f|X\| \ \text{ for } f \in X,$$

then the embedding $\tilde{J} : \tilde{X} \to L_p(\Omega, g\,\mathrm{d}\mu)$ has norm

$$\|\tilde{J} : \tilde{X} \to L_p(\Omega, g\,\mathrm{d}\mu)\| \le \pi_q(J' : L_\infty(\Omega,\mu) \to X').$$

In particular, under the assumptions of the Theorem, we obtain for $f \in X$ that

$$\begin{aligned}\Big(\int_\Omega |f|^p g^{1-p}\,\mathrm{d}\mu\Big)^{1/p} &= \|fg^{-1}|L_p(g\,\mathrm{d}\mu)\| = \|Mf|L_p(g\,\mathrm{d}\mu)\| \\ &\le \pi_q(J' : L_\infty(\Omega,\mu) \to X')\,\|Mf|\tilde{X}\| \\ &= \pi_q(J' : L_\infty(\Omega,\mu) \to X')\,\|f|X\|.\end{aligned}$$

5.1.2 Importance Sampling from Change of Density

Theorem 8. *Let H be a Hilbert space of functions with reproducing kernel K. Let ρ be a probability density such that $C^{\mathrm{init}} < \infty$ or, equivalently, the embedding $J_H : H \to L_1(\rho)$ is a bounded operator. Assume that H has full support with respect to the measure $\rho\,\mathrm{d}x$. Then*

$$C^{\text{imps}} \le \sqrt{\tfrac{\pi}{2}}\, \|J_H : H \to L_1(\rho)\|.$$

In particular, there exists a density function $\omega > 0$ such that the worst case error of importance sampling with density function ω for the integral

$$\text{INT}_d(f) = \int_D f(x)\rho(x)\,\mathrm{d}x$$

is bounded by

$$e_n \le \sqrt{\tfrac{\pi}{2}}\, \|J_H : H \to L_1(\rho)\|\, n^{-1/2}.$$

Remark. The restriction to densities ρ with full support is not essential. If ρ does not have full support we may restrict ρ to the support of $J_H(H)$ and renormalize. This increases $\|J_H : H \to L_1(\rho)\|$ but not the rate of convergence of e_n. This applies also to the upcoming Theorems 9 and 10.

Proof. The Little Grothendieck Theorem, see e.g. [51, 22.4.2] tells us that the dual operator $J_H' : L_\infty(\rho) \to H$ is 2-summing with

$$\pi_2(J_H') \le \sqrt{\tfrac{\pi}{2}}\, \|J_H\|.$$

Now the Change of Density Theorem 7 provides us with a measurable function $g > 0$ such that

$$\int_D g(x)\rho(x)\,\mathrm{d}x = 1$$

and

$$\Big(\int_D |f(x)|^2 g(x)^{-1}\rho(x)\,\mathrm{d}x\Big)^{1/2} \le \sqrt{\tfrac{\pi}{2}}\, \|J_H\|\, \|f\|_H \quad \text{for } f \in H.$$

Letting $\omega = g\rho$ we obtain $\omega > 0$ and $\int_D \omega = 1$ and

$$C(\omega) = \left(\sup_{\|f\|_H \le 1} \int_D \frac{f^2\rho^2}{\omega}\,\mathrm{d}x\right)^{1/2} \le \sqrt{\tfrac{\pi}{2}}\, \|J_H\|.$$

This completes the proof. □

The following theorem deals with the case that the reproducing kernel is nonnegative.

Theorem 9. *Let H be a Hilbert space of functions with nonnegative reproducing kernel K and let ρ be a probability density such that*

$$C^{\text{init}} = \left(\int_D \int_D K(x,y)\rho(x)\rho(y)\,\mathrm{d}x\mathrm{d}y\right)^{1/2} < \infty.$$

Assume that H has full support with respect to the measure $\rho\,\mathrm{d}x$. Then

$$C^{\mathrm{imps}} \leq \sqrt{\tfrac{\pi}{2}}\, C^{\mathrm{init}}.$$

In particular, there exists a density function $\omega > 0$ such that the worst case error of importance sampling with density function ω for the integral

$$\mathrm{INT}_d(f) = \int_D f(x)\rho(x)\,\mathrm{d}x$$

is bounded by

$$e_n \leq \sqrt{\tfrac{\pi}{2}}\, C^{\mathrm{init}}\, n^{-1/2}.$$

Proof. This follows immediately from Theorem 8 and the observation that for a nonnegative kernel the embedding $J_H : H \to L_1(\rho)$ is always bounded with operator norm $C^{\mathrm{norm}} = C^{\mathrm{init}}$. □

5.1.3 Tractability of Multivariate Integration

Theorems 8 and 9 have immediate consequences for the tractability of multivariate integration problems. For more details on the notion of tractability we refer to [47].

Let $\{K_d, \rho_d\}_{d=1}^{\infty}$ be a sequence of kernels and densities defined on $D_d \subset \mathbb{R}^d$ and consider the corresponding sequence of integration problems I_d from (21). Let H_d be the corresponding sequence of Hilbert spaces. Let $n(\varepsilon, d)$ be the minimal number of sample points necessary so that there exists a randomized algorithm with error $e_n \leq \varepsilon$ for the integration problem I_d, where the error is given by (26). Then the multivariate weighted integration problem is called polynomially tractable for the absolute error criterion in the randomized setting if there exist constants $c, a, b \geq 0$ such that

$$n(\varepsilon, d) \leq c\varepsilon^{-a} d^{b} \quad \text{for } \varepsilon \in (0,1) \text{ and } d = 1, 2, \ldots.$$

If $b = 0$ then the problem is called strongly polynomially tractable for the absolute error criterion. If we require $e_n / C^{\mathrm{init}} = e_n / \|I_d\| \leq \varepsilon$, the corresponding notions are (strong) polynomial tractability for the normalized error criterion.

The following result directly follows from Theorems 8 and 9.

Theorem 10. *With the above notation, let $\mathrm{CRI}_d = 1$ for the absolute error criterion and $\mathrm{CRI}_d = \|I_d\|$ for the relative error criterion. If the embeddings $J_{H_d} : H_d \to L_1(\rho_d)$ have full support with respect to the measures $\rho_d\,\mathrm{d}x$, and $\|J_{H_d}\|/\mathrm{CRI}_d$ is uniformly bounded then the multivariate weighted integration problem is strongly polynomially tractable in the randomized setting with exponent $a = 2$. This is in particular the case if all the kernels K_d are nonnegative.*

If the embeddings $J_{H_d} : H_d \to L_1(\rho_d)$ *have full support with respect to the measures* $\rho_d \, \mathrm{d}x$*, and the norms satisfy an estimate*

$$\frac{\|J_{H_d}\|}{\mathrm{CRI}_d} \leq c d^{\beta} \text{ for } d = 1, 2, \ldots.$$

for some constants $c, \beta \geq 0$*, then the multivariate weighted integration problem is polynomially tractable in the randomized setting with exponents* $a = 2$ *and* $b = 2\beta$*.*

We finally comment on the optimality of the exponent of tractability $a = 2$. Assume that we have that K_d is an integrable nonnegative kernel which is the d-fold tensor product of a one-dimensional *decomposable* kernel K_1. For the notion of decomposability we refer to [48, Chap. 11]. Then we obtain for the normalized error criterion that

$$n(\varepsilon, d) \leq \frac{\pi}{2} \varepsilon^{-2} \text{ for } 0 < \varepsilon < 1, d \in \mathbb{N}.$$

Novak and Woźniakowski showed in [49] that

$$n(\varepsilon, d) \geq \frac{1}{8} \varepsilon^{-2} \text{ for } 0 < \varepsilon < 1, d \geq \frac{2 \ln \varepsilon^{-1} - \ln 2}{\ln \alpha^{-1}},$$

where $\alpha < 1$ depends on the decomposable kernel K_1. These results are also treated in detail in [50, Chap. 23]

5.1.4 Examples

The first example which illustrates the difference between standard Monte-Carlo, the algorithm from [52] and the results of Sect. 5.1.2 is artificially constructed to show the main points clearly. It is the same example that is used in [52] to point out the difference between the conditions $C^{\mathrm{init}} < \infty$ and $C^{\mathrm{sqrt}} < \infty$.

The example is built on $D = [0, \infty)$ with the kernel

$$K(x, y) = \sum_{j=1}^{\infty} a_j^2 \, \mathbf{1}_j(x) \, \mathbf{1}_j(y),$$

where $\mathbf{1}_j$ stands for the indicator function of the interval $[j - 1, j)$. Moreover, the weight ρ is given by

$$\rho(x) = \sum_{j=1}^{\infty} r_j \, \mathbf{1}_j(x)$$

for some $r_j \geq 0$ with $\sum_j r_j = 1$. The functions $a_j \mathbf{1}_j$ are an orthonormal basis of H.

In this case,

$$C^{\mathrm{init}} = \Big(\sum_{j=1}^{\infty} a_j^2 r_j^2 \Big)^{1/2},$$

so the problem is well defined whenever $(a_j r_j) \in \ell_2$.

Since the kernel is nonnegative, Theorem 4 applies. However, it is easy to directly construct a weight ω such that

$$C(\omega) = C^{\mathrm{imps}} = C^{\mathrm{init}} = \Big(\sum_{j=1}^{\infty} a_j^2 r_j^2 \Big)^{1/2}.$$

Indeed, we can choose

$$\omega(x) = \big(C^{\mathrm{init}}\big)^{-2} \sum_{j=1}^{\infty} a_j^2 r_j^2 \, \mathbf{1}_j(x).$$

Moreover,

$$C^{\mathrm{sqrt}} = \int_0^{\infty} \sqrt{K(x,x)} \rho(x) \, \mathrm{d}x = \sum_{j=1}^{\infty} |a_j| r_j,$$

hence $C^{\mathrm{sqrt}} < \infty$ is satisfied iff $(a_j r_j) \in \ell_1$. Finally,

$$C^{\mathrm{std}} = \left(\int_0^{\infty} K(x,x) \rho(x) \, \mathrm{d}x \right)^{1/2} = \left(\sum_{j=1}^{\infty} a_j^2 r_j \right)^{1/2},$$

hence $C^{\mathrm{std}} < \infty$ is satisfied iff $(a_j^2 r_j) \in \ell_1$.

It is also worth to observe that H is a subspace of $L_2(\rho)$ (and the embedding is bounded) iff $(a_j^2 r_j) \in \ell_\infty$. So standard Monte-Carlo in this case still has error of order $n^{-1/2}$ even though C^{std} and C^{sqrt} might be infinite.

The second example uses the kernel $K(x,y) = \min\{x,y\}$ on $D = [0,\infty)$ which is the covariance kernel of the Wiener measure. In this case, the Hilbert space H obtained from the reproducing kernel K is the space of all absolutely continuous functions $f : [0,\infty) \to \mathbb{R}$ which satisfy $f' \in L_2[0,\infty)$ and $f(0) = 0$, with norm

$$\|f\|_H^2 = \int_0^{\infty} f'(x)^2 \, \mathrm{d}x < \infty.$$

We obtain

$$\begin{aligned} C^{\mathrm{init}} &= \big(2 \textstyle\int_0^{\infty} x\rho(x) \int_x^{\infty} \rho(y) \, \mathrm{d}y \, \mathrm{d}x\big)^{1/2} \\ C^{\mathrm{sqrt}} &= \big(\textstyle\int_0^{\infty} \sqrt{x}\rho(x) \mathrm{d}x\big)^{1/2} \\ C^{\mathrm{std}} &= \textstyle\int_0^{\infty} x\rho(x) \mathrm{d}x. \end{aligned}$$

The kernel is obviously nonnegative and H has full support, so Theorem 9 applies and $C^{\text{imps}} \le \sqrt{\pi/2}\, C^{\text{init}}$.

Let us first consider a polynomial weight of the form $\rho(x) = c_\alpha \min\{1, x^\alpha\}$ for $\alpha < -1$, where c_α is chosen so that ρ is a probability density. Then

$$\begin{aligned} C^{\text{init}} < \infty &\iff \textstyle\int_1^\infty x^{2+2\alpha}\,\mathrm{d}x < \infty \iff \alpha < -3/2 \\ C^{\text{sqrt}} < \infty &\iff \textstyle\int_1^\infty x^{1/2+\alpha}\,\mathrm{d}x < \infty \iff \alpha < -3/2 \\ C^{\text{std}} < \infty &\iff \textstyle\int_1^\infty x^{1+\alpha}\,\mathrm{d}x < \infty \iff \alpha < -2. \end{aligned}$$

So there is no difference between the conditions $C^{\text{init}} < \infty$ and $C^{\text{sqrt}} < \infty$ in this case, but $C^{\text{std}} < \infty$ is more restrictive.

This example is also interesting since for $\alpha = -2$ we already have infinite C^{std} but, nevertheless, H is still continuously embedded in $L_2(\rho)$ which means that standard Monte-Carlo still has worst case error of order $n^{-1/2}$. Indeed, if we set $g(x) = f'(x)$ for $f \in H$, then

$$\frac{f(x)}{x} = \frac{1}{x}\int_0^x g(y)\,\mathrm{d}y.$$

Then it follows from Hardy's inequality that $f' = g \in L_2(0,\infty)$ implies $f(x)/x \in L_2(0,\infty)$ and there exists a constant $C > 0$ such that

$$\int_0^\infty \frac{f(x)^2}{x^2}\,\mathrm{d}x \le C \int_0^\infty f'(x)^2\,\mathrm{d}x = \|f\|_H^2$$

which is what we claimed. Observe also that H is not a subset of $L_2(\rho)$ if $\alpha > -2$, so standard Monte-Carlo does not have finite error.

Now let us look more closely at the borderline case $\alpha = -3/2$. We consider a weight of the form $\rho(x) = c_\beta \min\{1, x^{-3/2}(\log(1+x))^\beta\}$ for $\beta \in \mathbb{R}$, where c_β is now chosen so that ρ is a probability density. In this case we obtain that $C^{\text{std}} = \infty$ and

$$\begin{aligned} C^{\text{init}} < \infty &\iff \textstyle\int_1^\infty x^{-1}(\log(1+x))^{2\beta}\,\mathrm{d}x < \infty \iff \beta < -1/2 \\ C^{\text{sqrt}} < \infty &\iff \textstyle\int_1^\infty x^{-1}(\log(1+x))^{\beta}\,\mathrm{d}x < \infty \iff \beta < -1. \end{aligned}$$

So in this case the difference between $C^{\text{init}} < \infty$ and $C^{\text{sqrt}} < \infty$ is again visible.

The next two examples show the application of the tractability result to uniform integration on weighted Sobolev spaces. That is, we take $D_d = [0,1]^d$ and $\rho_d(x) = 1$. Both examples were considered in [62] and in [69].

In both examples, the d-dimensional kernel

$$K_d(x,t) = \prod_{j=1}^d K^{\gamma_j}(x_j, t_j) \quad \text{for } x, t \in [0,1]^d$$

is a tensor product of one-dimensional weighted kernels K^γ for some $\gamma \geq 0$ with

$$K^\gamma(x,t) = 1 + \gamma \min(x,t) \quad \text{for } x,t \in [0,1] \tag{29}$$

in the first case and

$$K^\gamma(x,t) = 1 + \gamma(\min(x,t) - xt) \quad \text{for } x,t \in [0,1] \tag{30}$$

in the second case. The Hilbert spaces H_d are tensor products $H_d = \otimes_{j=1}^d H^{\gamma_j}$ of the one-dimensional Hilbert spaces H^γ consisting of absolutely continuous functions on $[0,1]$ whose first derivatives belong to $L_2[0,1]$ with norm

$$\|f\|_{H^\gamma}^2 = |f(0)|^2 + \frac{1}{\gamma}\int_0^1 |f'(x)|^2 \,\mathrm{d}x.$$

In the first case (29), which is called the non-periodic case, there is no further restriction on the functions f. In the second case (30), the periodic case, the functions f have the additional restriction $f(0) = f(1)$.

It is known that in both the non-periodic and the periodic case, the multivariate integration problem with the normalized error criterion is strongly polynomially tractable in the deterministic setting iff $\sum \gamma_j < \infty$ [61], and the standard Monte-Carlo algorithm is strongly polynomial iff $\sum \gamma_j^2 < \infty$ [62]. It is shown in [69] that importance sampling with the weight ω^* from (28) provides a strongly polynomial algorithm for the periodic case iff $\sum \gamma_j^3 < \infty$. Now Theorem 10 shows that the problem is strongly polynomially tractable (with importance sampling) without any condition on the weights γ_j in both the periodic and the non-periodic case.

5.2 *Explicit Importance Sampling Densities: Sampling with the Representer*

We further study the integration problem

$$\mathrm{INT}_d(f) = \int_{D_d} f(x)\rho_d(x)\,\mathrm{d}x$$

for functions $f : D_d \to \mathbb{R}$ from a reproducing kernel Hilbert space H_d of d-variate functions. Here $D_d \subset \mathbb{R}^d$ is Borel measurable and ρ_d is a probability density. The reproducing kernel is $K_d : D_d \times D_d \to \mathbb{R}$ of H_d is assumed to be positive.

The setting is the randomized setting and we want to find explicit importance sampling algorithms. That is, our algorithms have the form

$$Q_n(f) = \frac{1}{n}\sum_{k=1}^n \frac{\rho_d(x_k)}{\omega_d(x_k)} f(x_k)$$

where the $x_1, \ldots, x_n$ are iid sampled according to the importance sampling density ω_d.

Our problem is well defined if $H_d \subset L_1(\rho_d)$ or, equivalently, the embedding $J_d : H_d \to L_1(\rho_d)$ is a bounded operator. The initial error is the norm of the functional INT_d. In the case of a nonnegative kernel, we have

$$\|\mathrm{INT}_d\| = \|J_d : H_d \to L_1(\rho_d)\| = \left(\int_D \int_D K_d(x,y) \rho_d(x) \rho_d(y) \, \mathrm{d}x \mathrm{d}y \right)^{1/2}.$$

Now Theorem 8 shows that there exists a probability density ω_d such that

$$e_n \le \sqrt{\tfrac{\pi}{2}} \, \|\mathrm{INT}_d\| \, \frac{1}{\sqrt{n}}.$$

The proof is nonconstructive. It is desirable to have an explicit sampling density ω_d at least for some important cases of integration problems where the standard Monte Carlo approach with $\omega_d = 1$ does not work.

We provide such a density for the following example. The domain of integration is the unit cube $D_d = [0,1]^d$. We consider uniform integration, so the integration density is $\rho_d = 1$. K_d is a tensor product kernel of the form

$$K_d(x,y) = \prod_{j=1}^{d} \big(1 + \min(x_j, y_j)\big).$$

Then H_d is the Hilbert space tensor product of d copies of H_1 where H_1 consists of all absolutely continuous functions f on $[0,1]$ with finite norm $\|f\|_{H_1}$ which is the Hilbert space norm induced by the scalar product

$$\langle f, g \rangle_{H_1} = f(0)g(0) + \int_0^1 f'(x) g'(x) \, \mathrm{d}x.$$

Since the integration problem INT_d is also a tensor product of d one-dimensional integration problems

$$\mathrm{INT}_1(f) = \int_0^1 f(x) \, \mathrm{d}x$$

the Riesz representer h_d of the linear functional INT_d is also the tensor product of the one-dimensional representer h_1, which can be computed as

$$h_1(x) = 1 + x - \frac{x^2}{2}.$$

This representer is a positive function. We have

$$\|\mathrm{INT}_d\| = \|h_d\|_{H_d} = \left(\frac{4}{3}\right)^{d/2}.$$

Normalizing h_d yields a probability density

$$\omega_d = \frac{h_d}{\|h_d\|_{L_1}} = \frac{h_d}{\|h_d\|_{H_d}^2}.$$

The main result in this section is the following theorem.

Theorem 11. *Let H_d, ρ_d and Q_n be as above. Then $e_n \le \frac{\|\mathrm{INT}_d\|}{\sqrt{n}}$.*

Proof. We have to show that the $L_2(\omega_d)$-norm of the function f/ω_d is bounded by $\|\mathrm{INT}_d\|\,\|f\|_{H_d}$ for any $f \in H_d$, that is

$$\int_{[0,1]^d} \frac{f(x)^2}{\omega_d(x)}\,\mathrm{d}x \le \left(\frac{4}{3}\right)^d \|f\|_{H_d}^2.$$

Since everything is a tensor product, it is enough to prove the one-dimensional version

$$\int_0^1 \frac{f(x)^2}{\omega_1(x)}\,\mathrm{d}x \le \frac{4}{3}\left(f(0)^2 + \int_0^1 f'(x)^2\,\mathrm{d}x\right)$$

for every absolutely continuous function f on $[0,1]$ with square integrable derivative. Using the definition of ω_1, this is equivalent to the Sobolev type inequality

$$\int_0^1 \frac{f(x)^2}{1+x-x^2/2}\,\mathrm{d}x \le f(0)^2 + \int_0^1 f'(x)^2\,\mathrm{d}x.$$

To derive this inequality, we use the function

$$g(x) = \int_0^x f'(t)\,\mathrm{d}t$$

and use the weighted arithmetic geometric mean inequality to derive

$$\begin{aligned} f(x)^2 &= \big(f(0)+g(x)\big)^2 = f(0)^2 + 2f(0)g(x) + g(x)^2 \\ &\le f(0)^2 + \Big(x - \frac{x^2}{2}\Big) f(0)^2 + \frac{g(x)^2}{x - \frac{x^2}{2}} + g(x)^2 \\ &= \Big(1 + x - \frac{x^2}{2}\Big)\left(f(0)^2 + \frac{g(x)^2}{x - \frac{x^2}{2}}\right). \end{aligned}$$

Dividing by the expression $1 + x - x^2/2$ and integrating leads to

$$\int_0^1 \frac{f(x)^2}{1 + x - x^2/2}\,\mathrm{d}x \le f(0)^2 + \int_0^1 \frac{g(x)^2}{x - \frac{x^2}{2}}\,\mathrm{d}x.$$

So it remains to check the inequality

$$\int_0^1 \frac{g(x)^2}{x - \frac{x^2}{2}}\,\mathrm{d}x \le \int_0^1 f'(x)^2\,\mathrm{d}x.$$

To this end, we use the Cauchy-Schwarz inequality to conclude that

$$g(x)^2 = \left(\int_0^x f'(t)\,\mathrm{d}t\right)^2 \le \int_0^x (1-t)\,\mathrm{d}t \int_0^x \frac{f'(t)^2}{1-t}\,\mathrm{d}t = \left(x - \frac{x^2}{2}\right) \int_0^x \frac{f'(t)^2}{1-t}\,\mathrm{d}t.$$

Dividing by the expression $x - x^2/2$ and integrating leads to

$$\int_0^1 \frac{g(x)^2}{x - \frac{x^2}{2}}\,\mathrm{d}x \le \int_0^1 \int_0^x \frac{f'(t)^2}{1-t}\,\mathrm{d}t\,\mathrm{d}x = \int_0^1 \int_t^1 \frac{f'(t)^2}{1-t}\,\mathrm{d}x\,\mathrm{d}t = \int_0^1 f'(t)^2\,\mathrm{d}t,$$

which finishes the proof. □

We conclude with three by now natural open problems

Problem 8. For which integration problems as described in this section does importance sampling with the Riesz representer gives a randomized algorithm with

$$e_n \le C\,\frac{\|\mathrm{INT}_d\|}{\sqrt{n}}$$

for some C independent of the dimension d?

Problem 9. For which integration problems as described above can an importance sampling density ϱ_d be explicitly constructed such that

$$e_n \le C\,\frac{\|\mathrm{INT}_d\|}{\sqrt{n}}$$

for some C independent of the dimension d?

Problem 10. Can Theorem 8 also be shown for other linear functionals different from integration?

References

1. Aistleitner, C.: Covering numbers, dyadic chaining and discrepancy. J. Complexity **27**, 531–540 (2011)
2. van Aardenne-Ehrenfest, T.: Proof of the impossibility of a just distribution of an infinite sequence of points over an interval. Proc. K. Ned. Akad. Wetensch. **48**, 266–271 (1945)
3. Berthet, P., Mason, D.M.: Revisiting two strong approximation results of Dudley and Philipp. In: Giné, E., et al. (eds.) High Dimensional Probability. Lecture Notes Monograph Series, vol. 51, pp. 155–172. IMS, Beachwood (2006)
4. Billingsley, P.: Probability and Measure, 2nd edn. Wiley, New York (1986)
5. Bilyk, D.: On Roth's orthogonal function method in discrepancy theory. Unif. Distrib. Theory **6**, 143–184 (2011)
6. Bilyk, D., Lacey, M.T., Vagharshakyan, A.: On the small ball inequality in all dimensions. J. Funct. Anal. **254**, 2470–2502 (2008)
7. Chen, W.W.L., Skriganov, M.M.: On irregularities of distribution. Mathematika **27**, 153–170 (1981)
8. Chen, W.W.L., Skriganov, M.M.: Explicit constructions in the classical mean squares problem in irregularities of point distribution. J. Reine Angew. Math. **545**, 67–95 (2002)
9. van der Corput, J.G.: Verteilungsfunktionen I. Proc. Akad. Wetensch. Amst. **38**, 813–821 (1935)
10. van der Corput, J.G.: Verteilungsfunktionen II. Proc. Akad. Wetensch. Amst. **38**, 1058–1068 (1935)
11. Davenport, H.: Note on irregularities of distribution. Mathematika **3**, 131–135 (1956)
12. Dick, J., Pillichshammer, F.: Digital Nets and Sequences: Discrepancy and Quasi-Monte Carlo Integration. Cambridge University Press, Cambridge (2010)
13. Diestel, J., Jarchow, H., Tonge, A.: Absolutely Summing Operators. Cambridge University Press, Cambridge/New York (1995)
14. Doerr, B., Gnewuch, M., Srivastav, A.: Bounds and constructions for the star-discrepancy via δ-covers. J. Complexity **21**, 691–709 (2005)
15. Faure, H.: Discrépance de suites associées à un système de numération (en dimension s). Acta Arith. **41**, 337–351 (1982)
16. Frank, K., Heinrich, S.: Computing discrepancies of Smolyak quadrature rules. J. Complexity **12**, 287–314 (1996)
17. Giannopoulos, P., Knauer, C., Wahlström, M., Werner, D.: Hardness of discrepancy computation and ϵ-net verification in high dimension. J. Complexity **28**, 162–176 (2012)
18. Gnewuch, M.: Bounds for the average L_p-extreme and L_∞-extreme discrepancy. Electron. J. Combin. **12**, 11 (2005). Research Paper 54
19. Gnewuch, M.: Weighted geometric discrepancies and numerical integration on reproducing kernel Hilbert spaces. J. Complexity **28**, 2–17 (2012)
20. Gnewuch, M., Srivastav, A., Winzen, C.: Finding optimal volume subintervals with k-points and calculating the star discrepancy are NP-hard problems. J. Complexity **25**, 115–127 (2009)
21. Gnewuch, M., Wahlström, M., Winzen, C.: A new randomized algorithm to approximate the star discrepancy based on threshold accepting. SIAM J. Numer. Anal. **50**, 781–807 (2012)
22. Gräf, M., Potts, D., Steidl, G.: Quadrature rules, discrepancies and their relations to halftoning on the torus and the sphere. SIAM J. Sci. Comput. **34**, A2760–A2791 (2012)
23. Halász, G.: On Roth's method in the theory of irregularities of point distributions. In: Halberstam, H., Hooley, C. (eds.) Recent Progress in Analytic Number Theory, vol. 2, pp. 79–94. Academic, London/New York (1981)
24. Halton, J.H.: On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. Numer. Math. **2**, 84–90 (1960)
25. Halton, J.H., Zaremba, S.C.: The extreme and L^2 discrepancies of some plane sets. Monatsh. Math. **73**, 316–328 (1969)

26. Hammersley, J.M.: Monte Carlo methods for solving multivariable problems. Ann. N. Y. Acad. Sci. **86**, 844–874 (1960)
27. Heinrich, S.: Efficient algorithms for computing the L_2-discrepancy. Math. Comp. **65**, 1621–1633 (1996)
28. Heinrich, S., Novak, E., Wasilkowski, G., Woźniakowski, H.: The inverse of the star-discrepancy depends linearly on the dimension. Acta Arith. **96**, 279–302 (2001)
29. Hinrichs, A.: Covering numbers, Vapnik-Červonenkis classes and bounds for the star-discrepancy. J. Complexity **20**, 477–483 (2004)
30. Hinrichs, A.: Optimal importance sampling for the approximation of integrals. J. Complexity **26**, 125–134 (2010)
31. Hinrichs, A.: Discrepancy of Hammersley points in Besov spaces of dominating mixed smoothness. Math. Nachr. **283**, 478–488 (2010)
32. Hinrichs, A., Markhasin, L.: On lower bounds for the L_2-discrepancy. J. Complexity **27**, 127–132 (2011)
33. Hinrichs, A., Novak, E.: New bounds for the star discrepancy. Extended abstract of a talk at the Oberwolfach seminar "Discrepancy Theory and its Applications", report no. 13/2004, Mathematisches Forschungsinstitut Oberwolfach
34. Hinrichs, A., Weyhausen, H.: Asymptotic behavior of average L_p-discrepancies. J. Complexity **28**, 425–439 (2012)
35. Hlawka, E.: Über die Diskrepanz mehrdimensionaler Folgen mod. 1. Math. Z. **77**, 273–284 (1961)
36. Hlawka, E.: Funktionen von beschränkter Variation in der Theorie der Gleichverteilung. Ann. Mat. Pura Appl. **54**, 325–333 (1961)
37. Jameson, G.J.O.: Summing and Nuclear Norms in Banach Space Theory. London Mathematical Society Student Texts, vol. 8. Cambridge University Press, Cambridge (1987)
38. Johnson, W.B., Schechtman, G.: Finite dimensional subspaces of L_p. In: Johnson, W.B., Lindenstrauss, J. (eds.) Handbook of the Geometry of Banach Spaces, pp. 837–870. North-Holland, Amsterdam (2001)
39. Koksma, J.F.: Een algemeene stelling uit de theorie der gelijkmatige verdeeling modulo 1. Mathematica B **11**, 7–11 (1943)
40. Markhasin, L.: Discrepancy and integration in function spaces with dominating mixed smoothness. Dissertation, Friedrich-Schiller-University Jena (2012)
41. Markhasin, L.: Discrepancy of generalized Hammersley type point sets in Besov spaces with dominating mixed smoothness. Unif. Distrib. Theory **8**, 135–164 (2013)
42. Markhasin, L.: Quasi-Monte Carlo methods for integration of functions with dominating mixed smoothness in arbitrary dimension. J. Complexity **29**, 370–388 (2013)
43. Matoušek, J.: Geometric Discrepancy. Springer, Berlin (1999)
44. Maurey, B.: Théorèmes de factorisation pour les opérateurs linéaires à valeurs dans les espaces L^p, Astérisque, No. 11, Société Mathématique de France, Paris (1974)
45. Niederreiter, H.: Low-discrepancy and low-dispersion sequences. J. Number Theory **30**, 51–70 (1988)
46. Novak, E.: Deterministic and Stochastic Error Bounds in Numerical Analysis. LNiM, vol. 1349. Springer, Berlin (1988)
47. Novak, E., Woźniakowski, H.: Tractability of Multivariate Problems. Volume I: Linear Information. European Mathematical Society Publishing House, Zürich (2008)
48. Novak, E., Woźniakowski, H.: Tractability of Multivariate Problems. Volume II: Standard Information for Functionals. European Mathematical Society Publishing House, Zürich (2010)
49. Novak, E., Woźniakowski, H.: Lower bounds on the complexity for linear functionals in the randomized setting. J. Complexity **27**, 1–22 (2011)
50. Novak, E., Woźniakowski, H.: Tractability of Multivariate Problems. Volume III: Standard Information for Operators. European Mathematical Society Publishing House, Zürich (2012)
51. Pietsch, A.: Operator Ideals. North-Holland, Amsterdam/New York (1980)
52. Plaskota, L., Wasilkowski, G.W., Zhao, Y.: New averaging technique for approximating weighted integrals. J. Complexity **25**, 268–291 (2009)

53. Rosenthal, H.: On subspaces of L_p. Ann. of Math. **97**, 344–373 (1973)
54. Roth, K.F.: On irregularities of distribution. Mathematika **1**, 73–79 (1954)
55. Roth, K.F.: On irregularities of distribution II. Commun. Pure Appl. Math. **29**, 739–744 (1976)
56. Roth, K.F.: On irregularities of distribution III. Acta Arith. **35**, 373–384 (1979)
57. Roth, K.F.: On irregularities of distribution IV. Acta Arith. **37**, 67–75 (1980)
58. Schmidt, W.M.: Irregularities of distribution VII. Acta Arith. **21**, 45–50 (1972)
59. Schmidt, W.M.: Irregularities of distribution X. In: Zassenhaus, H. (ed.) Number Theory and Algebra, pp. 311–329. Academic, New York (1977)
60. Skriganov, M.M.: Harmonic analysis on totally disconnected groups and irregularities of point distributions. J. Reine Angew. Math. **600**, 25–49 (2006)
61. Sloan, I.H., Woźniakowski, H.: Tractability of integration in non-periodic and periodic weighted tensor product Hilbert spaces. J. Complexity **18**, 479–499 (2004)
62. Sloan, I.H., Woźniakowski, H.: When does Monte Carlo depend polynomially on the number of variables? In: Niederreiter, H. (ed.) Monte Carlo and Quasi-Monte Carlo Methods 2002, pp. 407–437. Springer, Berlin/Heidelberg (2004)
63. Sobol', I.M.: Distribution of points in a cube and approximate evaluation of integrals. Ž. Vyčisl. Mat. i Mat. Fiz. **7**, 784–802 (1967)
64. Steinerberger, S.: The asymptotic behavior of the average L_p-discrepancies and a randomized discrepancy. Electron. J. Combin. **17**, 18 (2010). Research Paper 106
65. Tomczak-Jaegermann, N.: Banach-Mazur Distances and Finite-Dimensional Operator Ideals. Pitman Monographs and Surveys in Pure and Applied Mathematics, vol. 38. Longman Scientific & Technical, Harlow (1989)
66. Traub, J.F., Wasilkowski, G.W., Woźniakowski, H.: Information-Based Complexity. Academic, Boston (1988)
67. Triebel, H.: Bases in Function Spaces, Sampling, Discrepancy, Numerical Integration. European Mathematical Society Publishing House, Zürich (2010)
68. Triebel, H.: Numerical integration and discrepancy, a new approach. Math. Nachr. **283**, 139–159 (2010)
69. Wasilkowski, G.W.: On polynomial-time property for a class of randomized quadratures. J. Complexity **20**, 624–637 (2004)
70. Zaremba, S.K.: Some applications of multidimensional integration by parts. Ann. Pol. Math. **21**, 85–96 (1968)

Noisy Information: Optimality, Complexity, Tractability

Leszek Plaskota

Abstract In this paper, we present selected old and new results on the optimal solution of linear problems based on noisy information, where the noise is bounded or random. This is done in the framework of *information-based complexity (IBC)*, and the main focus is on the following questions:

(i) What is an optimal algorithm for given noisy information?
(ii) What is the ε-complexity of a problem with noisy information?
(iii) When is a multivariate problem with noisy information tractable?

The answers are given for the worst case, average case, and randomized (Monte Carlo) settings. For (ii) and (iii) we present a computational model in which the cost of information depends on the noise level. For instance, for integrating a function $f : D \to \mathbb{R}$, available information may be given as

$$y_j = f(\mathrm{t}_j) + x_j, \qquad 1 \le j \le n,$$

with $x_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_j^2)$. For this information one pays $\sum_{j=1}^{n} c(\sigma_j)$ where $c : [0, \infty) \to [0, \infty]$ is a given *cost function*. We will see how the complexity and tractability of linear multivariate problems depend on the cost function, and compare the obtained results with noiseless case, in which $c \equiv 1$.

L. Plaskota (✉)
Faculty of Informatics, Mathematics, and Mechanics, University of Warsaw, ul. Banacha 2, 02-097 Warsaw, Poland
e-mail: L.Plaskota@mimuw.edu.pl

J. Dick et al. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2012*, Springer Proceedings in Mathematics & Statistics 65,
DOI 10.1007/978-3-642-41095-6_7, © Springer-Verlag Berlin Heidelberg 2013

1 Introduction

The purpose of this paper is to survey selected old and present some recent unpublished results on the optimal solution of linear problems for which available information is *partial*, *priced*, and contaminated by *noise*. The noise can be bounded in a norm, or can be a Gaussian random variable. Examples of such problems include multivariate integration or approximation of a function $f : [0,1]^d \to \mathbb{R}$, where available information is standard, i.e.,

$$y_j = f(\mathrm{t}_j) + x_j, \qquad 1 \le j \le n,$$

and the noise $|x_j| \le \delta$, or $x_j \sim \mathcal{N}(0, \sigma^2)$ with x_j mutually independent. The analysis is done in the framework of *information-based complexity (IBC)* [17, 29, 34], in the worst case, average case, and randomized (*Monte Carlo*) settings. We concentrate on the three main questions.

Optimality: What is an optimal algorithm for given noisy information?

We focus on the existence of optimal algorithms that are linear or affine, and give sufficient conditions for this to hold in different settings. It turns out that for approximating functionals from information with Gaussian noise, or noise bounded in a Hilbert norm, all the settings are (in a way) equivalent. In particular, optimal algorithms for bounded noise of level δ are optimal for Gaussian noise of some level σ, and vice-versa. These results are possible due to the existence of one-dimensional subproblems (for the worst and average case approximations and for bounded and Gaussian noise) that are as difficult as the original problem. This property was first observed and used in [3].

For approximating operators in Hilbert spaces, optimal algorithms are smoothing splines with parameter appropriately adjusted to the noise level.

Complexity: What is the ε-complexity of a problem with noisy information?

By ε-complexity of a problem, we mean the minimal cost of information from which it is possible to construct an approximation with error ε. We assume that the information cost depends on the noise level via a nonincreasing cost function c. For instance, the cost of a single observation of a functional with variance σ^2 may be given as $c(\sigma) = (1 + \sigma^{-1})^s$ for some $s \ge 0$. Observe that information gets cheaper as s decreases, so that the problem gets easier; moreover, $s = 0$ corresponds to exact information, in which $c \equiv 1$. Here, general results reveal an interesting property that the worst case setting with Gaussian noise is equivalent to the randomized setting with Gaussian noise. The reason is that the use of adaption for noisy information allows us to mimic any randomized algorithm.

As an illustration of this phenomenon, we study the worst case complexity of numerical integration in the Hölder function classes $C^r_\alpha([0,1]^d)$ for standard information, with noise either bounded or Gaussian. To give an example of our results, suppose that the noise is Gaussian and the dimension d is large enough,

$d > 2(r + \alpha)$. Then to obtain an ε-approximation using nonadaptively chosen sample points, one has to pay at least $\varepsilon^{-\left(\frac{d}{r+\alpha}\right)}$. Since this is the same as for exact information, one may think that it is impossible to do better. However, thanks to adaption, for $s > 0$ the ε-complexity of this problem is $\varepsilon^{-\left(\frac{d}{r+\alpha+d/2}\left(1-\frac{\bar{s}}{2}\right)+\bar{s}\right)}$ where $\bar{s} = \min(2, s)$. Hence the exponent of ε^{-1} is never larger than 2. We add that similar results can be obtained for integration in Sobolev classes of functions, but we do not pursue this subject here.

We also study complexity in the average case setting with Gaussian noise. Such problems are never easier than problems with exact information, and adaption usually does not help. We illustrate this by average case approximation with respect to a Gaussian prior, using unrestricted linear information with Gaussian noise. In this case, the ε-complexity basically behaves in two ways: as for exact information, or it is proportional to ε^{-2}.

Tractability: When is a multivariate problem with noisy information tractable?

Tractability of multivariate problems with exact information is already a well established area, as seen from the resent three volume monograph of [17–19]. However, the study of tractability of problems with noisy information is still in its initial stage. We give sample results on polynomial tractability of the two problems considered earlier. The results show that the polynomial tractability for exact information is equivalent to polynomial tractability for noisy information. Moreover, if the problem for exact information is sufficiently difficult then the exponents of tractability from exact information carry over to noisy information.

The paper is organized as follows. In the preliminary Sect. 2 we introduce the basic notions. Optimal algorithms for given information are analyzed in Sect. 3. In Sect. 4 we deal with complexity, and Sect. 5 is devoted to tractability.

2 Basics

Let

$$S : F \to G$$

be a linear operator acting between a linear space F and a normed space G with the norm $\|\cdot\|$. We usually think of F as a space of (multivariate) functions $f : D \to \mathbb{R}$ where D is a Lebesgue measurable subset of $\mathbb{R}^d$. The two prominent examples are

- *Integration:* $S = \mathrm{Int}$, where $G = \mathbb{R}$ and

$$\mathrm{Int}(f) = \int_D f(\mathrm{t})\,\mathrm{dt},$$

- *Approximation:* $S = \mathrm{App}$, where F is embedded into G, and

$$\mathrm{App}(f) = f.$$

The aim is to approximate values $S(f)$ for $f \in F$. The approximation is produced by an algorithm that uses some information about f. The information is *partial* and *noisy*.

2.1 Noisy Information

Typical examples of noisy information about a function f include inaccurate observations/evaluations of f at some points $\mathrm{t}_j \in D$, i.e.,

$$y_j = f(\mathrm{t}_j) + x_j,$$

or, more generally, inaccurate observations/evaluations of some linear functionals L_j on F,

$$y_j = L_j(f) + x_j.$$

Here, x_j is the *noise*. The noise can be uniformly bounded, e.g., $|x_j| \le \delta$, or it can be a random variable, e.g., $x_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$. We now give a formal definition of noisy information.

2.1.1 Information with Bounded Noise

An *information operator* is a mapping

$$\mathbb{N} : F \to 2^Y$$

where Y is a set of finite real sequences $Y \subset \sum_{n=0}^{\infty} \mathbb{R}^n$. That is, $\mathbb{N}(f)$ is a subset of Y. A sequence $\mathrm{y} \in Y$ is called *information about* f iff $\mathrm{y} \in \mathbb{N}(f)$.

This general definition admits *linear information with uniformly bounded noise*. In this case, $Y = \mathbb{R}^n$ and

$$\mathbb{N}(f) = \{N(f) + \mathrm{x} : \ \mathrm{x} \in X\}$$

where $N : F \to Y$ is a linear mapping and X is a bounded subset of Y. The vector $N(f)$ is called exact information about f, and x is the noise. If $X = \{0\}$ then information is said to be exact. Examples include

$$X = \{\mathrm{x} \in Y : \ |x_j| \le \delta_j : \ 1 \le j \le n\}$$

with $\delta_j \geq 0$, or

$$X = \{\mathrm{x} \in Y : \|\mathrm{x}\|_Y \leq \delta\}$$

with a Hilbert norm $\|\mathrm{x}\|_Y = \sqrt{\langle \Sigma^{-1}\mathrm{x}, \mathrm{x}\rangle_2}$ and a symmetric positive definite operator (matrix) $\Sigma = \Sigma^* > 0$. The parameter $\delta \geq 0$ is the *noise level*. The information is exact iff $\delta = 0$.

2.1.2 Information with Gaussian Noise

In this case, the information operator assigns to each f a probability measure π_f on the set Y. That is, Y is such that the sets $Y \cap \mathbb{R}^i$ are measurable for all i, and

$$\mathbb{N} : F \to \mathscr{P}_Y$$

where $\mathscr{P}_Y$ is the set of all probability measures on Y. A sequence $\mathrm{y} \in Y$ is *information about* f iff it is realization of a random variable distributed according to $\pi_f = \mathbb{N}(f)$.

An important example is given by *linear information with Gaussian noise*. In this case, $Y = \mathbb{R}^n$ and $\mathrm{y} = N(f) + \mathrm{x}$, where

$$\mathrm{x} \sim \mathscr{G}_n(0, \sigma^2 \Sigma),$$

$N : F \to \mathbb{R}^n$ is a linear mapping, and $\mathscr{G}_n(0, \sigma^2 \Sigma)$ is the zero mean n-dimensional Gaussian distribution on $\mathbb{R}^n$ whose covariance matrix is $\sigma^2 \Sigma$. For the noise level $\sigma = 0$, the information is exact (with probability one).

2.2 Algorithms and Errors

Having defined the information $\mathbb{N}$, an *algorithm* is now a mapping

$$\varphi : Y \to G.$$

That is, an approximation to $S(f)$ is given as $\varphi(\mathrm{y})$ where y is information about f.

The error of an algorithm φ using information $\mathbb{N}$ depends on the *setting*. The setting is determined by the assumptions on the elements $f \in F$ and information y about f.

2.2.1 Worst Case Setting

In this setting, the error of an algorithm φ using information $\mathbb{N}$ is defined as the worst case error with respect to a given set $B \subseteq F$. More specifically,

- For bounded noise:

$$e_B^{\text{w–w}}(S;\mathbb{N},\varphi) := \sup_{f\in B} \sup_{\mathrm{y}\in\mathbb{N}(f)} \|S(f)-\varphi(\mathrm{y})\|,$$

- For Gaussian noise:

$$e_B^{\text{w–a}}(S;\mathbb{N},\varphi) := \sup_{f\in B} \left(\int_Y \|S(f)-\varphi(\mathrm{y})\|^2 \pi_f(\mathrm{dy})\right)^{1/2}.$$

The set B is usually chosen such that the elements $f \in B$ are 'smooth' and bounded, e.g., some derivatives of f are uniformly bounded. If F is a normed space with norm $\|\cdot\|_F$ then we often take B to be the unit ball of F,

$$B = \{f \in F : \|f\|_F \le 1\}.$$

2.2.2 Average Case Setting

In this setting the error is defined as the average error with respect to a given probability measure μ on the space F. Specifically,

- For Gaussian noise:

$$e_\mu^{\text{a–a}}(S;\mathbb{N},\varphi) := \left(\int_F \int_Y \|S(f)-\varphi(\mathrm{y})\|^2 \pi_f(\mathrm{dy})\mu(\mathrm{d}f)\right)^{1/2},$$

- For bounded noise:

$$e_\mu^{\text{a–w}}(S;\mathbb{N},\varphi) := \left(\int_F \sup_{\mathrm{y}\in\mathbb{N}(f)} \|S(f)-\varphi(\mathrm{y})\|^2 \mu(\mathrm{d}f)\right)^{1/2}.$$

We will assume that F is a separable Banach space and the a priori measure μ is a Gaussian measure on the Borel subsets of F. This measure has mean zero and a covariance operator $C_\mu : F^* \to F$, i.e.,

$$L_1(C_\mu L_2) = \int_F L_1(f)L_2(f)\,\mu(\mathrm{d}f), \qquad \forall\, L_1, L_2 \in F^*.$$

See, e.g., [11] or [35] for the theory of Gaussian measures on Banach spaces.

2.2.3 Randomized Setting

In contrast to the deterministic settings considered above, in the randomized (or *Monte Carlo*) settings the information and/or algorithms are chosen at random.

We may have different randomized settings, depending on the assumptions on the problem elements f and information y about f. In this paper, we only concentrate on the *worst case randomized setting with Gaussian noise*. In this setting, the problem elements f belong to B, and the information y has a distribution π_f. For brevity, we will use the name *randomized setting*.

Formally, the information operator and algorithm are now families $\{\mathbb{N}_\omega\}$ and $\{\varphi_\omega\}$ parameterized by a random variable $\omega \in \Omega$. For technical reasons, we assume that Ω is a Polish space or the Cartesian product of countably many Polish spaces. The elements $S(f)$ are approximated by $\varphi_\omega(\mathrm{y})$ where the information y about f is given as

$$\mathrm{y} \sim \pi_f(\cdot|\omega) = \mathbb{N}_\omega(f)$$

and ω is drawn randomly from Ω according to some distribution. The error in the randomized setting is defined as

$$e_B^{\mathrm{ran}}(S; \{\mathbb{N}_\omega\}, \{\varphi_\omega\}) := \sup_{f \in B} \left(\mathbb{E}_\omega \int_{Y_\omega} \|S(f) - \varphi_\omega(\mathrm{y})\|^2 \, \pi_f(\mathrm{dy}|\omega) \right)^{1/2}$$

where $\mathbb{E}_\omega$ denotes the expectation with respect to ω.

Observe that in the randomized setting we have two sources of randomness, one coming from random selection of ω, and the other coming from the noise.

3 Optimality

In this section, we give some general results on optimal algorithms for given noisy information $\mathbb{N}$. An algorithm is optimal in a given setting iff it minimizes the error among all possible algorithms using $\mathbb{N}$.

3.1 Approximation of Functionals

We first consider the case when S is a linear functional, e.g., $S = \mathrm{Int}$. We concentrate our attention on the existence of optimal algorithms that are affine or linear. Note that widely used cubatures for the integration problem are linear algorithms.

3.1.1 Worst Case Setting

Assume information is linear with uniformly bounded noise,

$$\mathrm{y} = N(f) + \mathrm{x}, \qquad \mathrm{x} \in X. \tag{1}$$

Let $\mathrm{rad}_B^{\text{w–w}}(S;\mathbb{N})$, called the *radius of information*, be the minimal error that can be achieved using information $\mathbb{N}$. That is,

$$\mathrm{rad}_B^{\text{w–w}}(S;\mathbb{N}) = \inf_{\varphi} e_B^{\text{w–w}}(S;\mathbb{N},\varphi) \tag{2}$$

where the infimum is taken over all possible algorithms φ.

Theorem 1. *If $B \subset F$ and $X \subset Y$ are convex sets then there exists an affine algorithm $\varphi_{\text{aff}}^{\text{w–w}}$ that is optimal, i.e., such that*

$$e_B^{\text{w–w}}(S;\mathbb{N},\varphi_{\text{aff}}^{\text{w–w}}) = \mathrm{rad}_B^{\text{w–w}}(S;\mathbb{N}).$$

In addition, if B and X are balanced (symmetric about zero) then any optimal $\varphi_{\text{aff}}^{\text{w–w}}$ is linear.

Theorem 1 was first proven in [31] (see also [1]) for exact information and convex and balanced B, and then generalized in [14] for noisy information, and in [32] for exact information and convex B. The most general formulation was presented in [12].

The proof of Theorem 1 is non-constructive. However, in some cases, it is possible to construct optimal $\varphi_{\text{aff}}^{\text{w–w}}$. Assume that the noise x of information (1) belongs to the ball of radius $\delta > 0$ in a Hilbert norm, i.e., that

$$X = \{\mathrm{x} : \|\mathrm{x}\|_Y \le \delta\}$$

where $\|\mathrm{x}\|_Y = \sqrt{\langle \mathrm{x},\mathrm{x}\rangle_Y}$ and $\langle\cdot,\cdot\rangle_Y$ is an inner product in Y. Then we have the following elegant construction from [3], see also Sect. 2.4 in [29].

Let $r(\delta)$ be the radius (2) of our information with noise level δ, which can be conveniently written as

$$r(\delta) = \sup\{\, S(h) : \ h \in \mathrm{bal}(B),\ \|N(h)\|_Y \le \delta \,\} \tag{3}$$

where

$$\mathrm{bal}(B) := \tfrac{1}{2}(B - B) = \left\{ \tfrac{1}{2}(b_1 - b_2) : \ b_1, b_2 \in B \right\}.$$

The radius $r(\gamma)$ is a concave and nondecreasing function of $\gamma \ge 0$. Hence, it can be bounded from above by a straight line passing through $(\delta, r(\delta))$. In other words, there exists $d \ge 0$ such that

$$r(\gamma) \le r(\delta) + d(\gamma - \delta) \qquad \forall \gamma \ge 0.$$

Denote by $\partial r(\delta)$ the set of all such d. Suppose that the radius is achieved at h^*, i.e., $r(\delta) = S(h^*)$ where $h^* = (f_1^* - f_{-1}^*)/2$, $f_{-1}^*, f_1^* \in B$, $\|N(h^*)\|_Y \le \delta$. From the formula (2) it follows that our original problem is as difficult as the same problem, but with B replaced by the interval $I^* = [f_{-1}^*, f_1^*] \subset B$. That is,

$$\mathrm{rad}_B^{\text{w–w}}(S;\mathbb{N}) = \mathrm{rad}_{I^*}^{\text{w–w}}(S;\mathbb{N}).$$

We say that I^* is the *hardest one-dimensional subproblem* contained in B. To find all optimal affine algorithms it is enough to first find all optimal affine algorithms for the subproblem I^* and then remove those algorithms for which the error increases when taken over B. This leads to the following formulas. If $N(h^*) = 0$ then $\varphi_{\mathrm{aff}}^{\text{w–w}} \equiv 0$, and otherwise

$$\varphi_{\mathrm{aff}}^{\text{w–w}}(\mathrm{y}) = S(f_0^*) + d\,\langle \mathrm{y} - N(f_0^*), \mathbf{w}\rangle_Y, \tag{4}$$

where $f_0^* = (f_{-1}^* + f_1^*)/2$, $\mathbf{w} = N(h^*)/\|N(h^*)\|_Y$, and $d \in \partial r(\delta)$.

This construction cannot be applied for $\delta = 0$. However, optimal algorithms for exact information can be obtained by letting $\delta \to 0^+$.

We now switch to linear information with Gaussian noise, so that

$$\mathrm{x} \sim \mathscr{G}_n(0, \sigma^2\Sigma), \qquad \text{where} \quad \sigma > 0. \tag{5}$$

To see that optimal algorithms are *not* affine in this case, it is enough to have a short look at the one-dimensional problem of approximating f from an interval $B = [-\tau, \tau] \subset \mathbb{R}$ based on just one observation $y = f + x$ with noise $x \sim \mathscr{N}(0, \sigma^2)$. Indeed, the optimal affine algorithm is linear, given by $\varphi_{\mathrm{lin}}(y) = c_1 y$ with

$$c_1 = c_1(\tau, \sigma^2) = \frac{\tau^2}{\tau^2 + \sigma^2}.$$

For $|y| > \tau/c_1$ we have $\varphi_{\mathrm{lin}}(y) \notin B$, and therefore φ_{lin} cannot be optimal among arbitrary algorithms.

How much do we lose by using only linear approximations? This question was studied, e.g., in [6, 9], but the most precise answer was given in L.D. Brown and I. Feldman (The Minimax Risk for Estimating a Bounded Normal Mean, 1990, unpublished manuscript). Let $r_{\mathrm{arb}}(\tau, \sigma^2)$ and $r_{\mathrm{lin}}(\tau, \sigma^2)$ be the minimal errors for arbitrary nonlinear and linear algorithms. Then

$$\kappa_1 := \sup_{\tau,\sigma} \frac{r_{\mathrm{arb}}(\tau, \sigma^2)}{r_{\mathrm{lin}}(\tau, \sigma^2)} \;<\; 1.12\,.$$

Surprisingly, this result for the one-dimensional problem can be generalized to the approximation of arbitrary linear functionals S. Let $\mathrm{rad}_B^{\text{w–a}}(S;\mathbb{N})$ denote the radius of information in case of Gaussian noise (5). By an *optimal affine algorithm* we mean an algorithm that minimizes the error over all affine algorithms.

Theorem 2. *If B is a convex set then for the optimal affine algorithm $\varphi_{\mathrm{aff}}^{\text{w–a}}$ we have*

$$e_B^{\text{w–a}}(S;\mathbb{N}, \varphi_{\mathrm{aff}}^{\text{w–a}}) \le \kappa_1 \cdot \mathrm{rad}_B^{\text{w–a}}(S;\mathbb{N}).$$

In addition, if B is balanced then any such $\varphi_{\mathrm{aff}}^{\mathrm{w\text{–}a}}$ is linear.

The proof of Theorem 2 was first given in [3], see also Sect. 4.2 in [29]. It reveals a remarkable connection between Gaussian noise (5) and noise bounded in the Hilbert norm

$$\|\mathrm{x}\|_Y = \sqrt{\langle \Sigma^{-1}\mathrm{x}, \mathrm{x}\rangle_2} \le \delta. \tag{6}$$

This connection can be explained as follows. Let $\varphi_\delta^{\mathrm{w\text{–}w}}$ denote an optimal affine algorithm (4) in the worst case setting with bounded noise (6). Let $\delta = \delta(\sigma)$ be such that $d_\delta \in \partial r(\delta)$ and

$$d_\delta = \frac{\delta\, r(\delta)}{\sigma^2 + \delta^2}. \tag{7}$$

(If the equality (7) never holds, in which case $d_\infty = \lim_{\delta\to\infty} d_\delta$, we let $\delta = +\infty$.) Then $\varphi_\sigma^{\mathrm{w\text{–}a}} = \varphi_\delta^{\mathrm{w\text{–}w}}$ is an optimal affine algorithm for Gaussian noise of level σ and

$$e_B^{\mathrm{w\text{–}a}}(S;\mathbb{N},\varphi_\sigma^{\mathrm{w\text{–}a}}) = \frac{\sigma\, r(\delta)}{\sqrt{\sigma^2+\delta^2}}.$$

(If $\delta = \delta(\sigma) = +\infty$ then $e_B^{\mathrm{w\text{–}a}} = \sigma d_\infty$.)

Theorem 1 now follows from the following property: if we restrict ourselves only to affine algorithms then the hardest one-dimensional subproblem for the worst case with bounded noise of level $\delta = \delta(\sigma)$ is also the hardest one-dimensional subproblem for the worst case with Gaussian noise of level σ.

From what we already know, it is possible to infer the following fact about randomization. Suppose that information $\mathbb{N}$ with Gaussian noise (5) is given. Then randomization with respect to algorithms does not help much. Indeed, let I^* be the hardest one-dimensional subproblem for affine algorithms and information $\mathbb{N}$. Then there exists a least favorable probability distribution μ on I^* for which the minimal average case error with respect to μ is equal to the minimal worst case error with respect to I^*, see, e.g., [2]. Using this fact, along with the mean value theorem, we see that for any randomized algorithm $\{\varphi_\omega\}$ we have

$$\begin{aligned}
\left(e_B^{\mathrm{ran}}(S;\mathbb{N},\{\varphi_\omega\})\right)^2 &= \sup_{f\in B}\left(\mathbb{E}_\omega \int_Y |S(f)-\varphi_\omega(\mathrm{y})|^2 \pi_f(\mathrm{dy})\right)\\
&\ge \sup_{f\in I^*}\left(\mathbb{E}_\omega \int_Y |S(f)-\varphi_\omega(y)|^2 \pi_f(\mathrm{dy})\right)\\
&\ge \int_{I^*}\left(\mathbb{E}_\omega \int_Y |S(f)-\varphi_\omega(\mathrm{y})|^2 \pi_f(\mathrm{dy})\right)\mu(\mathrm{d}f)\\
&= \mathbb{E}_\omega\left(\int_{I^*}\int_Y |S(f)-\varphi_\omega(\mathrm{y})|^2 \pi_f(\mathrm{dy})\mu(\mathrm{d}f)\right)
\end{aligned}$$

$$\geq \int_{I^*} \int_Y |S(f) - \varphi_{\omega^*}(\mathrm{y})|^2 \pi_f(\mathrm{dy}) \mu(\mathrm{d}f)$$
$$\geq \left(\mathrm{rad}_{I^*}^{\text{w–a}}(S;\mathbb{N})\right)^2$$
$$\geq \kappa_1^{-2} \cdot \sup_{f \in B} \int_Y |S(f) - \varphi_{\mathrm{aff}}^{\text{w–a}}(\mathrm{y})|^2 \pi_f(\mathrm{dy})$$
$$\geq \kappa_1^{-2} \cdot \left(\mathrm{rad}_B^{\text{w–a}}(S;\mathbb{N})\right)^2.$$

3.1.2 Average Case Setting

We will see that parallel results to that in the worst case hold in the average case setting. Recall that F is now a separable Banach space and the a priori measure μ on the Borel sets of F is Gaussian with mean zero and a covariance operator $C_\mu : F^* \to F$. We also assume that the functional S is continuous, $S \in F^*$. The noisy information is $\mathrm{y} = N(f) + \mathrm{x}$, where

$$N(f) = [L_1(f), L_2(f), \ldots, L_n(f)], \qquad L_j \in F^*.$$

Let $\langle K, L\rangle_\mu = K(C_\mu L) = L(C_\mu K)$ denote the μ-semi-inner product on F^*, with $\|L\|_\mu = \sqrt{L(C_\mu L)}$ denoting the corresponding seminorm.

Consider first Gaussian noise $\mathrm{x} \sim \mathcal{G}_n(0, \sigma^2 \Sigma)$ with $\sigma > 0$. Then for any continuous S (not necessarily a functional), optimal algorithms are linear and rely on applying S to the conditional mean with respect to given information. For a functional S these algorithms take the following form, see Sect. 3.5 in [29].

Theorem 3. *The optimal algorithm is unique and is given as*

$$\varphi_{\mathrm{lin}}^{\text{a–a}}(\mathrm{y}) = \langle \mathrm{y}, \mathbf{w}\rangle_2,$$

where $\mathbf{w}$ *is the solution of the linear system*

$$(\sigma^2 \Sigma + G_N)\mathbf{w} = N(C_\mu S)$$

with matrix $G_N = \left[L_i(C_\mu L_j)\right]_{i,j=1}^n$. *Moreover,*

$$e_\mu^{\text{a–a}}(S;\mathbb{N},\varphi_{\mathrm{lin}}^{\text{a–a}}) = \mathrm{rad}_\mu^{\text{a–a}}(S;\mathbb{N}) = \sqrt{\|S\|_\mu^2 - \langle N(C_\mu S), \mathbf{w}\rangle_2}.$$

Similarly to the worst case, the concept of a one-dimensional subproblem is also important in the analysis of the average case. For a functional $K \in F^*$, let

$$P_K(f) = f - \frac{K(f)}{\|K\|_\mu^2} C_\mu K$$

be a projection onto $\ker K$. Then $\ker P_K = \operatorname{span}\{C_\mu K\}$. The measure μ can be then decomposed as

$$\mu = \int_{\ker K} \mu_K(\cdot|g)\, \mu P_K^{-1}(\mathrm{d}g)$$

where $\mu_K(\cdot|g)$ is the conditional measure on F given $g = P_K(f)$. The measure $\mu_K(\cdot|g)$ is Gaussian with mean g and correlation operator

$$A_K(L) = \frac{\langle L, K\rangle_\mu}{\|K\|_\mu^2} C_\mu K,$$

and it is concentrated on the line

$$P_K^{-1} g = \{g + \alpha C_\mu K : \alpha \in \mathbb{R}\}. \tag{8}$$

The family of one-dimensional subproblems induced by K is indexed by $g \in \ker K$. For each g, the subproblem relies on minimizing the average error $e^{\text{a–a}}_{\mu_K(\cdot|g)}(\mathbb{N}, \varphi)$ over all algorithms φ. (Equivalently, the subproblem is the original problem with additional information that f is on the line (8).)

It is clear that $\mathrm{rad}^{\text{a–a}}_{\mu_K(\cdot|g)}(S;\mathbb{N}) \le \mathrm{rad}^{\text{a–a}}_{\mu}(S;\mathbb{N})$. However, if

$$K_\sigma = S - \varphi^{\text{a–a}}_{\text{lin}}(N(\cdot)) = S - \langle N(\cdot), \mathbf{w}\rangle_2$$

then $\mu_{K_\sigma}(\cdot|g)$ is the hardest family of one-dimensional subproblems. For all g we have

$$\mathrm{rad}^{\text{a–a}}_{\mu_{K_\sigma}(\cdot|g)}(S;\mathbb{N}) = \mathrm{rad}^{\text{a–a}}_{\mu}(S;\mathbb{N}).$$

Moreover, the same linear algorithm $\varphi^{\text{a–a}}_{\text{lin}}$ is optimal for the original problem μ and for each subproblem $\mu_{K_\sigma}(\cdot|g)$.

Consider noise that is bounded in a Hilbert norm. For the one-dimensional problem of approximating $f \sim \mathcal{N}(0,\lambda)$ from one observation $y = f + x$ with $|x| \le \delta$, the optimal linear algorithm is $\varphi_{\text{lin}}(y) = c_2 y$ with

$$c_2 = c_2(\lambda,\delta) = \begin{cases} 1, & \delta^2 \le \frac{2}{\pi}\lambda, \\ \frac{\lambda - \delta\sqrt{2\lambda/\pi}}{\lambda + \delta^2 - 2\delta\sqrt{2\lambda/\pi}}, & \frac{2}{\pi}\lambda < \delta^2 < \frac{\pi}{2}\lambda, \\ 0, & \frac{\pi}{2}\lambda \le \delta^2. \end{cases} \tag{9}$$

Let κ_2 denote the maximum with respect to λ and δ of the ratio between the minimal errors of linear and nonlinear algorithms. Then

$$\kappa_2 < 1.5\,.$$

This generalizes to arbitrary functionals S. Let the noise $\|\mathrm{x}\|_Y = \sqrt{\langle \Sigma^{-1}\mathrm{x}, \mathrm{x}\rangle_2}$ satisfy $\|x\|_Y \le \delta$ for some $\delta > 0$.

Theorem 4. *The optimal linear algorithm $\varphi_{\mathrm{lin}}^{\mathrm{a\text{–}w}}$ satisfies*

$$e_\mu^{\mathrm{a\text{–}w}}(S; \mathbb{N}, \varphi_{\mathrm{lin}}^{\mathrm{a\text{–}w}}) \le \kappa_2 \cdot \mathrm{rad}_\mu^{\mathrm{a\text{–}w}}(S; \mathbb{N}).$$

The algorithm $\varphi_{\mathrm{lin}}^{\mathrm{a\text{–}w}}$ is given as follows. Let $\varphi_\sigma^{\mathrm{a\text{–}a}}$ be an optimal algorithm from Theorem 3 for the noise level σ. Let

$$\rho(\sigma) = \frac{\|K_\sigma\|_\mu}{\|N(C_\mu K_\sigma)\|_Y}.$$

If $\delta\|S\|_\mu \ge \sqrt{\pi/2}\|N(C_\mu S)\|_Y$ then the zero algorithm is optimal. Otherwise, the optimal linear algorithm $\varphi_\delta^{\mathrm{a\text{–}w}} = \varphi_\sigma^{\mathrm{a\text{–}a}}$ where $\sigma = \sigma(\delta)$ is such that

$$c_2 = c_2(1, \delta\rho(\sigma)) = \frac{1}{1 + \sigma^2\rho^2(\sigma)}.$$

The assertion of Theorem 4 now follows from the fact that if we restrict ourselves to only linear algorithms then the family determined by K_σ is the hardest family of one-dimensional subproblems for the average case with bounded noise of level δ. For details, see [25] and Sect. 4.2 in [29].

3.1.3 Equivalence of Different Settings

We showed close connections between Gaussian noise and noise bounded in a Hilbert norm, for both the worst case and average case settings. Actually, similar connections hold between the settings themselves.

Let F be a separable Banach space equipped with a zero mean Gaussian measure with positive definite covariance operator $C_\mu : F^* \to F$. Let $H_0 = C_\mu(F^*) \subseteq F$ be a pre-Hilbert space with inner product

$$\langle f_1, f_2\rangle_H = L_1(C_\mu L_2) = \int_F L_1(f)L_2(f)\mu(\mathrm{d}f)$$

for all $f_j = C_\mu L_j$, where $L_j \in F^*$, and let the Hilbert space H be the closure of H_0 with respect to $\langle\cdot,\cdot\rangle_H$. We let B denote the unit ball of H.

The pair (F, H) forms an *abstract Wiener space*. The name comes from the special case where F is the classical Wiener space of continuous functions $f : [0, 1] \to \mathbb{R}$ with $f(0) = 0$, norm $\|f\|_F = \max_{0\le t\le 1} |f(t)|$, and $H \subset F$ is the Hilbert space of absolutely continuous functions with the inner product

$$\langle f_1, f_2 \rangle_H = \int_0^1 f_1'(t)\, f_2'(t)\, \mathrm{d}t.$$

Equivalently, H is the reproducing kernel Hilbert space with the kernel $K(s,t) = \min(s,t)$, $0 \le s,t \le 1$. See, e.g., [11].

Consider now the problem of approximating a functional $S \in F^*$, so that

$$S(f) = \langle f, f^* \rangle_H, \qquad f^* = C_\mu S,$$

from noisy information $\mathrm{y} = N(f) + \mathrm{x}$, where

$$N(f) = [L_1(f), L_2(f), \ldots, L_n(f)]$$

with $L_j \in F^*$, so that

$$L_j(f) = \langle f, f_j \rangle_H, \qquad f_j = C_\mu L_j.$$

Suppose that for all four settings:

- Worst case with bounded noise: $f \in B$ and $\langle \Sigma^{-1}\mathrm{x}, \mathrm{x}\rangle_2 \le \delta^2$,
- Worst case with Gaussian noise: $f \in B$ and $\mathrm{x} \sim \mathscr{G}_n(0, \sigma^2\Sigma)$,
- Average case with Gaussian noise: $f \sim \mu$ and $\mathrm{x} \sim \mathscr{G}_n(0, \sigma^2\Sigma)$,
- Average case with bounded noise: $f \sim \mu$ and $\langle \Sigma^{-1}\mathrm{x}, \mathrm{x}\rangle_2 \le \delta^2$,

we apply the same algorithm $\varphi^* = \varphi_{\mathrm{lin}}^{\mathrm{a\text{–}a}}$. That is,

$$\varphi^*(\mathrm{y}) = \sum_{j=1}^n w_j\, y_j$$

with $(\sigma^2\Sigma + G_N)\mathrm{w} = N(f^*)$, $G_N = \left[\langle f_i, f_j\rangle_H\right]_{i,j=1}^n$. Then we have the following theorem, see Sect. 5.2.3 in [29].

Theorem 5. *Let $\delta = \sigma$. Then for $e^{\mathrm{sett}} \in \{e_B^{\mathrm{w\text{–}w}}, e_B^{\mathrm{w\text{–}a}}, e_\mu^{\mathrm{a\text{–}a}}, e_\mu^{\mathrm{a\text{–}w}}\}$ we have*

$$e^{\mathrm{sett}}(S; \mathbb{N}, \varphi^*) = \sqrt{S\left(f^* - \sum_{j=1}^n w_j\, f_j\right)} \le \kappa^{\mathrm{sett}} \cdot \inf_\varphi e^{\mathrm{sett}}(S; \mathbb{N}, \varphi)$$

where $(\kappa^{\mathrm{w\text{–}w}}, \kappa^{\mathrm{w\text{–}a}}, \kappa^{\mathrm{a\text{–}a}}, \kappa^{\mathrm{a\text{–}w}}) = (1.43,\ 1.59,\ 1.00,\ 2.13)$.

Observe that for 'small' noise levels σ the element $f^* - \sum_{j=1}^n w_j\, f_j$ is 'almost' H-orthogonal projection of f^* onto the $\mathrm{span}(f_1, \ldots, f_n)$.

3.2 Approximation of Operators

If S is a linear operator, but not a functional, then the results of Sect. 3.1 generally do not hold. It is even possible that the worst case radius of (exact) information is arbitrarily small, but the worst case error of any linear algorithm is infinite, see [38].

For problems such as function approximation, one commonly-used technique is based on smoothing splines. The general idea relies on constructing algorithms that are simultaneously smooth and adjusted to the given data. In some cases, smoothing splines lead to algorithms that are linear and (almost) optimal.

3.2.1 Worst Case Setting

Let F be equipped with a semi-norm $\|\cdot\|_F$ and $S : F \to G$ be an arbitrary linear operator. Let the set B be the unit ball in F and information be linear with uniformly bounded noise, so that $\mathbb{N}(f) = \{\mathrm{y} \in \mathbb{R}^n : \ \|\mathrm{y} - N(f)\|_Y \le \delta\}$.

For given information y, an *ordinary spline* $\mathrm{s}_o(\mathrm{y})$ is given as an element in F satisfying the following two conditions:

1. $\mathrm{y} \in \mathbb{N}(\mathrm{s}_o(\mathrm{y}))$.
2. $\|\mathrm{s}_o(\mathrm{y})\|_F = \inf\{\|f\|_F : \ f \in F,\ \mathrm{y} \in \mathbb{N}(f)\}$.

(For simplicity, we assume that the infimum is attained.) Then the spline algorithm is defined as $\varphi_o(\mathrm{y}) = S(\mathrm{s}_o(\mathrm{y}))$, and we have from [10] that

$$e_B^{\mathrm{w\text{–}w}}(S; \mathbb{N}, \varphi_o) \le 2 \cdot \mathrm{rad}_B^{\mathrm{w\text{–}w}}(S; \mathbb{N}),$$

where $\mathrm{rad}_B^{\mathrm{w\text{–}w}}(S; \mathbb{N})$ is the radius of information $\mathbb{N}$, as in (2). The algorithm φ_o is in general not linear.

We now pass to smoothing splines. We assume that F, Y, and G are Hilbert spaces, so that the norms $\|\cdot\|_F$, $\|\cdot\|_Y$, and $\|\cdot\|$ are induced by the corresponding inner products. We also assume that $S : F \to G$ and $N : F \to Y$ are continuous linear operators. For $0 < \gamma < \infty$, a *smoothing spline* is defined as an element $\mathrm{s}_\gamma(\mathrm{y}) \in F$ minimizing the functional

$$\gamma \cdot \|f\|_F^2 + \|\mathrm{y} - N(f)\|_Y^2 \qquad \text{over all} \quad f \in F.$$

The smoothing spline $\mathrm{s}_\gamma(\mathrm{y})$ is uniquely defined and it depends linearly on y. Equivalently, $\mathrm{s}_\gamma(\mathrm{y})$ can be defined as the result of *regularization* [33], i.e., as the solution of the linear equation

$$(\gamma\, I + N^* N)\, f = N^* \mathrm{y}$$

where $N^* : Y \to F$ is the operator adjoint to N, i.e., $\langle Nf, \mathrm{z}\rangle_Y = \langle f, N^*\mathrm{z}\rangle_F$ for all $f \in F$ and $\mathrm{z} \in Y$. We additionally set $\mathrm{s}_\infty(\mathrm{y}) = 0$, and

$$s_0(y) = \arg\min\{\|f\|_F : N(f) = P_N y\},$$

where $P_N y$ is the orthogonal projection of y onto the subspace $N(F) \subseteq Y$.

The *smoothing spline algorithm* is defined as

$$\varphi_\gamma(y) = S(s_\gamma(y)).$$

If the information is exact, i.e., $\delta = 0$, then φ_0 is optimal and

$$\mathrm{rad}_B^{\text{w-w}}(S; \mathbb{N}, \varphi_0) = \sup_{h \in B \cap \ker N} \|S(h)\|,$$

see, e.g., Sect. 5.7 in Chap. 4 of [34]. In the general case, taking $\gamma = \delta^2$ we obtain

$$e_B^{\text{w-w}}(S; \mathbb{N}, \varphi_{\delta^2}) \le \sqrt{2} \cdot \mathrm{rad}_B^{\text{w-w}}(S; \mathbb{N}).$$

However, the following stronger result holds, which was proven in [13].

Theorem 6. *If F, G, and Y are Hilbert spaces then there exists γ^* such that the smoothing spline algorithm φ_{γ^*} is optimal, i.e., such that*

$$e_B^{\text{w-w}}(S; \mathbb{N}, \varphi_{\gamma^*}) = \mathrm{rad}_B^{\text{w-w}}(S; \mathbb{N}).$$

The proof of Theorem 6 is not constructive. Nevertheless, the optimal value γ^* of the smoothing parameter can be found in some special cases.

Assume that $S : F \to G$ is a compact operator. By $S^* : G \to F$ and $N^* : Y \to F$ we mean the adjoint operators to S and N. Assume that S^*S and N^*N have a common complete orthonormal in F basis $\{\xi_j\}_{j \ge 1}$ of eigenelements, so that

$$S^*S\xi_j = \lambda_j \xi_j, \qquad N^*N\xi_j = \eta_j \xi_j, \qquad j \ge 1, \tag{10}$$

and $\lambda_1 \ge \lambda_2 \ge \cdots \ge 0$. (If $\dim(F) = d < \infty$ then we set $\lambda_j = \eta_j = 0$ for all $j \ge d + 1$.) Then we have the following result. Let

$$t = \min\{i : \eta_i = 0\} \qquad \text{and} \qquad s = \arg \max_{1 \le i \le t-1} \frac{\lambda_i - \lambda_t}{\eta_i}.$$

If the noise level is sufficiently small, $\delta^2 \le \min_{1 \le i \le t-1} \eta_i$, then

$$\gamma^* = \frac{\eta_s \lambda_t}{\lambda_s - \lambda_t}$$

and

$$e_B^{\text{w-w}}(S;\mathbb{N},\varphi_{\gamma^*}) = \text{rad}_B^{\text{w-w}}(S;\mathbb{N}) = \sqrt{\lambda_t + \delta^2 \frac{(\lambda_s - \lambda_t)}{\eta_s}}.$$

Note a rather surprising property that γ^* (the optimal algorithm) does not depend on the noise level δ. This is obviously a desirable property, especially when δ is not exactly known. For these results and the corresponding formulas for arbitrary δ, see Sect. 2.6.1 in [29].

Consider now an even more special case, in which $F = G = \mathbb{R}^d$ with the ordinary inner product, $S = I$ is the identity, and $\text{rank}(N) = d$. That is, we want to approximate a vector $f \in \mathbb{R}^d$ using information of full rank. Then $\lambda_i = 1$ and $\eta_i > 0$ for all $i = 1, 2, \ldots, d$. For $\delta^2 \le \min_{1\le i\le d} \eta_i$ we find that $\gamma^* = 0$, and that the optimal algorithm

$$\varphi_0(\mathrm{y}) = N^{-1} P_N \mathrm{y}$$

is nothing but the *least squares* algorithm. In addition, we have

$$e_B^{\text{w-w}}(I;\mathbb{N},\varphi_0) = \frac{\delta}{\min_{1\le i\le d} \sqrt{\eta_i}}.$$

We now switch to Gaussian noise, $\mathrm{x} \sim \mathscr{G}_n(0, \sigma^2 \Sigma)$. In this case, the situation is more complicated. A major step forward was done when importance of rectangular and ellipsoidal problems was discovered in [6] and [22]. This allowed us to find asymptotically optimal algorithms for L_2-approximation in the space of univariate functions f with $f^{(r)} \in L_2$, see [20]. In [4, 5] an approximation problem over Besov and Triebel bodies is considered, where non-linear algorithms turn out to be much better than linear algorithms. The following special results can be found in Sect. 4.3.2 of [29].

Consider again the case when S^*S and N^*N have a common orthonormal basis of eigenelements (10). Let

$$s = \min\{i : \lambda_{i+1} = 0 \text{ or } \eta_{i+1} = 0\}$$

and k be the smallest integer from $\{1, 2, \ldots, s\}$ for which

$$\sqrt{\lambda_{k+1}} \le \frac{\sigma^2 \sum_{j=1}^k (\sqrt{\lambda_j}\eta_j)^{-1}}{1 + \sigma^2 \sum_{j=1}^k \eta_j^{-1}}, \tag{11}$$

or $k = s + 1$ if (11) never holds. Define τ_i^2 for $i \ge 1$ as follows:

(i) If $1 \le k \le s$ then

$$\tau_i^2 = \begin{cases} \sigma^2 \frac{\lambda_i}{\eta_i} \left(\sqrt{\lambda_i} \left(\frac{1+\sigma^2 \sum_{j=1}^k \eta_j^{-1}}{\sigma^2 \sum_{j=1}^k (\sqrt{\lambda_j}\eta_j)^{-1}} \right) - 1 \right), & 1 \le i \le s, \\ 0, & i \ge s+1. \end{cases} \tag{12}$$

(ii) If $k = s+1$ then

$$\tau_i^2 = \begin{cases} \sigma^2 \frac{\lambda_i}{\eta_i} \left(\frac{\sqrt{\lambda_i}}{\sqrt{\lambda_{s+1}}} - 1 \right), & 1 \le i \le s, \\ \lambda_{s+1} - \sigma^2 \sqrt{\lambda_{s+1}} \sum_{j=1}^s \left(\frac{\sqrt{\lambda_j} - \sqrt{\lambda_{s+1}}}{\eta_j} \right), & i = s+1, \\ 0, & i \ge s+2. \end{cases} \tag{13}$$

The optimal linear algorithm is given as

$$\varphi_{\text{lin}}^{\text{w–a}}(\text{y}) = \sum_{j=1}^s \frac{\sigma^2 \tau_j^2}{\sigma^2 + \eta_j \tau_j^2} z_j \, S(\xi_j)$$

where $z_j = \eta_j^{-1} \langle N\xi_j, \Sigma^{-1}\text{y} \rangle_2$, $1 \le j \le s$, and

(i) For $1 \le k \le s$

$$e_B^{\text{w–a}}(S; \mathbb{N}, \varphi_{\text{lin}}^{\text{w–a}}) = \sigma \cdot \sqrt{\sum_{j=1}^k \frac{\lambda_j}{\eta_j} - \frac{\sigma^2 \left(\sum_{j=1}^k (\sqrt{\lambda_j}\eta_j)^{-1} \right)^2}{1 + \sigma^2 \sum_{j=1}^k \eta_j^{-1}}},$$

(ii) For $k = s+1$

$$e_B^{\text{w–a}}(S; \mathbb{N}, \varphi_{\text{lin}}^{\text{w–a}}) = \sqrt{\lambda_{s+1} + \sigma^2 \sum_{j=1}^s \frac{(\sqrt{\lambda_j} - \sqrt{\lambda_{s+1}})^2}{\eta_j}}.$$

Moreover, nonlinear algorithms can be at most $\kappa_1 < 1.12$ times better than $\varphi_{\text{lin}}^{\text{w–a}}$.

Consider again the special case when $F = G = \mathbb{R}^d$, $S = I$, and N has full rank. Then the formulas simplify to

$$\varphi_{\text{lin}}^{\text{w–a}}(\text{y}) = \frac{1}{1 + \sigma^2 \text{trace}((N^*N)^{-1})} \sum_{j=1}^d z_j \xi_j$$

where $z_j = \eta_j^{-1} \langle N\xi_j, \Sigma^{-1}\text{y} \rangle_2$, and

$$e_B^{\text{w–a}}(I; \mathbb{N}, \varphi_{\text{lin}}) = \sigma \cdot \sqrt{\frac{\text{trace}((N^*N)^{-1})}{1 + \sigma^2 \text{trace}((N^*N)^{-1})}}.$$

In particular, for $N = I$ and for a diagonal matrix Σ we have $\mathrm{z} = \mathrm{y}$. This means that $\varphi_{\mathrm{lin}}^{\mathrm{w\text{–}a}}$ is *not* a smoothing spline algorithm for any parameter γ since the latter uses coefficients $c_j = (1+\gamma/\eta_j)^{-1}$ at y_j. Note also that the least squares algorithm is in this case not optimal (unless $B = \mathbb{R}^d$).

3.2.2 Average Case Setting

We make the same assumptions as in Sect. 3.1.2 except that $S : F \to G$ is not a functional, but a continuous linear operator. As we have already mentioned, for linear information with Gaussian noise $\mathrm{x} \sim \mathscr{G}_n(0, \sigma^2\Sigma)$ optimal algorithms are the mean elements of appropriate conditional distributions in G. More precisely,

$$\varphi_{\mathrm{opt}}^{\mathrm{a\text{–}a}}(\mathrm{y}) = \sum_{j=1}^{n} z_j\, S(C_\mu L_j)$$

where $(\sigma^2\Sigma + G_N)\mathrm{z} = \mathrm{y}$, with $G_N = [L_i(C_\mu L_j)]_{i,j=1}^n$. Furthermore,

$$e_\mu^{\mathrm{a\text{–}a}}(S, \mathbb{N}, \varphi_{\mathrm{opt}}^{\mathrm{a\text{–}a}}) = \mathrm{rad}_\mu^{\mathrm{a\text{–}a}}(S, \mathbb{N}) = \sqrt{\mathrm{trace}(SC_\mu S^*) - \mathrm{trace}(\varphi_{\mathrm{opt}}^{\mathrm{a\text{–}a}}(NC_\mu S^*))}.$$

It turns out that this algorithm can be interpreted as smoothing spline algorithm. See [36] for smoothing splines in the reproducing kernel Hilbert spaces. For the following result, see Theorem 3.3 and Sect. 3.6.3 in [29].

Theorem 7. *For a Banach space F, let H be a Hilbert space H such that (H, F) is an abstract Wiener space. Then the optimal algorithm*

$$\varphi_{\mathrm{opt}}^{\mathrm{a\text{–}a}}(\mathrm{y}) = S(\mathrm{s}_{\sigma^2}(\mathrm{y})),$$

where $\mathrm{s}_{\sigma^2}(\mathrm{y})$ is the minimizer of

$$\sigma^2 \cdot \|f\|_H^2 + \|\mathrm{y} - N(f)\|_Y^2$$

over all $f \in H$.

Very little is known about approximation of operators in the average case with noise bounded in a Hilbert norm, $\|\mathrm{x}\|_Y \le \delta$. Therefore we mention only one result that corresponds to the examples from the former settings.

Suppose again that $F = G = \mathbb{R}^d$, $S = I$, and N is of full rank. Then, for sufficiently small $\delta > 0$ the least squares algorithm $\varphi_{ls}(\mathrm{y}) = N^{-1}P_N\mathrm{y}$ is optimal and linear, see Sect. 5.3 in [29] for details.

4 Complexity

In Sect. 3, we were looking for optimal algorithms φ for given noisy information $\mathbb{N}$. In practical computations, we usually have some freedom in choosing information and its accuracy. Then we want to compute an approximation within a given error ε with minimal cost. In this section, we formalize the notions of cost of approximation and ε-complexity of a problem, and find the complexity for some problems.

4.1 Cost of Information

Let Λ be a given class of linear functionals over F. We assume that information about $f \in F$ is collected by noisy evaluations of linear functionals from Λ at various levels of precision. We distinguish between nonadaptive information and adaptive information.

4.1.1 Bounded Noise

In case of bounded noise, *nonadaptive* information $\mathbb{N}$ consists of an (exact) information operator $N : F \to \mathbb{R}^n$ of the form

$$N(f) = [L_1(f), L_2(f), \ldots, L_n(f)], \qquad f \in F, \tag{14}$$

where $L_1, \ldots, L_n \in \Lambda$, and a *precision vector*

$$\Delta = [\delta_1, \delta_2, \ldots, \delta_n],$$

where $\delta_i \geq 0$ for $1 \leq i \leq n$. Then $\mathbb{N} : F \to 2^Y$ with $Y = \mathbb{R}^n$,

$$\mathbb{N}(f) = \{\mathrm{y} \in \mathbb{R}^n : \ (\mathrm{y} - N(f)) \in X(\Delta)\}.$$

The essence of this definition is that the 'size' of the noise depends on the precision used; the higher precision the smaller the noise. In this paper, we consider

$$X(\Delta) = \{\mathrm{x} = [x_1, \ldots, x_n] : \ |x_i| \leq \delta_i, \ 1 \leq i \leq n\},$$

which means that each $L_i(f)$ is evaluated with absolute error at most δ_i. For other possibilities and discussion, see Sect. 2.7 in [29].

Adaptive information has a richer structure. The decision about the choice of each successive functional L_i, each precision δ_i, and when to terminate is made based on the values $y_1, \ldots, y_{i-1}$ obtained from previous steps. That is, we first choose the linear functional $L_1 \in \Lambda$ and precision δ_1, and compute y_1 such that

$|y_1 - L_1(f)| \le \delta_1$. In the ith step we make a decision whether we terminate or proceed further with computations. If we decide to stop then $y_1, \ldots, y_{i-1}$ is our final information about f. Otherwise we choose $L_i \in \Lambda$ and δ_i based on $y_1, \ldots, y_{i-1}$ and compute y_i such that $|y_i - L_i(f)| \le \delta_i$.

More formally, we assume that the range Y of adaptive information $\mathbb{N} : F \to 2^Y$ satisfies the condition:

$$\text{for any } [y_1, y_2, \ldots] \in \mathbb{R}^\infty \text{ there is exactly one } n \text{ such that } [y_1, y_2, \ldots, y_n] \in Y. \tag{15}$$

Denote $Z_i = \{\, [y_1, \ldots, y_i] \in \mathbb{R}^i : \ [y_1, \ldots, y_j] \notin Y, \ 1 \le j \le i \,\}$. Then

$$\begin{aligned} \mathbb{N}(f) = \{\, [y_1, \ldots, y_n] \in Y : \qquad & \\ |y_i - L_i(f; y_1, \ldots, y_{i-1})| \le \delta_i(y_1, \ldots, y_{i-1}), \ & i = 1, 2, \ldots, n \,\} \end{aligned} \tag{16}$$

where each $\delta_i : Z_{i-1} \to [0, \infty)$ and each functional $L_i : F \times Z_{i-1} \to \mathbb{R}$ is such that for any fixed $[y_1, \ldots, y_{i-1}] \in Z_{i-1}$ we have $L_i(\cdot; y_1, \ldots, y_{i-1}) \in \Lambda$.

Note that (15) assures that the process of gaining information terminates, which happens when the condition $[y_1, \ldots, y_i] \in Y$ is met. Obviously, any nonadaptive information is also adaptive.

With computation of y_i we associate some cost. The cost depends on the precision δ_i via a *cost function*

$$c : [0, \infty) \to [0, \infty].$$

We assume that c is nonincreasing. For instance,

$$c(\delta) = \begin{cases} +\infty, & 0 \le \delta < \delta_0, \\ 1, & \delta_0 \le \delta, \end{cases}$$

which corresponds to a fixed noise level δ_0, and to the exact information when $\delta_0 = 0$. It seems natural to assume that the cost depends polynomially on δ^{-1}, e.g.,

$$c_s(\delta) = (1 + \delta^{-1})^s$$

for some $s \ge 0$, as this is what we usually have in numerical computations. Indeed, for problems for which we use a mesh of size h, the accuracy and cost depend inversely on h. Note that $s = 0$ corresponds to the exact information, $c \equiv 1$ (under the convention that $\infty^0 = 1$).

Let $\mathbb{N}$ be a given, in general adaptive, information (16). The cost of obtaining information $\mathrm{y} = [y_1, \ldots, y_n]$ about some $f \in F$, $\mathrm{y} \in \mathbb{N}(f)$, is given as

$$\mathrm{cost}(\mathbb{N}; \mathrm{y}) = \sum_{i=1}^{n} c(\delta_i(y_1, \ldots, y_{i-1})).$$

4.1.2 Gaussian Noise

In the case of Gaussian noise, the process of gaining information is basically the same as for bounded noise. The only difference is that instead of the bounds δ_i we choose variances σ_i^2 of successive observations.

Specifically, nonadaptive information consists of an exact information operator (14) and a precision (variance) vector $\Sigma = [\sigma_1^2, \ldots, \sigma_n^2]$. Then y is information about f iff

$$\mathrm{y} \sim \pi_f = \mathscr{G}_n(N(f), \Sigma) \quad \text{where } \Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2).$$

(Note that we use the same letter Σ to denote the precision vector and the covariance matrix.)

For adaptive information $\mathbb{N} : F \to \mathscr{P}_Y$, we assume that the range Y satisfies (15), and that the mappings $L_i(f;\cdot), \sigma_i(\cdot) : Y_{i-1} \to \mathbb{R}$, where $Y_{i-1} = Y \cap \mathbb{R}^{i-1}$, are measurable. The measure π_f is defined inductively on each Y_i as follows. For $i = 1$, we define

$$\pi_f(A_1) = \mathscr{N}\left(L_1(f), \sigma_1^2\right)(A_1), \qquad A_1 \in \mathscr{B}(Y_1),$$

where $\mathscr{B}(Y_1)$ denotes Borel measurable subsets of Y_1, and for $i = 2, 3, \ldots$ we have

$$\pi_f(A_i) = \int_{Z_{i-1}} \mathscr{N}\left(L_i(f;\mathrm{y}^{(i-1)}), \sigma_i^2(\mathrm{y}^{(i-1)})\right)\left(C_{\mathrm{y}^{(i-1)}}\right) \pi_f\left(\mathrm{dy}^{(i-1)}\right), \quad A_i \in \mathscr{B}(Y_i),$$

where $C_{\mathrm{y}^{(i-1)}} = \left\{y_i \in \mathbb{R} : \ [y^{(i-1)}, y_i] \in A_i\right\}$. Here $\mathscr{N}(a, \sigma^2)$ is the one-dimensional normal distribution with mean a and variance σ^2. Then, for any $A \in \mathscr{B}(Y)$, we define

$$\pi_f(A) = \sum_{i=1}^{\infty} \pi_f(A \cap \mathbb{R}^i).$$

The cost of obtaining information $\mathrm{y} = [y_1, \ldots, y_n]$ about f is defined as

$$\mathrm{cost}(\mathbb{N}; \mathrm{y}) = \sum_{i=1}^{n} c(\sigma_i(y_1, \ldots, y_{i-1})),$$

where $c(\sigma)$ is the cost of a single observation with variance σ^2.

4.2 Worst Case and Randomized Complexities

The *information ε-complexity* (or simply *ε-complexity*) of a problem $S : F \to G$ is defined as

$$\mathrm{comp}^{\mathrm{sett}}(S;\varepsilon) = \inf\left\{\mathrm{cost}^{\mathrm{sett}}(\mathbb{N}) : \quad \exists\varphi \text{ such that } e^{\mathrm{sett}}(S;\mathbb{N},\varphi) \le \varepsilon\right\}, \tag{17}$$

where $e^{\mathrm{sett}}(S;\mathbb{N},\varphi)$ is the error, and $\mathrm{cost}^{\mathrm{sett}}(\mathbb{N})$ is the cost of information $\mathbb{N}$ in a given setting. In the worst case setting, we have

- For bounded noise:

$$\mathrm{cost}_B^{\mathrm{w\text{–}w}}(\mathbb{N}) = \sup_{f\in B}\ \sup_{\mathrm{y}\in\mathbb{N}(f)} \mathrm{cost}(\mathbb{N};\mathrm{y}),$$

- For Gaussian noise:

$$\mathrm{cost}_B^{\mathrm{w\text{–}a}}(\mathbb{N}) = \sup_{f\in B}\int_Y \mathrm{cost}(\mathbb{N};\mathrm{y})\,\pi_f(\mathrm{dy}),$$

while in the randomized setting

$$\mathrm{cost}_B^{\mathrm{ran}}(\{\mathbb{N}_\omega\}) = \sup_{f\in B}\left(\mathbb{E}_\omega\int_{Y_\omega}\mathrm{cost}(\mathbb{N}_\omega;\mathrm{y})\,\pi_f(\mathrm{dy}|\omega)\right).$$

4.2.1 Adaption Versus Nonadaption

For a given setting, denote by

$$\overline{\mathrm{comp}}^{\mathrm{sett}}(S;\varepsilon)$$

the minimum cost of obtaining approximation within ε using only nonadaptive information. The problem of how $\overline{\mathrm{comp}}^{\mathrm{sett}}(S;\varepsilon)$ is related to $\mathrm{comp}^{\mathrm{sett}}(S;\varepsilon)$ has been extensively studied for exact information, see, e.g., [16] for a comprehensive survey.

In the worst case setting with bounded noise, we have the following theorem, which is a generalization of the corresponding result for exact information, see, e.g., Theorem 5.2.1 in Chap. 4 of [34] and Theorem 2.15 in [29].

Theorem 8. *Let the class $B \subset F$ be convex and balanced. Then*

$$\overline{\mathrm{comp}}_B^{\mathrm{w\text{–}w}}(S;\varepsilon) = \mathrm{comp}_B^{\mathrm{w\text{–}w}}(S;\alpha\varepsilon)$$

where $1 \le \alpha \le 2$. If S is a functional then $\alpha = 1$.

In the worst case setting with Gaussian noise, adaption can significantly help, even for convex balanced classes B. This is a marked contrast to the case of bounded noise. A simple explanation for this is as follows: Even though randomization is formally not allowed, we can mimic a random selection of information and algorithm by using the adaptive mechanism along with noise. Since randomized algorithms are much better than deterministic algorithms for many problems (the

main example being Monte Carlo for the integration problem over many function classes), adaption can help. Hence, for some problems S we have

$$\mathrm{comp}_B^{\text{w–a}}(S;\varepsilon) \ll \overline{\mathrm{comp}}_B^{\text{w–a}}(S;\varepsilon).$$

In short, the mechanism works as follows. We take arbitrary functional $L \in \Lambda$ and observe it twice with arbitrary precisions σ_i^2. We obtain $y_i = L_i(f) + x_i$, $i = 1, 2$. Now, the next functional L_3 can be chosen dependently on y_1, y_2 via $\omega = y_1 - y_2$. Then L_3 is formally chosen adaptively. However, since ω is the zero mean Gaussian random variable with variance $\sigma_1^2 + \sigma_2^2$, the selection of L_3 can also be viewed as random based on the value of ω. Hence we mimic randomization at cost of just two observations, see, e.g., [27] for details.

Actually, we have the following general result.

Theorem 9. *For any $B \subset F$ we have*

$$\mathrm{comp}_B^{\mathrm{ran}}(S;\varepsilon) \le \mathrm{comp}_B^{\text{w–a}}(S;\varepsilon) \le \mathrm{comp}_B^{\mathrm{ran}}(S;\varepsilon) + 2c_0$$

where $c_0 = \lim_{\sigma\to\infty} c(\sigma)$.

The proof for fixed precision σ can be found in [28], and it can be straightforwardly generalized to variable precision.

Note that even if the algorithm realizing the ε-complexity in the worst case with random noise is adaptive, the corresponding algorithm in the randomized setting can be nonadaptive. The question whether adaption helps in the randomized setting for linear problems over convex and balanced classes B is open.

4.2.2 Complexity of Integration in Hölder Classes

In this section, we find the ε-complexity of integration, $S = \mathrm{Int}$, for folded Hölder classes of functions defined on $D = [0,1]^d$. Specifically, we assume that the set $B = C_\alpha^r(D)$ consists of functions $f : D \to \mathbb{R}$ for which all partial derivatives of order up to r exist and satisfy Hölder condition with exponent α, where $0 \le \alpha < 1$. That is, for any multi-index $\mathrm{r} = (r_1, \ldots, r_d)$ with $|\mathrm{r}| = r_1 + \ldots + r_d \le r$ we have

$$|f^{(\mathrm{r})}(\mathrm{t}_1) - f^{(\mathrm{r})}(\mathrm{t}_2)| \le \|\mathrm{t}_1 - \mathrm{t}_2\|_\infty^\alpha, \qquad \text{for all} \quad \mathrm{t}_1, \mathrm{t}_2 \in D.$$

The class Λ of permissible functionals consists of standard function evaluations, i.e., $L \in \Lambda$ iff it is of the form

$$L(f) = f(\mathrm{t}) \qquad \forall f \in B$$

for some $\mathrm{t} \in D$.

In the sequel, we write $a(\varepsilon) \asymp b(\varepsilon)$ iff there exist $\varepsilon_0 > 0$ and $0 < c_1 \le c_2 < \infty$ such that

$$c_1\, b(\varepsilon) \le a(\varepsilon) \le c_2\, b(\varepsilon), \qquad \forall\, \varepsilon \in (0, \varepsilon_0].$$

For exact information the problem was analyzed, e.g., in [15] and [7]. In the deterministic setting, the ε-complexity is proportional to $\varepsilon^{-\left(\frac{d}{r+\alpha}\right)}$, and is achieved by cubatures $Q_n(f) = \mathrm{Int}(P_f)$ where P_f is the piecewise polynomial of degree $(r-1)$ with respect to each coordinate that interpolates f on the uniform grid. In the randomized (Monte Carlo) setting, the ε-complexity is proportional to $\varepsilon^{-\left(\frac{d}{r+\alpha+d/2}\right)}$, and it is achieved by $\tilde{Q}_n(f) = \mathrm{Int}(P_f) + \mathbf{MC}_n(f - P_f)$ where $\mathbf{MC}_n$ is the classical Monte Carlo. In particular, optimal algorithms are nonadaptive. We now generalize these results to noisy information.

Consider first the bounded noise, so that

$$|y_j - f(\mathrm{t}_j)| \le \delta_j \quad \text{for all } j.$$

Theorem 10. *Let the cost function*

$$c(\delta) \asymp \delta^{-s}.$$

For integration in the Hölder class $B = C^r_\alpha(D)$ we have

$$\mathrm{comp}^{\mathrm{w\text{–}w}}_B(\mathrm{Int};\varepsilon) \;\asymp\; \overline{\mathrm{comp}}^{\mathrm{w\text{–}w}}_B(\mathrm{Int};\varepsilon) \;\asymp\; \varepsilon^{-\left(\frac{d}{r+\alpha}+s\right)}.$$

Proof. The upper bound is easily obtained by nonadaptive algorithms that are optimal for exact information. That is, for given ε, let $n \asymp \varepsilon^{-d/(r+\alpha)}$ and $Q_n(f) = \sum_{j=1}^n a_j\, f(\mathrm{t}_j)$ be the cubature that is based on piecewise polynomial interpolation on the uniform grid, for which the worst case error for exact information is at most $\varepsilon/2$. Then $\sum_{j=1}^n |a_j| \le A_1$ for some A_1 independent of n, as shown in [28]. Taking the error of Q_n for information with noise bounded by δ we obtain

$$\Big|\int_D f(\mathrm{t})\,\mathrm{dt} - \sum_{j=1}^n a_j\,(f(\mathrm{t}_j) + x_j)\Big| \le \Big|\int_D f(\mathrm{t})\,\mathrm{dt} - \sum_{j=1}^n a_j\, f(\mathrm{t}_j)\Big| + \delta \sum_{j=1}^n |a_j|$$
$$\le \tfrac{\varepsilon}{2} + \delta\, A_1.$$

For $\delta = \varepsilon/(2A_1)$ the error is at most ε, and the cost is proportional to $n\, c(\delta) = \varepsilon^{-\left(\frac{d}{r+\alpha}+s\right)}$, as claimed.

To show the lower bound, we reduce the original problem to a simpler one (with respect to complexity). We choose a "bump" function $\psi \in C^r_\alpha(\mathbb{R}^d)$ that is supported on the cube D, with both, $a = \|\psi\|_\infty$ and $b = \int_D \psi(\mathrm{x})\mathrm{dx}$ being positive. Given $h = 1/m$, for all m^d multi-indices $\mathbf{i} = (i_1, \ldots, i_d)$, $0 \le i_j \le m-1$, $1 \le j \le d$, define the functions

$$\psi_{h,\mathbf{i}}(\mathrm{x}) = h^{r+\alpha}\psi\left(\frac{\mathrm{x}-\mathrm{x}_{\mathbf{i}}}{h}\right),$$

where $\mathrm{x}_{\mathbf{i}} = (i_1 h, \ldots, i_d h)$. The $\psi_{h,\mathbf{i}}$ are all in $C^r_\alpha(D)$ and have mutually disjoint supports, each on a cube of edge length h. Define the function class

$$B_h = \left\{ f = \sum_{\mathbf{i}} a_{\mathbf{i}}\psi_{\mathbf{i}} : \ |a_{\mathbf{i}}| \le 1 \ \forall\, \mathbf{i} \right\}.$$

Since $B_h \subset B$, the integration problem over B is not easier than the integration problem over B_h. Noting that $\|\psi_{h,\mathbf{i}}\|_\infty = ah^{r+\alpha}$ and that

$$\int_D f(\mathrm{t})\mathrm{dt} = \left(\frac{b}{a}\right) h^d \left(\sum_{\mathbf{i}} a_{\mathbf{i}}(ah^{r+\alpha})\right) \qquad \forall\, f \in B_h,$$

we conclude that the latter problem is not easier than the following problem.

(AP) For $k = h^{-d}$, $\beta = (b/a)h^d$, and $\tau = ah^{r+\alpha}$, approximate the sum

$$\beta \cdot \sum_{i=1}^{k} v_i$$

of a vector $\mathrm{v} = (v_1, v_2, \ldots, v_k)$ from the ball $\left\{\mathrm{v} \in \mathbb{R}^k : \ |v_j| \le \tau, \ 1 \le j \le k\right\}$. Available information functionals are noisy evaluations of the coefficients of v.

We bound from below the complexity of (AP) for a special choice of ε. In view of Theorem 8 we can restrict ourselves to nonadaptive approximations only. Since repetitive observations do not help, and observations with precisions $\delta \ge \tau$ are useless, we can also assume that each coefficient is observed at most once with precision at most τ. Suppose without loss of generality that we observe the first n coefficients where $n \le k$. Then the optimal approximation is

$$\varphi(y_1, \ldots, y_n) = [y_1, \ldots, y_n, \underbrace{0, \ldots, 0}_{k-n}]$$

and its error equals $\beta\Big(\big(\sum_{i=1}^n \delta_i\big) + (k-n)\tau\Big)$.

Suppose now that $\varepsilon = \tau\beta k/4 \asymp h^{r+\alpha}$. Then for the error to be at most ε, there must be at least $k/2 \asymp \varepsilon^{-\left(\frac{d}{r+\alpha}\right)}$ indices i with $\delta_i \le \tau/2 \asymp \varepsilon$. This yields that the cost is at least proportional to $k\, c(\varepsilon) \asymp \varepsilon^{-\left(\frac{d}{r+\alpha}+s\right)}$, as desired. Since $\varepsilon \to 0^+$ as $h \to 0^+$, the proof is complete. □

We now consider Gaussian noise,

$$(y_j - f(\mathrm{t}_j)) \sim \mathcal{N}(0, \sigma_j^2).$$

Theorem 11. *Let the cost function satisfy*

$$1 \le c(\sigma) \asymp \sigma^{-s}, \qquad \text{where} \quad s > 0.$$

For integration in the Hölder class $B = C_\alpha^r(D)$ we have:

(i) For nonadaptive approximations

$$\overline{\text{comp}}_B^{\text{w-a}}(\text{Int};\varepsilon) \asymp \begin{cases} \varepsilon^{-2}, & d \le 2(r+\alpha), \quad s \ge 2, \\ \varepsilon^{-\left(\frac{d}{r+\alpha}\left(1-\frac{s}{2}\right)+s\right)}, & d \le 2(r+\alpha), \quad 0 < s < 2, \\ \varepsilon^{-\left(\frac{d}{r+\alpha}\right)}, & d > 2(r+\alpha), \end{cases}$$

(ii) For adaptive approximations

$$\text{comp}_B^{\text{w-a}}(\text{Int};\varepsilon) \asymp \text{comp}_B^{\text{ran}}(\text{Int};\varepsilon) \asymp \begin{cases} \varepsilon^{-2}, & s \ge 2, \\ \varepsilon^{-\left(\frac{d}{r+\alpha+d/2}\left(1-\frac{s}{2}\right)+s\right)}, & 0 < s < 2. \end{cases}$$

Proof. We first show (i). Since we restrict ourselves to only nonadaptive approximations, the problem, by Theorem 5, is equivalent to the same problem, but with noise x bounded in the Hilbert norm, so that

$$\sum_{j=1}^{n} |x_j|^2/\delta_j^2 \le 1. \tag{18}$$

Then the upper bounds can be obtained as in the proof of Theorem 10. That is, take the cubature $Q_n(f) = \sum_{j=1}^n a_j f(\mathrm{t}_j)$ with error for exact information at most $\varepsilon/2$. Now $\sum_{j=1}^n |a_j|^2 \le A_2/n$ for some A_2 independent of n. Hence for $\delta_j = \delta$ such that $\delta n^{-1/2} = \varepsilon A_2^{-1/2}/2$ we have

$$\left| \int_D f(\mathrm{t})\mathrm{dt} - \sum_{j=1}^{n} a_j (f(\mathrm{t}_j) + x_j) \right| \le \frac{\varepsilon}{2} + \sum_{j=1}^{n} |a_j x_j| \le \frac{\varepsilon}{2} + \frac{\delta A_2^{1/2}}{\sqrt{n}} \le \varepsilon.$$

The upper bounds are achieved by taking

$$\begin{aligned} \delta \asymp 1 \ \text{ and } \ n \asymp \varepsilon^{-2} \qquad & \text{for} \quad d \le 2(r+\alpha), \ s \ge 2, \\ \delta \asymp \varepsilon^{\left(1-\frac{d}{r+\alpha}\right)} \ \text{ and } \ n \asymp \varepsilon^{-\left(\frac{d}{r+\alpha}\right)} \qquad & \text{for} \quad d \le 2(r+\alpha), \ 0 \le s < 2, \\ \delta \asymp 1 \ \text{ and } \ n \asymp \varepsilon^{-\left(\frac{d}{r+\alpha}\right)} \qquad & \text{for} \quad d > 2(r+\alpha). \end{aligned}$$

The lower bound in case $d > 2(r+\alpha)$ follows from the general property that the complexity for nonadaptive approximations with noisy information is not smaller than the complexity for exact information, see Lemma 1 in [28]. To show the lower

bound ε^{-2} for $s \geq 2$, it suffices to consider integration of constant functions f and note that, by convexity of the function $t \mapsto t^{s/2}$, the best strategy is to use repetitive observations with $\delta \asymp 1$.

To show the remaining lower bound, we proceed as in the corresponding part of the proof of Theorem 10 and arrive at the problem (AP) with noise (18). Assume for simplicity that the cost $c(\delta) = (1+\delta^{-2})^{s/2}$. Then repetitive observations do not help. Indeed, suppose that a coefficient is observed ℓ times with precisions $\delta_1, \ldots, \delta_\ell$. In terms of the error, this is equivalent to just one observation with precision $\delta = \left(\sum_{j=1}^{\ell} \delta_j^{-2}\right)^{-1/2}$. Comparing the costs in both situations we have

$$\sum_{i=1}^{\ell} c(\delta_i) = \sum_{i=1}^{\ell} (1+\delta_i^{-2})^{s/2} \geq \left(\sum_{i=1}^{\ell} (1+\delta_i^{-2})\right)^{s/2} \geq \left(1+\sum_{i=1}^{\ell} \delta_i^{-2}\right)^{s/2} = c(\delta)$$

where we used the fact that the function $t \mapsto t^{s/2}$ is concave for $s < 2$.

Thus we can restrict ourselves to at most one observation for each coordinate. Suppose we observe $v_1, \ldots, v_n$ with precisions $\delta_1 \leq \cdots \leq \delta_n$. Then the radius of the corresponding noisy information can be expressed as

$$\beta \cdot \max\left\{\sum_{i=1}^{k} v_i : \ |v_i| \leq \tau, \ \sum_{i=1}^{n} v_i^2 \delta_i^{-2} \leq 1\right\}.$$

Since we can reduce the error by neglecting a given observation and increasing the precisions of the previous observations, we may assume that $\delta_n^2 \leq \tau\left(\sum_{j=1}^{n} \delta_j^2\right)^{1/2}$. Then this radius equals

$$\beta\Big(\Big(\sum_{j=1}^{n} \delta_j^2\Big)^{1/2} + (k-n)\tau\Big).$$

Since the function $t \mapsto (1+t)^{s/2}$ is concave, we can replace those n observations with different precisions δ_i by n observations with the same precision $\delta^2 = (\sum_{j=1}^{n} \delta_j^2)/n$ obtaining information with the same radius and smaller cost. We arrive at the conclusion that to find the complexity of (AP) we have to minimize $n\, c(\delta)$ under the conditions that $\delta \leq \tau\sqrt{n}$ and

$$\sqrt{n}\,\delta + (k-n)\tau \leq \frac{\varepsilon}{\beta}.$$

Now, take $\varepsilon = \tau\beta k/2$. Then $n \geq k/2 \asymp \varepsilon^{-\left(\frac{d}{r+\alpha}\right)}$ and $\delta \leq \sqrt{k}\tau\beta/2 \asymp \varepsilon^{1-\frac{d}{2(r+\alpha)}}$, so that the ε-complexity is at least

$$\frac{1}{2} k (1+\delta^{-2})^{s/2} \asymp \varepsilon^{-\left(\frac{d}{r+\alpha}\left(1-\frac{s}{2}\right)+s\right)}$$

as claimed. This obviously holds for any cost function satisfying $1 \le c(\delta) \asymp \delta^{-s}$.

We now show (ii). The upper bound is obtained by applying the optimal randomized algorithm $\tilde{Q}_n$ with the squared error at most $\varepsilon^2/2$ for exact information. Its coefficients a_j satisfy $\sum_{j=1}^{2n} |a_j|^2 \le A_3/n$ for some A_3 independent of n. Taking the noise level $\sigma_j = \sigma$ such that $\sigma^2/n \le \varepsilon^2/(2A_3)$, the squared expected error for any $f \in B$ can be bounded as

$$\mathbb{E}\Big(\int_D f(\mathrm{t})\mathrm{dt} - \sum_{j=1}^{2n} a_j f(\mathrm{t}_j)\Big)^2 + \sigma^2\Big(\sum_{j=1}^{2n} |a_j|^2\Big) \le \frac{\varepsilon^2}{2} + \frac{\sigma^2 A_3}{n} \le \varepsilon^2.$$

For the case $s \ge 2$, we take $\sigma^2 \asymp 1$ and $n \asymp \varepsilon^{-2}$, whereas in the case $s < 2$ we take $\sigma^2 \asymp \varepsilon^{\left(2-\frac{d}{r+\alpha+d/2}\right)}$ and $n \asymp \varepsilon^{-\left(\frac{d}{r+\alpha+d/2}\right)}$.

The lower bound for $s \ge 2$ can be obtained by switching from the worst case to the average case with a normal distribution of variance λ placed on the constant functions. Then adaption and randomization do not help, see Sect. 4.3.1, and the best strategy is to use repetitive observations with variance $\sigma^2 \asymp 1$.

For $s < 2$, we use another averaging argument. The complexity can be bounded from below by the average case complexity of the problem (AP) with Gaussian noise, and with respect to a probability measure

$$\nu = \underbrace{\mu \times \cdots \times \mu}_{k}$$

where μ is placed on $[-\tau, \tau]$. We then have that the optimal φ is of the form $\varphi(\mathrm{y}) = \sum_{j=1}^{k} \varphi_j(\mathrm{y})$ where φ_j is the optimal average case approximation of v_j. Moreover, φ_j uses only those y_i that come from observations of v_j. This yields that the squared average error of φ equals

$$\begin{aligned}
\mathrm{err}^2 &= \beta^2 \cdot \underbrace{\int_{-\tau}^{\tau} \cdots \int_{-\tau}^{\tau}}_{k} \left(\int_Y \Big| \sum_{j=1}^{k} \beta v_j - \varphi_j(\mathrm{y}) \Big|^2 \pi_{\mathrm{v}}(\mathrm{dy}) \right) \mu(\mathrm{d}v_1) \cdots \mu(\mathrm{d}v_k) \\
&= \beta^2 \cdot \sum_{j=1}^{k} \int_Y \left(\int_{-\tau}^{\tau} |\beta v_j - \varphi_j(\mathrm{y})|^2 \, \mu(\mathrm{d}v_j|\mathrm{y}) \right) \mu_1(\mathrm{dy}) \\
&= \sum_{j=1}^{k} \mathrm{err}_j^2,
\end{aligned}$$

where

$$\mathrm{err}_j^2 = \beta^2 \cdot \left(\int_{-\tau}^{\tau} |\beta v_j - \varphi_j(\mathrm{y})|^2 \, \mu(\mathrm{d}v_j|\mathrm{y}) \right) \mu_1(\mathrm{dy}).$$

Here μ_1 is the (total) distribution of y on Y and $\mu(\cdot|\mathrm{y})$ is the conditional distribution of v_j given information y. Let $\mathrm{cost}_i(\mathrm{y})$ be the cost of all those y_i that come from observations of v_i. Then the average (total) cost can be expressed as

$$\sum_{j=1}^{k} \int_Y \left(\int_{-\tau}^{\tau} \mathrm{cost}_i(\mathrm{y})\, \mu(\mathrm{d}v_j|\mathrm{y}) \right) \mu_1(\mathrm{dy}).$$

Thus the error and cost are separated with respect to the successive coordinates v_j.

We now specify ν by putting equal mass $1/2$ on $\pm\tau$, and suppose that $\mathrm{err}^2 \le \varepsilon^2 = \tau^2\beta^2 k/4$. Then for at least $k/2$ indices j we have $\mathrm{err}_j^2 \le \tau^2\beta^2/2$. Since the initial squared error with respect to ν is $\tau^2\beta^2 k$, the cost of observing v_i must be at least proportional to τ^{-s}. Indeed, suppose that a parameter $v \in \{-1, 1\}$ is approximated with the average squared error $1/2$ at cost K using observations with precisions $\sigma_i = \sigma_i(y_1, \ldots, y_{i-1})$. By a simple change of variables, one can see that this is equivalent to approximating $v \in \{-\tau, \tau\}$ with the squared error $\tau^2/2$ at cost proportional to $K\tau^{-s}$ using observations with

$$\tilde{\sigma}_i(y_1, \ldots, y_{i-1}) = \tau \cdot \sigma_i\Big(\frac{y_1}{\tau}, \ldots, \frac{y_{i-1}}{\tau}\Big).$$

Finally, the cost of approximating v with error ε is at least proportional to

$$k\tau^{-s} \asymp \varepsilon^{-\left(\frac{d}{r+\alpha+d/2}\right)} \varepsilon^{-\left(\frac{s(r+\alpha)}{r+\alpha+d/2}\right)} = \varepsilon^{-\left(\frac{d}{r+\alpha+d/2}\left(1-\frac{s}{2}\right)+s\right)}$$

as claimed. The proof is complete. □

Note the difference between the bounded and Gaussian noise. Integration with bounded noise is always more difficult than without noise, and nonadaptive approximations are optimal. In the presence of Gaussian noise the situation is more complicated. Nonadaptive approximations are optimal only for $d \le 2(r+\alpha)$ and $s \ge 2$, and for $d > 2(r+\alpha)$ the complexity is of the same order as for exact information. Adaptive approximations (that are equivalent to randomized ones) are significantly better for $s \in (0, 2)$;

We also note the assumption $c(\sigma) \ge 1$ in case of random noise. Apart from the fact that it is quite natural, it is also necessary to avoid possibility of reducing the error to an arbitrarily small level by repetitive observations with large variance σ^2. This would happen when, e.g., $c(\sigma) = \sigma^{-s}$ with $s > 2$.

4.3 Average Case Complexity

We consider the average case setting with respect to a Gaussian measure μ whose mean element is zero and a covariance operator $C_\mu : F^* \to G$, and with information contaminated by Gaussian noise. Then the ε-complexity, $\mathrm{comp}_\mu^{\mathrm{a\text{–}a}}(S; \varepsilon)$, is defined by (17) with

$$\begin{aligned}\mathrm{cost}_{\mu}^{\text{a-a}}(\mathbb{N}) &= \int_F \int_Y \mathrm{cost}(\mathbb{N};\mathrm{y})\,\pi_f(\mathrm{dy})\mu(\mathrm{d}f)\\ &= \int_Y \mathrm{cost}(\mathbb{N};\mathrm{y})\,\mu_1(\mathrm{dy}),\end{aligned}$$

where μ_1 is the total distribution of y on Y.

4.3.1 Adaption Versus Nonadaption

The results on adaption versus nonadaption in the average case setting for exact information are presented in [37], see also Sect. 5.6 in Chap. 6 of [34]. They were generalized to noisy information in [23, 26], and also in Sects. 3.7.2 and 3.9.1 in [29]. One of the results is as follows.

Theorem 12. *Suppose that* $\overline{\mathrm{comp}}_{\mu}^{\text{a-a}}(S;\sqrt{\varepsilon})$ *is a semiconvex function of* ε*, i.e., there exist* $\varepsilon_0 > 0$*,* $\alpha > 0$*,* $\beta \ge \alpha$*, and a convex function* $\psi : [0,\infty) \to [0,\infty]$ *such that*

$$\alpha \cdot \psi(\varepsilon) \;\le\; \overline{\mathrm{comp}}_{\mu}^{\text{a-a}}(S;\sqrt{\varepsilon}) \qquad \forall\, \varepsilon \ge 0,$$

and

$$\overline{\mathrm{comp}}_{\mu}^{\text{a-a}}(S;\sqrt{\varepsilon}) \;\le\; \beta \cdot \psi(\varepsilon) \qquad \forall\, \varepsilon \in [0,\varepsilon_0].$$

Then

$$\mathrm{comp}_{\mu}^{\text{a-a}}(S;\varepsilon) \;\ge\; \left(\frac{\alpha}{\beta}\right) \cdot \overline{\mathrm{comp}}_{\mu}^{\text{a-a}}(S;\varepsilon) \qquad \forall \varepsilon \in [0,\varepsilon_0].$$

This theorem is applicable to many problems, including the one presented in Sect. 4.3.2. Examples of problems for which adaption significantly helps in the average case are given in [24].

4.3.2 Complexity of Approximation with $\Lambda = \Lambda^{\text{all}}$

As an example, we now present complexity results for an approximation problem, $S = \mathrm{App}$. This section is based on Sect. 3.10.1 in [29], see also [23, 26, 30] for some other results.

Specifically, we want to approximate elements $f \in F$ with error measured in the norm $\|\cdot\| = \sqrt{\langle\cdot,\cdot\rangle}$ of a Hilbert space G such that the embedding $S : F \to G$ is continuous. We assume that the class of permissible information functionals is defined as

$$\Lambda = \Lambda^{\text{all}} = \left\{ L \in F^* :\; \|L\|_{\mu}^2 = \int_F L^2(f)\,\mu(\mathrm{d}f) \le 1 \right\}.$$

Note that the restriction to functionals with uniformly bounded norm is necessary since we could otherwise observe functionals L with arbitrary μ-norm at constant cost, and this would lead to almost exact information.

Let $S^* : G \to F^*$ be the adjoint operator to the embedding S, i.e., $(S^*g)(f) = \langle f, g\rangle \ \forall\, f \in F$. Then $S\,C_\mu S^* : G \to G$ is a linear operator with finite trace. Denote by

$$\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_n \ge \ldots \ge 0$$

its eigenvalues, and by ξ_i the corresponding orthonormal eigenelements. Finally, let

$$K_i = \lambda_i^{-1/2} S^* \xi_i, \qquad i \ge 1,$$

where we put $K_i = 0$ if $\lambda_i = 0$. Note that $\|K_i\|_\mu = 1 \ \forall i$.

Let $R(T)$ be the minimal error of algorithms that use nonadaptive information with cost at most T. For exact information we have $R(T)^2 = \sum_{j=n+1}^{\infty} \lambda_j$ where $n = \lfloor T \rfloor$, and optimal information consists of observations of K_i for $i = 1, 2, \ldots, n$. For the noisy case, we have the following formulas.

Suppose that the cost function $c(\sigma) = (1 + \sigma^{-2})^{s/2}$. Then, assuming that

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \left(\frac{\lambda_j^{1/2}}{\lambda_n^{1/2}} - 1 \right) > 0,$$

(which holds, e.g., if $\lambda_j \asymp j^{-p}$ for some $p > 1$), we find that the ε-complexity for $s > 2$ is of the same order as the ε-complexity for $s = 2$.

For $s \le 2$ we have the following exact formulas:

$$R(T)^2 = \left(\frac{1}{T}\right)^{2/s} \left(\sum_{i=1}^{n} \lambda_i^r \right)^{1/r} + \sum_{j=n+1}^{\infty} \lambda_j$$

where $r = s/(s+2)$ and $n = n(T)$ is the largest integer satisfying

$$\left(1 + \sum_{i=1}^{n-1} \frac{\lambda_i^r}{\lambda_n^r} \right)^{1/r} - \left(\sum_{i=1}^{n-1} \frac{\lambda_i^r}{\lambda_n^r} \right)^{1/r} \le T^{2/s}.$$

Furthermore, $R(T)$ is attained by observing the functionals $K_1, \ldots, K_n$ with variances

$$\sigma_i^2 = \left(\lambda_i^{2/(2+s)} \left(\frac{T}{\sum_{j=1}^{n} \lambda_j^r} \right)^{2/s} - 1 \right)^{-1}, \qquad 1 \le i \le n.$$

Knowing $R(T)^2$, we can check whether $\overline{\mathrm{comp}}_\mu^{\mathrm{a\text{–}a}}(\mathrm{App};\sqrt{\varepsilon})$ is a convex function of ε. To be more concrete, we present the following result.

Theorem 13. *Let the eigenvalues*

$$\lambda_j \asymp \left(\frac{\ln^q j}{j}\right)^p$$

where $p > 1$ *and* $q \geq 0$. *Consider the cost function*

$$1 \leq c(\sigma) \asymp \sigma^{-s}.$$

Then, for the average case approximation we have

$$\mathrm{comp}_\mu^{\mathrm{a\text{–}a}}(\mathrm{App};\varepsilon) \asymp \begin{cases} \left(\frac{1}{\varepsilon}\right)^{\bar{s}}, & (p-1)\bar{s} > 2, \\ \left(\frac{1}{\varepsilon}\right)^{\frac{2}{p-1}} \left(\ln \frac{1}{\varepsilon}\right)^{\frac{(q+1)p}{p-1}}, & (p-1)\bar{s} = 2, \\ \left(\frac{1}{\varepsilon}\right)^{\frac{2}{p-1}} \left(\ln \frac{1}{\varepsilon}\right)^{\frac{q\,p}{p-1}}, & 0 \leq (p-1)\bar{s} < 2, \end{cases}$$

where $\bar{s} = \min(2, s)$.

It seems surprising that the complexity behaves roughly in only two different ways. If $p < 1 + 2/\bar{s}$ (the problem is 'difficult') then noise does not influence the exponent of ε^{-1}, and if $p > 1 + 2/\bar{s}$ (the problem is 'easy') then the complexity is proportional to $\varepsilon^{-\bar{s}}$.

For instance, consider the L_2-approximation with respect to the r-folded Wiener sheet measure placed on the space $C_0^{r,r,\ldots,r}([0,1]^d)$ of functions $f : [0,1]^d \to \mathbb{R}$ that are r times continuously differentiable with respect to all variables, and the derivative $f^{(j_1,j_2,\ldots,j_d)}(x_1, x_2, \ldots, x_d) = 0$ whenever at least one x_j is zero. From [21] we know that $\lambda_j \asymp \left(\frac{\ln^{d-1} j}{j}\right)^{2r+2}$. Therefore, for $\bar{s} > 1/(r+1/2)$ the ε-complexity is proportional to $\varepsilon^{-\bar{s}}$, and for $\bar{s} < 1/(r+1/2)$ we have

$$\mathrm{comp}_\mu^{\mathrm{a\text{–}a}}(\mathrm{App};\varepsilon) \asymp \left(\frac{1}{\varepsilon}\right)^{\frac{1}{r+1/2}} \left(\ln \frac{1}{\varepsilon}\right)^{\frac{(d-1)(r+1)}{r+1/2}}.$$

5 Tractability

Although a systematic study of tractability of multivariate problems was initiated only in 1994 in [39], there already exists a rich literature on the subject. The main reference is now the three-volume monograph [17–19]. Unfortunately (and fortunately for the author), all those results treat only exact information. In this section, we give some sample results on tractability of problems with noisy information.

5.1 Polynomial Tractability

There are many different notions of tractability. We concentrate on *polynomial tractability*. Let

$$S_d : F_d \to G_d$$

be a problem parameterized by $d = 1, 2, \ldots$, where the spaces F_d consist of functions of d variables. We say that, in a given setting, the problem is *polynomially tractable* iff there exist nonnegative C, p, q such that the inequality

$$\mathrm{comp}^{\mathrm{sett}}(S_d; \varepsilon e_d) \le C \cdot d^q \cdot \varepsilon^{-p}$$

holds for all d and $\varepsilon \in (0, 1)$. Here e_d is the initial error, i.e., the minimal error that can be achieved from zero information.

The choice of the cost function is now a more delicate question since it should depend not only on the precision, but also on the dimension. We do not want to start here a discussion on this. We only notice that for polynomial tractability the cost should depend at most polynomially on d. For instance,

$$c(d, x) = d^t (1 + x^{-1})^s \qquad \text{for some} \quad s, t \ge 0,$$

where x stands for δ in case of bounded noise, and for σ in case of random noise.

5.1.1 Worst Case Integration in Hölder Classes

Consider first the integration problem in Hölder classes of Sect. 4.2.2. In case of information with bounded noise the problem is *not* polynomially tractable, since the exponent $\left(\frac{d}{r+\alpha} + s\right)$ at ε^{-1} in the complexity formula of Theorem 10 grows to infinity as $d \to \infty$. This is not any surprise since problems with bounded noise are never easier than problems with no noise. Actually we know from [8] that the problem is *intractable*, i.e., for a fixed ε, the complexity grows exponentially fast with $d \to \infty$.

The situation is different for information with Gaussian noise. Recall that in this case adaptive deterministic algorithms are equivalent to randomized algorithms. Using classical Monte Carlo for f with n noisy observations of fixed variance $\sigma^2 = 1$ we easily obtain the (sharp) upper bound $\sqrt{2/n}$ for the error. On the other hand, the exponent at ε^{-1} of Theorem 11 equals $\frac{d}{r+\alpha+d/2}\left(1 - \frac{\bar{s}}{2}\right) + \bar{s}$ where $\bar{s} = \min(s, 2)$, which approaches 2 as $d \to \infty$. Thus we have the following result.

Theorem 14. *Consider the worst case integration in Hölder classes* $B = C_\alpha^r([0,1]^d)$ *with the cost function satisfying* $1 \le c(d, x) \asymp d^t x^{-s}$. *The problem is:*

(i) Not polynomially tractable for bounded noise,
(ii) Polynomially tractable for Gaussian noise, and the exponents are

$$p = 2 \quad \textit{and} \quad q = t$$

5.1.2 Average Case Approximation with $\Lambda = \Lambda^{\mathrm{all}}$

As a second example, we consider the approximation problem of Sect. 4.3.2. It is clear that if this problem with exact information is not polynomially tractable then it is not polynomially tractable for information with Gaussian noise. The converse is less obvious.

Theorem 15. *Consider the average case approximation problem with Gaussian noise, $\Lambda = \Lambda^{\mathrm{all}}$, and the cost function $1 \le c(d,\sigma) \asymp d^{\,t}\,\sigma^{-s}$. If the problem is polynomially tractable for exact information ($s = 0$) with the exponents p and q, then the problem is polynomially tractable for information with Gaussian noise ($s > 0$), and the exponents are correspondingly*

$$\begin{cases} p' = \bar{s}, & q' = \bar{s}\,q/p & \textit{for } \bar{s} > p, \\ p' = p^+, & q' = q^+ & \textit{for } \bar{s} = p, \\ p' = p, & q' = q & \textit{for } \bar{s} < p, \end{cases}$$

where $\bar{s} = \min(2, s)$, and p^+ and q^+ denote any numbers larger than p and q, respectively.

Proof. Let $\lambda_{d,1} \ge \lambda_{d,2} \ge \ldots$ denote the eigenvalues of $S_d C_\mu S_d^*$. Then $e_d^2 = \sum_{j=1}^{\infty} \lambda_{d,j}$ and

$$n(d,\varepsilon) = \min\left\{ n \ge 0 : \sum_{j=n+1}^{\infty} \lambda_{d,j} \le \varepsilon^2 e_d^2 \right\}$$

is the minimal number of functional evaluations that allow us to reduce the initial error by a factor of ε for exact information. Polynomial tractability yields $n = n(d,\varepsilon) \le C\, d^{q-t}\varepsilon^{-p}$ or, equivalently,

$$\sum_{j=n+1}^{\infty} \lambda_{d,j} \le \varepsilon^2 e_d^2 \le C^{2/p} d^{2(q-t)/p} n^{-2/p} e_d^2.$$

Since the monotonicity of $\lambda_{d,j}$ implies that $n\lambda_{d,2n} \le \sum_{j=n+1}^{\infty} \lambda_{d,j}$, the eigenvalues can be bounded as

$$\lambda_{d,n} \le C_1 e_d^2 d^{2(q-t)/p} n^{-(1+2/p)}$$

where C_1 is independent of d, ε, and n. To complete the proof, it is now enough to apply Theorem 13 with $\lambda_j \asymp j^{-(1+2/p)}$ and ε replaced by $\varepsilon d^{-(q-t)/p}$, and to multiply the result by d^t. □

We add that the general conditions on $\{\lambda_{d,j}\}$ for polynomial tractability of the problem in case of exact information are given in Sect. 6.1 of [17].

Finally, we note that the L_2-approximation with r-folded Wiener sheet measure from Sect. 5.1.2 is not polynomially tractable, because of the 'bad' dependence of the eigenvalues $\lambda_{d,j}$ on d. Actually, the problem is intractable, as shown in Sect. 3.2.3 of [17].

Acknowledgements This research was supported by the Ministry of Science and Higher Education of Poland under the research grant N N201 547738. The author highly appreciates valuable comments from Henryk Woźniakowski and two anonymous referees.

References

1. Bakhvalov, N.S.: On the optimality of linear methods for operator approximation in convex classes. Comput. Math. Math. Phys. **11**, 244–249 (1971)
2. Casella, G., Strawderman, W.E.: Estimating bounded normal mean. Ann. Statist. **9**, 870–878 (1981)
3. Donoho, D.L.: Statistical estimation and optimal recovery. Ann. Statist. **22**, 238–270 (1994)
4. Donoho, D.L., Johnstone, I.M.: Minimax estimation via wavelet shrinkage. Ann. Statist. **26**, 879–921 (1998)
5. Donoho, D.L., Johnstone, I.M., Kerkyacharian, G., Picard, D.: Wavelet shrinkage: asymptopia? J. Roy. Stat. Soc. ser. B **57**, 301–369 (1995)
6. Donoho, D.L., Liu, R.C., MacGibbon, K.B.: Minimax risk over hyperrectangles, and implications. Ann. Statist. **18**, 1416–1437 (1990)
7. Heinrich, S.: Random approximation in numerical analysis. In: Berstadt et al. (ed.) Proceedings of the Functional Analysis Conference, Essen 1991, pp. 123–171. Marcel Dekker, New York (1993)
8. Hinrichs, A., Novak, E., Ulrich, M., Woźniakowski, H.: The curse of dimensionalty for numerical integration of smooth functions. Submitted
9. Ibragimov I.A., Hasminski, R.Z.: On the nonparametric estimation of the value of a linear functional in Gaussian white noise. Theory Probab. Appl. **29**, 19–32 (1984). In Russian
10. Kacewicz, B.Z., Plaskota, L.: Noisy information for linear problems in the asymptotic setting. J. Complexity **7**, 35–57 (1991)
11. Kuo, H.H.: Gaussian measures in banach spaces. In: Lecture Notes in Math, vol. 463. Springer, Berlin (1975)
12. Magaril-Il'yaev G.G., Osipenko, K.Yu.: On optimal recovery of functionals from inaccurate data. Matem. Zametki **50**, 85–93 (1991). In Russian
13. Melkman, A.A., Micchelli, C.A.: Optimal estimation of linear operators in Hilbert spaces from inaccurate data. SIAM J. Numer. Anal. **16**, 87–105 (1979)
14. Micchelli, C.A., Rivlin, T.J.: A survey of optimal recovery. In: Estimation in Approximation Theory, pp. 1–54. Plenum, New York (1977)
15. Novak, E.: Deterministic and stochastic error bounds in numerical analysis. In: Lecture Notes in Mathematics, vol. 1349. Springer, Berlin (1988)
16. Novak, E.: On the power of adaption. J. Complexity **12**, 199–237 (1996)

17. Novak, E., Woźniakowski, H.: Tractability of multivariate problems. Volume I: linear information. In: EMS Tracts in Mathematics, vol. 6. European Mathematical Society, Zürich (2008)
18. Novak, E., Woźniakowski, H.: Tractability of multivariate problems. Volume II: standard information for functionals. In: EMS Tracts in Mathematics, vol. 6. European Mathematical Society, Zürich (2010)
19. Novak, E., Woźniakowski, H.: Tractability of multivariate problems. Volume III: standard information for operators. In: EMS Tracts in Mathematics, vol. 6. European Mathematical Society, Zürich (2012)
20. Nussbaum, M.: Spline smoothing in regression model and asymptotic efficiency in l_2. Ann. Stat. **13**, 984–997 (1985)
21. Papageorgiou, A., Wasilkowski, G.: On the average complexity of multivariate problems. J. Complexity **6**, 1–23 (1990)
22. Pinsker, M.S.: Optimal filtering of square integrable signals in Gaussian white noise. Probl. Inform. Transm. **16**, 52–68 (1980). In Russian
23. Plaskota, L.: On average case complexity of linear problems with noisy information. J. Complexity **6**, 199–230 (1990)
24. Plaskota, L.: A note on varying cardinality in the average case setting. J. Complexity **9**, 458–470 (1993)
25. Plaskota, L.: Average case approximation of linear functionals based on information with deterministic noise. J. Comput. Inform. **4**, 21–39 (1994)
26. Plaskota, L.: Average complexity for linear problems in a model with varying noise of information. J. Complexity **11**, 240–264 (1995)
27. Plaskota, L.: How to benefit from noise. J. Complexity **12**, 175–184 (1996)
28. Plaskota, L.: Worst case complexity of problems with random information noise. J. Complexity **12**, 416–439 (1996)
29. Plaskota, L.: Noisy Information and Computational Complexity. Cambridge University Press, Cambridge (1996)
30. Plaskota, L.: Average case uniform approximation in the presence of Gaussian noise. J. Approx. Theory **93**, 501–515 (1998)
31. Smolyak, S.A.: On optimal recovery of functions and functionals of them. Ph.D. thesis, Moscow State University (1965). In Russian
32. Sukharev, A.G.: On the existence of optimal affine methods for approximating linear functionals. J. Complexity **2**, 317–322 (1986)
33. Tikhonov, A.N., Goncharsky A.V., Stepanov, V.V., Yagola, A.G.: Numerical Methods for the Solution of Ill-Posed Problems. Kluwer, Dordrecht (1995)
34. Traub, J.F., Wasilkowski, G.W., Woźniakowski, H.: Information-Based Complexity. Academic, New York (1980)
35. Vakhania, N.N., Tarieladze, V.I., Chobanyan, S.A.: Probability Distributions on Banach Spaces. Reidel, Dordrecht (1987)
36. Wahba, G.: Spline models for observational data. In: CBMS-NSF Series in Applied Mathematics, vol. 59. SIAM, Philadelphia (1990)
37. Wasilkowski, G.W.: Information of varying cardinality. J. Complexity **2**, 204–228 (1986)
38. Werschulz, A.G., Woźniakowski, H.: Are linear algorithms always good for linear problems? Aequations Math. **30**, 202–212 (1986)
39. Woźniakowski, H.: Tractability and strong tractability of linear mulivariate problems. J. Complexity **10**, 96–128 (1994)

Part II
Tutorial

Quasi-Monte Carlo Image Synthesis in a Nutshell

Alexander Keller

Abstract This self-contained tutorial surveys the state of the art in quasi-Monte Carlo rendering algorithms as used for image synthesis in the product design and movie industry. Based on the number theoretic constructions of low discrepancy sequences, it explains techniques to generate light transport paths to connect cameras and light sources. Summing up their contributions on the image plane results in a consistent numerical algorithm, which due to the superior uniformity of low discrepancy sequences often converges faster than its (pseudo-) random counterparts. In addition, its deterministic nature allows for simple and efficient parallelization while guaranteeing exact reproducibility. The underlying techniques of parallel quasi-Monte Carlo integro-approximation, the high speed generation of quasi-Monte Carlo points, treating weak singularities in a robust way, and high performance ray tracing have many applications outside computer graphics, too.

1 Introduction

"One look is worth a thousand words" characterizes best the expressive power of images. Being able to visualize a product in a way that cannot be distinguished from a real photograph before realization can greatly help to win an audience. As ubiquitous in many movies, a sequence of such images can tell whole stories in a captive and convincing way. As a consequence of the growing demand and benefit of synthetic images, a substantial amount of research has been dedicated to finding more efficient rendering algorithms.

The achievable degree of realism depends on the physical correctness of the model and the consistency of the simulation algorithms. While modeling is beyond

A. Keller (✉)
NVIDIA, Fasanenstraße 81, 10623 Berlin, Germany
e-mail: keller.alexander@gmail.com

J. Dick et al. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2012*,
Springer Proceedings in Mathematics & Statistics 65,
DOI 10.1007/978-3-642-41095-6_8,

the focus of this tutorial, we review the fundamentals in Sect. 2. The paradigm of consistency is discussed in the next Sect. 1.1 as it is key to the quasi-Monte Carlo techniques in Sect. 3 that are at the heart of the deterministic rendering algorithms explored in Sect. 4.

On a historical note, the investigation of quasi-Monte Carlo methods in computer graphics goes back to Shirley [69] and Niederreiter [54], and received early industrial attention [60]. This comprehensive tutorial surveys the state of the art, includes new results, and is applicable far beyond computer graphics, as for example in financial mathematics and general radiation transport simulation.

1.1 Why Consistency Matters Most

Analytic solutions in light transport simulation are only available for problems too simple to be of practical relevance, although some of these settings are useful in understanding and testing algorithms [31]. In practical applications, functions are high-dimensional and contain discontinuities that cannot be located efficiently. Therefore approximate solutions are computed using numerical algorithms. In the following paragraphs, we clarify the most important notions, as they are often confused, especially in marketing.

1.1.1 Consistency

Numerical algorithms, whose approximation error vanishes as the sample size increases, are called consistent. Note that consistency is not a statement with respect to the speed of convergence. Within computer graphics, consistency guarantees image synthesis without persistent artifacts such as discretization artifacts introduced by a rendering algorithm; the results are consistent with the input model and in that sense the notion of consistency is understandable without any mathematical background. While many commercial implementations of rendering algorithms required expert knowledge to tweak a big set of parameters until artifacts due to intermediate approximations become invisible, the design of many recent rendering algorithms follows the paradigm of consistency. As a result, users can concentrate on content creation, because light transport simulation has become as simple as pushing the “render”-button in an application.

1.1.2 Unbiased Monte Carlo Algorithms

The bias of an algorithm using random numbers is the difference between the mathematical object and the expectation of the estimator of the mathematical object to be approximated. If this difference is zero, the algorithm is called unbiased. However, this property alone is not sufficient, because an estimator can

be unbiased but not consistent, thus even lacking convergence. In addition, biased but consistent algorithms can handle problems that unbiased algorithms cannot handle: For example, density estimation allows for efficiently handling the problem of "insufficient techniques" (for the details see Sect. 4.4.1).

The theory of many unbiased Monte Carlo algorithms is based on independent random sampling, which is used at the core of many proofs in probability theory and allows for simple parallelization and for estimating the variance as a measure of error.

1.1.3 Physically Based Modeling

Physically based modeling subsumes the creation of input for image synthesis algorithms, where physical entities such as measured data for light sources and optical properties of matter or analytic models thereof are used for the input specification. Modeling with such entities and relying on consistent light transport simulation to many users is much more natural as compared to tweaking lights and materials in order to deliver photorealistic results.

Although often confused in computer graphics, physically correct rendering is not equivalent to unbiased Monte Carlo algorithms: Even non-photorealistic images can be rendered using unbiased Monte Carlo algorithms. In addition, so far none of the physically based algorithms can claim to comply with all the laws of physics, because they are simply not able to efficiently simulate all effects of light transport and therefore cannot be physically correct.

1.1.4 Deterministic Consistent Numerical Algorithms

While independence and unpredictability characterize random numbers, these properties often are undesirable for computer simulations: Independence compromises the speed of convergence and unpredictability disallows the exact repetition of a computer simulation. Mimicking random numbers by pseudo-random numbers generated by deterministic algorithms, computations become exactly repeatable, however, arbitrarily jumping ahead in such sequences as required in scalable parallelization often is inefficient due to the goal of emulating unpredictability.

In fact, deterministic algorithms can produce samples that approximate a given distribution much better than random numbers can. By their deterministic nature, such samples must be correlated and predictable. The lack of independence is not an issue, because independence is not visible in an average anyhow and consistency can be shown using number theoretic arguments instead of probabilistic ones. In addition, partitioning such sets of samples and leaping in such sequences of samples can be highly efficient.

As it will be shown throughout the article, advantages of such deterministic consistent numerical algorithms are improved convergence, exact reproducibility, and simple communication-avoiding parallelization. Besides rendering physically

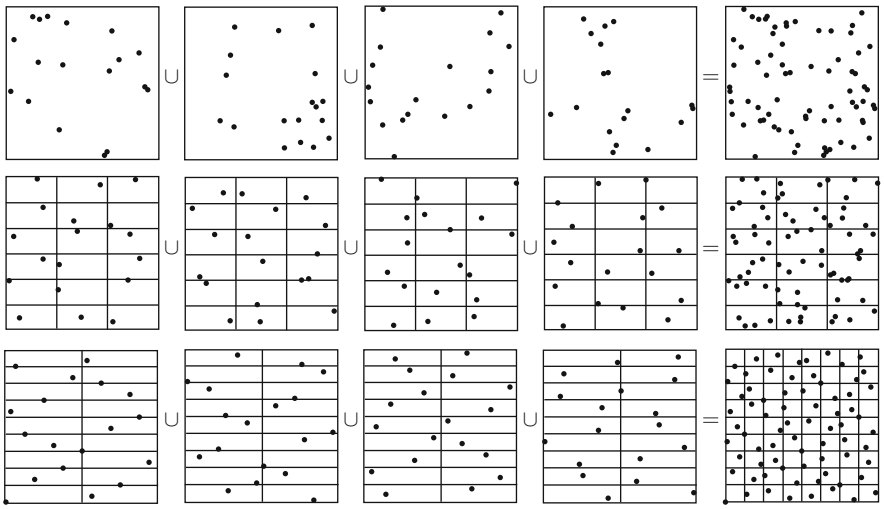


Fig. 1 Illustration of the difference between unbiased and deterministic consistent uniform sampling: The *top row* shows four independent sets of 18 points each and their union as generated by a pseudo-random number generator. The *middle row* shows independent realizations of so-called stratified samples with their union that result from uniformly partitioning the domain and independently sampling inside each resulting interval in order to increase uniformity. However, points can come arbitrarily close together along interval boundaries and there is no guarantee for their union to improve upon uniformity. The *bottom row* shows the union of four contiguous blocks of 18 points of the Halton sequence. As opposed to the pseudorandom number generator and stratified sampling, the samples of the Halton sequence are more uniform, nicely complement each other in the union, and provide a guaranteed minimum distance and intrinsic stratification along the sequence.

based models, these methods also apply to rendering non-physical models that often are chosen to access artistic freedom or to speed up the rendering process. The illustration in Fig. 1 provides some initial intuition of the concepts and facts discussed in this section.

2 Principles of Light Transport Simulation

Implementing the process of taking a photo on a computer involves the simulation of light transport. This in turn requires a mathematical model of the world: A boundary representation with attached optical properties describes the surfaces of the objects to be visualized. Such a model may be augmented by the optical properties of volumes, spectral properties, consideration of interference, and many more physical phenomena. Once the optical properties of the camera system and the light sources are provided, the problem specification is complete.

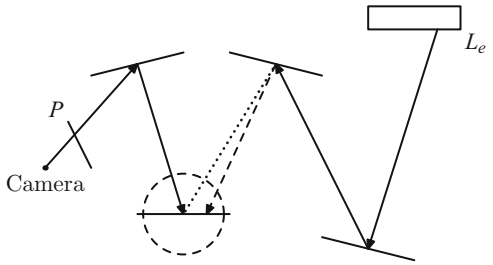


Fig. 2 Bidirectional generation of light transport paths: A path segment started from the camera and a path segment started from a light source L_e can be connected by a shadow ray (*dotted line*, see Sect. 4.4.2), which checks whether the vertices to connect are mutually visible. Alternatively, the basic idea of photon mapping (see Sect. 4.4.1) is to relax the precise visibility check by allowing for a connection of both path segments if their end points are sufficiently close as indicated by the *dashed circle*. Both techniques are illustrated for identical path length, which is the reason for the dashed prolongation of the light path segment for photon mapping.

The principles of light transport simulation are well covered in classic textbooks on computer graphics: Currently, [66] is the most updated standard reference, [16] is a classic reference available for free on the internet, and [70] can be considered a primer and kick start. Recent research is well surveyed in [6, 22, 82] along with profound investigations of numerical algorithms and their issues.

2.1 Light Transport Along Paths

Light transport simulation consists of identifying all paths that connect cameras and light sources and integrating their contribution to form the synthetic image. Figure 2 illustrates the principles of exploring path space.

One way of generating light transport paths is to follow the trajectories of photons emitted from the light sources along straight line segments between the interactions with matter. However, no computational device can simulate a number of photons sufficiently large to represent reality and hence the direct simulation often is not efficient.

When applicable, light transport paths can be reversed due to the Helmholtz reciprocity principle and trajectories can be traced starting from the camera sensor or eye. Most efficient algorithms connect such camera and light path segments and therefore are called bidirectional.

Vertices of paths can be connected by checking their mutual visibility with respect to a straight line or by checking their mutual distance with respect to a suitable metric. While checking the mutual visibility is precise, it does not allow for efficiently simulating some important contributions of light caused by surfaces that are highly specular and/or transmissive, which is known as the problem of

insufficient techniques [42]. In such cases, connecting paths by merging two vertices that are sufficiently close helps. The resulting bias can be controlled by the maximum distance allowed for merging vertices.

The interactions with matter need to be modeled: Bidirectional scattering distribution functions (BSDFs) describe the properties of optical interfaces, while scattering and absorption cross sections determine when to scatter in volume using the distribution given by a phase function [66]. Similarly, the optical properties of the light sources and sensors have to be mathematically modeled. For cameras, models range from a simple pinhole to complete lenses allowing for the simulation of depth of field and motion blur. Light sources often are characterized by so-called light profiles. All these physical properties can be provided in measured form, too, which in many cases provides quality superior to the current analytic models.

Beyond that, optical properties can be modeled as functions of wavelength across the spectrum of light in order to overcome the restriction of the common approach using only three selected wavelengths to represent red, green, and blue and to enable dispersion and fluorescence. The simulation of effects due to polarization and the wave character of light are possible to a certain extent, however, are subject to active research.

While modeling with real entities is very intuitive, it must be noted that certain violations of physics can greatly help the efficiency of rendering and/or help telling stories at the cost of systematic errors.

2.2 Accelerated Ray Tracing and Visibility

The boundary of the scene often is stored as a directed acyclic graph, which allows for referencing parts of the scene multiple times to instance them at multiple positions in favor of a compact representation. Complex geometry like for example hair, fur, foliage, or crowds often are generated procedurally, in which case the call graph implicitly represents the scene graph. Triangles, quadrangles, or multi-resolution surfaces, which include subdivision surfaces, are the most common geometric primitives used for boundary representation.

The vertices of a light transport path are connected by straight line segments. First, these can be found by tracing rays from a point x into a direction ω to identify the closest point of intersection $h(x, \omega)$ with the scene boundary. A second way to construct paths is to connect two vertices x and y of two different path segments. This can be accomplished by checking the mutual visibility $V(x, y)$, which is zero if the straight line of sight between the points x and y, a so-called shadow ray, is occluded, one otherwise. As a third operation, two vertices can be merged, if their distance with respect to a metric is less than a threshold. Efficient implementations of the three operations all are based on hierarchal culling (see [35, 39] for a very basic primer).

In order to accelerate ray tracing, the list of objects and/or space are recursively partitioned. Given a ray to be traced, traversal is started from the root node

descending into a subtree, whenever the ray intersects this part of the scene. Most parts of the scene thus are hierarchically culled and never touched. In case the cost of the construction of such an auxiliary acceleration hierarchy can be amortized over tracing many paths, it makes sense to store it partially or completely. Checking the mutual visibility by a shadow ray is even more efficient, since the traversal can be stopped upon any intersection with the boundary, while tracing a ray requires to find the intersection closest to its origin.

Efficiently merging vertices follows the same principle of hierarchical culling [35]: Given two sets of points in space, the points of the one set that are at a maximum given distance from the points of the other set are found by hierarchically subdividing space and pruning the search for partitions of space that cannot overlap within the given distance.

3 Principles of Quasi-Monte Carlo Integro-Approximation

Image synthesis can be considered an integro-approximation problem of the form

$$g(\mathbf{y}) := \int_X f(\mathbf{x}, \mathbf{y}) d\mu(\mathbf{x}) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(\mathbf{x}_i, \mathbf{y}), \tag{1}$$

where $f(\mathbf{x}, \mathbf{y})$ is the measurement contribution to a location $\mathbf{y}$ by a light transport path identified by $\mathbf{x}$. We will focus on deterministic linear algorithms [78] to consistently determine the whole image function g for all pixels $\mathbf{y}$ using one low discrepancy sequence $\mathbf{x}_i$ of deterministic sample points. The principles of such quasi-Monte Carlo methods have been introduced to a wide audience in [55], which started a series of MCQMC conferences, whose proceedings contain almost all recent developments in quasi-Monte Carlo methods. Many of the results and developments are summarized in recent books [10, 49, 72].

Before reviewing the algorithms to generate low discrepancy sequences in Sect. 3.3 and techniques resulting from their number theoretic construction in Sect. 3.4, error bounds are discussed with respect to measures of uniformity.

3.1 Uniform Sampling, Stratification, and Discrete Density Approximation

A common way to generate a discrete approximation of a density comprises the creation of uniformly distributed samples that are transformed [9, 25]. For many such transformations, an improved uniformity results in a better discrete density approximation. Measures of uniformity often follow from proofs of error bounds (see the next paragraph) as a result of the attempt to bound the error by a product of

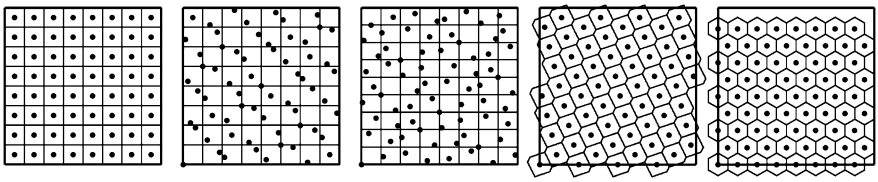


Fig. 3 Examples of $(\mathscr{M},\mu)$-uniform point sets with an outline of their partition $\mathscr{M}$ (from *left* to *right*): $n = 8 \times 8$ points of a Cartesian grid, the first $n = 64$ points of the Sobol' sequence, the first $n = 72$ points of the Halton sequence, and the maximized minimum distance rank-1 lattice with $n = 64$ points and generator vector $\mathbf{g} = (1, 28)$. The hexagonal grid with $n = 72$ point is shown for comparison, as it cannot tile the unit square due to its irrational basis.

properties of the sampling points and the function as used in for example Theorem 1. For the setting of computer graphics, where X is a domain of integration, $\mathscr{B}$ are the Borel sets over X, and μ the Lebesgue measure, a practical measure of uniformity is given by

Definition 1 (see [56]). Let $(X,\mathscr{B},\mu)$ be an arbitrary probability space and let $\mathscr{M}$ be a nonempty subset of $\mathscr{B}$. A point set P_n of n elements of X is called $(\mathscr{M},\mu)$-uniform if

$$\sum_{i=0}^{n-1} \chi_M(\mathbf{x}_i) = \mu(M)\cdot n \qquad \text{for all } M \in \mathscr{M},$$

where $\chi_M(\mathbf{x}_i) = 1$ if $\mathbf{x}_i \in M$, zero otherwise.

Figure 3 shows examples of $(\mathscr{M},\mu)$-uniform points from $X = [0,1)^2$ that obviously can only exist if the measures $\mu(M)$ are rational numbers with the same denominator n [56]. While the subset $\mathscr{M}$ may consist of the Voronoi regions of a lattice, it also may consist of axis aligned intervals of the form given by

Definition 2 (see [56]). An interval of the form

$$E(p_1,\ldots,p_s) := \prod_{j=1}^{s} \left[\frac{p_j}{b_j^{d_j}}, \frac{p_j+1}{b_j^{d_j}} \right) \subseteq [0,1)^s$$

for $0 \le p_j < b_j^{d_j}$ and integers $b_j, d_j \ge 0$ is called an *elementary interval*.

As compared to the original definition in [55, p. 48], which considers the special case of b-adic intervals, i.e. $b_j = b$ (for $b_j = 2$, the intervals are called dyadic), different bases b_j are allowed for each dimension j to include a wider variety of point sets [52, 56]. Representing numbers in base b_j, d_j can be thought of as the number of digits and fixes the resolution in dimension j, which allows for specifying an elementary interval by its coordinates $p_1,\ldots,p_s$.

Characterizations of uniformity beyond the stratification properties imposed by $(\mathscr{M},\mu)$-uniformity (see Fig. 3) include the maximum minimum distance

$$d_{\min}(P_n) := \min_{0 \le i < j < n} \|\mathbf{x}_j - \mathbf{x}_i\|_T$$

of the points on the torus $T = [0,1)^s$ [33], and their deviation from uniformity measured by various kinds of discrepancy [55].

In many applications, uniform points are transformed to approximate a continuous density. The quality of such a discrete density approximation can be judged by the star-discrepancy

$$D^*(p, P_n) := \sup_{A = \prod_{j=1}^{s} [0, a_j) \subset [0,1)^s} \left| \int_{[0,1)^s} \chi_A(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} - \frac{1}{n} \sum_{i=0}^{n-1} \chi_A(\mathbf{x}_i) \right|$$

with respect to the density p [25]. Discrepancies can be understood as integration errors, where the exact measure of a test set A with respect to p is compared to the average number of points in that set. For $p \equiv 1$, we have the so-called star-discrepancy $D^*(P_n) := D^*(1, P_n,)$ [55], which is of central importance for quasi-Monte Carlo methods: Low discrepancy point sequences have $D^*(P_n) \in \mathcal{O}\left(\frac{\log^s n}{n}\right)$, while uniform random numbers can only achieve an order of $\mathcal{O}\left(\sqrt{\frac{\log \log n}{n}}\right)$ manifesting the asymptotic inferiority of random sampling with respect to discrete density approximation.

3.2 Error Bounds

Using $(\mathcal{M}, \mu)$-uniformity, an error bound for the integro-approximation problem in Eq. 1 is given by

Theorem 1 (see [33]). *Let $(X, \mathcal{B}, \mu)$ be an arbitrary probability space and let $\mathcal{M} = \{M_1, \ldots, M_k\}$ be a partition of X with $M_j \in \mathcal{B}$ for $1 \le j \le k$. Then for any $(\mathcal{M}, \mu)$-uniform point set $\mathcal{P} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and any bounded function f, which restricted to X is μ-integrable, we have*

$$\left\| \frac{1}{n} \sum_{i=0}^{n-1} f(\mathbf{x}_i, \mathbf{y}) - \int_X f(\mathbf{x}, \mathbf{y}) d\mu(\mathbf{x}) \right\| \le \sum_{j=1}^{k} \mu(M_j) \left\| \sup_{\mathbf{x} \in M_j} f(\mathbf{x}, \mathbf{y}) - \inf_{\mathbf{x} \in M_j} f(\mathbf{x}, \mathbf{y}) \right\|$$

for any suitable norm $\|\cdot\|$.

In analogy to the Monte Carlo case [15], the above theorem has been derived in order to prove the convergence of quasi-Monte Carlo methods for Eq. 1 in the setting of computer graphics, where the only properties of f that are easily accessible are square integrability and boundedness. By omitting $\mathbf{y}$, the above theorem reduces to an error bound for quasi-Monte Carlo integration as originally developed in

[56, Theorem 2], which improved the derivation and results obtained for Riemann integrable functions [24].

Other than trivial worst case bounds, the theorem does not provide a rate of convergence, which is the price for its generality. However, including more knowledge about the function f by restricting the function class allows one to obtain much better error bounds along with measures of uniformity: The Koksma-Hlawka inequality [55] bounds the error by a product of the star discrepancy of the point set and the variation of the function in the sense of Hardy and Krause, bounds for functions with sufficiently fast decaying Fourier coefficients are found in [72], and the error for integrating Lipschitz functions can be bounded by a product of the Lipschitz constant and the maximum minimum distance of a lattice of points [8].

While quasi-Monte Carlo methods allow for improved convergence rates as compared to Monte Carlo methods, the variance of an estimate cannot be consistently computed due to the lack of independence. As a compromise to this issue, randomized quasi Monte Carlo methods [4, 62] have been introduced that sacrifice some uniformity of the sample points in order to control adaptive termination by unbiased variance estimation. As we focus on deterministic algorithms only, this is not an option and we refer to a deterministic variant of termination by comparing differences of norms of intermediate results as introduced in [65]. In computer graphics such norms should reflect the properties of the human visual system and often the L^2-norm [11, Sect. 3.5] is appropriate to measure error.

3.3 Algorithms for Low Discrepancy Sequences

Most known constructions of low discrepancy sequences imply sequences of $(\mathscr{M}, \mu)$-uniform point sets (see [56, Remark 1] and [33]) that guarantee Eq. 1 to converge. In the following, such mappings from $\mathbb{N}_0$ into the s-dimensional unit cube $[0, 1)^s$ are surveyed with respect to their algorithmic principles that enable the techniques reviewed in Sect. 3.4. Note that the Weyl sequence [85] is irrational and as such cannot fulfill the condition of $(\mathscr{M}, \mu)$-uniformity. It is therefore excluded from our considerations, since most proofs, and especially computers, rely on rational numbers.

3.3.1 Radical Inversion

A digital radical inverse

$$\Phi_{b,C} : \mathbb{N}_0 \to \mathbb{Q} \cap [0, 1)$$

$$i = \sum_{l=0}^{M-1} a_l(i) b^l \mapsto \left(b^{-1} \cdots b^{-M}\right) \left[C \begin{pmatrix} a_0(i) \\ \vdots \\ a_{M-1}(i) \end{pmatrix} \right] \tag{2}$$

in a prime power base b is computed using a generator matrix C, where the matrix-vector multiplications are performed in the finite field $\mathbb{F}_b$ (for the theory and mappings from and to $\mathbb{F}_b$ see [55]). While in theory these matrices are infinite-dimensional, in practice they are finite due to the finite precision of computer arithmetic. The inverse mapping $\Phi_{b,C}^{-1}$ exists, if C is regular. M is the number of digits, which allows for generating up to $N = b^M$ points.

The time to compute digital radical inverses is far from negligible in many applications. Efficient implementations use tables of precomputed terms [14], take advantage of bit vector arithmetic in $b = 2$ [82], or enumerate the radical inverses using the Gray-code order [67]. Cancellation errors of floating point arithmetic are avoided by ordering summations and computing in integers as long as possible. Note that already the conversion to floating point numbers by the multiplication with $\left(b^{-1} \cdots b^{-M}\right)$ causes collisions of numbers that were different in integer representation.

Van der Corput Sequence

Selecting the identity matrix I as generator matrix results in the points $\Phi_b(i) := \Phi_{b,I}(i) = \sum_{l=0}^{\infty} a_l(i) b^{-l-1}$ of the van der Corput sequence, which is the simplest radical inverse. The mapping reflects the digits $a_l(i)$ of the index i represented in base b at the decimal point. Obviously the computation is finite, as i has only finitely many digits $a_l(i) \neq 0$.

For $0 \leq i < b^m$, the mapping $b^m \Phi_b(i)$ is a permutation and hence the first b^m points of the sequence $\Phi_b(i)$ are equidistantly spaced with a distance of $\frac{1}{b^m}$. Furthermore, this implies that partitioning the van der Corput sequence into contiguous blocks of length b^m, the integer parts of the points within each block multiplied by b^m must be permutations, too. Many of the techniques described in this article rely on these properties and their generalizations.

Another interesting property of the van der Corput sequence is its intrinsic stratification [33]: For example, $\Phi_2(i) < \frac{1}{2}$ for even i and $\Phi_2(i) \geq \frac{1}{2}$ otherwise. In general,

$$\Phi_b(k + l \cdot b^m) \in [\Phi_b(k), \Phi_b(k) + b^{-m}) \text{ for } l \in \mathbb{N}_0.$$

While this property is very useful, it also is the reason why pseudo-random number generators cannot just be replaced by the van der Corput sequence (and radical inverses in general): Already a two-dimensional vector assembled by subsequent numbers from the van der Corput sequence is not uniformly distributed.

3.3.2 Scrambling

Scrambling a set of points on $H = [0, 1)$ comprises the following steps:

1. Partition H into b equal intervals $H_1, H_2, \ldots, H_b$.
2. Permute these intervals.
3. For $h \in \{1, 2, \ldots, b\}$, recursively repeat the procedure starting out with $H = H_h$.

Formalizing the scrambling of the i-th point of a sequence represented in base b as defined in Eq. 2 yields the scrambled digits

$$\begin{aligned} a'_{i,0} &:= \pi\,(a_0(i)) \\ a'_{i,1} &:= \pi_{a_0(i)}\,(a_1(i)) \\ &\vdots \\ a'_{i,l} &:= \pi_{a_0(i),a_1(i),\ldots,a_{l-1}(i)}\,(a_l(i))\,, \\ &\vdots \end{aligned}$$

where the l-th permutation $\pi_{a_0(i),a_1(i),\ldots,a_{l-1}(i)} : \{0, \ldots, b-1\} \to \{0, \ldots, b-1\}$ depends on the $l-1$ leading digits $a_0(i), a_1(i), \ldots, a_{l-1}(i)$. The mapping is bijective, because it is based on the sequential application of permutations.

While obviously this procedure becomes finite by the finite precision of computation, uniformly distributed points are mapped to uniformly distributed points. Originally, these properties combined with random permutations were introduced to randomize uniform points sets [62,63]. However, many deterministic optimizations of low discrepancy sequences in fact can be represented as scramblings with deterministic permutations. Note that using a regular generator matrix $C \neq I$ in Eq. 2 already can be considered a deterministic scrambling of the van der Corput sequence.

3.3.3 Halton Sequence and Hammersley Points

The Halton sequence [21]

$$\mathbf{x}_i = (\Phi_{b_1}(i), \ldots, \Phi_{b_s}(i))$$

has been constructed by using one van der Corput sequence for each component, where the bases b_j are relatively prime. Replacing one of the components by $\frac{i}{n}$ results in n points that form the Hammersley point set. As compared to the Halton sequence, where by construction subsequent points fill the largest holes in space,

the Hammersley points are even more uniformly distributed, however, at the price of not being extensible.

Although the Halton sequence is of low discrepancy, it has the undesirable property that projections are not as well distributed as they could be: For example, the first $\min\{b_1, b_2\}$ points of a two-dimensional Halton sequence $(\Phi_{b_1}(i), \Phi_{b_2}(i))$ lie on a straight line through the origin. Similar linear alignments appear over and over again in the sequence and the b_j can be large for high dimensional projections.

Therefore many improvements of the Halton sequence have been developed. In fact, all of them turn out to be deterministic scramblings (see Sect. 3.3.2): For example, Zaremba [88] used the simple permutation $\pi_{b_j}(a_l(i)) = (a_l(i) + l) \bmod b_j$ instead of directly using the digits $a_l(i)$ and later Faure [13] developed a set of permutations generalizing and improving Zaremba's results. A very efficient implementation can be found at http://gruenschloss.org/halton/halton.zip. While the modifications improve the constant of the order of discrepancy, they also improve upon the minimum distance [33].

Whenever the number of samples $n = \prod_{j=1}^{s} b_j^{n_j}$ is a product of power of the bases, the Halton sequence (including all its variants) is fully stratified.

3.3.4 Digital (t, s)-Sequences and (t, m, s)-Nets

Low discrepancy sequences can also be constructed from radical inverses using the same base $b_j = b$. They are based on b-adic elementary intervals as covered by Definition 2:

Definition 3 (see [55, Definition 4.1]). For integers $0 \leq t \leq m$, a (t, m, s)-*net in base* b is a point set of b^m points in $[0, 1)^s$ such that there are exactly b^t points in each b-adic elementary interval E with volume b^{t-m}.

Definition 4 (see [55, Definition 4.2]). For an integer $t \geq 0$, a sequence $\mathbf{x}_0, \mathbf{x}_1, \ldots$ of points in $[0, 1)^s$ is a (t, s)-*sequence in base* b if, for all integers $k \geq 0$ and $m > t$, the point set $\mathbf{x}_{kb^m}, \ldots, \mathbf{x}_{(k+1)b^m-1}$ is a (t, m, s)-net in base b.

The elementary intervals from Definition 2 use a resolution of b^{d_j} along dimension j. For a (t, m, s)-net in base b we then have $\sum_{j=1}^{s} d_j = m - t$, which relates the number of points determined by m and the quality parameter t. Since scrambling (see Sect. 3.3.2) permutes elementary intervals, it does not change the t parameter. Similar to the Halton sequence, any (t, s)-sequence can be transformed into a $(t, m, s + 1)$-net by concatenating a component $\frac{i}{b^m}$ [55].

According to Definition 3, a $(0, s)$-sequence is a sequence of $(0, m, s)$-nets, similar to what is illustrated for the Halton sequence in Fig. 1. This especially includes $(0, ms, s)$-nets, where in each hypercube-shaped elementary interval of side length b^{-m}, there is exactly one point. As the number of points of $(0, ms, s)$-nets is exponential in the dimension, this construction is only feasible in small dimensions.

(0, 1)-Sequences in Base b

The simplest example of a $(0, 1)$-sequence in base b is the van der Corput sequence. For regular generator matrices C, the radical inverses in Eq. 2 are $(0, 1)$-sequences, too, and Definition 4 guarantees all properties of the van der Corput sequence as described above for the more general $(0, 1)$-sequences.

Constructions of (t, s)-Sequences for $s > 1$

The digital construction

$$\mathbf{x}_i = (\Phi_{b,C_1}(i), \ldots, \Phi_{b,C_s}(i))$$

of (t, s)-sequence is based on the radical inverses from Eq. 2 with identical base b and consequently different generator matrices C_j for each coordinate j.

The most popular (t, s)-sequence is the Sobol' sequence [73] in base $b = 2$, because it can be implemented efficiently using bit-vector operations [17, 43, 82] to compute the radical inverses. The sequence can be constructed for any dimension and in fact each component is a $(0, 1)$-sequence in base 2 itself. Due to the properties of $(0, 1)$-sequences, the Sobol' sequence at $n = 2^m$ samples must be a Latin hypercube sample [61]. Other than Latin hypercube samples based on random permutations that would have to be stored in $\mathcal{O}(sn)$ memory, the permutations generated by the Sobol' sequence are infinite, can be computed on demand without storing them, and are guaranteed to be of low discrepancy. A description of how to compute the binary generator matrices can be found in [29, 30] and one good set of matrices can be downloaded at http://web.maths.unsw.edu.au/~fkuo/sobol/. In [75] Sobol' et al. introduced additional criteria for the selection of the generator matrices.

As the first two components of the Sobol' sequence form a $(0, 2)$-sequence in base 2, the first 2^{2m} two-dimensional points must be stratified such that there is exactly one point in each voxel of a $2^m \times 2^m$ regular grid over $[0, 1)^2$. This structure is very useful in image synthesis (see Sect. 4.1).

Since $(0, s)$-sequences can only exist for $s \leq b$ [55, Corollary 4.24], Faure [12] generalized Sobol's construction to higher bases. Following Sobol's idea, each component is constructed as $(0, 1)$-sequence. In fact both Sobol's and Faure's construction yield upper triangular generator matrices.

The construction of better generator matrices is an ongoing effort and various approaches have been taken [26]. In fact, there exist (t, m, s)-nets, which cannot be generated by radical inverses [18, Sect. 3]. This in connection with the observation that scrambling often improves the uniformity properties [33] of low discrepancy points alludes to conjecture that there are better low discrepancy sequences that are generated by general permutations instead of only generator matrices as in Eq. 2.

3.3.5 Rank-1 Lattice Sequences and Rank-1 Lattices

For a suitable generator vector $\mathbf{g} = (g_1, \ldots, g_s)$, rank-1 lattice sequences [23,50,51]

$$\mathbf{x}_i = \Phi_b(i)(g_1, \ldots, g_s) \bmod 1 \in (\mathbb{Q} \cap [0,1))^s$$

provide the simplest algorithm for generating a low discrepancy sequence in s dimensions. While in theory the components $g_j = \sum_{m=0}^{\infty} g_{j,m} b^m$ of the generator vector are represented by infinite sequences of digits $g_{j,m} \in \{0, \ldots, b-1\}$, in practice the components can be represented by positive integers due to the finite precision of computer arithmetic. Yet, there only exists a tiny number of constructions for the generator vectors [3, 84] and usually good generator vectors result from exhaustive computer searches [2]. An implementation of a variety of such methods is described in [48].

Lattice sequences resemble (t,s)-sequences, as contiguous blocks of b^m points form lattices, where the first lattice is anchored in the origin and the subsequent lattices are shifted copies. For $\gcd(g_i, b^m) = 1$ rank-1 lattices are instances of a Latin hypercube sample, which in addition provides a trivial lower bound on the minimum distance, because the one-dimensional projections are equidistantly spaced at $\frac{1}{b^m}$.

By allowing only generator vectors of the form $\mathbf{g} = (a^0, a^1, a^2, \ldots, a^{s-1})$, Korobov restricted the search space to one integer $a \in \mathbb{N}$ [72]. Note that for suitable a and b^m, the generator vector coincides with a multiplicative linear congruential pseudo-random number generator.

Hybrid Sequences

Besides the common properties of especially the Sobol' (t,s)-sequence and rank-1 lattice sequences in base $b = 2$, there even exist rank-1 lattices that are $(0,2,2)$-nets [8, Sect. 2.1]. There is an even closer relationship as stated by

Theorem 2. *Given b and g_j are relatively prime, the component $\Phi_b(i)g_j \bmod 1$ of a rank-1 lattice sequence is a $(0,1)$-sequence in base b.*

Proof. Φ_b is a $(0,1)$-sequence [55] and by Definition 4 each contiguous block of b^m points is a $(0,m,1)$-net in base b. As a consequence, the integer parts of such a $(0,m,1)$-net multiplied by b^m are a permutation. If now b and g_j are relatively prime, then for such a $(0,m,1)$-net the integers $g_j \lfloor b^m \Phi_b(i) \rfloor \bmod b^m$ form a permutation, too. Hence $\Phi_b(i)g_j \bmod 1$ is a $(0,1)$-sequence in base b. □

If now the generator matrix C is regular, a permutation exists that maps the elements of any $(0,m,1)$-net of $\Phi_b(i)g_j \bmod 1$ to $\Phi_{b,C}(i)$ and consequently $\Phi_{b,C}(i)$ and $\Phi_b(i)g_j \bmod 1$ are scrambled (see Sect. 3.3.2) versions of each other.

This close relationship allows one to combine components of (t,s)-sequences in base b with components of rank-1 lattice sequences using a radical inverse in base b.

While this is of theoretical interest [45,46], it also is of practical interest, especially in computer graphics: Rank-1 lattice sequences are cheap to evaluate, while (t,s)-sequences use the structure of b-adic elementary intervals [83].

3.4 Algorithms for Enumerating Low Discrepancy Sequences

The properties of radical inversion allow for enumerating low discrepancy sequences in different ways that are very useful building blocks of quasi-Monte Carlo methods. The enumeration schemes can be derived by equivalence transformations of integrals.

3.4.1 Enumeration in Elementary Intervals

Both the Halton and (t,s)-sequences are stratificatied with respect to elementary intervals (see Definition 2 and [52]). In [19] methods have been developed to efficiently enumerate the samples in a given elementary interval: Restricting the sequences to a given elementary interval yields a system of equations, whose solution results in an enumeration algorithm.

As the construction of the Halton sequence is based on the Chinese remainder theorem [21], enumerating the Halton sequence restricted to an elementary interval requires to solve a system of congruences. The solution of this system yields the indices $i + t \cdot \prod_{j=1}^{s} b_j^{d_j}, t \in \mathbb{N}_0$ to enumerate the Halton points in a given elementary interval. The initial offset i is uniquely identified by that elementary interval, while the subsequent points are found by jumping along the sequence with a stride that is a product of the prime powers of the bases b_j, where d_j fixes the resolution along dimension j.

For (t,s)-sequences, the system of linear equations is assembled by solving Eq. 2 for each dimension j for $M = d_j$ digits, where the d_j specify the size of the elementary interval as defined in Definition 2. The righthand side of the equation system then is given by each the first d_j digits of the coordinates p_j of the elementary interval and the number q of the point to be computed. In an implementation, the inverse system matrix can be stored and enumerating the points of an elementary interval is as expensive as computing the points of a (t,s)-sequence (see the code at http://gruenschloss.org/sample-enum/sample-enum-src.zip).

Typical applications of enumerating samples per elementary interval are problems, where the structure matches the stratification implied by elementary intervals. Such problems include integro-approximation and adaptive sampling [19, 40, 59], where the number of samples needs to be controlled per elementary interval. Enumerating samples per elementary interval also is a strategy for parallelization [19].

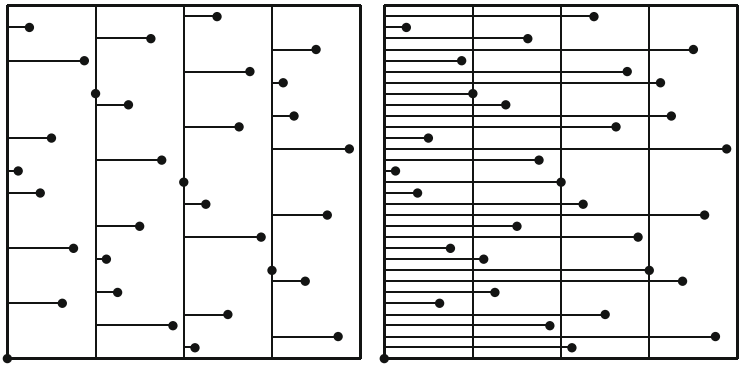


Fig. 4 Illustration of partitioned and nested low discrepancy sequences using the first 32 points of the Sobol' sequence. *Left*: Partitioning an $s+1$-dimensional low discrepancy sequence by its first component (along the x-axis) results in each one low discrepancy sequence in s dimensions as illustrated by the projections onto the partitioning lines parallel to the y-axis. *Right*: Thresholding the first component results in nested s-dimensional low discrepancy sequences, where each sequence with a smaller threshold is included in a sequence with a larger threshold.

3.4.2 Partitioning Low Discrepancy Sequences

Restricting a low discrepancy sequence to an axis-aligned subinterval does not change its order of discrepancy [55]. Similarly the order of discrepancy is not changed by omitting dimensions, i.e. projecting the points along canonical axis.

Using a characteristic function

$$\chi_j(x') := \begin{cases} 1 & j \le x' < j+1 \\ 0 & \text{otherwise,} \end{cases}$$

the equivalence transformation

$$\int_{[0,1)^s} f(\mathbf{x})\, d\mathbf{x} = \sum_{j=0}^{b^m-1} \int_{[0,1)} \int_{[0,1)^s} \chi_j(b^m \cdot x') \cdot f(\mathbf{x})\, d\mathbf{x}\, dx'$$

identifies the point set

$$P_j := \left\{\mathbf{x}_i : \chi_j(b^m \cdot x_{i,c}) = 1, i \in \mathbb{N}_0\right\} = \{\mathbf{x}_i : j \le b^m \cdot x_{i,c} < j+1, i \in \mathbb{N}_0\}$$

used to integrate the j-th summand when applying one $s+1$ dimensional quasi-Monte Carlo point sequence $(\mathbf{x}_i)_{i\ge 0}$ for integral estimation, where $x_{i,c}$ is the c-th component of the point $\mathbf{x}_i$. Enumerating the subsequences

$$P_{\Phi_b^{-1}(j/b^m)} = \left\{\mathbf{x}_{l\cdot b^m+j} : l \in \mathbb{N}_0\right\}$$

of a sequence partitioned along a component, which is a radical inverse, is as simple as leaping through the sequence with a stride of b^m elements and an offset j [36]. As mentioned before, the P_j must be of low discrepancy, too. The method is illustrated in Fig. 4, where Φ_2 is used to partition the two-dimensional Sobol' sequence.

It is important to note that the partitioning component must not be used to sample the integrand, because computations may diverge, as explained in [36]: This extra dimension is used to partition an $s + 1$-dimensional low discrepancy sequence into b^m s-dimensional low discrepancy sequences.

The main application of this scheme are communication-avoiding parallel quasi-Monte Carlo methods: Each thread, process, or job is assigned its own subsequence. Upon the reduction of the partial results, the ensemble of all samples forms the original low discrepancy sequence without any intermediate communication. Even if each thread, process, or job terminates adaptively, on the average the number of points consumed in each thread of process will be similar due to the low discrepancy of each of the subsequences. Due to the partitioning property the result is even independent of the number of processing elements; parallel or sequential execution yield identical results.

3.4.3 Nested Low Discrepancy Sequences

Similar to partitioning low discrepancy sequences, nested s-dimensional low discrepancy sequences are obtained by thresholding an additional component. As illustrated in Fig. 4, the threshold determines the fraction of samples selected from the original sequence. The sequences are nested in the sense that sequences resulting from a smaller threshold are always included in sequences resulting from a larger threshold.

Nested sequences can be used in consistent algorithms, where several problems use the samples of one sequence. Depending on the single problem, a threshold can be selected to control what fraction of samples is consumed. Similar to the previous section, the nested sequences can be enumerated by leaping with a stride of b^m for a threshold b^{-m}.

3.4.4 Splitting

If a problem is less sensitive in some dimensions as compared to others, efficiency often can be increased by concentrating samples in the more important dimensions of the problem. Trajectory splitting is one such technique that after a certain path length splits one particle into multiple and follows their individual trajectories as illustrated in Fig. 5.

The principle of a very simple and efficient quasi-Monte Carlo algorithm for trajectory splitting is based on rewriting the integral of f

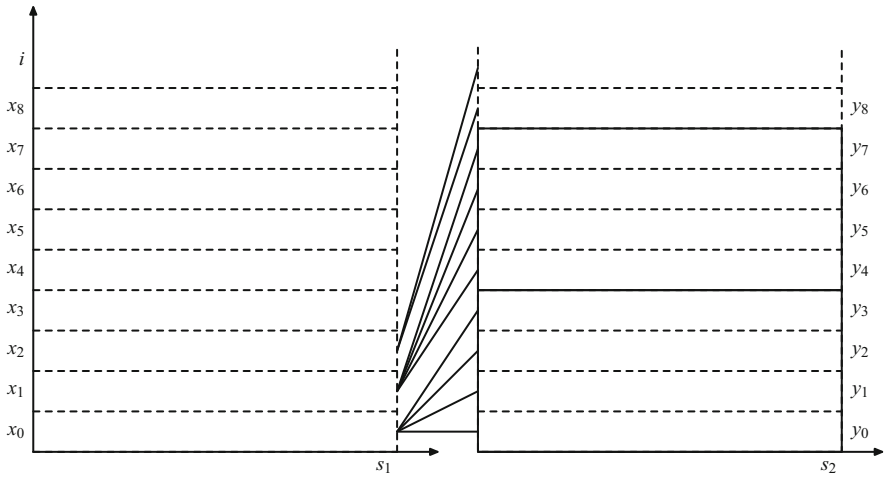


Fig. 5 Splitting can increase the efficiency by sampling the dimensions of y more than the dimensions of x. Using one low discrepancy sequence $(\mathbf{x}_i, \mathbf{y}_i)$, the dimensions of $\mathbf{x}_i$ are enumerated slower by a fixed factor as compared to the dimensions of $\mathbf{y}_i$.

$$\int_{[0,1)^s} f(\mathbf{x}, t)\, dt\, d\mathbf{x} = \int_{[0,1)^s} \sum_{j=0}^{b^m-1} \chi_{[\frac{j}{b^m}, \frac{j+1}{b^m})}(t)\, f(\mathbf{x}, b^m t - j)\, dt\, d\mathbf{x}$$

as an integral of b^m copies of f with respect to the dimension t, where the characteristic function $\chi_A(t)$ is one if $t \in A$ and zero otherwise. Applying a low discrepancy sequence $(\mathbf{x}_i, t_i)$, where the component t_i is a $(0, 1)$-sequence generated by an identity matrix scaled by an element from $\mathbb{F}_b \setminus \{0\}$, to compute the righthand side of the equivalence transformation yields:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \sum_{j=0}^{b^m-1} \chi_{[\frac{j}{b^m}, \frac{j+1}{b^m})}(t_i)\, f(\mathbf{x}_i, b^m t_i - j) \tag{3}$$

$$= \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \sum_{j=0}^{b^m-1} \chi_{[\frac{j}{b^m}, \frac{j+1}{b^m})}(\Phi_b(i))\, f(\mathbf{x}_i, b^m \Phi_b(i) - j)$$

$$= \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(\mathbf{x}_i, \Phi_b(\lfloor i/b^m \rfloor)) \underbrace{\sum_{j=0}^{b^m-1} \chi_{[\frac{j}{b^m}, \frac{j+1}{b^m})}(\Phi_b(i))}_{=1}$$

$$= \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(\mathbf{x}_i, t_{\lfloor i/b^m \rfloor}).$$

In Eq. 3, the characteristic function χ selects the summands with $t_i \in [\frac{j}{b^m}, \frac{j+1}{b^m})$ and therefore the index j is equal to the integer part of $b^m t_i$. Since $t_i = \Phi_b(i)$ is a $(0, 1)$-sequence in base b as defined in Eq. 2 generated by an identity matrix scaled by an element from $\mathbb{F}_b \setminus \{0\}$, the m least significant digits of i can only influence the m most significant digits of $\Phi_b(i)$. Therefore the fraction $b^m \Phi_b(i) - j$ can be computed by just removing the m least significant digits of the index i. In fact, $\Phi_b(\lfloor i/b^m \rfloor) = b^m \Phi_b(i) - j$, which becomes obvious by comparing the digits of both numbers in base b. As the term $\Phi_b(\lfloor i/b^m \rfloor))$ no longer does depend on j, it can be factored out of the sum. The remaining sum of characteristic functions is always one, because the whole unit interval is covered and $\Phi_b(i) \in [0, 1) = \cup_{j=0}^{b^m-1} [\frac{j}{b^m}, \frac{j+1}{b^m})$.

As a result, the implementation of the algorithm to sample one dimension at a rate b^m less than others can be as simple as shifting the index i to the right by m digits in base b. Numerical evidence leads to the conjecture that the algorithm also works for the component t_i being generated by an upper triangular generator matrix.

Although multiple splitting along a trajectory creates exponential work, the splitting scheme along one dimension can be applied to multiple dimensions. As a result, each dimension j can have its own sampling rate slow down factor b^{m_j}. Note that this includes components of rank-1 lattice sequences, where the generator and the base are relatively prime. For the Halton sequence, the splitting rate obviously must be a product of prime powers, which grows exponentially with dimension.

The new method replaces and improves upon previous approaches [32, 43] and has many applications in graphics, for example ambient occlusion, scattering, sampling environment maps and area light sources, and simulating motion blur.

4 Deterministic Consistent Image Synthesis

The consistency of quasi-Monte Carlo methods for light transport simulation (see Sect. 2.1) follows from Theorem 1 and allows one to use the building blocks developed in the previous Sect. 3 in deterministic consistent image synthesis algorithms.

The following sections describe how the subsequent components of a vector of a low discrepancy sequence are transformed in order to generate a light transport path: A path is started on the image plane, where the stratification properties of the first two dimensions are used to sample the image plane in Sect. 4.1, while the subsequent dimensions are applied in the simulation of a camera in Sect. 4.2. The path then is continued by repeated scattering as described in Sect. 4.3 and connected to the light sources. Section 4.4 considers more aspects of the opposite direction of assembling light transport paths by tracing photon trajectories from the light sources, their use in quasi-Monte Carlo density approximation, and the combination of paths both starting from the image plane and the light sources.

The resulting algorithms in principle all implement Eq. 1 and are progressively refining the results more and more over time. It is simple to interrupt and resume

computation at any time, because the current state of the computation is completely described by the index of the last sample taken within the problem domain. Therefore termination can be triggered by any criterion and the computation can be continued unless the result was satisfactory.

4.1 Sampling the Image Plane for Anti-aliasing

The rectangular picture elements of a display device match the structure of elementary intervals in two dimensions. Simultaneously determining the average color of each pixel is an integro-approximation problem that can be reduced to the form of Eq. 1. As described in Sect. 3.4.1, the computation of all pixels can be realized by one low discrepancy sequence covering the whole screen. Enumerating the samples per pixel offers several advantages:

- The samples in adjacent pixels are never the same, which at low sampling rates hides aliasing artifacts in noise.
- Keeping track of the last sample index per pixel allows for prioritized computation: Pixels of special interest can receive more samples per time, while others progress with relatively fewer samples [40]. Such regions of interest can be specified automatically or by user interaction. Nevertheless, the pixels remain consistent; they only converge at different speeds.
- The computation can be parallelized and load balanced by rendering each pixel in a separate thread. As the parallelization scheme is based on a domain partitioning, the computation is independent of the sequence and timing of the single tasks. Therefore the computation remains strictly deterministic and thus reproducible independent of the parallel execution environment.

The implementation details for the Sobol' $(0, 2)$-sequence and the Halton sequence are found in [19], while the code can be downloaded at http://gruenschloss.org/sample-enum/sample-enum-src.zip. After dedicating the first two dimensions of a low discrepancy sequence to sampling the image plane, the use of the subsequent dimensions for the construction of light transport paths is explored.

4.2 Depth of Field, Motion Blur, and Spectral Rendering

Light is colored and reaches the image plane through an optical system. Except for pinhole cameras, light passes a lens with an aperture, which both specify the focal plane and depth of field. With focal plane and aperture selected by the user, the simulation associated to a sample with index i on the image plane continues by using its next dimensions. For the example of a thin lens with the shape of a unit disk, components $x_{i,3}$ and $x_{i,4}$ can be used to uniformly select a point

Fig. 6 Simple mathematical models for bidirectional scattering distribution functions (BSDF) $f_s(\omega_o, x, \omega)$ use only a small number of basis functions, like for example the Phong model, and therefore cannot efficiently capture the variety and precision of measured data, like for example the silver metallic paint or the violet rubber. The two cylinders illustrate the direction ω of incidence from the *right* and the direction of perfect reflection exiting to the *left* (Images courtesy Ken Dahm partially using data from [53]).

$$\begin{pmatrix} \sqrt{x_{i,3}} \cos 2\pi x_{i,4} \\ \sqrt{x_{i,3}} \sin 2\pi x_{i,4} \end{pmatrix} \tag{4}$$

on the lens from which a ray is traced through the point in space identified by the thin lens law applied to the sample point $(x_{i,1}, x_{i,2})$ on the image plane. The above mapping from the unit square onto the unit disk has been derived using the multi-dimensional inversion method [74].

The simulation of more advanced camera models including spectral properties [77] follows the same principle: Using another dimension to select a wavelength, the samples are summed up weighted according to spectral response curves, which for example map to the color basis of the display device [22].

However, including the simulation of motion blur caused by camera and/or object motion by just adding another dimension to sample time during the open shutter is not efficient. Each sample with a different time would require to adjust all scene assets to that instant, invoking the loading of temporal data and rebuilding acceleration data structures. On the other hand, interpolating data to increase efficiency may result in inconsistent rendering algorithms: For example, once approximated by an insufficient number of linear spline segments, a rotating propeller will never get round during the course of computation. In addition the memory footprint increases linearly with the number of spline segments.

Unless rendering relativistic effects, the speed of light is much faster than the motion to be rendered and efficiency can be increased by selecting one instant in time for multiple light transport paths. This efficient approach is easily implemented as consistent deterministic algorithm using splitting as introduced in Sect. 3.4.4. A splitting rate in the order of the number of pixels on the display device causes temporal data to be prepared once per accumulated frame. A lower splitting rate results in interleaving and averaging lower resolution images [38].

4.3 Sampling Light Transport Path Segments

As determined by the sample on the image plane and the camera, a ray is traced into the scene, where it interacts with the scene boundary and media inside the volume.

Given a direction of incidence ω and a direction ω_o of observation, the fraction of light transported in a location x on the boundary is described by the bidirectional scattering distribution function (BSDF) $f_s(\omega_o, x, \omega)$ as illustrated in Fig. 6. Such densities are modeled as linear combinations of basis functions, which can be analytic or tabulated measurements [53].

For the simulation, one of the basis functions can be selected proportional to its linear combination weight. Then, given a direction of incidence, the direction of scattering is determined by transforming two uniformly distributed components of a low discrepancy sequence. If the selected basis function refers to measured data, the table of measurements can be transformed into a discrete cumulative density distribution function and a direction of scattering can be found using binary search. For analytic functions, the inversion method is applied, which of course requires the basis function to have an analytic integral that is invertible in closed form.

The observed radiance

$$L_o(x, \omega_o) = \int_{\mathscr{S}^2_-(x)} f_s(\omega_o, x, \omega) L_{in}(x, \omega) \cos\theta_x \, d\omega \tag{5}$$

results from the incident radiance L_{in} integrated over the hemisphere $\mathscr{S}^2_-(x)$ aligned by the surface normal in x attenuated by the BSDF f_s, where the cosine of the angle θ_x between the normal on the boundary and the direction of incidence accounts for the perpendicular incident radiance, i.e. the effective part.

For the example of a constant basis function, evaluating such integrals requires to sample the hemisphere with respect to the cosine weight. Such unit vectors

$$\begin{pmatrix} \sqrt{x_{i,j}} \cos 2\pi x_{i,j+1} \\ \sqrt{x_{i,j}} \sin 2\pi x_{i,j+1} \\ \sqrt{1 - x_{i,j}} \end{pmatrix}$$

are similar to uniform samples on the disk (see Eq. 4), as the z component just results from the constraint of unit norm. Alike transformations exist for many other analytic basis functions of BSDFs [66].

The efficient simulation of scattering within media is subject to active research [57, 58, 68, 87], especially the consistent deterministic simulation of inhomogenous media is an open challenge and beyond the scope of this tutorial.

4.3.1 Controlling Path Length

After a direction of scattering has been determined, the next ray can be traced and the procedure is repeated at each point of interaction to follow a path through the scene. The path length can be controlled by Russian roulette, where an extra dimension of the low discrepancy sequence is compared to the reflectivity/transmissivity of a surface in order to determine whether the path is continued or terminated by absorption [74].

Low discrepancy sequences, like for example the Sobol' sequence, are dimension extensible. Nevertheless, path length must be restricted to a maximum path length in an implementation in order to avoid the possibility of an infinite loop due to numerical issues in scattering and especially ray tracing.

4.3.2 Path Tracing

Path tracing generates samples on the image plane, traces a path through the camera, and determines a scattering direction upon interacting with the scene. Whenever a light source is encountered along the path, its contribution attenuated by the product of BSDFs along the path is recorded on the image plane. This first simple rendering algorithm is deterministic and consistent, as each path is completely determined by one vector of a low discrepancy sequence realizing Eq. 1.

Typical types of light sources are area light sources $L_e(x, \omega)$ and high dynamic range environment maps $L_{e,x}(\omega)$, which describe the light incident in one point x from all directions of the sphere ω. Besides analytic models describing the sky dome, environment maps often contain incident light, for example measured by a high dynamic range photograph of a perfect mirror ball. Similarly, area light sources can be modeled by analytic functions or can be given as measured data. For a given direction ω (and a location x for area light sources), the evaluation of the emission distribution function returns a spectral density.

Path tracing is efficient, as long as hitting a light source is likely as for example in product and car visualization, where objects are rendered enclosed by an environment light source. Whenever a ray is scattered off the object to be visualized and does not intersect the boundary any more, the light path is terminated and the direction of the ray is used to look up the spectral density in the environment map. In cases where the integrand exposes more variance in the dimensions used for sampling the environment map, it pays off to split (see Sect. 3.4.4) the camera path by sending multiple rays into the hemisphere.

4.3.3 Next Event Estimation

Path tracing is not efficient for small light sources as for example spots as used in interiors. However, as the position of such light sources is known, it is easy to check, whether they are visible. For so-called next event estimation, one component of a

low discrepancy vector is used to select a light source, while two more are used to determine a point on the selected light source. The visibility is checked by tracing a so-called shadow ray from the location to be illuminated towards the point on the light source. Unless the ray is occluded, the contribution of the light source is recorded.

If the light sources are small, visible, and at sufficient distance, next event estimation will be efficient. Otherwise, there are issues: Sampling the surface of the light sources, the contribution of the light source must be divided by the squared distance to account for the solid angle subtended by the area of the light source. If the point to be illuminated is close to the point of the light source, the division by the squared distance may result in overmodulation or even a division by zero. The corresponding integral over the area of the light source is therefore called weakly singular. This numerical problem can be robustly treated by combining both techniques of sampling the solid angle and the area by either partitioning the integral [44] or weighting the contributions [47, Sect. 4.1.5]. Both approaches bound the integrand and are simple to implement, while the latter approach performs slightly superior with respect to path tracing with next event estimation (see the example in Sect. 4.4.3).

Applying the splitting technique from Sect. 3.4.4 as illustrated in Fig. 5 overcomes the necessity of randomization as required in [43]. Testing multiple shadow rays for one location to be illuminated may increase efficiency.

With an increasing number of light sources and varying area and distance, the selection of contributing light sources becomes costly. Especially visibility cannot be efficiently predicted in a general way and must be tested. Typical such scenarios include architectural scenes, where light comes through door slits and corridors and many lights can be occluded by walls. Note that shadow rays are testing only geometry and therefore transparent or refractive surfaces report occlusion. For that reason, next event estimation will not transport light through glass.

4.3.4 Light Tracing

Instead of starting light transport paths on the image plane, it appears more natural to follow the trajectories of photons emitted by light sources, which requires the simulation of the emission distribution functions. Similar to the previous section, a light source is to be selected. For area light sources a point of emission needs to be determined in addition. The direction of emission results from transforming two more uniform components according to the emission distribution function. For the special case of environment maps, the procedure is described in detail in [7].

Once emitted, the photon trajectory can be followed as described before in Sect. 4.3. Similar to the issue of small light sources in path tracing, it is not very likely that photons pass the camera to hit the image plane and therefore shadow rays are traced from the light path vertices to the camera (in analogy to next event estimation, see Sect. 4.3.3).

Opposite to path tracing, light tracing with next event estimation can render caustics, which for example are caused by light projected through a glass onto a diffuse surface. Generating such light transport paths starting on the image plane is inefficient due to the low probability of hitting the light source after scattering on the diffuse surface, especially, if the light is small. However, if the caustic is seen through a mirror, light tracing with next event estimation fails, too, because the connection from the mirror to the camera realizes the reflection direction with probability zero.

4.4 Blockwise Sampling of Light Transport Paths

When path tracing and light tracing with or without next event estimation are not efficient due to a small probability of establishing a light transport path, starting path segments from both the light sources and the camera and connecting them can help. This requires two Markov chains to be simulated: One with the emission distribution function as initial distribution and the BSDF as transition probabilities and another one starting on the image plane using the BSDF as transition probabilities as well.

By partitioning one low discrepancy sequence along the components as illustrated in Fig. 7, both Markov chains can be realized by one vector $(\mathbf{x}_i, \mathbf{y}_i)$, where for example the odd components $\mathbf{x}_i$ determine the path segment starting from the image plane and the even components $\mathbf{y}_i$ determine a photon trajectory.

As illustrated in Fig. 2 and introduced in Sect. 2.1 the connections between path segments can be established by checking the mutual visibility of the end points of the path segments (see Sect. 4.4.2) or by proximity (see Sect. 4.4.1). Depending on the kind of the connection and the specific length of each of the two path segments, the same transport path may be generated by multiple such techniques (similar to the special case mentioned in Sect. 4.3.3). Their optimal combination is discussed in Sect. 4.4.3.

4.4.1 Connecting Path Segments by Proximity

The basic idea of photon mapping [27, 28] is to compute the transported light using density estimation [71]. The discrete density is stored as a point cloud called photon map, which results from tracing photon trajectories from the light sources and recording the incident energy at each interaction with the scene. In order to compute the radiance as given by Eq. 5, the contribution of the photons within a sphere around the point of interest is averaged.

As has been shown in [34], photon mapping can be realized as a deterministic consistent quasi-Monte Carlo method: Either a (t, s)- or a rank-1 lattice sequence $(\mathbf{x}_i, \mathbf{y}_i)$ in base b is progressively enumerated in blocks of size b^m. For each vector $(\mathbf{x}_i, \mathbf{y}_i)$ of the low discrepancy sequence, a light transport path segment from the camera is constructed using the dimensions of $\mathbf{x}_i$, while a photon trajectory is started

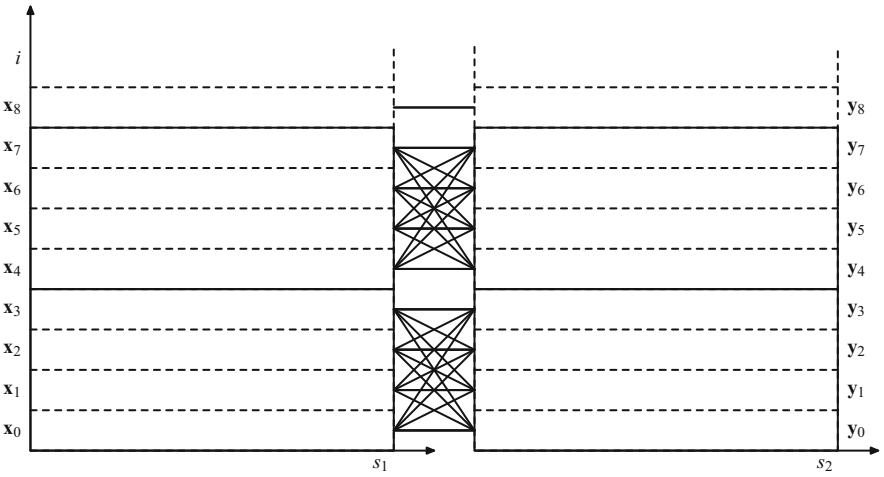


Fig. 7 In order to increase the efficiency of deterministic consistent density estimation (see Sect. 4.4.1), one low discrepancy sequence is partitioned along the components and enumerated in blocks. Within each block, each parameter $\mathbf{x}_i$ is combined with each parameter $\mathbf{y}_i$. It is instructive to compare the blockwise averaging to splitting as shown in Fig. 5.

from the lights using $\mathbf{y}_i$. Then all camera and light path segments within a block are combined to simultaneously compute the radiance

$$L_P = \lim_{n\to\infty} \frac{|P|}{n} \sum_{i=0}^{n-1} \chi_P(\mathbf{x}_i) W(\mathbf{x}_i) \frac{1}{b^m} \sum_{k=0}^{b^m-1} \frac{\chi_{\mathscr{B}(r(n))}(h(\mathbf{x}_i) - h(\mathbf{y}_{b^m\lfloor i/b^m\rfloor+k}))}{\pi r^2(n)} \cdot f_s(\omega(\mathbf{x}_i), h(\mathbf{x}_i), \omega(\mathbf{y}_{b^m\lfloor i/b^m\rfloor+k}))\phi(\mathbf{y}_{b^m\lfloor i/b^m\rfloor+k}) \tag{6}$$

through each pixel P on the image plane. Figure 7 illustrates how the low discrepancy sequence is used in the equation and how the sum over k enumerates all $\mathbf{y}_{b^m\lfloor i/b^m\rfloor+k}$ for each $\mathbf{x}_i$. Out of the camera paths, $\chi_P(\mathbf{x}_i)$ selects the ones contributing the pixel P. Their contribution is weighted by W, which is the product of all attenuations by interactions along the path segment until the query location $h(\mathbf{x}_i)$ is hit. The flux deposited in $h(\mathbf{y}_{b^m\lfloor i/b^m\rfloor+k})$ by a photon is ϕ. If now the difference of both hit points is in a ball $\mathscr{B}$ of radius $r(n)$, both path segments are considered connected by proximity and the product of weight, the flux, and the BSDF f_s is recorded for the pixel P. Assuming a locally planar surface around the query location, the contribution is averaged over the disk area $\pi r^2(n)$ and both ω denote the directions from which the end points of the path segments are hit.

Consistency requires the radius

$$r^2(n) = \frac{r_0^2}{n^\alpha} \text{ for } 0 < \alpha < 1$$

to decrease with a power of n. As shown in [37, Sect. 3.1], the radius vanishes arbitrarily slowly and the influence of the parameter α becomes negligible already after enumerating a few blocks. Consequently, the efficiency is controlled by the initial radius r_0 and the parameter m determining the block size b^m. The initial radius r_0 determines the ratio of how many photons can interact with a query location. Besides choosing r_0 constant, adaptive methods have been discussed in [37, 41].

Once all query locations and photons within a block have been stored as point clouds, space is subdivided hierarchically in order to prune sets of query locations and photons that cannot interact [35] within the distance of $r(n)$. For the light path segments that can be connected by proximity, the contribution is accumulated according to Eq. 6. Obviously, the block size should be chosen as large as memory permits in order to connect as many light path segments as possible within one block.

The algorithm as described, for example, can render caustics seen in a mirror or through a transparent object and thus overcomes the problem of "insufficient techniques" (see [42, Fig. 2] and Sect. 4.3.4). However, this comes at a price: Since connections are not precise but within a certain radius, images may appear blurred and light may leak through the boundary. Although these artifacts vanish due to consistency, they vanish arbitrarily slowly [37, Sect. 3.1], which underlines the importance of the choice of r_0.

4.4.2 Connecting Path Segments by Shadow Rays

Bidirectional path tracing (BDPT) [47, 79, 80] is a generalization of next event estimation (see Sect. 4.3.3), where any vertex of a camera path can be connected to any vertex of a light path segment by a shadow ray as illustrated in Fig. 2. The algorithm is complementary to photon mapping, because it still is limited by the problem of "insufficient techniques" (see previous section), however, lacks the transient artifacts of progressive photon mapping due to precisely testing the mutual visibility of vertices.

The mapping of a low discrepancy sequence to light transport path segments works as described before by partitioning along the dimensions. While the original approach connected vertices of one camera path segment with the vertices of the corresponding light path segment, now connections can be established within a block of path segments (see Fig. 7). Opposite to photon mapping, where the number of connections can be restricted by proximity, the number of shadow rays is quadratic in the block size multiplied by the product of camera and light path segment length. Besides complexity, the block size also determines the look of the transient artifacts. Using larger block sizes, camera paths in neighboring pixels are illuminated by the same light path vertices. These can be considered point light sources, resulting in sharp shadow contours that of course vanish due to consistency. As an extension, splitting (see Sect. 3.4.4) allows for controlling the ratio of camera and light paths to be connected.

With photon mapping, bidirectional path tracing, and all the variants of sampling, the question of which technique is best to use comes up and will be discussed in the next section.

4.4.3 Optimally Combining Techniques

As described in the previous two sections, connecting camera and light path segments by either shadow rays or proximity, the same path may be generated by multiple sampling techniques. While the mapping of low discrepancy sequences to camera and light path segments is shared over all techniques, their optimal combination is still subject of active research, because the efficiency of the rendering techniques may depend on the scene description. One of the key issues is the lack of efficient algorithms for predicting visibility and especially the discontinuities in the integrands of computer graphics.

Multiple Importance Sampling

Importance sampling aims to improve the efficiency by sampling the function more frequently in important regions of the integration domain. Given a density p representing importance, an integral

$$\int_{[0,1)^s} f(x)dx = \int_{[0,1)^s} f(x)\frac{p(x)}{p(x)}dx = \int_{[0,1)^s} \frac{f(x)}{p(x)}dP(x) = \int_{[0,1)^s} \frac{f(P^{-1}(x))}{p(P^{-1}(x))}dx \tag{7}$$

of a function f is transformed using the substitution $p(x) = \frac{dP(x)}{dx}$ and the formalism of the Riemann-Stieltjes integral. Assuming existence, in many cases the transformation P^{-1} of uniform samples into p-distributed samples can be realized by the multi-dimensional inversion method [25, 74]. Originally developed in Monte Carlo integration [74], the theory has been extended to cover quasi-Monte Carlo integration as well [76].

The observation that in light transport simulation the same path may be generated by different importance sampling techniques led to the idea of multiple importance sampling [47, 79, 80]: Given a function $f(x)$ to be integrated and m densities $p_i(x)$ that can be evaluated and sampled, defining the weights

$$w_i(x) := \frac{p_i^\beta(x)}{\sum_{j=0}^{m-1} p_j^\beta(x)} \qquad \text{with} \qquad \sum_{i=0}^{m-1} w_i(x) = 1 \tag{8}$$

allows for transforming the integral

$$\int_{[0,1)^s} f(x)dx = \int_{[0,1)^s} \sum_{i=0}^{m-1} w_i(x) f(x)dx = \sum_{i=0}^{m-1} \int_{[0,1)^s} w_i(x)\frac{f(x)}{p_i(x)}dP_i(x). \tag{9}$$

into a sum of integrals, where the i-th integral is evaluated using samples that are distributed according to p_i in analogy to Eq. 7. Hence for $m = 1$ the general formulation of multiple importance sampling coincides with importance sampling.

The convex combination using the set of weights $w_i(x)$ is called the power heuristic [79]. While for $\beta = 0$ the weights $w_i = \frac{1}{m}$ result in a uniform weighting, for $\beta > 0$ higher weights are assigned to techniques with higher density. Special cases are the balance heuristic for $\beta = 1$ with weights $w_i \sim p_i$ and the maximum heuristic for $\beta = \infty$, which selects the technique with the highest density $p_i(x)$. Among these and other heuristics, the power heuristic with $\beta = 2$ is slightly superior [79, Theorem 9.2].

Samples x for which $p_i(x) = 0$ obviously cannot be generated, which requires $w_i(x) = 0$ to make the method work. As a direct consequence, for any x at least one density must be positive. It may well happen that this cannot be guaranteed, which is called the problem of "insufficient techniques" [42]. A related issue is the situation, where the denominator is smaller than the numerator and samples may be overly amplified [64, Sect. 2.2], although their importance actually is small.

Example: Removing the Weak Singularity in Direct Illumination

Given an emission distribution function L_e and a BSDF f_s on the scene surface, the direct illumination

$$\begin{aligned} L_d(x, \omega_o) &= \int_{\mathscr{S}^2_-(x)} f_s(\omega_o, x, \omega) L_e(h(x, \omega), -\omega) \cos\theta_x d\omega \\ &= \int_A f_s(\omega_o, x, \omega) L_e(y, -\omega) \cos\theta_x V(x, y) \frac{\cos\theta_y}{|x - y|^2} dy \end{aligned} \tag{10}$$

is equivalently determined by either integrating over the hemisphere $\mathscr{S}^2_-(x)$ or the surface A of the light source L_e with area $|A|$, where the direction ω points from x towards the respective point y on the light source. The ray tracing function $h(x, \omega)$ and the visibility $V(x, y)$ are introduced in Sect. 2.2. Note that Eq. 10 is weakly singular due to the division by the squared distance $|x - y|^2$, which in this form causes numerical problems whenever x and y are sufficiently close.

Two resulting sampling techniques are simulating scattering directions according to

$$p_1 \equiv f_s(\omega_o, x, \omega) \cos\theta_x \qquad \text{and using} \qquad p_2 \equiv \frac{1}{|A|}$$

to generate uniform samples on the light source. For a given hit point $y := h(x, \omega)$, the visibility $V(x, y)$ is one and changing the measure from solid angle in x to a point y on the area of a light source and vice versa [47, Sect. 4.1.5] results in the densities

$$p_1 \frac{\cos\theta_y}{|x-y|^2} \qquad \text{and} \qquad p_2 \frac{|x-y|^2}{\cos\theta_y}.$$

Then the weights for the balance heuristic in Eq. 8 are

$$w_1 \equiv \frac{p_1}{p_1 + p_2 \frac{|x-y|^2}{\cos\theta_y}} = \frac{|A| f_s(\omega_o, x, \omega) \cos\theta_x \cos\theta_y}{|A| f_s(\omega_o, x, \omega) \cos\theta_x \cos\theta_y + |x-y|^2} \qquad \text{and}$$

$$w_2 \equiv \frac{p_2}{p_1 \frac{\cos\theta_y}{|x-y|^2} + p_2} = \frac{|x-y|^2}{|A| f_s(\omega_o, x, \omega) \cos\theta_x \cos\theta_y + |x-y|^2}.$$

While w_1 has been derived using densities with respect to the solid angle and w_2 has been using densities with respect to the area measure, the weights are ratios of densities with respect to the same measure and therefore have no unit. Using the transformation in Eq. 9 and the equivalence in Eq. 10, the direct illumination amounts to

$$\begin{aligned} &L_d(x, \omega_o) \qquad (11)\\ &= \int_{[0,1)^2} L_e(h(x,\omega), -\omega) \frac{|A| f_s(\omega_o, x, \omega) \cos\theta_x \cos\theta_y}{|A| f_s(\omega_o, x, \omega) \cos\theta_x \cos\theta_y + |x-y|^2} dP_1(\omega)\\ &\quad + \int_{[0,1)^2} L_e(y, -\omega) V(x,y) \frac{f_s(\omega_o, x, \omega) \cos\theta_x \cos\theta_y}{|A| f_s(\omega_o, x, \omega) \cos\theta_x \cos\theta_y + |x-y|^2} dP_2(y) \end{aligned}$$

in accordance with [47, Eq. 4.7, Sect. 4.1.5].

Assuming that directions can be generated with the density p_1, both integrands in Eq. 11 are bounded, because the weak singularity [44] has been removed by the transformation, which underlines one of the major advantages of multiple importance sampling.

Note that the undefined $\frac{0}{0}$ case needs to be handled explicitly: In order to avoid numerical exceptions, it is sufficient to test $f_s \cos\theta_x \cos\theta_y$ for zero explicitly, since then no radiation is transported. Comparing this term to a small, positive threshold, substantial amounts of transport may be missed for small distances $|x-y|^2$.

A Path Tracer with Next Event Estimation and Multiple Importance Sampling

For the purpose of this tutorial an efficient implementation of a path tracer with next event estimation (see Sect. 4.3.3) and multiple importance sampling is described [47, Sect. 4.1.5]. Light transport is modeled by a Fredholm integral equation of the second kind

$$L = L_e + T_{f_s} L,$$

where the integral operator

$$T_{f_s} L \equiv \int_{\mathcal{S}^2_+(x)} f_s(\omega_o, x, \omega) L(h(x, \omega), -\omega) \cos \theta_x d\omega$$

determines the transported radiance in analogy to Eq. 10. The radiance thus is the source radiance L_e plus transported radiance

$$T_{f_s} L = T_{f_s}(L_e + T_{f_s} L) = T_{f_s}((w_1 + w_2)L_e + T_{f_s} L) = T_{f_s}(w_1 L_e + T_{f_s} L) + T_{f_s} w_2 L_e,$$

which can be computed by first replacing one instance of the radiance by its definition and then inserting a linear combination of weights that always sums up to one as derived in the previous section. As a result the transported radiance is determined by two terms: The first term is evaluated by integration over the hemisphere, which comprises sampling a scattering direction, tracing a ray and summing up the weighted emission L_e and recursively computing the transported radiance. The second term uses a shadow ray towards the support of the light sources and the according weight as derived in Eq. 11 for the example of the balance heuristic.

The implementation can be realized as a simple loop without recursion, terminating the path started from the image plane by Russian roulette (see Sect. 4.3.2), and using the scattering direction both for path tracing and next event estimation with multiple importance sampling. For regions with visibility $V = 1$ the method converges much faster than the path tracer without multiple importance sampling although no additional rays need to be traced.

This simple but already quite powerful algorithm can be extended to bidirectional path tracing, where all vertices of a camera path are connected to all vertices of a light path. Using the principle of implicit importance sampling as before, the implementation is compact [1, 5]. As bidirectional path tracing suffers the problem of insufficient techniques [42], photon mapping can be added [34] by using multiple importance sampling as well [20, and references therein]. Using the quasi-Monte Carlo technique of blockwise enumeration as illustrated in Fig. 7, all image synthesis algorithms can be implemented as progressive, consistent, and deterministic algorithms.

Although multiple importance sampling takes care of optimally combining techniques, it does not consider visibility: For the simple example of a light bulb, all shadow rays will report occlusion by the glass around the glowing wire. Similarly, shadow rays are not efficient, when light enters a room or a car through a window. On the other hand, mostly diffuse scenes do not benefit from photon mapping, which raises the question, whether for a given scene description the relevant techniques can be determined algorithmically and whether visibility can be efficiently predicted.

5 State of the Art

Consistent quasi-Monte Carlo methods are easily parallelized and are perfectly reproducible due to their deterministic nature. In industry they take advantage of SIMD (single instruction multiple data) architectures and are perfectly suitable for latency hiding architectures, especially GPUs (graphics processing units). Besides computer graphics, other domains like for example finance, will benefit from parallel quasi-Monte Carlo methods as well.

Low discrepancy sequences can be generated at the speed of high-quality pseudo-random numbers and they offer a performance advantage due to better discrete density approximation [25]. On certain restricted function classes [8, 55, 72], quasi-Monte Carlo methods are roughly quadratically faster than Monte Carlo methods and it is known that quasi-Monte Carlo methods outperform Monte Carlo methods on the average [78, 86]. However, due to the deterministic nature of quasi-Monte Carlo methods, it is possible to construct theoretical worst cases, especially for the class of square integrable functions, where a Monte Carlo method can be expected to be better. For this reason the general Theorem 1 cannot provide a good rate of convergence on the class of square integrable functions.

Beyond the state of the art as surveyed in this article, there are still fundamental issues in image synthesis: While (multiple) importance sampling is deeply explored in the context of computer graphics, there are indications that the weights for combining bidirectional path tracing and photon mapping are not optimal and that there is no efficient deterministic method that can incorporate the prediction of visibility (see Sect. 4.4.2), yet. While the Metropolis light transport [79, 81] algorithm can efficiently handle boundaries with complex visibility, there does not exist a deterministic version and it is unknown how to benefit from low discrepancy.

For all known techniques, settings can be constructed that result in inefficient performance: For example, shadow rays do not work with transparent objects like glass and the Metropolis light transport algorithm is not efficient in simple settings. There is a desire to algorithmically determine which techniques are efficient for a given setting.

Besides lighting complexity, the amount of data to be rendered in one frame reaches amounts that require simplification to enable efficient processing. Such approaches relate to multi-level algorithms and function representations and level-of-detail representations. Finding such approximations is still a challenge, because changing visibility often dramatically changes the light transport and consequently the rendered image.

In conclusion, the paradigm of consistency has led to many new developments in quasi-Monte Carlo methods and numerous industrial rendering solutions apply quasi-Monte Carlo methods for light transport simulation.

Acknowledgements The author likes to thank Ian Sloan, Frances Kuo, Josef Dick, and Gareth Peters for the extraordinary opportunity to present this tutorial at the MCQMC 2012 conference and Pierre L'Ecuyer for the invitation to present an initial tutorial on "Monte Carlo and Quasi-Monte

Carlo Methods in Computer Graphics" at MCQMC 2008. In addition, the author is grateful to the anonymous reviewers, Nikolaus Binder, and Ken Dahm.

References

1. van Antwerpen, D.: Unbiased physically based rendering on the GPU. Master's thesis, Computer Graphics Research Group, Department of Software Technology Faculty EEMCS, Delft University of Technology, The Netherlands (2011)
2. Cools, R., Kuo, F., Nuyens, D.: Constructing embedded lattice rules for multivariate integration. SIAM J. Sci. Comput. **28**, 2162–2188 (2006)
3. Cools, R., Reztsov, A.: Different quality indexes for lattice rules. J. Complexity **13**, 235–258 (1997)
4. Cranley, R., Patterson, T.: Randomization of number theoretic methods for multiple integration. SIAM J. Numer. Anal. **13**, 904–914 (1976)
5. Dahm, K.: A comparison of light transport algorithms on the GPU. Master's thesis, Computer Graphics Group, Saarland University (2011)
6. Dammertz, H.: Acceleration Methods for Ray Tracing based Global Illumination. Ph.D. thesis, Universität Ulm (2011)
7. Dammertz, H., Hanika, J.: Plane sampling for light paths from the environment map. J. Graph. GPU Game Tools **14**, 25–31 (2009)
8. Dammertz, S., Keller, A.: Image synthesis by rank-1 lattices. In: Keller, A., Heinrich, S., Niederreiter, H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2006, pp. 217–236. Springer, Berlin/Heidelberg (2008)
9. Devroye, L.: Non-Uniform Random Variate Generation. Springer, New York (1986)
10. Dick, J., Pillichshammer, F.: Digital Nets and Sequences. Discrepancy Theory and Quasi-Monte Carlo Integration. Cambridge University Press, Cambridge (2010)
11. Edwards, D.: Practical Sampling for Ray-Based Rendering. Ph.D. thesis, The University of Utah (2008)
12. Faure, H.: Discrépance de suites associées à un système de numération (en dimension s). Acta Arith. **41**, 337–351 (1982)
13. Faure, H.: Good permutations for extreme discrepancy. J. Number Theory **42**, 47–56 (1992)
14. Friedel, I., Keller, A.: Fast generation of randomized low-discrepancy point sets. In: Fang, K.-T., Hickernell, F.J., Niederreiter, H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2000, pp. 257–273. Springer, Berlin/Heidelberg (2002)
15. Frolov, A., Chentsov, N.: On the calculation of certain integrals dependent on a parameter by the Monte Carlo method. Zh. Vychisl. Mat. Fiz. **2**, 714–717 (1962). (in Russian)
16. Glassner, A.: Principles of Digital Image Synthesis. Morgan Kaufmann, San Francisco (1995)
17. Grünschloß, L.: Motion Blur. Master's thesis, Ulm University (2008)
18. Grünschloß, L., Keller, A.: (t, m, s)-nets and maximized minimum distance, Part II. In: L'Ecuyer, P., Owen, A. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2008, pp. 395–409. Springer, Berlin/Heidelberg (2010)
19. Grünschloß, L., Raab, M., Keller, A.: Enumerating Quasi-Monte Carlo point sequences in elementary intervals. In: Plaskota, L., Woźniakowski, H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2010, pp. 399–408. Springer, Berlin (2012)
20. Hachisuka, T., Pantaleoni, J., Jensen, H.: A path space extension for robust light transport simulation. ACM Trans. Graph. **31**, 191:1–191:10 (2012)
21. Halton, J.: On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. Numer. Math. **2**, 84–90 (1960)
22. Hanika, J.: Spectral light transport simulation using a precision-based ray tracing architecture. Ph.D. thesis, Universität Ulm (2010)

23. Hickernell, F., Hong, H., L'Ecuyer, P., Lemieux, C.: Extensible lattice sequences for quasi-Monte Carlo quadrature. SIAM J. Sci. Comput. **22**, 1117–1138 (2001)
24. Hlawka, E.: Discrepancy and Riemann integration. In: Mirsky, L. (ed.) Studies in Pure Mathematics, pp. 121–129. Academic Press, New York (1971)
25. Hlawka, E., Mück, R.: Über eine Transformation von gleichverteilten Folgen II. Computing **9**, 127–138 (1972)
26. Hong, H.: Digital Nets and Sequences for Quasi-Monte Carlo Methods. Ph.D. thesis, Hong Kong Baptist University (2002)
27. Jensen, H.: Global illumination using photon maps. In: Rendering Techniques 1996: Proceedings of the 7th Eurographics Workshop on Rendering, Porto, pp. 21–30. Springer (1996)
28. Jensen, H.: Realistic Image Synthesis Using Photon Mapping. AK Peters, Natick (2001)
29. Joe, S., Kuo, F.: Remark on algorithm 659: Implementing Sobol's quasirandom sequence generator. ACM Trans. Math. Software **29**, 49–57 (2003)
30. Joe, S., Kuo, F.: Constructing Sobol' sequences with better two-dimensional projections. SIAM J. Sci. Comput. **30**, 2635–2654 (2008)
31. Keller, A.: Quasi-Monte Carlo methods in computer graphics: The global illumination problem. Lect. Appl. Math. **32**, 455–469 (1996)
32. Keller, A.: Trajectory splitting by restricted replication. Monte Carlo Methods Appl. **10**, 321–329 (2004)
33. Keller, A.: Myths of computer graphics. In: Niederreiter, H., Talay, D. (ed.) Monte Carlo and Quasi-Monte Carlo Methods 2004, pp. 217–243. Springer, Berlin/Heidelberg (2006)
34. Keller, A., Binder, N.: Deterministic consistent density estimation for light transport simulation. In: Dick, J., Kuo, F.Y., Peters, G.W., Sloan, I.H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2012, this volume 467–480. Springer, Berlin/Heidelberg (2013)
35. Keller, A., Droske, M., Grünschloß, L., Seibert, D.: A divide-and-conquer algorithm for simultaneous photon map queries. Poster at High-Performance Graphics in Vancouver (2011)
36. Keller, A., Grünschloß, L.: Parallel Quasi-Monte Carlo integration by partitioning low discrepancy sequences. In: Plaskota, L., Woźniakowski, H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2010, pp. 487–498. Springer, Berlin/Heidelberg (2012)
37. Keller, A., Grünschloß, L., Droske, M.: Quasi-Monte Carlo progressive photon mapping. In: Plaskota, L., Woźniakowsi, H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2010, pp. 499–509. Springer, Berlin/Heidelberg (2012)
38. Keller, A., Heidrich, W.: Interleaved sampling. In: Myszkowski, K., Gortler, S. (eds.) Rendering Techniques 2001: Proceedings of the 12th Eurographics Workshop on Rendering, London, pp. 269–276. Springer (2001)
39. Keller, A., Wächter, C.: Efficient ray tracing without auxiliary acceleration data structure. Poster at High-Performance Graphics in Vancouver (2011)
40. Keller, A., Wächter, C., Kaplan, M.: System, method, and computer program product for consistent image synthesis. United States Patent Application US20110025682 (2011)
41. Knaus, C., Zwicker, M.: Progressive photon mapping: A probabilistic approach. ACM Trans. Graph. (TOG) **25**, (2011)
42. Kollig, T., Keller, A.: Efficient bidirectional path tracing by randomized Quasi-Monte Carlo integration. In: Niederreiter, H., Fang, K., Hickernell, F. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2000, pp. 290–305. Springer, Berlin (2002)
43. Kollig, T., Keller, A.: Efficient multidimensional sampling. Comput. Graph. Forum (Proc. Eurographics 2002) **21**, 557–563 (2002)
44. Kollig, T., Keller, A.: Illumination in the presence of weak singularities. In: Niederreiter, H., Talay, D., (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2004, pp. 245–257. Springer, Berlin/Heidelberg (2006)
45. Kritzer, P.: On an example of finite hybrid quasi-Monte Carlo point sets. Monatsh. Math. **168**, 443–459 (2012)
46. Kritzer, P., Leobacher, G., Pillichshammer, F.: Component-by-component construction of hybrid point sets based on Hammersley and lattice point sets. In: Dick, J., Kuo, F.Y., Peters, G.W., Sloan, I.H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2012, this volume 501–515. Springer, Berlin/Heidelberg (2013)

47. Lafortune, E.: Mathematical models and Monte Carlo algorithms for physically based rendering. Ph.D. thesis, KU Leuven, Belgium (1996)
48. L'Ecuyer, P., Munger, D.: Latticebuilder: a general software tool for constructing rank-1 lattice rules. ACM Trans. Math. Software (2012, submitted)
49. Lemieux, C.: Monte Carlo and Quasi-Monte Carlo Sampling. Springer, New York (2009)
50. Maize, E.: Contributions to the Theory of Error Reduction in Quasi-Monte Carlo Methods. Ph.D. thesis, Claremont Graduate School (1980)
51. Maize, E., Sepikas, J., Spanier, J.: Accelerating the convergence of lattice methods by importance sampling-based transformations. In: Plaskota, L., Woźniakowski, H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2010, pp. 557–572. Springer, Berlin/Heidelberg (2012)
52. Matoušek, J.: On the L_2-discrepancy for anchored boxes. J. Complexity **14**, 527–556 (1998)
53. Matusik, W., Pfister, H., Brand, M., McMillan, L.: A data-driven reflectance model. ACM Tran. Graph. (Proc. SIGGRAPH 2003) **22**, 759–769 (2003)
54. Niederreiter, H.: Quasirandom sampling in computer graphics. In: Proceedings of the 3rd International Seminar on Digital Image Processing in Medicine, Remote Sensing and Visualization of Information, Riga, Latvia, pp. 29–34 (1992)
55. Niederreiter, H.: Random Number Generation and Quasi-Monte Carlo Methods. SIAM, Philadelphia (1992)
56. Niederreiter, H.: Error bounds for quasi-Monte Carlo integration with uniform point sets. J. Comput. Appl. Math. **150**, 283–292 (2003)
57. Novák, J., Nowrouzezahrai, D., Dachsbacher, C., Jarosz, W.: Progressive virtual beam lights. Comput. Graph. Forum (Proceedings of EGSR 2012) **31**, 1407–1413 (2012)
58. Novák, J., Nowrouzezahrai, D., Dachsbacher, C., Jarosz, W.: Virtual ray lights for rendering scenes with participating media. ACM Trans. Graph. (Proceedings of ACM SIGGRAPH 2012) **60**, (2012)
59. Nuyens, D., Waterhouse, B.: A global adaptive quasi-Monte Carlo algorithm for functions of low truncation dimension applied to problems of finance. In: Plaskota L., Woźniakowsi, H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2010, pp. 591–609. Springer, Berlin/Heidelberg (2012)
60. Ohbuchi, R., Aono, M.: Quasi-Monte Carlo rendering with adaptive sampling. IBM Tokyo Research Laboratory (1996)
61. Owen, A.: Orthogonal arrays for computer experiments, integration and visualization. Stat. Sin. **2**, 439–452 (1992)
62. Owen, A.: Randomly permuted (t, m, s)-nets and (t, s)-sequences. In: Niederreiter, H., Shiue, P.J.-S. (eds.) Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing. Lecture Notes in Statistics, vol. 106, pp. 299–315. Springer, New York (1995)
63. Owen, A.: Monte Carlo variance of scrambled net quadrature. SIAM J. Numer. Anal. **34**, 1884–1910 (1997)
64. Owen, A., Zhou, Y.: Safe and effective importance sampling. J. Amer. Statist. Assoc. **95**, 135–143 (2000)
65. Paskov, S.: Termination criteria for linear problems. J. Complexity **11**, 105–137 (1995)
66. Pharr, M., Humphreys, G.: Physically Based Rendering, 2nd edn. Morgan Kaufmann, San Francisco (2011)
67. Press, H., Teukolsky, S., Vetterling, T., Flannery, B.: Numerical Recipes in C. Cambridge University Press, Cambridge (1992)
68. Raab, M., Seibert, D., Keller, A.: Unbiased global illumination with participating media. In: Keller, A., Heinrich, S., Niederreiter, H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2006, pp. 669–684. Springer, Berlin (2007)
69. Shirley, P.: Discrepancy as a quality measure for sampling distributions. In: Eurographics '91, Vienna, pp. 183–194. Elsevier/North-Holland, Amsterdam (1991)
70. Shirley, P.: Realistic Ray Tracing. AK Peters, Natick (2000)
71. Silverman, B.: Density Estimation for Statistics and Data Analysis. Chapman & Hall/CRC, Boca Raton (1986)

72. Sloan, I., Joe, S.: Lattice Methods for Multiple Integration. Clarendon Press, Oxford (1994)
73. Sobol', I.: On the Distribution of points in a cube and the approximate evaluation of integrals. Zh. vychisl. Mat. mat. Fiz. **7**, 784–802 (1967)
74. Sobol', I.: Die Monte-Carlo-Methode. Deutscher Verlag der Wissenschaften (1991)
75. Sobol', I., Asotsky, D., Kreinin, A., Kucherenko, S.: Construction and comparison of high-dimensional Sobol' generators. WILMOTT Mag. **56**, 64–79 (2011)
76. Spanier, J., Maize, E.: Quasi-random methods for estimating integrals using relatively small samples. SIAM Rev. **36**, 18–44 (1994)
77. Steinert, B., Dammertz, H., Hanika, J., Lensch, H.: General spectral camera lens simulation. Comput. Graph. Forum **30**, 1643–1654 (2011)
78. Traub, J., Wasilkowski, G., Woźniakowski, H.: Information-Based Complexity. Academic Press, Boston (1988)
79. Veach, E.: Robust Monte Carlo methods for light transport simulation. Ph.D. thesis, Stanford University (1997)
80. Veach, E., Guibas, L.: Optimally combining sampling techniques for Monte Carlo rendering. In: Proceedings of the SIGGRAPH 1995, Annual Conference Series, Los Angeles, pp. 419–428 (1995)
81. Veach, E., Guibas, L.: Metropolis light transport. In: Whitted, T. (ed.) Proceedings of the SIGGRAPH 1997, Annual Conference Series, Los Angeles, pp. 65–76. ACM SIGGRAPH, Addison Wesley (1997)
82. Wächter, C.: Quasi-Monte Carlo light transport simulation by efficient ray tracing. Ph.D. thesis, Universität Ulm (2008)
83. Wächter, C., Keller, A.: System and process for improved sampling for parallel light transport simulation. ISF MI-12-0006-US0 filed as United States Patent Application US20130194268 (2013)
84. Wang, Y., Hickernell, F.: An historical overview of lattice point sets. In: Fang, K.-T., Hickernell, F.J., Niederreiter, H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2000, pp. 158–167. Springer, Berlin/Heidelberg (2002)
85. Weyl, H.: Über die Gleichverteilung von Zahlen mod. Eins. Math. Ann. **77**, 313–352 (1916)
86. Woźniakowski, H., Traub, J.: Breaking intractability. Sci. Am. **270**, 102–107 (1994)
87. Yue, Y., Iwasaki, K., Chen, B., Dobashi, Y., Nishita, T.: Unbiased, adaptive stochastic sampling for rendering inhomogeneous participating media. ACM Trans. Graph. **29**, 177 (2010)
88. Zaremba, S.: La discrépance isotrope et l'intégration numérique. Ann. Mat. Pura Appl. **87**, 125–136 (1970)

Part III
Contributed Articles

Conditional Sampling for Barrier Option Pricing Under the Heston Model

Nico Achtsis, Ronald Cools, and Dirk Nuyens

Abstract We propose a quasi-Monte Carlo algorithm for pricing knock-out and knock-in barrier options under the Heston (Rev Financ Stud 6(2):327–343, 1993) stochastic volatility model. This is done by modifying the LT method from Imai and Tan (J Comput Financ 10(2):129–155, 2006) for the Heston model such that the first uniform variable does not influence the stochastic volatility path and then conditionally modifying its marginals to fulfill the barrier condition(s). We show that this method is unbiased and never does worse than the unconditional algorithm. In addition, the conditioning is combined with a root finding method to also force positive payouts. The effectiveness of this method is shown by extensive numerical results.

1 Introduction

It is well known that the quasi-Monte Carlo method in combination with a good path construction method, like the LT method from Imai and Tan [10], can be a helpful tool in option pricing, see, e.g., [4, 13]. The integrand functions usually take the form $\max(f, 0)$ and a good path construction will somehow align the discontinuity in the derivative along the axes. However, as soon as other discontinuities, in the form of barrier conditions, are introduced, the performance of the quasi-Monte Carlo method degrades as a lot of sample paths might not contribute to the estimator anymore and are basically wasted, see [14] for an illustration and an alternative solution. This is also the case for the Monte Carlo method for which in [6] a conditional sampling method has been introduced to alleviate this problem.

N. Achtsis (✉) · R. Cools · D. Nuyens
Department of Computer Science, KU Leuven, B-3001 Heverlee, Belgium
e-mail: nico.achtsis@cs.kuleuven.be; ronald.cools@cs.kuleuven.be; dirk.nuyens@cs.kuleuven.be

J. Dick et al. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2012*,
Springer Proceedings in Mathematics & Statistics 65,
DOI 10.1007/978-3-642-41095-6_9, © Springer-Verlag Berlin Heidelberg 2013

A conditional sampling scheme will make certain all sample paths will adhere to the barrier condition and weights their contribution by the likelihood of its occurrence.

In previous work [1] we have introduced a conditional sampling method to deal with barrier conditions in the Black–Scholes setting that can be used in combination with a good path construction method like the LT method. In that paper we have shown that such a scheme always performs better than the unconditional method. Here we consider the more realistic Heston model [8], which has a stochastic volatility component, and derive an algorithm to do conditional sampling on barrier conditions under this model. We focus solely on the LT path construction which enables us to construct a good path construction for the payoff; excluding the maximum and barrier conditions which are handled by a root finding method (optional) and the conditional sampling proposed in this paper.

2 The LT Method for Heston Under Log Prices

Assume a Heston world [8] in which the risk-neutral dynamics of the asset are given by

$$\begin{aligned} dS(t) &= rS(t)dt + \sqrt{V(t)}S(t)dW^1(t), \\ dV(t) &= (\theta - V(t))\kappa dt + \sigma\sqrt{V(t)}dW^2(t), \\ dW^1(t)dW^2(t) &= \rho dt, \end{aligned}$$

where $S(t)$ denotes the price of the asset at time t, r is the risk-free interest rate, κ is the mean-reversion parameter of the volatility process, θ is the long run average price variance and σ is the volatility of the volatility. We assume the Feller condition $2\kappa\theta \geq \sigma^2$ such that the process $V(t)$ is strictly positive. The parameter ρ controls the correlation between the log-returns and the volatility. A useful observation is that one can write

$$W^1(t) = \rho W^2(t) + \sqrt{1-\rho^2}\, W^3(t),$$

where $W^2(t)$ and $W^3(t)$ are independent Brownian motions. This corresponds to the Cholesky decomposition of the correlation structure. When resorting to Monte Carlo techniques for pricing options under this model, asset paths need to be discretized. For simplicity we assume that time is discretized using m equidistant time steps $\Delta t = T/m$, but all results can be extended to the more general case. The notations $\hat{S}_k$ and $\hat{V}_k$ will be used for $\hat{S}(k\Delta t)$ and $\hat{V}(k\Delta t)$, respectively. We use the Euler–Maruyama scheme [12] to discretize the asset paths in log-space (see also [5, Sect. 6.5] w.r.t. transformations of variables) and sample the independent Brownian motions W^2 and W^3 by using independent standard normal variables Z^1 and Z^2; for $k = 0, \ldots, m-1$,

$$\log \hat{S}_{k+1} = \log \hat{S}_k + \left(r - \frac{\hat{V}_k}{2}\right)\Delta t + \sqrt{\hat{V}_k}\sqrt{\Delta t}\left(\rho Z^1_{k+1} + \sqrt{1-\rho^2}Z^2_{k+1}\right), \tag{1}$$

$$\hat{V}_{k+1} = \hat{V}_k + (\theta - \hat{V}_k)\kappa\Delta t + \sigma\sqrt{\hat{V}_k}\sqrt{\Delta t}Z^1_{k+1}. \tag{2}$$

For our method it is important that $\hat{V}$ is sampled solely from Z^1 and to switch to log-space. This will be explained in the next sections.

Write $\boldsymbol{Z} = (Z^1_1, Z^2_1, Z^1_2, Z^2_2, \ldots, Z^2_m)' \in \mathbb{R}^{2m}$, where the prime is used to denote the transpose of a vector. Then $\boldsymbol{Z}$ has multivariate standard normal distribution. Assuming a European option payoff represented as

$$\max\left(f(\boldsymbol{Z}), 0\right)$$

one usually simulates the function $f(\boldsymbol{Z})$ by mapping a uniform variate $\boldsymbol{u}$ in the unit cube to $\boldsymbol{Z}$ by applying the inverse cumulative distribution function Φ^{-1}. We will call this method the standard Monte Carlo method (MC). When using quasi-Monte Carlo (QMC), the uniform variates are replaced by a low-discrepancy point set. Our conditional sampling scheme will use the influence of the first uniform variable u_1 to try and force the barrier conditions to be met. For this we will employ the LT method. First, the uniformly sampled variate $\boldsymbol{u}$ is mapped to a standard normal variate $\boldsymbol{z}$ as in the MC method. The function $f(\boldsymbol{Z})$ is then sampled using the transformation $\boldsymbol{Z} = Q\boldsymbol{z}$ for a carefully chosen orthogonal matrix Q. This means that in (1) and (2) we take, for $k = 0, \ldots, m-1$,

$$Z^1_{k+1} = \sum_{n=1}^{2m} q_{2k+1,n} z_n \qquad \text{and} \qquad Z^2_{k+1} = \sum_{n=1}^{2m} q_{2k+2,n} z_n,$$

where $q_{i,j}$ denotes the element from the matrix Q at row i and column j. We remark that, for ease of notation, we will write $f(\boldsymbol{Z})$, $f(\boldsymbol{z})$, $f(\boldsymbol{u})$ or $f(\hat{S}_1, \ldots, \hat{S}_m)$ to denote the function f from above in terms of normal variates $\boldsymbol{Z}$ or $\boldsymbol{z}$, uniform variates $\boldsymbol{u}$ or just the discretized stock path $\hat{S}_1, \ldots, \hat{S}_m$.

In what follows the notation $Q_{\bullet k}$ denotes the kth column of Q and $Q_{k\bullet}$ denotes the kth row. The LT method [10] chooses the matrix Q according to the following optimization problem:

$$\begin{aligned} \underset{Q_{\bullet k} \in \mathbb{R}^{2m}}{\text{maximize}} \quad & \text{variance contribution of } f \text{ due to } k\text{th dimension} \\ \text{subject to} \quad & \|Q_{\bullet k}\| = 1, \\ & \langle Q^*_{\bullet j}, Q_{\bullet k}\rangle = 0, \quad j = 1, \ldots, k-1, \end{aligned}$$

where $Q^*_{\bullet j}$ denotes the columns of Q that have already been optimized in the previous iterations. The algorithm is carried out iteratively for $k = 1, 2, \ldots, 2m$ so that in the kth optimization step the objective function ensures that, given columns $Q^*_{\bullet j}$, $j = 1, \ldots, k-1$ which have already been determined in the previous iterations, the variance contribution due to the kth dimension is maximized while the constraints ensure orthogonality. Being able to express the variance contribution for each component analytically for general payoff functions f can be quite complicated. Therefore, Imai and Tan [10] propose to approximate the objective function by linearizing it using a first-order Taylor expansion for $z = \hat{z} + \Delta z$,

$$f(z) \approx f(\hat{z}) + \sum_{k=1}^{2m} \left.\frac{\partial f}{\partial z_k}\right|_{z=\hat{z}} \Delta z_k .$$

Using this expansion, the variance contributed due to the kth component is

$$\left(\left.\frac{\partial f}{\partial z_k}\right|_{z=\hat{z}} \right)^2 .$$

The expansion points are chosen as $\hat{z}_k = (1, \ldots, 1, 0, \ldots, 0)$, the vector with $k-1$ leading ones. Different expansion points will lead to different transformation matrices; this particular choice allows for an efficient construction. The optimization problem becomes

$$\begin{aligned} \underset{Q_{\bullet k} \in \mathbb{R}^{2m}}{\text{maximize}} \quad & \left(\left.\frac{\partial f}{\partial z_k}\right|_{z=\hat{z}_k} \right)^2 \\ \text{subject to} \quad & \| Q_{\bullet k} \| = 1, \\ & \langle Q^*_{\bullet j}, Q_{\bullet k} \rangle = 0, \quad j = 1, \ldots, k-1. \end{aligned} \tag{3}$$

The original Imai and Tan paper [10] considers a European call option to illustrate the computational advantage of the LT method under the Heston model. In their paper the stochastic volatility is described in [10, Sect. 4.2] and we will revisit their method in Sect. 4. For ease of illustration we also consider the payoff function inside the max-function to be that of a European call option

$$f(z) = \hat{S}_m - K$$

where K is the strike price. For notational ease, we introduce the following functions:

$$f_k^1 = \frac{\sqrt{\Delta t}}{2\sqrt{\hat{V}_k}} \left(\rho Z_{k+1}^1 + \sqrt{1-\rho^2} Z_{k+1}^2 \right) - \frac{\Delta t}{2},$$

$$f_k^2 = \sqrt{\hat{V}_k}\sqrt{\Delta t},$$
$$f_k^3 = 1 - \kappa\Delta t + \frac{\sigma\sqrt{\Delta t}}{2\sqrt{\hat{V}_k}} Z_{k+1}^1,$$
$$f_k^4 = \sigma\sqrt{\hat{V}_k}\sqrt{\Delta t}.$$

Note that all the above functions f^i depend on $\boldsymbol{Z}$. Similar to [10], to find the partial derivatives $\partial\hat{S}_m/\partial z_i$ needed for the optimization algorithm, we obtain the recursive relations (with initial conditions $\partial \log \hat{S}_0/\partial z_i = 0$ and $\partial\hat{V}_0/\partial z_i = 0$)

$$\frac{\partial \log \hat{S}_{k+1}}{\partial z_i} = \frac{\partial \log \hat{S}_k}{\partial z_i} + \frac{\partial \hat{V}_k}{\partial z_i} f_k^1 + \left(\rho q_{2k+1,i} + \sqrt{1-\rho^2} q_{2k+2,i}\right) f_k^2, \quad (4)$$
$$\frac{\partial \hat{V}_{k+1}}{\partial z_i} = \frac{\partial \hat{V}_k}{\partial z_i} f_k^3 + q_{2k+1,i} f_k^4, \quad (5)$$

where k goes from 0 to $m-1$. The chain rule is used to obtain

$$\frac{\partial \hat{S}_m}{\partial z_i} = \hat{S}_m \frac{\partial \log \hat{S}_m}{\partial z_i}.$$

We will use the following lemma to calculate the transformation matrix.

Lemma 1. *The recursion*

$$F_{k+1} = a_k F_k + b_k q_k,$$
$$G_{k+1} = c_k G_k + d_k q_k + e_k F_k,$$

with initial values $F_0 = G_0 = 0$ *can be written at index* $k+1$ *as a linear combination of the* q_ℓ, $\ell = 0, \ldots, k$, *as follows*

$$F_{k+1} = \sum_{\ell=0}^{k} q_\ell b_\ell \prod_{j=\ell+1}^{k} a_j,$$
$$G_{k+1} = \sum_{\ell=0}^{k} q_\ell \left(d_\ell \prod_{j=\ell+1}^{k} c_j + b_\ell \sum_{t=\ell+1}^{k} e_t \prod_{v=t+1}^{k} c_v \prod_{v=\ell+1}^{t-1} a_v \right).$$

Proof. The formula for F_{k+1} follows immediately by induction. For the expansion of G_{k+1} we first rewrite this formula in a more explicit recursive form

$$G_{k+1} = \sum_{\ell=0}^{k} q_\ell d_\ell \prod_{j=\ell+1}^{k} c_j + \sum_{\ell=0}^{k-1} q_\ell b_\ell \sum_{t=\ell+1}^{k} e_t \prod_{v=t+1}^{k} c_v \prod_{v=\ell+1}^{t-1} a_v$$
$$= \sum_{\ell=0}^{k} q_\ell d_\ell \prod_{j=\ell+1}^{k} c_j + \sum_{t=1}^{k} e_t \prod_{v=t+1}^{k} c_v \left(\sum_{\ell=0}^{t-1} q_\ell b_\ell \prod_{v=\ell+1}^{t-1} a_v \right).$$

The part in-between the braces equals F_t and the proof now follows by induction on k. □

A similar result is obtained if the second recursion is replaced by $G_{k+1} = c_k G_k + d_k q_k + d'_k q'_k + e_k F_k$. Furthermore the coefficients in the expansion for q_ℓ and q'_ℓ can cheaply be calculated recursively. Using this lemma, we can make the log-LT construction for the Heston model explicit in the following lemma.

Proposition 1. *The column vector $Q_{\bullet k}$ that solves the optimization problem (3) for a call option under the Heston model is given by $Q_{\bullet k} = \pm \boldsymbol{v}/\|\boldsymbol{v}\|$ where*

$$v_{2\ell+1} = \hat{S}_m f_\ell^2 \rho + \hat{S}_m f_\ell^4 \sum_{t=\ell+1}^{m-1} f_t^1 \prod_{v=\ell+1}^{t-1} f_v^3,$$
$$v_{2\ell+2} = \hat{S}_m f_\ell^2 \sqrt{1-\rho^2},$$

for $\ell = 0, \ldots, m-1$.

Proof. By Imai and Tan [10, Theorem 1] the solution to the optimization problem (3) is given by

$$Q_{\bullet k} = \pm \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|},$$

where $\boldsymbol{v}$ is determined from

$$Q'_{\bullet k} \boldsymbol{v} = \frac{\partial \hat{S}_m}{\partial z_k} = \hat{S}_m \frac{\partial \log \hat{S}_m}{\partial z_k}.$$

With the help of Lemma 1 we find from (4) and (5)

$$\frac{\partial \log \hat{S}_m}{\partial z_k} = \sum_{\ell=0}^{m-1} q_{2\ell+1,k} \left(\rho f_\ell^2 + f_\ell^4 \sum_{t=\ell+1}^{m-1} f_t^1 \prod_{v=\ell+1}^{t-1} f_v^3 \right) + \sum_{\ell=0}^{m-1} q_{2\ell+2,k} \sqrt{1-\rho^2} f_\ell^2,$$

from which the result now follows. □

Note that since $\hat{S}_m$ and all functions f^i depend on $\boldsymbol{Z}$, the vector $\boldsymbol{v}$ changes in each iteration step of (3) as the reference point $\hat{\boldsymbol{z}}$ is changed.

This construction can also be used for a put option with payoff

$$f(z) = K - \hat{S}_m.$$

In case of an arithmetic Asian option, the payoff is given by

$$f(z) = \frac{1}{m}\sum_{j=1}^{m} \hat{S}_j - K.$$

In that case the optimization problem (3) contains the sum of partial derivatives

$$\left.\frac{\partial f}{\partial z_k}\right|_{z=\hat{z}_k} = \frac{1}{m}\sum_{j=1}^{m} \left.\frac{\partial \hat{S}_j}{\partial z_k}\right|_{z=\hat{z}_k}.$$

It is thus straightforward to use the results for the call option in Proposition 1 to construct the transformation matrix for the arithmetic Asian option.

Crucial to our conditional sampling algorithm is that we modify the LT construction by forcing all odd elements in the first column of Q to zero, i.e., $q_{2k+1,1} = 0$ for $k = 0, \ldots, m-1$. This removes the influence of z_1 to Z_k^1 and thus $\hat{V}_k$ for all k. The LT algorithm then finds the orthogonal matrix Q which solves the optimization problem under this extra constraint (which fixes m elements of the $4m^2$). In the next section we will show this leads to an elegant conditional sampling scheme.

Lemma 2. *Under the condition that $q_{2\ell+1,1} = 0$ for $\ell = 0, \ldots, m-1$ we have that the elements $q_{2\ell+2,1}$ all have the same sign.*

Proof. From Proposition 1, for $k = 1$, we find that $q_{2\ell+2,1}$ is proportional to $v_{2\ell+2}$, i.e.,

$$v_{2\ell+2} = \hat{S}_m \sqrt{\hat{V}_\ell}\sqrt{\Delta t}\sqrt{1-\rho^2},$$

which is always positive, and $q_{2\ell+1,1} = v_{2\ell+1} = 0$. Following Proposition 1 we now take $\pm \boldsymbol{v}/\|\boldsymbol{v}\|$ from which the result follows. □

3 Conditional Sampling on Log-LT

For expository reasons assume for now an up-&-out option with barrier B,

$$g(\hat{S}_1, \ldots, \hat{S}_m) = \max\left(f(\hat{S}_1, \ldots, \hat{S}_m), 0\right) \mathbb{I}\left\{\max_k \hat{S}_k < B\right\}. \tag{6}$$

The condition at time t_{k+1} that the asset stays below the barrier can then be written, for $k = 0, \ldots, m-1$, as

$$\begin{aligned}
\log \hat{S}_{k+1} &= \log \hat{S}_k + \left(r - \frac{\hat{V}_k}{2}\right)\Delta t + \sqrt{\hat{V}_k}\sqrt{\Delta t}\left(\rho Z^1_{k+1} + \sqrt{1-\rho^2}Z^2_{k+1}\right) \\
&= \log S_0 + r(k+1)\Delta t - \Delta t \sum_{\ell=0}^{k} \frac{\hat{V}_\ell^2}{2} \\
&\quad + \sum_{\ell=0}^{k} \sqrt{\hat{V}_\ell}\sqrt{\Delta t} \sum_{n=2}^{2m} \left(\rho q_{2\ell+1,n} + \sqrt{1-\rho^2}\, q_{2\ell+2,n}\right) z_n \\
&\quad + z_1 \sqrt{\Delta t}\sqrt{1-\rho^2} \sum_{\ell=0}^{k} \sqrt{\hat{V}_\ell}\, q_{2\ell+2,1} \\
&< \log B,
\end{aligned}$$

where we have used $q_{2\ell+1,1} = 0$. For notational ease we define the function

$$\begin{aligned}
\Gamma_k(B, \boldsymbol{z}_{2:2m}) = &\frac{\log B/S_0 - r(k+1)\Delta t + \Delta t \sum_{\ell=0}^{k} \hat{V}_\ell^2/2}{\sqrt{\Delta t}\sqrt{1-\rho^2}\sum_{\ell=0}^{k}\sqrt{\hat{V}_\ell}\, q_{2\ell+2,1}} \\
&- \frac{\sum_{\ell=0}^{k}\sqrt{\hat{V}_\ell}\sqrt{\Delta t}\sum_{n=2}^{2m}\left(\rho q_{2\ell+1,n} + \sqrt{1-\rho^2}\, q_{2\ell+2,n}\right) z_n}{\sqrt{\Delta t}\sqrt{1-\rho^2}\sum_{\ell=0}^{k}\sqrt{\hat{V}_\ell}\, q_{2\ell+2,1}}. \qquad (7)
\end{aligned}$$

Here the notation $\boldsymbol{z}_{2:2m}$ is used to indicate the dependency on $z_2, \ldots, z_{2m}$, but not z_1. Note that Γ_k depends on all other market parameters as well, but this dependency is supressed not to clutter the formulas. Because of the assumption that $q_{2k+1,1} = 0$ for all k, $\hat{V}$ can be sampled independently of z_1. This means the barrier condition can be written as a single condition on z_1, i.e.,

$$z_1 < \min_k \Gamma_k(B, \boldsymbol{z}_{2:2m}) \quad \text{if all } q_{2\ell+2,1} > 0,$$

and

$$z_1 > \max_k \Gamma_k(B, \boldsymbol{z}_{2:2m}) \quad \text{if all } q_{2\ell+2,1} < 0.$$

The condition on z_1 was here derived for an up-&-out option for ease of exposition. The modifications for more complex barriers can easily be obtained from here. Table 1 gives an overview of the conditions on z_1 for the basic barrier types and shows that these conditions can easily be combined for more complex types.

Table 1 The barrier constraints on z_1 for different types of barriers: up-&-out (U&O), down-&-out (D&O), up-&-in (U&I), down-&-in (D&I) and some combinations.

Type	all $q_{2\ell+2,1} > 0$
U&O (B)	$z_1 < \min_k \Gamma_k(B, z_{2:2m})$
D&O (B)	$z_1 > \max_k \Gamma_k(B, z_{2:2m})$
U&I (B)	$z_1 > \min_k \Gamma_k(B, z_{2:2m})$
D&I (B)	$z_1 < \max_k \Gamma_k(B, z_{2:2m})$
U&O + D&O ($B_1 > B_2$)	$z_1 \in (\max_k \Gamma_k(B_2, z_{2:2m}), \min_k \Gamma_k(B_1, z_{2:2m}))$
U&O + D&I ($B_1 > B_2$)	$z_1 < \min\{\max_k \Gamma_k(B_2, z_{2:2m}), \min_k \Gamma_k(B_1, z_{2:2m})\}$
Type	**all $q_{2\ell+2,1} < 0$**
U&O (B)	$z_1 > \max_k \Gamma_k(B, z_{2:2m})$
D&O (B)	$z_1 < \min_k \Gamma_k(B, z_{2:2m})$
U&I (B)	$z_1 < \max_k \Gamma_k(B, z_{2:2m})$
D&I (B)	$z_1 > \min_k \Gamma_k(B, z_{2:2m})$
U&O + D&O ($B_1 > B_2$)	$z_1 \in (\max_k \Gamma_k(B_1, z_{2:2m}), \min_k \Gamma_k(B_2, z_{2:2m}))$
U&O + D&I ($B_1 > B_2$)	$z_1 > \max\{\max_k \Gamma_k(B_1, z_{2:2m}), \min_k \Gamma_k(B_2, z_{2:2m})\}$

We now show the main results on our conditional sampling scheme. Again, for expository reasons, specialized for the case of the up-&-out option from above. This result can easily be modified for other payout structures in the same spirit as the results in Table 1. The following theorem holds for both the Monte Carlo method as for a randomly shifted quasi-Monte Carlo rule.

Theorem 1. *For the up-&-out option (6) and assuming that we fixed $q_{2\ell+2,1} > 0$ for $\ell = 0, \ldots, m-1$ (see Lemma 2) the approximation based on sampling*

$$\hat{g}(z_1, \ldots, z_m) = \Phi\left(\min_k \Gamma_k(B, z_{2:2m})\right) \max\left(f(\hat{z}_1, z_2, \ldots, z_m), 0\right)$$

where, using the relation $z_1 = \Phi^{-1}(u_1)$,

$$\hat{z}_1 = \Phi^{-1}\left(u_1 \min_k \Gamma_k(B, z_{2:2m})\right), \tag{8}$$

is unbiased. Furthermore, if we denote the respective unconditional method by

$$g(z_1, \ldots, z_m) = \max\left(f(\hat{S}_1, \ldots, \hat{S}_m), 0\right) \mathbb{I}\left\{\max_k \hat{S}_k < B\right\},$$

where the $\hat{S}_1, \ldots, \hat{S}_m$ are obtained directly from $z_1, \ldots, z_m$ without using (8), then, when using the Monte Carlo method or a randomly shifted quasi-Monte Carlo method, the conditional sampling has reduced variance, i.e., $\mathrm{Var}[\hat{g}] \leq \mathrm{Var}[g]$. *Furthermore the inequality is strict if $\mathbb{P}[\max_k \hat{S}_k \geq B] > 0$ and $\mathbb{E}[g] > 0$, i.e., if there is any chance of knock-out and positive payoff.*

Proof. The proof can be constructed similar to [1, Theorems 3–5] from our previous work. □

The previous result shows that the proposed conditional algorithm can never do worse than its unconditional variant. Furthermore, the more chance there is on a knock-out the more effect the conditional algorithm will have. This can be observed in the examples in Sect. 5.

Remark. The conditional sampling was applied to z_1 (or, equivalently, to u_1) to keep the asset from knocking out (or in). Taking it one step further one could try to add an additional bound on z_1, keeping $\boldsymbol{z}_{2:2m}$ constant, in order to force a strictly positive payout. This is more involved than the barrier condition however as for more complicated payoffs than calls and puts there might not exist analytical formulae such as in Table 1 to condition z_1. It is interesting to note that for calls and puts the same formulas can be used as in Table 1, only now restricting the Γ_k functions to $\Gamma_m(K, \boldsymbol{z}_{2:2m})$. Adding this constraint to the existing barrier conditions is straightforward. Root finding methods can be employed for more complex payout structures. See our previous work [1] for a detailed analysis of root finding for Asian options.

4 The Original LT Method for Heston

We mentioned previously that it is essential for our method to switch to log prices. To illustrate the problem, we introduce the LT method for the Heston model as in [10] and we derive also an explicit form of the orthogonal matrix Q (cf. Proposition 1). However, the conditional sampling scheme from the previous section is not applicable. The Euler–Maruyama discretizations for $S(t)$ and $V(t)$ are given by

$$\hat{S}_{k+1} = \hat{S}_k + r\hat{S}_k\Delta t + \sqrt{\hat{V}_k}\hat{S}_k\sqrt{\Delta t}\left(\rho Z^1_{k+1} + \sqrt{1-\rho^2}Z^2_{k+1}\right),$$

$$\hat{V}_{k+1} = \hat{V}_k + (\theta - \hat{V}_k)\kappa\Delta t + \sigma\sqrt{\hat{V}_k}\sqrt{\Delta t}Z^1_{k+1},$$

compare with (1) and (2). For ease of notation, we introduce the following functions:

$$f^1_k = 1 + r\Delta t + \sqrt{\hat{V}_k}\sqrt{\Delta t}\left(\rho Z^1_{k+1} + \sqrt{1-\rho^2}Z^2_{k+1}\right),$$

$$f^2_k = \frac{\hat{S}_k\sqrt{\Delta t}}{2\sqrt{\hat{V}_k}}\left(\rho Z^1_{k+1} + \sqrt{1-\rho^2}Z^2_{k+1}\right),$$

$$f^3_k = \hat{S}_k\sqrt{\hat{V}_k}\sqrt{\Delta t},$$

$$f_k^4 = 1 - \kappa\Delta t + \frac{\sigma\sqrt{\Delta t}}{2\sqrt{\hat{V}_k}} Z_{k+1}^1,$$

$$f_k^5 = \sigma\sqrt{\hat{V}_k}\sqrt{\Delta t}.$$

Note that all the above functions f^i depend on $\boldsymbol{Z}$. The recursion relations for the partial derivatives become

$$\frac{\partial \hat{S}_{k+1}}{\partial z_i} = \frac{\partial \hat{S}_k}{\partial z_i} f_k^1 + \frac{\partial \hat{V}_k}{\partial z_i} f_k^2 + q_{2k+1,i}\rho f_k^3 + q_{2k+2,i}\sqrt{1-\rho^2} f_k^3,$$

$$\frac{\partial \hat{V}_{k+1}}{\partial z_i} = \frac{\partial \hat{V}_k}{\partial z_i} f_k^4 + q_{2k+1,i} f_k^5,$$

for $k = 0, \ldots, m-1$, and initial conditions $\partial \hat{S}_0/\partial z_i = 0$ and $\partial \hat{V}_0/\partial z_i = 0$. With this notation we obtain the LT construction for the Heston model in explicit form.

Proposition 2. *The column vector $\boldsymbol{v} = Q_{\bullet k}$ that maximizes the optimization problem (3) for a call option under the Heston model is given by $Q_{\bullet k} = \pm \boldsymbol{v}/\|\boldsymbol{v}\|$ where*

$$v_{2\ell+1} = f_\ell^3 \rho \prod_{j=\ell+1}^{m-1} f_j^1 + f_\ell^5 \sum_{t=\ell+1}^{m-1} f_t^2 \prod_{v=t+1}^{m-1} f_v^1 \prod_{v=\ell+1}^{t-1} f_v^4,$$

$$v_{2\ell+2} = f_\ell^3 \sqrt{1-\rho^2} \prod_{j=\ell+1}^{m-1} f_j^1,$$

for $\ell = 0, \ldots, m-1$.

Proof. The proof is similar to Proposition 1, again making use of Lemma 1. □

To show the advantage for conditional sampling of the log-LT method (as explained in Sects. 2 and 3) over this version we consider again the up-&-out option with payoff

$$g(\hat{S}_1, \ldots, \hat{S}_m) = \max\left(f(\hat{S}_1, \ldots, \hat{S}_m), 0\right) \mathbb{I}\left\{\max_k \hat{S}_k < B\right\}.$$

The barrier condition at an arbitrary time step t_{k+1} takes the following form:

$$\begin{aligned}\hat{S}_{k+1} &= \hat{S}_k \left(1 + r\Delta t + \sqrt{\hat{V}_k}\sqrt{\Delta t}\left(\rho Z_{k+1}^1 + \sqrt{1-\rho^2} Z_{k+1}^2\right)\right) \\ &= S_0 \prod_{\ell=0}^{k} \left(1 + r\Delta t + \sqrt{\hat{V}_\ell}\sqrt{\Delta t} \sum_{n=1}^{2m} \left(\rho q_{2\ell+1,n} + \sqrt{1-\rho^2} q_{2\ell+2,n}\right) z_n\right)\end{aligned}$$

$$< B.$$

Trying to condition on z_1, as we did in the log-LT model (assuming again $q_{2\ell+1,1} = 0$), leads to the following condition:

$$\prod_{\ell=0}^{k} \left(A_\ell + \sqrt{\hat{V}_\ell}\sqrt{\Delta t}\sqrt{1-\rho^2} q_{2\ell+2,1} z_1 \right) < \frac{B}{S_0}$$

where

$$A_\ell = 1 + r\Delta t + \sqrt{\hat{V}_\ell}\sqrt{\Delta t} \sum_{n=2}^{2m} \left(\rho q_{2\ell+1,n} + \sqrt{1-\rho^2} q_{2\ell+2,n} \right) z_n.$$

To satisfy the condition on z_1, a $k+1$-th order polynomial must be solved in order to find the regions where the above condition holds. To find the global condition, one has to solve polynomials of degrees 1 to m, and then find the overlapping regions where all conditions hold. This quickly becomes impractical and we therefore use the log-LT method which does not have this drawback.

5 Examples

5.1 Up-&-Out Call and Put

Consider the up-&-out call and put options with payoffs

$$P_c(\hat{S}_1, \ldots, \hat{S}_m) = \max\left(\hat{S}_m - K, 0\right) \mathbb{I}\left\{ \max_k \hat{S}_k < B \right\},$$

$$P_p(\hat{S}_1, \ldots, \hat{S}_m) = \max\left(K - \hat{S}_m, 0\right) \mathbb{I}\left\{ \max_k \hat{S}_k < B \right\}.$$

The fixed model parameters are $r = 0\,\%$ and $\kappa = 1$. Furthermore, time is discretized using $m = 250$ steps and thus our stochastic dimension is 500. The results for this example are calculated using a lattice sequence (with generating vector `exod8_base2_m13` from [9] constructed using the algorithm in [2]). The improvements of the standard deviations w.r.t. the Monte Carlo method for different choices of ρ, S_0, $V_0 = \theta = \sigma$, K and B are shown in Table 2. The results for the call and put option seem to be consistent over all choices of parameters: the new conditional scheme (denoted by QMC+LT+CS) improves significantly on the unconditional LT method (denoted by QMC+LT). Note that the QMC+LT method uses the construction of Proposition 2. Adding root finding (denoted by QMC+LT+CS+RF), to force a positive payout, further dramatically improves the

Table 2 Up-&-out call and put. The reported numbers are the standard deviations of the MC method divided by those of the QMC+LT+CS+RF, QMC+LT+CS and QMC+LT methods. The MC method uses 30,720 samples, while the QMC methods use 1,024 samples and 30 independent shifts. The rightmost column denotes the option value.

$(V_0 = \theta = \sigma, \rho, S_0, K, B)$	QMC+LT+CS+RF	QMC+LT+CS	QMC+LT	Value
Call				
$(0.2, -0.5, 90, 80, 100)$	405 %	148 %	98 %	0.09
$(0.2, -0.5, 100, 100, 120)$	502 %	173 %	90 %	0.09
$(0.2, -0.5, 110, 100, 150)$	463 %	231 %	117 %	1.25
$(0.2, 0.5, 90, 80, 100)$	474 %	120 %	124 %	0.08
$(0.2, 0.5, 100, 100, 125)$	446 %	130 %	99 %	0.16
$(0.2, 0.5, 110, 100, 140)$	454 %	166 %	136 %	0.56
$(0.3, -0.5, 90, 80, 100)$	623 %	160 %	82 %	0.05
$(0.3, -0.5, 100, 100, 120)$	590 %	160 %	144 %	0.06
$(0.3, -0.5, 110, 100, 150)$	429 %	246 %	141 %	0.77
$(0.3, 0.5, 90, 80, 100)$	360 %	191 %	106 %	0.05
$(0.3, 0.5, 100, 100, 125)$	353 %	141 %	81 %	0.10
$(0.3, 0.5, 110, 100, 140)$	367 %	142 %	104 %	0.34
Put				
$(0.2, -0.5, 90, 80, 100)$	367 %	331 %	184 %	9.02
$(0.2, -0.5, 100, 100, 105)$	279 %	235 %	126 %	7.76
$(0.2, -0.5, 110, 100, 112)$	298 %	263 %	123 %	4.44
$(0.2, 0.5, 90, 80, 100)$	361 %	376 %	148 %	6.05
$(0.2, 0.5, 100, 100, 105)$	326 %	298 %	131 %	5.33
$(0.2, 0.5, 110, 100, 112)$	317 %	325 %	149 %	2.98
$(0.3, -0.5, 90, 80, 100)$	383 %	348 %	137 %	10.3
$(0.3, -0.5, 100, 100, 105)$	260 %	243 %	144 %	8.65
$(0.3, -0.5, 110, 100, 112)$	214 %	187 %	129 %	5.38
$(0.3, 0.5, 90, 80, 100)$	380 %	294 %	160 %	6.44
$(0.3, 0.5, 100, 100, 105)$	304 %	272 %	174 %	5.57
$(0.3, 0.5, 110, 100, 112)$	305 %	279 %	124 %	3.33

results. The improvement of the QMC+LT+CS method for the put option is even larger than that for the call option. This difference should not come as a surprise: when using conditional sampling on a knock-out option, z_1 is modified such that the asset does not hit the barrier. In case of an up-&-out call option, the asset paths are essentially pushed down in order to achieve this. The payout of the call option however is an increasing function of $\hat{S}_m$, so that pushing the asset paths down has the side-effect of also pushing a lot of paths out of the money. For the put option the reverse is true: the payout is a decreasing function of $\hat{S}_m$, meaning that pushing the paths down will result in more paths ending up in the money. Root finding can be used to control this off-setting effect in case of the call option, this effect is clearly visible in Table 2. These numerical results are illustrated in terms of N in Fig. 1 for two parameter choices for the call option.

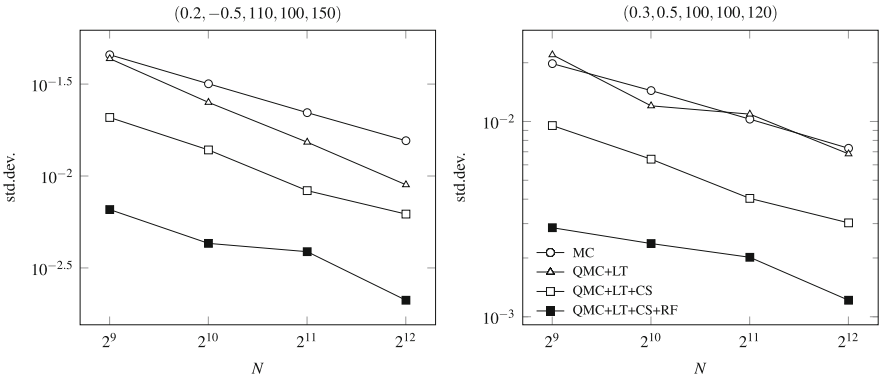


Fig. 1 Up-&-out call convergence plots for two options with different parameters. The fixed parameters are $r = 0\,\%$ and $\kappa = 1$. The different choices for $(V_0 = \theta = \sigma, \rho, S_0, K, B)$ are denoted above the figures.

5.2 Up-&-In Call

Consider an up-&-in call option with payoff

$$P(\hat{S}_1, \ldots, \hat{S}_m) = \max\left(\hat{S}_m - K, 0\right) \mathbb{I}\left\{\max_k \hat{S}_k > B\right\}.$$

The fixed model parameters are $r = 2\,\%$, $\kappa = 1$ and $\sigma = 0.2$. Again, $m = 250$. Here we use the Sobol' sequence with parameters from [11] and digital shifting [3]. The standard deviations for different choices of ρ, S_0, $V_0 = \theta$, K and B are shown in Table 3. The improvements of the conditional scheme are extremely high for this case. Note the impact of the correlation on the results: the improvement for $\rho = 0.5$ is even approximately twice that for $\rho = -0.5$. All parameter choices indicate that conditional sampling on the barrier condition greatly improves accuracy. Adding the additional condition of the payout itself (root finding) provides another serious reduction in the standard deviation.

5.3 Up-&-Out Asian

Consider an up-&-out Asian option with payoff

$$P(\hat{S}_1, \ldots, \hat{S}_m) = \max\left(\frac{1}{m}\sum_{k=1}^{m} \hat{S}_k - K, 0\right) \mathbb{I}\left\{\max_k \hat{S}_k < B\right\}.$$

The fixed model parameters are $r = 5\,\%$, $\kappa = 1$ and $\sigma = 0.2$. The number of time steps is fixed at $m = 250$. We use the Sobol' sequence as in the previous example and the results are shown in Table 4. The results are once more very satisfactory

Table 3 Up-&-in call. The reported numbers are the standard deviations of the MC method divided by those of the QMC+LT+CS+RF, QMC+LT+CS and QMC+LT methods. The MC method uses 30,720 samples, while the QMC+LT+CS+RF, QMC+LT+CS and QMC+LT methods use 1,024 samples and 30 independent shifts. The rightmost column denotes the option value.

$(V_0 = \theta, \rho, S_0, K, B)$	QMC+LT+CS+RF	QMC+LT+CS	QMC+LT	Value
$(0.1, -0.5, 90, 80, 160)$	2,158 %	1,515 %	242 %	5.47
$(0.1, -0.5, 100, 100, 180)$	2,377 %	1,542 %	240 %	5.05
$(0.1, -0.5, 110, 120, 200)$	2,572 %	1,545 %	250 %	4.74
$(0.1, 0.5, 90, 80, 160)$	1,557 %	654 %	341 %	17.4
$(0.1, 0.5, 100, 100, 180)$	1,564 %	644 %	354 %	16.9
$(0.1, 0.5, 110, 120, 200)$	1,556 %	640 %	373 %	16.6
$(0.15, -0.5, 90, 80, 160)$	2,044 %	1,247 %	366 %	10.6
$(0.15, -0.5, 100, 100, 180)$	2,243 %	1,262 %	420 %	10.1
$(0.15, -0.5, 110, 120, 200)$	2,391 %	1,236 %	349 %	9.72
$(0.15, 0.5, 90, 80, 160)$	1,570 %	568 %	421 %	23.3
$(0.15, 0.5, 100, 100, 180)$	1,622 %	567 %	418 %	23.0
$(0.15, 0.5, 110, 120, 200)$	1,649 %	562 %	366 %	22.9

Table 4 Up-&-out Asian call. The reported numbers are the standard deviations of the MC method divided by those of the QMC+LT+CS+RF, QMC+LT+CS and QMC+LT methods. The MC method uses 30,720 samples, while the QMC+LT+CS+RF, QMC+LT+CS and QMC+LT methods use 1,024 samples and 30 independent shifts. The rightmost column denotes the option value.

$(V_0 = \theta, \rho, S_0, K, B)$	QMC+LT+CS+RF	QMC+LT+CS	QMC+LT	Value
$(0.1, -0.5, 90, 80, 120)$	483 %	329 %	154 %	1.70
$(0.1, -0.5, 100, 100, 140)$	461 %	245 %	185 %	0.77
$(0.1, -0.5, 110, 120, 160)$	404 %	189 %	110 %	0.30
$(0.1, 0.5, 90, 80, 120)$	392 %	328 %	144 %	1.34
$(0.1, 0.5, 100, 100, 140)$	414 %	252 %	115 %	0.53
$(0.1, 0.5, 110, 120, 160)$	502 %	209 %	133 %	0.18
$(0.15, -0.5, 90, 80, 120)$	463 %	247 %	143 %	0.77
$(0.15, -0.5, 100, 100, 140)$	425 %	183 %	125 %	0.29
$(0.15, -0.5, 110, 120, 160)$	389 %	161 %	93 %	0.10
$(0.15, 0.5, 90, 80, 120)$	416 %	257 %	111 %	0.61
$(0.15, 0.5, 100, 100, 140)$	486 %	201 %	119 %	0.20
$(0.15, 0.5, 110, 120, 160)$	528 %	171 %	108 %	0.05

with similar results as for the up-&-out call and put options in Table 2. Figure 2 shows the convergence behaviour for two sets of parameter choices. As before, a significant variance reduction can be seen for our conditional sampling scheme and the root finding method further improves this result.

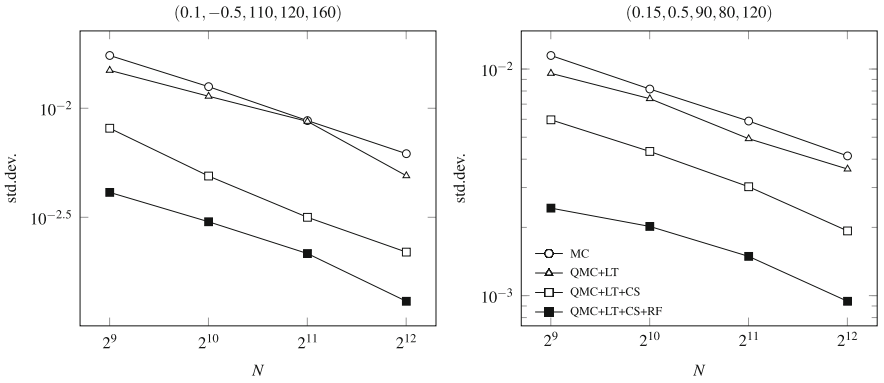


Fig. 2 Up-&-out Asian call convergence plots for two options with different parameters. The fixed parameters are $r = 5\,\%$, $\kappa = 1$ and $\sigma = 0.2$. The different choices for $(V_0 = \theta, \rho, S_0, K, B)$ are denoted above the figures.

6 Conclusion and Outlook

The conditional sampling scheme for the LT method introduced in [1] for the Black–Scholes model has been extended to the Heston model. This was done by considering log prices and making the sampling of the volatility process independent of z_1. We also obtained explicit constructions for the matrix Q of the LT method. The numerical results show that the method is very effective in reducing variance and outperforms the LT method by a huge margin. We only considered an Euler–Maruyama discretization scheme for the asset and volatility processes. It might be interesting to see if the theory and results carry over when other simulation methods are used, see [7] for an overview of other methods.

Acknowledgements This research is part of a project funded by the Research Fund KU Leuven. Dirk Nuyens is a fellow of the Research Foundation Flanders (FWO). This paper presents research results of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office.

References

1. Achtsis, N., Cools, R., Nuyens, D.: Conditional sampling for barrier option pricing under the LT method. SIAM J. Finan. Math. **4**, 327–352 (2013)
2. Cools, R., Kuo, F.Y., Nuyens, D.: Constructing embedded lattice rules for multivariate integration. SIAM J. Sci. Comput. **28**, 2162–2188 (2006)
3. Dick, J., Pillichshammer, F.: Digital Nets and Sequences: Discrepancy Theory and Quasi-Monte Carlo Integration. Cambridge University Press, New York (2010)

4. Giles, M.B., Kuo, F.Y., Sloan, I.H., Waterhouse, B.J.: Quasi-Monte Carlo for finance applications. ANZIAM J. **50**, 308–323 (2008)
5. Glasserman, P.: Monte Carlo Methods in Financial Engineering. Springer, New York (2003)
6. Glasserman, P., Staum, J.: Conditioning on one-step survival for barrier option simulations. Oper. Res. **49**, 923–937 (2001)
7. Van Haastrecht, A., Pelsser, A.A.J.: Efficient, almost exact simulation of the Heston stochastic volatility model. Int. J. Theor. Appl. Finance **31**, 1–43 (2010)
8. Heston, S.L.: A closed-form solution for options with stochastic volatility with applications to bond and currency options. Rev. Financ. Stud. **6**, 327–343 (1993)
9. http://people.cs.kuleuven.be/~dirk.nuyens/qmc-generators (27/07/2012)
10. Imai, J., Tan, K.S.: A general dimension reduction technique for derivative pricing. J. Comput. Finance **10**, 129–155 (2006)
11. Joe, S., Kuo, F.Y.: Constructing Sobol' sequences with better two-dimensional projections. SIAM J. Sci. Comput. **30**, 2635–2654 (2008)
12. Kloeden, P.E., Platen, E.: Numerical Solution of Stochastic Differential Equations. Springer, Berlin/New York (1992)
13. L'Écuyer, P.: Quasi-Monte Carlo methods with applications in finance. Finance Stoch. **13**, 307–349 (2009)
14. Nuyens, D., Waterhouse, B.J.: A global adaptive quasi-Monte Carlo algorithm for functions of low truncation dimension applied to problems from finance. In: Plaskota, L., Woźniakowski, H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2010, pp. 589–607. Springer, Berlin/Heidelber (2012)

Probabilistic Star Discrepancy Bounds for Double Infinite Random Matrices

Christoph Aistleitner and Markus Weimar

Abstract In 2001 Heinrich, Novak, Wasilkowski and Woźniakowski proved that the inverse of the discrepancy depends linearly on the dimension, by showing that a Monte Carlo point set $\mathscr{P}$ of N points in the s-dimensional unit cube satisfies the discrepancy bound $D_N^{*s}(\mathscr{P}) \leq c_{\text{abs}} s^{1/2} N^{-1/2}$ with positive probability. Later their results were generalized by Dick to the case of double infinite random matrices. In the present paper we give asymptotically optimal bounds for the discrepancy of such random matrices, and give estimates for the corresponding probabilities. In particular we prove that the $N \times s$-dimensional projections $\mathscr{P}_{N,s}$ of a double infinite random matrix satisfy the discrepancy estimate

$$D_N^{*s}(\mathscr{P}_{N,s}) \leq \left(2130 + 308 \frac{\ln \ln N}{s}\right)^{1/2} s^{1/2} N^{-1/2}$$

for all N and s with positive probability. This improves the bound $D_N^{*s}(\mathscr{P}_{N,s}) \leq (c_{\text{abs}} \ln N)^{1/2} s^{1/2} N^{-1/2}$ given by Dick. Additionally, we show how our approach can be used to show the existence of completely uniformly distributed sequences of small discrepancy which find applications in Markov Chain Monte Carlo.

C. Aistleitner (✉)
Institute of Mathematics A, Graz University of Technology, Steyrergasse 30, 8010 Graz, Austria

Current address: School of Mathematics and Statistics, University of New South Wales, Sydney, NSW, Australia
e-mail: aistleitner@math.tugraz.at

M. Weimar
Institute of Mathematics, Friedrich-Schiller-University Jena, Ernst-Abbe-Platz 2, 07743 Jena, Germany

Current address: Faculty of Mathematics and Computer Science, Workgroup Numerics and Optimization, Philipps-University Marburg, Hans-Meerwein-Straße, Lahnberge 35032 Marburg, Germany
e-mail: weimar@mathematik.uni-marburg.de

J. Dick et al. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2012*,
Springer Proceedings in Mathematics & Statistics 65,
DOI 10.1007/978-3-642-41095-6_10, © Springer-Verlag Berlin Heidelberg 2013

1 Introduction and Statement of Results

1.1 Uniform Distribution and Discrepancy

Let x, y be two elements of the s-dimensional unit cube $[0,1]^s$. We write $x \leq y$ if this inequality holds coordinatewise, and $x < y$ if all coordinates of x are smaller than the corresponding coordinates of y. Furthermore, $[x, y)$ denotes the set $\{z \in [0,1]^s \mid x \leq z < y\}$. We write 0 for the s-dimensional vector $(0, \ldots, 0)$, and thus $[0, x)$ denotes the set $\{z \in [0,1]^s \mid 0 \leq z < x\}$. Throughout the paper we will use the same notation for real numbers and for real vectors; it will be clear from the context what we mean. Moreover, by c_{abs} we will denote universal constants which may change at every occurrence.

A sequence $(x_n)_{n\in\mathbb{N}}$ of points from $[0,1]^s$ is called *uniformly distributed* (modulo 1) if for any $x \in [0,1]^s$ the asymptotic equality

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}_{[0,x)}(x_n) = \lambda([0,x)) \tag{1}$$

holds. Here $\mathbb{N}$ denotes the set of positive integers, and λ denotes the s-dimensional Lebesgue measure. By an observation of Weyl [23] a sequence is uniformly distributed if and only if

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(x_n) = \int_{[0,1]^s} f(x)\,\mathrm{d}x \tag{2}$$

for any continuous s-dimensional function f. This interrelation already suggests that uniformly distributed sequences can be used for numerical integration—an idea which is the origin of the so-called *Quasi-Monte Carlo* (QMC) *method* for numerical integration. The speed of convergence in (1) and (2) can be measured by means of the *star discrepancy* of the point sequence $(x_n)_{n\in\mathbb{N}} \subset [0,1]^s$, which is defined as

$$D_N^{*s}(x_1, \ldots, x_N) = \sup_{x\in[0,1]^s} \left| \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}_{[0,x)}(x_n) - \lambda([0,x)) \right|, \quad N \in \mathbb{N}. \tag{3}$$

A sequence is uniformly distributed if and only if the discrepancy of its first N elements tends to 0 as $N \to \infty$.

The Koksma-Hlawka inequality states that the deviation between the finite average

$$\frac{1}{N} \sum_{n=1}^{N} f(x_n)$$

and the integral of a function f can be estimated by the product of the star discrepancy of the point set $\{x_1, \ldots, x_N\}$ and the variation (in the sense of Hardy and Krause) of f; see [7, 15, 17] for details, as well as for a general introduction to uniform distribution theory and discrepancy theory. Thus, as a rule of thumb it is reasonable to perform Quasi-Monte Carlo integration by using point sets having small discrepancy. There exist many constructions of so–called *low-discrepancy point sets* and *low-discrepancy sequences*, where for many decades the main focus of research was set on finding point sets and sequences satisfying strong discrepancy bounds for large N and fixed s; however, recently, the problem asking for point sets having small discrepancy for a moderate number of points in comparison with the dimension has attracted some attention.

From a probabilistic point of view, a sequence $(x_n)_{n\in\mathbb{N}}$ is uniformly distributed if the corresponding sequence of empirical distribution functions converges to the uniform distribution. In particular, by the Glivenko-Cantelli theorem a random sequence is almost surely uniformly distributed.

1.2 The Inverse of the Discrepancy

Let $n^*(s, \varepsilon)$ denote the smallest possible size of a set of s-dimensional points having star discrepancy not exceeding ε. This quantity is called the *inverse of the discrepancy*. By a profound result of Heinrich, Novak, Wasilkowski and Woźniakowski [14] we know that

$$n^*(s, \varepsilon) \leq c_{\text{abs}} s\, \varepsilon^{-2}. \tag{4}$$

This upper bound is complemented by a lower bound of Hinrichs [13], stating that

$$n^*(s, \varepsilon) \geq c_{\text{abs}} s\, \varepsilon^{-1}. \tag{5}$$

Together, (4) and (5) give a complete description of the dependence of the inverse of the star discrepancy on the dimension s, while the precise dependence on ε is still an important open problem. For their proof Heinrich et al. use deep results of Haussler [12] and Talagrand [22]. In fact, what they exactly prove is that a randomly generated sequence satisfies (4) with positive probability. The upper bound in (4) is equivalent to the fact that for any N and s there exists a set of N points in $[0, 1]^s$ satisfying the discrepancy bound

$$D_N^{*s} \leq c_{\text{abs}} \frac{\sqrt{s}}{\sqrt{N}}. \tag{6}$$

For more details on the inverse of the discrepancy and on feasibility of Quasi-Monte Carlo integration we refer to [11, 18, 19].

1.3 Double Infinite Matrices

Dick [4] observed that the probabilities of the exceptional sets in the argument of Heinrich et al. to prove (6) are summable over s and N, if the factor N is replaced by $N \ln N$. More precisely, he proved that with positive probability all the $N \times s$-dimensional projections of a randomly generated double infinite matrix $(X_{n,i})_{n,i\in\mathbb{N}}$ satisfy

$$D_N^{*s} \leq c_{\text{abs}} \sqrt{\ln N} \frac{\sqrt{s}}{\sqrt{N}}. \tag{7}$$

Dick's result has been slightly improved by Doerr, Gnewuch, Kritzer and Pillichshammer [6], again for randomly generated matrices. It is clear that such a randomly generated matrix cannot achieve the discrepancy bound (6) uniformly in s and N, since by Philipp's law of the iterated logarithm [20] for any sequence $(X_n)_{n\in\mathbb{N}}$ of independent, uniformly distributed random vectors

$$\limsup_{N\to\infty} \frac{\sqrt{N} D_N^{*s}(X_1, \ldots, X_N)}{\sqrt{\ln \ln N}} = \frac{1}{\sqrt{2}} \quad \text{almost surely.} \tag{8}$$

Thus, the factor $\sqrt{\ln N}$ in (7) cannot be reduced to a function from the class $o\left(\sqrt{\ln \ln N}\right)$, since by (8) no positive probability can exist for a random matrix satisfying such an asymptotic discrepancy bound. However, there exists a double infinite matrix constructed in a hybrid way (that is, consisting of both random and deterministic entries) whose $N \times s$-dimensional projections satisfy

$$D_N^{*s} \leq c_{\text{abs}} \frac{\sqrt{s}}{\sqrt{N}}$$

uniformly in N and s, see [2]. The purpose of the present paper is to find optimal discrepancy bounds which hold for *random* double infinite matrices with positive probability, and to give estimates for the corresponding probabilities.

1.4 Complete Uniform Distribution and Markov Chain Monte Carlo

A sequence $(x_n)_{n\in\mathbb{N}}$ of numbers from $[0, 1]$ is called *completely uniformly distributed* (c.u.d.), if for any s the sequences

$$((x_n, \ldots, x_{n+s-1}))_{n\in\mathbb{N}} \subset [0, 1]^s$$

are uniformly distributed. This property was suggested by Knuth as a test for pseudorandomness of sequences in volume II of his celebrated monograph on

The Art of Computer Programming. However, in our context it is more sensible to use an non-overlapping version of the above construction, namely to use the first Ns elements of an infinite sequence $(x_n)_{n\in\mathbb{N}}$ to construct N points $u_1^{(s)}, \dots, u_N^{(s)} \in [0,1]^s$ in the form

$$\begin{aligned} u_1^{(s)} &= (x_1, \dots, x_s), \\ u_2^{(s)} &= (x_{s+1}, \dots, x_{2s}), \\ &\vdots \\ u_N^{(s)} &= (x_{(N-1)s+1}, \dots, x_{Ns}). \end{aligned} \tag{9}$$

These two notions of complete uniform distribution are equivalent insofar as a sequence is c.u.d. in the first sense if and only if it is c.u.d. in the second sense.

In many practical applications, e.g., in financial mathematics, a general integral of the form

$$\int_\Omega f(y)\,\mathrm{d}\mu(y)$$

for some measure space Ω and some measure μ can be transferred to the form

$$\int_{[0,1]^s} \hat{f}(y)\,\mathrm{d}y. \tag{10}$$

That is, the original function f (which can be, for example, the payoff-function of some financial derivative, where the properties of the underlying problem are described by μ) has to be replaced by a new function $\hat{f}$, which contains all the information about the change of measure from μ to λ. If $\hat{f}$ can be easily calculated and is a well-behaved function, then the integral in (10) can be directly computed using classical QMC methods. However, in many cases the function $\hat{f}$ will be difficult to handle, and it is computationally easier to directly calculate the integral

$$\int_\Omega f(y)\,\mathrm{d}\mu(y) = \int_\Omega f(y)\pi(y)\,\mathrm{d}y, \tag{11}$$

where π is the density function of μ, by sampling random variables having density π. In other words, it is necessary to sample random variables having density π, which may not be directly possible by standard methods. This problem can be solved by using *Markov Chain Monte Carlo* (MCMC). Here y_0 is a (random) starting element, and the other samples y_n are constructed iteratively in the form $y_n = \Phi(y_{n-1}, u_n)$, where $u_n \in [0,1]^s$ and Φ is an appropriate function. The distribution of $(y_n|y_0, \dots, y_{n-1})$ is the same as the distribution of $(y_n|y_{n-1})$, which means that the sequence $(y_n)_{n\in\mathbb{N}}$ has the Markov property. Then, if π is the density of the stationary distribution of $(y_n)_{n\in\mathbb{N}}$ under Φ, the integral (11) can be estimated by

$$\frac{1}{N}\sum_{n=1}^{N} f(y_n).$$

For more background information on MCMC we refer to [16, 21].

Traditionally, the points $(u_n)_{n\in\mathbb{N}} \subset [0,1]^s$ in the aforementioned construction are sampled randomly. However, Chen, Dick and Owen [3] recently showed that it is also possible to use quasi-random points instead, namely by choosing $u_n = u_n^{(s)}$, $n \in \mathbb{N}$, constructed out of a completely uniformly distributed sequence $(x_n)_{n\in\mathbb{N}}$ according to (9). Then, under some regularity assumptions, the MCMC sampler consistently samples points having density π, provided the discrepancy of the c.u.d.-sequence is sufficiently small. The results of [3] are of a merely qualitative nature, stating that certain MCMC-methods are consistent if the discrepancy of the QMC-points, constructed according to (9), tends to zero. However, it is natural to assume that the speed of convergence of these MCMC samplers can be estimated by the speed of decay of the discrepancy of $\left\{u_n^{(s)} \mid 1 \le n \le N\right\}$, and therefore it is desirable to find sequences $(x_n)_{n\in\mathbb{N}}$ for which this discrepancy is small. In [3] it is noted that Dick's proof from [4] can be modified to prove the existence of a sequence $(x_n)_{n\in\mathbb{N}}$ for which

$$D_N^{*s}\left(u_1^{(s)},\dots,u_N^{(s)}\right) \le c_{\mathrm{abs}}\sqrt{\ln N}\,\frac{\sqrt{s}}{\sqrt{N}},$$

uniformly in N and s. In the present paper we will show that the factor $\sqrt{\ln N}$ can be reduced to $\sqrt{c_{\mathrm{abs}} + (\ln\ln N)/s}$, which is already very close to the upper bound of Heinrich et al. in (6). We tried to find a hybrid construction achieving (6), similar to the hybrid construction of a double infinite matrix mentioned at the end of the previous section, but due to the complicated dependence between the diverse coordinates of the point sets $\left\{u_n^{(s_1)} \mid 1 \le n \le N_1\right\}$ and $\left\{u_n^{(s_2)} \mid 1 \le n \le N_2\right\}$ for different s_1, s_2 and N_1, N_2 this seems to be hopeless. The discrepancy bound in our Theorem 2 below is the strongest known discrepancy bound for c.u.d.-sequences (which is valid uniformly in N and s) at present. Furthermore, Dick's result is of limited practical use as it involves unknown constants, while our results are completely explicit and even allow to calculate the probability of a random sequence satisfying the desired discrepancy bounds.

1.5 Results

Let $X = (X_{n,i})_{n,i\in\mathbb{N}}$ be a double infinite array of independent copies of some uniformly $[0,1]$-distributed random variable. For positive integers N and s set

$$\mathscr{P}_{N,s} = \left\{X^{(1)},\dots,X^{(N)}\right\},$$

where $X^{(n)} = (X_{n,1}, \ldots, X_{n,s}) \in [0,1]^s$ for $n = 1, \ldots, N$. Hence, $\mathscr{P}_{N,s}$ is the projection of X onto its first $N \times s$ entries. As in (3) let $D_N^{*s}(\mathscr{P}_{N,s})$ denote the s-dimensional star discrepancy of these N points.

The main technical tool of the present paper is the following Lemma 1, which will be used to derive our theorems.

Lemma 1. *Let $\alpha \geq 1$ and $\beta \geq 0$ be given. Moreover, for $M, s \in \mathbb{N}$ set*

$$\Omega_{M,s} = \left\{ \max_{2^M \leq N < 2^{M+1}} N \cdot D_N^{*s}(\mathscr{P}_{N,s}) > \sqrt{\alpha A + \beta B \, \frac{\ln M}{s}} \sqrt{s \cdot 2^M} \right\},$$

where $A = 1165$ and $B = 178$. Then we have for all natural numbers M and s

$$\mathbb{P}(\Omega_{M,s}) < \frac{1}{(1+s)^\alpha} \frac{1}{M^\beta}.$$

The proof of Lemma 1, which is given in Sect. 2 below, essentially follows the lines of [1]. In addition we use a Bernstein type inequality which can be found, e.g., in Einmahl and Mason [8, Lemma 2.2]:

Lemma 2 (Maximal Bernstein inequality). *For $M \in \mathbb{N}$ let Z_n, $1 \leq n \leq 2^{M+1}$, be independent random variables with zero mean and variance $\mathbb{V}(Z_n)$. Moreover, assume $|Z_n| \leq C$ for some $C > 0$ and all $n \in \{1, \ldots, 2^{M+1}\}$. Then for every $t \geq 0$*

$$\mathbb{P}\left(\max_{1 \leq N \leq 2^{M+1}} \sum_{n=1}^{N} Z_n > t \right) \leq \exp\left(-t^2 / \left(2 \sum_{n=1}^{2^{M+1}} \mathbb{V}(Z_n) + 2Ct/3 \right) \right).$$

At the end of Sect. 2, we will conclude the following two theorems from Lemma 1. Here ζ denotes the Riemann Zeta function.

Theorem 1. *Let $\gamma \geq \zeta^{-1}(2) \approx 1.73$ be arbitrarily fixed. Then with probability strictly larger than $1 - (\zeta(\gamma) - 1)^2 \geq 0$ we have for all $s \in \mathbb{N}$ and every $N \geq 2$*

$$D_N^{*s}(\mathscr{P}_{N,s}) \leq \sqrt{\gamma} \cdot \sqrt{1165 + 178 \frac{\ln \log_2 N}{s}} \cdot \sqrt{\frac{s}{N}}.$$

In particular, there exists a positive probability that a random matrix X satisfies for all $s \in \mathbb{N}$ and every $N \geq 2$

$$D_N^{*s}(\mathscr{P}_{N,s}) \leq \sqrt{2130 + 308 \frac{\ln \ln N}{s}} \cdot \sqrt{\frac{s}{N}}.$$

In our second theorem, we show how our method can be applied to obtain discrepancy bounds for completely uniformly distributed sequences. To this end let $X = (X_n)_{n \in \mathbb{N}}$ be a sequence of independent, identically distributed random

variables having uniform distribution on $[0,1]$. For any $s \in \mathbb{N}$ and $N \geq 2$ define a sequence

$$\begin{aligned} U_1^{(s)} &= (X_1, \dots, X_s), \\ U_2^{(s)} &= (X_{s+1}, \dots, X_{2s}), \\ &\vdots \\ U_N^{(s)} &= (X_{(N-1)s+1}, \dots, X_{Ns}). \end{aligned}$$

Furthermore, let

$$\mathscr{U}_{N,s} = \left\{ U_1^{(s)}, \dots, U_N^{(s)} \right\}. \tag{12}$$

Theorem 2. *Let $\gamma \geq \zeta^{-1}(2)$ be arbitrarily fixed. Then with probability strictly larger than $1 - (\zeta(\gamma) - 1)^2 \geq 0$ we have for all $s \in \mathbb{N}$ and every $N \geq 2$*

$$D_N^{*s}(\mathscr{U}_{N,s}) \leq \sqrt{\gamma} \cdot \sqrt{1165 + 178 \frac{\ln \log_2 N}{s}} \cdot \sqrt{\frac{s}{N}}.$$

In particular, there exists a positive probability that a random sequence X is completely uniformly distributed and satisfies for all $s \in \mathbb{N}$ and every $N \geq 2$

$$D_N^{*s}(\mathscr{U}_{N,s}) \leq \sqrt{2130 + 308 \frac{\ln \ln N}{s}} \cdot \sqrt{\frac{s}{N}}.$$

Our results are essentially optimal in two respects. On the one hand, for any N and s satisfying $N \leq \exp(\exp(c_{\text{abs}} s))$ our Theorem 1 gives (with positive probability) a discrepancy estimate of the form

$$D_N^{*s}(\mathscr{P}_{N,s}) \leq c_{\text{abs}} \frac{\sqrt{s}}{\sqrt{N}} \tag{13}$$

and by this means resembles the aforementioned result of Heinrich et al. (note that a discrepancy estimate of the form (13) is not of much use if $N > \exp(\exp(c_{\text{abs}} s))$, since in this case the well-known bounds for low-discrepancy sequences are much smaller). Hence, any improvement of our Theorem 1 (up to the values of the constants) would require an improvement of (4). Furthermore, recent research of Doerr [5] shows that the expected value of the star-discrepancy of a set of N points in $[0,1]^s$ is of order $s^{1/2} N^{-1/2}$. The probability estiamtes in [5] can be used to show that for an i.i.d. random matrix X there exist absolute constants K_1, K_2 such that for every s the probability, that the $N \times s$-dimensional projections of X have a discrepancy bounded by $K_1 s^{1/2} N^{-1/2}$ for all $N \leq \exp(\exp(K_2 s))$, is zero. This means that for N in this range our Theorem 1 is essentially optimal. It should also

be mentioned that it is possible that (4) is already optimal and cannot be improved. On the other hand, for fixed s and large N our discrepancy estimate is of the form

$$D_N^{*s}(\mathscr{P}_{N,s}) \leq c(s)\frac{\sqrt{\ln\ln N}}{\sqrt{N}}.$$

This discrepancy bound is asymptotically optimal for random matrices (up to the value of the constants c(s)), since in view of the law of the iterated logarithm (8) no random construction can achieve a significantly better rate of decay of the discrepancy with positive probability.

We note that constructing a sequence of elements of $[0,1]$ satisfying good discrepancy bounds in the sense of complete uniform distribution is much more difficult than constructing point sets in $[0,1]^s$ for a fixed number of points and fixed dimension s. There exist constructions of sequences having good c.u.d.-behavior, but usually the corresponding discrepancy bounds are only useful if N is much larger than s; it is possible that the discrepancy estimates in Theorem 2 are optimal in the sense that they give good results uniformly for *all* possible values of N and s, and that Theorem 2 cannot be significantly improved (up to the values of the constants) in this regard.

2 Proofs

Proof of Lemma 1. Since the proof is somewhat technical we split it into different steps. The main ingredients in the proof are a dyadic decomposition of the unit cube, which was introduced in [1], a maximal version of Bernstein's inequality (Lemma 2), which is also used to prove the law of the iterated logarithm in probability theory, and Dick's observation from [4] that the exceptional probabilities are exponentially decreasing in s and are therefore summable over s. Our results could not be proved using the method in [14] (which is tailor-made for fixed N and s), since there does not exist a maximal version of Talagrand's large deviations inequality for empirical processes, which is the crucial ingredient in [14].

Step 1. Let $M, s \in \mathbb{N}$ be fixed. Without loss of generality we can assume

$$\frac{1}{2}\sqrt{\alpha A + \beta B\,\frac{\ln M}{s}}\sqrt{\frac{s}{2^M}} < 1 \tag{14}$$

because otherwise

$$\Omega_{M,s} \subseteq \left\{ \max_{2^M \leq N < 2^{M+1}} N \cdot D_N^{*s}(\mathscr{P}_{N,s}) > 2^{M+1} \right\} = \emptyset.$$

For a moment assume $L \geq 2$ to be given. Let $(a_k)_{k=-1}^{L}$ and $(b_k)_{k=-1}^{L}$ be two non-negative, non-increasing sequences such that

$$A \geq 2\left(\sum_{k=-1}^{L}\sqrt{a_k}\right)^2 \quad \text{and} \quad B \geq 2\left(\sum_{k=-1}^{L}\sqrt{b_k}\right)^2 \tag{15}$$

and set

$$y_k = \alpha a_k + \beta b_k \frac{\ln M}{s} \quad \text{and} \quad t_k = \sqrt{y_k}\sqrt{s\cdot 2^M}. \tag{16}$$

Hence, using (14), as well as (15), we have

$$1 > \frac{1}{\sqrt{2}}\sqrt{\alpha a_{-1} + \beta b_{-1}\frac{\ln M}{s}}\sqrt{\frac{s}{2^M}} = 2\cdot\frac{1}{2\sqrt{2}}\sqrt{y_{-1}}\sqrt{\frac{s}{2^M}}.$$

If we choose $L \in \mathbb{N}$ such that

$$\frac{1}{2}\left(\frac{1}{2\sqrt{2}}\sqrt{y_{-1}}\sqrt{\frac{s}{2^M}}\right) < 2^{-L} \leq \frac{1}{2\sqrt{2}}\sqrt{y_{-1}}\sqrt{\frac{s}{2^M}} \tag{17}$$

this implies $L \geq 2$.

Since the square root function is sublinear and concave we may use Jensen's inequality to obtain

$$\begin{aligned}\sum_{k=-1}^{L}\sqrt{y_k} &\leq \sum_{k=-1}^{L}\sqrt{\alpha a_k} + \sum_{k=-1}^{L}\sqrt{\beta b_k\frac{\ln M}{s}}\\ &= \sqrt{\alpha\left(\sum\nolimits_{k=-1}^{L}\sqrt{a_k}\right)^2} + \sqrt{\beta\left(\sum\nolimits_{k=-1}^{L}\sqrt{b_k}\right)^2\frac{\ln M}{s}}\\ &\leq \sqrt{\alpha 2\left(\sum\nolimits_{k=-1}^{L}\sqrt{a_k}\right)^2 + \beta 2\left(\sum\nolimits_{k=-1}^{L}\sqrt{b_k}\right)^2\frac{\ln M}{s}}\\ &\leq \sqrt{\alpha A + \beta B\frac{\ln M}{s}}\end{aligned} \tag{18}$$

out of (15) and the definition of y_k.

Step 2. In what follows we use a decomposition of the s-dimensional unit cube in terms of δ-covers and δ-bracketing covers. A detailed description of this decomposition can be found in [1]. We briefly sketch the main points.

For any given $\delta \in (0,1]$ a finite set Γ of points in $[0,1]^s$ is called a *δ-cover* of $[0,1]^s$ if for every $y \in [0,1]^s$ there exist two elements $x, z \in \Gamma$ such that $x \leq y \leq z$ and $\lambda([0,z)\backslash[0,x)) \leq \delta$. Furthermore, a finite set Δ of pairs of points from $[0,1]^s$ is called a *δ-bracketing cover* if for every pair $(x,z) \in \Delta$ we have $\lambda([0,z)\backslash[0,x)) \leq \delta$, and if for every $y \in [0,1]^s$ there exists a pair $(x,z) \in \Delta$

such that $x \le y \le z$. The concepts of δ-covers and δ-bracketing covers were investigated in detail in [9].
For $1 \le k < L$ let Γ_k denote a 2^{-k}-cover of $[0,1]^s$. Moreover, let Δ_L denote a 2^{-L}-bracketing cover of $[0,1]^s$. For notational convenience we set

$$\Gamma_L = \left\{ p_L \in [0,1]^s \,\middle|\, (p_L, p_{L+1}) \in \Delta_L \text{ for some } p_{L+1} \right\},$$
$$\Gamma_{L+1} = \left\{ p_{L+1} \in [0,1]^s \,\middle|\, (p_L, p_{L+1}) \in \Delta_L \text{ for some } p_L \right\}$$

and $p_0 = 0 \in [0,1]^s$. Furthermore, for points $a \le b$ in $[0,1]^s$ we define

$$\overline{[a,b)} = \begin{cases} [0,b) \setminus [0,a), & \text{if } a \neq 0, \\ [0,b), & \text{if } a = 0 \text{ and } b \neq 0, \end{cases}$$

as well as $\overline{[0,b)} = \emptyset$ if $b = 0$.
By the definition of a δ-bracketing cover, to every $x \in [0,1]^s$ we can assign a set $\overline{[p_L(x), p_{L+1}(x))}$, such that $p_k \in \Gamma_k$, $k = L, L+1$, and $\lambda(\overline{[p_L(x), p_{L+1}(x))}) \le 2^{-L}$. Now, by the definition of a δ-cover, we can assign a point $p_{L-1}(x) \in \Gamma_{L-1} \cup \{0\}$ such that $p_{L-1}(x) \le p_L(x)$ and $\lambda(\overline{[p_{L-1}(x), p_L(x))}) \le 2^{-L+1}$. Next we assign a point $p_{L-2}(x) \in \Gamma_{L-2} \cup \{0\}$ such that $p_{L-2}(x) \le p_{L-1}(x)$ and $\lambda(\overline{[p_{L-2}(x), p_{L-1}(x))}) \le 2^{-L+2}$. Proceeding inductively, also for every $k = 1, \ldots, L-3$ we find a point $p_k(x) \in \Gamma_k \cup \{0\}$ such that $p_k(x) \le p_{k+1}(x)$ and $\lambda(\overline{[p_k(x), p_{k+1}(x))}) \le 2^{-k}$. Finally for every $k = 1, \ldots, L+1$ we have assigned points $p_k(x)$, $1 \le k \le L+1$, belonging to $\Gamma_k \cup \{0\}$ for each k, such that, writing $I_k(x) = \overline{[p_k(x), p_{k+1}(x))}$, $1 \le k \le L$, and setting $I_0(x) = \overline{[0, p_1(x))}$, we have

$$\bigcup_{k=0}^{L-1} I_k(x) \subset [0,x) \subset \bigcup_{k=0}^{L} I_k(x). \tag{19}$$

and

$$\lambda(I_k(x)) \le 2^{-k}, \quad k \in \{0, \ldots, L\}.$$

For every $k \in \{0, \ldots, L\}$, let $\mathscr{A}_k = \{I_k(x) \,|\, x \in [0,1]^s\}$ denote the collection of all possible sets $I_k(x)$, as x runs through the whole unit cube $[0,1]^s$. Then the cardinality of these sets is bounded by $\#\Gamma_{k+1}$. Using Theorem 1.15 from Gnewuch [9] we see that we can choose our 2^{-k}-covers Γ_k such that

$$\begin{aligned} \#\mathscr{A}_k \le \#\Gamma_{k+1} &\le 2^s \frac{s^s}{s!} (2^{k+1}+1)^s < \frac{1}{2}\sqrt{2/\pi} \left(2e(2^{k+1}+1)\right)^s \\ &\le \frac{1}{2} \exp\left(\ln\sqrt{2/\pi} + \alpha s \ln\left(2e(2^{k+1}+1)\right)\right), \qquad k \in \{0, \ldots, L-2\}, \end{aligned} \tag{20}$$

where we used Stirling's formula, as well as $s \geq 1$, and $\alpha \geq 1$. Similarly, we can choose the 2^{-L}-bracketing cover Δ_L in a way that

$$\#\mathscr{A}_k \leq \#\Gamma_L = \#\Delta_L \leq 2^{s-1}\frac{s^s}{s!}(2^L+1)^s < \frac{1}{2}\sqrt{1/(2\pi)}\left(2e(2^L+1)\right)^s$$
$$\leq \frac{1}{2}\exp\left(\ln\sqrt{1/(2\pi)} + \alpha s \ln\left(2e(2^L+1)\right)\right), \qquad k \in \{L-1, L\}.$$

Step 3. Given the decomposition from Step 2 we define for $k = 0, \ldots, L$ and $I \in \mathscr{A}_k$

$$E_k(I) = \left\{ \max_{2^M \leq N < 2^{M+1}} \left| \sum_{n=1}^{N} \mathbb{1}_I\left(X^{(n)}\right) - N\lambda(I) \right| > t_k \right\},$$

where the numbers t_k were defined in (16). Moreover, let $E_k = \bigcup_{I\in\mathscr{A}_k} E_k(I)$ and $E = \bigcup_{k=0}^{L} E_k$. If we can show that independently of $I \in \mathscr{A}_k$ we have

$$2^{k+1}(1+s)^{\alpha} M^{\beta} \cdot \#\mathscr{A}_k \cdot \mathbb{P}(E_k(I)) < 1 \qquad \text{for} \qquad k \in \{0, \ldots, L\}, \tag{21}$$

then this leads to

$$\mathbb{P}(E) \leq \sum_{k=0}^{L} \sum_{I\in\mathscr{A}_k} \mathbb{P}(E_k(I)) \leq \sum_{k=0}^{L} 2^{-(k+1)}(1+s)^{-\alpha} M^{-\beta} < \frac{1}{(1+s)^{\alpha}} \frac{1}{M^{\beta}}. \tag{22}$$

In order to show (21) for fixed $I \in \mathscr{A}_k$ with $k \in \{0, \ldots, L\}$ let us define the random variables $Z_n = \mathbb{1}_I(X^{(n)}) - \lambda(I)$, where $n = 1, \ldots, 2^{M+1}$. Obviously, all the Z_n are independent and bounded by

$$|Z_n| \leq C = \max\{\lambda(I), 1-\lambda(I)\}.$$

Furthermore, we have $\mathbb{E}(Z_n) = 0$ and $\mathbb{V}(Z_n) = \lambda(I)(1-\lambda(I))$. From Lemma 2 applied to $\pm Z_n$ and $t = t_k$ we conclude

$$\mathbb{P}(E_k(I)) \leq 2\exp\left(-\frac{t_k^2/2^M}{4\,\lambda(I)\,(1-\lambda(I)) + 2\,C\,t_k/(3\cdot 2^M)}\right). \tag{23}$$

Since $t_k = \sqrt{y_k}\sqrt{s\cdot 2^M} \leq \sqrt{y_{-1}}\sqrt{s\cdot 2^M}$ estimate (17) implies

$$2\frac{t_k}{3\cdot 2^M} < 2\frac{4\sqrt{2}}{3}2^{-L} < \begin{cases} 2^{-k+1}, & \text{for } k = 0, \ldots, L-1, \\ 2^{-L+2}, & \text{for } k = L. \end{cases}$$

For $k \in \{0,1\}$ we use $\lambda(I)(1-\lambda(I)) \le 1/4$ and $C \le 1$ to estimate the denominator in (23) by 3 and 2, respectively. If $L > 2$ and $k \in \{2,\dots,L-1\}$ it is easy to see that maximizing $4\lambda(1-\lambda) + 2^{-k+1}(1-\lambda)$ subject to $\lambda \in [0, 2^{-k}]$ gives the bound $3 \cdot 2^{-k+1}(1-2^{-k})$. For $K = L$ a similar argument shows that the denominator in (23) is less than $2^{-L+3}(1-2^{-L})$. Hence, we have

$$\mathbb{P}(E_k(I)) \le 2\exp(-y_k s/\Lambda_k), \quad \text{where} \quad \Lambda_k = \begin{cases} 3, & k = 0, \\ 2^{-k+3}(1-2^{-k}), & k > 0. \end{cases} \tag{24}$$

Using $s \ge 1$ it is easily seen that for any $k \ge 0$

$$2^{k+1}(1+s)^{\alpha} M^{\beta} \le \exp\left(\ln 2^{k+1} + \alpha s \ln 2 + \beta \ln M\right) \tag{25}$$

such that we can conclude (21) if we choose a_k and b_k in the right way. We explain the necessary arguments for the case $k = 0$ explicitly. The case $k > 0$ then works in the same manner. Combining the estimates (20), (24) and (25) we have because of $\alpha s \ge 1$ that

$$\begin{aligned} &2^1(1+s)^{\alpha} M^{\beta} \cdot \#\mathscr{A}_0 \cdot \mathbb{P}(E_0(I)) \\ &< \exp\left(\ln\left(2\sqrt{2/\pi}\right) + \alpha s \ln\left(4e(2^1+1)\right) + \beta \ln(M) - \alpha s a_0/\Lambda_0 - \beta \ln(M) b_0/\Lambda_0\right) \\ &\le \exp\left(\alpha s\left(\ln\left(24e\sqrt{2/\pi}\right) - a_0/\Lambda_0\right) + \beta \ln(M)\,(1 - b_0/\Lambda_0)\right) \le \exp(0) = 1, \end{aligned}$$

if we choose

$$b_0 = \Lambda_0 \quad \text{and} \quad a_0 = \Lambda_0 \ln\left(24e\sqrt{2/\pi}\right). \tag{26}$$

Similarly, we choose

$$b_k = \Lambda_k \quad \text{and} \quad a_k = \Lambda_k \cdot \ln\left(2^{k+3} e \sqrt{2/\pi}(2^{k+1}+1)\right) \tag{27}$$

to obtain (21) also for $k = 1,\dots,L$.

Step 4. To conclude the main statement of Lemma 1 by applying (22) it remains to show that $\Omega_{M,s} \subseteq E$. To this end let $N \in [2^M, 2^{M+1})$ be arbitrary, but fixed. Due to the definitions in Step 3 for every $\omega \in E^C = \bigcap_{k=0}^{L} \bigcap_{I \in \mathscr{A}_k} E_k(I)^C$ and $x_n = X^{(n)}(\omega) \in [0,1]^s$ (for $n = 1,\dots,N$) we have

$$\left| \sum_{n=1}^{N} \mathbb{1}_I(x_n) - N\lambda(I) \right| \le t_k \qquad \text{for all } k \in \{0,\dots,L\} \text{ and every } I \in \mathscr{A}_k.$$

Thus, from (19) we conclude for every $x \in [0,1]^s$

$$\begin{aligned}\sum_{n=1}^{N} \mathbb{1}_{[0,x)}(x_n) &\le \sum_{k=0}^{L} \sum_{n=1}^{N} \mathbb{1}_{\overline{[p_k(x),p_{k+1}(x))}}(x_n) \\ &\le \sum_{k=0}^{L} \left(N\lambda\left(\overline{[p_k(x), p_{k+1}(x))}\right) + t_k \right) \\ &= N\lambda([0,x)) + N\lambda\left(\overline{[x, p_{L+1}(x))}\right) + \sum_{k=0}^{L} t_k.\end{aligned}$$

Since $p_L(x) \le x \le p_{L+1}(x)$ the volume of the set $\overline{[x, p_{L+1}(x))}$ can be estimated from above by 2^{-L} what is no larger than $1/2\sqrt{y_{-1}}\sqrt{s/2^M}$ due to (17). Hence, because of $N < 2 \cdot 2^M$, the second term in the above sum is less than $\sqrt{y_{-1}}\sqrt{s \cdot 2^M}$. Consequently,

$$\begin{aligned}\sum_{n=1}^{N} \mathbb{1}_{[0,x)}(x_n) &< N\lambda([0,x)) + \sqrt{s \cdot 2^M} \sum_{k=-1}^{L} \sqrt{y_k} \\ &\le N\lambda([0,x)) + \sqrt{\alpha A + \beta B \frac{\ln M}{s}} \sqrt{s \cdot 2^M},\end{aligned}$$

where we used (18) from Step 1 for the last estimate. In a similar way we obtain the corresponding lower bound

$$\begin{aligned}\sum_{n=1}^{N} \mathbb{1}_{[0,x)}(x_n) &\ge \sum_{k=0}^{L-1} \sum_{n=1}^{N} \mathbb{1}_{\overline{[p_k(x),p_{k+1}(x))}}(x_n) \\ &\ge N\lambda([0,x)) - N\lambda\left(\overline{[x, p_{L+1}(x))}\right) - \sum_{k=0}^{L-1} t_k \\ &> N\lambda([0,x)) - \sqrt{\alpha A + \beta B \frac{\ln M}{s}} \sqrt{s \cdot 2^M}.\end{aligned}$$

Both the estimates, together with the definition of D_N^{*s}, imply

$$N \cdot D_N^{*s}(x_1, \ldots, x_N) \le \sqrt{\alpha A + \beta B \frac{\ln M}{s}} \sqrt{s \cdot 2^M}$$

since $x \in [0,1]^s$ was arbitrary. Due to the fact that this holds for all $N \in [2^M, 2^{M+1})$ and for every $\omega \in E^C$ we have shown that $E^C \subseteq (\Omega_{M,s})^C$, i.e., $\Omega_{M,s} \subseteq E$.

Step 5. Finally, we need to check that the sequences $(a_k)_{k=-1}^{L}$ and $(b_k)_{k=-1}^{L}$, which were defined in (26), (27), and (24), satisfy the assumptions made at the

beginning of Sect. 2. We already checked that $L \geq 2$, see (17). Moreover, it is obvious that both sequences are non-negative and non-increasing for $k \geq 0$. Hence, we define $a_{-1} = a_0$ and $b_{-1} = b_0$ to guarantee that this holds for all k. It remains to show (15). To this end we calculate for $c \in \{a, b\}$

$$2\left(\sum_{k=-1}^{L} \sqrt{c_k}\right)^2 \leq 2\left(2\sqrt{c_0} + \sum_{k=1}^{\infty} \sqrt{c_k}\right)^2 \leq \begin{cases} 1164.87, & \text{if } c = a, \\ 177.41, & \text{if } c = b. \end{cases}$$

This completes the proof choosing $A = 1165$ and $B = 178$. □

Proof of Theorem 1. Let ζ denote the Riemann Zeta function, and let $\gamma \geq \zeta^{-1}(2)$. Due to the choice of $A > 9/2$ in Lemma 1 we have $\Omega_{1,s} = \emptyset$ for all $s \in \mathbb{N}$. Hence, for $\alpha = \beta = \gamma$ it follows

$$\mathbb{P}\left(\bigcup_{s\geq 1}\bigcup_{M\geq 1} \Omega_{M,s}\right) \leq \sum_{s\geq 1}\sum_{M\geq 2} \mathbb{P}(\Omega_{M,s}) < \sum_{s\geq 1}\sum_{M\geq 2} \frac{1}{(1+s)^\gamma}\frac{1}{M^\gamma} = (\zeta(\gamma)-1)^2 \leq 1.$$

In particular, this implies $\mathbb{P}\left(\left(\bigcup_{s\geq 1}\bigcup_{M\geq 1} \Omega_{M,s}\right)^C\right) > 0$. Since $\zeta^{-1}(2) < 1.73$ and $B\frac{\ln \log_2 N}{s} \leq B\frac{\ln \ln N}{s} - B \ln \ln 2 \leq 66 + B\frac{\ln \ln N}{s}$, we can choose $\gamma = 1.73$ and obtain

$$\begin{aligned} D_N^{*s}(\mathscr{P}_{N,s}) &\leq \sqrt{1.73}\sqrt{1165 + 66 + 178\frac{\ln \ln N}{s}} \cdot \sqrt{\frac{s}{N}} \\ &\leq \sqrt{2130 + 308\frac{\ln \ln N}{s}} \cdot \sqrt{\frac{s}{N}} \end{aligned}$$

with positive probability. This proves Theorem 1. □

Proof of Theorem 2. For any fixed N and s, the point set $\mathscr{U}_{N,s}$ is an array of $N \times s$ i.i.d. uniformly distributed random variables, just like $\mathscr{P}_{N,s}$ in the assumptions of Lemma 1. For given $\alpha \geq 1$, $\beta \geq 0$, $M, s \in \mathbb{N}$, as well as $A = 1165$ and $B = 178$, set

$$\Omega_{M,s} = \left\{ \max_{2^M \leq N < 2^{M+1}} N \cdot D_N^{*s}(\mathscr{U}_{N,s}) > \sqrt{\alpha A + \beta B \frac{\ln M}{s}}\sqrt{s \cdot 2^M} \right\}$$

where $\mathscr{U}_{N,s}$ now is defined in (12). Then Lemma 1 yields

$$\mathbb{P}(\Omega_{M,s}) < \frac{1}{(1+s)^\alpha}\frac{1}{M^\beta}.$$

With this estimate for the probabilities of the exceptional sets, the rest of the proof of Theorem 2 can be carried out in exactly the same way as the proof of Theorem 1. □

Note that the choice of the constants $\alpha = \beta = \gamma$ in the above proofs is not essential. Alternatively, it would be sufficient to take any pair of parameters $1 < \alpha, \beta < \infty$ such that $(\zeta(\alpha) - 1)(\zeta(\beta) - 1) \leq 1$. Using this trade-off it is possible to fine-tune the absolute constants in our theorems in order to minimize the discrepancy bounds for given N and s. Moreover, better estimates on the size of the used δ-(bracketing) covers may lead to (minor important) improvements of these constants. For details we refer to [1] and the conjectures in Gnewuch [10].

Acknowledgements The first author is supported by the Austrian Research Foundation (FWF), Project S9603-N23.

The questions considered in the present paper arose during discussions with several colleagues at the MCQMC 2012 conference. We particularly want to thank Josef Dick for pointing out to us the connection between the discrepancy of random matrices, complete uniform distribution and Markov Chain Monte Carlo. Finally, we wish to express our gratitude to the two anonymous referees who helped to improve the presentation of the paper.

References

1. Aistleitner, C.: Covering numbers, dyadic chaining and discrepancy. J. Complexity **27**, 531–540 (2011)
2. Aistleitner, C.: On the inverse of the discrepancy for infinite dimensional infinite sequences. J. Complexity **29**, 182–194 (2013)
3. Chen, S., Dick, J., Owen, A.B.: Consistency of Markov chain quasi-Monte Carlo on continuous state spaces. Ann. Statist. **39**, 673–701 (2011)
4. Dick, J.: A note on the existence of sequences with small star discrepancy. J. Complexity **23**, 649–652 (2007)
5. Doerr, B.: A lower bound for the discrepancy of a random point set. J. Complexity (2013). doi:10.1016/j.jco.2013.06.001
6. Doerr, B., Gnewuch, M., Kritzer, P., Pillichshammer, F.: Component-by-component construction of low-discrepancy point sets of small size. Monte Carlo Methods Appl. **14**, 129–149 (2008)
7. Drmota, M., Tichy, R.F.: Sequences, Discrepancies and Applications. Lecture Notes in Mathematics, vol. 1651. Springer, Berlin (1997)
8. Einmahl, U., Mason, D.M.: Some universal results on the behavior of increments of partial sums. Ann. Probab. **24**, 1388–1407 (1996)
9. Gnewuch, M.: Bracketing numbers for axis-parallel boxes and applications to geometric discrepancy. J. Complexity **24**, 154–172 (2008)
10. Gnewuch, M.: Construction of minimal bracketing covers for rectangles. Electron. J. Combin. **15** (2008), Research Paper 95, 20pp
11. Gnewuch, M.: Entropy, randomization, derandomization, and discrepancy. In: Plaskota, L., Woźniakowski, H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2010, pp. 43–78. Springer, Berlin/Heidelberg (2012)
12. Haussler, D.: Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension. J. Combin. Theory Ser. A **69**, 217–232 (1995)
13. Hinrichs, A.: Covering numbers, Vapnik-Červonenkis classes and bounds for the star-discrepancy. J. Complexity **20**, 477–483 (2004)
14. Heinrich, S., Novak, E., Wasilkowski, G.W., Woźniakowski, H.: The inverse of the star-discrepancy depends linearly on the dimension. Acta Arith. **96**, 279–302 (2001)

15. Kuipers, L., Niederreiter, H.: Uniform Distribution of Sequences. Pure and Applied Mathematics. Wiley-Interscience [Wiley], New York (1974)
16. Liu, J.S.: Monte Carlo Strategies in Scientific Computing. Springer Series in Statistics. Springer, New York (2008)
17. Matoušek, J.: Geometric Discrepancy. An Illustrated Guide. Algorithms and Combinatorics, vol. 18. Springer, Berlin (2010). Revised paperback reprint of the 1999 original
18. Novak, E., Woźniakowski, H.: Tractability of Multivariate Problems. Volume I: Linear Information. EMS Tracts in Mathematics, vol. 6. European Mathematical Society (EMS), Zürich (2008)
19. Novak, E., Woźniakowski, H.: Tractability of Multivariate Problems. Volume II: Standard Information for Functionals. EMS Tracts in Mathematics, vol. 12. European Mathematical Society (EMS), Zürich (2010)
20. Philipp, W.: Mixing Sequences of Random Variables and Probablistic Number Theory. Memoirs of the American Mathematical Society, vol. 114. American Mathematical Society, Providence (1971)
21. Robert, C.P., Casella, G.: Monte Carlo Statistical Methods. Springer Texts in Statistics, 2nd edn. Springer, New York (2004)
22. Talagrand, M.: Sharper bounds for Gaussian and empirical processes. Ann. Probab. **22**, 28–76 (1994)
23. Weyl, H.: Über die Gleichverteilung von Zahlen mod. Eins. Math. Ann. **77**, 313–352 (1916)

The L^2 Discrepancy of Irrational Lattices

Dmitriy Bilyk

Abstract It is well known that, when α has bounded partial quotients, the lattices $\{(k/N, \{k\alpha\})\}_{k=0}^{N-1}$ have optimal extreme discrepancy. The situation with the L^2 discrepancy, however, is more delicate. In 1956 Davenport established that a *symmetrized* version of this lattice has L^2 discrepancy of the order $\sqrt{\log N}$, which is the lowest possible due to the celebrated result of Roth. However, it remained unclear whether this holds for the original lattices without any modifications. It turns out that the L^2 discrepancy of the lattice depends on much finer Diophantine properties of α, namely, the alternating sums of the partial quotients. In this paper we extend the prior work to arbitrary values of α and N. We heavily rely on Beck's study of the behavior of the sums $\sum (\{k\alpha\} - \frac{1}{2})$.

1 Introduction

The present note is a sequel to the papers of the author with Temlyakov and Yu [7, 8] – we continue the study of the L^2 discrepancy of two-dimensional lattices of the form $\mathscr{L}_N(\alpha) := \left\{ (k/N, \{k\alpha\}) \right\}_{k=0}^{N-1}$. Historically these lattices play a very important role in discrepancy theory. It has been known for a long time (cf., Lerch [12, 1904]) that, when α has bounded partial quotients of the continued fraction (α is badly approximable), the extreme discrepancy of these lattices satisfies the inequality

$$\left\| D_{\mathscr{L}_N(\alpha)} \right\|_\infty \leq C_1(\alpha) \log N, \tag{1}$$

D. Bilyk (✉)
School of Mathematics, University of Minnesota, Minneapolis, MN 55455, USA
e-mail: dbilyk@math.umn.edu

J. Dick et al. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2012*,
Springer Proceedings in Mathematics & Statistics 65,
DOI 10.1007/978-3-642-41095-6_11, © Springer-Verlag Berlin Heidelberg 2013

which is best possible in view of the famous result of Schmidt [16, 1972]. For $(x, y) \in [0, 1)^2$, the discrepancy function is defined as

$$D_{\mathscr{L}_N(\alpha)}(x, y) = \#\big(\mathscr{L}_N(\alpha) \cap [0, x) \times [0, y)\big) - Nxy. \tag{2}$$

Regarding the L^2 discrepancy, Davenport [10, 1956] has shown that the symmetrized lattice $\mathscr{L}_N^{sym}(\alpha) := \mathscr{L}_N(\alpha) \cup \mathscr{L}_N(-\alpha)$ consisting of $2N$ points satisfies the inequality

$$\big\| D_{\mathscr{L}_N^{sym}(\alpha)} \big\|_2 \le C_2(\alpha) \sqrt{\log(2N)}, \tag{3}$$

complementing the celebrated lower bound obtained by Roth [14, 1954] slightly earlier. Similar inequalities also hold for the rational approximations of irrational lattices (see [7, 8, 13]). Later Roth [15, 1979] established that random shifts of lattices also achieve the optimal order of the L^2 discrepancy.

Nevertheless, it still remained a mystery whether these modifications are indeed necessary and whether the original lattices have asymptotically minimal L^2 discrepancy. At least a couple of standard references in discrepancy theory erroneously stated without proof that $\big\| D_{\mathscr{L}_N(\alpha)} \big\|_2 \ge C''_\alpha \log N$.

The belief in this bound was partially justified by the fact that it holds for another classical low-discrepancy distribution – the Van der Corput set, while its modifications (symmetrizations, translations, digit shifts) have L^2 discrepancy of the order $\sqrt{\log N}$, i.e. in this case the modifications are really necessary.

However, in 1982 Sós and Zaremba [17] proved that if all the partial quotients of the (finite or infinite) continued fraction are equal, then $\big\| D_{\mathscr{L}_N(\alpha)} \big\|_2 \le C'_\alpha \sqrt{\log N}$. This result, in particular, applied to $\alpha = 1 + \sqrt{2}$, the golden section $\alpha = \frac{1+\sqrt{5}}{2}$, the ratio of consecutive Fibonacci numbers $\alpha = \frac{F_n}{F_{n+1}}$. Unfortunately, the paper went largely unnoticed in the subject and the generalizations of this result only appeared recently. It turns out that the L^2 discrepancy estimates for lattices depend on much finer Diophantine properties than just boundedness of partial quotients.

We introduce some notation. For $\alpha \in \mathbb{R}$ consider its continued fraction expansion

$$\alpha = [a_0; a_1, a_2, \ldots] = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \ldots}}} \tag{4}$$

with the partial quotients $a_0 \in \mathbb{Z}$, $a_k \in \mathbb{N}$, $k \ge 1$. This expansion is finite if α is rational, and infinite otherwise. We denote by p_n/q_n the nth order convergents of α, i.e. $p_n/q_n = [a_0; a_1, \ldots, a_n]$. We say that $A \approx B$ if $A = \mathscr{O}(B)$ and vice versa.

In this note we prove the following theorem:

Theorem 1. *Assume that $\alpha = [a_0; a_1, a_2, \ldots]$ has bounded partial quotients and let p_n/q_n be its nth order convergent. Then, for $q_{n-1} < N \le q_n$ we have*

$$\big\| D_{\mathscr{L}_N(\alpha)} \big\|_2 \approx \max\left\{ \left| \sum_{k=1}^{n} (-1)^k a_k \right|, \sqrt{\log N} \right\}, \tag{5}$$

in particular,

$$\left\| D_{\mathscr{L}_N(\alpha)} \right\|_2 \approx \sqrt{\log N} \quad \text{if and only if} \quad \left| \sum_{k=0}^{n} (-1)^k a_k \right| \le C(\alpha)\sqrt{n}. \tag{6}$$

(If $\alpha = p_{n*}/q_{n*}$ is rational, we additionally assume that $N \le q_{n*}$.)

The classical recurrence relation $q_{n+1} = a_{n+1}q_n + q_{n-1}$ easily implies that q_n grows exponentially and thus whenever $q_{n-1} < N \le q_n$, we have $n \approx \log N$. Therefore, the first expression in the estimate above is at most of the order $\log N$.

We note that this theorem obviously includes the aforementioned result of Sós and Zaremba. In addition, a partial case of this theorem has been obtained by the author with Temlyakov and Yu [8] – this case deals with the situation when the rational $\alpha = p_n/q_n$ is the nth convergent of a badly approximable number θ and the number of points $N = q_n$. This case, in particular, takes care of the famous Fibonacci lattice $\mathscr{F}_n = \left\{(k/F_n, \{kF_{n-1}/F_n\})\right\}_{k=0}^{F_n-1}$. Aicke Hinrichs (private communication) conjectures that the Fibonacci lattice has the lowest L^2 discrepancy among all lattices with F_n points. For more information on the Fibonacci lattice and its relation to discrepancy and numerical integration see [7, 8, 18–20].

We briefly mention some other values of α which yield a lattice with an optimal order of L^2 discrepancy. First of all, for any integer of the form $m = p^2 + 1$, we have $\sqrt{m} = [p; \overline{2p}]$. Hence it follows already from the Sós–Zaremba result that $\mathscr{L}_N(\sqrt{m})$ has L^2 discrepancy of order $\sqrt{N}$. Therefore, $\mathscr{L}_N(\sqrt{2})$ is optimal, while $\mathscr{L}_N(\sqrt{3})$ is not, since $\sqrt{3} = [1; \overline{1, 2}]$ and the alternating sums grow linearly. We can also construct other examples. It is well known that quadratic irrationalities have periodic continued fraction expansions. Notice that if the length of the period is odd, then the alternating sums $\sum_{k=1}^{n}(-1)^k a_k$ stay bounded and the L^2 discrepancy is bounded by $\sqrt{\log N}$. We list the first few values of m (excluding $m = p^2 + 1$) such that the expansions of $\sqrt{m}$ have periods of odd length: 13, 29, 41, 53, 58, 61, 73, 74, 85, 89, 97. Notice that the periodicity implies an interesting dichotomy: for any quadratic irrational β, the L^2 discrepancy of $\mathscr{L}_N(\beta)$ is either of the order $\log N$ or $\sqrt{\log N}$. In general, it is not had to construct α so that $\mathscr{L}_N(\alpha)$ has any intermediate rate of the L^2 discrepancy.

We add a few words about the methods. Both the original paper of Davenport [10], and the work of Bilyk, Temlyakov, and Yu [7, 8] used the Fourier series analysis of the discrepancy function. However, Davenport looked at discrepancy as a function of y and obtained estimates independent of x, while the author and collaborators considered the two-dimensional Fourier series, which for a rational lattice are supported on a very sparse set. In both cases, the main problem comes from the zero-order term of the Fourier expansion (the integral); indeed, both Davenport's symmetrization and Roth's translation are intended to handle this term. In this paper, we revert to Davenport's method.

2 Preliminaries

Consider the 1-periodic *sawtooth* function $\psi(x) = \{x\} - \frac{1}{2}$. It will be crucial for us to understand the behavior of the sums

$$S_m(\alpha) = \sum_{k=0}^{m} \psi(k\alpha). \tag{7}$$

These objects have been extensively studied by Beck [2–4] (I would like to thank Nir Lev for pointing out these references to me). In particular, it turns out that the Cesaro mean of these sums

$$T_N(\alpha) := \frac{1}{N}\sum_{m=0}^{N-1} S_m(\alpha) = \sum_{m=0}^{N-1}\left(1 - \frac{m}{N}\right)\psi(m\alpha) \tag{8}$$

satisfies the following (see Theorem 3.2 in [3])

$$T_N(\alpha) = \frac{1}{12}\sum_{k=1}^{n}(-1)^k a_k \ + \ \mathcal{O}(\max_{1\le i\le n} a_i), \tag{9}$$

where n is the smallest index such that $q_n \ge N$. It can also be shown (see [3]) that the second moment of these sums satisfy

$$V_N(\alpha) := \frac{1}{N}\sum_{m=0}^{N-1}\big(S_m(\alpha) - T_N(\alpha)\big)^2 \approx \sum_{m:\, q_m \le N} a_m^2. \tag{10}$$

In addition, the Central Limit Theorem holds for the sums $S_n(\alpha)$. The CLT takes the following form (see Theorem 4.1 in [3])

$$\frac{1}{N}\cdot \#\left\{0 \le m \le N-1 : \ \frac{S_m(\alpha) - T_N(\alpha)}{\sqrt{V_N(\alpha)}} \le \lambda\right\} \longrightarrow \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\lambda} e^{-t^2/2}dt \quad \text{as } N\to\infty \tag{11}$$

provided that $a_k^2/(\sum_{i=1}^{k} a_i^2) \to 0$ as $k \to \infty$.

This statement is applicable, in particular, when a_k's are bounded. In this case it follows from (10) that

$$V_N(\alpha) \le \max a_k^2 \cdot \#\{m : \ q_m \le N\} \le C_\alpha \log N \tag{12}$$

for some absolute constant $C_\alpha > 0$, since, as noted earlier, $q_{n-1} < N \le q_n$ implies $n \approx \log N$.

Now the CLT easily implies that

$$\left\| \frac{S_m(\alpha) - T_N(\alpha)}{\sqrt{V_N(\alpha)}} \right\|_{\ell^2(N)} = \mathscr{O}(1) \tag{13}$$

as $N \to \infty$, where $\|x\|_{\ell^2(N)} = \left(\frac{1}{N} \sum_{m=0}^{N-1} |x(m)|^2 \right)^{1/2}$. Indeed, if x satisfies the CLT (11), then

$$\begin{aligned} \frac{1}{N} \sum_{m=0}^{N-1} |x(m)|^2 &\le \sum_{k \in \mathbb{Z}} \frac{\#\{m : 2^{k-1} < |x(m)| \le 2^k\}}{N} \cdot 2^{2k} \\ &\approx \sum_{k \in \mathbb{Z}} \frac{2^{2k}}{\sqrt{2\pi}} \int_{2^{k-1}}^{2^k} e^{-t^2/2} dt \le \frac{4}{\sqrt{2\pi}} \int_0^\infty t^2 \cdot e^{-t^2/2} dt \end{aligned}$$

when N is large. Therefore,

$$T_N(\alpha) \le \left(\frac{1}{N} \sum_{m=0}^{N-1} S_m^2(\alpha) \right)^{1/2} \le K_\alpha \big(T_N(\alpha) + \sqrt{\log N} \big) \tag{14}$$

for some constant $K_\alpha > 0$. The first inequality is obvious by Cauchy–Schwartz, while the second one is a corollary of (13) and (12). This estimate will be crucial in the proof of Theorem 1.

In the end we would like to note that the mean values of $S_m(\alpha)$ arise naturally with respect to discrepancy. It is easy to check that

$$\int_{[0,1)^2} D_{\mathscr{L}_N(\alpha)}(x, y)\, dx\, dy = \sum_{m=0}^{N-1} \left(1 - \frac{m}{N}\right)\big(1 - \{m\alpha\}\big) - \frac{N}{4} = -T_N(\alpha) + \frac{1}{4}. \tag{15}$$

This, together with Roth's theorem, immediately implies the lower bound in (5) since $\|f\|_2 \ge |\int f|$. Estimate (14) for the quadratic mean of $S_m(\alpha)$ will arise in the proof of the upper bound.

In the case considered in [8] when $\alpha = p/q$ is rational and $N = q$, the integral above equals $\mathscr{D}(p, q) + \frac{1}{2}$, where

$$\mathscr{D}(p, q) = \sum_{k=0}^{q-1} \frac{k}{q} \cdot \psi\left(k \frac{p}{q} \right) \tag{16}$$

is the *Dedekind sum*. The fact that its behavior is controlled by the alternating sums of partial quotients of p/q has been known independently of Beck's work (e.g. [1, 11]) and has been used in the present setting in [8].

3 The Proof of Theorem 1 (Upper Bound)

We follow Davenport's approach. For a moment, let us fix $x \in [0,1)$ and set $U = U(x) = \lceil Nx - 1 \rceil$. It is well known (see [10,15]) that the discrepancy function may be approximated as $D_{\mathscr{L}_N(\alpha)}(x,y) = M_U(y) + \mathscr{O}(1)$, where

$$M_U(y) = \sum_{k=0}^{U} \big(\psi(k\alpha - y) - \psi(k\alpha)\big) = \frac{1}{2\pi i} \sum_{m\neq 0} \frac{1}{m} \Big(\sum_{k=0}^{U} e^{2\pi i m k \alpha} \Big) \big(1 - e^{-2\pi i m y}\big), \tag{17}$$

where the equality is understood in the L^2 sense. We have used the Fourier expansion $\psi(x) \sim -\sum_{m\neq 0} \frac{e^{2\pi i m y}}{2\pi i m}$. Using Parseval's identity one obtains:

$$\big\| M_U \big\|^2_{L^2(dy)} \leq |\widehat{M_U}(0)|^2 + C \sum_{m=1}^{\infty} \frac{1}{m^2} \Big| \sum_{k=0}^{U} e^{2\pi i m k \alpha} \Big|^2. \tag{18}$$

The sum above is bounded by a constant multiple of $\log U \leq \log N$ (see [9,10] for details – this estimate was the heart of Davenport's proof). The zero-order Fourier coefficient (the constant term) is

$$\widehat{M_U}(0) = \frac{1}{2\pi i} \sum_{m\neq 0} \frac{1}{m} \Big(\sum_{k=0}^{U} e^{2\pi i m k \alpha} \Big) = -\sum_{k=0}^{U} \psi(k\alpha) = -S_U(\alpha). \tag{19}$$

We thus arrive to

$$\big\| M_U \big\|^2_{L^2(dy)} \leq S_U^2(\alpha) + C'_\alpha \log N. \tag{20}$$

We now integrate estimate (20) over $x \in [0,1)$. Notice that as x runs over $[0,1)$, the discrete parameter $U = U(x)$ changes between 0 and $N-1$, hence the first term results in

$$\frac{1}{N} \sum_{U=0}^{N-1} S_U^2(\alpha) \leq C''_\alpha (T_N^2(\alpha) + \log N) \tag{21}$$

according to (14). Putting together these estimates and (9) we find that

$$\big\| M_{U(x)}(y) \big\|^2_{L^2(dx\,dy)} \leq K_1(\alpha) \log N + K_2(\alpha) \Big| \sum_{k:\, q_k \leq N} (-1)^k a_k \Big|^2, \tag{22}$$

for some constants $K_1(\alpha)$ and $K_2(\alpha)$, which yields the upper bound in (5) and finishes the proof of Theorem 1. □

We would like to make a concluding remark. It seems to be a recurrent feature that whenever a well-distributed set fails to meet the optimal L^2 discrepancy bounds, the problem is always already in the constant term, i.e. the integral of the discrepancy function [5, 6, 8, 10, 15]. We conjecture that this should be true in general, in other words the following statement should hold: *there exist constants* $C_1, C_2, C_3 > 0$ *such that whenever* $\mathscr{P}_N \subset [0,1)^2$, $\#\mathscr{P}_N = N$ *satisfies* $\|D_{\mathscr{P}_N}\|_\infty \leq C_1 \log N$ *and* $\|D_{\mathscr{P}_N}\|_2 \geq C_2 \log N$, *it should also satisfy*

$$\left| \int_{[0,1)^2} D_{\mathscr{P}_N}(x,y)\, dx\, dy \right| \geq C_3 \log N. \tag{23}$$

References

1. Barkan, P.: Sur les sommes de Dedekind et les fractions continues finies. Comptes Rendus de l'Académie des Sciences, Paris, Sér. A **284**, 923–926 (1977)
2. Beck, J.: From probabilistic diophantine approximation to quadratic fields. In: Random and Quasi-Random Point Sets. Lecture Notes in Statistics, vol. 138, pp. 1–48. Springer, New York (1998)
3. Beck, J.: Randomness in lattice point problems. Discrete Math. **229**, 29–55 (2001)
4. Beck, J.: Lattice point problems: crossroads of number theory, probability theory and Fourier analysis. In: Fourier Analysis and Convexity. Applied and Numerical Harmonic Analysis, pp. 1–35. Birkhäuser, Boston (2004)
5. Bilyk, D.: Cyclic shifts of the van der Corput Set. Proc. Amer. Math. Soc. **137**, 2591–2600 (2009)
6. Bilyk, D., Lacey, M., Parissis, I., Vagharshakyan, A.: Exponential squared integrability of the discrepancy function in two dimensions. Mathematika **55**, 1–27 (2009)
7. Bilyk, D., Temlyakov, V.N., Yu, R.: Fibonacci sets and symmetrization in discrepancy theory. J. Complexity **28**, 18–36 (2012)
8. Bilyk, D., Temlyakov, V.N., Yu, R.: The L_2 discrepancy of two-dimensional lattices. In: Bilyk, D., De Carli, L., Petukhov, A., Stokolos, A.M., Wick, B.D. (eds.) Recent Advances in Harmonic Analysis and Applications. Proceedings in Mathematic and Statistics, 25, pp. 63–77. Springer, New York/Heidelberg/Dordrecht/London (2013)
9. Chen, W.W.L.: Fourier techniques in the theory of irregularities of point distribution. In: Fourier Analysis and Convexity. Applied and Numerical Harmonic Analysis, pp. 59–82. Birkhäuser, Boston (2004)
10. Davenport, H.: Note on irregularities of distribution. Mathematika **3**, 131–135 (1956)
11. Hall, R.R., Huxley, M.N.: Dedekind sums and continued fractions. Acta Arith. **63**, 79–90 (1993)
12. Lerch, M.: Question 1547. L'Intermédiaire des Mathématiciens. **11**, 144–145 (1904)
13. Niederreiter, H.: Random Number Generation and Quasi-Monte Carlo Methods. SIAM, Philadelphia (1992)
14. Roth, K.F.: On irregularities of distribution. Mathematika **1**, 73–79 (1954)
15. Roth, K.F.: On irregularities of distribution. III. Acta Arith. **35**, 373–384 (1979)
16. Schmidt, W.M.: Irregularities of distribution. VII. Acta Arith. **2**, 45–50 (1972)
17. Sós, V.T., Zaremba, S.K.: The mean-square discrepancies of some two-dimensional lattices. Studia Sci. Math. Hungar. **14**, 255–271 (1982)

18. Temlyakov, V.N.: Error estimates for Fibonacci quadrature formulas for classes of functions with bounded mixed derivative. Tr. Mat. Inst. Steklova **200**, 327–335 (1991). (English translation in Proceedings of the Steklov Institute of Mathematics **2**, (1993))
19. Zaremba, S.C.: Good lattice points, discrepancy, and numerical integration. Annali di Matematica Pura ed Applicata **73**, 293–317 (1966)
20. Zaremba, S.K.: A remarkable lattice generated by fibonacci numbers. Fibonacci Quart. **8**, 185–194 (1970)

Complexity of Banach Space Valued and Parametric Integration

Thomas Daun and Stefan Heinrich

Abstract We study the complexity of Banach space valued integration. The input data are assumed to be r-smooth. We consider both definite and indefinite integration and analyse the deterministic and the randomized setting. We develop algorithms, estimate their error, and prove lower bounds. In the randomized setting the optimal convergence rate turns out to be related to the geometry of the underlying Banach space. Then we study the corresponding problems for parameter dependent scalar integration. For this purpose we use the Banach space results and develop a multilevel scheme which connects Banach space and parametric case.

1 Introduction

While complexity of integration in the scalar case is well-studied, the Banach space case has not been investigated before. We consider both definite and indefinite integration, develop randomized algorithms and analyse their convergence. We also prove lower bounds and this way estimate the complexity of the integration problems. The results are related to the geometry of the underlying Banach space. It turns out that the bounds are matching and the algorithms are of optimal order for special spaces, including the L_p spaces. For general Banach spaces an arbitrarily small gap in the exponent of upper and lower bounds remains. We also study the deterministic case and show that for arbitrary Banach spaces our methods are of optimal order for any fixed choice of the random parameters.

The study of Banach space valued problems turns out to be crucial for the development of algorithms and the complexity analysis for parameter dependent problems, since such problems can be viewed as special cases of this general

T. Daun (✉) · S. Heinrich
University of Kaiserslautern, D-67653 Kaiserslautern, Germany
e-mail: daun@informatik.uni-kl.de; heinrich@informatik.uni-kl.de

J. Dick et al. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2012*,
Springer Proceedings in Mathematics & Statistics 65,
DOI 10.1007/978-3-642-41095-6_12, © Springer-Verlag Berlin Heidelberg 2013

context. To apply our Banach space results we need a way of passing from Banach space valued to scalar information (function values). This is achieved by a multilevel scheme which is based on the ideas of [2, 6]. As a result, we obtain multilevel algorithms for the parametric problems and show that they are of optimal order (in some cases up to a logarithmic factor).

The paper is organized as follows. In Sect. 2 we provide the needed notation and technical tools. Section 3 contains algorithms for definite and indefinite Banach space valued integration, their analysis and lower bounds. In Sect. 4 we present the multilevel approach and in Sect. 5 we apply the previous results to the parametric problems.

2 Preliminaries

Let $\mathbb{N} = \{1, 2, \dots\}$ and $\mathbb{N}_0 = \{0, 1, 2, \dots\}$. We introduce some notation and concepts from Banach space theory needed in the sequel. For a Banach space X the closed unit ball is denoted by B_X, the identity mapping on X by I_X, and the dual space by X^*. Given another Banach space Y, we let $\mathscr{L}(X, Y)$ be the space of bounded linear mappings $T : X \to Y$ endowed with the canonical norm. If $X = Y$, we write $\mathscr{L}(X)$ instead of $\mathscr{L}(X, X)$. Throughout the paper the norm of X is denoted by $\|\cdot\|$. Other norms are usually distinguished by subscripts. We assume all considered Banach spaces to be defined over the same scalar field $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$.

Let $Q = [0, 1]^d$ and let $C^r(Q, X)$ be the space of all r-times continuously differentiable functions $f : Q \to X$ equipped with the norm

$$\|f\|_{C^r(Q,X)} = \max_{0 \le j \le r, t \in Q} \|f^{(j)}(t)\|.$$

For $r = 0$ we write $C^0(Q, X) = C(Q, X)$, which is the space of continuous X-valued functions on Q. If $X = \mathbb{K}$, we write $C^r(Q)$ and $C(Q)$.

Let $1 \le p \le 2$. A Banach space X is said to be of (Rademacher) type p, if there is a constant $c > 0$ such that for all $n \in \mathbb{N}$ and $x_1, \dots, x_n \in X$

$$\mathbb{E} \left\| \sum_{i=1}^{n} \varepsilon_i x_i \right\|^p \le c^p \sum_{k=1}^{n} \|x_i\|^p, \tag{1}$$

where $(\varepsilon_i)_{i=1}^n$ is a sequence of independent Bernoulli random variables with $\mathbb{P}\{\varepsilon_i = -1\} = \mathbb{P}\{\varepsilon_i = +1\} = 1/2$ (we refer to [7, 9] for this notion and related facts). The smallest constant satisfying (1) is called the type p constant of X and is denoted by $\tau_p(X)$. If there is no such $c > 0$, we put $\tau_p(X) = \infty$. The space $L_{p_1}(\mathscr{N}, \nu)$ with $(\mathscr{N}, \nu)$ an arbitrary measure space and $p_1 < \infty$ is of type p with $p = \min(p_1, 2)$. Furthermore, there is a constant $c > 0$ such that $\tau_2(\ell_\infty^n) \le c(\log(n + 1))^{1/2}$ for all $n \in \mathbb{N}$. We will use the following result (see [7], Proposition 9.11).

Lemma 1. *Let $1 \le p \le 2$, let X be a Banach space, $n \in \mathbb{N}$ and $(\theta_i)_{i=1}^n$ be a sequence of independent X-valued random variables with $E\|\theta_i\|^p < \infty$ and $\mathbb{E}\,\theta_i = 0$ $(i = 1, \dots, n)$. Then*

$$\left(\mathbb{E}\left\|\sum_{i=1}^n \theta_i\right\|^p\right)^{1/p} \le 2\tau_p(X)\left(\sum_{k=1}^n \mathbb{E}\,\|\theta_i\|^p\right)^{1/p}.$$

We need some notation and facts on tensor products of Banach spaces. For details and proofs we refer to [1] and [8]. Let $X \otimes Y$ be the algebraic tensor product of Banach spaces X and Y. For $z = \sum_{i=1}^n x_i \otimes y_i \in X \otimes Y$ define

$$\lambda(z) = \sup_{u \in B_{X^*},\, v \in B_{Y^*}} \left|\sum_{i=1}^n \langle x_i, u\rangle \langle y_i, v\rangle\right|.$$

The injective tensor product $X \otimes_\lambda Y$ is defined as the completion of $X \otimes Y$ with respect to the norm λ. We use the canonical isometric identification

$$C(Q, X) = X \otimes_\lambda C(Q), \tag{2}$$

valid for arbitrary Banach spaces X, and in particular, for $d > 1$

$$C([0,1]^d) = C([0,1]) \otimes_\lambda C([0,1]^{d-1}) = C([0,1]) \otimes_\lambda \cdots \otimes_\lambda C([0,1]).$$

Given Banach spaces X_1, X_2, Y_1, Y_2 and operators $T_1 \in \mathscr{L}(X_1, Y_1)$, $T_2 \in \mathscr{L}(X_2, Y_2)$, the algebraic tensor product $T_1 \otimes T_2 : X_1 \otimes X_2 \to Y_1 \otimes Y_2$ extends to a bounded linear operator $T_1 \otimes T_2 \in \mathscr{L}(X_1 \otimes_\lambda X_2, Y_1 \otimes_\lambda Y_2)$ with

$$\|T_1 \otimes T_2\|_{\mathscr{L}(X_1 \otimes_\lambda X_2, Y_1 \otimes_\lambda Y_2)} = \|T_1\|_{\mathscr{L}(X_1,Y_1)} \|T_2\|_{\mathscr{L}(X_2,Y_2)}. \tag{3}$$

For $r, m \in \mathbb{N}$ we let $P_m^{r,1} \in \mathscr{L}(C([0,1]))$ be composite with respect to the partition of $[0,1]$ into m intervals of length m^{-1} Lagrange interpolation of degree r. Let

$$P_m^{r,d} = \otimes^d P_m^{r,1} \in \mathscr{L}(C([0,1]^d))$$

be its d-dimensional version. Setting $\Gamma_k^d = \left\{\frac{i}{k} : 0 \le i \le k\right\}^d$ for $k \in \mathbb{N}$, it follows that $P_m^{r,d}$ interpolates on Γ_{rm}^d. Given a Banach space X, the X-valued versions of the operators above are defined in the sense of identification (2) as

$$P_m^{r,d,X} = I_X \otimes P_m^{r,d}. \tag{4}$$

This means that if $P_m^{r,d}$ is represented as

$$P_m^{r,d} f = \sum_{s \in \Gamma_{rm}^d} f(s)\varphi_s \quad (f \in C(Q))$$

for some $\varphi_s \in C(Q)$, then $P_m^{r,d,X}$ has the representation

$$P_m^{r,d,X} f = \sum_{s \in \Gamma_{rm}^d} f(s)\varphi_s \quad (f \in C(Q, X)).$$

We can obviously consider $P_m^{r,d,X}$ also as an operator from $\ell_\infty(\Gamma_{rm}^d, X)$ to $C(Q, X)$. Given $r \in \mathbb{N}_0$ and $d \in \mathbb{N}$, there are constants $c_1, c_2 > 0$ such that for all Banach spaces X and all $m \in \mathbb{N}$

$$\| P_m^{r,d,X} \|_{\mathscr{L}(C(Q,X))} \le c_1, \qquad \sup_{f \in B_{C^r(Q,X)}} \| f - P_m^{r,d,X} f \|_{C(Q,X)} \le c_2 m^{-r}. \tag{5}$$

The scalar case of (5) is well-known, which in turn readily implies the Banach space case by considering functions $f_u \in C(Q)$ given for $f \in C(Q, X)$ and $u \in B_{X^*}$ by $f_u(t) = \langle f(t), u \rangle$ $(t \in Q)$.

We will work in the setting of information-based complexity theory (IBC), see [10, 12]. For the precise notions used here we also refer to [3, 4]. An abstract numerical problem is described by a tuple $\mathscr{P} = (F, G, S, K, \Lambda)$. The set F is the set of input data, G is a normed linear space and $S : F \to G$ an arbitrary mapping, the solution operator, which maps the input $f \in F$ to the exact solution Sf. K is an arbitrary set and Λ is a set of mappings from F to K – the class of admissible information functionals.

A randomized algorithm for $\mathscr{P}$ is a family $A = (A_\omega)_{\omega \in \Omega}$, where $(\Omega, \Sigma, \mathbb{P})$ is the underlying probability space and each A_ω is a mapping $A_\omega : F \to G$. For ω fixed, $A_\omega : F \to G$ is a deterministic algorithm, that is, stands for a deterministic process (depending on ω) which uses values of information functionals on $f \in F$ in an adaptive way. The result of the algorithm, $A_\omega f$, is the approximation to Sf. The parameter ω incorporates all randomness used in the algorithm $A = (A_\omega)_{\omega \in \Omega}$. The error of A is defined as

$$e(S, A, F) = \sup_{f \in F} \mathbb{E} \| Sf - A_\omega f \|_G.$$

Let $\mathrm{card}(A_\omega, f)$ be the number of information functionals used by A_ω at input f. We define the cardinality of A as

$$\mathrm{card}(A, F) = \sup_{f \in F} \mathbb{E}\, \mathrm{card}(A_\omega, f).$$

The central notion of IBC is the n-th minimal error, which is defined for $n \in \mathbb{N}_0$ as

$$e_n^{\mathrm{ran}}(S,F) = \inf_{\mathrm{card}(A,F)\le n} e(S,A,F).$$

So $e_n^{\mathrm{ran}}(S,F)$ is the minimal possible error among all randomized algorithms that use (on the average) at most n information functionals.

We can introduce respective notions for the deterministic setting as a special case of the above by considering only one-point probability spaces $\Omega = \{\omega_0\}$, which means that there is no dependence on randomness. Let $e_n^{\mathrm{det}}(S,F)$ denote the n-th minimal error in this setting.

The complexity of definite scalar integration has been studied in numerous papers, see [10–12] and the references therein. The complexity of scalar indefinite integration was considered only recently in [5]. Let us summarize these known results.

Let $r \in \mathbb{N}_0$, $\iota \in \{0,1\}$, and let S_ι be the operator of definite ($\iota = 0$), respectively indefinite ($\iota = 1$) scalar integration (for the precise definitions see (8)–(9) and the line after (10)). Then there are constants $c_{1-4} > 0$ such that for $n \in \mathbb{N}$ the following hold. The deterministic n-th minimal error satisfies

$$c_1 n^{-r/d} \le e_n^{\mathrm{det}}(S_\iota, B_{C^r(Q)}) \le c_2 n^{-r/d}, \tag{6}$$

while the randomized n-th minimal errors fulfills

$$c_3 n^{-r/d-1/2} \le e_n^{\mathrm{ran}}(S_\iota, B_{C^r(Q)}) \le c_4 n^{-r/d-1/2}. \tag{7}$$

The Banach space cases of both problems have not been studied before. The complexity of parametric definite integration was analysed in [6] (this result is stated as part of Theorem 2 below), parametric indefinite integration has not been investigated before.

Throughout the paper $c, c_1, c_2, \ldots$ are constants, which depend only on the problem parameters r, d, but depend neither on the algorithm parameters n, l etc. nor on the input f. We emphasize that they do not depend on X either. The same symbol may denote different constants, even in a sequence of relations.

3 Banach Space Valued Integration

Let X be a Banach space, $r \in \mathbb{N}_0$, and let the definite integration operator $S_0^X : C(Q,X) \to X$ be given by

$$S_0^X f = \int_Q f(t)dt. \tag{8}$$

Put $F = B_{C^r(Q,X)}$, $G = X$, let $K = X$ and $\Lambda = \Lambda(Q,X) = \{\delta_t : t \in Q\}$ with $\delta_t(f) = f(t)$. So here we consider X-valued information functionals. This describes the definite integration problem

$$\mathscr{P}_0 = (B_{C^r(Q,X)}, X, S_0^X, X, \Lambda(Q,X)).$$

The indefinite integration operator $S_1^X : C(Q,X) \to C(Q,X)$ is given by

$$(S_1^X f)(t) = \int_{[0,t]} f(u)du \quad (t \in Q), \tag{9}$$

with $[0,t] = \prod_{i=1}^d [0,t_i]$ for $t = (t_i)_{i=1}^d \in Q$. Here we take $G = C(Q,X)$, while F, K, and Λ are the same as above, so the indefinite integration problem is

$$\mathscr{P}_1 = (B_{C^r(Q,X)}, C(Q,X), S_1^X, X, \Lambda(Q,X)).$$

Note that in the sense of identification (2) we have

$$S_\iota^X = I_X \otimes S_\iota \quad (\iota = 0,1), \tag{10}$$

where S_ι is the scalar version of S_ι^X, with $X = \mathbb{K}$.

Now we present algorithms for the two integration problems (8) and (9). We start with definite integration. Let $n \in \mathbb{N}$ and let $\xi_i : \Omega \to Q (i = 1,\dots,n)$ be independent, uniformly distributed on Q random variables on some complete probability space $(\Omega, \Sigma, \mathbb{P})$. Set for $f \in C(Q,X)$

$$A_{n,\omega}^{0,0,X} f = \frac{1}{n} \sum_{i=1}^n f(\xi_i(\omega)) \tag{11}$$

and, if $r \geq 1$, put $k = \lceil n^{1/d} \rceil$ and

$$A_{n,\omega}^{0,r,X} f = S_0^X (P_k^{r,d,X} f) + A_{n,\omega}^{0,0,X} (f - P_k^{r,d,X} f). \tag{12}$$

We write $A_{n,\omega}^{0,r}$ for the scalar case $A_{n,\omega}^{0,r,\mathbb{K}}$. Finally we set $A_n^{0,r,X} = \left(A_{n,\omega}^{0,r,X}\right)_{\omega\in\Omega}$. In the scalar case for $r = 0$ this is just the standard Monte Carlo method and for $r \geq 1$ the Monte Carlo method with separation of the main part. Note that for $r \in \mathbb{N}_0$, $n \in \mathbb{N}$, $\omega \in \Omega$

$$A_{n,\omega}^{0,r,X} = I_X \otimes A_{n,\omega}^{0,r}. \tag{13}$$

Let us turn to the error analysis for this algorithm. Fixing the random parameter $\omega \in \Omega$ means that we obtain a deterministic method, the error of which we also consider.

Proposition 1. *Let $r \in \mathbb{N}_0$ and $1 \leq p \leq 2$. Then there are constants $c_{1-3} > 0$ such that for all Banach spaces X, $n \in \mathbb{N}$, $\omega \in \Omega$ we have* $\operatorname{card}\left(A_{n,\omega}^{0,r,X}\right) \leq c_1 n$ *and for all $f \in C^r(Q,X)$*

$$\|S_0^X f - A_{n,\omega}^{0,r,X} f\| \le c_2 n^{-r/d} \|f\|_{C^r(Q,X)} \tag{14}$$

$$\left(\mathbb{E}\,\|S_0^X f - A_{n,\omega}^{0,r,X} f\|^p\right)^{1/p} \le c_3 \tau_p(X) n^{-r/d-1+1/p} \|f\|_{C^r(Q,X)}. \tag{15}$$

Proof. Let $r = 0$ and $f \in C(Q, X)$. With

$$\eta_i(\omega) = \int_Q f(t)dt - f(\xi_i(\omega))$$

we have $\mathbb{E}\,\eta_i(\omega) = 0$,

$$S_0^X f - A_{n,\omega}^{0,0,X} f = \frac{1}{n} \sum_{i=1}^{n} \eta_i(\omega)$$

and

$$\|\eta_i(\omega)\| \le 2\|f\|_{C(Q,X)}.$$

This implies (14) and, together with Lemma 1, also (15). The case $r \ge 1$ follows directly from the case $r = 0$ and relation (5), since

$$S_0^X f - A_{n,\omega}^{0,r,X} f = S_0^X (f - P_k^{r,d,X} f) - A_{n,\omega}^{0,0,X} (f - P_k^{r,d,X} f).$$

This completes the proof. □

Next we consider indefinite integration. First we assume $r = 0$ and present the Banach space version of the algorithm from Sect. 4 of [5]. It is a combination of the Smolyak algorithm with the Monte Carlo method. Fix any $m \in \mathbb{N}$, $m \ge 2$ and $L \in \mathbb{N}_0$. For $\bar{l} = (l_1, \ldots, l_d) \in \mathbb{N}_0^d$ we set $|\bar{l}| = l_1 + \cdots + l_d$ and define $U_{\bar{l}}, V_L \in \mathscr{L}(C(Q))$ by

$$U_{\bar{l}} = (P_{m^{l_1}}^{1,1} - P_{m^{l_1-1}}^{1,1}) \otimes \cdots \otimes (P_{m^{l_d-1}}^{1,1} - P_{m^{l_d-1-1}}^{1,1}) \otimes P_{m^{l_d}}^{1,1}, \tag{16}$$

with the understanding that $P_{m^{-1}}^{1,1} := 0$. Furthermore, put

$$V_L = \sum_{\bar{l} \in \mathbb{N}_0^d,\, |\bar{l}| = L} U_{\bar{l}} \tag{17}$$

and let

$$U_{\bar{l}}^X = I_X \otimes U_{\bar{l}}, \quad V_L^X = I_X \otimes V_L \tag{18}$$

be the respective Banach space versions. Set

$$\bar{1} = (\underbrace{1, \ldots, 1}_{d}), \quad m^{\bar{l}} = (m^{l_1}, \ldots, m^{l_d}), \quad \Gamma_{m^{\bar{l}}} = \Gamma_{m^{l_1}}^1 \times \cdots \times \Gamma_{m^{l_d}}^1,$$

and for $\bar{i} = (i_1, \dots, i_d) \in \mathbb{N}^d$ with $\bar{1} \le \bar{i} \le m^{\bar{l}}$ (component-wise inequalities)

$$Q_{\bar{l},\bar{i}} = \left[\frac{i_1 - 1}{m^{l_1}}, \frac{i_1}{m^{l_1}}\right] \times \cdots \times \left[\frac{i_d - 1}{m^{l_d}}, \frac{i_d}{m^{l_d}}\right].$$

So $(Q_{\bar{l},\bar{i}})_{\bar{1} \le \bar{i} \le m^{\bar{l}}}$ is the partition of Q corresponding to the grid $\Gamma_{m^{\bar{l}}}$. Let $\xi_{\bar{l},\bar{i}} : \Omega \to Q_{\bar{l},\bar{i}}$ $(|\bar{l}| = L, \bar{1} \le \bar{i} \le m^{\bar{l}})$ be independent random variables on a complete probability space $(\Omega, \Sigma, \mathbb{P})$ such that $\xi_{\bar{l},\bar{i}}$ is uniformly distributed on $Q_{\bar{l},\bar{i}}$. Define $g_{\bar{l},\omega} \in \ell_\infty(\Gamma_{m^{\bar{l}}}, X)$ by

$$g_{\bar{l},\omega}(t) = \sum_{\bar{j} : Q_{\bar{l},\bar{j}} \subseteq [0,t]} |Q_{\bar{l},\bar{j}}| f(\xi_{\bar{l},\bar{j}}(\omega)) \quad (t \in \Gamma_{m^{\bar{l}}}), \tag{19}$$

with the convention that $g_{\bar{l},\omega}(t) = 0$ if there is no $\bar{j}$ with $Q_{\bar{l},\bar{j}} \subseteq [0,t]$ (that is, if some component of t is zero). Finally we put

$$L = 2d - 1 \tag{20}$$

and, given $n \in \mathbb{N}$,

$$m = \left\lceil (n+1)^{\frac{1}{L}} \right\rceil. \tag{21}$$

If $r = 0$, we define

$$A^{1,0,X}_{n,\omega} f := \sum_{\bar{l} \in \mathbb{N}_0^d,\, |\bar{l}| = L} U^X_{\bar{l}}\, g_{\bar{l},\omega}. \tag{22}$$

In the case $r \ge 1$ we put $k = \lceil n^{1/d} \rceil$ and

$$A^{1,r,X}_{n,\omega} f = S^X_1 (P^{r,d,X}_k f) + A^{1,0,X}_{n,\omega}(f - P^{r,d,X}_k f). \tag{23}$$

Finally set $A^{1,r,X}_n = \left(A^{1,r,X}_{n,\omega}\right)_{\omega \in \Omega}$. Similarly to (13) we have for $r \in \mathbb{N}_0$, $n \in \mathbb{N}$, $\omega \in \Omega$

$$A^{1,r,X}_{n,\omega} = I_X \otimes A^{1,r}_{n,\omega}, \tag{24}$$

with $A^{1,r}_{n,\omega} = A^{1,r,\mathbb{K}}_{n,\omega}$. The scalar case of the following result for $r = 0$ has been shown in [5]. We use the tensor product technique to carry over parts of the proof.

Proposition 2. *Let $r \in \mathbb{N}_0$, $1 \le p \le 2$. Then there are constants $c_{1-3} > 0$ such that for all Banach spaces X, $n \in \mathbb{N}$, $\omega \in \Omega$ we have* $\operatorname{card}\left(A^{1,r,X}_{n,\omega}\right) \le c_1 n$ *and for all $f \in C^r(Q, X)$*

$$\|S^X_1 f - A^{1,r,X}_{n,\omega} f\|_{C(Q,X)} \le c_2 n^{-r/d} \|f\|_{C^r(Q,X)} \tag{25}$$

$$(\mathbb{E}\, \|S^X_1 f - A^{1,r,X}_{n,\omega} f\|^p_{C(Q,X)})^{1/p} \le c_3 \tau_p(X) n^{-r/d - 1 + 1/p} \|f\|_{C^r(Q,X)}. \tag{26}$$

Proof. We start with the case $r = 0$, where we have

$$\begin{aligned}&\|S_1^X f - A_{n,\omega}^{1,0,X} f\|_{C(Q,X)}\\ &\le \|S_1^X f - V_L^X S_1^X f\|_{C(Q,X)} + \|V_L^X S_1^X f - A_{n,\omega}^{1,0,X} f\|_{C(Q,X)}.\end{aligned} \tag{27}$$

The first term can be estimated using

$$\|S_1^X - V_L^X S_1^X\|_{\mathscr{L}(C(Q,X))} \le c m^{-L+d-1}, \tag{28}$$

the scalar case of which is Lemma 4.2 of [5]. The Banach space case follows by taking tensor products and using (10) and (18). Now we consider the second term. We have

$$\|V_L^X S_1^X f - A_{n,\omega}^{1,0,X} f\|_{C(Q,X)} \le \sum_{\bar{l}\in\mathbb{N}_0^d,\,|\bar{l}|=L} \|U_{\bar{l}}^X S_1^X f - U_{\bar{l}}^X g_{\bar{l},\omega}\|_{C(Q,X)} \tag{29}$$

and

$$\begin{aligned}&\|U_{\bar{l}}^X S_1^X f - U_{\bar{l}}^X g_{\bar{l},\omega}\|_{C(Q,X))}\\ &\le \|U_{\bar{l}}^X\|_{\mathscr{L}\left(\ell_\infty(\Gamma_{m^{\bar{l}}}),C(Q,X)\right)} \left\|(S_1^X f)|_{\Gamma_{m^{\bar{l}}}} - g_{\bar{l},\omega}\right\|_{\ell_\infty(\Gamma_{m^{\bar{l}}})}\\ &\le c \max_{t\in\Gamma_{m^{\bar{l}}}} \left\| \int_{[0,t]} f(t)dt - \sum_{\bar{j}:Q_{\bar{l},\bar{j}}\subseteq[0,t]} |Q_{\bar{l},\bar{j}}| f(\xi_{\bar{l},\bar{j}}) \right\| = c \max_{\bar{1}\le\bar{i}\le m^{\bar{l}}} \left\| \sum_{\bar{1}\le\bar{j}\le\bar{i}} \eta_{\bar{l},\bar{j}} \right\|\end{aligned} \tag{30}$$

with

$$\eta_{\bar{l},\bar{j}} = \int_{Q_{\bar{l},\bar{j}}} f(t)dt - |Q_{\bar{l},\bar{j}}| f(\xi_{\bar{l},\bar{j}}) \quad (\bar{1} \le \bar{j} \le m^{\bar{l}}). \tag{31}$$

The random variables $\{\eta_{\bar{l},\bar{j}} : \bar{1} \le \bar{j} \le m^{\bar{l}}\}$ are independent, of mean zero, and satisfy

$$\|\eta_{\bar{l},\bar{j}}\| \le 2|Q_{\bar{l},\bar{j}}| \|f\|_{C(Q,X)} = 2m^{-L}\|f\|_{C(Q,X)}. \tag{32}$$

Combining (20)–(21) and (27)–(32), we obtain (25) for $r = 0$.

For $p > 1$ we get from Lemma 4.3 of [5] (a simple generalization of Doob's inequality, the proof of which literally carries over to the Banach space case)

$$\left(\mathbb{E} \max_{\bar{1}\le\bar{i}\le m^{\bar{l}}} \left\| \sum_{\bar{1}\le\bar{j}\le\bar{i}} \eta_{\bar{l},\bar{j}} \right\|^p\right)^{1/p} \le c\left(\mathbb{E} \left\| \sum_{\bar{1}\le\bar{j}\le\bar{l}} \eta_{\bar{l},\bar{j}} \right\|^p\right)^{1/p}. \tag{33}$$

Moreover, Lemma 1 gives

$$\Big(\mathbb{E}\Big\|\sum_{\bar{1}\le\bar{j}\le\bar{l}}\eta_{\bar{l},\bar{j}}\Big\|^p\Big)^{1/p}\le 2\tau_p(X)\Big(\sum_{\bar{1}\le\bar{j}\le\bar{l}}\mathbb{E}\,\|\eta_{\bar{l},\bar{j}}\|^p\Big)^{1/p}. \tag{34}$$

From (33) and (34) we conclude for $p>1$

$$\Big(\mathbb{E}\max_{\bar{1}\le\bar{i}\le m^{\bar{l}}}\Big\|\sum_{\bar{1}\le\bar{j}\le\bar{i}}\eta_{\bar{l},\bar{j}}\Big\|^p\Big)^{1/p}\le c\tau_p(X)\Big(\sum_{\bar{1}\le\bar{j}\le\bar{l}}\mathbb{E}\,\|\eta_{\bar{l},\bar{j}}\|^p\Big)^{1/p}. \tag{35}$$

The same relation also holds for $p=1$ by the triangle inequality. We obtain from (29)–(30), (32), and (35)

$$(\mathbb{E}\,\|V_L^X S_1^X f-A_{n,\omega}^{1,0,X}f\|_{C(Q,X)}^p)^{1/p}\le c\tau_p(X)m^{-(1-1/p)L}\|f\|_{C(Q,X)}. \tag{36}$$

Now relation (26) for $r=0$ follows from (20)–(21), (27)–(28), and (36).

As in the proof of Proposition 1 the case $r\ge 1$ follows from the case $r=0$ and (5), since

$$S_1^X f-A_{n,\omega}^{1,r,X}f=S_1^X(f-P_k^{r,d,X}f)-A_{n,\omega}^{1,0,X}(f-P_k^{r,d,X}f).$$

By (16)–(17) and (19)–(23) the number of function values used in $A_{n,\omega}^{1,r,X}f$ is

$$ck^d+c\sum_{|\bar{l}|=L}m^{l_1}\dots m^{l_d}\le cn.$$

□

Theorem 1. *Let $r\in\mathbb{N}_0$, $\iota\in\{0,1\}$, $1\le p\le 2$. Then there are constants $c_{1-4}>0$ such that for all Banach spaces X and $n\in\mathbb{N}$ the following hold. The deterministic n-th minimal error satisfies*

$$c_1 n^{-r/d}\le e_n^{\mathrm{det}}(S_\iota^X,B_{C^r(Q,X)})\le c_2 n^{-r/d}.$$

Moreover, if X is of type p and p_X is the supremum of all p_1 such that X is of type p_1, then the randomized n-th minimal errors fulfills

$$c_3 n^{-r/d-1+1/p_X}\le e_n^{\mathrm{ran}}(S_\iota^X,B_{C^r(Q,X)})\le c_4\tau_p(X)n^{-r/d-1+1/p}.$$

Proof. The upper bounds follow from Propositions 1 and 2. Since definite integration is a particular case of indefinite integration in the sense that $S_0^X f=(S_1^X f)(\bar{1})$, it suffices to prove the lower bound for S_0^X. The lower bounds for the deterministic setting and for the randomized setting with $p_X=2$ follow from the respective scalar cases (6) and (7), since trivially every Banach space X over $\mathbb{K}$ contains an isometric copy of $\mathbb{K}$.

It remains to show the lower bound for the randomized setting for Banach spaces with $p_X < 2$. Any such Banach space must be infinite dimensional (a finite dimensional space X always has $p_X = 2$). Let $n \in \mathbb{N}$ and let $k \in \mathbb{N}$ be such that

$$(k-1)^d < 8n \le k^d. \tag{37}$$

The Maurey-Pisier Theorem (see [9], Theorem 2.3) implies that for every $k \in \mathbb{N}$ there is a subspace $E_k \subset X$ of dimension k^d and an isomorphism $T : \ell_{p_X}^{k^d} \to E_k$ with $\|T\| \le 1$ and $\|T^{-1}\| \le 2$. Let $x_i = Te_i$, where $(e_i)_{i=1}^{k^d}$ is the unit vector basis of $\ell_{p_X}^{k^d}$. Let $\psi \in C^\infty(\mathbb{R}^d)$ be such that $\psi(t) > 0$ for $t \in (0,1)^d$ and $\operatorname{supp}\psi \subset [0,1]^d$. Let $(Q_i)_{i=1}^{k^d}$ be the partition of Q into closed cubes of side length k^{-1} of disjoint interior, let t_i be the point in Q_i with minimal coordinates and define $\psi_i \in C(Q)$ by

$$\psi_i(t) = \psi(k(t - t_i)) \quad (i = 1, \dots, k^d).$$

It is readily checked that there is a constant $c_0 > 0$ such that for all $(\alpha_i)_{i=1}^{k^d} \in [-1,1]^{k^d}$

$$c_0 k^{-r} \sum_{i=1}^{k^d} \alpha_i x_i \psi_i \in B_{C^r(Q,X)}.$$

Put $f_i = c_0 k^{-r} x_i \psi_i$ and $\sigma = \int_Q \psi(t)dt$. Then for $(\alpha_i)_{i=1}^{k^d} \in \mathbb{R}^{k^d}$

$$\begin{aligned}\Big\| \sum_{i=1}^{k^d} \alpha_i S_0^X f_i \Big\| &= c_0 k^{-r} \Big\| \sum_{i=1}^{k^d} \alpha_i x_i \int_Q \psi_i(t)dt \Big\| \\ &= c_0 \sigma k^{-r-d} \Big\| \sum_{i=1}^{k^d} \alpha_i x_i \Big\| \ge c k^{-r-d} \left(\sum_{i=1}^{k^d} |\alpha_i|^{p_X} \right)^{1/p_X}.\end{aligned}$$

Next we use Lemmas 5 and 6 of [3] with $K = X$ (Lemma 6 is formulated for $K = \mathbb{R}$, but directly carries over to $K = X$) and (37) to obtain

$$\begin{aligned}e_n^{\mathrm{ran}}(S_0^X, B_{C^r(Q,X)}) &\ge \frac{1}{4} \min_{I \subseteq \{1,\dots,k^d\}, |I| \ge k^d - 4n} \mathbb{E} \Big\| \sum_{i \in I} \varepsilon_i S_0^X f_i \Big\| \\ &\ge c k^{-r-(1-1/p_X)d} \ge c n^{-r/d-1+1/p_X},\end{aligned}$$

where $(\varepsilon_i)_{i=1}^{k^d}$ is a sequence of independent centered Bernoulli random variables. □

Note that the bounds in the randomized cases of Theorem 1 are matching up to an arbitrarily small gap in the exponent. In some cases, they are even of matching order.

Corollary 1. *Let $r \in \mathbb{N}_0$, $1 \le p \le 2$, $\iota \in \{0,1\}$. Then there are constants $c_1, c_2 > 0$ such that the following hold. Let X be a Banach space which is of type p and moreover, satisfies $p_X = p$ (that is, the supremum of types is attained). Then for all $n \in \mathbb{N}$*

$$c_1 n^{-r/d-1+1/p} \le e_n^{\mathrm{ran}}(S_\iota^X, B_{C^r(Q,X)}) \le c_2 \tau_p(X) n^{-r/d-1+1/p}.$$

This holds, in particular, for spaces of type 2 with $p = 2$ and, if $1 \le p_1 < \infty$, for spaces $X = L_{p_1}(\mathcal{N}, \nu)$, where $(\mathcal{N}, \nu)$ is some measure space, with $p = \min(p_1, 2)$.

For general Banach spaces X upper and lower bounds of matching order of $e_n^{\mathrm{ran}}(S_\iota^X, B_{C^r(Q,X)})$ $(\iota = 0, 1)$ remain an open problem.

4 A Multilevel Procedure

In the previous section we considered Banach space valued information functionals. Now we develop a scheme which will serve as a bridge between the Banach space and the scalar case. It is based on the multilevel Monte Carlo approach from [2, 6]. Assume that a Banach space Y is continuously embedded into the Banach space X, and let J be the embedding map. We shall identify elements of Y with their images in X. For $r, \varrho \in \mathbb{N}_0$ we consider integration of functions from the set

$$B_{C^r(Q,X)} \cap B_{C^\varrho(Q,Y)}.$$

Let $(T_l)_{l=0}^\infty \subset \mathscr{L}(X)$ (this is intended to be a sequence which approximates the embedding J) and set for $l \in \mathbb{N}_0$

$$R_l = T_l \otimes I_{C(Q)} \in \mathscr{L}(C(Q, X)). \tag{38}$$

The operator R_l is just the pointwise application of T_l in the sense that for $f \in C(Q, X)$ and $t \in Q$ we have $(R_l f)(t) = T_l f(t)$. Fix any $l_0, l_1 \in \mathbb{N}_0$, $l_0 \le l_1$, $n_{l_0}, \dots, n_{l_1} \in \mathbb{N}$ and define for $\iota \in \{0,1\}$ and $f \in C(Q, X)$ an approximation $A_\omega^{(\iota)} f$ to $S_\iota^X f$ as follows:

$$A_\omega^{(\iota)} f = A_{n_{l_0},\omega}^{\iota,r,X} R_{l_0} f + \sum_{l=l_0+1}^{l_1} A_{n_l,\omega}^{\iota,\varrho,X} (R_l - R_{l-1}) f \tag{39}$$

and $A^{(\iota)} = (A_\omega^{(\iota)})_{\omega \in \Omega}$. It follows from (13), (24), and (38) that

$$A_\omega^{(\iota)} = T_{l_0} \otimes A_{n_{l_0},\omega}^{\iota,r} + \sum_{l=l_0+1}^{l_1} (T_l - T_{l-1}) \otimes A_{n_l,\omega}^{\iota,\varrho}. \tag{40}$$

Furthermore, put

$$X_l = \mathrm{cl}_X(T_l(X)) \quad (l \in \mathbb{N}_0), \quad X_{l-1,l} = \mathrm{cl}_X((T_l - T_{l-1})(X)) \quad (l \in \mathbb{N}), \tag{41}$$

where cl_X denotes the closure in X. In particular, X_l and $X_{l-1,l}$ are endowed with the norm induced by X. Given a Banach space Z, we introduce the notation $G_0(Z) = Z$ and $G_1(Z) = C(Q,Z)$. Now we estimate the error of $A_\omega^{(\iota)}$ on $B_{C^r(Q,X)} \cap B_{C^\varrho(Q,Y)}$.

Proposition 3. *Let $1 \le p \le 2$, $r, \varrho \in \mathbb{N}_0$, and $\iota \in \{0,1\}$. Then there are constants $c_1, c_2 > 0$ such that for all Banach spaces X, Y, and operators $(T_l)_{l=0}^\infty$ as above, for all $l_0, l_1 \in \mathbb{N}_0$ with $l_0 \le l_1$, and for all $(n_l)_{l=l_0}^{l_1} \subset \mathbb{N}$ the so-defined algorithm $A_\omega^{(\iota)}$ satisfies*

$$\begin{aligned}
&\sup_{f \in B_{C^r(Q,X)} \cap B_{C^\varrho(Q,Y)}} \|S_\iota^X f - A_\omega^{(\iota)} f\|_{G_\iota(X)} \\
&\le \|J - T_{l_1} J\|_{\mathscr{L}(Y,X)} + c_1 \|T_{l_0}\|_{\mathscr{L}(X)} \, n_{l_0}^{-r/d} \\
&\quad + c_1 \sum_{l=l_0+1}^{l_1} \|(T_l - T_{l-1}) J\|_{\mathscr{L}(Y,X)} \, n_l^{-\varrho/d} \quad (\omega \in \Omega)
\end{aligned} \tag{42}$$

and

$$\begin{aligned}
&\sup_{f \in B_{C^r(Q,X)} \cap B_{C^\varrho(Q,Y)}} \left(\mathbb{E} \, \|S_\iota^X f - A_\omega^{(\iota)} f\|_{G_\iota(X)}^p \right)^{1/p} \\
&\le \|J - T_{l_1} J\|_{\mathscr{L}(Y,X)} + c_2 \tau_p(X_{l_0}) \|T_{l_0}\|_{\mathscr{L}(X)} n_{l_0}^{-r/d-1+1/p} \\
&\quad + c_2 \sum_{l=l_0+1}^{l_1} \tau_p(X_{l-1,l}) \|(T_l - T_{l-1}) J\|_{\mathscr{L}(Y,X)} n_l^{-\varrho/d-1+1/p}.
\end{aligned} \tag{43}$$

Proof. Let $f \in B_{C^r(Q,X)} \cap B_{C^\varrho(Q,Y)}$. From (39) we get

$$\begin{aligned}
&\|S_\iota^X f - A_\omega^{(\iota)} f\|_{G_\iota(X)} \\
&\le \|S_\iota^X f - S_\iota^X R_{l_1} f\|_{G_\iota(X)} + \|S_\iota^X R_{l_0} f - A_{n_{l_0},\omega}^{\iota,r,X} R_{l_0} f\|_{G_\iota(X_{l_0})} \\
&\quad + \sum_{l=l_0+1}^{l_1} \|S_\iota^X (R_l - R_{l-1}) f - A_{n_l,\omega}^{\iota,\varrho,X} (R_l - R_{l-1}) f\|_{G_\iota(X_{l-1,l})}.
\end{aligned} \tag{44}$$

We have

$$\begin{aligned}&\|S_\iota^X f - S_\iota^X R_{l_1} f\|_{G_\iota(X)} \le \|S_\iota^X\|_{\mathscr{L}(C(Q,X),G_\iota(X))}\|f - R_{l_1} f\|_{C(Q,X)}\\ &\le \|J - T_{l_1} J\|_{\mathscr{L}(Y,X)}\|f\|_{C(Q,Y)} \le \|J - T_{l_1} J\|_{\mathscr{L}(Y,X)}.\end{aligned} \tag{45}$$

Furthermore, by Propositions 1 and 2

$$\begin{aligned}&\|S_\iota^X R_{l_0} f - A^{\iota,r,X}_{n_{l_0},\omega} R_{l_0} f\|_{G_\iota(X_{l_0})} \le c n_{l_0}^{-r/d}\|R_{l_0} f\|_{C^r(Q,X_{l_0})}\\ &\le c n_{l_0}^{-r/d}\|T_{l_0}\|_{\mathscr{L}(X)}\|f\|_{C^r(Q,X)} \le c\|T_{l_0}\|_{\mathscr{L}(X)} n_{l_0}^{-r/d},\end{aligned} \tag{46}$$

and similarly,

$$\mathbb{E}\left(\|S_\iota^X R_{l_0} f - A^{\iota,r,X}_{n_{l_0},\omega} R_{l_0} f\|^p_{G_\iota(X_{l_0})}\right)^{1/p} \le \quad c\tau_p(X_{l_0})\|T_{l_0}\|_{\mathscr{L}(X)} n_{l_0}^{-r/d-1+1/p}. \tag{47}$$

For $l_0 < l \le l_1$ we obtain

$$\begin{aligned}&\|S_\iota^X (R_l - R_{l-1}) f - A^{\iota,\varrho,X}_{n_l,\omega}(R_l - R_{l-1}) f\|_{G_\iota(X_{l-1,l})}\\ &\le c n_l^{-\varrho/d}\|(R_l - R_{l-1}) f\|_{C^\varrho(Q,X_{l-1,l})}\\ &\le c n_l^{-\varrho/d}\|(T_l - T_{l-1}) J\|_{\mathscr{L}(Y,X)}\|f\|_{C^\varrho(Q,Y)} \le c\|(T_l - T_{l-1}) J\|_{\mathscr{L}(Y,X)} n_l^{-\varrho/d}\end{aligned} \tag{48}$$

and

$$\begin{aligned}&\mathbb{E}\left(\|S_\iota^X (R_l - R_{l-1}) f - A^{\iota,\varrho,X}_{n_l,\omega}(R_l - R_{l-1}) f\|^p_{G_\iota(X_{l-1,l})}\right)^{1/p}\\ &\le c\tau_p(X_{l-1,l})\|(T_l - T_{l-1}) J\|_{\mathscr{L}(Y,X)} n_l^{-\varrho/d-1+1/p}.\end{aligned} \tag{49}$$

Combining (44)–(49) yields the result. □

5 Scalar Parametric Case

In this section we apply the previous results to parametric definite and indefinite integration. Let $d, d_1 \in \mathbb{N}$, $Q_1 = [0,1]^{d_1}$. We consider numerical integration of functions depending on a parameter $s \in Q_1$. The definite parametric integration operator $\mathscr{S}_0 : C(Q_1 \times Q) \to C(Q_1)$ is given by

$$(\mathscr{S}_0 f)(s) = \int_Q f(s,t)dt \quad (s \in Q_1).$$

We put $F = B_{C^r(Q_1\times Q)}$, the set K is the scalar field $\mathbb{K}$, and Λ is the following class of information functionals $\Lambda(Q_1 \times Q, \mathbb{K}) = \{\delta_{s,t} : \ s \in Q_1, t \in Q\}$ where $\delta_{s,t}(f) = f(s,t)$. This is just standard information consisting of values of f. Hence, the definite parametric integration problem is

$$\Pi_0 = (B_{C^r(Q_1\times Q)}, C(Q_1), \mathscr{S}_0, \mathbb{K}, \Lambda(Q_1 \times Q, \mathbb{K})).$$

The indefinite parametric integration operator $\mathscr{S}_1 : C(Q_1 \times Q) \to C(Q_1 \times Q)$ is given by

$$(\mathscr{S}_1 f)(s,t) = \int_{[0,t]} f(s,u)du \quad (s \in Q_1, t \in Q).$$

Here F, K, Λ are chosen to be the same as above, so the indefinite parametric integration problem is described by

$$\Pi_1 = (B_{C^r(Q_1\times Q)}, C(Q_1 \times Q), \mathscr{S}_1, \mathbb{K}, \Lambda(Q_1 \times Q, \mathbb{K})).$$

We can relate these problems to the previously considered Banach space valued ones as follows. Setting $X = C(Q_1)$, we have

$$C(Q_1 \times Q) = C(Q_1) \otimes_\lambda C(Q) = X \otimes_\lambda C(Q) = C(Q, X)$$

and $\mathscr{S}_\iota = S_\iota^{C(Q_1)} (\iota = 0, 1)$. Moreover, referring to the notation of Sect. 4, we put $Y = C^r(Q_1)$ and $\varrho = 0$, which gives

$$B_{C^r(Q_1\times Q)} \subseteq B_{C^r(Q,C(Q_1))} \cap B_{C(Q,C^r(Q_1))} = B_{C^r(Q,X)} \cap B_{C(Q,Y)}.$$

Let $r_1 = \max(r, 1)$ and define for $l \in \mathbb{N}_0$

$$T_l = P_{2^l}^{r_1,d_1} \in \mathscr{L}(C(Q_1)). \tag{50}$$

By (5),

$$\|T_l\|_{\mathscr{L}(C(Q_1))} \le c_1, \quad \|J - T_l J\|_{\mathscr{L}(C^r(Q_1),C(Q_1))} \le c_2 2^{-rl}, \tag{51}$$

where $J : C^r(Q_1) \to C(Q_1)$ is the embedding. The algorithms $A_\omega^{(\iota)}$ defined in (39) and equivalently (40) turn into

$$A_\omega^{(\iota)} = P_{2^{l_0}}^{r_1,d_1} \otimes A_{n_{l_0},\omega}^{\iota,r} + \sum_{l=l_0+1}^{l_1} \left(P_{2^l}^{r_1,d_1} - P_{2^{l-1}}^{r_1,d_1}\right) \otimes A_{n_l,\omega}^{\iota,0}. \tag{52}$$

Let us note that (52) together with the definitions of $P_m^{r_1,d_1}$ and $A_{n,\omega}^{\iota,r}$ imply the following representation of $A_\omega^{(\iota)}$. There are $s_{l,i} \in Q_1$, $t_{l,j,\omega} \in Q$, $\varphi_{l,i} \in C(Q_1)$, $\psi_{l,j,\omega}^{(\iota)} \in \mathbb{K}$ if $\iota = 0$, $\psi_{l,j,\omega}^{(\iota)} \in C(Q)$ if $\iota = 1$, $M_l \le c2^{d_1 l}$, and $N_l \le cn_l$ such that

$$A_\omega^{(\iota)} f = \sum_{l=l_0}^{l_1} \sum_{i=1}^{M_l} \sum_{j=1}^{N_l} f(s_{l,i}, t_{l,j,\omega})\, \varphi_{l,i} \otimes \psi_{l,j,\omega}^{(\iota)} \quad (f \in C(Q_1 \times Q),\ \omega \in \Omega). \tag{53}$$

The particular shape of these functions can be read from the Definitions (11)–(12) and (16)–(23), for more details in the case $\iota = 1$ see also [5]. It follows that

$$\operatorname{card}(A_\omega^{(\iota)}) \le c \sum_{l=l_0}^{l_1} n_l 2^{d_1 l} \quad (\omega \in \Omega). \tag{54}$$

Now we estimate the error of $A_\omega^{(\iota)}$. Recall the notation $G_0(C(Q_1)) = C(Q_1)$ and $G_1(C(Q_1)) = C(Q_1 \times Q)$.

Proposition 4. *Let $r \in \mathbb{N}_0$, $d, d_1 \in \mathbb{N}$, $\iota \in \{0,1\}$. There are constants $c_{1-4} > 0$ such that the following hold. For each $n \in \mathbb{N}$ there are $l_0 \in \mathbb{N}_0$ and $n_{l_0} \in \mathbb{N}$ such that with $l_1 = l_0$ we have $\operatorname{card}(A_\omega^{(\iota)}) \le c_1 n$ and*

$$\sup_{f \in B_{C^r(Q_1 \times Q)}} \| \mathscr{S}_\iota f - A_\omega^{(\iota)} f \|_{G_\iota(C(Q_1))} \le c_2 n^{-\frac{r}{d_1+d}} \tag{55}$$

for all $\omega \in \Omega$. Moreover, for each $n \in \mathbb{N}$ with $n \ge 2$ there is a choice of $l_0, l_1 \in \mathbb{N}_0$, $(n_l)_{l=l_0}^{l_1} \subset \mathbb{N}$ such that $\operatorname{card}(A_\omega^{(\iota)}) \le c_3 n$ $(\omega \in \Omega)$ and

$$\begin{aligned} &\sup_{f \in B_{C^r(Q_1 \times Q)}} \left(\mathbb{E} \| \mathscr{S}_\iota f - A_\omega^{(\iota)} f \|_{G_\iota(C(Q_1))}^2 \right)^{1/2} \\ &\le c_4 \begin{cases} n^{-\frac{2r+d}{2(d_1+d)}} (\log n)^{\frac{1}{2}} & \text{if} \quad r/d_1 > 1/2 \\ n^{-\frac{1}{2}} (\log n)^2 & \text{if} \quad r/d_1 = 1/2 \\ n^{-\frac{r}{d_1}} (\log n)^{\frac{r}{d_1}} & \text{if} \quad r/d_1 < 1/2. \end{cases} \end{aligned} \tag{56}$$

Proof. Let $n \in \mathbb{N}$ and put

$$l^* = \left\lceil \frac{\log_2 n}{d_1} \right\rceil, \quad l_0 = \left\lfloor \frac{d_1}{d_1 + d} l^* \right\rfloor. \tag{57}$$

Furthermore, let $l_1 \in \mathbb{N}_0$, $l_0 \le l_1 \le l^*$, $\delta_0, \delta_1 \ge 0$ to be fixed later on and define

$$n_{l_0} = 2^{d_1(l^* - l_0)}, \quad n_l = \left\lceil 2^{d_1(l^*-l) - \delta_0(l-l_0) - \delta_1(l_1-l)} \right\rceil \quad (l = l_0+1, \dots, l_1). \tag{58}$$

Then by (54) the cost fulfills

$$\begin{aligned}\operatorname{card}(A_\omega^{(\iota)}) &\le c\sum_{l=l_0}^{l_1} n_l 2^{d_1 l} \le c2^{d_1 l^*} + c\sum_{l=l_0+1}^{l_1} 2^{d_1 l^* - \delta_0(l-l_0)-\delta_1(l_1-l)} \\ &\le c\begin{cases} n & \text{if} \quad \delta_0 > 0 \text{ or } \delta_1 > 0 \\ n & \text{if} \quad \delta_0 = \delta_1 = 0 \text{ and } l_1 = l_0 \\ n\log n & \text{if} \quad \delta_0 = \delta_1 = 0 \text{ and } l_1 > l_0. \end{cases} \end{aligned} \tag{59}$$

To show (55), we put $l_1 = l_0$ and get from (42) of Proposition 3, (51), and (57)–(58)

$$\begin{aligned}&\sup_{f\in B_{C^r(Q_1\times Q)}} \|\mathscr{S}_\iota f - A_\omega^{(\iota)} f\|_{G_\iota(C(Q_1))} \\ &\le c2^{-rl_0} + cn_{l_0}^{-\frac{r}{d}} \le c2^{-r\frac{d_1}{d+d_1}l^*} + c2^{-\frac{r}{d}d_1(l^*-l_0)} \le c2^{-\frac{rd_1l^*}{d+d_1}} \le cn^{-\frac{r}{d+d_1}},\end{aligned}$$

which together with (59) gives (55).

Now we turn to the proof of (56) and assume, in addition, that $n \ge 2$. Observe that by (41) and (50)

$$X_l = P_{2^l}^{r_1,d_1}(C(Q_1)) = P_{2^l}^{r_1,d_1}(\ell_\infty(\Gamma_{r_1 2^l}^{d_1})) \tag{60}$$

and $P_{2^l}^{r_1,d_1} : \ell_\infty(\Gamma_{r_1 2^l}^{d_1}) \to X_l$ is an isomorphism which satisfies

$$\left\|P_{2^l}^{r_1,d_1}\right\| \le c_1, \quad \left\|\left(P_{2^l}^{r_1,d_1}\right)^{-1}\right\| = 1.$$

Indeed, the first estimate is just the first part of (5), the second estimate is a consequence of the fact that the inverse of the interpolation operator is just the restriction of functions in X_l to $\Gamma_{r_1 2^l}^{d_1}$. It follows that

$$\tau_2(X_l) \le c\,\tau_2\big(\ell_\infty\big(\Gamma_{r_1 2^l}^{d_1}\big)\big) \le c(l+1)^{1/2}. \tag{61}$$

By (60), $X_{l-1} \subseteq X_l$ for $l \ge 1$, therefore (41) implies that we also have $X_{l-1,l} \subseteq X_l$, thus

$$\tau_2(X_{l-1,l}) \le c(l+1)^{1/2}. \tag{62}$$

For brevity we denote

$$E := \sup_{f\in B_{C^r(Q_1\times Q)}} \left(\mathbb{E}\,\|\mathscr{S}_\iota f - A_\omega^{(\iota)} f\|_{G_\iota(C(Q_1))}^2\right)^{1/2}.$$

By (43) of Proposition 3, (51), and (61)–(62)

$$E \le c2^{-rl_1} + c(l_0+1)^{1/2} n_{l_0}^{-r/d-1/2} + c\sum_{l=l_0+1}^{l_1} (l+1)^{1/2} 2^{-rl} n_l^{-1/2}$$

$$\le c2^{-rl_1} + c(l^*+1)^{1/2} 2^{-(r/d+1/2)d_1(l^*-l_0)} + c(l^*+1)^{1/2} \sum_{l=l_0+1}^{l_1} 2^{-\mu(l)}, \tag{63}$$

where we defined

$$\mu(l) = rl + (d_1(l^*-l) - \delta_0(l-l_0) - \delta_1(l_1-l))/2 \qquad (l_0 \le l \le l_1). \tag{64}$$

We have from (57)

$$\frac{rd_1}{d}(l^*-l_0) \ge \frac{rd_1}{d} \cdot \frac{d}{d_1+d} l^* = r\frac{d_1}{d_1+d} l^* \ge rl_0,$$

consequently,

$$2^{-(r/d+1/2)d_1(l^*-l_0)} \le 2^{-rl_0 - d_1(l^*-l_0)/2} \le 2^{-\mu(l_0)},$$

which together with (63) gives

$$E \le c2^{-rl_1} + c(l^*+1)^{1/2} \sum_{l=l_0}^{l_1} 2^{-\mu(l)}. \tag{65}$$

We rewrite (64) as

$$\mu(l) = rl_0 + d_1(l^*-l_1)/2 + (r-\delta_0/2)(l-l_0) + (d_1-\delta_1)(l_1-l)/2. \tag{66}$$

If $r > d_1/2$, we set $\delta_1 = 0, l_1 = l^*$ and choose $\delta_0 > 0$ in such a way that $r - \delta_0/2 > d_1/2$. From (57), (65), and (66) we obtain

$$E \le c2^{-rl^*} + c(l^*+1)^{1/2} \sum_{l=l_0}^{l^*} 2^{-rl_0 - (r-\delta_0/2)(l-l_0) - d_1(l^*-l)/2}$$

$$\le c2^{-rl^*} + c(l^*+1)^{1/2} 2^{-rl_0 - d_1(l^*-l_0)/2}$$

$$\le c2^{-rl^*} + c(l^*+1)^{1/2} 2^{-\frac{(r+d/2)d_1}{d_1+d} l^*} \tag{67}$$

$$\le c(l^*+1)^{1/2} 2^{-\frac{(r+d/2)d_1}{d_1+d} l^*} \le cn^{-\frac{r+d/2}{d_1+d}} (\log n)^{1/2}, \tag{68}$$

where in the step from (67) to (68) we used $\frac{(r+d/2)d_1}{d_1+d} < r$, which follows from the assumption $r > d_1/2$. This together with (59) proves (56) for $r > d_1/2$.

If $r = d_1/2$, we set $\delta_0 = \delta_1 = 0$, $l_1 = l^*$ and get from (57), (65), and (66)

$$E \le c2^{-rl^*} + c(l^*+1)^{1/2} \sum_{l=l_0}^{l^*} 2^{-rl_0 - r(l-l_0) - d_1(l^*-l)/2}$$
$$\le c(l^*+1)^{3/2} 2^{-d_1 l^*/2} \le cn^{-1/2} (\log n)^{3/2}.$$

Combining this with (59) and transforming $n \log n$ into n gives the respective estimate (56) in this case.

Finally, if $r < d_1/2$, we set $\delta_0 = 0$, choose $\delta_1 > 0$ in such a way that $(d_1 - \delta_1)/2 > r$ and put

$$l_1 = l^* - \left\lceil d_1^{-1} \log_2(l^*+1) \right\rceil. \tag{69}$$

Consequently,

$$\log_2(l^*+1) \le d_1(l^* - l_1) < \log_2(l^*+1) + d_1. \tag{70}$$

Also observe that there is a constant $c_0 \in \mathbb{N}$ such that for $n \ge c_0$

$$l_0 \le l^* - \left\lceil d_1^{-1} \log_2(l^*+1) \right\rceil \le l^*. \tag{71}$$

Since for $n < c_0$ the estimate (56) follows trivially from (65) by a suitable choice of the constant, we can assume $n \ge c_0$, and thus (71). By (57), (65)–(66), (69), and (70)

$$E \le c2^{-rl_1} + c(l^*+1)^{1/2} \sum_{l=l_0}^{l_1} 2^{-rl_0 - d_1(l^*-l_1)/2 - r(l-l_0) - (d_1-\delta_1)(l_1-l)/2}$$
$$\le c2^{-rl_1} + c(l^*+1)^{1/2} 2^{-rl_0 - d_1(l^*-l_1)/2 - r(l_1-l_0)}$$
$$\le c2^{-rl_1} + c(l^*+1)^{1/2} 2^{-rl_1 - (\log_2(l^*+1))/2}$$
$$\le c2^{-rl_1} = c2^{-rl^* + r(l^*-l_1)} \le c2^{-rl^* + (r/d_1)\log_2(l^*+1)}$$
$$= c2^{-rl^*} (l^*+1)^{r/d_1} \le cn^{-r/d_1} (\log n)^{r/d_1}.$$

With this, (56) is now a consequence of (59). □

The following theorem gives the complexity of parametric integration. The case of definite parametric integration is already contained in [6] (with a slightly better upper bound in the limit case $r/d_1 = 1/2$: $(\log n)^{3/2}$ instead of $(\log n)^2$). The case of indefinite parametric integration is new.

Theorem 2. *Let $r \in \mathbb{N}_0$, $d, d_1 \in \mathbb{N}$, $\iota \in \{0, 1\}$. Then there are constants $c_{1-8} > 0$ such that for all $n \in \mathbb{N}$ with $n \ge 2$ the deterministic n-th minimal error satisfies*

$$c_1 n^{-\frac{r}{d_1+d}} \le e_n^{\mathrm{det}}(\mathscr{S}_\iota, B_{C^r(Q_1\times Q)}) \le c_2 n^{-\frac{r}{d_1+d}}.$$

For the randomized n-th minimal error we have the following: If $r/d_1 > 1/2$*, then*

$$c_3 n^{-\frac{2r+d}{2(d_1+d)}} (\log n)^{\frac{1}{2}} \le e_n^{\mathrm{ran}}(\mathscr{S}_\iota, B_{C^r(Q_1\times Q)}) \le c_4 n^{-\frac{2r+d}{2(d_1+d)}} (\log n)^{\frac{1}{2}},$$

if $r/d_1 = 1/2$*, then*

$$c_5 n^{-\frac{1}{2}} (\log n)^{\frac{1}{2}} \le e_n^{\mathrm{ran}}(\mathscr{S}_\iota, B_{C^r(Q_1\times Q)}) \le c_6 n^{-\frac{1}{2}} (\log n)^2$$

and if $r/d_1 < 1/2$*, then*

$$c_7 n^{-\frac{r}{d_1}} (\log n)^{\frac{r}{d_1}} \le e_n^{\mathrm{ran}}(\mathscr{S}_\iota, B_{C^r(Q_1\times Q)}) \le c_8 n^{-\frac{r}{d_1}} (\log n)^{\frac{r}{d_1}}.$$

Proof. The upper bounds follow from Proposition 4. For the lower bounds it suffices to consider parametric definite integration. But these are contained in Theorem 2.4 of [6] (note a misprint there, case $r < d_1/2$: d_2 is to be replaced by d_1). □

Let us finally note that the choice of $Y = C^r(Q_1)$ and $\varrho = 0$ in this section was motivated by our application to the class $C^r(Q_1 \times Q)$, but is, of course, not the only interesting one. We leave other cases to future consideration.

References

1. Defant, A., Floret, K.: Tensor Norms and Operator Ideals. North Holland, Amsterdam (1993)
2. Heinrich, S.: Monte Carlo complexity of global solution of integral equations. J. Complexity **14**, 151–175 (1998)
3. Heinrich, S.: Monte Carlo approximation of weakly singular integral operators. J. Complexity **22**, 192–219 (2006)
4. Heinrich, S.: The randomized information complexity of elliptic PDE. J. Complexity **22**, 220–249 (2006)
5. Heinrich, S., Milla, B.: The randomized complexity of indefinite integration. J. Complexity **27**, 352–382 (2011)
6. Heinrich, S., Sindambiwe, E.: Monte Carlo complexity of parametric integration. J. Complexity **15**, 317–341 (1999)
7. Ledoux, M., Talagrand, M.: Probability in Banach Spaces. Springer, Berlin (1991)
8. Light, W.A., Cheney, W.: Approximation Theory in Tensor Product Spaces. Lecture Notes in Mathematics 1169. Springer, Berlin (1985)
9. Maurey, B., Pisier, G.: Series de variables aléatoires vectorielles independantes et propriétés geométriques des espaces de Banach. Studia Mathematica **58**, 45–90 (1976)
10. Novak, E.: Deterministic and Stochastic Error Bounds in Numerical Analysis. Lecture Notes in Mathematics 1349. Springer, Berlin (1988)
11. Novak, E., Woźniakowski, H.: Tractability of Multivariate Problems, Volume 2, Standard Information for Functionals. European Mathematical Society, Zürich (2010)
12. Traub, J.F., Wasilkowski, G.W., Woźniakowski, H.: Information-Based Complexity. Academic, New York (1988)

Extended Latin Hypercube Sampling for Integration and Simulation

Rami El Haddad, Rana Fakhereddine, Christian Lécot, and Gopalakrishnan Venkiteswaran

Abstract We analyze an extended form of Latin hypercube sampling technique that can be used for numerical quadrature and for Monte Carlo simulation. The technique utilizes random point sets with enhanced uniformity over the s-dimensional unit hypercube. A sample of $N = n^s$ points is generated in the hypercube. If we project the N points onto their ith coordinates, the resulting set of values forms a stratified sample from the unit interval, with one point in each subinterval $[(k-1)/N, k/N)$. The scheme has the additional property that when we partition the hypercube into N subcubes $\prod_{i=1}^{s}[(\ell_i - 1)/n, \ell_i/n)$, each one contains exactly one point. We establish an upper bound for the variance, when we approximate the volume of a subset of the hypercube, with a regular boundary. Numerical experiments assess that the bound is tight. It is possible to employ the extended Latin hypercube samples for Monte Carlo simulation. We focus on the random walk method for diffusion and we show that the variance is reduced when compared with classical random walk using ordinary pseudo-random numbers. The numerical comparisons include stratified sampling and Latin hypercube sampling.

R. El Haddad (✉)
Faculté des Sciences, Département de Mathématiques, Université Saint-Joseph,
BP 11-514 Riad El Solh, Beyrouth 1107 2050, Liban
e-mail: rami.haddad@fs.usj.edu.lb

R. Fakhereddine · C. Lécot
Laboratoire de Mathématiques, Université de Savoie, UMR 5127 CNRS,
Campus scientifique, 73376 Le Bourget-du-Lac, France
e-mail: Rana.Fakhereddine@univ-savoie.fr; Christian.Lecot@univ-savoie.fr

G. Venkiteswaran
Department of Mathematics, Birla Institute of Technology and Science, Pilani,
Jhunjhunu 333 031, Rajasthan, India
e-mail: gvenki.bits@gmail.com

J. Dick et al. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2012*,
Springer Proceedings in Mathematics & Statistics 65,
DOI 10.1007/978-3-642-41095-6_13, © Springer-Verlag Berlin Heidelberg 2013

1 Introduction

Approximating integrals is a basic problem of numerical analysis and may be a component in a more complex computation. Two families of techniques have been developed: deterministic methods and Monte Carlo. We only consider here random algorithms, which are parts of stochastic simulation methods used in applied sciences. Monte Carlo (MC) methods are known to converge slowly, with respect to the number of random points used. Various techniques have been developed, in order to reduce the variance of the approximation, including stratified sampling and Latin hypercube sampling [5,7,9].

Let $s \geq 1$ be a given dimension; then $I^s := [0,1)^s$ is the s-dimensional half-open unit hypercube and λ_s denotes the s-dimensional Lebesgue measure. If g is a square-integrable function defined on I^s, we want to approximate

$$\mathscr{I} := \int_{I^s} g(x) d\lambda_s(x). \tag{1}$$

For the usual MC approximation, $\{U_1, \ldots, U_N\}$ are independent random variables uniformly distributed over I^s. Then

$$X := \frac{1}{N} \sum_{k=1}^{N} g(U_k) \tag{2}$$

is an unbiased estimator of $\mathscr{I}$. A simple stratified sampling (SSS) method was proposed in [10]. Let $\{D_1, \ldots, D_N\}$ be a partition of I^s, so that $\lambda_s(D_1) = \cdots = \lambda_s(D_N) = 1/N$. Let $\{V_1, \ldots, V_N\}$ be independent random variables, with V_ℓ uniformly distributed over D_ℓ. Then

$$Y := \frac{1}{N} \sum_{\ell=1}^{N} g(V_\ell) \tag{3}$$

is another unbiased estimator of $\mathscr{I}$ and for a regular g, one has $\mathrm{Var}(Y) \leq \mathrm{Var}(X)$: we refer to [1, 2, 10] for variance reduction analyses. Latin hypercube sampling (LHS) was introduced in [15]. Let $I_\ell := [(\ell-1)/N, \ell/N)$ for $1 \leq \ell \leq N$ and $\{V_1^i, \ldots, V_N^i\}$ be independent random variables, where V_ℓ^i is uniformly distributed over I_ℓ. If $\{\pi^1, \ldots, \pi^s\}$ are independent random permutations of $\{1, \ldots, N\}$, put $W_\ell := (V^1_{\pi^1(\ell)}, \ldots, V^s_{\pi^s(\ell)})$. Then

$$Z := \frac{1}{N} \sum_{\ell=1}^{N} g(W_\ell) \tag{4}$$

is another unbiased estimator of $\mathscr{I}$. McKay et al. [15] showed that if g is a monotonic function of each of its argument then one has $\mathrm{Var}(Z) \leq \mathrm{Var}(X)$. The analysis in [21] established that for any square-integrable g, LHS does reduce the variance relative to simple random sampling in an asymptotic sense ($N \to \infty$). A proposition in [20] implied that an N-point Latin hypercube sample never leads to a variance greater than that of simple MC with $N - 1$ points. LHS stratifies only the one-dimensional marginals of the uniform distribution over the unit hypercube. Orthogonal array (OA)-based LHS was proposed in [18, 22]. This method generalizes LHS by stratifying low-dimensional ($\leq r$ for OA-based LHS with a corresponding orthogonal array of strength r) marginal distributions. Variance formulas of order $\mathscr{O}(N^{-1})$ were given by Owen [19, 20].

We analyze here a hybrid of SSS and LHS, where the random samples retain some uniformity properties of the nets used in quasi-Monte Carlo methods [17]. More precisely, we construct $N = n^s$ random points in I^s such that in every interval

$$I^{i-1} \times \left[\frac{k-1}{N}, \frac{k}{N}\right) \times I^{s-i} \text{ (for } 1 \leq i \leq s \text{ and } 1 \leq k \leq N)$$

or

$$I_\ell := \prod_{i=1}^{s} \left[\frac{\ell_i - 1}{n}, \frac{\ell_i}{n}\right) \text{ (for } 1 \leq i \leq s \text{ and } 1 \leq \ell_i \leq n)$$

lies only one point of the set (property $\mathscr{P}$): an example is shown on Fig. 1. We call this approach *extended Latin hypercube sampling* (ELHS). In contrast with OA-based LHS, ELHS achieves full (s-dimensional) stratification and also stratifies the one-dimensional marginals. The construction of extended Latin hypercube samples is elementary and requires only random permutations. Both methods are similar in the two-dimensional case.

In Sect. 2 we analyze a MC method using ELHS for numerical integration. Since we have experienced that some simulation methods can be reduced to numerical integration of indicator functions of subdomains of I^s, we focus here on the approximation of the volume of subsets of the unit hypercube. We prove a bound for the variance and we show through numerical experiments that the orders obtained are precise. We compare the variance of the following methods: usual MC, SSS, LHS and ELHS. In Sect. 3, we propose a random walk algorithm for one-dimensional diffusion. Each step of the simulation is formulated as a numerical integration in I^2. In order to benefit from the great uniformity of extended Latin hypercube samples, the particles are sorted by position before performing MC quadrature. The results of a numerical experiment show that the use of ELHS leads to reduced variance, when compared with usual MC, SSS or LHS.

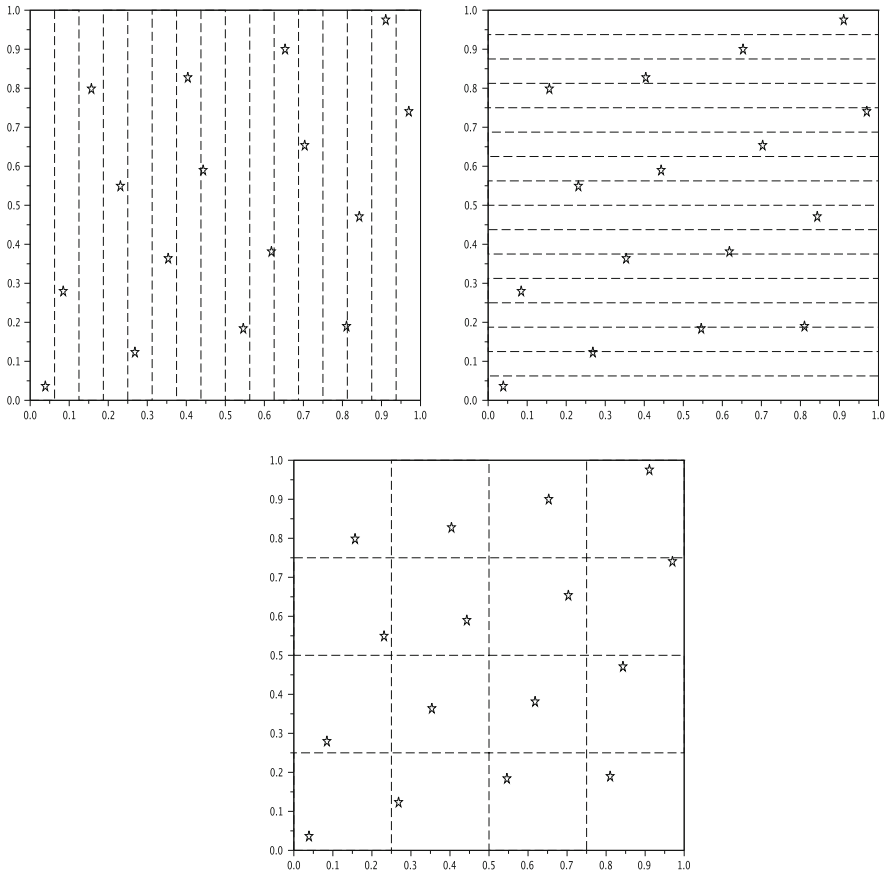


Fig. 1 An extended Latin hypercube sample of 4^2 points ($\star$) in dimension $s = 2$.

2 Numerical Integration

We consider the problem of evaluating integrals like (1) when $g = 1_A$, for some measurable $A \subset I^s$. For usual MC approximation (2), one has

$$\mathrm{Var}(X) = \frac{1}{N}\lambda_s(A)\big(1 - \lambda_s(A)\big) \leq \frac{1}{4N}. \tag{5}$$

We analyze here ELHS using samples of $N = n^s$ points. If $x := (x_1, \ldots, x_s)$, we put $\hat{x}_i := (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_s)$. Let $\sigma^1, \ldots, \sigma^s$ be random bijections

$$\{1, \ldots, n\}^{s-1} \to \{1, \ldots, n^{s-1}\}$$

and $u^1, \ldots, u^s$ be random variables uniformly distributed on I^N; we assume that all these variables are mutually independent. Then we put

$$W_\ell = (W_\ell^1, \ldots, W_\ell^s) \quad \text{with} \quad W_\ell^i := \frac{\ell_i - 1}{n} + \frac{\sigma^i(\hat{\ell}_i) - 1 + u_\ell^i}{N}, \tag{6}$$

for $\ell := (\ell_1, \ldots, \ell_s)$ with $1 \le \ell_i \le n$. Then the point set $\{W_\ell : 1 \le \ell_i \le n\}$ has property $\mathscr{P}$. For $\ell = (\ell_1, \ldots, \ell_s)$ with $1 \le \ell_i \le n$ and $m = (m_1, \ldots, m_s)$ with $1 \le m_i \le n^{s-1}$, let

$$I_{\ell,m} := \prod_{i=1}^{s} \left[\frac{\ell_i - 1}{n} + \frac{m_i - 1}{N}, \frac{\ell_i - 1}{n} + \frac{m_i}{N} \right);$$

then $I_\ell = \bigcup_m I_{\ell,m}$. We have

$$E[1_A(W_\ell)] = \frac{1}{n^{s(s-1)}} \sum_m \int_{I^s} 1_A \left(\frac{\ell_1 - 1}{n} + \frac{m_1 - 1 + u_1}{N}, \ldots, \frac{\ell_s - 1}{n} + \frac{m_s - 1 + u_s}{N} \right) du$$

where the sum extends over all $m = (m_1, \ldots, m_s)$ with $1 \le m_i \le n^{s-1}$. Hence

$$E[1_A(W_\ell)] = N \int_{I_\ell} 1_A(u) du = N \lambda_s(I_\ell \cap A). \tag{7}$$

Consequently, if Z is defined by (4), it is an unbiased estimator of $\mathscr{I}$; we want to estimate $\mathrm{Var}(Z)$.

Proposition 1. *Let $A \subset I^s$ be such that, for all i, with $1 \le i \le s$,*

$$A = \{(u_1, \ldots, u_s) \in I^s : u_i < f_i(\hat{u}_i)\},$$

where $\hat{u}_i := (u_1, \ldots, u_{i-1}, u_{i+1}, \ldots, u_s)$ and f_i are Lipschitz continuous functions $\overline{I}^{s-1} \to \overline{I}$. Let $\{W_\ell : 1 \le \ell_i \le n\}$ be defined by (6). If

$$Z := \frac{1}{N} \sum_\ell 1_A(W_\ell),$$

then

$$\mathrm{Var}(Z) \le \left(\frac{k+2}{4} + 2s(k+2)^2 \right) \frac{1}{N^{1+1/s}},$$

where k is a Lipschitz constant (for the maximum norm) for all the f_i.

Proof. We may write

$$\mathrm{Var}(Z) = \frac{1}{N^2} \sum_\ell \mathrm{Var}(1_A(W_\ell)) + \frac{1}{N^2} \sum_{\ell \ne \ell'} \mathrm{Cov}(1_A(W_\ell), 1_A(W_{\ell'})).$$

From (7) we obtain

$$\frac{1}{N^2}\sum_{\ell}\mathrm{Var}(1_A(W_\ell)) = \sum_{\ell}\mathscr{V}_0(\ell),$$

where

$$\mathscr{V}_0(\ell) = \frac{1}{n^s}\lambda_s(I_\ell \cap A) - (\lambda_s(I_\ell \cap A))^2.$$

Since $\mathscr{V}_0(\ell) = 0$ whenever $I_\ell \subset A$ or $I_\ell \cap A = \emptyset$ and since $0 \le n^s\lambda_s(I_\ell \cap A) \le 1$, we have

$$\sum_{\ell}|\mathscr{V}_0(\ell)| \le \frac{1}{4n^{2s}}\#\{\ell : I_\ell \not\subset A \text{ and } I_\ell \cap A \neq \emptyset\}.$$

Here, $\#E$ denotes the number of elements of a set E. Similarly, we have

$$\frac{1}{N^2}\sum_{\ell\neq\ell'}\mathrm{Cov}(1_A(W_\ell), 1_A(W_{\ell'})) = \sum_{i=1}^{s}\sum_{\substack{\hat{\ell}_i=\hat{\ell}'_i\\ \ell_i\neq\ell'_i}}\mathscr{V}_i(\ell,\ell') + \sum_{\hat{\ell}_j\neq\hat{\ell}'_j}\mathscr{V}_{s+1}(\ell,\ell'),$$

where

$$\mathscr{V}_i(\ell,\ell') = \frac{n^{s(s-1)}}{(n^{s-1}-1)^{s-1}}\sum_{\substack{m_i=m'_i\\ m_j\neq m'_j}}\lambda_s(I_{\ell,m}\cap A)\lambda_s(I_{\ell',m'}\cap A) - \lambda_s(I_\ell\cap A)\lambda_s(I_{\ell'}\cap A),$$

$$\mathscr{V}_{s+1}(\ell,\ell') = \frac{n^{s(s-1)}}{(n^{s-1}-1)^{s}}\sum_{m_j\neq m'_j}\lambda_s(I_{\ell,m}\cap A)\lambda_s(I_{\ell',m'}\cap A) - \lambda_s(I_\ell\cap A)\lambda_s(I_{\ell'}\cap A).$$

And so

$$\sum_{\substack{\hat{\ell}_i=\hat{\ell}'_i\\ \ell_i\neq\ell'_i}}|\mathscr{V}_i(\ell,\ell')| \le$$

$$\frac{1}{n^{2s}}\#\{(\ell,\ell') : \hat{\ell}_i = \hat{\ell}'_i, \ell_i \neq \ell'_i,\ I_\ell \not\subset A,\ I_\ell\cap A\neq\emptyset,\ I_{\ell'}\not\subset A,\ I_{\ell'}\cap A\neq\emptyset\},$$

$$\sum_{\hat{\ell}_j\neq\hat{\ell}'_j}|\mathscr{V}_{s+1}(\ell,\ell')| \le$$

$$\frac{s}{n^{3s-1}}\#\{(\ell,\ell') : \hat{\ell}_j \neq \hat{\ell}'_j,\ I_\ell \not\subset A,\ I_\ell\cap A\neq\emptyset,\ I_{\ell'}\not\subset A,\ I_{\ell'}\cap A\neq\emptyset\}.$$

Consequently

$$\begin{aligned}\mathrm{Var}(Z) \;&\le \frac{1}{4n^{2s}}\#\{\ell : I_\ell \not\subset A \text{ and } I_\ell \cap A \neq \emptyset\}\\ &+\frac{1}{n^{2s}}\sum_{i=1}^{s}\#\{(\ell,\ell') : \hat{\ell}_i = \hat{\ell}'_i, \ell_i \neq \ell'_i,\ I_\ell \not\subset A,\ I_\ell \cap A \neq \emptyset,\ I_{\ell'} \not\subset A,\ I_{\ell'} \cap A \neq \emptyset\}\\ &+\frac{s}{n^{3s-1}}\#\{(\ell,\ell') : \hat{\ell}_j \neq \hat{\ell}'_j,\ I_\ell \not\subset A,\ I_\ell \cap A \neq \emptyset,\ I_{\ell'} \not\subset A,\ I_{\ell'} \cap A \neq \emptyset\}.\end{aligned}$$

Let us note

$$\hat{I}_{\ell,i} = \prod_{\substack{j=1\\ j\neq i}}^{s}\left[\frac{\ell_j - 1}{n}, \frac{\ell_j}{n}\right).$$

We have the following inferences:

- If $I_\ell \not\subset A$, there exists $\hat{u}_{\ell,i} \in \hat{I}_{\ell,i}$ such that $nf_i(\hat{u}_{\ell,i}) < \ell_i$,
- If $I_\ell \cap A \neq \emptyset$, there exists $\hat{v}_{\ell,i} \in \hat{I}_{\ell,i}$ such that $\ell_i < nf_i(\hat{v}_{\ell,i}) + 1$.

Hence

$$\begin{aligned}&\#\{\ell : I_\ell \not\subset A \text{ and } I_\ell \cap A \neq \emptyset\} \le n^{s-1}(k+2),\\ &\#\{(\ell,\ell') : \hat{\ell}_i = \hat{\ell}'_i, \ell_i \neq \ell'_i,\ I_\ell \not\subset A,\ I_\ell \cap A \neq \emptyset,\ I_{\ell'} \not\subset A,\ I_{\ell'} \cap A \neq \emptyset\} \le n^{s-1}(k+2)^2,\\ &\#\{(\ell,\ell') : \hat{\ell}_j \neq \hat{\ell}'_j,\ I_\ell \not\subset A,\ I_\ell \cap A \neq \emptyset,\ I_{\ell'} \not\subset A,\ I_{\ell'} \cap A \neq \emptyset\} \le n^{2(s-1)}(k+2)^2,\end{aligned}$$

and the result follows. □

The variance bound represents a gain in accuracy of the factor $N^{-1/s} = 1/n$ as compared with simple MC. The gain is of diminishing importance as s becomes large and limits the use of the present approach to problems of moderate dimension. This is precisely the case in some MC particle simulations, such as the random walk proposed in Sect. 3. A variance bound with the same order was established in [14]. The differences are as follows. Firstly a two-dimensional analysis in the context of the simulation of Markov chains was conducted in [14] and a possible generalization to higher-dimensional settings was discussed. Secondly the point set used in [14] was obtained by simple stratified sampling over the unit square, with one point in each subsquare $\prod_{i=1}^{2}[(\ell_i - 1)/n, \ell_i/n)$ (without the LHS property).

We use a simple example to illustrate the previous analysis. We consider the subset of the unit ball:

$$Q := \{u \in I^s : \|u\|_2 < 1\},$$

where $\|u\|_2$ denotes the Euclidean norm of u. In order to estimate the variance of the MC, SSS, LHS and ELHS approximations, we replicate the quadrature independently M times and compute the sample variance. We use $M = 100, \ldots, 1{,}000$ and

we only see small differences between the estimates. The results (for $M = 1{,}000$) are displayed in Fig. 2. It appears that the better accuracy due to ELHS goes beyond an improved convergence order: the slope of the curve given by the series of ELHS points is steeper than the slope of the corresponding curve for the MC or LHS points; in addition, the starting ELHS point is below the starting MC or LHS points. The computation times are given in the same figure; one can see that for obtaining the smallest variance achieved by usual MC or LHS, the ELHS approach needs less time. Assuming $\text{Var} = \mathcal{O}(N^{-\alpha})$, linear regression can be used to evaluate α and the outputs are listed in Table 1. The values obtained for ELHS are very close to the orders of the bounds given in Proposition 1, despite the fact that the hypothesis on the boundary of A is not satisfied, since the functions

$$f_i : \hat{u}_i \to \sqrt{1 - u_1^2 - \cdots - u_{i-1}^2 - u_{i+1}^2 - \cdots - u_s^2}$$

are not Lipschitz continuous on $\overline{I}^{s-1}$. This suggests that the hypothesis is too strong.

Table 1 Order α of the variance of the calculation of $\lambda_s(Q)$.

Dimension	MC	LHS	SSS	ELHS
$s = 2$	0.99	1.00	1.48	1.50
$s = 3$	1.00	1.00	1.34	1.33
$s = 4$	1.01	1.00	1.26	1.24

3 Simulation of Diffusion

In many physical applications, there is a need to simulate plain diffusion problems. These problems are frequently encountered as sub-problems while solving more complicated ones. MC simulation has proved a valuable tool for investigating processes involving the diffusion of substances [6,8,23]. In this section we consider a particle method for solving the initial value problem

$$\frac{\partial c}{\partial t}(x,t) = D\frac{\partial^2 c}{\partial x^2}(x,t), \quad x \in \mathbf{R},\ t > 0, \tag{8}$$

$$c(x,0) = c_0(x), \quad x \in \mathbf{R}, \tag{9}$$

with diffusion coefficient $D > 0$. We assume that the initial data satisfies

$$c_0 \geq 0, \quad \int_{\mathbf{R}} c_0(x)dx = 1. \tag{10}$$

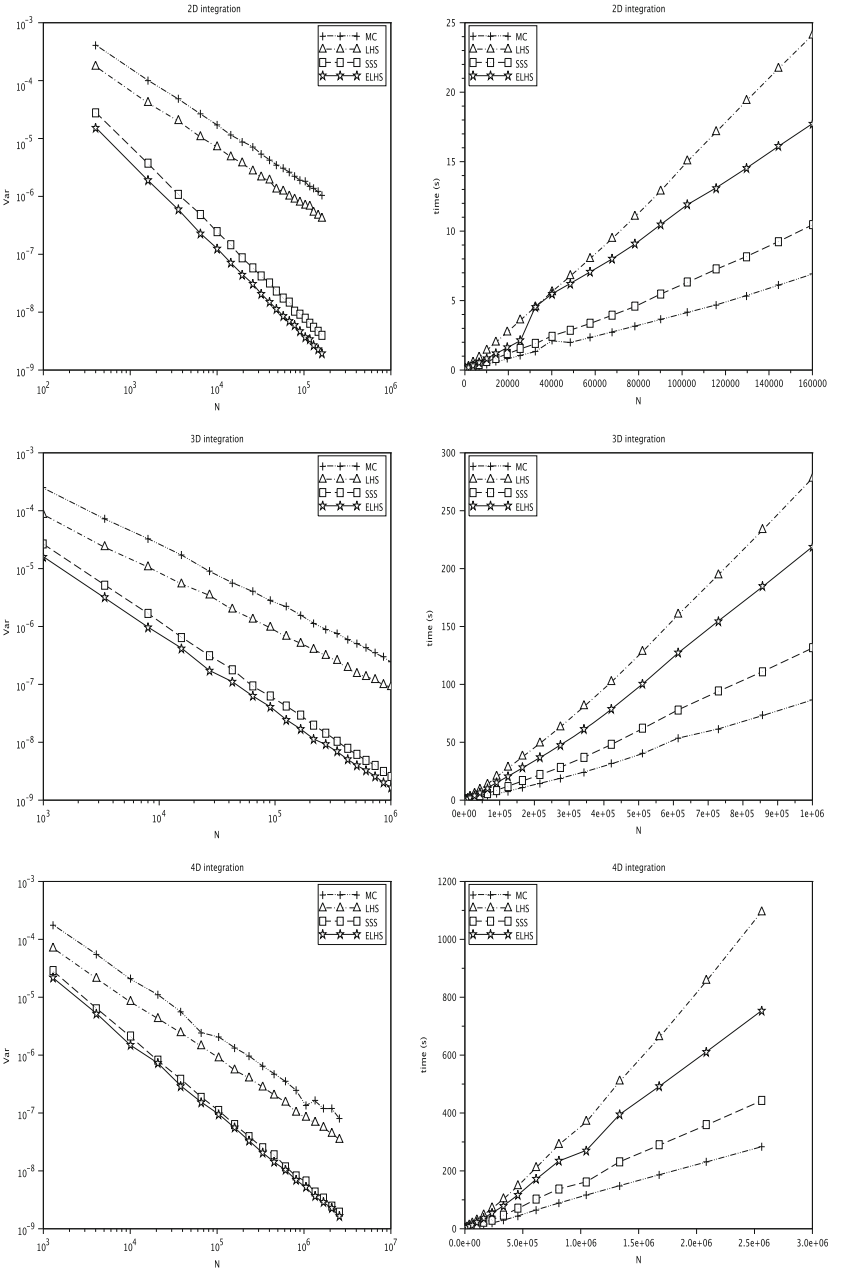


Fig. 2 Sample variance of $M = 1{,}000$ independent copies of the calculation of $\lambda_s(Q)$ as a function of N (*left*, log–log plot) and CPU time in seconds for the sample variance (*right*). Comparison of MC (+), LHS (△), SSS (□) and ELHS methods (⋆) outputs for $s = 2$ and $20^2 \leq N \leq 400^2$ (*top*), $s = 3$ and $10^3 \leq N \leq 100^3$ (*middle*), $s = 4$ and $6^4 \leq N \leq 40^4$ (*bottom*).

The solution possesses the conservation property

$$\forall t > 0 \quad \int_{\mathbf{R}} c(x,t)dx = 1. \tag{11}$$

The fundamental solution for the heat operator $\dfrac{\partial}{\partial t} - D\dfrac{\partial^2}{\partial x^2}$ is

$$E(x,t) := \frac{1}{\sqrt{4\pi Dt}} \mathrm{e}^{-x^2/4Dt}, \quad x \in \mathbf{R},\ t > 0.$$

For any $\tau \geq 0$ the solution of (8) satisfies

$$c(x,t) = \int_{\mathbf{R}} E(x-y, t-\tau)c(y,\tau)dy, \quad x \in \mathbf{R},\ t > \tau. \tag{12}$$

For the numerical approximation of the solution we choose an integer n and we put $N = n^2$. The first step of the simulation involves approximating the initial data u_0 with a sum of Dirac delta functions (particles),

$$c^0(x) := \frac{1}{N}\sum_{k=1}^{N} \delta(x - x_k^0).$$

One has to sample $x_1^0, \ldots, x_N^0$ according to the density function c_0; this may be done by inversion method

$$x_k^0 := C_0^{-1}\left(\frac{2k-1}{2N}\right), \quad 1 \leq k \leq N,$$

where C_0 is the cumulative distribution function associated with c_0. Let Δt be a time step, put $t_p := p\Delta t$ and $c_p(x) := c(x, t_p)$. Given particles at positions x_k^p and the approximate solution

$$c^p(x) := \frac{1}{N}\sum_{k=1}^{N} \delta(x - x_k^p)$$

at time t_p, the solution at time t_{p+1} is obtained as follows.

Generate an extended Latin hypercube sample, as is done in Sect. 2

$$\{W_\ell : 1 \leq \ell_1 \leq n, 1 \leq \ell_2 \leq n\} \subset I^2.$$

Relabel the particles. We order the particles by position:

$$x_1^p \leq x_2^p \leq \cdots \leq x_N^p. \tag{13}$$

This type of sorting was initiated in [11] and used in the context of simulation of diffusion in [12, 16]. Since each step of the random walk algorithm may be described by a numerical integration (see below), the sorting reverts to minimizing the amplitude of the jumps of the function to be integrated.

Diffusion of particles. Using (12), one obtains an approximation to the solution at time t_{p+1}:

$$\tilde{c}^{p+1}(x) := \frac{1}{N}\sum_{k=1}^{N} E(x - x_k^p, \Delta t).$$

Let

$$f(u) := \sqrt{2D\Delta t}\Phi^{-1}(u), \quad u \in (0,1),$$

where Φ denotes the standard normal cumulative distribution function. If $A \subset \mathbf{R}$, denote by 1_A the indicator function. For any measurable $A \subset \mathbf{R}$, one has

$$\int_{\mathbf{R}} \tilde{c}^{p+1}(x) 1_A(x) dx = \frac{1}{N}\sum_{k=1}^{N} \int_I 1_A(x_k^p + f(u)) du. \tag{14}$$

For $1 \le k \le N$, let 1_{I_k} denote the indicator function of $I_k := [(k-1)/N, k/N)$. We associate to any measurable $A \subset \mathbf{R}$ the following indicator function:

$$C_A^{p+1}(u) := \sum_{k=1}^{N} 1_{I_k}(u_1) 1_A(x_k^p + f(u_2)), \quad u = (u_1, u_2) \in I \times (0,1).$$

It is easy to verify that

$$\int_{\mathbf{R}} \tilde{c}^{p+1}(x) 1_A(x) dx = \int_{I^2} C_A^{p+1}(u) du. \tag{15}$$

We recover an approximate solution at time t_{p+1} by performing a MC quadrature using the extended Latin hypercube sample defined above: for any measurable $A \subset \mathbf{R}$

$$\int_{\mathbf{R}} 1_A(x) c^{p+1}(x) = \frac{1}{N}\sum_{\ell_1=1}^{n}\sum_{\ell_2=1}^{n} C_A^{p+1}(W_\ell).$$

The algorithm may be summarized as follows. Let $\lfloor x \rfloor$ denote the greatest integer $\le x$ and put $k(u) := \lfloor Nu \rfloor$. The positions of the particles are updated according to

$$x_{k(W_\ell^1)}^{p+1} = x_{k(W_\ell^1)}^{p} + f(W_\ell^2), \quad 1 \le \ell_1 \le n,\ 1 \le \ell_2 \le n. \tag{16}$$

For any $\ell := (\ell_1, \ell_2)$, the first projection W_ℓ^1 selects the particle number $k(W_\ell^1)$ and the second projection W_ℓ^2 gives the random displacement $f(W_\ell^2)$ of the selected

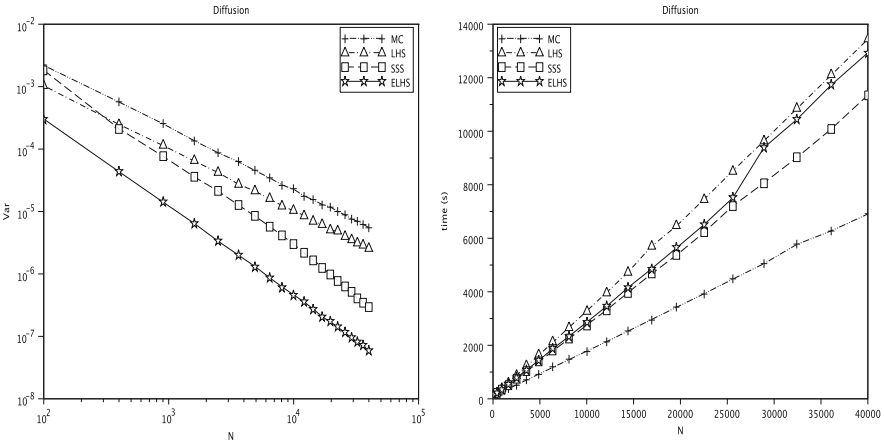


Fig. 3 Sample variance of $M = 5{,}000$ independent copies of the calculation of $\int_0^a c(x,T)dx$ as a function of N (*left*, log–log plot) and CPU time in seconds for the sample variance (*right*). Comparison of MC (+), LHS (△), SSS (□) and ELHS methods (⋆) outputs for $10^2 \leq N \leq 200^2$.

particle. In this algorithm, we may replace extended Latin hypercube samples with simple stratified samples or Latin hypercube samples. The classical random walk algorithm works as follows: there is no reordering of the particles and

$$x_k^{p+1} = x_k^p + f(U_k), \quad 1 \leq k \leq N. \tag{17}$$

Here $U_1, \ldots, U_N$ are independent random samples drawn from the uniform distribution on I.

We compare the approaches in a simple situation. We solve (8)–(9) with $D = 1.0$ and

$$c_0(x) := \frac{1}{\sqrt{\pi}} \mathrm{e}^{-x^2}, \quad x \in \mathbf{R}.$$

We approximate the integral

$$\int_0^a c(x,T)dx,$$

for $a = 4.0$ and $T = 1.0$. The time step is chosen to be $\Delta t := 1/100$. We replicate the computation independently $M = 5{,}000$ times to calculate the sample variance of the MC, SSS, LHS and ELHS approximations. The results are displayed in Fig. 3. As before (Sect. 2), the ELHS method produces better accuracy and improved convergence rate for the variance. The computation times are given in the same figure; one can see that, for the same calculation time, the ELHS technique has a

smaller variance than the other methods. If we assume Var $= \mathcal{O}(N^{-\beta})$, we can estimate β using linear regression; the outputs are listed in Table 2.

Table 2 Order β of the variance of the calculation of $\int_0^a c(x,T)dx$.

MC	LHS	SSS	ELHS
1.00	1.00	1.44	1.43

Although not reported, supplementary results lead to the following remarks. Firstly, it is useless to reorder the particles by position, when using the simple random walk algorithm (with ordinary pseudo-random numbers). Secondly, if we employ ELHS without reordering the particles, the variance is larger than the variance of the simple random walk, and the convergence order (estimated by linear regression) is the same.

4 Conclusion

We have analyzed an extended LHS technique that produces random points which are evenly spread over the unit cube. We have established that for approximate calculation of the measure of some subsets of the hypercube, the technique has a reduced variance, when compared to usual Monte Carlo, simple stratified sampling or Latin hypercube sampling, and a better convergence order.

Then we have modified the classical random walk method for simulation of diffusion. We reorder the particles by position in every time step, and we replace pseudo-random numbers with simple stratified samples, Latin hypercube samples or extended Latin hypercube samples. In an example, we have shown that the method using extended Latin hypercube samples produces lower variance with improved convergence order than the other strategies.

For approximate integration, the hypothesis made on the subsets of the unit hypercube could be relaxed. For the simulation procedure, a bound of the variance is not available: it certainly deserves future work. Another way of progress is in applications of the method to more complex diffusion problems [4] or to Markov chains, as it was done for QMC [3] or randomized QMC methods [13, 14].

References

1. Cheng, R.C.H., Davenport, T.: The problem of dimensionality in stratified sampling. Management Science **35**, 1278–1296 (1989)
2. El-Haddad, R., Fakhreddine, R., Lécot, C.: Stratified Monte Carlo integration. In: Sabelfeld, K.K., Dimov, I. (eds.) Monte Carlo Methods and Applications, pp. 105–113. De Gruyter, Berlin (2013)

3. El-Haddad, R., Lécot, C., L'Ecuyer, P.: Quasi-Monte Carlo simulation of discrete-time Markov chains on multidimensional state spaces. In: Keller, A., Heinrich, S., Niederreiter H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2006, pp. 413–429. Springer, Berlin/Heidelberg (2008)
4. El-Haddad, R., Lécot, C., Venkiteswaran, G.: Quasi-Monte Carlo simulation of diffusion in a spatially nonhomogeneous medium. In: L'Ecuyer, P., Owen, A.B. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2008, pp. 339–354. Springer, Berlin/Heidelberg (2010)
5. Evans, M., Swartz, T.: Approximating Integrals via Monte Carlo and Deterministic Methods. Oxford University Press, Oxford (2000)
6. Farnell, L., Gibson, W.G.: Monte Carlo simulation of diffusion in a spatially nonhomogeneous medium: correction to the Gaussian steplength. J. Comput. Phys. **198**, 65–79 (2004)
7. Fishman, G.S.: Monte Carlo. Springer, New York (1996)
8. Ghoniem, A.F., Sherman, F.S.: Grid-free simulation of diffusion using random walk methods. J. Comput. Phys. **61**, 1–37 (1985)
9. Glasserman, P.: Monte Carlo Methods in Financial Engineering. Springer, New York (2004)
10. Haber, S.: A modified Monte-Carlo quadrature. Math. Comp. **20**, 361–368 (1966)
11. Lécot, C.: A Direct Simulation Monte Carlo scheme and uniformly distributed sequences for solving the Boltzmann equation. Computing **41**, 41–57 (1988)
12. Lécot, C., El-Khettabi, F.: Quasi-Monte Carlo simulation of diffusion. J. Complexity **15**, 342–359 (1999)
13. Lécot, C., Tuffin, B.: Quasi-Monte Carlo methods for estimating transient measures of discrete time Markov chains. In: Niederreiter H. (ed.) Monte Carlo and Quasi-Monte Carlo Methods 2002, pp. 329–343. Springer, Berlin/Heidelberg (2004)
14. L'Ecuyer, P., Lécot, C., Tuffin, B.: A randomized quasi-Monte Carlo simulation method for Markov chains. Oper. Res. **56**, 958–975 (2008)
15. McKay, M.D., Beckman, R.J., Conover, W.J.: A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics **21**, 239–245 (1979)
16. Morokoff, W.J., Caflisch, R.E.: A Quasi-Monte Carlo approach to particle simulation of the heat equation. SIAM J. Numer. Anal. **30**, 1558–1573 (1993)
17. Niederreiter, H.: Random Number Generation and Quasi-Monte Carlo Methods. SIAM, Philadelphia (1992)
18. Owen, A.B.: Orthogonal arrays for computer experiments, integration and visualization. Statist. Sinica **2**, 439–452 (1992)
19. Owen, A.B.: Lattice sampling revisited: Monte Carlo variance of means over randomized orthogonal arrays. Ann. Statist. **22**, 930–945 (1994)
20. Owen, A.B.: Monte Carlo variance of scrambled net quadrature. SIAM J. Numer. Anal. **34**, 1884–1910 (1997)
21. Stein, M.: Large sample properties of simulations using Latin hypercube sampling. Technometrics **29**, 143–151 (1987)
22. Tang, B.: Orthogonal array-based Latin hypercubes. J. Amer. Statist. Assoc. **88**, 1392–1397 (1993)
23. Venkiteswaran, G., Junk, M.: A QMC approach for high dimensional Fokker-Planck equations modelling polymeric liquids. Math. Comput. Simulation **68**, 45–56 (2005)

A Kernel-Based Collocation Method for Elliptic Partial Differential Equations With Random Coefficients

Gregory E. Fasshauer and Qi Ye

Abstract This paper is an extension of previous work where we laid the foundation for the kernel-based collocation solution of stochastic partial differential equations (SPDEs), but dealt only with the simpler problem of right-hand-side Gaussian noises. In the present paper we show that kernel-based collocation methods can be used to approximate the solutions of high-dimensional elliptic partial differential equations with potentially non-Gaussian random coefficients on the left-hand-side. The kernel-based method is a meshfree approximation method, which does not require an underlying computational mesh. The kernel-based solution is a linear combination of a reproducing kernel derived from the related random differential and boundary operators of SPDEs centered at collocation points to be chosen by the user. The random expansion coefficients are obtained by solving a system of random linear equations. For a given kernel function, we show that the convergence of our estimator depends only on the fill distance of the collocation points for the bounded domain of the SPDEs when the random coefficients in the differential operator are random variables. According to our numerical experiments, the kernel-based method produces well-behaved approximate probability distributions of the solutions of SPDEs.

G.E. Fasshauer
Department of Applied Mathematics, Illinois Institute of Technology, Chicago, IL 60616, USA
e-mail: fasshauer@iit.edu

Q. Ye (✉)
Department of Mathematics, Syracuse University, Syracuse, NY 13244, USA
e-mail: qiye@syr.edu

J. Dick et al. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2012*,
Springer Proceedings in Mathematics & Statistics 65,
DOI 10.1007/978-3-642-41095-6_14, © Springer-Verlag Berlin Heidelberg 2013

1 Introduction

Stochastic partial differential equations (SPDEs) represent a recent, fast growing research area which has frequent applications in physics, biology, geology, meteorology and finance. However, in many cases it is difficult to obtain an explicit form of the solutions of these SPDEs. There has been a great interest in devising numerical methods to treat these problems resulting in various techniques such as stochastic Galerkin finite element methods, stochastic collocation methods and Monte Carlo and Quasi-Monte Carlo methods [1, 2, 13, 14]. All of these methods require a finite element approximation restricted to a suitable regular spatial grid in the domain space and the probability space. The *kernel-based* approximation method (*meshfree* approximation method) is a relatively new numerical tool for high-dimensional problems. In our recent papers [5, 11, 16], we use a kernel-based collocation method to approximate the solutions of high-dimensional SPDEs with random noises on the right-hand side. The kernel-based method is very flexible so that its collocation points can be placed at rather arbitrarily scattered locations allowing for the use of either deterministic or random designs, e.g., Halton or Sobol' points. How to find a good design for various PDEs or SPDEs is still a popular open problem which we will not address in this paper. Both stochastic Galerkin finite element methods and stochastic collocation methods are based on truncating the Karhunen-Loéve (KL) expansion of the stochastic fields. To get the KL expansion one is required to know the eigenvalues and the eigenfunctions of the stochastic fields associated with SPDEs. The kernel-based collocation method can avoid the use of KL expansion just as the Quasi-Monte Carlo method does [13, 14]. The Quasi-Monte Carlo method, however, needs to evaluate high-dimensional integrals, whose integral dimensions are equal to the number of nodes of the triangular finite element mesh. As is common with (stochastic) Galerkin methods, one has to project the stochastic field onto the finite element basis. This process is potentially very computationally expensive. For the kernel-based method we simulate the stochastic field at the collocation points directly and solve a system of random linear equations, whose collocation matrix can be exactly obtained. While the kernel-based systems are usually dense systems, the finite element solution can be obtained by a sparse linear system. However, the sparse matrix usually needs to be approximated in the stochastic fields. The kernel-based collocation method requires the solutions of SPDEs to be smooth enough such that the approximate solution is well-behaved at each collocation point while the finite element method is able to solve non-smooth problems.

Our previous papers [5, 11, 16] focus only on solving elliptic SPDEs with right-hand-side Gaussian noises because a parabolic SPDE derived by the white noises can be discretized into several elliptic SPDEs with Gaussian noises. In the classical sense, an SPDE is introduced by the white noise and one refers to a PDE as being *random* when its random part is associated with random coefficients. For convenience, we call a PDE *stochastic* if its solution is a stochastic field. This means that, using our terminology, a PDE whose random part is only dependent

on random differential operators is also called an SPDE. In this paper, we want to extend the results of the kernel-based collocation method stated in [5] to solve the same elliptic PDEs with non-Gaussian random coefficients as discussed in [1, 14], but avoiding finite element constructions and KL expansions. The kernel-based collocation solution is a linear combination of a reproducing kernel derived from the related random differential and boundary operators of the SPDE centered at chosen collocation points (see Eq. (9)). The covariance matrix (collocation matrix) becomes random when the random parts of SPDEs appear on the left-hand side. Moreover, we prove the weak convergence of the kernel-based method and show that the convergence rate depends only on the fill distance of collocation points for the bounded domain of SPDEs for the case when the random coefficients of the differential operator are random variables (see Sect. 4.1) – a topic not covered at all in [5]. Finally, we show this new method to be well-behaved for a two-dimensional stochastic Poisson differential equation with Dirichlet boundary conditions driven by a random coefficient of the elliptic differential operators on the left-hand side (see Sect. 5).

1.1 Problem Setting

Let a vector of random coefficient functions $\mathbf{a} := (a_1, \cdots, a_s)^T$ with $s \in \mathbb{N}$ be defined on a regular open bounded domain $\mathscr{D} \subset \mathbb{R}^d$ and a probability space $(\Omega_{\mathbf{a}}, \mathscr{F}_{\mathbf{a}}, \mathbb{P}_{\mathbf{a}})$, i.e., $\mathbf{a} : \mathscr{D} \times \Omega_{\mathbf{a}} \to \mathbb{R}^s$. Suppose that $\mathbf{a}$ has a nonzero mean and is nonzero almost surely, i.e., $\mathbb{E}(\mathbf{a}) \not\equiv \mathbf{0}$ and $\mathbb{P}_{\mathbf{a}}(\mathbf{a} \not\equiv \mathbf{0}) = 1$, and that $\mathbb{P}_{\mathbf{a}}\left(\left\{a_j \in \mathrm{C}(\overline{\mathscr{D}})\right\}_{j=1}^{s}\right) = 1$. Consider an elliptic partial differential equation driven by the random coefficients $a_1, \ldots, a_s$

$$\begin{cases} L_{\mathbf{a}} u = f, & \text{in } \mathscr{D}, \\ Bu = g, & \text{on } \partial\mathscr{D}, \end{cases} \tag{1}$$

where $L_{\mathbf{a}}$ is a linear elliptic differential operator with the random coefficients $\mathbf{a}$, B is a boundary operator for Dirichlet or Neumann boundary conditions, and $f : \mathscr{D} \to \mathbb{R}$, $g : \partial\mathscr{D} \to \mathbb{R}$ are deterministic functions. For example, $L_{\mathbf{a}} := a\Delta$ or $L_{\mathbf{a}} := a_1\Delta - a_2 I$ and $B := I|_{\partial\mathscr{D}}$ (see Formulas (4)).

Remark 1. For each realization of the random coefficients $\mathbf{a}$, we can obtain the realization of the random differential operator $L_{\mathbf{a}}$ in order to introduce different deterministic partial differential equations from the SPDE (1).

In what follows we assume that the probability structure of the random coefficients $\mathbf{a}$ is given, e.g., if $\mathbf{a} = a$ (scalar coefficient) is a random variable then we know its cumulative distribution function $F_a : \mathbb{R} \to [0, 1]$, and if $\mathbf{a} = a$ is a Gaussian field then its mean $m_a : \mathscr{D} \to \mathbb{R}$ and covariance kernel $R_a : \mathscr{D} \times \mathscr{D} \to \mathbb{R}$ are known.

We assume that the elliptic SPDE (1) for suitable left-hand sides $\mathbf{a}$ and right-hand sides f, g has a unique solution $u \in \mathscr{H}^m(\mathscr{D})$ almost surely (see [4]), where $\mathscr{H}^m(\mathscr{D})$ is the classical L_2-based Sobolev space of order $m > \mathscr{O}(L_{\mathbf{a}}) + d/2$ and $\mathscr{O}(L_{\mathbf{a}})$ is the order of $L_{\mathbf{a}}$, which is independent of the order of differential derivatives.

The kernel-based collocation method for solving the SPDE (1) can be described as follows:

1. We firstly choose a finite collection of predetermined *collocation points*

$$X_{\mathscr{D}} := \{\mathbf{x}_1, \cdots, \mathbf{x}_N\} \subset \mathscr{D}, \quad X_{\partial\mathscr{D}} := \{\mathbf{x}_{N+1}, \cdots, \mathbf{x}_{N+M}\} \subset \partial\mathscr{D},$$

   and fix a reproducing kernel $K : \mathscr{D} \times \mathscr{D} \to \mathbb{R}$, whose reproducing kernel Hilbert space $\mathscr{H}_K(\mathscr{D})$ is equivalent to $\mathscr{H}^m(\mathscr{D})$.
2. Next we sample the values $\{y_j\}_{j=1}^N$ and $\{y_{N+j}\}_{j=1}^M$ from the SPDE (1) at the collocation points $X_{\mathscr{D}}$ and $X_{\partial\mathscr{D}}$, i.e.,

$$y_j := f(\mathbf{x}_j), \quad j = 1, \ldots, N, \quad y_{N+j} := g(\mathbf{x}_{N+j}), \quad j = 1, \ldots, M, \tag{2}$$

   and simulate the random coefficients $\mathbf{a} = (a_1, \cdots, a_s)^T$ at the collocation points $X_{\mathscr{D}}$, i.e.,

$$\mathbf{a}_{\mathbf{x}_1}, \ldots, \mathbf{a}_{\mathbf{x}_N} \sim \text{some joint probability distributions.}$$

   For example, when $\mathbf{a} = a$ is a random variable with cumulative distribution function F_a then

$$a_{\mathbf{x}_1} = \cdots = a_{\mathbf{x}_N} = a \sim F_a^{-1}(U), \quad U \sim \text{Unif}[0, 1],$$

   where F_a^{-1} is the inverse of F_a and U is a random variable with uniform distribution on $[0, 1]$. If, on the other hand, $\mathbf{a} = a$ is a Gaussian field with mean m_a and covariance kernel R_a, then $a_{\mathbf{x}_1}, \ldots, a_{\mathbf{x}_N}$ have joint multi-normal distributions with mean $\mathbf{m}_X$ and covariance matrix R_X, i.e.,

$$(a_{\mathbf{x}_1}, \cdots, a_{\mathbf{x}_N})^T \sim \mathscr{N}(\mathbf{m}_X, \mathsf{R}_X),$$

   where $\mathbf{m}_X := (m_a(\mathbf{x}_1), \cdots, m_a(\mathbf{x}_N))^T$ and $\mathsf{R}_X := \left(R_a(\mathbf{x}_j, \mathbf{x}_k)\right)_{j,k=1}^{N,N}$ (see [12, Chapter 2]).
3. Finally, we approximate the solution u of the SPDE (1) using a kernel-based collocation method written as

$$u(\mathbf{x}) \approx \hat{u}(\mathbf{x}) := \sum_{k=1}^{N} c_k L_{\mathbf{a}_{\mathbf{x}_k},2} \overset{*}{K}(\mathbf{x}, \mathbf{x}_k) + \sum_{k=1}^{M} c_{N+k} B_2 \overset{*}{K}(\mathbf{x}, \mathbf{x}_{N+k}), \quad \mathbf{x} \in \mathscr{D}, \tag{3}$$

where $\overset{*}{K}$ is an integral-type kernel of K (see Formula (5)). Here $L_{\mathbf{a}_\mathbf{x},2}$ and B_2 mean that we differentiate with respect to the second argument, i.e., $L_{\mathbf{a}_{\mathbf{x}_k},2}\overset{*}{K}(\mathbf{x},\mathbf{x}_k) = L_{\mathbf{a}_\mathbf{y},2}\overset{*}{K}(\mathbf{x},\mathbf{y})|_{\mathbf{y}=\mathbf{x}_k}$ and $B_2\overset{*}{K}(\mathbf{x},\mathbf{x}_k) = B_\mathbf{y}\overset{*}{K}(\mathbf{x},\mathbf{y})|_{\mathbf{y}=\mathbf{x}_k}$. The unknown random coefficients $c_1,\ldots,c_{N+M}$ are obtained by solving a system of random linear equations (with deterministic right-hand side and random matrix that varies with each realization of the random coefficients). Details are provided in Sect. 4.

2 Reproducing Kernels and Matérn Functions

Definition 1 ([15, Definition 10.1]). A Hilbert space $\mathscr{H}_K(\mathscr{D})$ consisting of functions $f : \mathscr{D} \to \mathbb{R}$ is called a *reproducing-kernel Hilbert space* with a *reproducing kernel* $K : \mathscr{D} \times \mathscr{D} \to \mathbb{R}$ if (i) $K(\cdot,\mathbf{y}) \in \mathscr{H}_K(\mathscr{D})$ and (ii) $f(\mathbf{y}) = (f, K(\cdot,\mathbf{y}))_{\mathscr{H}_K(\mathscr{D})}$ for all $f \in \mathscr{H}_K(\mathscr{D})$ and all $\mathbf{y} \in \mathscr{D}$.

In our recent papers [9, 10, 16] we show that reproducing kernels suitable for the numerical solution of (S)PDEs can be computed from the *Matérn functions* (Sobolev splines) and that their reproducing kernel Hilbert spaces are equivalent to certain Sobolev spaces. According to [9, Example 5.7] and [16, Example 4.4], the Matérn function with shape parameter $\theta > 0$ and degree $m > d/2$

$$G_\theta(\mathbf{x}) := \frac{2^{1-m-d/2}}{\pi^{d/2}\Gamma(m)\theta^{2m-d}}(\theta\,\|\mathbf{x}\|_2)^{m-d/2}K_{d/2-m}(\theta\,\|\mathbf{x}\|_2), \quad \mathbf{x} \in \mathbb{R}^d,$$

is a full-space Green function of the differential operator $L := \left(\theta^2 I - \Delta\right)^m$, where $t \mapsto K_\nu(t)$ is the modified Bessel function of the second kind of order ν. The reproducing kernel related to G_θ,

$$K_\theta(\mathbf{x},\mathbf{y}) := G_\theta(\mathbf{x}-\mathbf{y}), \quad \mathbf{x},\mathbf{y} \in \mathbb{R}^d,$$

is positive definite. Moreover, its reproducing-kernel Hilbert space is equivalent to the L_2-based Sobolev space of order m, i.e.,

$$\mathscr{H}_{K_\theta}(\mathbb{R}^d) \cong \mathscr{H}^m(\mathbb{R}^d)$$

and its inner product has the explicit form

$$(f,g)_{\mathscr{H}_{K_\theta}(\mathbb{R}^d)} := \int_{\mathbb{R}^d} \mathbf{P}f(\mathbf{x})^T\mathbf{P}g(\mathbf{x})\mathrm{d}\mathbf{x},$$

where $\mathbf{P} := \left(\mathbf{Q}_0^T, \mathbf{Q}_1^T, \cdots, \mathbf{Q}_m^T\right)^T$ and

$$\mathbf{Q}_j := \begin{cases} \kappa_j \Delta^k, & j = 2k, \\ \kappa_j \Delta^k \nabla^T, & j = 2k+1, \end{cases} \qquad \kappa_j := \sqrt{\frac{m!\theta^{2m-2j}}{j!(m-j)!}}, \quad k \in \mathbb{N}_0, \quad j = 0, 1, \ldots, m.$$

According to [3, Theorem 1.4.6], the reproducing-kernel Hilbert space $\mathscr{H}_{K_\theta}(\mathscr{D})$ on the bounded domain $\mathscr{D}$ is endowed with the reproducing-kernel norm

$$\|f\|_{\mathscr{H}_{K_\theta}(\mathscr{D})} := \inf_{\tilde{f} \in \mathscr{H}_{K_\theta}(\mathbb{R}^d)} \left\{ \left\| \tilde{f} \right\|_{\mathscr{H}_{K_\theta}(\mathbb{R}^d)} : \tilde{f}|_{\mathscr{D}} = f \right\}.$$

If the open bounded domain $\mathscr{D} \subset \mathbb{R}^d$ is regular, then $\mathscr{H}_{K_\theta}(\mathscr{D})$ is again equivalent to the L_2-based Sobolev space of order m, i.e.,

$$\mathscr{H}_{K_\theta}(\mathscr{D}) \cong \mathscr{H}^m(\mathscr{D}).$$

3 Constructing Gaussian Fields by Reproducing Kernels with Differential and Boundary Operators

For the reader's convenience we repeat the theoretical results from [5, 16] that are essential to our discussion later on.

Definition 2 ([3, Definition 3.28]). A stochastic process $S : \mathscr{D} \times \Omega \to \mathbb{R}$ is said to be *Gaussian* with mean $\mu : \mathscr{D} \to \mathbb{R}$ and covariance kernel $R : \mathscr{D} \times \mathscr{D} \to \mathbb{R}$ on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ if, for any pairwise distinct points $X := \{\mathbf{x}_1, \cdots, \mathbf{x}_N\} \subset \mathscr{D}$, the random vector

$$\mathbf{S}_X := (S_{\mathbf{x}_1}, \cdots, S_{\mathbf{x}_N})^T \sim \mathscr{N}(\boldsymbol{\mu}_X, \mathsf{R}_X),$$

is a multi-normal random vector with mean $\boldsymbol{\mu}_X := (\mu(\mathbf{x}_1), \cdots, \mu(\mathbf{x}_N))^T$ and covariance matrix $\mathsf{R}_X := \left(R(\mathbf{x}_j, \mathbf{x}_k)\right)_{j,k=1}^{N,N}$.

Let a *differential operator* $L : \mathscr{H}^m(\mathscr{D}) \to L_2(\mathscr{D})$ and a *boundary operator* $B : \mathscr{H}^m(\mathscr{D}) \to L_2(\partial\mathscr{D})$ be linear combinations of derivatives D^α with $\alpha \in \mathbb{N}_0^d$ and nonconstant coefficients defined on $\mathscr{D}$ and $\partial\mathscr{D}$ respectively, i.e.,

$$L := \sum_{|\alpha| \le m} c_\alpha D^\alpha, \quad B := \sum_{|\alpha| \le m-1} b_\alpha D^\alpha|_{\partial\mathscr{D}}, \tag{4}$$

where $c_\alpha \in C(\overline{\mathscr{D}})$ and $b_\alpha \in C(\partial\mathscr{D})$. Moreover, the orders of these operators are given by

$$\mathscr{O}(L) := \max\{|\alpha| : c_\alpha \not\equiv 0\}, \quad \mathscr{O}(B) := \max\{|\alpha| : b_\alpha \not\equiv 0\}.$$

The following theorem shows that we can view the reproducing kernel Hilbert space $\Omega_K := \mathscr{H}_K(\mathscr{D})$ as a sample space and its Borel σ-field $\mathscr{F}_K := \mathscr{B}(\mathscr{H}_K(\mathscr{D}))$ as a σ-algebra to set up the probability spaces $\mathbb{P}^\mu$ so that the stochastic fields $LS_{\mathbf{x}}(\omega) := L\omega(\mathbf{x})$ and $BS_{\mathbf{x}} := B\omega(\mathbf{x})$ are Gaussian, where $\omega \in \Omega_K$.

Theorem 1 ([5, Theorem 3.1] and [16, Theorem 7.2]). *Suppose that the reproducing kernel Hilbert space $\mathscr{H}_K(\mathscr{D})$ is embedded into the Sobolev space $\mathscr{H}^m(\mathscr{D})$ with $m > d/2$. Further assume that the differential operator L and the boundary operator B have the orders $\mathscr{O}(L) < m - d/2$ and $\mathscr{O}(B) < m - d/2$. Given a function $\mu \in \mathscr{H}_K(\mathscr{D})$ there exists a probability measure (Gaussian measure) $\mathbb{P}^\mu$ defined on the measurable space $(\Omega_K, \mathscr{F}_K) := (\mathscr{H}_K(\mathscr{D}), \mathscr{B}(\mathscr{H}_K(\mathscr{D})))$ such that the stochastic processes LS, BS given by*

$$LS_{\mathbf{x}}(\omega) = LS(\mathbf{x}, \omega) := (L\omega)(\mathbf{x}), \quad \mathbf{x} \in \mathscr{D}, \quad \omega \in \Omega_K = \mathscr{H}_K(\mathscr{D}),$$
$$BS_{\mathbf{x}}(\omega) = BS(\mathbf{x}, \omega) := (B\omega)(\mathbf{x}), \quad \mathbf{x} \in \partial\mathscr{D}, \quad \omega \in \Omega_K = \mathscr{H}_K(\mathscr{D}),$$

are Gaussian with means $L\mu$, $B\mu$ and covariance kernels

$$L_1 L_2 \overset{*}{K}(\mathbf{x}, \mathbf{y}) = \int_{\mathscr{D}} L_1 K(\mathbf{x}, \mathbf{z}) L_1 K(\mathbf{y}, \mathbf{z}) \mathrm{d}\mathbf{z}, \quad \mathbf{x}, \mathbf{y} \in \mathscr{D},$$
$$B_1 B_2 \overset{*}{K}(\mathbf{x}, \mathbf{y}) = \int_{\mathscr{D}} B_1 K(\mathbf{x}, \mathbf{z}) B_1 K(\mathbf{y}, \mathbf{z}) \mathrm{d}\mathbf{z}, \quad \mathbf{x}, \mathbf{y} \in \partial\mathscr{D},$$

defined on $(\Omega_K, \mathscr{F}_K, \mathbb{P}^\mu)$ respectively, where the integral-type kernel $\overset{*}{K}$ of the reproducing kernel K is given by

$$\overset{*}{K}(\mathbf{x}, \mathbf{y}) := \int_{\mathscr{D}} K(\mathbf{x}, \mathbf{z}) K(\mathbf{y}, \mathbf{z}) \mathrm{d}\mathbf{z}, \quad \mathbf{x}, \mathbf{y} \in \mathscr{D}. \tag{5}$$

When $L := I$ then we rewrite $LS = S$ which indicates that $S_{\mathbf{x}}(\omega) = \omega(\mathbf{x})$. (Here L_1, B_1 and L_2, B_2 denote the differential and boundary operators with respect to the first and second arguments, respectively.)

Moreover, the probability measure $\mathbb{P}^\mu$ is only related to μ and $\overset{*}{K}$, which indicates that it is independent of L and B.

4 Kernel-Based Collocation Methods

For convenience, we only discuss the case when the random coefficients $\mathbf{a}$ of the differential operator $L_{\mathbf{a}}$ of the SPDE (1) consist of single a scalar random function a, i.e., $\mathbf{a} = a$. This means that the random differential operator only has one random coefficient a. We therefore denote the operator by L_a. Moreover, if we know the means and the covariance structures of a vector of Gaussian coefficients, then it is

not difficult to generalize the case of scalar random coefficients presented here to that of multiple random coefficients. In this section, we only solve an SPDE with a scalar random coefficient. This is similar to what was done in [1, 2, 13]. We intend to investigate non-Gaussian multiple random coefficients in our future research.

4.1 Random Coefficients as Random Variables

We firstly consider the random coefficient a of L_a as a random variable, which means that $z = a(\omega)$ is a real constant for any sample $\omega \in \Omega_a$. So L_z becomes a deterministic elliptic differential operator for any fixed real constant $z \in \mathbb{R}$.

We use the Gaussian fields $L_z S$, BS with means $L_z\mu$, $B\mu$ and covariance kernels $L_{z,1}L_{z,2}\overset{*}{K}$, $B_1B_2\overset{*}{K}$ (see Theorem 1), respectively, to construct the kernel-based solution $\hat{u}$ to estimate the solution u of the SPDE (1). Here the covariance kernels are defined by $L_{z,1}L_{z,2}\overset{*}{K}(\mathbf{x}_j, \mathbf{x}_k) = L_{z,\mathbf{x}}L_{z,\mathbf{y}}\overset{*}{K}(\mathbf{x}, \mathbf{y})|_{\mathbf{x}=\mathbf{x}_j,\mathbf{y}=\mathbf{x}_k}$ and $B_1B_2\overset{*}{K}(\mathbf{x}_j, \mathbf{x}_k) = B_{\mathbf{x}}B_{\mathbf{y}}\overset{*}{K}(\mathbf{x}, \mathbf{y})|_{\mathbf{x}=\mathbf{x}_j,\mathbf{y}=\mathbf{x}_k}$.

Remark 2. Since we want to interpolate the values of the differential equation at the collocation points, $L_z\omega(\mathbf{x})$ needs to be well-defined pointwise for each available solution $\omega \in \mathscr{H}_K(\mathscr{D}) \cong \mathscr{H}^m(\mathscr{D}) \subset \mathrm{C}^2(\overline{\mathscr{D}})$. This requires the Sobolev space $\mathscr{H}^m(\mathscr{D})$ to be smooth for second-order elliptic differential operators. If we just need a weak solution as for the finite element method, then the order needs to satisfy $m \geq 2$ only.

We define the vector

$$\mathbf{y}_0 := (y_1, \cdots, y_N, y_{N+1} \cdots, y_{N+M})^T,$$

where the values $\{y_j\}_{j=1}^N$ and $\{y_{N+j}\}_{j=1}^M$ are given in Eq. (2), and we also define the product space

$$\Omega_{Ka} := \Omega_K \times \Omega_a, \quad \mathscr{F}_{Ka} := \mathscr{F}_K \otimes \mathscr{F}_a, \quad \mathbb{P}_a^{\mu} := \mathbb{P}^{\mu} \otimes \mathbb{P}_a,$$

where the probability measure $\mathbb{P}^{\mu}$ is defined on $(\mathscr{H}_K(\mathscr{D}), \mathscr{B}(\mathscr{H}_K(\mathscr{D}))) = (\Omega_K, \mathscr{F}_K)$ independent of the differential and boundary operators as in Theorem 1. The probability space $(\Omega_a, \mathscr{F}_a, \mathbb{P}_a)$ comes from the SPDE (1). We extend the random variables defined on the original probability spaces to random variables on the new probability space in the natural way: if random variables $V_1 : \Omega_K \to \mathbb{R}$ and $V_2 : \Omega_a \to \mathbb{R}$ are defined on $(\Omega_K, \mathscr{F}_K, \mathbb{P}^{\mu})$ and $(\Omega_a, \mathscr{F}_a, \mathbb{P}_a)$, respectively, then

$$V_1(\omega_1, \omega_2) := V_1(\omega_1), \; V_2(\omega_1, \omega_2) := V_2(\omega_2), \quad \text{for each } \omega_1 \in \Omega_K \text{ and } \omega_2 \in \Omega_a.$$

Note that in this case the random variables have the same probability distributional properties, and they are independent on $\left(\Omega_{Ka}, \mathscr{F}_{Ka}, \mathbb{P}_a^{\mu}\right)$. This implies that the stochastic processes $L_z S, S$ and a can be extended to the product space $\left(\Omega_{Ka}, \mathscr{F}_{Ka}, \mathbb{P}_a^{\mu}\right)$ while preserving the original probability distributional properties, and that $(L_z S, S)$ and a are independent.

Therefore, [5, Corollary 3.2] and [16, Corollary 7.3] show that the random vector

$$\mathbf{S}_{X,a} := \left(L_a S_{\mathbf{x}_1}, \cdots, L_a S_{\mathbf{x}_N}, B S_{\mathbf{x}_{N+1}} \cdots, B S_{\mathbf{x}_{N+M}}\right)^T$$

conditioned on $a = z$ has a *multi-normal* distribution with mean

$$\boldsymbol{\mu}_{X,z} := (L_z \mu(\mathbf{x}_1), \cdots, L_z \mu(\mathbf{x}_N), B\mu(\mathbf{x}_{N+1}), \cdots, B\mu(\mathbf{x}_{N+M}))^T$$

and covariance matrix (collocation matrix)

$$\overset{*}{\mathsf{K}}_{X,z} := \begin{pmatrix} \left(L_{z,1} L_{z,2} \overset{*}{K}(\mathbf{x}_j, \mathbf{x}_k)\right)_{j,k=1}^{N,N}, & \left(L_{z,1} B_2 \overset{*}{K}(\mathbf{x}_j, \mathbf{x}_{N+k})\right)_{j,k=1}^{N,M} \\ \left(B_1 L_{z,2} \overset{*}{K}(\mathbf{x}_{N+j}, \mathbf{x}_k)\right)_{j,k=1}^{M,N}, & \left(B_1 B_2 \overset{*}{K}(\mathbf{x}_{N+j}, \mathbf{x}_{N+k})\right)_{j,k=1}^{M,M} \end{pmatrix} \in \mathbb{R}^{N+M,N+M},$$

where $X := X_{\mathscr{D}} \cup X_{\partial\mathscr{D}}$. We can check that $\overset{*}{\mathsf{K}}_{X,z}$ is always positive semi-definite. This collocation matrix is used to set up the random linear equations one needs to solve for the random coefficients of the kernel-based solution.

Fix any $\mathbf{x} \in \mathscr{D}$. By Bayes' rule, we can obtain the *conditional probability density function* $p_{\mathbf{x}}^{\mu}$ of $S_{\mathbf{x}}$ given $\mathbf{S}_{X,a}$ and a, i.e., for any $\mathbf{v} \in \mathbb{R}^{N+M}$ and any $z \in \mathbb{R}$,

$$p_{\mathbf{x}}^{\mu}(v|\mathbf{v}, z) := \frac{1}{\sigma(\mathbf{x}|z)\sqrt{2\pi}} \exp\left(-\frac{(v - m_{\mathbf{x}}^{\mu}(\mathbf{v}|z))^2}{2\sigma(\mathbf{x}|z)^2}\right), \quad v \in \mathbb{R}, \tag{6}$$

where

$$\begin{aligned} m_{\mathbf{x}}^{\mu}(\mathbf{v}|z) &:= \mu(\mathbf{x}) + \overset{*}{\mathbf{k}}_{X,z}(\mathbf{x})^T \overset{*}{\mathsf{K}}_{X,z}{}^{\dagger} \left(\mathbf{v} - \boldsymbol{\mu}_{X,z}\right), \\ \sigma(\mathbf{x}|z)^2 &:= \overset{*}{K}(\mathbf{x}, \mathbf{x}) - \overset{*}{\mathbf{k}}_{X,z}(\mathbf{x})^T \overset{*}{\mathsf{K}}_{X,z}{}^{\dagger} \overset{*}{\mathbf{k}}_{X,z}(\mathbf{x}), \end{aligned} \tag{7}$$

and $\overset{*}{\mathbf{k}}_{X,z}(\mathbf{x}) := \left(L_{z,2} \overset{*}{K}(\mathbf{x}, \mathbf{x}_1), \cdots, L_{z,2} \overset{*}{K}(\mathbf{x}, \mathbf{x}_N), B_2 \overset{*}{K}(\mathbf{x}, \mathbf{x}_{N+1}), \cdots, B_2 \overset{*}{K}(\mathbf{x}, \mathbf{x}_{N+M})\right)^T$. In particular, $S_{\mathbf{x}}$ conditioned on $\mathbf{S}_{X,a} = \mathbf{y}_0$ and $a = z$ has the probability density function $p_{\mathbf{x}}^{\mu}(\cdot|\mathbf{y}_0, z)$. Here the dagger † denotes the pseudo inverse. The vector $\overset{*}{\mathbf{k}}_{X,z}$ contains the basis for the kernel-based solution and $\sigma(\mathbf{x}|z)^2$ is used in our derivation of the error bounds for the kernel-based solution.

4.1.1 Constructing Kernel-Based Solutions

We use techniques from statistics to construct the kernel-based solution $\hat{u}$. Fix any $\mathbf{x} \in \mathscr{D}$. Let

$$\mathscr{A}_{\mathbf{x}}(v) := \{\omega_1 \times \omega_2 \in \Omega_{Ka} : \omega_1(\mathbf{x}) = v\},$$

and

$$\begin{aligned}\mathscr{A}_{X,a}(\mathbf{y}_0, z) := \{&\omega_1 \times \omega_2 \in \Omega_{Ka} : L_z\omega_1(\mathbf{x}_1) = y_1, \ldots, L_z\omega_1(\mathbf{x}_N) = y_N,\\ &B\omega_1(\mathbf{x}_{N+1}) = y_{N+1}, \ldots, B\omega_1(\mathbf{x}_{N+M}) = y_{N+M} \text{ and } a(\omega_2) = z\}.\end{aligned}$$

Since $L_z S_{\mathbf{x}}(\omega_1, \omega_2) = L_z S_{\mathbf{x}}(\omega_1) = L_z\omega_1(\mathbf{x})$, $BS_{\mathbf{x}}(\omega_1, \omega_2) = BS_{\mathbf{x}}(\omega_1) = B\omega_1(\mathbf{x})$ and $a(\omega_1, \omega_2) = a(\omega_2)$ for each $\omega_1 \in \Omega_K$ and $\omega_2 \in \Omega_a$, we can use the same methods as in [5] and [16, Chapter 7] to obtain that

$$\mathbb{P}_a^{\mu}(\mathscr{A}_{\mathbf{x}}(v)|\mathscr{A}_{X,a}(\mathbf{y}_0, z)) = \mathbb{P}_a^{\mu}(S_{\mathbf{x}} = v|\mathbf{S}_{X,a} = \mathbf{y}_0, a = z) = p_{\mathbf{x}}^{\mu}(v|\mathbf{y}_0, z).$$

We can view z as a realization of a and simulate the random coefficient a by its cumulative distribution function F_a, i.e.,

$$a \sim F_a^{-1}(U), \quad U \sim \text{Unif}[0, 1],$$

where Unif means the standard uniform distribution. We now obtain our estimator $\hat{u}(\mathbf{x})$ by solving an optimization problem that is reminiscent of the maximum likelihood method. The estimator is given by

$$\hat{u}(\mathbf{x}, \omega_2) := \operatorname*{argmax}_{v \in \mathbb{R}} \sup_{\mu \in \mathscr{H}_K(\mathscr{D})} p_{\mathbf{x}}^{\mu}(v|\mathbf{y}_0, a(\omega_2)), \quad \omega_2 \in \Omega_2. \tag{8}$$

Assume that the random covariance matrix $\overset{*}{K}_{X,a}$ is nonsingular almost surely. One optimal solution of the maximization problem (8) has the form

$$\hat{u}(\mathbf{x}) = \sum_{k=1}^{N} c_k L_{a,2} \overset{*}{K}(\mathbf{x}, \mathbf{x}_k) + \sum_{k=1}^{M} c_{N+k} B_2 \overset{*}{K}(\mathbf{x}, \mathbf{x}_{N+k}) = \overset{*}{\mathbf{k}}_{X,a}(\mathbf{x})^T \overset{*}{\mathsf{K}}_{X,a}{}^{-1} \mathbf{y}_0. \tag{9}$$

This means that the random coefficients of the estimator can be computed from the system of random linear equations

$$\overset{*}{\mathsf{K}}_{X,a}\mathbf{c} = \mathbf{y}_0,$$

where $\mathbf{c} := (c_1, \cdots, c_{N+M})^T$ and the matrix $\overset{*}{\mathsf{K}}_{X,a}$ was defined in the previous section.

The estimator $\hat{u}$ also satisfies the interpolation conditions almost surely, i.e., $L_a\hat{u}(\mathbf{x}_1) = y_1, \ldots, L_a\hat{u}(\mathbf{x}_N) = y_N$ and $B\hat{u}(\mathbf{x}_{N+1}) = y_{N+1}, \ldots, B\hat{u}(\mathbf{x}_{N+M}) = y_{N+M}$. It is obvious that $\hat{u}(\cdot, \omega_2) \in \mathscr{H}_K(\mathscr{D})$ for each $\omega_2 \in \Omega_a$. Since the random part of $\hat{u}(\mathbf{x})$ is only related to the random variable a, we can formally rewrite $\hat{u}(\mathbf{x}, \omega_2)$ as $\hat{u}(\mathbf{x}, a)$ and $\hat{u}(\mathbf{x})$ can be transferred to a random variable defined on the one-dimensional probability space $(\mathbb{R}, \mathscr{B}(\mathbb{R}), \mathrm{d}F_a)$.

4.1.2 Convergence Analysis

Since $u(\cdot, \omega_2)$ belongs to $\mathscr{H}_K(\mathscr{D}) \cong \mathscr{H}^m(\mathscr{D})$ almost surely for $\omega_2 \in \Omega_a$, u can be seen as a map from Ω_a into $\mathscr{H}_K(\mathscr{D})$. So we have $u \in \Omega_{Ka} = \Omega_K \times \Omega_a$.

We fix any $\mathbf{x} \in \mathscr{D}$ and any $\epsilon > 0$. Let the subset

$$\begin{aligned}\mathscr{E}_{\mathbf{x}}(\epsilon; X, a) := \{&\omega_1 \times \omega_2 \in \Omega_{Ka} : |\omega_1(\mathbf{x}) - \hat{u}(\mathbf{x}, \omega_2)| \geq \epsilon, \text{ such that } L_z\omega_1(\mathbf{x}_1) = y_1, \\ &\ldots, L_z\omega_1(\mathbf{x}_N) = y_N, B\omega_1(\mathbf{x}_{N+1}) = y_{N+1}, \ldots, B\omega_1(\mathbf{x}_{N+M}) = y_{N+M} \\ &\text{and } a(\omega_2) = z\},\end{aligned}$$

and

$$m_a := \mathbb{E}(a).$$

Since z represents a linear coefficient of the differential operator L_z and $\overset{*}{\mathsf{K}}_{X,z}$ is always positive semi-definite, $\frac{d^2}{dz^2}\sigma(\mathbf{x}|z) \leq 0$ for all $z \in \mathbb{R}$. This indicates that the variance $\sigma(\mathbf{x}|\cdot) : \mathbb{R} \to \mathbb{R}$ induced by L_z (see Eq. (7)) is a concave function. Thus we can deduce that

$$\begin{aligned}\mathbb{P}_a^{\mu}\left(\mathscr{E}_{\mathbf{x}}(\epsilon; X, a)\right) &= \mathbb{P}_a^{\mu}\left(|S_{\mathbf{x}} - \hat{u}(\mathbf{x})| \geq \epsilon \text{ such that } \mathbf{S}_{X,a} = \mathbf{y}_0\right) \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}^{N+M}} \int_{|v - \hat{u}(\mathbf{x},a)| \geq \epsilon} p_{\mathbf{x}}^{\mu}(v|\mathbf{v}, z)\mathrm{d}v\delta_{\mathbf{y}_0}(\mathrm{d}\mathbf{v})\mathrm{d}F_a(z) \\ &= \int_{\mathbb{R}} \operatorname{erfc}\left(\frac{\epsilon}{\sqrt{2}\sigma(\mathbf{x}|z)}\right)\mathrm{d}F_a(z) \\ &\leq \int_{\mathbb{R}} \frac{\sqrt{2}\sigma(\mathbf{x}|z)}{\epsilon}\mathrm{d}F_a(z) \\ &= \frac{\sqrt{2}}{\epsilon}\mathbb{E}\left(\sigma(\mathbf{x}|a)\right) \leq \frac{\sqrt{2}}{\epsilon}\sigma(\mathbf{x}|m_a),\end{aligned}$$

where $\delta_{\mathbf{y}_0}$ is a Dirac delta function at $\mathbf{y}_0$ and erfc is the complementary error function.

The reader may note that the form of the expression for the variance $\sigma(\mathbf{x}|m_a)^2$ (see Eq. (7)) is analogous to that of the *power function* [6, 15], and we can therefore use the same techniques as in the proofs from [5, 6, 15, 16] to obtain a formula for

the order of $\sigma(\mathbf{x}|m_a)$. When L_z is a second-order elliptic differential operator and B is a Dirichlet boundary operator, then we have

$$\sigma(\mathbf{x}|m_a) = \mathscr{O}\left(h_{X,\mathscr{D}}^{m-2-d/2}\right),$$

where

$$h_{X,\mathscr{D}} = \sup_{\mathbf{x}\in\mathscr{D}} \min_{\mathbf{x}_j\in X} \left\|\mathbf{x}-\mathbf{x}_j\right\|_2$$

is the *fill distance* of X for $\mathscr{D}$. This implies that

$$\sup_{\mu\in\mathscr{H}_K(\mathscr{D}),\mathbf{x}\in\mathscr{D}} \mathbb{P}_a^{\mu}\left(\mathscr{E}_{\mathbf{x}}(\epsilon;X,a)\right) = \mathscr{O}\left(\frac{h_{X,\mathscr{D}}^{m-2-d/2}}{\epsilon}\right).$$

Since $|u(\mathbf{x},\omega_2)-\hat{u}(\mathbf{x},\omega_2)| \geq \epsilon$ if and only if $u \in \mathscr{E}_{\mathbf{x}}(\epsilon;X)$, we conclude that

$$\sup_{\mu\in\mathscr{H}_K(\mathscr{D})} \mathbb{P}_a^{\mu}\left(\|u-\hat{u}\|_{\mathrm{L}_\infty(\mathscr{D})} \geq \epsilon\right) \leq \sup_{\mu\in\mathscr{H}_K(\mathscr{D}),\mathbf{x}\in\mathscr{D}} \mathbb{P}_a^{\mu}\left(\mathscr{E}_{\mathbf{x}}(\epsilon;X,a)\right) \to 0,$$

when $h_{X,\mathscr{D}} \to 0$. Therefore we say that the estimator $\hat{u}$ converges to the exact solution u of the SPDE (1) in all probabilities $\mathbb{P}_a^{\mu}$ when $h_{X,\mathscr{D}}$ goes to 0.

Remark 3. The error bounds for kernel-based collocation solutions can also be described in terms of the number of collocation points and the dimension of the domain spaces as is typical for the Quasi-Monte Carlo method. In particular, when K is a Gaussian kernel and X satisfies the condition for an optimal sampling scheme as defined in [7, 8], then [7, Theorem 5.2] and [8, Theorem 1] show that the convergence rate of the kernel-based collocation estimator is dimension-independent, i.e.,

$$\sup_{\mu\in\mathscr{H}_K(\mathscr{D})} \mathbb{P}_a^{\mu}\left(\|u-\hat{u}\|_{\mathrm{L}_\infty(\mathscr{D})} \geq \epsilon\right) = \mathscr{O}\left(\frac{(N+M)^{-p}}{\epsilon}\right), \quad \text{for some } p>0,$$

where $N+M$ denotes the combined number of interior and boundary collocation points. We want to emphasize again that the number of collocation points employed in this paper has a different meaning than the number of sample points used as designs in the probability space as discussed in [1, 14]. In our future research, we will try to find other kernel functions to construct kernel-based estimators with dimension-independent errors in terms of the number of collocation points similar as done in [7, 8].

4.2 Random Coefficients as Stochastic Fields

Now we can also generalize the random coefficient a of the SPDE (1) to be a stochastic field defined on the domain $\mathscr{D}$ and the probability space $(\Omega_a, \mathscr{F}_a, \mathbb{P}_a)$.

If the stochastic field a is Gaussian with nonzero mean $m_a : \mathscr{D} \to \mathbb{R}$ and covariance kernel $R_a : \mathscr{D} \times \mathscr{D} \to \mathbb{R}$, then we can also use the kernel-based collocation method to approximate its solution.

Since a is Gaussian with a known correlation structure, we can simulate the value a at the collocation points $X_{\mathscr{D}}$, i.e., $a_{\mathbf{x}_1}, \ldots, a_{\mathbf{x}_N}$ have multi-normal distributions with mean $\mathbf{m}_X$ and covariance matrix R_X, i.e.,

$$(a_{\mathbf{x}_1}, \cdots, a_{\mathbf{x}_N})^T \sim \mathscr{N}(\mathbf{m}_X, \mathsf{R}_X),$$

where $\mathbf{m}_X := (m_a(\mathbf{x}_1), \cdots, m_a(\mathbf{x}_N))^T$ and $\mathsf{R}_X := \left(R_a(\mathbf{x}_j, \mathbf{x}_k)\right)_{j,k=1}^{N,N}$.

Similar to before, the covariance matrix can be rewritten as

$$\overset{*}{\mathsf{K}}_{X,a} := \begin{pmatrix} \left(L_{a_{\mathbf{x}_j},1} L_{a_{\mathbf{x}_k},2} \overset{*}{K}(\mathbf{x}_j, \mathbf{x}_k)\right)_{j,k=1}^{N,N}, & \left(L_{a_{\mathbf{x}_j},1} B_2 \overset{*}{K}(\mathbf{x}_j, \mathbf{x}_{N+k})\right)_{j,k=1}^{N,M} \\ \left(B_1 L_{a_{\mathbf{x}_k},2} \overset{*}{K}(\mathbf{x}_{N+j}, \mathbf{x}_k)\right)_{j,k=1}^{M,N}, & \left(B_1 B_2 \overset{*}{K}(\mathbf{x}_{N+j}, \mathbf{x}_{N+k})\right)_{j,k=1}^{M,M} \end{pmatrix}.$$

If $\overset{*}{\mathsf{K}}_{X,a}$ is nonsingular almost surely, then the kernel-based solution is given by

$$\hat{u}(\mathbf{x}) = \sum_{k=1}^{N} c_k L_{a_{\mathbf{x}_k},2} \overset{*}{K}(\mathbf{x}, \mathbf{x}_k) + \sum_{k=1}^{M} c_{N+k} B_2 \overset{*}{K}(\mathbf{x}, \mathbf{x}_{N+k}),$$

and the random coefficients $\mathbf{c} = (c_1, \cdots, c_{N+M})^T$ are obtained by solving the system of random linear equations

$$\overset{*}{\mathsf{K}}_{X,a} \mathbf{c} = \mathbf{y}_0,$$

where $\mathbf{y}_0 := (y_1, \cdots, y_{N+M})^T$ is defined in Eq. (2).

Even when the nonconstant coefficient a is non-Gaussian, we can still use the kernel-based collocation method to set up the estimator. We assume that the joint cumulative distribution function $F_{\mathbf{x}_1,\ldots,\mathbf{x}_N} : \mathbb{R}^n \to [0,1]$ of $a_{\mathbf{x}_1}, \ldots, a_{\mathbf{x}_N}$ is known for any finite set of collocation points $X_{\mathscr{D}} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ in $\mathscr{D}$.

Moreover, we can use the Markov chain rule to generate a realization of the random numbers $a_{\mathbf{x}_1}, \ldots, a_{\mathbf{x}_N}$. When $N = 1$ then we can simulate $a_{\mathbf{x}_1}$ by the inverse transform method, i.e.,

$$a_{\mathbf{x}_1} \sim F_{\mathbf{x}_1}^{-1}(U_1), \quad U_1 \sim \text{Unif}[0,1].$$

When $N > 1$ we can simulate the random numbers inductively. Suppose that we have already got the random numbers $a_{\mathbf{x}_1}, \ldots, a_{\mathbf{x}_{N-1}}$. Finally, we start to generate the remainder $a_{\mathbf{x}_N}$, i.e.,

$$a_{\mathbf{x}_N} \sim F^{-1}_{\mathbf{x}_N|\mathbf{x}_1,\ldots,\mathbf{x}_{N-1}}(U_N|a_{\mathbf{x}_1}, \ldots, a_{\mathbf{x}_{N-1}}), \quad U_N \sim \text{Unif}[0,1],$$

where the conditional probability distribution

$$F_{\mathbf{x}_N|\mathbf{x}_1,\ldots,\mathbf{x}_{N-1}}(z_N|z_1,\ldots,z_{N-1}) := \frac{\frac{\partial^{N-1}}{\partial z_1 \cdots \partial z_{N-1}} F_{\mathbf{x}_1,\ldots,\mathbf{x}_N}(z_1,\ldots,z_{N-1},z_N)}{\frac{\partial^{N-1}}{\partial z_1 \cdots \partial z_{N-1}} F_{\mathbf{x}_1,\ldots,\mathbf{x}_{N-1}}(z_1,\ldots,z_{N-1})}.$$

Thus we can obtain the kernel-based collocation solution in the same manner as in the Gaussian cases. We plan to include a convergence analysis for this approach in a future research paper.

5 Numerical Examples

In this section we present a simple numerical experiment. Let the domain $\mathscr{D} := (0,1)^2 \subset \mathbb{R}^2$ and let the scalar random variable a have a Gamma distribution with $\lambda_1 > 0$ and $\lambda_2 > 0$, i.e.,

$$F_a(z) = \frac{\gamma\left(\lambda_1, \lambda_2^{-1} z\right)}{\Gamma(\lambda_1)},$$

where $\gamma(k_1, k_2)$ is the lower incomplete gamma function and $\Gamma(k)$ is the standard gamma function. We use the deterministic functions

$$f(\mathbf{x}) := \exp\left(\sin(\pi x_1) + \cos(\pi x_1) + \sin(\pi x_2) + \cos(\pi x_2)\right), \quad \mathbf{x} = (x_1, x_2) \in \mathscr{D},$$

and

$$g(\mathbf{x}) := \begin{cases} \sin(2\pi x_1), & 0 < x_1 < 1,\ x_2 = 0, \\ -\sin(2\pi x_2), & x_1 = 1,\ 0 < x_2 < 1, \\ 0, & \text{otherwise on } \partial\mathscr{D}, \end{cases}$$

and the random coefficient $\mathbf{a} = a$ to set up a stochastic Poisson equation with Dirichlet boundary condition, i.e.,

$$\begin{cases} -a\Delta u = f, & \text{in } \mathscr{D}, \\ u = g, & \text{on } \partial\mathscr{D}, \end{cases} \tag{10}$$

where $\Delta := \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2}$ is the Laplace differential operator.

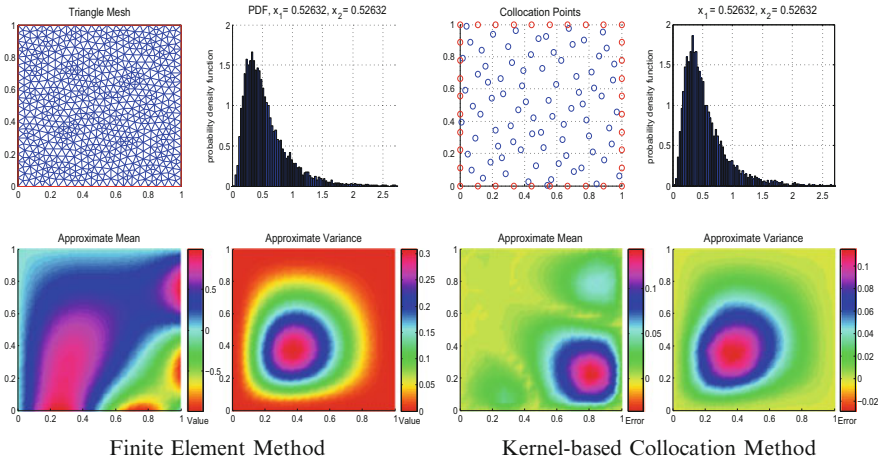


Fig. 1 Numerical Experiments of the probability distributions for the SPDE (10) with $\lambda_1 = 10$ and $\lambda_2 = 10$. *Left*: the stochastic Galerkin finite element methods with maximum mesh parameter $h = 0.05$ of the triangulation. *Right*: the kernel-based collocation method using fill distance $h_{X,\mathscr{D}} = 0.0965$ ($X_{\mathscr{D}}$-Halton points of $N = 81$ and $X_{\partial\mathscr{D}}$-uniform grid points of $M = 36$), and shape parameter $\theta = 2$. Error: relative point-wise absolute error.

We firstly use the stochastic Galerkin finite element method to compute a benchmark solution of the SPDE (10) on a triangulation with a small maximum mesh parameter h. This solution will serve as a stand-in for the exact solution of the SPDE (10) in this section (see the left-hand side of Fig. 1).

For the numerical experiments, we approximate the mean and variance of the arbitrary random variable V by its sample mean and sample variance based on $n_s := 10{,}000$ simulated sample paths, i.e.,

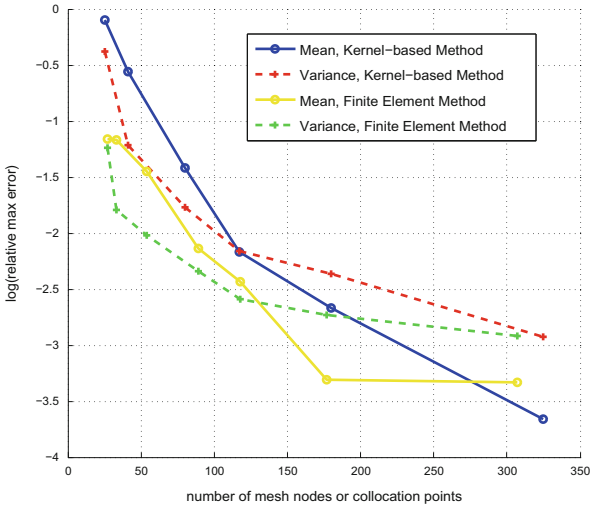
$$\mathbb{E}(V) \approx \frac{1}{n_s}\sum_{k=1}^{n_s} V(\omega_k), \quad \mathrm{Var}(V) \approx \frac{1}{n_s}\sum_{k=1}^{n_s}\left(V(\omega_k) - \frac{1}{n_s}\sum_{j=1}^{n_s} V(\omega_j)\right)^2.$$

For the kernel-based methods, we use the C^4-Matérn function (radial basis function) with shape parameter $\theta > 0$

$$G_\theta(r) := (3 + 3\theta r + \theta^2 r^2)e^{-\theta r}, \quad r > 0,$$

to construct the reproducing kernel (Sobolev-spline kernel)

$$K_\theta(\mathbf{x}, \mathbf{y}) := G_\theta(\|\mathbf{x} - \mathbf{y}\|_2), \quad \mathbf{x}, \mathbf{y} \in \mathscr{D}$$



Kernel-based collocation methods: Sobolev-spline kernel with $m = 3 + \frac{1}{2}$ and $\theta = 2$,

Fig. 2 Convergence of the SPDE (10) with $\lambda_1 = 10$ and $\lambda_2 = 10$ by kernel-based methods and finite element methods.

because $u \in \mathscr{H}_K(\mathscr{D}) \cong \mathscr{H}^{3+1/2}(\mathscr{D})$ (see Sect. 2). Then we can compute the integral-type $\overset{*}{K}_\theta(\mathbf{x}, \mathbf{y}) = \int_0^1 \int_0^1 K_\theta(\mathbf{x}, \mathbf{z}) K_\theta(\mathbf{y}, \mathbf{z}) \mathrm{d}z_1 \mathrm{d}z_2$. To fix our choice of collocation points we choose Halton points in $\mathscr{D}$ and uniform grid points on $\partial\mathscr{D}$. However, this choice is rather arbitrary and we do not claim that it is a particularly good choice. We can also simulate the random coefficient a by

$$a \sim F_a^{-1}(U), \quad U \sim \mathrm{Unif}[0, 1].$$

Using the kernel-based collocation method, we can obtain the approximation $\hat{u}$ via Eq. (9).

According to the numerical results in the right-hand side of Fig. 1, the kernel-based collocation method is well-behaved for the approximate probability distributions. Figure 2 shows that the approximate mean and the approximate variance are convergent as the fill distance $h_{X,\mathscr{D}}$ is refined (see Sect. 4.1.2). According to Fig. 2, we find that the convergence of the kernel-based method we used here seems to be comparable to that of the finite element method.

Remark 4. If the random part of the SPDE is given on the right-hand side as in [5, 11, 16], then the covariance matrix (collocation matrix) is deterministic and we only compute its inverse once to get the kernel-based solution at any event from the sample space. The resulting method is more efficient than the left-hand-side case discussed in this paper.

References

1. Babuška, I., Nobile, F., Tempone, R.: A stochastic collocation method for elliptic partial differential equations with random input data. SIAM Rev. **52**, 317–355 (2010)
2. Babuška, I., Tempone, R., Zouraris, G.E.: Galerkin finite element approximations of stochastic elliptic partial differential equations. SIAM J. Numer. Anal. **42**, 800–825 (2004)
3. Berlinet, A., Thomas-Agnan, C.: Reproducing Kernel Hilbert Spaces in Probability and Statistics. Academic Publishers, Boston, MA (2004)
4. Chow, P-L.: Stochastic Partial Differential Equations, Chapman & Hall/CRC, Taylor & Francis, Boca Raton (2007)
5. Cialenco, I., Fasshauer, G.E., Ye, Q.: Approximation of stochastic partial differential equations by a kernel-based collocation method. Int. J. Comput. Math. **89**, 2543–2561 (2012)
6. Fasshauer, G.E.: Meshfree Approximation Methods with MATLAB, World Scientific, Hackensack (2007)
7. Fasshauer, G.E., Hickernell, F.J., Woźniakowki, H.: On dimension-independent rates of convergence for function approximation with Gaussian kernels. SIAM J. Numer. Anal. **50**, 247–271 (2012)
8. Fasshauer, G.E., Hickernell, F.J., Woźniakowki, H.: Average case approximation: convergence and tractability of Gaussian kernels. In: Plaskota, L., Woźniakowski (ed.) Monte Carlo and Quasi-Monte Carlo Methods 2010, pp. 329–344. Springer, Berlin/Heidelberg (2012)
9. Fasshauer, G.E., Ye, Q.: Reproducing kernels of generalized Sobolev spaces via a green function approach with distributional operators. Numer. Math. **119**, 585–611 (2011)
10. Fasshauer, G.E., Ye, Q.: Reproducing kernels of Sobolev spaces via a green kernel approach with differential operators and boundary operators. Adv. Comput. Math. **38**, 891–921 (2013)
11. Fasshauer, G.E., Ye, Q.: Kernel-based collocation methods versus Galerkin finite element methods for approximating elliptic stochastic partial differential equations. In: Schweitzer, M.A. (ed.) Meshfree Methods for Partial Differential Equations VI, pp. 155–170. Springer, Heidelberg (2013)
12. Glasserman, P.: Monte Carlo Methods in Financial Engineering. Springer, New York (2004)
13. Graham, I.G., Kuo, F.Y., Nuyens, D., Scheichl, R., Sloan, I.H.: Quasi-Monte Carlo methods for elliptic PDEs with random coefficients and applications. J. Comput. Phys. **230**, 3668–3694 (2011)
14. Kuo, F.Y., Schwab, C., Sloan, I.H.: Quasi-Monte Carlo finite element methods for a class of elliptic partial differential equations with random coefficient. SIAM J. Numer. Anal. **50**, 3351–3374 (2012)
15. Wendland, H.: Scattered Data Approximation. Cambridge University Press, Cambridge/New York (2005)
16. Ye, Q.: Analyzing reproducing kernel approximation methods via a Green function approach. Ph.D. thesis, Illinois Institute of Technology (2012)

Polynomial Accelerated MCMC and Other Sampling Algorithms Inspired by Computational Optimization

Colin Fox

Abstract Polynomial acceleration methods from computational optimization can be applied to accelerating MCMC. For example, a geometrically convergent MCMC may be accelerated to be a perfect sampler in special circumstances. An equivalence between Gibbs sampling of Gaussian distributions and classical iterative methods can be established using matrix splittings, allowing direct application of Chebyshev acceleration. The conjugate gradient method can also be adapted to give an accelerated sampler for Gaussian distributions, that is perfect in exact arithmetic.

1 Introduction

Standard Markov chain Monte Carlo (MCMC) algorithms simulate a *homogeneous* Markov chain by performing a stationary linear iteration on the space of probability distributions. The repeated application of a *fixed* kernel results in geometric convergence of the Markov chain, just as it does for the *stationary* linear iterative solvers used to solve systems of linear equations. Stationary linear solvers were state-of-the-art in the 1950s, but are now considered very slow precisely because they are geometrically convergent.

In this paper, methods for accelerating stationary linear iterations developed in the field of numerical computation are applied to accelerating MCMC, both in the general setting of a Markov chain designed to target an arbitrary distribution π (Sect. 2), and also in the specific setting of Gibbs sampling from the multivariate Gaussian distribution $\mathrm{N}\left(0, \mathbf{A}^{-1}\right)$ with known precision matrix $\mathbf{A}$ (Sects. 4 and 5). We will see that polynomial acceleration of a geometrically convergent MCMC can, in certain cases, generate perfect samples in finite time.

C. Fox (✉)
Physics Department, University of Otago, PO Box 56, Dunedin 9054, New Zealand
e-mail: fox@physics.otago.ac.nz

J. Dick et al. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2012*,
Springer Proceedings in Mathematics & Statistics 65,
DOI 10.1007/978-3-642-41095-6_15, © Springer-Verlag Berlin Heidelberg 2013

The special case of Gibbs sampling applied to Gaussian distributions is precisely equivalent to classical iterative methods for solving linear systems understood in terms of matrix splittings, as shown in Sect. 3. Chebyshev acceleration, which is optimal in a certain sense for matrix-splitting methods, can therefore be used to optimally accelerate the Gibbs sampler, as demonstrated in Sect. 4. The conjugate gradient optimization algorithm may also be viewed as a polynomial acceleration in which the eigenvalues of the iteration operator are estimated within the iteration. A 'conjugate gradient sampler' for Gaussian distributions is presented in Sect. 5.

This work takes place within our ongoing efforts in computational (Bayesian) inference that utilizes sampling methods, specifically MCMC. In these problems one wishes to evaluate expectations with respect to a given (posterior) target distribution π over a typically high-dimensional state space. Since the statistics over π are analytically intractable, the best current technology is Monte Carlo integration with importance sampling using samples drawn from π via a random-walk MCMC. That can be very slow. By identifying sampling with optimization, at mathematical and algorithmic levels, we look to adapt the sophisticated methods developed for accelerating computational optimization to computational sampling.

We were also curious about Gibbs sampling being referred to as "stochastic relaxation" in [11], and whether this was related to the "relaxation" methods of numerical analysis in an intuitive sense or in a more formal mathematical sense.

Throughout this paper it is taken as understood that the tasks of computational optimization and solution of systems of equations are equivalent; the normal equations for the optimization form the system to be solved. The terms *solve* and *optimize* are used interchangeably.

2 Polynomial Acceleration of MCMC

This section provides a cartoon of polynomial acceleration of distributional convergence in standard MCMC, to convey the ideas behind polynomial acceleration that can get hidden in a more formal presentation. The weighted-subsampling scheme in Sect. 2.2 does not necessarily lead to a practical technique, but does show the remarkable speedup possible.

2.1 Errors and Convergence in Standard MCMC

The algorithmic mainstay of MCMC methods is the simulation of a homogeneous Markov chain $\{X_0, X_1, \ldots\}$ that tends to some desired target distribution π. The chain is homogeneous because the Markov chain is constructed by repeatedly simulating a *fixed* transition kernel $\mathscr{P}$ constructed so that π is invariant, i.e.,

$$\pi \mathscr{P} = \pi,$$

typically using Metropolis-Hastings (MH) dynamics that ensures that $\mathscr{P}$ and π are in detailed balance.

When the chain is initialized with $X_0 \sim \pi^{(0)}$, the n-step distribution (over X_n) is

$$\pi^{(n)} = \pi^{(n-1)}\mathscr{P} = \pi^{(0)}\mathscr{P}^n.$$

The difference between this distribution and the target distribution π,

$$\pi^{(n)} - \pi = \left(\pi^{(0)} - \pi\right)\mathscr{P}^n, \tag{1}$$

is called the n-step distribution error. Note how the magnitude of the error goes to zero according to the initial distribution error multiplied by the polynomial $\mathscr{P}^n$ of the transition kernel.

All iteration schemes lead to a n-step distribution error of this form, i.e. the initial error multiplied by an n-th order polynomial P_n of the transition kernel. In numerical analysis it is usual to write this *error polynomial* as a polynomial in $I - \mathscr{P}$. Hence the error polynomial in this case is

$$P_n(I - \mathscr{P}) = \mathscr{P}^n = (I - (I - \mathscr{P}))^n \quad \text{or} \quad P_n(\lambda) = (1 - \lambda)^n. \tag{2}$$

All error polynomials satisfy $P_n(0) = 1$, since $\mathscr{P} = I$ leaves the iterate (and error) unchanged. This error polynomial has only one (repeated) zero at $\lambda = 1$.

The second form in Eq. (2) emphasizes that the error polynomial may be evaluated over the eigenvalues of $I - \mathscr{P}$. Since $\mathscr{P}$ is a stochastic kernel, all eigenvalues of $I - \mathscr{P}$ are contained in $[0, 2]$. The error tends to zero when the eigenvalues of $I - \mathscr{P}$ in directions other than π are bounded away from 0 and 2, as is guaranteed by standard results for a *convergent* MCMC.

Thus, a homogeneous MCMC produces a sample correctly distributed as π either after one step (when all eigenvalues of $I - \mathscr{P}$ in directions other than π equal 1), or in the limit $n \to \infty$ (when any eigenvalue in a direction other than π is not 1). In the latter case, the distributional error in Eq. (1) will be dominated by the error in the direction of the eigenvalue of $I - \mathscr{P}$ furthest from 1, λ_*, hence decays as $(1 - \lambda_*)^n$, and the convergence is *geometric*.

2.2 Acceleration by Weighted Subsampling

The key idea in polynomial acceleration is to modify the iteration so that the error polynomial is 'better' than the stationary case in Eqs. (1) and (2), in the sense of smaller error. A simple way to modify the iteration in the setting of MCMC is to subsample with weights. This does not allow complete freedom in choosing the error polynomial, hence there is room for improvement. (Finding an optimal modification is an open problem.) The recipe I will use is: run n steps of a standard MCMC starting at $x^{(0)} \sim \pi^{(0)}$ to produce the realization $\{x^{(1)}, x^{(2)}, \ldots x^{(n)}\}$ and then choose

$x = x^{(i)}$ w.p. (with probabilities) $\{\alpha_i\}_{i=1}^n$ (where $\alpha_i \geq 0$ and $\sum_i \alpha_i = 1$). The resulting sample is distributed as the mixture model

$$x \sim \pi^{(0)} \sum_{i=1}^{n} \alpha_i \mathscr{P}^i$$

with the individual distributions related by increasing powers of $\mathscr{P}$. Weighted subsampling is also considered by Łatuszyński and Roberts [16]. The associated error polynomial is then

$$Q_n = \sum_{i=1}^{n} \alpha_i (1-\lambda)^i$$

which is an n-th order Lorentz polynomial. Since we choose the coefficients $\{\alpha_i\}_{i=1}^n$ we have some freedom in choosing the error polynomial. In special circumstances, it is possible to choose an error polynomial that is zero at the eigenvalues of $I - \mathscr{P}$ other than $\lambda = 0$, in which case subsampling with weights generates a *perfect* sample from π. That is possible, for example, when the sample space is finite, with s states. Then $I - \mathscr{P}$ has at most s distinct eigenvalues and when the $s-1$ eigenvalues other than 0 can be the zeros of a Lorentz polynomial it is possible to choose Q_n to give zero distribution error.

Consider the simple example in which we want to sample from a state-space with $s = 3$ states with target pmf $\pi = (1/3, 1/3, 1/3)$. A Markov chain that targets π can be generated by repeatedly simulating the transition matrix

$$\mathscr{P} = \begin{pmatrix} \frac{1}{48} & \frac{11}{24} & \frac{25}{48} \\ \frac{11}{24} & \frac{1}{12} & \frac{11}{24} \\ \frac{25}{48} & \frac{11}{24} & \frac{1}{48} \end{pmatrix}$$

which can easily be seen to be in detailed balance with π and gives a chain that is irreducible and aperiodic. Note that convergence is geometric, and that

$$\mathscr{P}^2 = \begin{pmatrix} \frac{185}{384} & \frac{55}{192} & \frac{89}{384} \\ \frac{55}{192} & \frac{41}{96} & \frac{55}{192} \\ \frac{89}{384} & \frac{55}{192} & \frac{185}{384} \end{pmatrix}, \mathscr{P}^3 = \begin{pmatrix} \frac{805}{3{,}072} & \frac{539}{1{,}536} & \frac{1{,}189}{3{,}072} \\ \frac{539}{1{,}536} & \frac{229}{768} & \frac{539}{1{,}536} \\ \frac{1{,}189}{3{,}072} & \frac{539}{1{,}536} & \frac{805}{3{,}072} \end{pmatrix}, \ldots, \mathscr{P}^\infty = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$$

so that as $n \to \infty$ the chain converges to a sample from π that is independent of the starting state. This chain can be accelerated by weighted subsampling, as follows:

1. Start with (any) $x^{(0)}$, simulate three steps with $\mathscr{P}$ to get $x^{(1)}, x^{(2)}, x^{(3)}$.
2. Sample x from $(x^{(1)}, x^{(2)}, x^{(3)})$ w.p. $\left(\frac{1}{11} \; \frac{14}{33} \; \frac{16}{33}\right)$.

The resulting x is an exact draw from π, and independent of the starting state, because $\frac{1}{11}\mathscr{P} + \frac{14}{33}\mathscr{P}^2 + \frac{16}{33}\mathscr{P}^3 = \mathscr{P}^\infty$. It is left as an exercise to explicitly construct the error polynomial to see how the example was constructed.

As mentioned above, there are a few practical difficulties with this simple subsampling scheme. An obvious limitation is that the zeros of a Lorentz polynomial only occur for eigenvalues that decorrelate the chain, in which case the polynomial 'acceleration' that draws exact (and i. i. d.) samples actually *increases* the variance in a CLT (see e.g. [16]). However, one might argue that distributional convergence is improved, which may be important in some settings. A further difficulty occurs when n needs to be large since we really want to specify the zeros of the error polynomial yet these are not a stable numerical function of the $\{\alpha_i\}$. Furthermore, n must be chosen in advance which is typically not convenient. All these difficulties may be circumvented in the case of a Gaussian target by using a second-order iteration.

3 Gibbs Sampling of Gaussians is Gauss-Seidel Iteration

The Gibbs sampling algorithm [11] repeatedly samples from (block) conditional distributions of π. We consider the simplest, and original, version of Gibbs sampling in which one iteration consists of conditional sampling along each coordinate direction in sequence, see e.g. Turčin [21], also known as Glauber dynamics [12], the local heat-bath algorithm [5], and the sequential updating method.

3.1 Normal Distributions

We now focus on the case of Gibbs sampling from the multivariate Normal (or Gaussian) distribution $\mathrm{N}\left(0, \mathbf{A}^{-1}\right)$ with known precision matrix $\mathbf{A}$. This situation commonly occurs in (hierarchical) Bayesian analyses when spatial dependencies are modelled via neighbourhood relationships, leading to a Gaussian Markov random field (GMRF) with sparse precision matrix [15]. Both $\mathbf{A}$ and the covariance matrix $\Sigma = \mathbf{A}^{-1}$ are symmetric positive definite. In d dimensions the density function is (written in the *natural parametrization*)

$$\pi(\mathbf{x}) = \sqrt{\frac{\det(\mathbf{A})}{2\pi^d}} \exp\left\{-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} + \mathbf{b}^{\mathrm{T}}\mathbf{x}\right\}. \tag{3}$$

The mean vector $\bar{\mathbf{x}}$ satisfies

$$\mathbf{A}\bar{\mathbf{x}} = \mathbf{b} \tag{4}$$

which gives the first indication that solution of linear equations is relevant to Gaussian distributions.

Cholesky factorization is the preferred method for solving moderately sized linear systems with symmetric and positive definite coefficient matrix, and also for sampling from moderate dimension Gaussian distributions [19] (also called global

heat bath [5]). We are interested in the case where the state-space dimension d is large and $\mathbf{A}$ is sparse. Then, iterative methods such as the Gibbs sampler are attractive as the main cost per iteration is operation by the precision matrix $\mathbf{A}$, which is cheap, and memory requirements are low.

The Gibbs sampler updates components via the conditional distributions, which are also Gaussian. Hence choosing $\pi^{(0)}$ to be Gaussian results in a sequence of Gaussian n-step distributions. Since these n-step distributions converge to π, the sequence of n-step covariance matrices converge to Σ, i.e., $\Sigma^{(n)} \to \Sigma$. One of the motivations for this work was to understand what decomposition of the matrix Σ this sequence is effectively performing. Many matrix decompositions are known in numerical analysis and we were curious to see if Gibbs sampling was effectively performing one of them.

3.2 Matrix Formulation of Gibbs Sampling From $\mathbf{N(0, A^{-1})}$

Let $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ denote the state of the Gibbs sampler. Componentwise Gibbs updates each component in sequence from the (normal) conditional distributions. One 'sweep' over all n components can be written [14]

$$\mathbf{y}^{(k+1)} = -\mathbf{D}^{-1}\mathbf{L}\mathbf{y}^{(k+1)} - \mathbf{D}^{-1}\mathbf{L}^T\mathbf{y}^{(k)} + \mathbf{D}^{-1/2}\mathbf{z}^{(k)} \tag{5}$$

where $\mathbf{D} = \text{diag}(\mathbf{A})$, $\mathbf{L}$ is the strictly lower triangular part of $\mathbf{A}$, and $\mathbf{z}^{(k-1)} \sim \text{N}(\mathbf{0}, \mathbf{I})$. Since $\mathbf{D}$ is invertible, the iteration can be written as the stochastic AR(1) process

$$\mathbf{y}^{(k+1)} = \mathbf{G}\mathbf{y}^{(k)} + \mathbf{c}^{(k)}$$

where $\mathbf{c}^{(k)}$ are i. i. d. draws from a 'noise' distribution with zero mean and finite covariance.

3.3 Matrix Splitting Form of Stationary Iterative Methods

Since about 1965, the *matrix splitting* formalism has been the standard for formulating and understanding the classical iteration schemes used to solve linear systems of equations, as in Eq. (4). The *splitting* $\mathbf{A} = \mathbf{M} - \mathbf{N}$ converts the linear system to $\mathbf{Mx} = \mathbf{Nx} + \mathbf{b}$. When $\mathbf{M}$ is invertible, this may be written

$$\mathbf{x} = \mathbf{M}^{-1}\mathbf{N}\mathbf{x} + \mathbf{M}^{-1}\mathbf{b}.$$

Classical iterative methods compute successive approximations to the solution by repeatedly applying the iteration

$$\begin{aligned}\mathbf{x}^{(k+1)} &= \mathbf{M}^{-1}\mathbf{N}\mathbf{x}^{(k)} + \mathbf{M}^{-1}\mathbf{b} \\ &= \mathbf{G}\mathbf{x}^{(k)} + \mathbf{g}.\end{aligned}$$

The iteration is *convergent* if the sequence of iterates converge for any $\mathbf{x}^{(0)}$.

Many splittings use terms in $\mathbf{A} = \mathbf{L} + \mathbf{D} + \mathbf{U}$ where $\mathbf{L}$ is the strictly lower triangular part of $\mathbf{A}$, $\mathbf{D}$ is the diagonal of $\mathbf{A}$, and $\mathbf{U}$ is the strictly upper triangular part of $\mathbf{A}$. For example, Gauss-Seidel iteration, that sequentially solves for each component using the most recent values, corresponds to the splitting $\mathbf{M} = \mathbf{L} + \mathbf{D}$. The resulting iteration for a sweep over all components in sequence is

$$\mathbf{x}^{(k+1)} = -\mathbf{D}^{-1}\mathbf{L}\mathbf{x}^{(k+1)} - \mathbf{D}^{-1}\mathbf{L}^{\mathrm{T}}\mathbf{x}^{(k)} + \mathbf{D}^{-1}\mathbf{b}. \tag{6}$$

The similarity between Gauss-Seidel iteration in Eq. (6) and the matrix formulation of Gibbs sampling in Eq. (5) is obvious. The only difference is that whereas in each iteration of Gauss-Seidel the constant vector $\mathbf{D}^{-1}\mathbf{b}$ is added, in Gibbs sampling the i. i. d. random vector $\mathbf{D}^{-1/2}\mathbf{z}^{(k)}$ is added. This equivalence has been known for some time; it was explicitly stated in Amit and Grenander [2] and is implicit in Adler [1].

3.4 Matrix Splittings Give Generalized Gibbs Samplers

The standard Gibbs sampler in Eq. (6) and Gauss-Seidel iteration in Eq. (5) are equivalent in the sense that they correspond to the same splitting of the precision matrix. In fact any splitting of the precision matrix leads to a (generalized) Gibbs sampler for $\mathrm{N}(0, \mathbf{A}^{-1})$. What makes this equivalence interesting and useful is that the generalized Gibbs sampler converges (in distribution) if and only if the stationary linear iteration converges (in value); hence convergent Gibbs samplers are equivalent to convergent matrix splittings. The following theorem formalizes this statement.

Theorem 1. *Let* $\mathbf{A} = \mathbf{M} - \mathbf{N}$ *be a splitting with* $\mathbf{M}$ *invertible. The stationary linear solver*

$$\begin{aligned}\mathbf{x}^{(k+1)} &= \mathbf{M}^{-1}\mathbf{N}\mathbf{x}^{(k)} + \mathbf{M}^{-1}\mathbf{b} \\ &= \mathbf{G}\mathbf{x}^{(k)} + \mathbf{M}^{-1}\mathbf{b}\end{aligned} \tag{7}$$

converges, if and only if the random iteration

$$\begin{aligned}\mathbf{y}^{(k+1)} &= \mathbf{M}^{-1}\mathbf{N}\mathbf{y}^{(k)} + \mathbf{M}^{-1}\mathbf{c}^{(k)} \\ &= \mathbf{G}\mathbf{y}^{(k)} + \mathbf{M}^{-1}\mathbf{c}^{(k)}\end{aligned} \tag{8}$$

converges in distribution. Here $\mathbf{c}^{(k)} \overset{\text{iid}}{\sim} \pi_{\text{n}}$ *is any 'noise' distribution that has zero mean and finite variance.*

Proof. (outline) Each converges iff the spectral radius $\varrho(\mathbf{G}) < 1$. □

A complete proof is given in Fox and Parker [9]. (A more general theory allowing $\mathbf{G}$ to be random can be found in [6].) We first saw this result in one direction in Goodman and Sokal [14] and Galli and Gao [10]. Further, it can be shown [9] that the mean converges with asymptotic convergence factor $\varrho(\mathbf{G})$, and covariance with $\varrho(\mathbf{G})^2$ (see also [18]). Thus, the *rate* of convergence is also the same for both the Gibbs sampler and the linear solver derived from a splitting. Hence the optimal solver leads to the optimal Gibbs sampler, and vice versa.

3.5 *Some (Not So Common) Gibbs Samplers for* $\mathbf{N}(\mathbf{0}, \mathbf{A}^{-1})$

There are many matrix splittings known in the numerical analysis community, with conditions for convergence being well established. Most introductory texts on numerical analysis cover the topic of stationary iterative methods and give several classical splittings. Some of these are tabulated in Table 1 with increasing sophistication and (roughly) speed listed from top to bottom. Conditions that guarantee convergence, taken from the numerical analysis literature, are also listed for the case where $\mathbf{A}$ is symmetric positive-definite.

Table 1 Some classical matrix splittings and the derived Gibbs samplers. Conditions for convergence are given in the right-most column, for $\mathbf{A}$ symmetric positive definite. Jacobi iteration converges when $\mathbf{A}$ is strictly diagonally dominant (SDD).

Splitting/sampler	$\mathbf{M}$	$\mathbf{Var}(\mathbf{c}^{(k)}) = \mathbf{M}^T + \mathbf{N}$	Converge if
Richardson	$\frac{1}{\omega}\mathbf{I}$	$\frac{2}{\omega}\mathbf{I} - \mathbf{A}$	$0 < \omega < \frac{2}{\varrho(\mathbf{A})}$
Jacobi	$\mathbf{D}$	$2\mathbf{D} - \mathbf{A}$	$\mathbf{A}$ SDD
GS/Gibbs	$\mathbf{D} + \mathbf{L}$	$\mathbf{D}$	always
SOR/B&F	$\frac{1}{\omega}\mathbf{D} + \mathbf{L}$	$\frac{2-\omega}{\omega}\mathbf{D}$	$0 < \omega < 2$
SSOR/REGS	$\frac{\omega}{2-\omega}\mathbf{M}_{\text{SOR}}\mathbf{D}^{-1}\mathbf{M}_{\text{SOR}}^{\text{T}}$	$\frac{\omega}{2-\omega}\left(\mathbf{M}_{\text{SOR}}\mathbf{D}^{-1}\mathbf{M}_{\text{SOR}}^{T} + \mathbf{N}_{\text{SOR}}^{T}\mathbf{D}^{-1}\mathbf{N}_{\text{SOR}}\right)$	$0 < \omega < 2$

The convenience of a splitting depends on being able to cheaply solve systems of the form $\mathbf{Mu} = \mathbf{r}$ given any vector $\mathbf{r}$. When the splitting is used to generate a Gibbs sampler, as in Eq. (8), it is also necessary to draw realizations of the noise $\mathbf{c}^{(k)} \sim \mathrm{N}(0, \mathbf{M}^T + \mathbf{N})$, so the covariance matrix $\mathbf{M}^T + \mathbf{N}$ needs to have some convenient form.

It is interesting to note that the simplest splittings – Richardson and Jacobi – give simple stationary iterative solvers because it is cheap to operate by $\mathbf{M}^{-1}$ in these cases. However, the required noise covariance matrix is not necessarily simple and so these splittings don't give particularly useful Gibbs samplers.

The Gauss-Seidel (GS) splitting, that gives the standard component-wise Gibbs sampler, hits a 'sweet-spot' in terms of simplicity of the required matrix solution and noise sampling problems. The matrix $\mathbf{M}$ is lower-triangular, so operation by $\mathbf{M}^{-1}$ is straightforward by *forward substitution*, while the noise covariance is diagonal which presents a simple sampling problem. It is no surprise, therefore, that the standard Gibbs sampler was the first of these methods to be discovered. We see from the right column in Table 1 that the Gauss-Seidel iteration is unconditionally convergent, hence Theorem 1 guarantees that so is the component-wise Gibbs sampler – but we already knew this from standard convergence results for the Gibbs sampler.

An early method for accelerating the Gauss-Seidel iteration, due to Young and Frankel in 1950, introduces a *relaxation parameter* ω and modifies the iteration to $\mathbf{x}^{(k+1)} = (1-\omega)\,\mathbf{x}^{(k)} + \omega\left(\mathbf{G}\mathbf{x}^{(k)} + \mathbf{M}^{-1}\mathbf{b}\right)$. This *successive over-relaxation* (SOR) method effectively uses the splitting shown on the row labeled SOR in Table 1. It can be shown that the method converges for $0 < \omega < 2$, though finding values of ω that actually increase convergence speed is problem-specific and can be difficult. The equivalent accelerated Gibbs sampler has been discovered a few times: initially by Adler in 1981 [1] in the physics literature, later in the statistics literature by Barone and Frigessi in 1990 [4] who subsequently referred to it (immodestly) as the 'method of Barone and Frigessi', and in Amit and Grenander [2].

A *symmetric* splitting, for which $\mathbf{M}$ and hence $\mathbf{N}$ is symmetric, has the desirable property that the iteration operator $\mathbf{G}$ has real eigenvalues. A simple way to achieve this is to perform a forwards then backwards sweep of SOR giving the *symmetric successive over-relaxation* (SSOR) method introduced by Young [22]. The effective splitting is listed in Table 1. The equivalent Gibbs sampler was introduced by Roberts and Sahu [18] as a reversible kernel produced by a forward then backward sweep of the standard Gibbs sampler, under the title of the REGS sampler. Polynomial acceleration of this sampler is developed in the next section.

4 Polynomial Acceleration of Gibbs Sampling

Sampling from $\mathrm{N}(\bar{\mathbf{x}}, \mathbf{A}^{-1})$, where $\mathbf{A}\bar{\mathbf{x}} = \mathbf{b}$, using the matrix splitting $\mathbf{A} = \mathbf{M} - \mathbf{N}$, with $\mathbf{M}$ invertible, determines the iteration operator $\mathbf{G} = \mathbf{M}^{-1}\mathbf{N}$ and noise distribution $\mathbf{c}^{(k)} \overset{\text{iid}}{\sim} \mathrm{N}(0, \mathbf{M}^T + \mathbf{N})$. One sweep of the resulting Gibbs sampler is the matrix iteration

$$\mathbf{y}^{(k+1)} = \mathbf{G}\mathbf{y}^{(k)} + \mathbf{M}^{-1}(\mathbf{c}^{(k)} + \mathbf{b}) \tag{9}$$

that combines Eqs. (7) and (8) to converge in both mean and covariance.

4.1 A Closer Look at Convergence

Since both the mean and covariance are invariant under the iteration in Eq. (9), the n-step error in the mean is

$$\mathrm{E}\left(\mathbf{y}^{(n)}\right) - \bar{\mathbf{x}} = \mathbf{G}^n \left[\mathrm{E}\left(\mathbf{y}^{(0)}\right) - \bar{\mathbf{x}}\right],$$

and the error in variance is

$$\mathrm{Var}\left(\mathbf{y}^{(n)}\right) - \mathbf{A}^{-1} = \mathbf{G}^n \left[\mathrm{Var}\left(\mathbf{y}^{(0)}\right) - \mathbf{A}^{-1}\right] \mathbf{G}^n.$$

Both these error terms show that the n-step error is the initial error operated on by the n-th order (matrix) polynomial $\mathbf{G}^n$. Hence, the asymptotic average convergence factor is $\varrho(\mathbf{G})$ for the mean, and $\varrho(\mathbf{G})^2$ for the covariance. These results also appear in Roberts and Sahu [18].

Thus, the error polynomial for the iteration is

$$P_n\left(\mathbf{I} - \mathbf{G}\right) = \left(\mathbf{I} - \left(\mathbf{I} - \mathbf{G}\right)\right)^n = \left(\mathbf{I} - \mathbf{M}^{-1}\mathbf{A}\right)^n \quad \text{or} \quad P_n(\lambda) = (1-\lambda)^n$$

which has the same form as in Eq. (2) because this iteration is also stationary, though now the eigenvalues are of the matrix $\mathbf{M}^{-1}\mathbf{A}$.

In particular, the solver and sampler have exactly the same error polynomial. This is a very important observation, since it means that methods for improving the error polynomial of the solver will also improve convergence of the generalized Gibbs sampler. Further, since the solver and sampler have exactly the same asymptotic average convergence factor, the optimal solver will also be the optimal sampler. Thus, the task of finding a fast Gibbs sampler (for Gaussian distributions) is reduced to the task of consulting the numerical linear algebra literature to find a fast linear iterative solver.

4.2 Chebyshev Acceleration

Golub and Varga [13] introduced the splitting

$$\mathbf{A} = \frac{1}{\tau}\mathbf{M} + \left(1 - \frac{1}{\tau}\right)\mathbf{M} - \mathbf{N},$$

with parameter τ, that the gives the iteration operator

$$\mathbf{G}_\tau = \left(\mathbf{I} - \tau\mathbf{M}^{-1}\mathbf{A}\right). \tag{10}$$

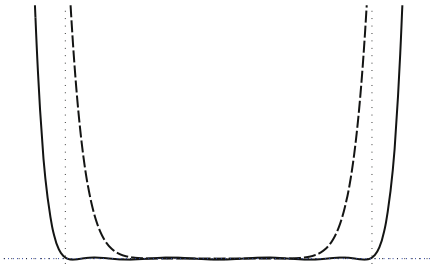


Fig. 1 The default error polynomial (*dashed*) and Chebyshev error polynomial (*solid*) after 10 iterations. *Vertical dotted lines* show the minimum and maximum eigenvalues of $\mathbf{M}^{-1}\mathbf{A}$.

Repeated iteration using this splitting results in the error polynomial $P_n(\lambda) = (1-\tau\lambda)^n$, while n iterations using the *sequence* of parameters $\tau_1, \tau_2, \ldots, \tau_n$ results in the error polynomial

$$P_n(\lambda) = \prod_{l=1}^{n} (1-\tau_l\lambda).$$

Note that the zeros of P_n can be chosen; they are just $1/\tau_1, 1/\tau_2, \ldots, 1/\tau_n$. The resulting iteration is non-stationary (because the iteration operator changes each iteration), hence the derived Gibbs sampler simulates a non-homogeneous Markov chain.

When estimates of the extreme eigenvalues $\lambda_{\min}$ and $\lambda_{\max}$ of $\mathbf{M}^{-1}\mathbf{A}$ are available ($\lambda_{\min}$ and $\lambda_{\max}$ are real when $\mathbf{M}$ is symmetric), the error polynomial may be chosen to be optimal in the sense that it has minimum maximum modulus over the interval $[\lambda_{\min}, \lambda_{\max}]$. The solution is the well-known scaled Chebyshev polynomial with zeros

$$\frac{1}{\tau_l} = \frac{\lambda_{\max}+\lambda_{\min}}{2} + \frac{\lambda_{\max}-\lambda_{\min}}{2}\cos\left(\pi\frac{2l+1}{2n}\right) \quad l = 0, 1, 2, \ldots, n-1. \quad (11)$$

The potential improvement in rate of convergence achievable by the Chebyshev error polynomial is shown in Fig. 1 that shows the Chebyshev (solid) and default (dashed) error polynomials for a random covariance over $d = 10$ variables, after $n = 10$ iterations.

The largest value of the default error polynomial occurs at the extreme eigenvalues of $\mathbf{M}^{-1}\mathbf{A}$, as we expect from standard MCMC convergence theory. The Chebyshev polynomial achieves a much lower maximum value over the interval, at the expense of some 'ripple' in the interval that is of no consequence for convergence. In this case the Chebyshev acceleration gives a factor of 300 improvement in convergence, i.e. the distribution error is 300 times smaller, after just 10 iterations.

An explicit calculation of the maximum of the scaled Chebyshev polynomial over the interval $[\lambda_{\min}, \lambda_{\max}]$ shows that the asymptotic average reduction factor (see e.g. Axelsson [3]) is

$$\sigma = \frac{1-\sqrt{\lambda_{\min}/\lambda_{\max}}}{1+\sqrt{\lambda_{\min}/\lambda_{\max}}},$$

and that this is necessarily better (smaller) than the per-iteration error reduction factor of the un-accelerated iteration.

4.3 Second-Order Accelerated Sampler

The first-order polynomial-accelerated iteration turns out to be numerically unstable, because the iteration operators in Eq. (10) may have spectral radius greater than 1, and also suffers from having to choose the number of iterations n in advance. Numerical stability, and optimality at each step, is given by the second-order iteration [3]

$$\mathbf{y}^{(k+1)} = (1-\alpha_k)\mathbf{y}^{(k-1)} + \alpha_k \mathbf{y}^{(k)} + \alpha_k \tau_k \mathbf{M}^{-1}(\mathbf{c}^{(k)} - \mathbf{A}\mathbf{y}^{(k)}) \tag{12}$$

with α_k and τ_k chosen so the error polynomial satisfies the Chebyshev recursion.

Theorem 2. *If* $\{\alpha_k\}$ *and* $\{\tau_k\}$ *are such that the* 2nd*-order solver converges, then the* 2nd*-order sampler in Eq. (12) converges. Further, the error polynomial is optimal, at each step, for both mean and covariance.*

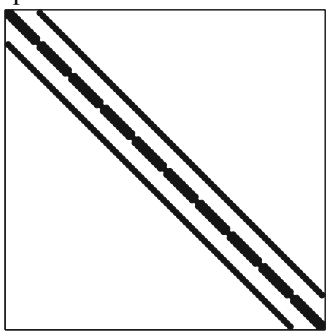
A proof of this theorem and details of a practical second-order Chebyshev accelerated Gibbs sampling algorithm are given in Fox and Parker [8].

4.3.1 An Example with $d = 10 \times 10$

Consider the locally-linear Gaussian distribution defined by the precision matrix [15]

$$[\mathbf{A}]_{ij} = 10^{-4}\delta_{ij} + \begin{cases} n_i & \text{if } i = j, \\ -1 & \text{if } i \neq j \text{ and } ||s_i - s_j||_2 \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

We compute an example on the square 10×10 lattice, so the problem dimension is $d = 100$. The precision matrix inherits the neighbourhood structure of the lattice, so is sparse, with non-zero pattern:



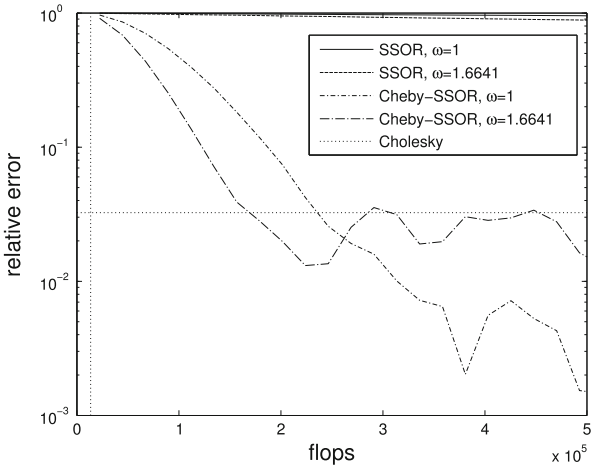


Fig. 2 Convergence of n-step covariance as a function of computational work, for plain and accelerated Gibbs samplers applied to $d = 100$ dimensional problem. The work and error for the Cholesky factorization is shown as *dotted lines*, for reference.

The convergence in n-step covariance of various Gibbs samplers applied to this distribution is shown in Fig. 2. The dashed line shows the SSOR (or REGS) sampler using the optimal SOR parameter of $\omega = 1.6641$. The solid curve shows the standard REGS (forward and backward sweep of Gibbs) sampler ($\omega = 1$). Dash-dot lines show the Chebyshev accelerated SSOR sampler. It is clear that the Chebyshev accelerated sampler is considerably faster than standard Gibbs sampling, in this case $\approx 10^4$ times faster. The dotted lines in Fig. 2 show the work and error for a sample drawn using the Cholesky factorization of $\mathbf{A}$, and confirm that Cholesky factoring is the method of choice for moderately-sized problems.

4.3.2 An Example with $d = 10^6$

Figure 3 shows a sample from a locally linear Gaussian random field, with the same definition of the precision matrix as the previous example, on the 3-dimensional lattice with $d = 100 \times 100 \times 100$, computed using the Chebyshev accelerated SSOR sampler. This problem has $d = 10^6$ which is much larger than could be calculated using a Cholesky factorization. However, the iterative structure of the Gibbs sampler is able to take advantage of the sparse precision matrix, which is the only special structure exploited here. (The Fourier transform is also applicable in this case because the GMRF is stationary.)

5 A Conjugate Gradient Sampling Algorithm

The conjugate gradient (CG) optimization method may be viewed as a polynomial acceleration in which the optimal error polynomial is chosen by also calculating the eigenvalues of the iteration operator within the procedure. However, we present the

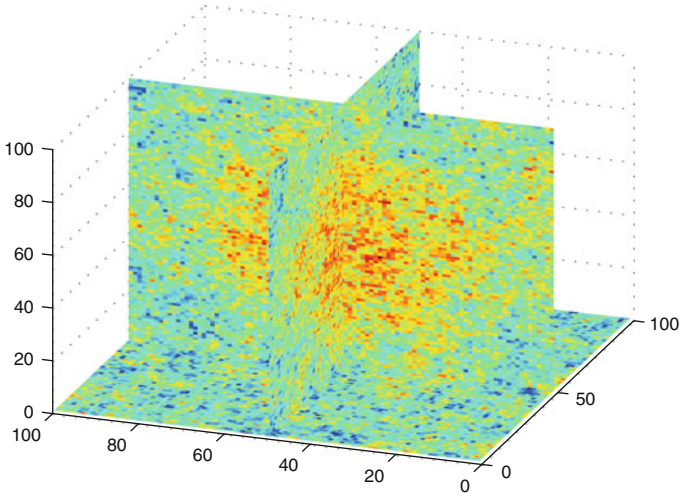


Fig. 3 Slices through a sample on the 3-dimensional lattice with $d = 100 \times 100 \times 100$.

method here by focusing on the mutually **A**-conjugate directions that are generated at each iteration.

Figure 4 shows a schematic of the iterative structure implemented by Gauss-Seidel (left) and conjugate gradient optimization (right) of a quadratic function in $d = 2$ dimensions. The sequence of *search directions* is depicted by dashed lines. The Gauss-Seidel iteration performs optimization along each coordinate direction, in sequence. As we have seen, this implements exactly the same iteration structure as the Gibbs sampler, depicted by solid lines with the sequence of conditional samples denoted $\mathbf{x}^{(0)}$, $\mathbf{x}^{(1)}$, etc. In contrast the CG algorithm uses a sequence of search directions that are mutually **A**-conjugate, seeded by the gradient at each iterate, as depicted in the right panel of Fig. 4. By performing conditional sampling along this sequence of directions, as opposed to 1-dimensional optimization, we get the conjugate gradient sampler (solid lines).

Mutually conjugate vectors (with respect to **A**) are independent directions for $\mathrm{N}(0, \mathbf{A}^{-1})$, since

$$\mathbf{V}^{\mathrm{T}}\mathbf{A}\mathbf{V} = \mathbf{D} \Rightarrow \mathbf{A}^{-1} = \mathbf{V}\mathbf{D}^{-1}\mathbf{V}^{\mathrm{T}}$$

where **V** has mutually conjugate columns and **D** is a diagonal matrix. Hence, if $\mathbf{z} \sim \mathrm{N}(0, I)$ then $\mathbf{x} = \mathbf{V}\sqrt{\mathbf{D}^{-1}}\mathbf{z} \sim \mathrm{N}(0, \mathbf{A}^{-1})$. Thus the problem of sampling from $\mathrm{N}(0, \mathbf{A}^{-1})$ is reduced to sampling from standard normal distributions. Both the Cholesky factorization and eigen-decomposition are examples of sets of mutually conjugate vectors [7].

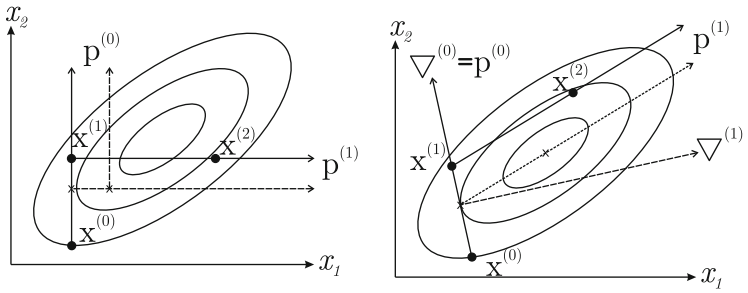


Fig. 4 Schematic in $d = 2$ dimensions depicting the path taken by the Gauss-Seidel iteration and Gibbs sampler (*left*) and the CG optimizer and sampler (*right*). Contours of the quadratic objective function and log-target density are also shown. Path of the optimizer shown in *dotted lines*, sampler shown in *solid lines*. Search directions are $\mathbf{p}^{(0)}, \mathbf{p}^{(1)}, \ldots$, iterates are $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \ldots$, while $\nabla^{(0)}, \nabla^{(1)}, \ldots$ show the direction of gradients at iterates.

Algorithm 1 (CD sampler producing $\mathbf{x} \sim \mathrm{N}(0, \mathbf{A}^{-1})$). *Initialize* $\mathbf{x}$ *and* $\mathbf{b}$ *(*$\mathbf{Ax} \neq \mathbf{b}$*)*

1. $\mathbf{r} \leftarrow \mathbf{b} - \mathbf{Ax}$
2. $\mathbf{p} \leftarrow \mathbf{r}$
3. *for* $k = 1$ *to* n *do:*
4. $\quad \mathbf{q} \leftarrow \mathbf{Ap}$
5. $\quad$ *set* $d \leftarrow \mathbf{q^T p}$, $e \leftarrow \mathbf{q^T x}/d$, $f \leftarrow \mathbf{p^T b}/d$
6. $\quad$ *draw* $z \sim \mathrm{N}(0,1)$ *and set* $\alpha \leftarrow z/\sqrt{d}$
7. $\quad \mathbf{x} \leftarrow \mathbf{x} + (\alpha - e)\,\mathbf{p}$
8. $\quad \mathbf{b} \leftarrow \mathbf{b} + (\alpha - f)\,\mathbf{q}$
9. $\quad \mathbf{r} \leftarrow \mathbf{r} - (f - e)\mathbf{q}$
10. $\quad \mathbf{p} \leftarrow \mathbf{r} - \dfrac{\mathbf{r^T q}}{d}\mathbf{p}$

The sequential conjugate-direction algorithm given in Fox [7] is shown in Algorithm 1. This algorithm operates locally, so can potentially be generalized to non-Gaussian targets. An earlier Krylov-space method was presented in Schneider and Willsky [20]. Ceriotti et al. [5] gave an algorithm that solves $\mathbf{Ax} = \mathbf{b}$ by standard linear CG and separately accumulates the sample $\mathbf{y}$. They mitigated problems associated with loss of conjugacy and degenerate eigenspaces by a combination of random restarts and orthogonalization over a small set of vectors. Parker and Fox [17] presented a convergence criterion based on the residual, also for an algorithm that solves $\mathbf{Ax} = \mathbf{b}$ by standard linear CG and separately accumulates the sample $\mathbf{y}$. They also established that, after k steps, $\mathrm{Var}(\mathbf{y}^k)$ is the CG polynomial, and gave following best-approximation property:

Theorem 3 (Parker 2009). *The covariance matrix*

$$\mathrm{Var}(\mathbf{y}^k | \mathbf{x}^0, \mathbf{b}^0) = V_k T_k^{-1} V_k^T$$

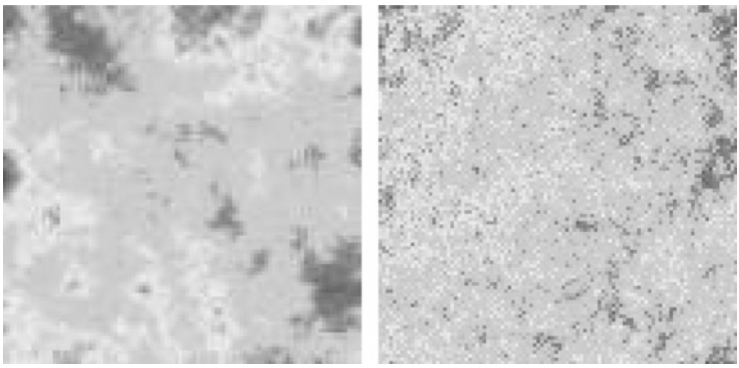


Fig. 5 A CG sample (*left panel*) $\mathbf{x} \sim \mathrm{N}(0, \mathbf{A}^{-1})$ from a 100×100-dimensional Gaussian field with a second-order locally linear precision matrix. The realized variance $\mathrm{Var}(\mathbf{x})$ accounts for 80 % of the variability in $\mathbf{A}^{-1}$. A Cholesky sample is shown on the right panel.

has k non-zero eigenvalues which are the Lanczos estimates of the eigenvalues of $\mathbf{A}^{-1}$. *The eigenvectors of* $\mathrm{Var}(\mathbf{y}^k|\mathbf{x}^0, \mathbf{b}^0)$ *are the Ritz vectors* $V_k v^i$ *which estimate the eigenvectors of* $\mathbf{A}$.

That is, the k-step variance $\mathrm{Var}(\mathbf{y}^k|\mathbf{x}^0, \mathbf{b}^0)$ approximates $\mathbf{A}^{-1}$ in the eigenspaces corresponding to the extreme and well separated eigenvalues of $\mathbf{A}$.

Figure 5 shows two samples drawn from a 100×100-dimensional Gaussian field with a second-order locally linear precision matrix. The left panel was drawn using the CG sampler of Parker and Fox [17], while the right sample was evaluated using the Cholesky factorization of $\mathbf{A}$. Loss of conjugacy in the CG algorithm means that the algorithm terminates before sampling all d-dimensions of the problem. For typical covariance functions, this results in over smooth samples as can be seen in the left panel of Fig. 5. However, the connection with iterative solvers immediately suggests the efficient solution which is to initialize the (accelerated) Gibbs sampler with the CG sample. This plays to the strengths of each method; the CG sampler efficiently calculates smooth structures in the Gaussian field, while relaxation techniques such as Gauss-Seidel (hence Gibbs sampling) are efficient in removing high-frequency errors.

6 Discussion

The motivating query of whether "stochastic relaxation" is formally equivalent to "relaxation" has been answered in the affirmative, in the Gaussian setting; Gibbs sampling is precisely equivalent to Gauss-Seidel iteration. This result generalizes to any splitting of the precision matrix, to give both a "stochastic relaxation" and a "relaxation" with identical conditions for convergence, rates of convergence, and error polynomial.

Hence, existing efficient solvers (multigrid, fast multipole, parallel tools) can all be used to perform sampling from Gaussian distributions; indeed, these 'best' solvers are necessarily the 'best' samplers for Gaussian distributions.

As was shown in Sect. 2, polynomial acceleration may also be applied to the Markov chain that targets a non-Gaussian distribution. The example presented, while rather special and not of practical use, did demonstrate that polynomial acceleration of a geometrically convergent chain can lead to an algorithm that draws 'perfect' samples in finite compute time.

For general target distributions, Chebyshev acceleration of convergence in mean and covariance is also not limited to Gaussian targets. The requirement of explicitly knowing the precision matrix $\mathbf{A}$ may be circumvented by *adapting* to it [8]. Applications in the setting of diffusion tomography show good results, though no proof of convergence exists for the accelerated adaptive algorithm.

Acknowledgements Polynomial acceleration of Gibbs sampling is the brainchild of Al Parker, to whom I am indebted. This research was supported by Marsden contract UOO1015.

References

1. Adler, S.L.: Over-relaxation method for the Monte Carlo evaluation of the partition function for multiquadratic actions. Phys. Rev. D **23**, 2901–2904 (1981)
2. Amit, Y., Grenander, U.: Comparing sweep strategies for stochastic relaxation. J. Multivariate Anal. **37**, 197–222 (1991)
3. Axelsson, O.: Iterative Solution Methods. Cambridge University Press, Cambridge (1996)
4. Barone, P., Frigessi, A.: Improving stochastic relaxation for Gaussian random fields. Probab. Engrg. Inform. Sci. **4**, 369–384 (1990)
5. Ceriotti, M., Bussi, G., Parrinello, M.: Conjugate gradient heat bath for ill-conditioned actions. Phys. Rev. E **76**, 026707-1–8 (2007)
6. Diaconis, P., Freedman, D.: Iterated random functions. SIAM Rev. **41**, 45–76 (1999)
7. Fox, C.: A conjugate direction sampler for normal distributions, with a few computed examples. Technical Reports from the Electronics Group, University of Otago (2008)
8. Fox, C., Parker, A.: Convergence in variance of Chebyshev accelerated Gibbs samplers. SIAM J. Sci. Comput. (2013, to appear)
9. Fox, C., Parker, A.: Gibbs sampling of normal distributions using matrix splittings and polynomial acceleration. (2013, in preparation)
10. Galli, A., Gao, H.: Rate of convergence of the Gibbs sampler in the Gaussian case. Mathematical Geology **33**, 653–677 (2001)
11. Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. **6**, 721–741 (1984)
12. Glauber, R.: Time dependent statistics of the Ising model. J. Math. Phys. **4**, 294–307 (1963)
13. Golub, G.H., Varga, R.S.: Chebyshev semi-iterative methods, successive over-relaxation iterative methods, and second-order Richardson iterative methods, parts I and II. Numer. Math. **3**, 147–156, 157–168 (1961)
14. Goodman, J., Sokal, A.D.: Multigrid Monte Carlo method. Conceptual foundations. Phys. Rev. D **40**, 2035–2071 (1989)
15. Higdon, D.: A primer on space-time modelling from a Bayesian perspective. In: Finkenstadt, B., Held, L., Isham, V. (eds.) Statistics of Spatio-Temporal Systems, pp. 217–279. Chapman & Hall/CRC, New York (2006)

16. Łatuszyński, K., Roberts, G.O.: CLTs and asymptotic variance of time-sampled Markov chains. Methodol. Comput. Appl. Probab. **15**, 237–247 (2013)
17. Parker, A., Fox, C.: Sampling Gaussian distributions in Krylov spaces with conjugate gradients. SIAM J. Sci. Comput. **34**, B312–B334 (2012)
18. Roberts, G.O., Sahu, S.K.: Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. J. R. Stat. Soc. Ser. B. **59**, 291–317 (1997)
19. Rue, H.: Fast sampling of Gaussian Markov random fields. J. R. Stat. Soc. Ser. B. **63**, 325–338 (2001)
20. Schneider, M.K., Willsky, A.S.: A Krylov subspace method for covariance approximation and simulation of random processes and fields. Multidimens. Syst. Signal Process. **14**, 295–318 (2003)
21. Turchin, V.F.: On the computation of multidimensional integrals by the Monte Carlo method. Theory Probab. Appl. **16**, 720–724 (1971)
22. Young, D.M.: Iterative Solution of Large Linear Systems. Academic, New York (1971)

Antithetic Multilevel Monte Carlo Estimation for Multidimensional SDEs

Michael B. Giles and Lukasz Szpruch

Abstract In this paper we develop antithetic multilevel Monte Carlo (MLMC) estimators for multidimensional SDEs driven by Brownian motion. Giles has previously shown that if we combine a numerical approximation with strong order of convergence $O(\Delta t)$ with MLMC we can reduce the computational complexity to estimate expected values of Lipschitz functionals of SDE solutions with a root-mean-square error of ϵ from $O(\epsilon^{-3})$ to $O(\epsilon^{-2})$. However, in general, to obtain a rate of strong convergence higher thnan $O(\Delta t^{1/2})$ requires simulation, or approximation, of Lévy areas. Recently, Giles and Szpruch [5] constructed an antithetic multilevel estimator thnnat avoids thnne simulation of Lévy areas and still achieves an MLMC correction variance which is $O(\Delta t^2)$ for smooth payoffs and almost $O(\Delta t^{3/2})$ for piecewise smooth payoffs, even though there is only $O(\Delta t^{1/2})$ strong convergence. This results in an $O(\epsilon^{-2})$ complexity for estimating the value of financial European and Asian put and call options. In this paper, we extend these results to more complex payoffs based on the path minimum. To achieve this, an approximation of the Lévy areas is needed, resulting in $O(\Delta t^{3/4})$ strong convergence. By modifying the antithetic MLMC estimator we are able to obtain $O(\epsilon^{-2}\log(\epsilon)^2)$ complexity for estimating financial barrier and lookback options.

1 Introduction

In his original MLMC paper [4], Giles showed that one could obtain a good MLMC variance for smooth payoffs by using a numerical approximation with good strong convergence properties. This is in contrast to the standard Monte Carlo approach to simulations of SDEs, where only a good weak order of convergence is required.

M.B. Giles · L. Szpruch
Mathematical Institute, University of Oxford, Oxford, UK
e-mail: mike.giles@maths.ox.ac.uk; szpruch@maths.ox.ac.uk

J. Dick et al. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2012*,
Springer Proceedings in Mathematics & Statistics 65,
DOI 10.1007/978-3-642-41095-6_16, © Springer-Verlag Berlin Heidelberg 2013

For multidimensional SDEs, to obtain good strong convergence, simulation of the Lévy areas is required. Indeed, Clark and Cameron [1] proved for a particular SDE that it is impossible to achieve a better order of strong convergence than the Euler-Maruyama discretisation when using just the discrete increments of the underlying Brownian motion. The analysis was extended by Müller-Gronbach [8] to general SDEs. As a consequence, if we use the standard MLMC method with the Milstein scheme without simulating the Lévy areas the complexity will remain the same as for Euler-Maruyama. Recently, Giles and Szpruch [5] constructed an antithetic MLMC estimator, enabling one to neglect the Lévy areas and still obtain a multilevel correction estimator with a variance which decays at the same rate as the scalar Milstein estimator. They achieved an $O(\Delta t^2)$ MLMC variance for smooth payoffs and almost an $O(\Delta t^{3/2})$ variance for piecewise smooth payoffs, even though there is only $O(\Delta t^{1/2})$ strong convergence. This results in an $O(\epsilon^{-2})$ complexity for estimating the value of European and Asian put and call options.

The question remains whether the approach can be extended to more complex payoffs such as those based on the minimum of the path over the simulation interval. For scalar SDEs with the Milstein discretisation, Giles [4] obtained $O(\epsilon^{-2})$ complexity for such payoffs by combining MLMC with conditional Monte Carlo methods. In this paper, we extend these results to the multidimensional case. Unlike the previous multidimensional work, we find that a suitable approximation to the Lévy areas is required. By a suitable modification of the antithetic MLMC estimator we are able to obtain $O(\epsilon^{-2}\log(\epsilon)^2)$ complexity for payoffs corresponding to financial lookback and barrier options. We focus on simulations of Clark and Cameron's SDE since it captures the essence of simulations requiring Lévy area simulation to obtain higher that $O(\Delta t^{1/2})$ strong convergence property. Our results are supported by numerical experiments.

2 MLMC

Multilevel Monte Carlo simulation uses a number of levels of resolution, $\ell = 0, 1, \ldots, L$, with $\ell = 0$ being the coarsest, and $\ell = L$ being the finest. In the context of an SDE simulation, level 0 may have just one timestep for the whole time interval $[0, T]$, whereas level L might have 2^L uniform timesteps $\Delta t_L = 2^{-L}T$. If P denotes the payoff (or other output functional of interest), and P_ℓ denote its approximation on level ℓ, then the expected value $\mathbb{E}[P_L]$ on the finest level is equal to the expected value $\mathbb{E}[P_0]$ on the coarsest level plus a sum of corrections which give the difference in expectation between simulations on successive levels,

$$\mathbb{E}[P_L] = \mathbb{E}[P_0] + \sum_{\ell=1}^{L} \mathbb{E}[P_\ell - P_{\ell-1}]. \tag{1}$$

Let Y_0 be an estimator for $\mathbb{E}[P_0]$ using N_0 samples, and let Y_ℓ, $\ell > 0$, be an estimator for $\mathbb{E}[P_\ell - P_{\ell-1}]$ using N_ℓ samples. The simplest estimator is a mean of N_ℓ independent samples, which for $\ell > 0$ is

$$Y_\ell = N_\ell^{-1} \sum_{i=1}^{N_\ell} (P_\ell^i - P_{\ell-1}^i). \tag{2}$$

The key point is that $P_\ell^i - P_{\ell-1}^i$ should come from two discrete approximations for the same underlying stochastic sample.

We recall the Theorem from [5]:

Theorem 1. *Let P denote a functional of the solution of a stochastic differential equation, and let P_ℓ denote the corresponding level ℓ numerical approximation. If there exist independent estimators Y_ℓ based on N_ℓ Monte Carlo samples, and positive constants $\alpha, \beta, \gamma, c_1, c_2, c_3$ such that $\alpha \geq \frac{1}{2} \min(\beta, \gamma)$ and*

(i) $|\mathbb{E}[P_\ell - P]| \leq c_1 \, 2^{-\alpha \ell}$

(ii) $\mathbb{E}[Y_\ell] = \begin{cases} \mathbb{E}[P_0], & \ell = 0 \\ \mathbb{E}[P_\ell - P_{\ell-1}], & \ell > 0 \end{cases}$

(iii) $\mathbb{V}[Y_\ell] \leq c_2 \, N_\ell^{-1} 2^{-\beta \ell}$

(iv) $C_\ell \leq c_3 \, N_\ell \, 2^{\gamma \ell}$, *where C_ℓ is the computational complexity of Y_ℓ*

then there exists a positive constant c_4 such that for any $\epsilon < e^{-1}$ there are values L and N_ℓ for which the multilevel estimator $Y = \sum_{\ell=0}^{L} Y_\ell$, has a mean-square-error with bound $MSE \equiv \mathbb{E}\left[(Y - \mathbb{E}[P])^2\right] < \epsilon^2$ with a computational complexity C with bound

$$C \leq \begin{cases} c_4 \, \epsilon^{-2}, & \beta > \gamma, \\ c_4 \, \epsilon^{-2} (\log \epsilon)^2, & \beta = \gamma, \\ c_4 \, \epsilon^{-2-(\gamma-\beta)/\alpha}, & 0 < \beta < \gamma. \end{cases}$$

In (2) we have used the same estimator for the payoff P_ℓ on every level ℓ, and therefore (1) is a trivial identity due to the telescoping summation. However, in [3] Giles explained that it can be better to use different estimators for the finer and coarser of the two levels being considered, P_ℓ^f when level ℓ is the finer level, and P_ℓ^c when level ℓ is the coarser level. In this case, we require that

$$\mathbb{E}[P_\ell^f] = \mathbb{E}[P_\ell^c] \quad \text{for } \ell = 0, \ldots, L-1, \tag{3}$$

so that $\mathbb{E}[P_L^f] = \mathbb{E}[P_0^f] + \sum_{\ell=1}^{L} \mathbb{E}[P_\ell^f - P_{\ell-1}^c]$. The MLMC Theorem is still applicable to this modified estimator. The advantage is that it gives the flexibility to construct approximations for which $P_\ell^f - P_{\ell-1}^c$ is much smaller than the original $P_\ell - P_{\ell-1}$, giving a larger value for β, the rate of variance convergence in condition

(iii) in the theorem. In the next sections we demonstrate how suitable choice of P_ℓ^f and P_ℓ^c can dramatically increase the convergence of the variance of the MLMC estimator.

2.1 Milstein Scheme

Let $(\Omega, \mathscr{F}, \{\mathscr{F}_t\}_{t\geq 0}, \mathbb{P})$ be a complete probability space with a filtration $\{\mathscr{F}_t\}_{t\geq 0}$ satisfying the usual conditions, and let $w(t)$ be a m-dimensional Brownian motion defined on the probability space. We consider the numerical approximation of SDEs of the form

$$\mathrm{d}x(t) = f(x(t))\,\mathrm{d}t + g(x(t))\,\mathrm{d}w(t), \tag{4}$$

where $x(t) \in \mathbb{R}^d$ for each $t \geq 0$, $f \in C^2(\mathbb{R}^d, \mathbb{R}^d)$, $g \in C^2(\mathbb{R}^d, \mathbb{R}^{d\times m})$ have bounded first and second derivatives, and for SIMPLICITY we assume a fixed initial value $x_0 \in \mathbb{R}^d$.

For Lipschitz continuous payoffs that depend on finite number of times $t_n^\ell = n\Delta t_\ell$, the MLMC variance can be estimated from the strong convergence of the numerical scheme, that is

$$\left(\mathbb{E}\Big[\sup_{0\leq n\leq 2^\ell} \big\|x(t_n^\ell) - X_n^\ell\big\|^p\Big]\right)^{1/p} = O(\Delta t_\ell^\xi) \quad \text{for} \quad p \geq 2.$$

For partition $\mathscr{P}_{\Delta t_\ell} := \{n\Delta t_\ell : n = 0, 1, 2, \ldots, 2^\ell = N\}$, where $\Delta t_\ell = T/N$, we consider the Milstein approximation X_n^ℓ with $i\,th$ component of the form

$$\begin{aligned} X_{i,n+1}^\ell =& X_{i,n}^\ell + f_i(X_n^\ell)\,\Delta t_\ell + \sum_{j=1}^m g_{ij}(X_n^\ell)\,\Delta w_{j,n}^\ell \\ &+ \sum_{j,k=1}^m h_{ijk}(X_n^\ell)\left(\Delta w_{j,n}^\ell \Delta w_{k,n}^\ell - \delta_{j,k}\,\Delta t_\ell - [A_{jk}^\ell]_{t_n}^{t_{n+1}}\right) \end{aligned} \tag{5}$$

where $h_{ijk}(x) = \frac{1}{2}\sum_{l=1}^d g_{lk}(x)\,\frac{\partial g_{ij}}{\partial x_l}(x)$, $\delta_{j,k}$ is a Kronecker delta, $\Delta w_n^\ell = w((n+1)\Delta t_\ell) - w(n\Delta t_\ell)$ and $[A_{jk}^\ell]_{t_n}^{t_{n+1}}$ is the Lévy area defined as

$$[A_{jk}^\ell]_{t_n}^{t_{n+1}} = \int_{t_n^\ell}^{t_{n+1}^\ell} \Big(w_j(t) - w_j(t_n^\ell)\Big)\,\mathrm{d}w_k(t) - \int_{t_n^\ell}^{t_{n+1}^\ell} \Big(w_k(t) - w_k(t_n^\ell)\Big)\,\mathrm{d}w_j(t). \tag{6}$$

For the Milstein scheme $\xi = 1$ and therefore $\beta = 2$ for smooth payoffs, and hence MLMC has complexity $O(\varepsilon^{-2})$. However, there is no method for simulating Lévy areas with a cost per timestep similar to that of Brownian increments, apart from

in dimension 2 [2, 9, 10]. Furthermore, within computational finance, options are often based on the continuously-monitored minimum (or maximum) or the path. The Milstein scheme gives an improved rate of convergence at the simulation times, but to maintain the strong order of convergence for such path-dependent options we use Brownian Bridge interpolation within each timestep $[t_n^\ell, t_{n+1}^\ell]$

$$\tilde{X}^\ell(t) = X_n^\ell + \lambda_\ell \, (X_{n+1}^\ell - X_n^\ell) + g(X_n^\ell) \, \left(w(t) - w(t_n^\ell) - \lambda \, \Delta w_n^\ell\right) \tag{7}$$

where $\lambda_\ell \equiv (t - t_n^\ell)/\Delta t_\ell$. Using this interpolant, we have the result [7]

$$\mathbb{E}\Big[\sup_{0\le t\le T} \left\| x(t) - \tilde{X}^\ell(t) \right\|^p\Big] = O(|\, \Delta t_\ell \log(\Delta t_\ell)\, |^p).$$

3 Antithetic MLMC Estimator

The idea for the antithetic estimator is to exploit the flexibility of the more general MLMC estimator by defining $P_{\ell-1}^c$ to be the usual payoff $P(X^c)$ coming from a level $\ell-1$ coarse simulation X^c, and define P_ℓ^f to be the average of the payoffs $P(X^f)$, $P(X^a)$ coming from an antithetic pair of level ℓ simulations, X^f and X^a.

X^f will be defined in a way which corresponds naturally to the construction of X^c. Its antithetic "twin" X^a will be defined so that it has exactly the same distribution as X^f, conditional on X^c, which ensures that $\mathbb{E}[P(X^f)] = \mathbb{E}[P(X^a)]$ and hence (3) is satisfied, but at the same time $\left(X^f - X^c\right) \approx -(X^a - X^c)$ and therefore $\left(P(X^f) - P(X^c)\right) \approx -(P(X^a) - P(X^c))$, so that $\frac{1}{2}\left(P(X^f) + P(X^a)\right) \approx P(X^c)$. This leads to $\frac{1}{2}\left(P(X^f) + P(X^a)\right) - P(X^c)$ having a much smaller variance than the standard estimator $P(X^f) - P(X^c)$. It was proved in [5], that if $\left\| \frac{\partial P}{\partial x} \right\| \le L_1$, $\left\| \frac{\partial^2 P}{\partial x^2} \right\| \le L_2$. then for $p \ge 2$,

$$\begin{aligned}
&\mathbb{E}\left[\left(\tfrac{1}{2}(P(X^f) + P(X^a)) - P(X^c)\right)^p\right] \\
&\quad \le 2^{p-1} L_1^p \, \mathbb{E}\left[\left\| \tfrac{1}{2}(X^f + X^a) - X^c \right\|^p\right] + 2^{-p-1} L_2^p \, \mathbb{E}\left[\left\| X^f - X^a \right\|^{2p}\right].
\end{aligned}$$

In the multidimensional SDE we will show that the Milstein approximation with the Lévy areas set to zero, combined with the antithetic construction, leads to $X^f - X^a = O(\Delta t^{1/2})$ but $\frac{1}{2}(X^f + X^a) - X^c = O(\Delta t)$. Hence, the variance $\mathbb{V}[\frac{1}{2}(P_\ell^f + P_\ell^a) - P_{\ell-1}^c]$ is $O(\Delta t^2)$ for smooth payoffs, which is the same order obtained for scalar SDEs using the Milstein discretisation with its first order strong convergence.

4 Clark-Cameron Example

The Clark and Cameron model problem [1] is

$$\mathrm{d}x_1(t) = \mathrm{d}w_1(t), \quad \mathrm{d}x_2(t) = x_1(t)\,\mathrm{d}w_2(t), \tag{8}$$

with $x_1(0) = x_2(0) = 0$, and zero correlation between the two Brownian motions $w_1(t)$ and $w_2(t)$. These equations can be integrated exactly over a time interval $[t_n, t_{n+1}]$, where $t_n = n\,\Delta t$, to give

$$\begin{aligned} x_1(t_{n+1}) &= x_1(t_n) + \Delta w_{1,n} \\ x_2(t_{n+1}) &= x_2(t_n) + x_1(t_n)\Delta w_{2,n} + \tfrac{1}{2}\Delta w_{1,n}\Delta w_{2,n} + \tfrac{1}{2}[A_{12}]_{t_n}^{t_{n+1}} \end{aligned} \tag{9}$$

where $\Delta w_{i,n} \equiv w_i(t_{n+1}) - w_i(t_n)$, and $[A_{12}]_{t_n}^{t_{n+1}}$ is the Lévy area defined in (6). This corresponds exactly to the Milstein discretisation presented in (5), so for this simple model problem the Milstein discretisation is exact. The point of Clark and Cameron's paper is that for *any* numerical approximation $X(T)$ based solely on the set of discrete Brownian increments Δw, $\mathbb{E}[(x_2(T) - X_2(T))^2] \geq \frac{1}{4}\,T\,\Delta t$. Since in this section we use superscript f, a, c for fine X^f, antithetic X^a and coarse X^c approximations, respectively, we drop the superscript ℓ for the clarity of notation.

We define a coarse path approximation X^c with timestep Δt, and times $t_n \equiv n\,\Delta t$, by neglecting the Lévy area terms to give

$$\begin{aligned} X^c_{1,n+1} &= X^c_{1,n} + \Delta w^{\ell-1}_{1,n} \\ X^c_{2,n+1} &= X^c_{2,n} + X^c_{1,n}\Delta w^{\ell-1}_{2,n} + \tfrac{1}{2}\,\Delta w^{\ell-1}_{1,n}\,\Delta w^{\ell-1}_{2,n} \end{aligned} \tag{10}$$

This is equivalent to replacing the true Brownian path by a piecewise linear approximation as illustrated in Fig. 1. Similarly, we define the corresponding two half-timesteps of the first fine path approximation X^f. Using

$$\begin{aligned} \Delta w^{\ell-1}_{n+1} &\equiv (w(t_{n+1}) - w(t_n)) \\ &= (w(t_{n+1}) - w(t_{n+1/2})) + (w(t_{n+1/2}) - w(t_n)) \;\equiv\; \Delta w^{\ell}_{n+1/2} + \Delta w^{\ell}_{n}, \end{aligned}$$

we can combine two half-timestep approximations to obtain an equation for the increment over the coarse timestep,

$$\begin{aligned} X^f_{1,n+1} &= X^f_{1,n} + \Delta w^{\ell-1}_{1,n} \\ X^f_{2,n+1} &= X^f_{2,n} + X^f_{1,n}\,\Delta w^{\ell-1}_{2,n} + \tfrac{1}{2}\,\Delta w^{\ell-1}_{1,n}\,\Delta w^{\ell-1}_{2,n} \\ &\qquad + \tfrac{1}{2}\left(\Delta w^{\ell}_{1,n}\,\Delta w^{\ell}_{2,n+1/2} - \Delta w^{\ell}_{2,n}\,\Delta w^{\ell}_{1,n+1/2}\right). \end{aligned} \tag{11}$$

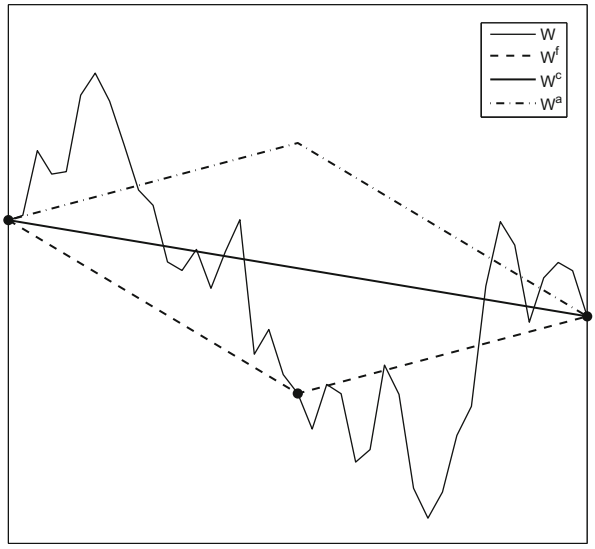


Fig. 1 Brownian path w, its piecewise linear interpolations w^c and w^f, and the antithetic w^a, for a single coarse timestep. The *circles* denote the points at which the Brownian path is sampled.

The antithetic approximation X_n^a is defined by exactly the same discretisation except that the Brownian increments $\Delta w_{n+1/2}^\ell$ and Δw_{n+1}^ℓ are swapped, as illustrated in Fig. 1. This gives

$$\begin{aligned} X_{1,n+1}^a &= X_{1,n}^a + \Delta w_{1,n}^{\ell-1}, \\ X_{2,n+1}^a &= X_{2,n}^a + X_{1,n}^a\,\Delta w_{2,n}^{\ell-1} + \tfrac{1}{2}\,\Delta w_{1,n}^{\ell-1}\,\Delta w_{2,n}^{\ell-1} \\ &\quad - \tfrac{1}{2}\left(\Delta w_{1,n}^\ell\,\Delta w_{2,n+1/2}^\ell - \Delta w_{2,n}^\ell\,\Delta w_{1,n+1/2}^\ell\right). \end{aligned} \tag{12}$$

Swapping Δw_n^ℓ and $\Delta w_{n+1/2}^\ell$ does not change the distribution of the driving Brownian increments, and hence X^a has exactly the same distribution as X^f. Note also the change in sign in the last term in (11) compared to the corresponding term in (12). This is important because these two terms cancel when the two equations are averaged.

In [5] Giles and Szpruch proved the following result:

Lemma 1. *If X_n^f, X_n^a and X_n^c are as defined above, then*

$$X_{1,n}^f = X_{1,n}^a = X_{1,n}^c, \quad \tfrac{1}{2}\left(X_{2,n}^f + X_{2,n}^a\right) = X_{2,n}^c, \quad n = 1, 2, \ldots, N \equiv 2^{\ell-1}.$$

and

$$\mathbb{E}\left[\left|X_{2,N}^f - X_{2,N}^a\right|^p\right] = O(\Delta t^{p/2}) \quad \text{for } p \geq 2.$$

This allows us to prove that for payoffs which are a smooth function of the final state the MLMC variance

$$\mathbb{V}\left[\tfrac{1}{2}\left(P(X_N^f) + P(X_N^a)\right) - P(X_N^c)\right]$$

has an $O(\Delta t^2)$ upper bound and therefore the complexity of the MLMC estimator is $O(\epsilon^2)$. This matches the convergence rate and complexity for the multilevel method for scalar SDEs using the standard first order Milstein discretisation, and is much better than the $O(\Delta t)$ MLMC convergence obtained with the Euler-Maruyama discretisation. Very few financial payoff functions are twice differentiable on the entire domain, but Giles and Szpruch have proved that for piecewise smooth put and call options the variance converges with rate $O(\Delta t^{3/2})$, assuming local boundedness of the density of the SDE solution (4) near the strike [5].

To perform numerical experiments we closely follow the algorithm prescribed in [4, Sect. 5] with predefined root-mean-square errors $\epsilon = [1, 2, 4, 8, 16] \times 10^{-4}$:

1. Start with level $\ell = 0$
2. Estimate variance using initial 10^4 samples
3. Evaluate optimal number of samples on each level as in [4, Sect. 5]
4. If $L \geq 2$, test for convergence [4, Sect. 5]
5. If $L < 2$ or not converged, set $\ell := \ell + 1$ and go to 2.

In addition, we used 10^6 samples to generate the plots where we estimate the rate of the strong and weak errors.

Figure 2 presents results for the payoff function $P = \max(0, x_2(1) - 1)$ applied to Clark and Camerson model problem with initial conditions $x_1(0) = x_2(0) = 1$. The top left plot with the superimposed reference slope with rate 1.5 shows that the variance $P_\ell - P_{\ell-1}$ is $O(\Delta t_\ell^{1.5})$. The top right plot shows that $\mathbb{E}[P_\ell - P_{\ell-1}] = O(\Delta t_\ell)$. The bottom left plot shows $\epsilon^2\, C$ where C is the computational complexity as defined in Theorem 1. The plot is versus ϵ, and the nearly horizontal line confirms that the MLMC complexity is $O(\epsilon^{-2})$, whereas the standard Monte Carlo approach has complexity $O(\epsilon^{-3})$. For accuracy $\varepsilon = 10^{-4}$, the antithetic MLMC is approximately 500 times more efficient than standard Monte Carlo. The bottom right plot shows that $\mathbb{V}[X_{2,N}^\ell - X_{2,N}^{\ell-1}] = O(\Delta t_\ell)$, corresponding to the standard strong convergence of order 0.5.

5 Subsampling of Levy Areas

Consider now a Brownian path $w(t)$ on the interval $[0, T]$ with N sub-intervals of size $\Delta t = T/N$. We define $w_n \equiv w(n\Delta t)$ and $\Delta w_n \equiv w_{n+1} - w_n$.

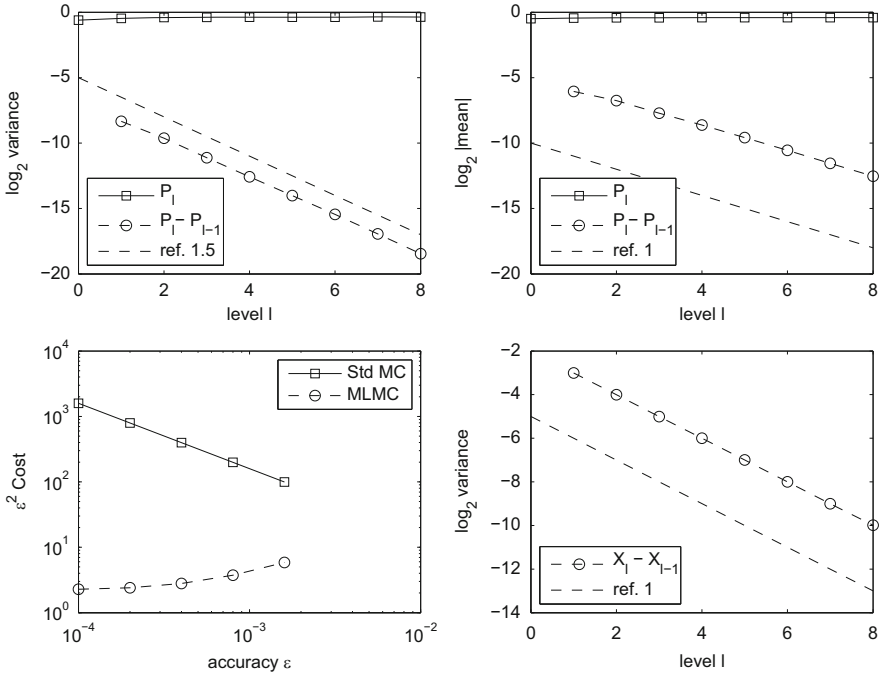


Fig. 2 Clark/Cameron model problem with payoff $\max(0, x_2(1) - 1)$.

Lemma 2. *The Lévy area for $w(t)$ can be expressed as*

$$[A_{jk}]_0^T = \sum_{n=0}^{N-1} \left((w_{j,n} - w_{j,0})\, \Delta w_{k,n} - (w_{k,n} - w_{k,0})\, \Delta w_{j,n} + [A_{jk}^s]_{n\Delta t}^{(n+1)\Delta t} \right)$$

where $[A_{jk}^s]_{n\Delta t}^{(n+1)\Delta t}$ is the Lévy area for the sub-interval $[n\Delta t, (n+1)\Delta t]$.

Proof. This follows from the definition of the Lévy area by expressing the integral over $[0, T]$ as the sum of integrals over each of the sub-intervals, and using the identity $w(t) - w(0) = (w_n - w_0) + (w(t) - w_n)$ to evaluate the integral on the n^{th} sub-interval. □

Ignoring the sub-interval Lévy areas $[A_{jk}^s]_{n\Delta t}^{(n+1)\Delta t}$, which corresponds to using the expected value of $[A_{jk}]_0^T$ conditional on $\{w(n\Delta t)\}_{0 \le n \le N}$, gives the Lévy area approximation:

$$[L_{jk}]_0^T = \sum_{n=0}^{N-1} \left((w_{j,n} - w_{j,0})\, \Delta w_{k,n} - (w_{k,n} - w_{k,0})\, \Delta w_{j,n} \right).$$

We denote by $[L^a_{jk}]_0^T$ the corresponding antithetic quantity generated by reversing the order of the Brownian increments $\Delta w_N, \Delta w_{N-1}, \ldots, \Delta w_1$. The antithetic label is due to the following lemma:

Lemma 3.

$$[L^a_{jk}]_0^T = -[L_{jk}]_0^T.$$

Proof.

$$\begin{aligned}
[L^a_{jk}]_0^T &= \sum_{n=0}^{N-1} \sum_{m=0}^{n-1} \Delta w_{j,N-1-m} \Delta w_{k,N-1-n} - \Delta w_{k,N-1-m} \Delta w_{j,N-1-n} \\
&= \sum_{m'=0}^{N-1} \sum_{n'=m'+1}^{N-1} \Delta w_{j,n'} \Delta w_{k,m'} - \Delta w_{k,n'} \Delta w_{j,m'} \\
&= -\sum_{n'=0}^{N-1} \sum_{m'=0}^{n'-1} \Delta w_{j,m'} \Delta w_{k,n'} - \Delta w_{k,m'} \Delta w_{j,n'} \\
&= -[L_{jk}]_0^T
\end{aligned}$$

The second line in the proof uses the substitutions $m' = N-1-n$, $n' = N-1-m$, and the third line simply switches the order of summation. □

5.1 Antithetic Subsampling

In Sect. 4 we showed that by setting the Lévy area to zero and using a suitable antithetic treatment we obtained an MLMC variance with the same order as the Milstein scheme for scalar SDEs. However, to obtain similarly good results for payoffs which depend on the path minimum (or maximum) we are not able to completely neglect the Lévy areas. Instead, for reasons which would require a lengthy explanation and will be addressed in future work, we need to improve the rate of strong convergence from $1/2$ to $3/4$ by approximating the Lévy areas by sub-sampling the driving Brownian path. Let M_f denote the number of subsamples required to approximate the Lévy area on the fine timestep. The subsampling timestep is given by $\delta_\ell = 2^{-\ell} T/M_f$. Since we want to obtain

$$\mathbb{E}\left\| [L_{jk}]_{t_n}^{t_n+1/2} - [A_{jk}]_{t_n}^{t_n+1/2} \right\|^2 = O((2^{-\ell})^{3/2}),$$

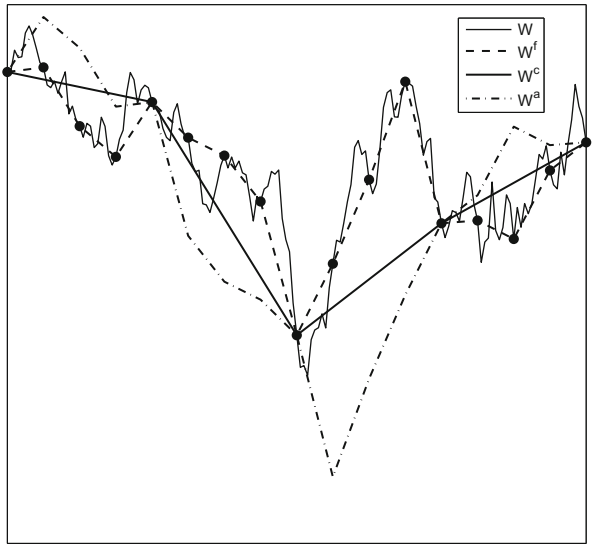


Fig. 3 Brownian path w, its piecewise linear interpolations w^c and w^f, and the antithetic w^a, for a single coarse timestep. The *circles* denote the points at which the Brownian path is sampled.

we need to take $M_f \approx 2^{\ell/2}$ sub-samples in each fine timestep. By the same reasoning we take $M_c \approx 2^{(\ell-1)/2}$, with sub-sampling timestep $\delta_{\ell-1} = 2^{-(\ell-1)}T/M_c$. For implementation we round the exponents, using $M_f = 2^{\lceil \ell/2 \rceil}$ and $M_c = 2^{\lceil (\ell-1)/2 \rceil}$.

Figure 3 illustrates a case in which w^c has $M_c = 4$ sub-sampling intervals within each coarse timestep, and w^f has $M_f = 8$ sub-sampling intervals within each fine timestep (this corresponds to level $\ell = 5$). With sub-sampling, the piecewise linear antithetic fine path w^a is defined by a time-reversal of the Brownian increments within each of the coarse sub-sampling intervals. In the case illustrated, the first coarse sub-sampling interval contains 4 fine sub-sampling intervals, so these 4 increments $\Delta w^{f,1}, \Delta w^{f,2}, \Delta w^{f,3}, \Delta w^{f,4}$ are re-ordered as $\Delta w^{f,4}, \Delta w^{f,3}, \Delta w^{f,2}, \Delta w^{f,1}$ to give the increments for w^a.

First we represent the Lévy area approximation on the coarse time interval as a sum of two approximations each with $M_c/2$ subsamples

$$[L^c_{jk}]^{t_{n+1}}_{t_n} = [L^c_{jk}]^{t_{n+1/2}}_{t_n} + [L^c_{jk}]^{t_{n+1}}_{t_{n+1/2}} + \left(\Delta w^\ell_{1,n+1/2}\, \Delta w^\ell_{2,n+1} - \Delta w^\ell_{2,n+1/2}\, \Delta w^\ell_{1,n+1} \right). \tag{13}$$

We can represent the Lévy area approximation for the first fine timestep within a coarse timestep as

$$[L_{jk}^{f}]_{t_n}^{t_n+1/2} = [L_{jk}^{c}]_{t_n}^{t_n+1/2} + \sum_{s=0}^{M_c/2-1} [L_{jk}^{f}]_{t_n+s\delta_{\ell-1}}^{t_n+(s+1)\delta_{\ell-1}}, \tag{14}$$

where $[L_{jk}^{f}]_{t_n+s\delta_{\ell-1}}^{t_n+(s+1)\delta_{\ell-1}}$ are Lévy area approximatiosn with $\frac{2M_f}{M_c}$ subsamples. Notice that $\frac{M_c}{2}\delta_{\ell-1} = 2^{-\ell}$. In the same way, we represent the Lévy area approximation for the antithetic path as

$$[L_{jk}^{a}]_{t_n}^{t_n+1/2} = [L_{jk}^{c}]_{t_n}^{t_n+1/2} + \sum_{s=0}^{M_c/2-1} [L_{jk}^{a}]_{t_n+s\delta_{\ell-1}}^{t_n+(s+1)\delta_{\ell-1}}. \tag{15}$$

Due to Proposition 3, $[L_{jk}^{a}]_{t_n+s\delta_{\ell-1}}^{t_n+(s+1)\delta_{\ell-1}} = -[L_{jk}^{f}]_{t_n+s\delta_{\ell-1}}^{t_n+(s+1)\delta_{\ell-1}}$. Hence

$$\left([L_{jk}^{f}]_{t_n}^{t_n+1/2} - [L_{jk}^{c}]_{t_n}^{t_n+1/2}\right) = -\left([L_{jk}^{a}]_{t_n}^{t_n+1/2} - [L_{jk}^{c}]_{t_n}^{t_n+1/2}\right),$$

which is the key antithetic property required for higher order MLMC variance convergence. We derive the analogous approximation for the second fine timestep within a coarse timestep. Returning to the Clark-Cameron example, where we focus only on the equation for x_2, with Lévy area approximation using M_c and M_f subsamples respectively, we obtain

$$X_{2,n+1}^{c} = X_{2,n}^{c} + X_{1,n}^{c}\Delta w_{2,n}^{\ell-1} + \tfrac{1}{2}\Delta w_{1,n}^{\ell-1}\Delta w_{2,n}^{\ell-1} + \tfrac{1}{2}[L_{jk}^{c}]_{t_n}^{t_{n+1}} \tag{16}$$

$$\begin{aligned} X_{2,n+1}^{f} = {} & X_{2,n}^{f} + X_{1,n}^{f}\Delta w_{2,n}^{\ell-1} + \tfrac{1}{2}\Delta w_{1,n}^{\ell-1}\Delta w_{2,n}^{\ell-1} + \tfrac{1}{2}[L_{jk}^{f}]_{t_n}^{t_n+1/2} + \tfrac{1}{2}[L_{jk}^{f}]_{t_n+1/2}^{t_{n+1}} \\ & + \tfrac{1}{2}\left(\Delta w_{1,n}^{\ell}\Delta w_{2,n+1/2}^{\ell} - \Delta w_{2,n}^{\ell}\Delta w_{1,n+1/2}^{\ell}\right), \end{aligned} \tag{17}$$

$$\begin{aligned} X_{2,n+1}^{a} = {} & X_{2,n}^{a} + X_{1,n}^{a}\Delta w_{2,n}^{\ell-1} + \tfrac{1}{2}\Delta w_{1,n}^{\ell-1}\Delta w_{2,n}^{\ell-1} + \tfrac{1}{2}[L_{jk}^{a}]_{t_n}^{t_n+1/2} + \tfrac{1}{2}[L_{jk}^{a}]_{t_n+1/2}^{t_{n+1}} \\ & + \tfrac{1}{2}\left(\Delta w_{1,n}^{\ell}\Delta w_{2,n+1/2}^{\ell} - \Delta w_{2,n}^{\ell}\Delta w_{1,n+1/2}^{\ell}\right), \end{aligned} \tag{18}$$

where we use Lévy areas approximations (13) and (14). We present a lemma that can be proved in a similar way to Lemma 3.1 in [5]:

Lemma 4. *If X_n^f, X_n^a and X_n^c are as defined above, and $N = 2^{\ell-1}$, then*

$$X_{1,n}^{f} = X_{1,n}^{a} = X_{1,n}^{c}, \quad \tfrac{1}{2}\left(X_{2,n}^{f} + X_{2,n}^{a}\right) = X_{2,n}^{c}, \quad n = 1, 2, \ldots, N$$

and

$$\sup_{0\le n\le N} \mathbb{E}\left[\left|X^f_{2,n} - X^a_{2,n}\right|^p\right] = O(\Delta t^{\frac{3}{4}p})$$

Our numerical experiments show that for lookback and barrier options the MLMC variance

$$\mathbb{V}\left[\tfrac{1}{2}\left(P(X^f) + P(X^a)\right) - P(X^c)\right]$$

has an $O(\Delta t^{3/2})$ upper bound. Since we use subsampling to approximate the Lévy areas, the computational cost corresponds to $\gamma = 3/2$ in Theorem 1, and as a consequence the complexity of the MLMC estimator is $O(\epsilon^{-2}(\log\epsilon)^2)$, whereas the standard Monte Carlo simulation complexity is $O(\epsilon^{-3})$.

6 Lookback and Barrier Options

6.1 Lookback Options

Lookback options are based on the minimum (or maximum) of the simulated path. As a specific example, we consider the payoff $P = x_2(T) - \min_{0<t<T} x_2(t)$, based on the second component x_2 of the Clark and Cameron model problem.

To improve the convergence we use $\tilde{X}^\ell(t)$ defined in (7). We have

$$\min_{0\le t<T} \tilde{X}^\ell_2(t) = \min_{0\le n<2^{\ell-1}} X^\ell_{2,n,min},$$

where the minimum of the fine approximation over the fine timestep $[t^{\ell-1}_n, t^{\ell-1}_{n+1/2}]$ is given by [6]

$$X^\ell_{2,n,min} = \tfrac{1}{2}\left(X^\ell_{2,n} + X^\ell_{2,n+1/2} - \sqrt{\left(X^\ell_{2,n+1/2} - X^\ell_{2,n}\right)^2 - 2\,g_2(X^\ell_n)^2\,\Delta t_\ell \log U_n}\right), \tag{19}$$

where U_n is a uniform random variable on the unit interval. The minima for the antithetic path are defined similarly, using the same uniform random numbers U_n.

For the coarse path, we do something slightly different. Using the same Brownian interpolation, we use Eq. (7) to define $\tilde{X}^{\ell-1}_{n+1/2} \equiv \tilde{X}^{\ell-1}((n+\frac{1}{2})\Delta t_{\ell-1})$. Given this interpolated value, the minimum value over the coarse interval can then be taken to be the smaller of the minima for the two fine intervals

$$X^{\ell-1}_{2,n,min}=\tfrac{1}{2}\left(X^{\ell-1}_{2,n}+\tilde{X}^{\ell-1}_{2,n+1/2}-\sqrt{\left(\tilde{X}^{\ell-1}_{2,n+1/2}-X^{\ell-1}_{2,n}\right)^2-2\,g_2(X^{\ell-1}_n)^2\,\Delta t_\ell\log U_n}\right),$$

$$X^{\ell-1}_{2,n+1/2,min}$$
$$=\tfrac{1}{2}\left(\tilde{X}^{\ell-1}_{2,n+1/2}+X^{\ell-1}_{2,n+1}-\sqrt{\left(X^{\ell-1}_{2,n+1}-\tilde{X}^{\ell-1}_{2,n+1/2}\right)^2-2\,g_2(X^{\ell-1}_n)^2\,\Delta t_\ell\log U_{n+1/2}}\right). \quad (20)$$

Note that we use $g_2(X^{\ell-1}_n)$ for both fine timesteps, because we have used the Brownian Bridge with diffusion term $g_2(X^{\ell-1}_n)$ to derive both minima. If we changed $g_2(X^{\ell-1}_n)$ to $g_2(\tilde{X}^{\ell-1}_{n+1/2})$ in $X^{\ell-1}_{2,n+1/2,min}$, this would mean that we used different Brownian Bridge on the first and second half of the coarse timestep and as a consequence we would violate (3). Note also the re-use of the same uniform random numbers U_n and $U_{n+1/2}$ used to compute the fine path minima. To perform numerical experiments we closely follow the algorithm prescribed in [4]. The results in Fig. 4 are for the Clark and Cameron model problem with this lookback payoff. The top left plot shows the behaviour of the variance of both P_ℓ and $P_\ell - P_{\ell-1}$. The superimposed reference slope with rate 1.5 indicates that the variance $V_\ell = \mathbb{V}[P_\ell - P_{\ell-1}] = O(\Delta t_\ell^{1.5})$, corresponding to $O(\epsilon^{-2}(\log\epsilon)^2)$ computational complexity for the antithetic MLMC estimator. The top right plot shows that $\mathbb{E}[P_\ell - P_{\ell-1}] = O(\Delta t_\ell)$. The bottom left plot shows computational complexity C (as defined in Theorem 1) with desired accuracy ϵ. The plot is of $\epsilon^2\,C$ versus ϵ, because we expect to see that $\epsilon^2\,C$ is only weakly dependent on ϵ for MLMC. For standard Monte Carlo without subsampling of the Lévy areas, theory predicts that $\epsilon^2\,C$ should be proportional to the number of timesteps on the finest level, which in turn is roughly proportional to ϵ^{-1} due to the weak convergence order. For accuracy $\varepsilon = 10^{-4}$, the antithetic MLMC is over 100 times more efficient than standard Monte Carlo. The bottom right plot shows that $\mathbb{V}[X^\ell_2 - X^{\ell-1}_2] = O(\Delta t_\ell^{3/2})$. This corresponds to the standard strong convergence of order 3/4.

6.2 Barrier Options

The barrier option which is considered is a down-and-out option for which the payoff is a Lipschitz function of the value of the underlying at maturity, provided the underlying has never dropped below a value B, i.e. $P = f(x_2(T))\,\mathbf{1}_{\{\tau>T\}}$, where the crossing time τ is defined as $\tau = \inf\{t : x_2(t) < B\}$. Using the Brownian Bridge interpolation, we can approximate $\mathbf{1}_{\{\tau>T\}}$ by $\prod_{n=0}^{2^{\ell-1}-1/2}\mathbf{1}_{\{X^\ell_{2,n,min}\geq B\}}$, where $X^\ell_{2,n,min}$ is defined in Eq. (19). This suggests following the lookback approximation

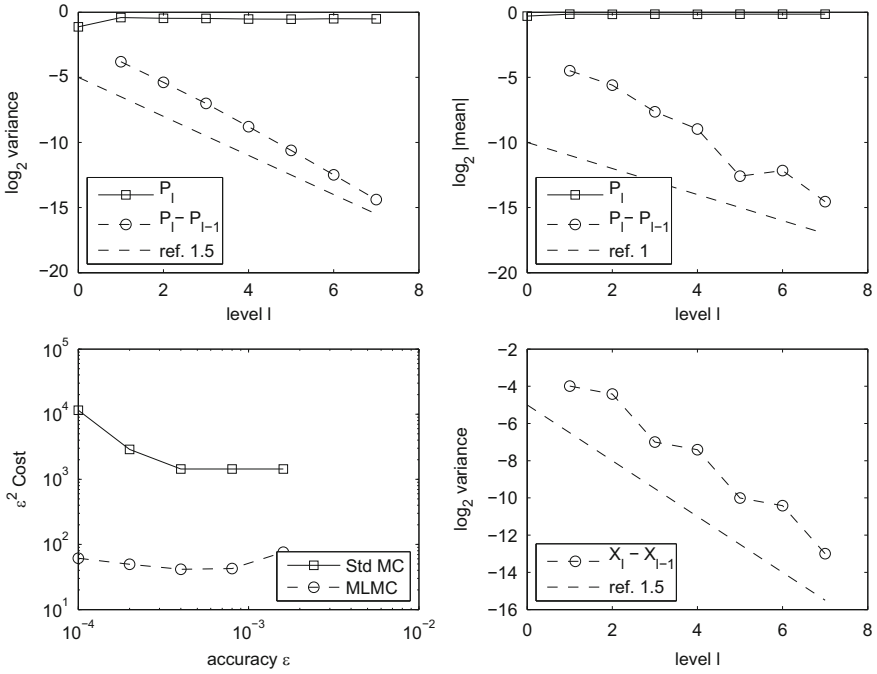


Fig. 4 Clark/Cameron model with payoff $x_2(1) - \min x_2(t)$.

in computing the minimum of both the fine and coarse paths. However, the variance would be larger in this case because the payoff is a discontinuous function of the minimum. A better treatment, which is the one used in [3], is to use the conditional Monte Carlo approach to further smooth the payoff. Since the process X_n^ℓ is Markovian, we have

$$\begin{aligned}
\mathbb{E}\left[f(X_{2,N}^\ell) \prod_{n=0}^{N-1/2} \mathbf{1}_{\{X_{n,min}^\ell \geq B\}} \right] &= \mathbb{E}\left[f(X_{2,N}^\ell)\, \mathbb{E}\left[\prod_{n=0}^{N-1/2} \mathbf{1}_{\{X_{2,n,min}^\ell \geq B\}} \mid X_0^\ell, \ldots, X_N^\ell \right] \right] \\
&= \mathbb{E}\left[f(X_{2,N}^\ell) \prod_{n=0}^{N-1/2} \mathbb{E}\big[\mathbf{1}_{\{X_{2,n,min}^\ell \geq B\}} \mid X_n^\ell, X_{n+1/2}^\ell\big] \right] \\
&= \mathbb{E}\left[f(X_{2,N}^\ell) \prod_{n=0}^{N-1/2} (1 - p_n^\ell) \right],
\end{aligned}$$

where

$$p_n^\ell = \mathbb{P}\left(\inf_{t_n<t<t_{n+1/2}} \tilde{X}_2^\ell(t) < B \mid X_n^\ell, X_{n+1/2}^\ell\right) = \exp\left(\frac{-2\,(X_{2,n}^\ell - B)^+ (X_{2,n+1/2}^\ell - B)^+}{g_2(X_n^\ell)^2\,\Delta t_\ell}\right)$$

The antithetic path is treated similarly. For the payoff for the coarse path we subsample $\tilde{X}_{n+1/2}^{\ell-1}$, as we did for the lookback option, to obtain

$$\mathbb{E}\left[f(X_{2,N}^{\ell-1}) \prod_{n=0}^{N-1/2} \mathbf{1}_{\{X_{2,n,min}^{\ell-1} \geq B\}}\right] = \mathbb{E}\left[f(X_{2,N}^{\ell-1}) \prod_{n=0}^{N-1/2} (1 - p_n^{\ell-1})\right],$$

where, for integer n,

$$p_n^{\ell-1} = \exp\left(\frac{-2\,(X_{2,n}^{\ell-1} - B)^+ (\tilde{X}_{n2,+1/2}^{\ell-1} - B)^+}{g_2(X_n^{\ell-1})^2\,\Delta t_\ell}\right),$$

$$p_{n+1/2}^{\ell-1} = \exp\left(\frac{-2\,(\tilde{X}_{2,n+1/2}^{\ell-1} - B)^+ (X_{2,n+1}^{\ell-1} - B)^+}{g_2(X_n^{\ell-1})^2\,\Delta t_\ell}\right).$$

Note that the same $g_2(X_n^{\ell-1})$ is used to calculate both probabilities for the same reason as for the lookback option.

The results in Fig. 5 are for barrier option with barrier $B = 0.1$. The top left plot shows the behaviour of the variance of both P_ℓ and $P_\ell - P_{\ell-1}$. The superimpose reference slope with rate 1.5 indicates that the variance $V_\ell = \mathbb{V}[P_\ell - P_{\ell-1}] = O(\Delta t_\ell^{1.5})$. This corresponds to an $O(\epsilon^2(\log\epsilon)^2)$ computational complexity for the antithetic MLMC, due to the additional cost of the sub-sampling to approximate the Lévy areas. The top right plot shows that $\mathbb{E}[P_\ell - P_{\ell-1}] = O(\Delta t_\ell)$. The bottom left plot shows the variation of the computational complexity C with desired accuracy ϵ. For standard Monte Carlo without subsampling of the Lévy areas, theory predicts that $\epsilon^2\,C$ should be proportional to the number of timesteps on the finest level, which in turn is roughly proportional to ϵ^{-1} due to the weak convergence order. For accuracy $\varepsilon = 10^{-4}$, antithetic MLMC is almost 10 times more efficient than standard Monte Carlo. The bottom right plot shows that $\mathbb{V}[X_2^\ell - X_2^{\ell-1}] = O(\Delta t_\ell^{3/2})$. This corresponds to standard strong convergence of order 3/4.

7 Conclusions

In this paper we extended results from [3] and [5] to lookback and barrier options for multidimensional SDEs. By suitable modification of the antithetic MLMC estimator, using sub-sampling of the driving Brownian path to approximate the Lévy areas, we obtained $O(\epsilon^{-2}\log(\epsilon)^2)$ complexity for barrier and lookback options.

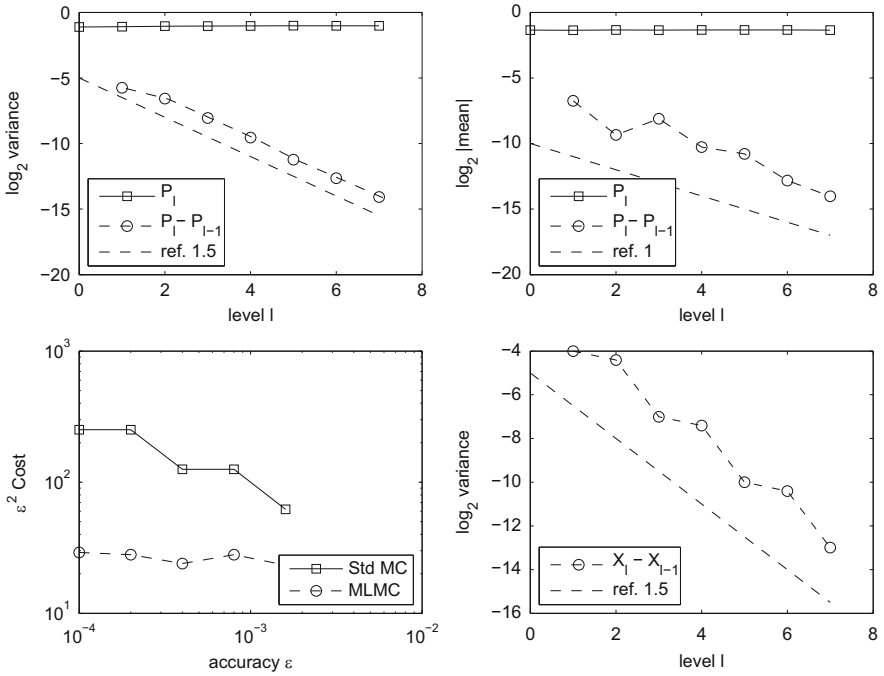


Fig. 5 Clark/Cameron model with payoff $\min(x_2(1) - 1, 0)\ \mathbf{1}_{\min x_2(t) > 0.1}$.

Similar results have also been obtained for digital options which are a discontinuous function of the final state, but they have been omitted here due to lack of space.

References

1. Clark, J.M.C., Cameron, R.J.: The maximum rate of convergence of discrete approximations for stochastic differential equations. In: Grigelionis, B. (ed.) Stochastic Differential Systems Filtering and Control. Lecture Notes in Control and Information Sciences, vol. 25, 162–171. Springer, Berlin/Heidelberg (1980)
2. Gaines, J.G., Lyons, T.J.: Random generation of stochastic integrals. SIAM J. Appl. Math. **54**, 1132–1146 (1994)
3. Giles, M.B.: Improved multilevel Monte Carlo convergence using the Milstein scheme. In: Keller, A., Heinrich, S., Niederreiter, H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2006, pp. 343–358. Springer, Berlin/Heidelberg (2008)
4. Giles, M.B.: Multilevel Monte Carlo path simulation. Oper. Res. **56**, 607–617 (2008)
5. Giles, M.B., Szpruch, L.: Antithetic multilevel Monte Carlo estimation for multi-dimensional SDEs without Lévy area simulation. Ann. Appl. Probab. (2013, to appear).
6. Glasserman, P.: Monte Carlo Methods in Financial Engineering. Springer, New York (2004)
7. Müller-Gronbach, T.: The optimal uniform approximation of systems of stochastic differential equations. Ann. Appl. Probab. **12**, 664–690 (2002)

8. Müller-Gronbach, T.: Strong approximation of systems of stochastic differential equations. Habilitation thesis, TU Darmstadt (2002)
9. Rydén, T., Wiktorsson, M.: On the simulation of iterated Itô integrals. Stochastic Process. Appl. **91**, 151–168 (2001)
10. Wiktorsson, M.: Joint characteristic function and simultaneous simulation of iterated Itô integrals for multiple independent Brownian motions. Ann. Appl. Probab. **11**, 470–487 (2001)

On the Convergence of Quantum and Sequential Monte Carlo Methods

François Giraud and Pierre Del Moral

Abstract Sequential and Quantum Monte Carlo methods, as well as genetic type search algorithms can be interpreted as a mean field and interacting particle approximations of Feynman-Kac models in distribution spaces. The performance of these population Monte Carlo algorithms is related to the stability properties of nonlinear Feynman-Kac semigroups. In this paper, we analyze these models in terms of Dobrushin ergodic coefficients of the reference Markov transitions and the oscillations of the potential functions. Sufficient conditions for uniform concentration inequalities w.r.t. time are expressed explicitly in terms of these two quantities. Special attention is devoted to the particular case of Boltzmann-Gibbs measures' sampling. In this context, we design an explicit way of tuning the temperature schedule with the number of Markov Chain Monte Carlo iterations.

1 Introduction

Sequential and Quantum Monte Carlo methods (*abbreviate SMC and QMC*) are stochastic algorithms to sample from complex high-dimensional probability distributions. These stochastic simulation techniques are of current use in numerical physics [1, 2, 21] to compute ground state energies. They are also used in statistics, signal processing and information sciences [4, 10, 12, 14] to compute posterior

F. Giraud (✉)
CEA-CESTA, 33114 Le Barp, France
e-mail: francois.giraud@ens-cachan.org

P. Del Moral
INRIA Bordeaux Sud-Ouest, Domaine Universitaire, 351, cours de la Liberation,
33405 Talence Cedex, France

CMAP, Applied Mathematical Department, Ecole Polytechnique, Paris, France
e-mail: pierre.del_moral@inria.fr

J. Dick et al. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2012*,
Springer Proceedings in Mathematics & Statistics 65,
DOI 10.1007/978-3-642-41095-6_17, © Springer-Verlag Berlin Heidelberg 2013

distributions of partially observed signal or unknown parameters. In evolutionary computing literature, these Monte Carlo methods are used as natural population search algorithms for solving optimization problems. From the pure mathematical viewpoint, these advanced Monte Carlo methods coincide with mean field particle interpretations of Feynman-Kac (*abbreviate FK*) models. For a thorough discussion on FK models we refer the reader to the monograph [11], and references therein. The principle (see also [12] and the references therein) is to approximate a sequence of target probability distributions $(\eta_n)_n$ by a large cloud of random samples termed particles or walkers. The algorithm starts with N independent samples from η_0 and then alternates two types of steps: an acceptance-rejection scheme equipped with a recycling mechanism, and a sequence of free exploration of the state space.

In the recycling stage, the current cloud of particles is transformed by randomly duplicating and eliminating particles in a suitable way, similarly to a selection step in models of population genetics. In the Markov evolution step, particles move independently of each other (mutation step).

This method is often used for solving sequential problems, such as filtering (see e.g., [10]). In other interesting problems, these algorithms also turn out to be efficient to sample from a single target measure η. In this context, the central idea is to find a judicious interpolating sequence of measures $(\eta_k)_{0\leq k\leq n}$ with increasing sampling complexity, starting from some initial distribution η_0, up to the terminal one $\eta_n = \eta$. Consecutive measures η_k and η_{k+1} are sufficiently similar to allow for efficient importance sampling and/or acceptance-rejection sampling. The sequential aspect of the approach is then an "artificial way" to introduce the difficulty of sampling gradually. Large population sizes allow to cover several modes simultaneously. This is an advantage compared to standard MCMC methods. These sequential samplers have been used with success in several application domains, including rare events simulation (see [5]), stochastic optimization and Boltzmann-Gibbs measures sampling [12].

Up to now, SMC and QMC algorithms have been mostly analyzed using asymptotic (i.e. when number of particles N tends to infinity) techniques, notably through central limit theorems and large deviation principles (see for instance [4, 7, 9, 10, 13, 14, 16, 17, 23] and [11] for an overview). Our work relates to less studied non-asymptotic problems, and follows those based on Markov kernels' mixing properties (see for instance [6, 17] and [11]). We emphasize that other independent approaches, such as Whiteley's [27] or Schweizer's [26], based on, e.g., drift conditions, hyper-boundedness, or spectral gaps, lead to convergence results that may also apply to non-compact state spaces. To our knowledge, these techniques are restricted to non-asymptotic variance theorems and they cannot be used to derive uniform and exponential concentration inequalities.

The present work consists in estimating explicitly the stability properties of FK semigroup in terms of the Dobrushin ergodic coefficient of the reference Markov chain and the oscillations of the potential functions. We combine these techniques with non-asymptotic theorems on L^p error bounds [17] and some useful concentration inequalities [18]. Another contribution is to provide parameter tuning strategies that allow to deduce some useful uniform concentration inequalities

w.r.t. the time parameter. These results also apply to non-homogeneous FK models associated with cooling temperature parameters.

In a preliminary section, we recall a few essential notions related to Dobrushin coefficients or FK semigroups, as well as a couple of important non-asymptotic results we use in the further development of the article. The second part is concerned with the semigroup stability analysis of these models. We also provide a couple of uniform L^p-deviations and concentration estimates. We end the article with an application of these results to Boltzmann-Gibbs models associated with a decreasing temperature schedule. In this context, SMC and QMC algorithms can be interpreted as a sequence of interacting simulated annealing (*abbreviate ISA*) algorithms. The detailed proofs of the results presented in this article will be presented in a forthcoming publication dedicated to adaptive particle algorithms (see [20] for a preliminary version).

2 Preliminaries

2.1 Notations

Let (E, r) be a complete, separable metric space and let $\mathscr{E}$ be the σ-algebra of Borel subsets of E. Denote by $\mathscr{P}(E)$ the space of probability measures on E. Let $\mathscr{B}(E)$ be the space of bounded, measurable, real-valued functions on E. Let $\mathscr{B}_1(E) \subset \mathscr{B}(E)$ be the subset of all bounded by 1 functions.

If $\mu \in \mathscr{P}(E)$, $f \in \mathscr{B}(E)$ and K, K_1, K_2 are Markov kernels on E, then $\mu(f)$ denotes the quantity $\int_E f(x)\mu(dx)$, $K_1.K_2$ denotes the Markov kernel defined by

$$K_1.K_2(x, A) = \int_E K_1(x, dy)K_2(y, A),$$

$K.f$ denotes the function defined by

$$K.f(x) = \int_E K(x, dy)\, f(y)$$

and $\mu.K$ denotes the probability measure defined by

$$\mu.K(A) = \int_E K(x, A)\mu(dx).$$

For any $f \in \mathscr{B}(E)$, denote by $\mathrm{osc}(f)$ the quantity $(f_{\max} - f_{\min})$. For any $x \in E$, the Dirac measure centered on x is designated by δ_x.

2.2 The Feynman-Kac Measure-Valued Model

Consider a sequence of probability measures $(\eta_n)_n$, defined by an initial measure η_0 and recursive relations:

$$\forall f \in \mathscr{B}(E), \quad \eta_n(f) = \frac{\eta_{n-1}(G_n \times M_n.f)}{\eta_{n-1}(G_n)}$$

for positive functions $G_n \in \mathscr{B}(E)$ and Markov kernels M_n with $M_n(x,\cdot) \in \mathscr{P}(E)$ and $M_n(\cdot, A) \in \mathscr{B}_1(E)$. This is the sequence of measures we wish to approximate with the SMC algorithm. In an equivalent way, $(\eta_n)_n$ can be defined by the relation:

$$\eta_n = \phi_n(\eta_{n-1})$$

where $\phi_n : \mathscr{P}(E) \to \mathscr{P}(E)$ is the FK transformation associated with potential function G_n and Markov kernel M_n and defined by

$$\phi_n(\eta_{n-1}) = \psi_{G_n}(\eta_{n-1}).M_n$$

with

$$\psi_{G_n}(\eta_{n-1})(dx) := \frac{1}{\eta_{n-1}(G_n)}\, G_n(x)\, \eta_{n-1}(dx)$$

The next formula provides an interpretation of the Boltzmann-Gibbs transformation in terms of a nonlinear Markov transport equation

$$\Psi_{G_n}(\eta_{n-1})(dy) = \left(\eta_{n-1} S_{n,\eta_{n-1}}\right)(dy) := \int \eta_{n-1}(dx) S_{n,\eta_{n-1}}(x, dy)$$

with the Markov transition S_{n,η_n} defined below

$$S_{n,\eta_{n-1}}(x, dy) = \varepsilon_n.G_n(x)\, \delta_x(dy) + (1 - \varepsilon_n.G_n(x))\ \Psi_{G_n}(\eta_{n-1})(dy),$$

(for any constant $\varepsilon_n > 0$ so that $\varepsilon_n.G_n \leq 1$). This implies

$$\eta_n = \eta_{n-1} K_{n,\eta_{n-1}} \quad \text{with} \quad K_{n,\eta_{n-1}} = S_{n,\eta_{n-1}} M_n$$

Therefore, η_n can be interpreted as the distributions of the random states $\overline{X}_n$ of a Markov chain whose Markov transitions

$$\mathbb{P}\left(\overline{X}_{n+1} \in dy \mid \overline{X}_n = x\right) := K_{n+1,\eta_n}(x, dy)$$

depend on the current distribution $\eta_n = \text{Law}\left(\overline{X}_n\right)$.

An important point (FK semigroup structure, see e.g., [17]) is that the semigroup transformations

$$\phi_{p,n} := \phi_n \circ \phi_{n-1} \circ \ldots \circ \phi_{p+1}$$

admit a comparable structure as each of the ϕ_k, i.e. for any integers $p < n$, there exist a positive function $G_{p,n} \in \mathscr{B}(E)$ and a Markov kernel $P_{p,n}$ so that:

$$\forall f \in \mathscr{B}(E), \quad \forall \mu \in \mathscr{P}(E), \quad \phi_{p,n}(\mu).f = \frac{\mu\left(G_{p,n} \times P_{p,n}.f\right)}{\mu(G_{p,n})} \tag{1}$$

2.3 The Associated Interacting Particle System

In SMC and QMC algorithms, we approximate the measures η_n by simulating an interacting particle system $(\zeta_n)_n = \left(\zeta_n^1, \ldots, \zeta_n^N\right)_n$ of size N so that

$$\eta_n^N = \frac{1}{N} \sum_{1 \le i \le N} \delta_{\zeta_n^i} \to_{N \uparrow \infty} \eta_n$$

Of course, the main issue is to make precise and to quantify this convergence.

We start with N independent samples $\zeta_0 = (\zeta_0^1, \ldots, \zeta_0^N)$ from η_0. The particle dynamics alternates two genetic type transitions.

During the first step, every particle ζ_n^i evolves to a new particle $\hat{\zeta}_n^i$ randomly chosen with the distribution

$$S_{\eta_n^N}(\zeta_n^i, dx) := \varepsilon_{n+1}.G_{n+1}(\zeta_n^i)\, \delta_{\zeta_n^i}(dx) + \left(1 - \varepsilon_{n+1}.G_{n+1}(\zeta_n^i)\right)\, \Psi_{G_{n+1}}(\eta_n^N)(dx)$$

with the updated measures

$$\Psi_{G_{n+1}}(\eta_n^N) = \sum_{j=1}^{N} \frac{G_{n+1}(\zeta_n^j)}{\sum_{k=1}^{N} G_{n+1}(\zeta_n^k)} \delta_{\zeta_n^j}$$

This transition can be interpreted as an acceptance-rejection scheme with a recycling mechanism. In the second step, the selected particles $\hat{\zeta}_n^i$ evolve randomly according to the Markov transitions M_{n+1}. In other words, for any $1 \le i \le N$, we sample a random state ζ_{n+1}^i with distribution $M_{n+1}\left(\hat{\zeta}_n^i, dx\right)$.

Denote respectively by $\mathbb{P}(\cdot)$ and $\mathbb{E}(\cdot)$ probabilities and expectations taken with respect to the random variables $(\zeta_n^i)_{n,i}$ and $(\hat{\zeta}_n^i)_{n,i}$.

2.4 Dobrushin Ergodic Coefficients

The Dobrushin coefficient $\beta(K) \in [0, 1]$ of a Markov kernel K on E, is defined by:

$$\beta(K) = \sup\{K(x, A) - K(y, A) \mid x, y \in E,\ A \in \mathscr{E}\},$$

or in an equivalent way:

$$\beta(K) = \sup\{\|K(x,\cdot) - K(y,\cdot)\|_{tv} \mid x, y \in E\}$$

where $\|\mu - \nu\|_{tv}$ denotes the total variation distance between the measures μ and ν.

The parameter $\beta(K)$ caracterizes mixing properties of the Markov kernel K. Note that the function β is an operator norm, in the sense that $\beta(K_1.K_2) \leq \beta(K_1).\beta(K_2)$, for any couple of Markov kernels K_1, K_2. Further details on these ergodic coefficients can be found in the monograph [11].

Estimating these coefficients is generally a difficult task (related to the large field of Markov chains' stability), since their definition involves a supremum over every pair $(x, y) \in E^2$ and every set $A \in \mathscr{E}$. However, here is a first remark: if a Markov kernel K satisfies the ergodic total variation convergence $K^m(x,\cdot) \to \mu$ uniformly w.r.t. $x \in E$ when m tends to infinity, then $\beta(K^m)$ tends to zero.

In the particular case of a finite state space E, the Dobrushin ergodic coefficient of a Markov kernel K on E is given by the formula

$$\beta(K) = \frac{1}{2} \sup\left\{\sum_{l\in E} |K(i,\{l\}) - K(j,\{l\})| \ ; \ i, j \in E\right\},$$

which implies it is calculable as soon as the probability of the elementary transitions $K(i,\{l\})$ are known. This formula can provide a approximation of $\beta(K)$ in the case of an infinite but simple (low dimensional) state space E, that one can discretize.

In practice, the property $\beta(K) < 1$ is easily met as soon as the state space E is compact. Typically, any Markov kernel of the form

$$K(x, dy) = h(x, y)m(dy)$$

where h is a positive, continuous function on E^2 and m a reference measure on E, satisfies $\beta(K) < 1$. Otherwise, in some other particular situations, one can explicitly estimate $\beta(K)$. For instance, if $E = \mathbb{R}^d$ and

$$K(x, dy) \propto e^{-\alpha|y-a(x)|}dy$$

for some $\alpha > 0$ and some bounded function $a : E \to E$, then for all $x, x' \in E$ we have

$$\frac{K(x, dy)}{K(x', dy)} = e^{\alpha(|y-a(x')|-|y-a(x)|)} \leq e^{\alpha\cdot osc(a)}$$

$\Rightarrow K(x, dy) \leq e^{-\alpha\cdot osc(a)}K(x', dy)$. This clearly implies $\beta(K) \leq (1 - e^{-\alpha\cdot osc(a)})$.

The reader will also find in [17] an estimate of $\beta(K^2)$ in the following case

$$K(x, dy) \propto e^{-\frac{1}{2}|y-a(x)|^2}dy,$$

when the function a is constant outside some compact set $F \subset E$. Finally, the case of the Metropolis-Hastings kernel will be recalled page 395.

2.5 Some Non-Asymptotic Results

To quantify the FK semigroup stability properties, it is convenient to introduce the following parameters.

Definition 1. For any integers $p < n$, we set

$$b_n := \beta(M_n) \quad \text{and} \quad b_{p,n} := \beta(P_{p,n}).$$

$$g_n := \sup_{x,y\in E} \frac{G_n(x)}{G_n(y)} \quad \text{and} \quad g_{p,n} := \sup_{x,y\in E} \frac{G_{p,n}(x)}{G_{p,n}(y)}.$$

The quantities $g_{p,n}$, and respectively $b_{p,n}$, reflect the oscillations of the potential functions $G_{p,n}$, and respectively the mixing properties of the Markov transition $P_{p,n}$ associated with the FK semigroup $\phi_{p,n}$ described in (1). Several contraction inequalities of $\phi_{p,n}$ w.r.t. the total variation norm or different types of relative entropies can be derived in terms of these two quantities (see for instance [11]).

The performance analysis developed in this article is partly based on the two non-asymptotic inequalities presented below.

The following L^p error bound for all $f \in \mathscr{B}_1(E)$ is proved in [17]:

$$\mathbb{E}\left(\left|\eta_n^N(f) - \eta_n(f)\right|^p\right)^{1/p} \leq \frac{B_p}{\sqrt{N}} \sum_{k=0}^{n} g_{k,n} b_{k,n} \tag{2}$$

where B_p designates an universal constant.

In the further development of the article we also use the following exponential concentration inequality derived in [18]. For all $f \in \mathscr{B}_1(E)$ and any $\varepsilon > 0$ we have:

$$\frac{-1}{N} \log \mathbb{P}\left(|\eta_n^N(f) - \eta_n(f)| \geq \frac{r_n}{N} + \varepsilon\right) \geq \frac{\varepsilon^2}{2}\left[b_n^\star \overline{\beta}_n + \frac{\sqrt{2} r_n}{\sqrt{N}} + \varepsilon\left(2r_n + \frac{b_n^\star}{3}\right)\right]^{-1} \tag{3}$$

where r_n, $\overline{\beta}_n$ and $b_n^\star$ are constants so that:

$$\begin{cases} r_n & \leq \sum_{p=0}^{n} 4 g_{p,n}^3 b_{p,n} \\ \overline{\beta}_n^2 & \leq \sum_{p=0}^{n} 4 g_{p,n}^2 b_{p,n}^2 \\ b_n^\star & \leq \sup_{0\leq p\leq n} 2 g_{p,n} b_{p,n} \end{cases}$$

3 General Feynman-Kac semigroup analysis

Equations (2) and (3) provide explicit non-asymptotic estimates in terms of the quantities $g_{p,n}$ and $b_{p,n}$. Written this way, they hardly apply to any SMC' parameters tuning decision, since the only known or calculable objects are generally the reference Markov chain M_p and the elementary potential functions G_p. We thus have to estimate $g_{p,n}$ and $b_{p,n}$ in terms of the g_p and b_p.

By construction (see Lemma 2.1 in [15]), $G_{p,n}$ and $P_{p,n}$ satisfy the following backward relations:

$$\begin{cases} G_{p-1,n} = G_p \times M_p.G_{p,n} \\ \\ P_{p-1,n}.f = \dfrac{M_p.(G_{p,n} \times P_{p,n}.f)}{M_p.G_{p,n}} \end{cases}$$

with the initial definitions $G_{n,n} = 1$ and $P_{n,n} = Id$. By combining these formulae with Dobrushin ergodic coefficient estimation techniques, we obtain the following lemma:

Lemma 1. *For any integers $p \leq n$, we have:*

$$\begin{aligned} g_{p,n} - 1 &\leq \sum_{k=p+1}^{n} (g_k - 1) \prod_{i=p+1}^{k-1} (b_i g_i) \\ b_{p,n} &\leq \prod_{k=p+1}^{n} b_k.g_{k,n} \end{aligned} \tag{4}$$

To obtain uniform bounds w.r.t. time n (in the case of the L^p norm), we notice that

$$\sum_{p=0}^{n} \prod_{k=p+1}^{n} b_k g_{k-1,n} < +\infty \Longrightarrow \sum_{p=0}^{n} g_{p,n} b_{p,n} < +\infty$$

This naturally leads to a sufficient condition of the following type:

$$b_k \times g_{k-1,n} \leq a \quad \text{with} \quad 0 < a < 1$$

for any $k < n$, which ensures:

$$\forall f \in \mathscr{B}_1(E), \quad \mathbb{E}\left(\left|\eta_n^N(f) - \eta_n(f)\right|^p\right)^{1/p} \leq \frac{B_p}{\sqrt{N}} \frac{1}{1-a} \tag{5}$$

More generally, this condition ensures uniform bounds for $\overline{\beta_n}^2$, $b_n^\star$ and r_n :

$$\overline{\beta_n}^2 \leq \frac{4}{1-a^2} \qquad b_n^\star \leq 2 \qquad r_n \leq \frac{4}{1-a} \cdot \sup_{p,n} g_{p,n}^2$$

bearing in mind that the $g_{p,n}$ are bounded in the cases of interest. We then fix $0 < a < 1$ and the objective is to find conditions on the b_p so that $b_k g_{k-1,n} \leq a$. This parameter a is to be chosen according to the error we allow ourselves to commit, and the number N of particles involved, with bounds explicited above. In order to explicit relevant and applicable conditions, we study two typical cases of assumptions on the potential functions G_p. The first one being that the g_p are bounded (Theorem 1), the second one being that the g_p tend to 1 (Theorem 2).

Theorem 1. *Under the assumption $\forall p \in \mathbb{N}, g_p \leq M$, where M is a constant, condition*

$$b_p \leq \frac{a}{M(1+a)} \tag{6}$$

ensures the L^p error bound (5), as well as the following concentration inequality:

$$\forall y \geq 0, \quad \forall f \in \mathscr{B}_1(E), \quad \mathbb{P}\left(|\eta_n^N(f) - \eta_n(f)| \geq \frac{r_1^\star N + r_2^\star y}{N^2}\right) \leq e^{-y}$$

with

$$\begin{cases} r_1^\star = \frac{9}{2}\frac{M^2}{(1-a)^3} + \sqrt{\frac{8}{\sqrt{1-a^2}} + \frac{18M^2}{(1-a)^2\sqrt{N}}} \\ r_2^\star = 18\frac{M^2}{(1-a)^2} + \sqrt{\frac{8}{\sqrt{1-a^2}} + \frac{18M^2}{(1-a)^2\sqrt{N}}} \end{cases}$$

Let us now consider the case where g_p tends decreasingly to 1. We define

$$\alpha = \frac{a}{1-a} > 0 \quad \text{so that} \quad a = \frac{\alpha}{1+\alpha}$$

Theorem 2. *Under the assumption $g_p \longrightarrow_{p\to\infty} 1$ (decreasingly), if the sequence b_p satisfies for any $p \geq 1$,*

$$b_p \leq \frac{g_p^\alpha - 1}{g_p^{\alpha+1} - 1} \longrightarrow a \quad \text{and} \quad b_p \leq \frac{a}{g_p^{\alpha+1}} \longrightarrow a$$

then the L^p error bound (5) is satisfied, as well as the following concentration inequalities :

$$\forall y \geq 0, \quad \forall f \in \mathscr{B}_1(E), \quad \mathbb{P}\left(|\eta_n^N(f) - \eta_n(f)| \geq \frac{r_3^\star(n).N + r_4^\star(n).y}{N^2}\right) \leq e^{-y}$$

In the above displayed formulae $r_3^\star(n)$ *and* $r_4^\star(n)$ *are defined below in terms of a sequence* u_n *which tends to* 1*, as n tends to* ∞*:*

$$\begin{cases} r_3^\star(n) = \frac{9.u_n}{2(1-a)} + \sqrt{\frac{8}{\sqrt{1-a^2}} + \frac{18.u_n}{\sqrt{N}}} \\ r_4^\star(n) = \frac{18.u_n}{1-a} + \sqrt{\frac{8}{\sqrt{1-a^2}} + \frac{18.u_n}{\sqrt{N}}} \end{cases}$$

Such conditions on the b_p can appear to be difficult to reach since the Markov kernels may be imposed by the application under study. However, we can deal with this problem as soon as we can simulate a Markov kernel K_n such that $\eta_n.K_n = \eta_n$. Indeed, the algorithm designer can add MCMC evolution steps next to each M_n-mutation step, to stabilize the system. From the formal viewpoint, the target sequence $(\eta_n)_n$ is clearly also solution of the FK measure-valued equations associated with the Markov kernels $M_n' = M_n.K_n^{m_n}$, where iteration numbers m_n are to be chosen loosely. This system is more stable since the corresponding b_p' satisfy:

$$b_p' \leq b_p.\beta(K_p^{m_p}) \leq b_p.\beta(K_p)^{m_p}.$$

In such cases, Theorems 1 and 2 provide sufficient conditions on iteration numbers m_p to ensure the convergence of the algorithm.

4 The Particular Case of Boltzmann-Gibbs Measures, Interacting Simulated Annealing

Let $V \in \mathscr{B}(E)$. For all $\beta \geq 0$, denote Boltzmann-Gibbs probability measure associated with "temperature" β and potential function V by:

$$\mu_\beta(dx) = \frac{1}{Z_\beta} e^{-\beta.V(x)} m(dx),$$

where m is a reference measure, and Z_β a normalizing constant. It is well known that Boltzmann-Gibbs measures' sampling is related to the problem of minimizing the potential function V, since μ_β tends to concentrate on V's minimizers as temperature β tends to infinity. One illustration is the following inequality, satisfied for all $0 < \varepsilon' < \varepsilon$:

$$\mu_\beta\left(V \geq V_{\min} + \varepsilon\right) \leq \frac{e^{-\beta(\varepsilon-\varepsilon')}}{m_{\varepsilon'}} \tag{7}$$

where $m_{\varepsilon'} = m\left(V \leq V_{\min} + \varepsilon'\right) > 0$.

Besides, let fix a "temperature schedule", being a strictly increasing sequence β_n so that $\beta_n \longrightarrow +\infty$. The sequence $(\eta_n)_n := (\mu_{\beta_n})_n$ admits a FK structure associated with potential functions $G_n = e^{-(\beta_n-\beta_{n-1}).V}$ and Markov kernels M_n

chosen as being MCMC dynamics for the current target distributions. In this context, the SMC algorithm, used as a strategy to minimize V, can be interpreted as a sequence of interacting simulated annealing (*abbreviate ISA*) algorithms.

We propose in this section to turn the previously raised conditions on b_p and g_p into conditions on the temperature schedule to use, and the number of MCMC steps. We will then combine the concentration results of Sect. 3 with inequality (7) to obtain results in terms of optimization performance.

Let us fix a "temperature schedule" (β_n) and denote:

- $\eta_n(dx) = \mu_{\beta_n}(dx) = \frac{1}{Z_{\beta_n}} e^{-\beta_n V(x)} m(dx)$;
- $G_p(x) = e^{-\Delta_p . V(x)}$;
- And then $g_p = e^{\Delta_p . \mathrm{osc}(V)}$.

where Δ_p are the increments of temperature $\Delta_p = \beta_p - \beta_{p-1}$. At a fixed temperature β, let us consider the simulated annealing Markov kernel, designated by K_β. It involves a proposition kernel $K(x, dy)$, assumed here as being fixed, according to the following formulae (written here in the case where K is symmetric, see [3]):

$$K_\beta(x, dy) = K(x, dy). \min\left(1, e^{-\beta(V(y)-V(x))}\right) \qquad \forall y \neq x$$

$$K_\beta(x, \{x\}) = 1 - \int_{y \neq x} K(x, dy). \min\left(1, e^{-\beta(V(y)-V(x))}\right)$$

Under the assumption $K^{k_0}(x, \cdot) \geq \delta \nu(\cdot)$ for some integer k_0, some measure ν and some $\delta > 0$, one can show (see [3]) that:

$$\beta(K_\beta^{k_0}) \leq \left(1 - \delta e^{-\beta \overline{\Delta V}(k_0)}\right) \tag{8}$$

where $\overline{\Delta V}(k_0)$ is the maximum potential gap one can obtain making k_0 movements with K. This quantity is bounded by $\mathrm{osc}(V)$. To let the b_p's tuning be possible, it is out of the question to choose $M_p = K_{\beta_p}$, but $M_p = K_{\beta_p}^{k_0 . m_p}$, the simulated annealing kernel iterated $k_0 . m_p$ times, to obtain suitable mixing properties. The algorithm's user then has a choice to make on two parameters: the temperature schedule β_p, and the kernels $K_{\beta_p}^{k_0}$ iteration numbers m_p. Note that for all $b \in (0, 1)$, condition $b_p \leq b$ is turned into $\left(1 - \delta e^{-\beta_p \overline{\Delta V}(k_0)}\right)^{m_p} \leq b$, which can also be written:

$$m_p \geq \frac{log(\frac{1}{b}) e^{\overline{\Delta V}(k_0) . \beta_p}}{\delta}$$

Then, combining the concentration inequality (7), the theorems of Sect. 3 (taken with indicator function $f = \mathbf{1}_{\{V \geq V_{\min} + \varepsilon\}}$), and the Dobrushin ergodic coefficient estimation (8) we obtain the following theorem:

Theorem 3. *Let us fix $a \in (0,1)$. If the temperature schedule (β_p) and the iteration numbers m_p satisfy one of these two conditions:*

1. *Δ_p bounded by Δ (e.g. linear temperature schedule) and*

$$m_p \geq \frac{log(\frac{e^{\Delta.osc(V)}(1+a)}{a}) e^{\overline{\Delta V}(k_0).\beta_p}}{\delta}$$

2. *$\Delta_p \longrightarrow 0$ (decreasingly) and $m_p \geq \left(osc(V).\Delta_p + log(\frac{1}{a}) \right) \frac{e^{\overline{\Delta V}(k_0).\beta_p}}{\delta}$*

then for all $\varepsilon > 0$, and all $\varepsilon' < \varepsilon$, the proportion $p_n^N(\varepsilon)$ of particles (ζ_n^i) so that $V(\zeta_n^i) \geq V_{\min} + \varepsilon$ satisfies the inequality:

$$\forall y \geq 0, \quad \mathbb{P}\left(p_n^N(\varepsilon) \geq \frac{e^{-\beta_n(\varepsilon-\varepsilon')}}{m_{\varepsilon'}} + \frac{r_i^\star N + r_j^\star y}{N^2} \right) \leq e^{-y}$$

where $(i,j) = (1,2)$ in the case of bounded Δ_p (taken with $M = e^{\Delta.osc(V)}$) and $(i,j) = (3,4)$ in the second one.

We then clearly distinguish two error terms: the first one, $\left(\frac{e^{-\beta_n(\varepsilon-\varepsilon')}}{m_{\varepsilon'}} \right)$, estimating the Boltzmann-Gibbs measure's concentration around V's minimizers, and the second one, $\left(\frac{r_i^\star N + r_j^\star y}{N^2} \right)$, estimating the occupation measure's concentration around this Boltzmann-Gibbs theoretical measure. More than providing tunings which ensure convergence, this last concentraion inequality explicits the relative impact of other parameters, such as probabilistic precision y, threshold t on the proportion of particles possibly out of the area of interest, final temperature β_n or population size N. A simple equation, deduced from this last theorem, such as $\left(\frac{e^{-\beta_n(\varepsilon-\varepsilon')}}{m_{\varepsilon'}} = \frac{r_i^\star N + r_j^\star y}{N^2} = \frac{t}{2} \right)$ may be applied to the global tuning of an Interacting Simulated Annealing algorithm, which is generally a difficult task.

5 Conclusion

It is instructive to compare the estimates of Theorem 3 with the performance analysis of the traditional simulated annealing model (*abbreviate SA*). Firstly, most of the literature on SA models is concerned with the weak convergence of the law of the random states of the algorithm. When the initial temperature of the scheme is greater than some critical value, using a logarithmic cooling schedule, it is well known that the probability for the random state to be in the global extrema levels

tends to 1, as the time parameter tends to ∞. The cooling schedule presented in Theorem 3 is again a logarithmic one. In contrast to the SA model, Theorem 3 allows to quantify the performance analysis of the ISA model in terms of uniform concentration inequalities, that does not depend on a critical parameter.

Like most rigorous and non-asymptotic tuning theorems, our results may not be applied directly. They highlight important principles (as uniform accessibility of all the state space after a given number of mutations) and the type of dependence in some parameters. Otherwise, to our knowledge, our work presently provides the most explicit non-asymptotic ISA convergence results, at least in the case $|E| = \infty$.

Nevertheless, the models we studied involve a deterministic sequence β_n, while choosing the sequence of increments $\Delta_n = (\beta_n - \beta_{n-1})$ in advance can cause computational problems. In practice, adaptive strategies, where increment Δ_n depends on the current set of particles ζ_{n-1}, are of common use in the engineering community (see for instance [8, 19, 22, 24, 25]). In a forthcoming paper (see [20] for a preliminary version), we try to adapt the present work to analyze one of these adaptive tuning strategies.

References

1. Assaraf, R., Caffarel, M.: A pedagogical introduction to quantum Monte Carlo. In: Mathematical Models and Methods for Ab Initio Quantum Chemistry, pp. 45–73. Springer, Berlin/Heidelberg (2000).
2. Assaraf, R., Caffarel, M., Khelif, A.: Diffusion Monte Carlo with a fixed number of walkers. Phys. Rev. E, **61**, 4566 (2000)
3. Bartoli, N., Del Moral, P.: Simulation & Algorithmes Stochastiques. Cépaduès éditons (2001)
4. Cappé, O., Moulines, E., Rydén, T.: Inference in Hidden Markov Models. Springer, New York (2005)
5. Cérou, F., Del Moral, P., Furon, T., Guyader, A.: Sequential Monte Carlo for rare event estimation. Stat. Comput. **22**, 795–808 (2012).
6. Cérou, F., Del Moral, P., Guyader, A.: A nonasymptotic variance theorem for unnormalized Feynman-Kac particle models. Ann. Inst. Henri Poincaré, Probab. Stat. **47**, 629–649 (2011)
7. Chopin, N.: Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. Ann. Statist. **32**, 2385–2411 (2004)
8. Clapp, T.: Statistical methods in the processing of communications data. Ph.D. thesis, Cambridge University Engineering Department (2000)
9. Dawson D.A., Del Moral P.: Large deviations for interacting processes in the strong topology. In: Duchesne, P., Rémillard, B. (eds.) Statistical Modeling and Analysis for Complex Data Problem, pp. 179–209. Springer US (2005)
10. Del Moral, P.: Nonlinear filtering: interacting particle solution. Markov Process. Related Fields **2**, 555–579 (1996)
11. Del Moral, P.: Feynman-Kac Formulae. Genealogical and Interacting Particle Approximations. Series: Probability and Applications, 575p. Springer, New York (2004)
12. Del Moral, P., Doucet, A., Jasra, A.: Sequential Monte Carlo samplers. J. R. Stat. Soc. Ser. B **68**, 411–436 (2006)
13. Del Moral P., Guionnet A.: Large deviations for interacting particle systems: applications to non linear filtering problems. Stochastic Process. Appl. **78**, 69–95 (1998)

14. Del Moral P., Guionnet A.: A central limit theorem for non linear filtering using interacting particle systems. Ann. Appl. Probab. **9**, 275–297 (1999)
15. Del Moral, P., Guionnet, A.: On the stability of interacting processes with applications to filtering and genetic algorithms. Ann. Inst. Henri Poincaré **37**, 155–194 (2001)
16. Del Moral P., Ledoux M.: On the convergence and the applications of empirical processes for interacting particle systems and nonlinear filtering. J. Theoret. Probab. **13**, 225–257 (2000)
17. Del Moral, P., Miclo, L.: Branching and interacting particle systems approximations of Feynman-Kac formulae with applications to nonlinear filtering. In: Séminaire de Probabilités XXXIV. Lecture Notes in Mathematics, vol. 1729, pp. 1–145. Springer, Berlin (2000)
18. Del Moral, P., Rio, E.: Concentration inequalities for mean field particle models. Ann. Appl. Probab. **21**, 1017–1052 (2011)
19. Deutscher, J., Blake, A., Reid, I.: Articulated body motion capture by annealed particle filtering. In: IEEE Conference on Computer Vision and Pattern Recognition, Hilton Head, vol. 2, pp. 126–133 (2000)
20. Giraud, F., Del Moral, P.: Non-asymptotic analysis of adaptive and annealed Feynman-Kac particle models (2012). arXiv math.PR/12095654
21. Hetherington, J.H.: Observations on the statistical iteration of matrices. Phys. Rev. A **30**, 2713–2719 (1984)
22. Jasra, A., Stephens, D., Doucet, A., Tsagaris, T.: Inference for Lévy driven stochastic volatility models via adaptive sequential Monte Carlo. Scand. J. Stat. **38**, 1–22 (2011)
23. Künsch, H.R.: Recursive Monte-Carlo filters: algorithms and theoretical analysis. Ann. Statist. **33**, 1983–2021 (2005)
24. Minvielle, P., Doucet, A., Marrs, A., Maskell, S.: A Bayesian approach to joint tracking and identification of geometric shapes in video squences. Image and Visaion Computing **28**, 111–123 (2010)
25. Schäfer, C., Chopin, N.: Sequential Monte Carlo on large binary sampling spaces. Stat. Comput. **23**, 163–184 (2013)
26. Schweizer, N.: Non-asymptotic error bounds for sequential MCMC and stability of Feynman-Kac propagators. Working Paper, University of Bonn (2012)
27. Whiteley, N.: Sequential Monte Carlo samplers: error bounds and insensitivity to initial conditions. Working Paper, University of Bristol (2011)

Lower Error Bounds for Randomized Multilevel and Changing Dimension Algorithms

Michael Gnewuch

Abstract We provide lower error bounds for randomized algorithms that approximate integrals of functions depending on an unrestricted or even infinite number of variables. More precisely, we consider the infinite-dimensional integration problem on weighted Hilbert spaces with an underlying anchored decomposition and arbitrary weights. We focus on randomized algorithms and the randomized worst case error. We study two cost models for function evaluation which depend on the number of active variables of the chosen sample points. Multilevel algorithms behave very well with respect to the first cost model, while changing dimension algorithms and also dimension-wise quadrature methods, which are based on a similar idea, can take advantage of the more generous second cost model. We prove the first non-trivial lower error bounds for randomized algorithms in these cost models and demonstrate their quality in the case of product weights. In particular, we show that the randomized changing dimension algorithms provided in Plaskota and Wasilkowski (J Complex 27:505–518, 2011) achieve convergence rates arbitrarily close to the optimal convergence rate.

1 Introduction

Integrals over functions with an unbounded or infinite number of variables are important in physics, quantum chemistry or in quantitative finance, see, e.g., [8, 25] and the references therein. In the last few years a large amount of research was dedicated to design new algorithms as, e.g., multilevel and changing dimension algorithms or dimension-wise quadrature methods, to approximate such integrals

M. Gnewuch (✉)
School of Mathematics and Statistics, University of New South Wales, Sydney, Australia

Fachbereich Mathematik, Technische Universität Kaiserslautern, Kaiserslautern, Germany
e-mail: gnewuch@mathematik.uni-kl.de

J. Dick et al. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2012*,
Springer Proceedings in Mathematics & Statistics 65,
DOI 10.1007/978-3-642-41095-6_18, © Springer-Verlag Berlin Heidelberg 2013

efficiently. Multilevel algorithms were introduced by Heinrich in [12] in the context of integral equations and by Giles in [8] in the context of stochastic differential equations. Changing dimension algorithms were introduced by Kuo et al. in [16] in the context of infinite-dimensional integration on weighted Hilbert spaces and dimension-wise quadrature methods were introduced by Griebel and Holtz in [11] for multivariate integration; changing dimension and dimension-wise quadrature algorithms are based on a similar idea.

Here we want to study the complexity of numerical integration on a weighted Hilbert space of functions with infinitely many variables as it has been done in [2, 4, 6, 9, 13, 14, 16, 18, 21]. The Hilbert space we consider here allows for a so-called anchored function space decomposition. For a motivation of this specific function space setting and connections to problems in stochastics and mathematical finance see, e.g., [13, 18]. We derive lower error bounds for randomized algorithms to solve the infinite-dimensional integration problem. Notice that the complexity of integration problems is not as well understood in the randomized setting as in the deterministic setting (where only deterministic algorithms are permitted and the deterministic worst case error is considered), see, e.g., the comments in [20, p. 487].

Our error bounds are for the randomized worst case error and are expressed in terms of the cost of a randomized algorithm. Here we solely take account of function evaluations, i.e., the cost of function sampling, and disregard other cost as, e.g., combinatorial cost. Notice that this makes the statements of our lower bounds only stronger. To evaluate the cost of sampling, we consider two sampling models: the nested subspace sampling model (introduced in [5], where it was called variable subspace sampling model) and the unrestricted subspace sampling model (introduced in [16]). Our lower error bounds are the first non-trivial lower bounds in these settings, cf. also the comments in the introductions of [13, 21]. Due to space restrictions, we do not provide new constructive upper error bounds. For the same reason we refer for a formal definition of multilevel algorithms and changing dimension algorithms for the infinite-dimensional integration problem on weighted Hilbert spaces to [9, 13, 18] and [16, 21], respectively. In this article we only compare our lower bounds to already known upper bounds. In particular, we show that the randomized changing dimension algorithms provided for product weights in [21] achieve convergence rates arbitrarily close to the optimal rate of convergence.

Let us mention that similar general lower error bounds for infinite-dimensional integration on weighted Hilbert spaces are provided in [6] in the determistic setting for the anchored decomposition and in [4] in the randomized setting for underlying ANOVA-type decompositions (to treat the latter decompositions, a technically more involved analysis is necessary).

The article is organized as follows: In Sect. 2 the setting we want to study is introduced. In Sect. 3 we prove new lower bounds for the complexity of randomized algorithms for solving the infinite-dimensional integration problem on weighted Hilbert spaces. In Sect. 3.1 we provide the most general form of our lower bounds which is valid for arbitrary weights. In Sect. 3.2 we state the simplified form of our lower bounds for specific classes of weights. In particular, we show in Sect. 3.2.1

that the randomized changing dimension algorithms from [21] are essentially optimal.

2 The General Setting

In this section we describe the precise setting we want to study. A comparison with the (slightly different) settings described in the papers [9,16,21] is provided in paper [10]; we refer to the same paper and to [13,14] for rigorous proofs of the statements on the function spaces we consider here.

2.1 Notation

For $n \in \mathbb{N}$ we denote the set $\{1, \ldots, n\}$ by $[n]$. If u is a finite set, then its size is denoted by $|u|$. We put $\mathscr{U} := \{u \subset \mathbb{N} \,|\, |u| < \infty\}$. We use the common Landau symbol O, and additionally for non-negative functions $f, g : [0, \infty) \to [0, \infty)$ the notation $f = \Omega(g)$ if $g = O(f)$.

2.2 The Function Spaces

As spaces of integrands of infinitely many variables, we consider *reproducing kernel Hilbert spaces*; our standard reference for these spaces is [1].

We start with univariate functions. Let $D \subseteq \mathbb{R}$ be a Borel measurable set of $\mathbb{R}$ and let $k : D \times D \to \mathbb{R}$ be a measurable reproducing kernel with anchor $a \in D$, i.e., $k(a, a) = 0$. This implies $k(\cdot, a) \equiv 0$. We assume that k is non-trivial, i.e., $k \neq 0$. We denote the reproducing kernel Hilbert space with kernel k by $H = H(k)$ and its scalar product and norm by $\langle \cdot, \cdot \rangle_H$ and $\| \cdot \|_H$, respectively. Additionally, we denote its norm unit ball by $B(k)$. We use corresponding notation for other reproducing kernel Hilbert spaces. If g is a constant function in H, then the reproducing property implies $g = g(a) = \langle g, k(\cdot, a) \rangle_H = 0$. Let ρ be a probability measure on D. We assume that

$$M := \int_D k(x, x)\, \rho(\mathrm{d}x) < \infty. \tag{1}$$

For arbitrary $\mathbf{x}, \mathbf{y} \in D^{\mathbb{N}}$ and $u \in \mathscr{U}$ we define

$$k_u(\mathbf{x}, \mathbf{y}) := \prod_{j \in u} k(x_j, y_j),$$

where by convention $k_\emptyset \equiv 1$. The Hilbert space with reproducing kernel k_u will be denoted by $H_u = H(k_u)$. Its functions depend only on the coordinates $j \in u$. If it is convenient for us, we identify H_u with the space of functions defined on D^u determined by the kernel $\prod_{j\in u} k(x_j, y_j)$, and write $f_u(\mathbf{x}_u)$ instead of $f_u(\mathbf{x})$ for $f_u \in H_u$, $\mathbf{x} \in D^{\mathbb{N}}$, and $\mathbf{x}_u := (x_j)_{j\in u} \in D^u$. For all $f_u \in H_u$ and $\mathbf{x} \in D^{\mathbb{N}}$ we have

$$f_u(\mathbf{x}) = 0 \quad \text{if } x_j = a \text{ for some } j \in u. \tag{2}$$

This property yields an *anchored decomposition* of functions, see, e.g., [17].

Let now $\boldsymbol{\gamma} = (\gamma_u)_{u\in\mathscr{U}}$ be weights, i.e., a family of non-negative numbers. We assume

$$\sum_{u\in\mathscr{U}} \gamma_u M^{|u|} < \infty. \tag{3}$$

Let us define the domain $\mathscr{X}$ of functions of infinitely many variables by

$$\mathscr{X} := \left\{ \mathbf{x} \in D^{\mathbb{N}} \;\middle|\; \sum_{u\in\mathscr{U}} \gamma_u k_u(\mathbf{x}, \mathbf{x}) < \infty \right\}.$$

Let μ be the infinite product probability measure of ρ on $D^{\mathbb{N}}$. Due to our assumptions we have $\mu(\mathscr{X}) = 1$, see [13, Lemma 1] or [10]. We define

$$K_{\boldsymbol{\gamma}}(\mathbf{x}, \mathbf{y}) := \sum_{u\in\mathscr{U}} \gamma_u k_u(\mathbf{x}, \mathbf{y}) \quad \text{for all } \mathbf{x}, \mathbf{y} \in \mathscr{X}.$$

$K_{\boldsymbol{\gamma}}$ is well-defined and, since $K_{\boldsymbol{\gamma}}$ is symmetric and positive semi-definite, it is a reproducing kernel on $\mathscr{X} \times \mathscr{X}$, see [1]. We denote the corresponding reproducing kernel Hilbert space by $\mathscr{H}_{\boldsymbol{\gamma}} = H(K_{\boldsymbol{\gamma}})$, its norm by $\|\cdot\|_{\boldsymbol{\gamma}}$, and its norm unit ball by $B_{\boldsymbol{\gamma}} = B(K_{\boldsymbol{\gamma}})$. For the next lemma see [14, Corollary 5] or [10].

Lemma 1. *The space $\mathscr{H}_{\boldsymbol{\gamma}}$ consists of all functions $f = \sum_{u\in\mathscr{U}} f_u$, $f_u \in H_u$, that have a finite norm*

$$\|f\|_{\boldsymbol{\gamma}} = \left(\sum_{u\in\mathscr{U}} \gamma_u^{-1} \|f_u\|_{H_u}^2 \right)^{1/2}.$$

For $u \in \mathscr{U}$ let P_u denote the orthogonal projection $P_u : \mathscr{H}_{\boldsymbol{\gamma}} \to H_u$, $f \mapsto f_u$ onto H_u. Then each $f \in \mathscr{H}_{\boldsymbol{\gamma}}$ has a unique representation

$$f = \sum_{u\in\mathscr{U}} f_u \quad \text{with } f_u = P_u(f) \in H_u, u \in \mathscr{U}.$$

2.3 Infinite-Dimensional Integration

For a given $f \in \mathscr{H}_{\boldsymbol{\gamma}}$ we want to approximate the integral

$$I(f) := \int_{\mathscr{X}} f(\mathbf{x})\, \mu(\mathrm{d}\mathbf{x}).$$

Due to (3), I is continuous on $\mathscr{H}_{\boldsymbol{\gamma}}$ and its representer $h \in \mathscr{H}_{\boldsymbol{\gamma}}$ is given by

$$h(\mathbf{x}) = \int_{\mathscr{X}} K_{\boldsymbol{\gamma}}(\mathbf{x}, \mathbf{y}) \mu(\mathrm{d}\mathbf{y}).$$

The operator norm of the integration functional I is given by

$$\|I\|^2_{\mathscr{H}_{\boldsymbol{\gamma}}} = \|h\|^2_{\boldsymbol{\gamma}} = \sum_{u \in \mathscr{U}} \gamma_u C_0^{|u|} < \infty, \tag{4}$$

where

$$C_0 := \int_D \int_D k(x, y)\, \rho(\mathrm{d}x)\, \rho(\mathrm{d}y).$$

We have $C_0 \le M$. We assume that I is non-trivial, i.e., that $C_0 > 0$ and $\gamma_u > 0$ for at least one $u \in \mathscr{U}$. For $u \in \mathscr{U}$ and $f \in \mathscr{H}_{\boldsymbol{\gamma}}$ we define $I_u := I \circ P_u$, i.e.,

$$I_u(f) = \int_{D^u} f_u(\mathbf{x}_u)\, \rho^u(\mathrm{d}\mathbf{x}_u),$$

and the representer h_u of I_u in $\mathscr{H}_{\boldsymbol{\gamma}}$ is given by $h_u(\mathbf{x}) = P_u(h)(\mathbf{x})$. Thus we have

$$h = \sum_{u \in \mathscr{U}} h_u \quad \text{and} \quad I(f) = \sum_{u \in \mathscr{U}} I_u(f_u) \quad \text{for all } f \in \mathscr{H}_{\boldsymbol{\gamma}}.$$

Furthermore,

$$\|h_u\|^2_{\boldsymbol{\gamma}} = \gamma_u C_0^{|u|} \quad \text{for all } u \in \mathscr{U}. \tag{5}$$

2.4 Randomized Algorithms, Cost Models, and Errors

As in [13], we assume that algorithms for approximation of $I(f)$ have access to the function f via a subroutine ("oracle") that provides values $f(\mathbf{x})$ for points $\mathbf{x} \in D^{\mathbb{N}}$. For convenience we define $f(\mathbf{x}) = 0$ for $\mathbf{x} \in D^{\mathbb{N}} \setminus \mathscr{X}$.

We now present the cost models introduced in [5] and [16]: In both models we only consider the cost of function evaluations. To define the cost of a function evaluation, we fix a monotone increasing function $\$: \mathbb{N}_0 \to [1, \infty]$. For our lower

error bounds we will later assume that $\$(v) = \Omega(v^s)$ for some $s \geq 0$. For each $v \in \mathscr{U}$ we define the finite-dimensional affine subspace $\mathscr{X}_{v,a}$ of $\mathscr{X}$ by

$$\mathscr{X}_{v,a} := \{\mathbf{x} \in D^{\mathbb{N}} \,|\, x_j = a \ \text{ for all } \ j \in \mathbb{N} \setminus v\}.$$

In the *nested subspace sampling model* function evaluations can be done in a sequence of affine subspaces $\mathscr{X}_{v_1,a} \subset \mathscr{X}_{v_2,a} \subset \cdots$ for a strictly increasing sequence $\mathrm{v} = (v_i)_{i\in\mathbb{N}}$ of sets $\emptyset \neq v_i \in \mathscr{U}$, and the cost for a function evaluation in some point $\mathbf{x}$ is given by the cost function

$$c_{\mathrm{v},a}(\mathbf{x}) := \inf\{\$(|v_i|) \,|\, \mathbf{x} \in \mathscr{X}_{v_i,a}\}, \tag{6}$$

where we use the standard convention that $\inf \emptyset = \infty$. Let C_{nest} denote the set of all cost functions of the form (6). The nested subspace sampling model was introduced in [5], where it was actually called "variable subspace sampling model". We prefer the name "nested subspace sampling model" to clearly distinguish this model from the following cost model, which is even more "variable":

In the *unrestricted subspace sampling model* we are allowed to sample in any subspace $\mathscr{X}_{u,a}, u \in \mathscr{U}$, without any restriction. The cost for each function evaluation is given by the cost function

$$c_a(\mathbf{x}) := \inf\{\$(|u|) \,|\, \mathbf{x} \in \mathscr{X}_{u,a},\ u \in \mathscr{U}\}. \tag{7}$$

Put $C_{\mathrm{unr}} := \{c_a\}$. The unrestricted subspace sampling model was introduced in [16], where it did not get a particular name. Obviously, the unrestricted subspace sampling model is more generous than the nested subspace sampling model.

We consider randomized algorithms for integration of functions $f \in \mathscr{H}_{\boldsymbol{\gamma}}$. For a formal definition we refer to [5, 19, 23, 24]. Here we require that a randomized algorithm Q yields for each $f \in \mathscr{H}_{\boldsymbol{\gamma}}$ a square-integrable random variable $Q(f)$. (More precisely, a randomized algorithm Q is a map $Q : \Omega \times \mathscr{H}_{\boldsymbol{\gamma}} \to \mathbb{R}$, $(\omega, f) \mapsto Q(\omega, f)$, where Ω is some suitable probability space. But for convenience we will not specify the underlying probability space Ω and suppress any reference to Ω or $\omega \in \Omega$. We use this convention also for other random variables.) Furthermore, we require that the cost of a randomized algorithm Q, which is defined to be the sum of the cost of all function evaluations, is a random variable, which may depend on the function f. That is why we denote this random variable by $\mathrm{cost}_c(Q, f)$, c the relevant cost function from C_{nest} or C_{unr}.

We denote the class of all randomized algorithms for numerical integration on $\mathscr{H}_{\boldsymbol{\gamma}}$ that satisfy the very mild requirements stated above by $\mathscr{A}^{\mathrm{ran}}$. For unrestricted subspace sampling we additionally consider a subclass $\mathscr{A}^{\mathrm{res}}$ of $\mathscr{A}^{\mathrm{ran}}$. We say that an algorithm $Q \in \mathscr{A}^{\mathrm{ran}}$ is in $\mathscr{A}^{\mathrm{res}}$ if there exist an $n \in \mathbb{N}_0$ and sets $v_1, \ldots, v_n \in \mathscr{U}$ such that for every $f \in \mathscr{H}_{\boldsymbol{\gamma}}$ the algorithm Q performs exactly n function evaluations of f, where the ith sample point is taken from $\mathscr{X}_{v_i,a}$, and $\mathbb{E}(\mathrm{cost}_{c_a}(Q, f)) = \sum_{i=1}^{n} \$(|v_i|)$. If additionally $|v_1|, \ldots, |v_n| \leq \omega$ for some $\omega \in \mathbb{N}$, we say that

$Q \in \mathscr{A}^{\mathrm{res}-\omega}$. Notice that the classes $\mathscr{A}^{\mathrm{ran}}$, $\mathscr{A}^{\mathrm{res}}$, and $\mathscr{A}^{\mathrm{res}-\omega}$ contain in particular non-linear and adaptive algorithms.

The *worst case cost* of a randomized algorithm Q on a class of integrands F is

$$\mathrm{cost}_{\mathrm{nest}}(Q, F) := \inf_{c \in C_{\mathrm{nest}}} \sup_{f \in F} \mathbb{E}(\mathrm{cost}_c(Q, f))$$

in the nested subspace sampling model and

$$\mathrm{cost}_{\mathrm{unr}}(Q, F) := \sup_{f \in F} \mathbb{E}(\mathrm{cost}_{c_a}(Q, f))$$

in the unrestricted subspace sampling model. The *randomized (worst case) error* $e(Q, F)$ of approximating the integration functional I by Q on F is defined as

$$e(Q, F) := \left(\sup_{f \in F} \mathbb{E}\left((I(f) - Q(f))^2 \right) \right)^{1/2}.$$

For $N \in \mathbb{R}$, $\mathrm{mod} \in \{\mathrm{nest}, \mathrm{unr}\}$, and $* \in \{\mathrm{ran}, \mathrm{res}, \mathrm{res}-\omega\}$ let us define the corresponding *N th minimal error* by

$$e^*_{\mathrm{mod}}(N, F) := \inf\{e(Q, F) \mid Q \in \mathscr{A}^* \text{ and } \mathrm{cost}_{\mathrm{mod}}(Q, F) \le N\}.$$

2.5 Strong Polynomial Tractability

Let $\omega \in \mathbb{N}$, $\mathrm{mod} \in \{\mathrm{nest}, \mathrm{unr}\}$, and $* \in \{\mathrm{ran}, \mathrm{res}, \mathrm{res}-\omega\}$. The ε-complexity of the infinite-dimensional integration problem I on $\mathscr{H}_{\boldsymbol{\gamma}}$ in the considered cost model with respect to the class of admissable randomized algorithms $\mathscr{A}^*$ is the minimal cost among all admissable algorithms, whose randomized errors are at most ε, i.e.,

$$\mathrm{comp}^*_{\mathrm{mod}}(\varepsilon, B_{\boldsymbol{\gamma}}) := \inf\left\{\mathrm{cost}_{\mathrm{mod}}(Q, B_{\boldsymbol{\gamma}}) \mid Q \in \mathscr{A}^* \text{ and } e(Q, B_{\boldsymbol{\gamma}}) \le \varepsilon\right\}. \quad (8)$$

The integration problem I is said to be *strongly polynomially tractable* if there are non-negative constants C and p such that

$$\mathrm{comp}^*_{\mathrm{mod}}(\varepsilon, B_{\boldsymbol{\gamma}}) \le C\, \varepsilon^{-p} \qquad \text{for all } \varepsilon > 0. \quad (9)$$

The *exponent of strong polynomial tractability* is given by

$$p^*_{\mathrm{mod}} = p^*_{\mathrm{mod}}(\boldsymbol{\gamma}) := \inf\{p \mid p \text{ satisfies (9) for some } C > 0\}.$$

Essentially, $1/p^*_{\mathrm{mod}}$ is the *convergence rate* of the Nth minimal error $e^*_{\mathrm{mod}}(N, B_{\boldsymbol{\gamma}})$. In particular, we have for all $p > p^*_{\mathrm{mod}}$ that $e^*_{\mathrm{mod}}(N, B_{\boldsymbol{\gamma}}) = O(N^{-1/p})$.

3 Lower Bounds

We start in Sect. 3.1 by proving lower bounds for general weights. In Sect. 3.2 we show how these bounds simplify for several specific classes of weights.

3.1 Results for General Weights

Let $\boldsymbol{\gamma} = (\gamma_u)_{u\in\mathscr{U}}$ be a given family of weights that satisfy (3). We denote by $\hat{\boldsymbol{\gamma}}$ the family of weights defined by

$$\hat{\gamma}_u := \gamma_u C_0^{|u|} \quad \text{for all } u \in \mathscr{U}. \tag{10}$$

Recall that (3) implies $\sum_{u\in\mathscr{U}} \hat{\gamma}_u < \infty$. Weights $\boldsymbol{\gamma}$ are called *finite-order weights of order* ω if there exists an $\omega \in \mathbb{N}$ such that $\gamma_u = 0$ for all $u \in \mathscr{U}$ with $|u| > \omega$. Finite-order weights were introduced in [7] for spaces of functions with a finite number of variables. The following definition is taken from [9].

Definition 1. For weights $\boldsymbol{\gamma}$ and $\sigma \in \mathbb{N}$ let us define the *cut-off weights* of order σ

$$\boldsymbol{\gamma}^{(\sigma)} = (\gamma_u^{(\sigma)})_{u\in\mathscr{U}} \quad \text{via} \quad \gamma_u^{(\sigma)} = \begin{cases} \gamma_u & \text{if } |u| \le \sigma, \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

Cut-off weights of order σ are in particular finite-order weights of order σ. Let us denote by $u_1(\sigma), u_2(\sigma), \ldots$, the distinct non-empty sets $u \in \mathscr{U}$ with $\gamma_u^{(\sigma)} > 0$ for which $\hat{\gamma}_{u_1(\sigma)}^{(\sigma)} \ge \hat{\gamma}_{u_2(\sigma)}^{(\sigma)} \ge \cdots$. Let us put $u_0(\sigma) := \emptyset$. We can make the same definitions for $\sigma = \infty$; then we have obviously $\boldsymbol{\gamma}^{(\infty)} = \boldsymbol{\gamma}$. For convenience we will often suppress any reference to σ in the case where $\sigma = \infty$. For $\sigma \in \mathbb{N} \cup \{\infty\}$ let us define

$$\mathrm{decay}_{\boldsymbol{\gamma},\sigma} := \sup\left\{ p \in \mathbb{R} \,\middle|\, \lim_{j\to\infty} \hat{\gamma}_{u_j(\sigma)}^{(\sigma)} j^p = 0 \right\}.$$

Due to assumption (3) the weights we consider always satisfy $\mathrm{decay}_{\boldsymbol{\gamma},\sigma} \ge 1$ for all $\sigma \in \mathbb{N} \cup \{\infty\}$. The following definition is from [9].

Definition 2. For $\sigma \in \mathbb{N} \cup \{\infty\}$ let $t_\sigma^* \in [0,\infty]$ be defined as

$$t_\sigma^* := \inf\left\{ t \ge 0 \,|\, \exists\, C_t > 0 \ \forall\, v \in \mathscr{U} : |\{i \in \mathbb{N} \,|\, u_i(\sigma) \subseteq v\}| \le C_t |v|^t \right\}.$$

Let $\sigma \in \mathbb{N}$. Since $|u_i(\sigma)| \le \sigma$ for all $i \in \mathbb{N}$, we have obviously $t_\sigma^* \le \sigma$. On the other hand, if we have an infinite sequence $(u_j(\sigma))_{j\in\mathbb{N}}$, it is not hard to verify that $t_\sigma^* \ge 1$, see [9].

For $v_1, \ldots, v_n \in \mathscr{U}$ we use the short hand $\{v_i\}$ for $(v_i)_{i=1}^n$. Put $v := \cup_{i=1}^n v_i$ and define the mapping

$$\Psi_{\{v_i\},a} := \sum_{j\,;\,\exists i\in[n]:u_j\subseteq v_i} P_{u_j}. \tag{12}$$

The operator $\Psi_{\{v_i\},a}$ is the orthogonal projection of $\mathscr{H}_{\boldsymbol{\gamma}}$ onto the subspace

$$H_{\{v_i\},a} := \sum_{j\,;\,\exists i\in[n]:u_j\subseteq v_i} H_{u_j}.$$

Put

$$\mathrm{b}_{\{v_i\},a} := \sup_{f\in B_{\boldsymbol{\gamma}}} |I(f) - I(\Psi_{\{v_i\},a} f)|.$$

In the case where $n = 1$ and $v = v_1$, we simply write $\Psi_{v,a}$ and $\mathrm{b}_{v,a}$. In that case we have, due to (2),

$$(\Psi_{v,a}(f))(\mathbf{x}) = f(\mathbf{x}_v;\mathbf{a}) \quad \text{for all } f\in\mathscr{H}_{\boldsymbol{\gamma}} \text{ and } \mathbf{x}\in\mathscr{X}, \tag{13}$$

where the jth component of $(\mathbf{x}_v;\mathbf{a})$ is defined by

$$(\mathbf{x}_v;\mathbf{a})_j := \begin{cases} x_j & \text{if } j\in v,\\ a & \text{otherwise.}\end{cases}$$

Lemma 2. *Let $v_1,\dots,v_n\in\mathscr{U}$. Then*

$$\mathrm{b}^2_{\{v_i\},a} = \sum_{j\,;\,\forall i\in[n]:u_j\not\subseteq v_i} \hat{\gamma}_{u_j}.$$

Proof. Let $h_{\{v_i\},a}$ denote the representer of the continuous functional $I\circ\Psi_{\{v_i\},a}$. Due to (12) we get

$$h_{\{v_i\},a} = \sum_{j\,;\,\exists i\in[n]:u_j\subseteq v_i} h_{u_j}.$$

Since $h - h_{\{v_i\},a}$ is the representer of $I - I\circ\Psi_{\{v_i\},a}$ in $\mathscr{H}_{\boldsymbol{\gamma}}$, we obtain with (5)

$$\mathrm{b}^2_{\{v_i\},a} = \|h-h_{\{v_i\},a}\|^2_{\boldsymbol{\gamma}} = \left\| \sum_{j\,;\,\forall i\in[n]:u_j\not\subseteq v_i} h_{u_j} \right\|^2_{\boldsymbol{\gamma}} = \sum_{j\,;\,\forall i\in[n]:u_j\not\subseteq v_i} \|h_{u_j}\|^2_{\boldsymbol{\gamma}} = \sum_{j\,;\,\forall i\in[n]:u_j\not\subseteq v_i} \hat{\gamma}_{u_j}.$$

This completes the proof. □

Lemma 3. *Let $\theta\in(1/2,1]$ and $v_1,\dots,v_n\in\mathscr{U}$. Let the randomized algorithm $Q\in\mathscr{A}^{\mathrm{ran}}$ satisfy $\mathbb{P}(Q(f) = Q(\Psi_{\{v_i\},a} f))\ge\theta$ for all $f\in B_{\boldsymbol{\gamma}}$. Then*

$$e(Q,B_{\boldsymbol{\gamma}})^2 \ge (2\theta-1)\mathrm{b}^2_{\{v_i\},a}.$$

Proof. Since $\Psi_{\{v_i\},a}$ is an orthogonal projection, we have for all $f \in B_{\boldsymbol{\gamma}}$ that $g := f - \Psi_{\{v_i\},a} f \in B_{\boldsymbol{\gamma}}$. Furthermore, $\Psi_{\{v_i\},a}(g) = \Psi_{\{v_i\},a}(-g) = 0$. Let $A := \{Q(g) = Q(-g)\}$. Then $\{Q(g) = Q(\Psi_{\{v_i\},a} g)\} \cap \{Q(-g) = Q(\Psi_{\{v_i\},a}(-g))\} \subseteq A$, and hence $\mathbb{P}(A) \geq 2\theta - 1$. Therefore

$$\begin{aligned} e(Q, B_{\boldsymbol{\gamma}})^2 &\geq \max\left\{\mathbb{E}\left((I(g) - Q(g))^2\right), \mathbb{E}\left((I(-g) - Q(-g))^2\right)\right\} \\ &\geq \max\left\{\int_A (I(g) - Q(g))^2 \, \mathrm{d}\mathbb{P}, \int_A (I(-g) - Q(-g))^2 \, \mathrm{d}\mathbb{P}\right\} \\ &\geq (2\theta - 1)|I(g)|^2 = (2\theta - 1)|I(f) - I(\Psi_{\{v_i\},a} f)|^2. \end{aligned}$$

Hence $e(Q, B_{\boldsymbol{\gamma}})^2 \geq (2\theta - 1) \sup_{f \in B_{\boldsymbol{\gamma}}} |I(f) - I(\Psi_{\{v_i\},a} f)|^2 = (2\theta - 1)\mathrm{b}^2_{\{v_i\},a}$. □

Further Assumptions. We assume for the rest of this article that $\$(v) = \Omega(v^s)$ for some $s \in (0, \infty)$. Furthermore, we assume that $\gamma_{\{1\}} > 0$ and that there exists an $\alpha > 0$ such that for univariate integration in $H(\gamma_{\{1\}} k)$ the Nth minimal error satisfies

$$e^{\mathrm{ran}}(N, B(\gamma_{\{1\}} k)) = \Omega(N^{-\alpha/2}). \tag{14}$$

(Note that in the univariate situation the nested and the unrestricted subspace sampling models are equal; that is why we suppress the reference to unr or nest.) Since $B(\gamma_{\{1\}} k) \subseteq B_{\boldsymbol{\gamma}}$, assumption (14) implies in particular

$$e^{\mathrm{ran}}_{\mathrm{nest}}(N, B_{\boldsymbol{\gamma}}) = \Omega(N^{-\alpha/2}) \quad \text{and} \quad e^{\mathrm{res}-\omega}_{\mathrm{unr}}(N, B_{\boldsymbol{\gamma}}) \geq e^{\mathrm{res}}_{\mathrm{unr}}(N, B_{\boldsymbol{\gamma}}) = \Omega(N^{-\alpha/2}). \tag{15}$$

Theorem 1. *Consider the nested subspace sampling model. To achieve strong polynomial tractability for the class $\mathscr{A}^{\mathrm{ran}}$ it is necessary that the weights satisfy*

$$\mathrm{decay}_{\boldsymbol{\gamma},\sigma} > 1 \quad \textit{for all } \sigma \in \mathbb{N}. \tag{16}$$

If (16) holds, we have

$$p^{\mathrm{ran}}_{\mathrm{nest}} \geq \max\left\{\frac{2}{\alpha}, \sup_{\sigma \in \mathbb{N}} \frac{2s/t^*_\sigma}{\mathrm{decay}_{\boldsymbol{\gamma},\sigma} - 1}\right\}.$$

As we will see in Sect. 3.2, for product weights and finite-order weights condition (16) is equivalent to $\mathrm{decay}_{\boldsymbol{\gamma}} = \mathrm{decay}_{\boldsymbol{\gamma},\infty} > 1$.

Proof. Let $Q \in \mathscr{A}^{\mathrm{ran}}$ with $\mathrm{cost}_{\mathrm{nest}}(Q, B_{\boldsymbol{\gamma}}) \leq N$. Then there exists an increasing sequence $\mathrm{v} = (v_i)_{i \in \mathbb{N}}$, $\emptyset \neq v_i \in \mathscr{U}$, such that $\mathbb{E}(\mathrm{cost}_{\mathrm{v},a}(Q, f)) \leq N + 1$ for every $f \in B_{\boldsymbol{\gamma}}$. Let m be the largest integer satisfying $\$(|v_m|) \leq 4(N + 1)$. This implies for all $f \in B_{\boldsymbol{\gamma}}$ that $\mathbb{P}(Q(f) = Q(\Psi_{v_m,a} f)) \geq 3/4$, see (13). Due to Lemmas 2 and 3 we get

$$e(Q, B_{\boldsymbol{\gamma}})^2 \geq \frac{1}{2} \sum_{j\,;\, u_j \not\subseteq v_m} \hat{\gamma}_{u_j}.$$

Let us now assume that $\boldsymbol{\gamma}$ are weights of finite order ω. Then we get for $t > t^*_\omega$ and a suitable constant $C_t > 0$

$$\tau_m := |\{j \mid u_j \subseteq v_m\}| \leq C_t |v_m|^t = O(N^{t/s}),$$

since $N = \Omega(|v_m|^s)$. Hence we get for $p_\omega > \mathrm{decay}_{\boldsymbol{\gamma},\omega} = \mathrm{decay}_{\boldsymbol{\gamma}} \geq 1$

$$e(Q, B_{\boldsymbol{\gamma}})^2 \geq \frac{1}{2} \sum_{j=\tau_m+1}^{\infty} \hat{\gamma}_{u_j} = \Omega(\tau_m^{1-p_\omega}) = \Omega\left(N^{\frac{t}{s}(1-p_\omega)}\right).$$

For general weights $\boldsymbol{\gamma}$, $\sigma \in \mathbb{N}$, and cut-off weights $\boldsymbol{\gamma}^{(\sigma)}$ we have $e(Q, B_{\boldsymbol{\gamma}}) \geq e(Q, B_{\boldsymbol{\gamma}^{(\sigma)}})$, see also [9, Remark 3.3]. Since the cut-off weights $\boldsymbol{\gamma}^{(\sigma)}$ are weights of finite order σ, we get for all $p_\sigma > \mathrm{decay}_{\boldsymbol{\gamma},\sigma}$ and $t_\sigma > t^*_\sigma$

$$e(Q, B_{\boldsymbol{\gamma}})^2 = \Omega\left(N^{\frac{t_\sigma}{s}(1-p_\sigma)}\right). \tag{17}$$

Since (15) holds, the inequality for the exponent of tractability follows.

Now assume that the infinite-dimensional integration problem I is strongly polynomially tractable. Let $\sigma \in \mathbb{N}$. Then we get from inequality (17) that $p_\sigma \geq 1 + 2s/(t_\sigma\, p^{\mathrm{ran}}_{\mathrm{nest}})$. Hence

$$\mathrm{decay}_{\boldsymbol{\gamma},\sigma} \geq 1 + \frac{2s/t^*_\sigma}{p^{\mathrm{ran}}_{\mathrm{nest}}}.$$

Thus we have $\mathrm{decay}_{\boldsymbol{\gamma},\sigma} > 1$ for all $\sigma \in \mathbb{N}$. □

Theorem 2. *Consider the unrestricted subspace sampling model. To achieve strong polynomial tractability for the class $\mathscr{A}^{\mathrm{res}}$ it is necessary that the weights satisfy*

$$\mathrm{decay}_{\boldsymbol{\gamma},\sigma} > 1 \quad \textit{for all } \sigma \in \mathbb{N}.$$

If this is the case, we have

$$p^{\mathrm{res}}_{\mathrm{unr}} \geq \max\left\{\frac{2}{\alpha}, \sup_{\sigma\in\mathbb{N}} \frac{2\min\{1, s/t^*_\sigma\}}{\mathrm{decay}_{\boldsymbol{\gamma},\sigma} - 1}\right\}.$$

Proof. Let $Q \in \mathscr{A}^{\mathrm{res}}$ have $\mathrm{cost}_{\mathrm{unr}}(Q, B_{\boldsymbol{\gamma}}) \leq N$. Then there exists an $n \in \mathbb{N}$ and coordinate sets $v_1, \ldots, v_n$ such that Q selects randomly n sample points $\mathbf{x}_1 \in \mathscr{X}_{v_1,a}, \ldots, \mathbf{x}_n \in \mathscr{X}_{v_n,a}$ and $\sum_{i=1}^n \$(|v_i|) \leq N$. Since $Q(f) = Q(\Psi_{\{v_i\},a} f)$ for all $f \in B_{\boldsymbol{\gamma}}$, we obtain from Lemmas 2 and 3

$$e(Q, B_{\boldsymbol{\gamma}})^2 \geq \sum_{j;\, \forall i \in [n]: u_j \not\subseteq v_i} \hat{\gamma}_{u_j}.$$

Let us first assume that $\boldsymbol{\gamma}$ are weights of finite order ω. Then we get with Jensen's inequality for $t > t^*_\omega$ and suitable constants $C_t, c > 0$

$$\begin{aligned} |\{j \mid \exists i \in [n] : u_j \subseteq v_i\}| &\leq \sum_{i=1}^{n} |\{j \mid u_j \subseteq v_i\}| \leq \sum_{i=1}^{n} C_t |v_i|^t \\ &\leq C_t \left(\sum_{i=1}^{n} |v_i|^s \right)^{1/\min\{1,s/t\}} \leq C_t (cN)^{1/\min\{1,s/t\}}. \end{aligned}$$

Hence we obtain for $S := \lceil C_t (cN)^{1/\min\{1,s/t\}} \rceil$ and all $p_\omega > \mathrm{decay}_{\boldsymbol{\gamma},\omega}$

$$e(Q, B_{\boldsymbol{\gamma}})^2 \geq \sum_{j=S+1}^{\infty} \hat{\gamma}_{u_j} = \Omega(S^{1-p_\omega}) = \Omega\left(N^{\frac{1-p_\omega}{\min\{1,s/t\}}} \right).$$

If we have general weights $\boldsymbol{\gamma}$, then we obtain for $\sigma \in \mathbb{N}$ and the cut-off weights $\boldsymbol{\gamma}^{(\sigma)}$ that $e(Q, B(K_{\boldsymbol{\gamma}})) \geq e(Q, B(K_{\boldsymbol{\gamma}^{(\sigma)}}))$. From this and (15) the inequality for the exponent of tractability follows. Similarly as in the proof of Theorem 1, the necessity of condition (16) is easily established. □

Theorem 3. *Let $\omega \in \mathbb{N}$ be fixed. We have for the exponent of tractability $p^{\mathrm{res}-\omega}_{\mathrm{unr}}$ in the unrestricted subspace sampling setting*

$$p^{\mathrm{res}-\omega}_{\mathrm{unr}} \geq \max\left\{ \frac{2}{\alpha}, \sup_{\sigma \in \mathbb{N}} \frac{2}{\mathrm{decay}_{\boldsymbol{\gamma},\sigma} - 1} \right\}.$$

Proof. We follow the lines of the proof of Theorem 2, and use the same notation. The difference is that this time Q selects randomly n sample points $\mathbf{x}_1 \in \mathscr{X}_{v_1,a}, \ldots, \mathbf{x}_n \in \mathscr{X}_{v_n,a}$, where $|v_i| \leq \omega$ for all $i \in [n]$, and that we therefore can make the estimate $|\{j \mid \exists i \in [n] : u_j \subseteq v_i\}| \leq 2^\omega n = O(N)$, since $\$(|v_i|) \geq 1$ for all $i \in [n]$ by definition of the function $\$$. Hence we get this time for $p > \mathrm{decay}_{\boldsymbol{\gamma}}$

$$e(Q, B_{\boldsymbol{\gamma}})^2 \geq \sum_{j=2^\omega n+1}^{\infty} \hat{\gamma}_{u_j} = \Omega(N^{1-p}).$$

This completes the proof. □

A comparison of Theorems 2 and 3 indicates that there are cost functions and classes of finite-order weights for which changing dimension algorithms cannot achieve convergence rates that are arbitrarily close to the optimal rate. Let us recall that for weights of finite order ω, changing dimension algorithms as defined

in [16, Proof of Theorem 5] would only use sample points from sample spaces $\mathcal{X}_{u,a}$ with $|u| \leq \omega$; see also the comment at the beginning of Sect. 4 in [16]. Examples of such cost functions and finite-order weights would be $\$(k) = \Omega(k^s)$ and lexicographically-ordered weights of order $\omega > s$, see Sect. 3.2.3. (A similar observation was made for the deterministic setting, see [9, Theorem 3.2 and Sect. 3.2.3].)

3.2 Results for Specific Classes of Weights

Here we consider some example classes of weights and show how our bounds from Sect. 3.1 simplify in those settings.

3.2.1 Product Weights and Finite-Product Weights

Definition 3. Let $(\gamma_j)_{j\in\mathbb{N}}$ be a sequence of non-negative real numbers satisfying $\gamma_1 \geq \gamma_2 \geq \ldots$. With the help of this sequence we define for $\omega \in \mathbb{N} \cup \{\infty\}$ weights $\boldsymbol{\gamma} = (\gamma_u)_{u\in\mathscr{U}}$ by

$$\gamma_u = \begin{cases} \prod_{j\in u} \gamma_j & \text{if } |u| \leq \omega, \\ 0 & \text{otherwise,} \end{cases} \tag{18}$$

where we use the convention that the empty product is 1. In the case where $\omega = \infty$, we call such weights *product weights*, in the case where ω is finite, we call them *finite-product weights of order* ω.

Product weights were introduced in [22] and have been studied extensively since then. Finite-product weights were considered in [9]. Observe that for $\sigma \in \mathbb{N}$ the cut-off weights $\boldsymbol{\gamma}^{(\sigma)}$ of product weights $\boldsymbol{\gamma}$ are finite-product weights of order σ.

Let us assume that $\boldsymbol{\gamma}$ are product or finite-product weights. As shown in [9, Lemma 3.8], we have

$$\mathrm{decay}_{\boldsymbol{\gamma},1} = \mathrm{decay}_{\boldsymbol{\gamma},\sigma} \quad \text{for all } \sigma \in \mathbb{N} \cup \{\infty\}. \tag{19}$$

(Actually, [9, Lemma 3.8] states identity (19) only for all $\sigma \in \mathbb{N}$. But the proof provided in [9] is also valid for the case $\sigma = \infty$.) In particular, we see that for strong polynomial tractability with respect to the nested subspace sampling model and the class $\mathscr{A}^{\mathrm{ran}}$ or with respect to the unrestricted subspace sampling model and the class $\mathscr{A}^{\mathrm{res}}$ it is necessary that $\mathrm{decay}_{\boldsymbol{\gamma}} = \mathrm{decay}_{\boldsymbol{\gamma},\infty} > 1$. Since $t_1^* = 1$, we obtain from Theorems 1 to 3

$$p^{\mathrm{ran}}_{\mathrm{nest}} \geq \max\left\{ \frac{2}{\alpha}, \frac{2s}{\mathrm{decay}_{\boldsymbol{\gamma},1} - 1} \right\}, \tag{20}$$

and

$$p^{\mathrm{res}}_{\mathrm{unr}} \geq \max\left\{\frac{2}{\alpha}\,,\,\frac{2\min\{1,s\}}{\mathrm{decay}_{\boldsymbol{\gamma},1}-1}\right\}\,,\qquad p^{\mathrm{res}-\omega}_{\mathrm{unr}} \geq \max\left\{\frac{2}{\alpha}\,,\,\frac{2}{\mathrm{decay}_{\boldsymbol{\gamma},1}-1}\right\}\,. \tag{21}$$

Note that the bounds for finite-product weights are the same as for product weights.

Remark 1. For *product and order-dependent (POD) weights* $(\gamma_u)_{u\in\mathscr{U}}$, which were recently introduced in [15] and are of the form

$$\gamma_u = \Gamma_{|u|}\prod_{j\in u}\gamma_j\,,\qquad \text{where } \gamma_1\geq\gamma_2\geq\cdots\geq 0, \text{ and } \Gamma_0=\Gamma_1=1,\Gamma_2,\Gamma_3,\ldots\geq 0,$$

identity (19) still holds; for a proof see [6]. Thus (20) and (21) are also valid for POD weights. Product and finite-product weights are, in particular, POD weights.

Let us assume that there exist constants $c, C, \kappa > 0$, $\alpha_1 \geq 0$, and $\alpha_2 \in [0,1]$ such that for all $\emptyset \neq u \in \mathscr{U}$ and all $n \geq 1$ there exist randomized algorithms $Q_{n,u}$ using for all $f_u \in H_u$ at most n function values of f_u with

$$\mathbb{E}\left(|I_u(f_u)-Q_{n,u}(f_u)|^2\right) \leq \frac{c\,C^{|u|}}{(n+1)^{\kappa}}\left(1+\frac{\ln(n+1)}{(|u|-1)^{\alpha_2}}\right)^{\alpha_1(|u|-1)^{\alpha_2}}\|f_u\|^2_{H_u}.$$

Note that necessarily $\kappa \leq \alpha$. Let us further assume that $\mathrm{decay}_{\boldsymbol{\gamma},1} > 1$ and the cost function \$ satisfies $\$(\nu) = O(e^{r\nu})$ for some $r \geq 0$. Plaskota and Wasilkowski proved in [21] with the help of randomized changing dimension algorithms that

$$p^{\mathrm{res}}_{\mathrm{unr}} \leq \max\left\{\frac{2}{\kappa},\frac{2}{\mathrm{decay}_{\boldsymbol{\gamma},1}-1}\right\}\,.$$

Hence, if $\Omega(\nu) = \$(\nu) = O(e^{r\nu})$ and $\kappa = \alpha$, our lower bound (21) is sharp and the randomized algorithms from [21] exhibit essentially the optimal convergence rate.

Let us consider a more specific example, namely the case where $D = [0,1]$, k is the Wiener kernel given by $k(x,y) = \min\{x,y\}$, and ρ is the restriction of the Lebesgue measure to D. In this case the anchor a is zero. The space $H(k)$ is the Sobolev space anchored at zero, and its elements are the absolutely continuous functions f with $f(0) = 0$ and square-integrable first weak derivative. It is known that $\kappa = 3 = \alpha$, see [26, Example 1 and Proposition 3] (or [21, Example 2]) and [19, Sect. 2.2.9, Proposition 1]. Thus the upper bound from [21] and our lower bound (21) establish for $\Omega(\nu) = \$(\nu) = O(e^{r\nu})$ that

$$p^{\mathrm{res}}_{\mathrm{unr}} = \max\left\{\frac{2}{3},\frac{2}{\mathrm{decay}_{\boldsymbol{\gamma},1}-1}\right\}\,.$$

For the same specific example Hickernell et al. showed for the case $\$(\nu) = \Theta(\nu)$ with the help of multilevel Monte Carlo algorithms that

$$p^{\mathrm{ran}}_{\mathrm{nest}} \le \max\left\{2, \frac{2}{\mathrm{decay}_{\boldsymbol{\gamma},1} - 1}\right\} \quad \text{for } \mathrm{decay}_{\boldsymbol{\gamma},1} > 1,$$

see [13, Corollary 5]. Hence

$$p^{\mathrm{ran}}_{\mathrm{nest}} = \frac{2}{\mathrm{decay}_{\boldsymbol{\gamma},1} - 1} \quad \text{for } \mathrm{decay}_{\boldsymbol{\gamma},1} \in (1, 2].$$

Similarly as in the deterministic setting [6, 9, 18] or in the randomized setting with underlying ANOVA-type decomposition [4], our lower bound for $p^{\mathrm{ran}}_{\mathrm{nest}}$ is sharp for sufficiently large $\mathrm{decay}_{\boldsymbol{\gamma},1}$. This may be proved by using multilevel algorithms based on the integration algorithms provided in [26, Sect. 4] (cf. also [21, Sect. 3.2]) or on scrambled quasi-Monte Carlo algorithms similar to those discussed in [3], but providing a rigorous proof for this claim is beyond the scope of this article.

3.2.2 Finite-Intersection Weights

We restate Definition 3.5 from [9].

Definition 4. Let $\rho \in \mathbb{N}$. Finite-order weights $(\gamma_u)_{u \in \mathscr{U}}$ are called *finite-intersection weights* with *intersection degree* at most ρ if we have

$$|\{v \in \mathscr{U} \mid \gamma_v > 0\,,\, u \cap v \neq \emptyset\}| \le 1 + \rho \quad \text{for all } u \in \mathscr{U} \text{ with } \gamma_u > 0. \tag{22}$$

For finite-intersection weights of order ω it was observed in [9] that $t^*_\sigma = 1$ for all $\sigma \in \mathbb{N}$, resulting in the lower bounds (20) and (21) with $\mathrm{decay}_{\boldsymbol{\gamma},1}$ replaced by $\mathrm{decay}_{\boldsymbol{\gamma},\omega}$.

3.2.3 Lexicographically-Ordered Weights

To every set $u \subset \mathbb{N}$ with $|u| = \ell$ we may assign a word $\varphi(u) := i_1 i_2 \ldots i_\ell$, where for $j \in [\ell]$ the number i_j is the jth-largest element of u. On the set of all finite words over the alphabet $\mathbb{N}$ we have the natural lexicographical order $\prec_{\mathrm{lex}}$, where by convention the empty word should be the first (or "smallest") word.

Definition 5. We call weights $\boldsymbol{\gamma}$ *lexicographically-ordered weights* of *order* ω if $\gamma_\emptyset = 1$, $\gamma_u > 0$ for all $u \subset \mathbb{N}$ with $|u| \le \omega$, and

$$\varphi(u_i) \prec_{\mathrm{lex}} \varphi(u_j) \quad \text{for all } i, j \in \mathbb{N} \text{ satisfying } i < j\,.$$

Lexicographically-ordered weights were introduced in [9]. Their properties complement the properties of the other classes of weights considered before, see [9] for more information. For lexicographically-ordered weights of order ω we have $t_\sigma^* = \min\{\sigma, \omega\}$. Hence we get from Theorems 1 to 3 the lower bounds

$$p_{\mathrm{nest}}^{\mathrm{ran}} \geq \max\left\{\frac{2}{\alpha}, \frac{2s/\omega}{\mathrm{decay}_{\boldsymbol{\gamma},\omega} - 1}\right\},$$

and

$$p_{\mathrm{unr}}^{\mathrm{res}} \geq \max\left\{\frac{2}{\alpha}, \frac{2\min\{1, s/\omega\}}{\mathrm{decay}_{\boldsymbol{\gamma},\omega} - 1}\right\}, \qquad p_{\mathrm{unr}}^{\mathrm{res}-\omega} \geq \max\left\{\frac{2}{\alpha}, \frac{2}{\mathrm{decay}_{\boldsymbol{\gamma},\omega} - 1}\right\}.$$

The lower bounds indicate that in the setting where $\omega > s$ and $\mathrm{decay}_{\boldsymbol{\gamma},\omega}$ is only moderate, changing dimension algorithms (which are algorithms of the class $\mathscr{A}^{\mathrm{res}-\omega}$) cannot achieve the optimal rate of convergence and can be outperformed by multilevel algorithms (which can exhibit a behavior similar to the lower bound for $p_{\mathrm{nest}}^{\mathrm{ran}}$ above). For the deterministic setting and the Wiener kernel $k(x, y) = \min\{x, y\}$ on $D = [0, 1]$ this was rigorously proved in [9] by lower bounds for changing dimension algorithms and upper bounds for multilevel algorithms, see [9, Theorem 3.2 and Sect. 3.2.3].

Acknowledgements The author gratefully acknowledges support by the German Science Foundation (DFG) under grant GN 91/3-1 and by the Australian Research Council (ARC).

References

1. Aronszajn, N.: Theory of reproducing kernels. Trans. Am. Math. Soc. **68**, 337–404 (1950)
2. Baldeaux, J.: Scrambled polynomial lattice rules for infinite-dimensional integration. In: Plaskota, L., Woźniakowski, H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2010, pp. 255–263. Springer, Berlin/Heidelberg (2012)
3. Baldeaux, J., Dick, J.: A construction of polynomial lattice rules with small gain coefficients. Numer. Math. **119**, 271–297 (2011)
4. Baldeaux, J., Gnewuch, M.: Optimal randomized multilevel algorithms for infinite-dimensional integration on function spaces with ANOVA-type decomposition. arXiv:1209.0882v1 [math.NA] (2012, preprint)
5. Creutzig, J., Dereich, S., Müller-Gronbach, T., Ritter, K.: Infinite-dimensional quadrature and approximation of distributions. Found. Comput. Math. **9**, 391–429 (2009)
6. Dick, J., Gnewuch, M.: Infinite-dimensional integration in weighted Hilbert spaces: anchored decompositions, optimal deterministic algorithms, and higher order convergence. arXiv:1210.4223v1 [math.NA], (2012, preprint)
7. Dick, J., Sloan, I.H., Wang, X., Woźniakowski, H.: Good lattice rules in weighted Korobov spaces with general weights. Numer. Math. **103**, 63–97 (2006)
8. Giles, M.B.: Multilevel Monte Carlo path simulation. Oper. Res. **56**, 607–617 (2008)

9. Gnewuch, M.: Infinite-dimensional integration on weighted Hilbert spaces. Math. Comp. **81**, 2175–2205 (2012)
10. Gnewuch, M., Mayer, S., Ritter, K.: On weighted Hilbert spaces and integration of functions of infinitely many variables. J. Complexity. doi:10.1016/j.jco.2013.05.004 (2013)
11. Griebel, M., Holtz, M.: Dimension-wise integration of high-dimensional functions with applications to finance. J. Complexity **26**, 455–489 (2010)
12. Heinrich, S.: Monte Carlo complexity of global solution of integral equations. J. Complexity **14**, 151–175 (1998)
13. Hickernell, F.J., Müller-Gronbach, T., Niu, B., Ritter, K.: Multi-level Monte Carlo algorithms for infinite-dimensional integration on $\mathbb{R}^{\mathbb{N}}$. J. Complexity **26**, 229–254 (2010)
14. Hickernell, F.J., Wang, X.: The error bounds and tractability of quasi-Monte Carlo algorithms in infinite dimension. Math. Comp. **71**, 1641–1661 (2001)
15. Kuo, F.Y., Schwab, C., Sloan, I.H.: Quasi-Monte Carlo finite element methods for a class of elliptic partial differential equations with random coefficients. SIAM J. Numer. Anal. **50**, 3351–3374 (2012)
16. Kuo, F.Y., Sloan, I.H., Wasilkowski, G.W., Woźniakowski, H.: Liberating the dimension. J. Complexity **26**, 422–454 (2010)
17. Kuo, F.Y., Sloan, I.H., Wasilkowski, G.W., Woźniakowski, H.: On decompositions of multivariate functions. Math. Comp. **79**, 953–966 (2010)
18. Niu, B., Hickernell, F.J., Müller-Gronbach, T., Ritter, K.: Deterministic multi-level algorithms for infinite-dimensional integration on $\mathbb{R}^{\mathbb{N}}$. J. Complexity **27**, 331–351 (2011)
19. Novak, E.: Deterministic and Stochastic Error Bounds in Numerical Analysis. Lecture Notes in Mathematics, vol. 1349. Springer, Berlin (1988)
20. Novak, E., Woźniakowski, H.: Tractability of Multivariate Problems, vol. II. European Mathematical Society, Zürich (2010)
21. Plaskota, L., Wasilkowski, G.W.: Tractability of infinite-dimensional integration in the worst case and randomized setting. J. Complexity **27**, 505–518 (2011)
22. Sloan, I.H., Woźniakowski, H.: When are quasi-Monte Carlo algorithms efficient for high dimensional integrals? J. Complexity **14**, 1–33 (1998)
23. Traub, J.F., Wasilkowski, G.W., Woźniakowski, H.: Information-Based Complexity. Academic, New York (1988)
24. Wasilkowski, G.W.: Randomization for continuous problems. J. Complexity **5**, 195–218 (1989)
25. Wasilkowski, G.W., Woźniakowski, H.: On tractability of path integration. J. Math. Phys. **37**, 2071–2088 (1996)
26. Wasilkowski, G.W., Woźniakowski, H.: The power of standard information for multivariate approximation in the randomized setting. Math. Comp. **76**, 965–988 (2007)

A Non-empirical Test on the Second to the Sixth Least Significant Bits of Pseudorandom Number Generators

Hiroshi Haramoto, Makoto Matsumoto, Takuji Nishimura, and Yuki Otsuka

Abstract Lagged Fibonacci generators are widely used random number generators. Some implementations discard the least significant bit of their outputs, because their weight distribution has a strong deviation. But the degree of the improvement is unclear.

In this paper, we give a method to compute the weight distribution of the n-th least significant bit of several pseudo random number generators for arbitrary n, generalizing the weight discrepancy test which was possible only for $n = 1$. The method is based on the MacWilliams identity over $\mathbb{Z}/2^n$, and predicts the sample size for which the bit stream fails in a statistical test. These tests are effective to lagged Fibonacci generators such as `random()` in BSD-C library. For example, we show that the second least significant bit of `random()` will be rejected if the sample size is of order 10^4, while the sixth bit will be rejected for the sample size around 10^7.

H. Haramoto (✉)
Faculty of Education, Ehime University, Ehime, Japan
e-mail: haramoto@ehime-u.ac.jp

M. Matsumoto · Y. Otsuka
Department of Mathematics, Graduate School of Science, Hiroshima University, Higashi-Hiroshima, Japan
e-mail: m-mat@math.sci.hiroshima-u.ac.jp

T. Nishimura
Department of Mathematical Sciences, Faculty of Science, Yamagata University, Yamagata, Japan
e-mail: nisimura@sci.kj.yamagata-u.ac.jp

J. Dick et al. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2012*,
Springer Proceedings in Mathematics & Statistics 65,
DOI 10.1007/978-3-642-41095-6_19, © Springer-Verlag Berlin Heidelberg 2013

1 Introduction

Consider a pseudo random number generator (PRNG) based on the following recursion

$$\mathbf{x}_{i+31} = \mathbf{x}_{i+28} + \mathbf{x}_i \bmod 2^{32} \quad (i = 0, 1, 2, \ldots) \tag{1}$$

from randomly chosen initial values of $(\mathbf{x}_0, \ldots, \mathbf{x}_{31})$. Such a generator is called a lagged-Fibonacci generator, and still widely used, for example in the C-library of BSD UNIX (called `random()`). It is known that its least significant bit has a strong deviation, hence some applications discard the least significant bit.

In order to investigate the effect of this discarding, one usually applies statistical tests to the output sequence, which compute the deviation of the empirical distribution of the sequence from the theoretical distribution. Many statistical tests for PRNGs are proposed [4, 6].

However, statistical tests have some problems. One lies in the instability, i.e. every experiment shows a different test result, which depends on the choice of the initial value. Another is that new defects may be discovered when the sample-size is increased in accordance with the increase of computational power. Therefore, we want a test whose results are independent of the initial value.

The second and the third authors introduced a theoretical test on the distribution of 1's and 0's in the bits of the sequence, named *weight discrepancy test* [11]. This is not an empirical test but a figure of merit defined on the generator, like the spectral test [1] or the k-distribution test [2]. The weight discrepancy test gives the sample size for which the generator is rejected by the weight distribution test [8], which is a classical empirical test equivalent to a random walk test. However, their method is limited to linear generators based on $\mathbb{Z}/2$. Thus, it is applicable only to the least significant bit for the generators such as (1), since for $n > 1$, the n-th least significant bit in (1) is not $\mathbb{Z}/2$-linear.

The aim of this paper is to generalize their method to $\mathbb{Z}/2^n$ for arbitrary integer n, and which enables us to apply the weight discrepancy test to the n-th bit. In our method, the computation time increases at least exponentially with respect to n, but still we can execute the weight discrepancy test on the second to the sixth least significant bits of two lagged Fibonacci generators. The results show the degree of improvement by such discarding. Section 2 reviews the weight discrepancy test. Section 3 deals with how the output of a PRNG is regarded as a random variable, and how to compute its distribution. Section 4 shows the results of the weight discrepancy test, and compares it with the results of the weight distribution test, which is a statistical test.

2 χ^2-Discrepancy and Weight Discrepancy Tests

This section recalls the weight discrepancy test. See [11] for details. Consider a set of events $\{0, 1, 2, \ldots, \nu\}$. Let $(p_k)_{k=0,1,\ldots,\nu}$ be a probability distribution on $\{0, 1, 2, \ldots, \nu\}$, i.e.,

$$0 < p_k \leq 1 \text{ and } \sum_{k=0}^{\nu} p_k = 1.$$

Let $(q_k)_{k=0,1,\ldots,\nu}$ be another probability distribution.

Definition 1. We define the χ^2-discrepancy δ between the two distributions (p_k) and (q_k) by

$$\delta = \sum_{k=0}^{\nu} (q_k - p_k)^2 / p_k.$$

This value measures the amount of discrepancy between two distributions. Suppose that we make a null hypothesis that one trial of a probabilistic event conforms to the distribution p_k, and the different trials are independently identically distributed. To test this null hypothesis, we perform N trials, and count the number Y_k of occurrences of each event $k \in \{0, 1, \ldots, \nu\}$. The χ^2-value $\mathscr{X}$ of this experiment is defined as

$$\mathscr{X} := \sum_{k=0}^{\nu} (Y_k - Np_k)^2 / Np_k.$$

It is known that $\mathscr{X}$ approximately conforms to the χ^2-distribution with ν degrees of freedom under the null hypothesis, if Np_k is large enough for each k. Let X be a random variable with χ^2-distribution with ν degrees of freedom.

Recall that the (left) p-value p corresponding to the observed χ^2-value $\mathscr{X}$ is defined by

$$p = \Pr(X < \mathscr{X}).$$

If the p-value is too high like >0.99, then we reject the null hypothesis with significance level >0.99. (The null hypothesis is also rejected if the p-value is too small, but this is not treated in this manuscript.)

Suppose that the above null hypothesis is not correct, and the actual distribution is $(q_k)_{k=0,\ldots,\nu}$. It is shown that the expectation of $\mathscr{X}$ is approximated by

$$E(\mathscr{X}) \sim \nu + N\delta.$$

Thus, if $N\delta$ is large, then $E(\mathscr{X})$ is large, and $\mathscr{X}$ tends to be large so that the p-value is large. For $0 < p < 1$, we define N_p as the value N which satisfies

$$p = \Pr(X < E(\chi)).$$

Roughly speaking, this means that if we take N_p samples, then on average the χ^2-value of the χ^2-test by the wrong null hypothesis gives a p-value around p. To make explanations simple, we call $N_{.75}$ the safe sample size which corresponds to $p = 0.75$ (thus the sample tends to pass the χ^2-test with p-value around 0.75) and the risky sample size $N_{.99}$ which corresponds to $p = 0.99$ (and hence tends to fail the test).

Theorem 1. *Let ν be moderately large, say $\nu \geq 5$. (For $\nu < 5$, we need to consult a table of χ^2-distribution.)*

1. The safe sample size $N_{.75}$ is approximated by

$$N_{.75} \approx \frac{\sqrt{2\nu}\, x_p + \frac{2}{3}(x_p^2 - 1)}{\delta} \quad \textit{for} \quad x_p = 0.674.$$

2. The risky sample size $N_{.99}$ is approximated by

$$N_{.99} \approx \frac{\sqrt{2\nu}\, x_p + \frac{2}{3}(x_p^2 - 1)}{\delta} \quad \textit{for} \quad x_p = 2.33.$$

By this theorem, the χ^2-discrepancy δ provides us with an estimation of the sample size for which χ^2-test reveals the defect of the generator, as well as the size for which it does not.

Now return to a pseudo random number generator G, which consists of a state space S and the output symbol O. We can regard the output sequence as a function of the initial state. We assume that the initial state is uniformly randomly chosen from S. Then, the output sequence is a random variable. In particular, we treat the case where $O = \{0, 1\}$, namely, a pseudorandom bit generator. We fix an integer m. For an initial state $s \in S$, consider the output sequence $(x_0(s), x_1(s), \ldots, x_{m-1}(s))$ of length m, and let $W(s)$ be the weight of this sequence, namely, the number of 0's among the m bits (usually the number of 1's, but by symmetry it does not matter for testing randomness). If the bit sequence is really random, the weight should conform to the binomial distribution, whereas $W(s)$ with a random choice of s may not. As usual in applying χ^2-test, we categorize $\{0, 1, \ldots, m\}$ to several intervals so that each interval has enough probability for approximation by χ^2-distribution.

Definition 2. A weight discrepancy test means to obtain the χ^2-discrepancy δ for the weight $W(s)$ for the generator under the null hypothesis that it is the binomial distribution $B(m, 1/2)$, and to compute the safe sample size $N_{0.75}$ and the risky sample size $N_{0.99}$.

This test is not an empirical test but a theoretical test similar to the spectral test [1] or to the k-distribution test [2], since the value is determined by the PRNG only, independent of the choice of the initial state.

3 Computing the Probability by Enumeration

In this section, we shall mainly explain the weight discrepancy test on the second least significant bits of the sequence generated by (1).

Let $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \ldots$ be a pseudo random sequence generated by (1). The least significant two bits satisfy

$$\mathbf{x}_{i+31} = \mathbf{x}_{i+28} + \mathbf{x}_i \bmod 2^2 \quad (i = 0, 1, 2, \ldots).$$

Assume that exactly m outputs are consumed in one simulation. Let W denote the number of 0's appeared in the second least significant bits in $\mathbf{x}_0, \mathbf{x}_1, \ldots \mathbf{x}_{m-1}$, i.e., the weight.

In this case, we may regard the initial state space as $S = (\mathbb{Z}/4)^{31}$, the initial state is $(\mathbf{x}_0, \cdots, \mathbf{x}_{30})$, and the map

$$O_G : S \to (\mathbb{Z}/4)^m \quad (\mathbf{x}_0, \cdots, \mathbf{x}_{30}) \mapsto (\mathbf{x}_0, \cdots, \mathbf{x}_{m-1})$$

is an abelian group homomorphism.

Let $\mathbf{w} = (w_1, w_2, \ldots, w_m) \in (\mathbb{Z}/4)^m$. For $0 \leq i \leq 3$, we denote the number of i's among $w_1, w_2, \ldots, w_m$ by $\mathrm{wt}_i(\mathbf{w})$. Under a random choice of the initial state, wt_i are random variables.

Let us denote by q_k the probability $\Pr(W = k)$, where $0 \leq k \leq m$. By the definition, we have

$$q_k = \frac{\{s \in S : \mathrm{wt}_0(O_G(s)) + \mathrm{wt}_1(O_G(s)) = k\}}{\#S}.$$

For an ideal PRNG, q_k should be the binomial distribution $p_k := \binom{m}{k}/2^m$, but the generator (1) has a huge deviation for $m > \dim(S)$.

Let $C := O_G(S) \subset (\mathbb{Z}/4)^m$ be the image of S. Because S is a group and $O_G : S \to (\mathbb{Z}/4)^m$ is a group homomorphism, a uniform choice from S gives a uniform choice from $C := O_G(S)$. Thus, we have

$$q_k = \frac{\{\mathbf{w} \in C : \mathrm{wt}_0(\mathbf{w}) + \mathrm{wt}_1(\mathbf{w}) = k\}}{\#C}.$$

For non-negative integers j_0, j_1, j_2, j_3 with $j_0 + j_1 + j_2 + j_3 = m$, we define

$$A_{j_0,j_1,j_2,j_3} := \#\{\mathbf{w} \in C \ : \ \mathrm{wt}_i(\mathbf{w}) = j_i \ \ (i = 0, 1, 2, 3)\},$$

and we call the list of integers A_{j_0,j_1,j_2,j_3} the weight enumeration of C. Using this, q_k can be written as

$$q_k = \frac{\sum_{j_0+j_1=k} A_{j_0,j_1,j_2,j_3}}{\#C}.$$

In general, the computation of the list of integers A_{j_0,j_1,j_2,j_3} is NP-complete [14], so it is intractable. However, if its orthogonal space $C^\perp$ is not too large, the MacWilliams identity gives a solution as follows. We define a standard inner product on $(\mathbb{Z}/4)^m$ by $\langle (x_1,\cdots,x_m),(y_1,\cdots,y_m)\rangle := \sum_{i=1}^m x_i y_i \in \mathbb{Z}/4$, and the orthogonal space $C^\perp \subset (\mathbb{Z}/4)^m$ by

$$C^\perp := \{\mathbf{y} \in (\mathbb{Z}/4)^m : \langle \mathbf{x}, \mathbf{y}\rangle = 0 \text{ for all } \mathbf{x} \in C\}.$$

The polynomial

$$W_C(X_0,X_1,X_2,X_3) := \sum A_{j_0,j_1,j_2,j_3} X_0^{j_0} X_1^{j_1} X_2^{j_2} X_3^{j_3},$$

where $\sum$ means the sum of all tuples (j_0, j_1, j_2, j_3) of non-negative integers with $j_0+j_1+j_2+j_3 = m$, in the indeterminants X_0, X_1, X_2, and X_3 is called the weight enumerator polynomial of C.

Theorem 2 (the MacWilliams identity, $\mathbb{Z}/4$-version [5]).

$$\begin{aligned} &W_C(X_0,X_1,X_2,X_3) = \\ &\frac{1}{\#C^\perp} W_{C^\perp}(X_0+X_1+X_2+X_3,\ \ X_0+\sqrt{-1}X_1-X_2-\sqrt{-1}X_3, \\ &\qquad X_0-X_1+X_2-X_3,\ \ X_0-\sqrt{-1}X_1-X_2+\sqrt{-1}X_3) \end{aligned}$$

This identity enables us to compute the weight enumeration of C from that of $C^\perp$. If $C^\perp$ is not too large, then we can compute the weight enumeration of $C^\perp$ by an exhaustive check, and the weight enumerator polynomial $W_{C^\perp}(X_0,X_1,X_2,X_3)$, and by the substitution described on the right hand side in the theorem, we get $W_C(X_0,X_1,X_2,X_3)$. Then the coefficient of $X_0^{j_0}X_1^{j_1}X_2^{j_2}X_3^{j_3}$ is A_{j_0,j_1,j_2,j_3}. Take the standard basis $\mathbf{e}_1 = (1,0,\ldots,0),\ldots,\mathbf{e}_{31} = (0,0,\ldots,1)$ of $S = (\mathbb{Z}/4)^{31}$. Their images $O_G(\mathbf{e}_1),\ldots,O_G(\mathbf{e}_{31})$ generate C by linear combination. A vector $\mathbf{x}$ in $(\mathbb{Z}/4)^m$ belongs to $C^\perp$ if and only if $\mathbf{x}$ is perpendicular to $O_G(\mathbf{e}_i)$ for $1 \le i \le 31$. This is the solution of a linear equation described by a $(31\times m)$ matrix. By Gaussian elimination, we obtain a basis of this kernel, and we can enumerate vectors in $C^\perp$ as all possible linear combinations of this basis.

For the weight discrepancy test on the second least significant bit, we don't need each value of A_{j_0,j_1,j_2,j_3} but only the sums $\sum_{j_0+j_1=k} A_{j_0,j_1,j_2,j_3}$ for $0 \le k \le m$. Substituting X_0 for X_1 and X_2 for X_3 in the weight enumerator polynomial of C yields

$$W_C(X_0,X_0,X_2,X_2) = \sum_{0\le j_0+j_1=k\le m} A_{j_0,j_1,j_2,j_3} X_0^k X_2^{m-k},$$

and hence reduces

$$\sum_{0\le j_0+j_1=k\le m} A_{j_0,j_1,j_2,j_3} X_0^k X_2^{m-k}$$

$$= \frac{1}{\#C^\perp} W_{C^\perp}(2(X_0+X_2), (1+\sqrt{-1})(X_0-X_2), 0, (1-\sqrt{-1})(X_0-X_2)).$$

by the above MacWilliams identity. This substitution reduces the number of variables from four to two, therefore the right hand side can be expanded effectively.

In the same way, we can compute the weight distribution of the l-th least significant bits by using the MacWilliams identity over $\mathbb{Z}/2^l$.

4 The Results of Tests

The first example is a lagged Fibonacci generator, based on (1). As usual in a χ^2-test, the set $\{0, 1, \ldots, m\}$ is categorized into:

$$\begin{aligned} S_0 &= \{0, 1, \ldots, s_0\}, \\ S_k &= \{s_0 + k\} \quad (1 \le k \le m - 2s_0 - 1), \\ S_\nu &= \{m - s_0, m - s_0 + 1, \ldots, m\} \end{aligned} \tag{2}$$

for suitably chosen s_0 so that each category has moderate probability. Then the degree of freedom is $\nu = m - 2s_0$. We conduct the weight discrepancy test on the second least significant bits of this generator for these categories. The result is shown in Table 1.

The row δ shows the χ^2-discrepancy, the rows $N_{.75}, N_{.99}$ respectively show the safe, risky sample size implied by Theorem 1.

Table 1 Weight discrepancy test on the second least significant bit of $\mathbf{x}_{i+31} = \mathbf{x}_{i+28} + \mathbf{x}_i$ as in (1).

m	34	35	36	37	38
ν	14	15	16	17	18
δ	3.1×10^{-4}	5.3×10^{-4}	7.9×10^{-4}	1.1×10^{-3}	1.4×10^{-3}
$N_{.75}$	10,293	6,250	4,350	3,303	2,653
$N_{.99}$	49,113	29,513	20,349	15,314	12,206

We also empirically test the same generator by the weight distribution test [8], which we shall briefly explain. Fix the sample size N and choose an initial state. Generate m pseudo random bits by the generator. Let W_1 be the number of 0's in these m bits. Then again generate m words, count the number of 0's and let W_2 be this number. Iterate this N times. We have $W_1, W_2, \ldots, W_N$ each of which should conform to the binomial distribution $B(m, 1/2)$. We apply the χ^2-test to these N

samples, using the categories (2). We obtain one value of χ^2-statistics, and then obtain the corresponding probability value. This is the weight distribution test.

Table 2 shows the result of the weight distribution test in the case of $m = 34$. We choose five different initial values randomly, and test the generator for three different sample sizes, namely, 10,000, 50,000, and 100,000. The empirical results of the five tests are in good accordance with the theoretical expectation shown in Table 1.

Table 2 The p-values of the weight distribution tests on the second least significant bit with $m = 34$ (unit: %).

Sample size \trial	1st	2nd	3rd	4th	5th
$N_{.75} \approx 10{,}000$	95.2	50.0	2.6	96.2	45.8
$N_{.99} \approx 50{,}000$	93.4	99.0	99.6	99.9	92.6
$2 \times N_{.99} \approx 100{,}000$	88.0	99.9	100.0	99.9	99.8

Table 3 shows δ, $N_{.75}$ and $N_{.99}$ of the first to the sixth least significant bits of the distribution of (1). We observed that each δ becomes almost $1/4$ times smaller than the δ of the previous bit. Accordingly, $N_{.75}$ and $N_{.99}$ become four times larger.

Table 3 Weight discrepancy test for the 1st to 6th bits of $\mathbf{x}_{i+31} = \mathbf{x}_{i+28} + \mathbf{x}_i$ as in (1).

bit	1st	2nd	3rd	4th	5th	6th
δ	1.2×10^{-3}	3.1×10^{-4}	7.8×10^{-5}	1.9×10^{-5}	4.9×10^{-6}	1.2×10^{-6}
$N_{.75}$	2,566	10,293	41,195	164,806	659,250	2,637,030
$N_{.99}$	12,248	49,113	196,568	786,390	3,145,680	12,582,800

The second example is another type of a lagged Fibonacci generator, which is recommended by Knuth [1], based on

$$\mathbf{x}_{i+100} = -\mathbf{x}_{i+63} + \mathbf{x}_i \bmod 2^{30} \quad (i = 0, 1, 2, \ldots). \tag{3}$$

Table 4 shows the result of the weight discrepancy tests on the second least significant bit of the generator (3), Table 5 shows the result of the weight distribution tests on the same bit, and Table 6 shows the result of the weight discrepancy tests for the first to the sixth least significant bits for $m = 103$. This last table shows that the rate of decrease of δ is almost four. Although the order of δ of (3) is different from that of (1), the rate of decrease looks similar, but we have no mathematical explanation so far.

The CPU time consumed for the weight discrepancy test on the sixth bit is about 3 days by Mathematica 8 on the Intel Core i5 at 3.1 GHz, with 4 GB of memory. It seems more difficult to test the higher bits.

We also compute the χ^2-discrepancies of the second least significant bits of several generators. The following eight $\mathbb{F}_2$-linear generators were tested: a 13-term

Table 4 Weight discrepancy test on the 2nd bit of the generator $\mathbf{x}_{i+100} = -\mathbf{x}_{i+63} + \mathbf{x}_i$ as in (3).

m	$100+3$	$100+4$	$100+5$	$100+6$	$100+7$
v	43	44	45	46	47
δ	1.3×10^{-5}	2.2×10^{-5}	3.3×10^{-5}	4.7×10^{-5}	6.2×10^{-5}
$N_{.75}$	462,952	271,382	180,906	130,784	99,977
$N_{.99}$	1,931,530	1,129,910	751,697	542,374	413,834

Table 5 The p-values of the weight distribution tests on the generator (3) with $m = 103$ (unit: %).

Sample size \trial	1st	2nd	3rd	4th	5th
$4.6\times10^5 \approx N_{.75}$	50.1	47.9	40.8	35.5	80.5
$2.0\times10^6 \approx N_{.99}$	98.4	98.9	99.5	99.9	69.3
$4.0\times10^6 \approx 2\times N_{.99}$	100.0	99.2	99.9	99.9	99.9

Table 6 Weight discrepancy test for the 1st to 6th bits of $\mathbf{x}_{i+100} = -\mathbf{x}_{i+63} + \mathbf{x}_i$ as in (3).

bits	1st	2nd	3rd	4th	5th	6th
δ	5.1×10^{-5}	1.3×10^{-5}	3.2×10^{-6}	7.9×10^{-7}	2.0×10^{-7}	5.0×10^{-8}
$N_{.75}$	115,728	462,951	1,851,846	7,407,423	29,629,733	118,518,973
$N_{.99}$	482,844	1,931,532	7,726,286	30,905,312	123,621,410	494,485,798

and a 15-term linear recurrence generator with 100-dimensional state space, two toy models of Mersenne Twister [10], LFSR113 [3], XorShift [7, 12], TT800 [9], and WELL512 [13]. The dimension of the state space and the order of δ are shown in Table 7.

Table 7 Comparison of δ of several $\mathbb{F}_2$-generators.

Name	$\dim(S)$	$\lfloor\log_{10}\delta\rfloor$	Name	$\dim(S)$	$\lfloor\log_{10}\delta\rfloor$
$13-$term	100	-16	LFSR113	113	-34
$15-$term	100	-18	XorShift	128	-38
MT89	89	-26	TT800	800	-124
MT127	127	-32	WELL512	512	-152

5 Conclusion and Future Works

We computed the weight distribution of the second to the sixth least significant bits from lagged Fibonacci generators. The risky sample size is multiplied by a factor of 4, when the examined bit is shifted to the left. Tables 3 and 6 show that even the 6th least significant bit of the tested lagged Fibonacci generators have observable deviation, so more than 6 bits need to be discarded.

Lagged Fibonacci generators are one of the special types of multiple recursive generators (MRGs). The same method can apply to MRGs when the modulus is a power of 2, which is left for future work.

Acknowledgements This research has been supported in part by Grants-in-Aid for Scientific Research #24654019, #22740075 and #21654017.

References

1. Knuth, D.E.: The Art of Computer Programming, Volume 2: Seminumerical Algorithms, 3rd edn. Addison-Wesley, Reading (1997)
2. L'Ecuyer, P.: Uniform random number generation. Ann. Oper. Res. **53**, 77–120 (1994)
3. L'Ecuyer, P.: Tables of maximally-equidistributed combined LFSR generators. Math. Comp. **68**, 261–269 (1999)
4. L'Ecuyer, P., Simard, R.: TestU01: a C library for empirical testing of random number generators. ACM Trans. Math. Software **33**, Art. 22, 40 (2007)
5. MacWilliams, F.J., Sloane, N.J.A.: The Theory of Error-Correcting Codes. I. North-Holland, Amsterdam (1977)
6. Marsaglia, G.: DIEHARD: A battery of tests of randomness (1996). http://stat.fsu.edu/~geo/diehard.html
7. Marsaglia, G.: Xorshift rngs. J. Statist. Software **8**, 1–6 (2003)
8. Matsumoto, M., Kurita, Y.: Twisted GFSR generators. ACM Trans. Model. Comput. Simul. **2**, 179–194 (1992)
9. Matsumoto, M., Kurita, Y.: Twisted GFSR generators ii. ACM Trans. Model. Comput. Simul. **4**, 254–266 (1994)
10. Matsumoto, M., Nishimura, T.: Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. ACM Trans. Model. Comput. Simul. **8**, 3–30 (1998)
11. Matsumoto, M., Nishimura, T.: A nonempirical test on the weight of pseudorandom number generators. In: Fang, K., Hickernell, F.J., Niederreiter, H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods, 2000 (Hong Kong), pp. 381–395. Springer, Berlin (2002)
12. Panneton, F., L'Ecuyer, P.: On the xorshift random number generators. ACM Trans. Model. Comput. Simul. **15**, 346–361 (2005)
13. Panneton, F., L'Ecuyer, P., Matsumoto, M.: Improved long-period generators based on linear recurrences modulo 2. ACM Trans. Math. Software **32**, 1–16 (2006)
14. Vardy, A.: The intractability of computing the minimum distance of a code. IEEE Trans. Inform. Theory **43**, 1757–1766 (1997)

A Finite-Row Scrambling of Niederreiter Sequences

Roswitha Hofer and Gottlieb Pirsic

Abstract This paper introduces a scrambling matrix that modifies the generating matrices of the classical Niederreiter sequences to so-called finite-row generating matrices. The method used for determining this scrambling matrix also allows us to construct the inverse matrices of these generating matrices of Niederreiter. The question for *finite-row digital* (t, s)*-sequences* is motivated in the context of Niederreiter-Halton sequences, where—inspired by the Halton sequences—Niederreiter sequences in different bases are combined to a higher dimensional sequence. The investigation of the discrepancy of the Niederreiter-Halton sequences is a difficult task and still in its infancy. Results achieved for special examples raised the idea that the combination of finite-row generating matrices in different bases may be interesting. This paper also contains experiments that compare the performance of some Niederreiter-Halton sequences to the performance of Faure and Halton sequences and corroborate this idea.

1 Introduction

The interest in low-discrepancy sequences as research topic originally comes from number theory, more precisely Diophantine approximation, as exemplified by the Kronecker sequences (see e.g., Chap. XXIII in the classical textbook [6]). In the last decades, the application side has gained much importance in the field of what may be called number-theoretical analysis, here especially in quasi-Monte Carlo methods, with uses in many diverse fields such as mathematical finance, particle transport simulation and computer imaging. Particularly the task of numerically

R. Hofer (✉) · G. Pirsic
Institute of Financial Mathematics, Johannes Kepler University Linz, Altenbergerstrasse 69, 4040 Linz, Austria
e-mail: roswitha.hofer@jku.at; gottlieb.pirsic@jku.at

J. Dick et al. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2012*,
Springer Proceedings in Mathematics & Statistics 65,
DOI 10.1007/978-3-642-41095-6_20, © Springer-Verlag Berlin Heidelberg 2013

integrating highly multivariate function has gained importance, since the asymptotic error bound (e.g., via the Koksma-Hlawka inequality) is of order $1/N$ compared to the purely Monte-Carlo bound $1/\sqrt{N}$. (The interested reader, unfamiliar with (quasi-)Monte Carlo methods, discrepancy theory, and low-discrepancy sequences, is referred to [1, 14] for comprehensive introductions.)

Apart from the above mentioned Kronecker sequences a further example of a low-discrepancy sequence is the *Halton sequence* [5], which uses the notion of digit expansion: let φ_b be the *radical inverse*, i.e., the operation that reflects the b-adic expansion of a nonnegative integer at the (b-adic) decimal point. Then $((\varphi_{p_i}(n))_{i=1}^s)_{n\geq 0}$, where p_i are distinct primes, is an s-dimensional low-discrepancy sequence.

In a further application of the concept of using digit expansions, *digital* (t, s)*-sequences (over* $\mathbb{F}_q$*)* have been defined by Niederreiter [12], employing the *digital method* (earlier forms can be found in [2] and [15]). Briefly explained, the strategy is as follows: using s infinite matrices $C^{(i)} \in \mathbb{F}_q^{\mathbb{N}\times\mathbb{N}}$, the q-adic digit vector $\mathbf{v}_n \in \mathbb{F}_q^{\mathbb{N}}$ of $n \in \mathbb{N}$ is transformed into s vectors

$$\mathbf{x}_n^{(i)} := (C^{(i)}\mathbf{v}_n) \in \mathbb{F}_q^{\mathbb{N}}, \quad i \in \{1, \ldots, s\}.$$

These vectors are then considered as fractional digits of real numbers $\xi_n^{(i)} \in [0, 1)$, so in the end we gain the sequence $\omega = ((\xi_n^{(i)})_{i=1}^s)_{n\geq 0} \in ([0, 1)^s)^{\mathbb{N}_0}$. Here we implicitly used an enumeration of the elements of the finite field, i.e., a set bijection between $\mathbb{F}_q$ and the set of digits, $\{0, \ldots, q-1\}$. It is usually required that this bijection maps the zero in $\mathbb{F}_q$ to 0 and that the columns of the matrices contain only finitely many nonzero entries.

In these sequences, the parameter $t \in \mathbb{N}_0$ points to the quality of distribution of the point set; it occurs as an exponential factor in the Koksma-Hlawka bound, so low values of t are a goal, with 0 being the optimum. From the generating matrices the parameter t can be derived as follows: ω as defined above is a digital (t, s)-sequence, if: for all $m \in \mathbb{N}$ and all partitions $d_1, \ldots, d_s$ in $\mathbb{N}_0$ of $m - t$ the set of $m - t$ truncated (to m components) row vectors,

$$\{tru_m(\mathbf{c}_j^{(i)}) \in \mathbb{F}_q^m,\ i \in \{1, \ldots, s\},\ j \in \{1, \ldots, d_i\}\}$$

is linearly independent.

One particular construction of matrices was given by Niederreiter [13]: Let $\mathbb{F}_q$ be a finite field with $q = p^r$ elements, where p is prime and $r \geq 1$. For a given dimension s we choose $p_1, \ldots, p_s \in \mathbb{F}_q[x]$ to be distinct monic nonconstant irreducible polynomials over $\mathbb{F}_q$ of degrees $e_i := \deg p_i$ for $1 \leq i \leq s$. Now the jth row of the ith generating matrix $C^{(i)}$, denoted by $\rho_j^{(i)}$, is defined as follows. We choose $l \in \mathbb{N}$ and $m \in \{0, \ldots, e_i - 1\}$ such that $j = e_i l - m$ and consider the expansion

$$\frac{x^m}{p_i(x)^l} = \sum_{r\geq 0} a^{(i)}(l,m,r)x^{-r-1} \in \mathbb{F}_q((x^{-1}))$$

and set $\rho_j^{(i)} = (a^{(i)}(l,m,0), a^{(i)}(l,m,1), \ldots)$. It is easy to check that the generating matrices are nonsingular upper triangular (NUT) matrices over $\mathbb{F}_q$ and by (a small adaption of) [13] they generate a digital (t,s)-sequence over $\mathbb{F}_q$, with $t = \sum_{i=1}^{s}(e_i - 1)$.

The special case of $s \leq q$ with q prime and the choice of p_i as linear polynomials is well-known by the name of *Faure sequence* [2]. In this case, the generating matrices are powers of P, where P is the Pascal matrix of binomial coefficients modulo q,

$$P := \left(\binom{j-1}{i-1} \bmod q\right)_{i\geq 1, j\geq 1}.$$

Going back to Halton sequences, a characteristic distinguishing those sequences from digital (t,s)-sequences is the use of different bases in each coordinate. The construction principle using φ_b is however closely related to the digital method since using the identity matrix has the same effect as φ_b. A natural and obvious question now is the behavior of digital (t,s)-sequences with different bases in different components (or, from another viewpoint, of the direct product of sequences over different $\mathbb{F}_{q_i}$). For such sequences, first defined in [9] and earlier suggested in [3], the notion *Niederreiter-Halton sequence* was coined [7]. The investigation of the distribution of these sequences appears to be a difficult task. As can be seen by a particular choice of matrices, there are some deeper problems obstructing a straight-forward generalization: choose 2 and 3 as bases, let the generating matrix for base 3 equal the identity matrix and assume the first row of the matrix for base 2 only contains 1s (as is the case, e.g., with the Pascal matrix). Then one is lead immediately to the investigation of the binary digit sum of multiples of 3 which turns out to be uniformly distributed modulo 2 but with very slow convergence. Results in [10] imply that the discrepancy of, e.g., the five-dimensional sequence that is built by combining the Faure sequence in base 2 and the Faure sequence in base 3, which we call the *5-dimensional Faure-Halton sequence in bases* 2 *and* 3, satisfies a lower bound

$$ND_N \geq cN^{\log_4(3)} \text{ for all } N \in \mathbb{N}, \tag{1}$$

where the constant c does not depend on N. This result as well as some further investigations gave the impression, that too many nonzero entries in the generating matrices may cause problems when mixing digital sequences in different bases, and motivated the quest for *finite-row generating matrices*, i.e., matrices where each row contains only finitely many nonzero entries. In [10] first nontrivial examples of such matrices were given, which were further investigated in [11]. These sequences can be described as Faure-Tezuka scramblings [4] (a specific type of reordering) of Faure sequences. A direct construction of finite-row generating matrices of digital $(0,s)$-sequences over a finite field $\mathbb{F}_q$ was elaborated in [8]. So far, all results were limited to the quality parameter $t = 0$ and therefore limited to dimensions $s \leq q$.

Our aim in this paper is to give a scrambling matrix for Niederreiter sequences that transforms their generating matrices into finite-row matrices. This is done in Sect. 2. The method used therein also provides a tool to determine the inverse matrices of, e.g., the generating matrices proposed by Niederreiter. In Sect. 3 we briefly discuss results on the discrepancy of sequences that are built by combining sequences thus constructed and perform a numerical comparison to other constructions.

2 Scrambling of Faure and Tezuka and Finite-Row Digital (t, s)-Sequences

First we recall an easy result about linear scrambling of digital (t, s)-sequences.

Lemma 1 ([4, Proposition 1]). *Let $C^{(1)}, \dots, C^{(s)}$ be the generating matrices of a digital (t, s)-sequence over $\mathbb{F}_q$ and S be a non-singular upper triangular (NUT) $\mathbb{N} \times \mathbb{N}$-matrix over $\mathbb{F}_q$. Then $C^{(1)}S, \dots, C^{(s)}S$ are generating matrices of a digital (t, s)-sequence over $\mathbb{F}_q$.*

Proof. From the nonsingularity of S follows the nonsingularity of the square submatrices S_m consisting of the first m rows and columns for any $m > 0$. Suppose $d_1, \dots, d_s$ is a partition of $m - t$ and $\{tru_m(\mathbf{c}_j^{(i)}),\ i \in \{1, \dots, s\},\ j \in \{1, \dots, d_i\}\}$ is a linearly independent set of row vectors of the original generating matrices, as in the condition for a (t, s)-sequence. Consider the $(m - t) \times m$ matrix C' with these vectors as row vectors. Then, by $tru_m(\mathbf{v}S_m) = tru_m(\mathbf{v})S_m$ for any vector $\mathbf{v} \in \mathbb{F}_q^m$, the matrix $C'S_m$ is associated to the matrices $C^{(i)}S$ and has the same rank, $m - t$. Therefore all linear independence conditions translate from the matrices $C^{(i)}$ to $C^{(i)}S$ and the resulting net has at worst the same quality parameter t. □

We now aim for scrambling matrices S such that the generating matrices based on monic irreducible polynomials introduced by Niederreiter and already mentioned in the introduction yields finite-row matrices. In the following the quantity L_d is the maximal row length of the first d rows. More precisely, taking the matrix consisting of the first d rows of each of the generating matrices, L_d is the index of the last nonzero column (or ∞, if none exists). In [10] the notation $L(d, \dots, d)$ was used.

Theorem 1. *Let $C^{(1)}, \dots, C^{(s)}$ be the generating matrices associated to the distinct monic nonconstant irreducible polynomials $p_1, \dots, p_s$ with degrees $e_1, \dots, e_s$. We set $v := \operatorname{lcm}(e_1, \dots, e_s)$ and define the matrix S as follows. For $k \in \mathbb{N}$, the kth column c_k of S, is given by $c_k = (b_0, b_1, \dots, b_{k-1}, 0, \dots)^T$ where the b_u are the coefficients of the following monic polynomial of degree $k - 1$,*

$$p(x) = x^{r_1} \prod_{i=1}^{s} p_i(x)^{(l_i + l_{i+1} + \cdots + l_s)v/e_i} = \sum_{u \geq 0} b_u x^u.$$

Here the l_i and r_i are defined as follows

$$\begin{aligned} k-1 &= svl_s + r_s, & r_s &\in \{0,\dots,vs-1\} \\ r_s &= (s-1)vl_{s-1} + r_{s-1}, & r_{s-1} &\in \{0,\dots,v(s-1)-1\} \\ &\vdots & &\vdots \\ r_2 &= vl_1 + r_1, & r_1 &\in \{0,\dots,v-1\}. \end{aligned}$$

Then, the matrices $C^{(1)}S,\dots,C^{(s)}S$ generate a digital (t,s)-sequence, with $t = \sum_{i=1}^{s}(e_i-1)$ satisfying $L_d \le sv\left\lceil \frac{d}{v}\right\rceil \le sd + (v-1)s$.

Proof. It is easy to check that $p(x)$ is monic and has degree $k-1$ and therefore S is a NUT matrix. Hence Lemma 1 ensures the assertion on the quality parameter t. We prove the bounds on the row lengths. Let $\rho_j^{(i)}$ denote the jth row of the matrix $C^{(i)}$ and consider $\rho_j^{(i)}c_k$. We use that this product equals the coefficient of x^{-1} in the Laurent series of

$$\frac{x^r}{p_i(x)^l}x^{r_1}\prod_{i=1}^{s}p_i(x)^{(l_i+l_{i+1}+\cdots+l_s)v/e_i}, \tag{2}$$

where by definition $l \in \mathbb{N}$ and $r \in \{0,\dots,e_i-1\}$ are such that $j = e_i l - r$. To prove that the matrix $C^{(i)}S$ satisfies the upper bound on its lengths, i.e., $L_d \le sv\lceil d/v\rceil$, we have to ensure that for every positive integer d, for all $k > sv\lceil d/v\rceil$ and for all $1 \le j \le d$, the coefficient of x^{-1} in the Laurent series of (2) is 0.

Using the definition of l and r we obtain $l \le \lceil d/e_i\rceil$ for all $j \le d$. The definition of l_s and r_s together with the assumption $k > sv\lceil d/v\rceil$ yields $l_s \ge \lceil d/v\rceil$. The inequalities $l \le \lceil d/e_i\rceil \le \lceil d/v\rceil v/e_i \le l_s v/e_i$ (the second inequality follows from $v/e_i \in \mathbb{N}$) ensure that (2) is a polynomial and therefore in the Laurent series the coefficient of x^{-1} is 0. □

Remark 1. The row lengths obtained by S in Theorem 1 are asymptotically best possible. Proposition 1 in [10] says that if $C^{(1)},\dots,C^{(s)}$ are generating matrices of a digital $(0,s)$-sequence then for every $d \in \mathbb{N}$ there exists an $i \in \{1,\dots,s\}$ such that $L_d \ge sd$. For all $d \equiv 0 \pmod v$ we already obtain equality.

Example 1. If $s = q$, with q prime and $p_i(x) = x - i + 1$ then it is easily verified that the scrambling matrix is built columnwise by the coefficients of the falling factorials $x(x-1)(x-2)\cdots(x-(k-2))$, which are related to the signed Stirling numbers of the first kind and therefore

$$S = \left(\begin{bmatrix} r-1 \\ j-1 \end{bmatrix}(-1)^{r-j}\right)_{j\ge 1, r\ge 1} \in \mathbb{F}_p^{\mathbb{N}\times\mathbb{N}}$$

(the square brackets are Karamata-Knuth notation for Stirling numbers). This scrambling matrix was already introduced in [11], where nice formulas for the

generating matrices were identified, namely $C^{(i)}S = SQ^{i-1}$, where Q is a very thin NUT band matrix with bandwidth 2.

Open Problem 1. Taking a closer look at the example above one sees that the scrambling matrix related to the Stirling numbers almost diagonalizes simultaneously the generating matrices of the Faure sequences. In this case scrambling to finite rows means changing to a common base of the generating matrices. Of course an interesting question would be if such a scrambling, effecting a simultaneous almost diagonalization, exists for all examples of generating matrices of Niederreiter sequences; or whether it is already possible to identify similar formulas for $C^{(1)}S, \ldots, C^{(s)}S$ using the scrambling matrix suggested in Theorem 1.

As a corollary to the method of proof used for Theorem 1 we can determine the inverse matrices to the generating matrices of Niederreiter sequences. These inverse matrices can be useful e.g., in geometric applications, specifically in the fast determination of subsequences lying in specific elementary intervals.

We use the notation $\lfloor \cdot \rfloor$ and $\{\cdot\}$ for the polynomial and fractional part of a Laurent series in $1/x$.

Proposition 1. *Let $C^{(i)}$ be the generating matrices based on the monic distinct nonconstant irreducible polynomials $p_1, \ldots, p_s$ of degrees $e_1, \ldots, e_s$. Then their inverses, denoted by $C^{(-i)}$, can be determined as follows. For $k \in \mathbb{N}$ we denote the kth column of $C^{(-i)}$ by $c^{(-i)}{}_k$, choose $v \in \mathbb{N}_0, w \in \{0, \ldots, e_i - 1\}$ such that $k - 1 = e_i v + w$, regard the polynomial*

$$p_i(x)^v \left\lfloor \frac{p_i(x)}{x^{e_i - w}} \right\rfloor = \sum_{u \geq 0} b(v, w, u) x^u,$$

and set $c^{(-i)}{}_k = (b(v, w, 0), b(v, w, 1), \ldots)^T$.

Proof. We abbreviate the jth row of $C^{(i)}$ to $\rho_j^{(i)}$. The crux of the following proof is that $\rho_j^{(i)} \cdot c^{(-i)}{}_k$ equals the coefficient of x^{-1} in the Laurent series of

$$\frac{x^m}{p_i(x)^l} p_i(x)^v \left\lfloor \frac{p_i(x)}{x^{e_i - w}} \right\rfloor = \sum_{p \geq 0} d_p x^p + \sum_{r \geq 0} f_r x^{-r-1}$$

with $d_p, f_r \in \mathbb{F}_q$ and where m, l satisfies $m \in \{0, \ldots, e_i - 1\}$ and $j = l e_i - m$. Hence it suffices to consider f_0 for different values of k and j and to prove that $f_0 = 1$ if $k = j$ and zero otherwise. We make a distinction of cases with regard to v and l:

- $v \geq l$ and therefore $k > j$: In this case $\frac{x^m}{p_i(x)^l} p_i(x)^v \left\lfloor \frac{p_i(x)}{x^{e_i - w}} \right\rfloor$ is a polynomial over $\mathbb{F}_q$ and therefore the coefficient of x^{-1} is $0 \in \mathbb{F}_q$.
- $v + 1 < l$: In this case the leading coefficient in the Laurent series belongs to $x^{m - l e_i + v e_i + e_i - (e_i - w)} = x^{m + w + (v - l) e_i}$. Now it is easy to see that $0 \leq m + w \leq$

$2(e_i - 1)$ and $(v + l)e_i \leq -2e_i$ and therefore $m + w + (v - l)e_i \leq -2$. Thus the coefficient of x^{-1} is $0 \in \mathbb{F}_q$.

– $v + 1 = l$: We write $p_i(x) = x^{e_i} + \alpha_{e_i-1}x^{e_i-1} + \cdots + \alpha_1 x + \alpha_0$ and regard

$$\frac{x^m}{p_i(x)} \left\lfloor \frac{p_i(x)}{x^{e_i-w}} \right\rfloor = \frac{x^m(x^w + \alpha_{e_i-1}x^{w-1} + \cdots + \alpha_{e_i-w+1}x + \alpha_{e_i-w})}{x^{e_i} + \alpha_{e_i-1}x^{e_i-1} + \cdots + \alpha_1 x + \alpha_0}.$$

If $m + w = e_i - 1$ the coefficient of x^{-1} is $1 \in \mathbb{F}_q$, and if $m + w < e_i - 1$ the coefficient of x^{-1} is $0 \in \mathbb{F}_q$. Finally if $m + w > e_i - 1$ we consider the equality

$$\frac{x^m}{p_i(x)} \left\lfloor \frac{p_i(x)}{x^{e_i-w}} \right\rfloor = \frac{x^m}{p_i(x)} \frac{p_i(x)}{x^{e_i-w}} - \frac{x^m}{p_i(x)} \left\{ \frac{p_i(x)}{x^{e_i-w}} \right\}.$$

We know that the first term of the right hand side is a polynomial and the Laurent series of the second term is of the form $\sum_{r\geq 1} f_r x^{-r-1}$. This yields that the coefficient of x^{-1} is $0 \in \mathbb{F}_q$.

Since $j = k$ is equivalent to the only nonzero case, i.e., $v + 1 = l$ together with $m + w = e_i - 1$, the proof is concluded. □

3 Discrepancy and Experiments

As constructions of finite-row generating matrices of digital (t, s)-sequences with roughly lowest possible row lengths are available now, the next step is to investigate the discrepancy of Niederreiter-Halton sequences, that are built by combining finite-row digital (t, s)-sequences in different bases. Here we reduce our discussions to the five-dimensional *finite-row Faure-Halton sequence in bases* 2 *and* 3, which is built by juxtaposing the components of the 2-dimensional scrambled Faure sequence in base 2 and the components of the 3-dimensional scrambled Faure sequence in base 3 of Example 1. One method near at hand for the investigation of the discrepancy of this finite-row Faure-Halton sequence is to proceed analogously to papers that estimate the discrepancy of the Halton sequence, which are using as a main tool the Chinese Remainder Theorem. This was done in [9] and produced a bound that depends on the row lengths, see [9, Theorem 3.1]. As a consequence, the bound achieved for the finite-row Faure-Halton sequence in bases 2 and 3 is also valid for all other Niederreiter-Halton sequences that share the same bounds on the row lengths, as for example the one generated by the following two matrices over $\mathbb{F}_2$,

$$\begin{pmatrix} 1\,0\,0\,0\,0\,0\,0\,0\,\ldots \\ 0\,0\,1\,0\,0\,0\,0\,0\,\ldots \\ 0\,0\,0\,0\,1\,0\,0\,0\,\ldots \\ 0\,0\,0\,0\,0\,0\,1\,0\,\ldots \\ \vdots\,\vdots\,\vdots\,\vdots\,\vdots\,\vdots\,\vdots\,\vdots\,\ddots \end{pmatrix} \text{ and } \begin{pmatrix} 0\,1\,0\,0\,0\,0\,0\,0\,\ldots \\ 0\,0\,0\,1\,0\,0\,0\,0\,\ldots \\ 0\,0\,0\,0\,0\,1\,0\,0\,\ldots \\ 0\,0\,0\,0\,0\,0\,0\,1\,\ldots \\ \vdots\,\vdots\,\vdots\,\vdots\,\vdots\,\vdots\,\vdots\,\vdots\,\ddots \end{pmatrix},$$

and the following three matrices over $\mathbb{F}_3$,

$$\begin{pmatrix} 1\,0\,0\,0\,0\,0\,0\,0\,\ldots \\ 0\,0\,0\,1\,0\,0\,0\,0\,\ldots \\ 0\,0\,0\,0\,0\,0\,0\,1\,\ldots \\ \vdots\,\vdots\,\vdots\,\vdots\,\vdots\,\vdots\,\vdots\,\vdots\,\ddots \end{pmatrix}, \begin{pmatrix} 0\,1\,0\,0\,0\,0\,0\,0\,\ldots \\ 0\,0\,0\,0\,1\,0\,0\,0\,\ldots \\ 0\,0\,0\,0\,0\,0\,0\,1\,\ldots \\ \vdots\,\vdots\,\vdots\,\vdots\,\vdots\,\vdots\,\vdots\,\vdots\,\ddots \end{pmatrix} \text{ and } \begin{pmatrix} 0\,0\,1\,0\,0\,0\,0\,0\,\ldots \\ 0\,0\,0\,0\,0\,1\,0\,0\,\ldots \\ 0\,0\,0\,0\,0\,0\,0\,0\,\ldots \\ \vdots\,\vdots\,\vdots\,\vdots\,\vdots\,\vdots\,\vdots\,\vdots\,\ddots \end{pmatrix}.$$

We call this sequence *trivial finite-row sequence in bases* 2 *and* 3. Its component sequences do not have good distribution properties and, therefore, we do not expect that the bound achieved in [9, Theorem 3.1] is sharp for the finite-row Faure-Halton sequences. Nevertheless so far there are no methods known to get better discrepancy bounds for these sequences.

In order to get an impression about the quality of this sequence we carried out an experiment, where we approximated the integral of several functions using different sequences and compared the convergence of the integration error.

We used the following 5-dimensional sequences:

Halton:	the Halton sequence (with respect to the primes $2, 3, 5, 7, 11$).
Faure:	the quinary Faure sequence.
FinFH:	finite-row Faure-Halton sequence in bases 2 and 3.
FaureH:	the 5-dimensional Faure-Halton sequence in bases 2 and 3.
TrivFinH:	the trivial finite-row sequence in bases 2 and 3.
MC:	the average of integration errors using ten instances of a pseudo-random sequence (the internal MATHEMATICA [16] routine `Random[]`).

For numerical investigation of the performance of those point sequences in an integration task, we chose the following 5-dimensional test functions:

f_1: a polynomial: $(2(x_1 + x_2 + x_3 + x_4 + x_5) - 5)^4$.
f_2: a Gaussian distribution function: $\exp(\sum_i (x_i - 1/i)^2)$.
f_3: an oscillatory function with decreasing frequencies in successive components: $\cos(\sum_i x_i/i)$.
f_4: an oscillatory function with increasing frequencies in successive components: $\cos(e + 2\pi \sum_i i x_i)$.
f_5: a rational function:

$$\frac{((\sum_i x_i)^2 - 1)^2}{1 + ((\sum_i x_i)^2 - \sum_i (x_i^2))/2}.$$

Figures 1–5 exhibit the absolute numerical integration errors in log-log graphs.

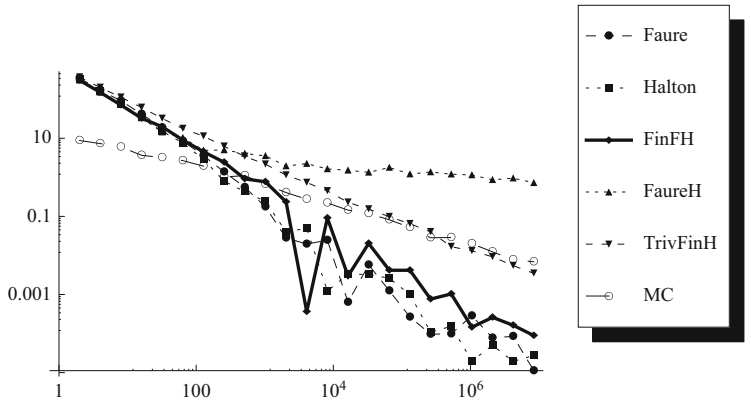


Fig. 1 f_1: a polynomial.

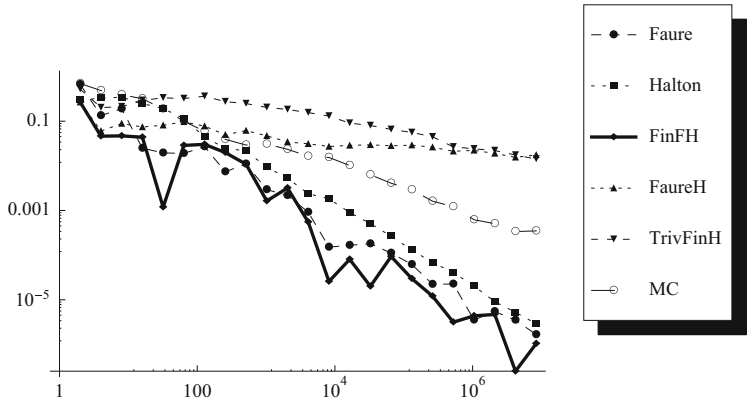


Fig. 2 f_2: a Gaussian distribution function.

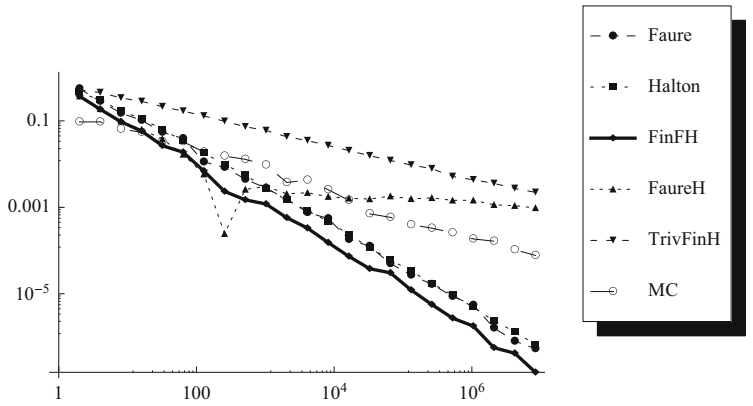


Fig. 3 f_3: an oscillatory function with decreasing frequencies.

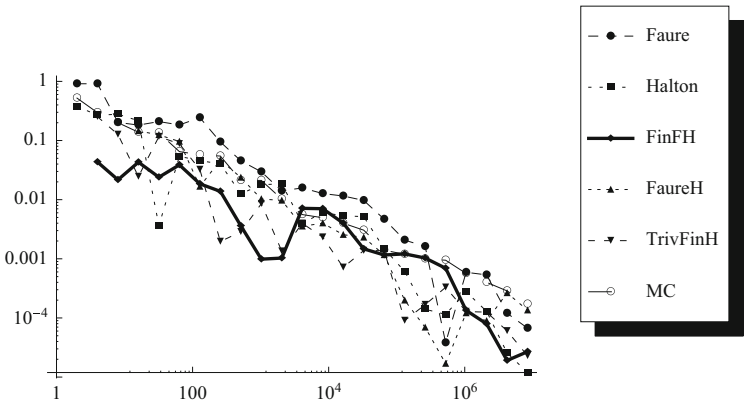


Fig. 4 f_4: an oscillatory function with increasing frequencies.

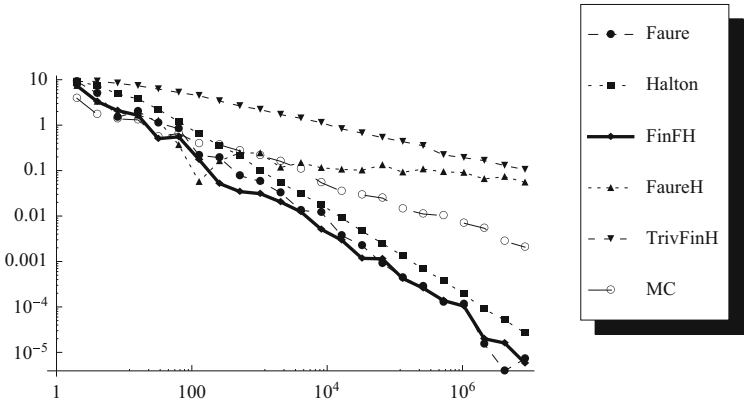


Fig. 5 f_5: a rational function.

3.1 Summary of the Numerical Experiments

Where QMC rules cannot prevail over MC rules due to the high variation of the function, see, e.g., function f_4, the finite-row Faure-Halton point sets are not worse (nor better). All other functions are relatively smooth. Note that even in this relatively small-scale data set it can be seen that the numerical integration with the classical MC and QMC rules (compare Faure and Halton) is according to expectation, so that a comparison is justified. We also observe a performance of the Faure-Halton sequence according to the large lower discrepancy bound (1) and also the integration error using the trivial finite-row sequence converges slowly as expected. Surprisingly, the QMC rule that is based on the finite-row Faure-Halton sequence behaves in a similar way as the ones that use the well-known low-discrepancy sequences. Comparison of this rule with the one based on the trivial

example, indicates that the bound in [9, Theorem 3.1] is not sharp. In summary, we consider the investigation of finite-row Niederreiter-Halton sequences as an interesting task for future research.

Acknowledgements The first author was supported by the Austrian Science Fund (FWF), Project P21943 and the second author by the Austrian Science Fund (FWF), Projects S9606 and P23285-N18. The first author would like to thank Christian Irrgeher for his advice for the numerical experiments.

References

1. Dick, J., Pillichshammer, F.: Digital Nets and Sequences. Discrepancy Theory and Quasi-Monte Carlo Integration. Cambridge University Press, Cambridge (2010)
2. Faure, H.: Discrépance de suites associées à un système de numération (en dimension s). Acta Arith. **41**, 337–351 (1982)
3. Faure, H.: Multidimensional quasi-Monte Carlo methods. (Méthodes quasi-Monte-Carlo multidimensionnelles.) Theor. Comput. Sci. **123**, 131–137 (1994)
4. Faure, H., Tezuka, S.: Another random scrambling of digital (t, s)-sequences. In: Fang, K.-T., Hickernell, F.J., Niederreiter, H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2000, pp. 242–256. Springer, Berlin/Heidelberg (2002)
5. Halton, J.H.: On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. Numer. Math. **2**, 84–90 (1960)
6. Hardy, G.H., Wright, E.M.: An Introduction to the Theory of Numbers. Oxford University Press, New York (1979)
7. Hofer, R.: On the distribution properties of Niederreiter-Halton sequences. J. Number Theory **129**, 451–463 (2009)
8. Hofer, R.: A construction of digital $(0, s)$-sequences involving finite-row generator matrices. Finite Fields Appl. **18**, 587–596 (2012)
9. Hofer, R., Kritzer, P., Larcher, G., Pillichshammer, F.: Distribution properties of generalized van der Corput-Halton sequences and their subsequences. Int. J. Number Theory **5**, 719–746 (2009)
10. Hofer, R., Larcher, G.: On existence and discrepancy of certain digital Niederreiter-Halton sequences. Acta Arith. **141**, 369–394 (2010)
11. Hofer, R., Pirsic, G.: An explicit construction of finite-row digital (0,s)-sequences. Unif. Distrib. Theory **6**, 13–30 (2011)
12. Niederreiter, H.: Point sets and sequences with small discrepancy. Monatsh. Math. **104**, 273–337 (1987)
13. Niederreiter, H.: Low-discrepancy and low-dispersion sequences. J. Number Theory **30**, 51–70 (1988)
14. Niederreiter, H.: Random Number Generation and Quasi-Monte Carlo Methods. CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 63. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (1992)
15. Sobol', I.M.: On the distribution of points in a cube and the approximation evaluation of integrals (Russian). Ž. Vyčisl. Mat. i Mat. Fiz. **7**, 784–802 (1967)
16. Wolfram Research, Inc.: Mathematica, Version 7.0, Champaign (2008)

Reconstructing Multivariate Trigonometric Polynomials by Sampling Along Generated Sets

Lutz Kämmerer

Abstract The approximation of problems in d spatial dimensions by sparse trigonometric polynomials supported on known or unknown frequency index sets $I \subset \mathbb{Z}^d$ is an important task with a variety of applications. The use of a generalization of rank-1 lattices as spatial discretizations offers a suitable possibility for sampling such sparse trigonometric polynomials. Given an index set of frequencies, we construct corresponding sampling sets that allow a stable and unique discrete Fourier transform. Applying the one-dimensional non-equispaced fast Fourier transform (NFFT) enables the fast evaluation and reconstruction of the multivariate trigonometric polynomials.

1 Introduction

Given a spatial dimension $d \in \mathbb{N}$, we consider Fourier series of continuous functions $f(\mathbf{x}) = \sum_{\mathbf{k}\in\mathbb{Z}^d} \hat{f}_{\mathbf{k}} \mathrm{e}^{2\pi \mathrm{i}\mathbf{k}\cdot\mathbf{x}}$ mapping the d-dimensional torus $[0,1)^d$ into the complex numbers $\mathbb{C}$, where $\left(\hat{f}_{\mathbf{k}}\right)_{\mathbf{k}\in\mathbb{Z}^d} \subset \mathbb{C}$ are the Fourier coefficients. A sequence $\left(\hat{f}_{\mathbf{k}}\right)_{\mathbf{k}\in\mathbb{Z}^d}$ with a finite number of nonzero elements specifies a trigonometric polynomial. We call the index set of the nonzero elements the frequency index set of the corresponding trigonometric polynomial. For a fixed index set $I \subset \mathbb{Z}^d$ with a finite cardinality $|I|$, $\Pi_I = \operatorname{span}\{\mathrm{e}^{2\pi \mathrm{i}\mathbf{k}\cdot\mathbf{x}} : \mathbf{k} \in I\}$ is called the space of trigonometric polynomials with frequencies supported by I.

Assuming the index set I is of finite cardinality and a suitable discretization in frequency domain for approximating functions, e.g. functions of dominating mixed smoothness, cf. [13], we are interested in evaluating the corresponding trigonometric polynomials at sampling nodes and reconstructing the Fourier

L. Kämmerer
Faculty of Mathematics, Chemnitz University of Technology, 09107 Chemnitz, Germany
e-mail: kaemmerer@mathematik.tu-chemnitz.de

J. Dick et al. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2012*,
Springer Proceedings in Mathematics & Statistics 65,
DOI 10.1007/978-3-642-41095-6_21, © Springer-Verlag Berlin Heidelberg 2013

coefficients $\left(\hat{f}_{\mathbf{k}}\right)_{\mathbf{k}\in I}$ from samples. Accordingly, we consider (sparse) multivariate trigonometric polynomials

$$f(\mathbf{x}) = \sum_{\mathbf{k}\in I} \hat{f}_{\mathbf{k}} \mathrm{e}^{2\pi \mathrm{i}\mathbf{k}\cdot\mathbf{x}}$$

and assume the frequency index set I is given.

For different specific index sets I there has been done some related work using rank-1 lattices as spatial discretizations [6, 11]. A multivariate trigonometric polynomial evaluated at all nodes of a rank-1 lattice essentially simplifies to a one-dimensional fast Fourier transform of the length of the cardinality of the rank-1 lattice, cf. [10]. Allowing for some oversampling one can find a rank-1 lattice, which even allows the reconstruction of the trigonometric polynomial from the samples at the rank-1 lattice nodes. A suitable strategy to search for such reconstructing rank-1 lattices can be adapted from numerical integration. In particular, a modification of the component-by-component constructions of lattice rules based on various weighted trigonometric degrees of exactness described in [2] allows one to find adequate rank-1 lattices in a relatively fast way, cf. [6]. The search strategy specified in [6] uses discrete optimization techniques.

In this paper we consider so-called generated sets, which generalize the concept of rank-1 lattices. The structure of these spatial discretizations allows for the evaluation of multivariate trigonometric polynomials by means of some simple precomputations and a one-dimensional non-equispaced discrete Fourier transform (NDFT). The fast computation can be realized by using the non-equispaced fast Fourier transform (NFFT), cf. [7]. The stability of the computation mainly depends on the Fourier matrices of this one-dimensional NFFT. Similar to the approaches known from rank-1 lattices, we have to search for suitable generating vectors guaranteeing a Fourier matrix of full column rank and, in addition, stability. In contrast to searching for suitable rank-1 lattices, we can use continuous optimization methods. Our search algorithm is based on the minimization of an upper bound of the maximum Gerschgorin circle radii, cf. [4], via a simplex search method.

The paper is organized as follows: In Sect. 2 we define generated sets, explain their advantages in computation, and give a basic example. To estimate the stability of the corresponding discrete Fourier transform, we specify an upper bound on the condition number of the involved Fourier matrices in Sect. 3. Algorithm 1 describes how to compute this upper bound in a simple and fast way. We optimize the generating vector by applying a nonlinear optimization technique as described in [12]. In practice, we (locally) minimize the theoretical number of samples needed to achieve at least a fixed stability. Some numerical examples can be found in Sect. 4.

The given examples include frequency index sets called weighted hyperbolic crosses

$$H_N^{d,\boldsymbol{\gamma}} := \{\mathbf{h} \in \mathbb{Z}^d : \prod_{s=1}^{d} \max(1, \gamma_s^{-1}|h_j|) \leq N\}$$

with parameters $d \in \mathbb{N}$, $N \in \mathbb{R}$, $\boldsymbol{\gamma} \subset \mathbb{R}^{\mathbb{N}}$, $1 \geq \gamma_1 \geq \gamma_2 \geq \ldots \geq 0$, and $0^{-1} := \infty$. More general index sets called generalized hyperbolic crosses were discussed in [5, 8]. Note that our approach is universally applicable. Accordingly, the theoretical statements of this paper also treat the frequency index sets specified in [5, 8].

2 Generated Sets

For given $M \in \mathbb{N}$ and $\mathbf{r} \in \mathbb{R}^d$ we define the *generated set*

$$\Lambda(\mathbf{r}, M) := \{\mathbf{x}_j = j\mathbf{r} \bmod 1, j = 0, \ldots, M-1\}$$

as a generalization of rank-1 lattices. We stress the fact that the restriction of the generating vector $\mathbf{r} \in M^{-1}\mathbb{Z}^d$ results in rank-1 lattices, cf. [2, 3, 14, 15]. Note, in contrast to rank-1 lattices, generated sets do not retain the group structure of the sampling sets, i.e. in general we have $\mathbf{x}_j \neq \mathbf{x}_{j+M}$.

However, we take advantage of the rank-1 structure of the generated set. In a similar way as described in [10], the evaluation of the trigonometric polynomial $f \in \Pi_I$ at all nodes $\mathbf{x}_j \in \Lambda(\mathbf{r}, M)$ simplifies to a one-dimensional NDFT. For $\mathscr{Y} = \{\mathbf{k} \cdot \mathbf{r} \bmod 1 : \mathbf{k} \in I\}$ is the set of all scalar products of the elements of the frequency index set I with the generating vector $\mathbf{r}$ we obtain

$$f(\mathbf{x}_j) = \sum_{\mathbf{k} \in I} \hat{f}_{\mathbf{k}} \mathrm{e}^{2\pi \mathrm{i} j \mathbf{k} \cdot \mathbf{r}} = \sum_{y \in \mathscr{Y}} \left(\sum_{\mathbf{k} \cdot \mathbf{r} \equiv y \pmod 1} \hat{f}_{\mathbf{k}} \right) \mathrm{e}^{2\pi \mathrm{i} j y}.$$

We evaluate f at all nodes $\mathbf{x}_j \in \Lambda(\mathbf{r}, M)$, $j = 0, \ldots, M-1$, by the precomputation of all $\hat{g}_y := \sum_{\mathbf{k} \cdot \mathbf{r} \equiv y \pmod 1} \hat{f}_{\mathbf{k}}$ together with a one-dimensional NFFT in $\mathscr{O}(M \log M + (|\log \varepsilon| + d)|I|)$ floating point operations, [7]. The parameter ε determines the accuracy of the computation and is independent of the dimension d.

As the fast evaluation of trigonometric polynomials at all sampling nodes $\mathbf{x}_j$ of the generated set $\Lambda(\mathbf{r}, M)$ is guaranteed, we draw our attention to the reconstruction of a trigonometric polynomial f with frequencies supported on I using function values at the nodes $\mathbf{x}_j$ of a generated set $\Lambda(\mathbf{r}, M)$. We consider the corresponding Fourier matrix $\mathbf{A}$ and its adjoint $\mathbf{A}^*$,

$$\mathbf{A} := \left(\mathrm{e}^{2\pi \mathrm{i} \mathbf{k} \cdot \mathbf{x}}\right)_{\mathbf{x} \in \Lambda(\mathbf{r}, M),\, \mathbf{k} \in I} \quad \text{and} \quad \mathbf{A}^* := \left(\mathrm{e}^{-2\pi \mathrm{i} \mathbf{k} \cdot \mathbf{x}}\right)_{\mathbf{k} \in I,\, \mathbf{x} \in \Lambda(\mathbf{r}, M)},$$

to determine necessary and sufficient conditions on generated sets $\Lambda(\mathbf{r}, M)$ allowing for a unique reconstruction of all Fourier coefficients of $f \in \Pi_I$. Assuming a full column rank matrix $\mathbf{A}$, the reconstruction of the Fourier coefficients $\hat{\mathbf{f}} = (\hat{f}_{\mathbf{k}})_{\mathbf{k} \in I}$ from sampling values $\mathbf{f} = (f(\mathbf{x}))_{\mathbf{x} \in \Lambda(\mathbf{r}, M)}$ can be realized by solving $\mathbf{A}^* \mathbf{A} \hat{\mathbf{f}} = \mathbf{A}^* \mathbf{f}$

using a standard conjugate gradient method, see [1, Chap. 11]. In particular, we aim to find generated sets $\Lambda(\mathbf{r}, M)$ that even allow for a stable reconstruction of the Fourier coefficients of specific trigonometric polynomials.

For that reason we consider the spectral condition number of the matrix $\mathbf{B} = M^{-1}\mathbf{A}^*\mathbf{A}$, which is defined as

$$\mathrm{cond}_2(\mathbf{B}) = \frac{\lambda_{\max}}{\lambda_{\min}},$$

where $\lambda_{\max}$ and $\lambda_{\min}$ are the largest and smallest eigenvalues of $\mathbf{B}$, respectively. Note that $\mathbf{B}$ is a symmetric, positive semidefinite matrix with eigenvalues $0 \leq \lambda_{\min} \leq \lambda_{\max}$. In particular, the condition number of $\mathbf{B}$ is bounded below by one.

Besides the stability, the considered condition number measures the speed of convergence of the conjugate gradient method used to reconstruct the trigonometric polynomial f, cf. [1, Chap. 13]. The lower the condition number the faster our reconstruction algorithm converges.

Of course, one can consider the condition number as a function of different variables. Our approach fixes the frequency index set I, which results in a functional

$$\kappa(\mathbf{r}, M) := \mathrm{cond}_2(\mathbf{B}(\mathbf{r}, M))$$

depending on the generating vector $\mathbf{r}$ and the number of samples M, where $\mathbf{B}(\mathbf{r}, M) = M^{-1}\left(\mathbf{A}(\mathbf{r}, M)\right)^* \mathbf{A}(\mathbf{r}, M)$ and $\mathbf{A}(\mathbf{r}, M) = \left(\mathrm{e}^{2\pi\mathrm{i}\mathbf{k}\cdot\mathbf{x}}\right)_{\mathbf{x}\in\Lambda(\mathbf{r},M),\,\mathbf{k}\in I}$. Now we are interested in a generating vector $\mathbf{r}$ which minimizes the functional κ for fixed M. For relatively small cardinalities $|I|$ one can evaluate this condition number exactly. Thus, we can minimize the functional κ using nonlinear optimization techniques such as nonlinear simplex methods. The vectors (1) in the following example were determined in this way.

Example 1. We consider the weighted hyperbolic cross $H^{d,\boldsymbol{\gamma}}_{256}$ with $\boldsymbol{\gamma} = \left(4^{1-s}\right)_{s\in\mathbb{N}}$ and fix the number of sampling points $M = 16{,}381 < 16{,}384 = \lfloor\gamma_1 N\rfloor\lfloor\gamma_2 N\rfloor$. Hence, for $d \geq 2$ Lemma 2.1 in [6] yields that there does not exist any sampling scheme of $M = 16{,}381$ nodes that allows for a perfectly stable reconstruction, i.e. it is proven that $\mathrm{cond}_2(\mathbf{B}) > 1$.

Nevertheless, we ask for a sampling scheme of cardinality $M = 16{,}381$ with a stable Fourier matrix $\mathbf{A}$. Generated sets are our first choice because of the easy possibility of the fast evaluation and reconstruction. In fact, the vectors

$$\mathbf{r}_2 = \begin{pmatrix} 0.508425953824 \\ 0.058509185871 \end{pmatrix} \qquad \text{and} \qquad \mathbf{r}_5 = \begin{pmatrix} 0.075119519237 \\ 0.285056619170 \\ 0.500703041738 \\ 0.970811563102 \\ 0.568203958723 \end{pmatrix} \tag{1}$$

generate the two-dimensional set $\Lambda_2 = \Lambda(\mathbf{r}_2, 16{,}381)$ and the five-dimensional set $\Lambda_5 = \Lambda(\mathbf{r}_5, 16{,}381)$. The corresponding condition numbers $\mathrm{cond}_2(\mathbf{B}_s)$ with $\mathbf{B}_s = M^{-1}\,(\mathbf{A}_s)^*\,\mathbf{A}_s$ and Fourier matrices $\mathbf{A}_s = \left(\mathrm{e}^{2\pi\mathrm{i}\mathbf{k}\cdot\mathbf{x}}\right)_{\mathbf{x}\in\Lambda_s,\,\mathbf{k}\in H_N^{s,\gamma}}$, $s = 2, 5$, are

$$\mathrm{cond}_2(\mathbf{B}_s) \approx \begin{cases} 3.9177, & \text{for } s = 2, \\ 11.934, & \text{for } s = 5. \end{cases}$$

Note that the corresponding matrices are square matrices with $\mathbf{B}_2 \in \mathbb{C}^{1{,}761\times 1{,}761}$ and $\mathbf{B}_5 \in \mathbb{C}^{2{,}187\times 2{,}187}$, respectively. □

Considering frequency index sets and corresponding generated sets of larger cardinalities, we cannot compute exact condition numbers efficiently. For that reason, we want to estimate the condition numbers from above.

3 Gerschgorin Circle Theorem and Generated Sets

In the following, we consider the Fourier matrix $\mathbf{A}(\mathbf{r}, M)$ and its adjoint $\mathbf{A}^*(\mathbf{r}, M)$ like above and apply the Gerschgorin circle theorem to the matrix $\mathbf{B}(\mathbf{r}, M)$. Let us consider the elements

$$(\mathbf{B}(\mathbf{r}, M))_{\mathbf{h},\mathbf{k}} = \frac{1}{M}\sum_{j=0}^{M-1} \mathrm{e}^{2\pi\mathrm{i}j(\mathbf{k}-\mathbf{h})\cdot\mathbf{r}} = \frac{1}{M}\sum_{j=0}^{M-1} \mathrm{e}^{2\pi\mathrm{i}j(y_\mathbf{k}-y_\mathbf{h})} =: K_M(y_\mathbf{k} - y_\mathbf{h}) \quad (2)$$

of the matrix $\mathbf{B}(\mathbf{r}, M)$. We define $y_\mathbf{h} = \mathbf{h}\cdot\mathbf{r} \bmod 1$ for all $\mathbf{h} \in I$ and therefore we can regard K_M as a univariate trigonometric kernel, which obviously is a Dirichlet kernel. Now we adapt some results from [9, Theorem 4.1] and formulate the following

Theorem 1. *We fix $\mathbf{r} \in \mathbb{R}^d$ and $I \subset \mathbb{Z}^d$ of finite cardinality. Let $y_\mathbf{h} = \mathbf{h}\cdot\mathbf{r} \bmod 1$ for all $\mathbf{h} \in I$. Moreover, let us assume that we have sorted the sequence of $y_\mathbf{h}$'s in ascending order, i.e. $0 \le y_{\mathbf{h}_1} \le y_{\mathbf{h}_2} \le \ldots \le y_{\mathbf{h}_{|I|}} < 1$. In addition, we define the sequence of gaps* $\mathbf{g}$

$$g_j = \begin{cases} 1 + y_{\mathbf{h}_1} - y_{\mathbf{h}_{|I|}}, & \text{for } j = 1, \\ y_{\mathbf{h}_j} - y_{\mathbf{h}_{j-1}}, & \text{for } j = 2, \ldots, |I|. \end{cases}$$

Then, for $M \in \mathbb{N}$ the interval $\left[1 - M^{-1}\Delta(\mathbf{r}), 1 + M^{-1}\Delta(\mathbf{r})\right]$ with

$$\Delta\colon \mathbb{R}^d \to \mathbb{R}, \quad \mathbf{r} \mapsto \Delta(\mathbf{r}) := \sum_{k=1}^{\left\lfloor \frac{|I|}{2} \right\rfloor} \left(\sum_{t=1}^{k} g_{\pi(t)} \right)^{-1}, \quad (3)$$

and π being a permutation of $\{1,\dots,|I|\}$ ordering the gaps $0 \le g_{\pi(1)} \le g_{\pi(2)} \le \dots \le g_{\pi(|I|)}$ contains all eigenvalues of the matrix

$$(\mathbf{B}(\mathbf{r},M))_{\mathbf{h},\mathbf{k}\in I} = (K_M(y_{\mathbf{k}}-y_{\mathbf{h}}))_{\mathbf{h},\mathbf{k}\in I}\,.$$

Proof. We consider the sequence $\left(g_{\pi(t)}\right)_{t=1}^{|I|}$. For $g_{\pi(1)}=0$ we obtain at least one pair $\mathbf{k},\mathbf{h}\in I$, $\mathbf{h}\neq\mathbf{k}$ with $y_{\mathbf{h}}\equiv y_{\mathbf{k}} \pmod 1$. Accordingly, the matrix $\mathbf{B}(\mathbf{r},M)$ contains at least two identical columns and thus is not of full rank. So, a unique solution of $\mathbf{B}(\mathbf{r},M)\mathbf{x}=\mathbf{b}$ is not guaranteed. The smallest eigenvalue of the matrix $\mathbf{B}(\mathbf{r},M)$ is zero. On the other hand, the corresponding upper bound $M^{-1}\Delta(\mathbf{r})$ of the Gerschgorin circle radius of $\mathbf{B}(\mathbf{r},M)$ is infinite. Certainly, the interval $[-\infty,\infty]$ contains all eigenvalues of $\mathbf{B}(\mathbf{r},M)$.

Now let us assume $g_{\pi(1)}>0$. Obviously, the diagonal elements of the considered matrices $\mathbf{B}(\mathbf{r},M)$ are all ones. Let λ_* be an arbitrary eigenvalue of $\mathbf{B}(\mathbf{r},M)$. Following the Gerschgorin circle theorem, there exists at least one index $j\in\{1,\dots,|I|\}$ with

$$|\lambda_*-1| \le \sum_{l=1;\,l\neq j}^{|I|} |K_M(y_{\mathbf{h}_j}-y_{\mathbf{h}_l})|.$$

For $x\in\mathbb{R}\setminus\mathbb{Z}$ we obtain

$$K_M(x) = \frac{1}{M}\sum_{j=0}^{M-1} \mathrm{e}^{2\pi \mathrm{i} j x} = \frac{1}{M}\frac{\mathrm{e}^{2\pi \mathrm{i} M x}-1}{\mathrm{e}^{2\pi \mathrm{i} x}-1} = \frac{\mathrm{e}^{\pi \mathrm{i} M x}}{\mathrm{e}^{\pi \mathrm{i} x}}\frac{\sin \pi M x}{M\sin \pi x}. \tag{4}$$

Due to $2x\le \sin \pi x$ for $x\in(0,1/2]$ we estimate

$$|K_M(x)| = \left|\frac{\sin \pi M x}{M\sin \pi x}\right| \le \frac{1}{|M\sin \pi x|} \le \frac{1}{2Mx}$$

for all $x\in(0,1/2]$. Moreover, we have $|K_M(x)|=|K_M(-x)|$ for $x\in\mathbb{R}$.

We split the index set $J=\{1,\dots,|I|\}\setminus\{j\}$ in the following two subsets

$$J_1 = \{l\in J:\ 0 < y_{\mathbf{h}_j}-y_{\mathbf{h}_l} \bmod 1 \le \frac{1}{2}\}$$

$$\text{and}\quad J_2 = \{l\in J:\ \frac{1}{2} < y_{\mathbf{h}_j}-y_{\mathbf{h}_l} \bmod 1 < 1\}.$$

This yields

$$\begin{aligned}\sum_{l=1;\,l\neq j}^{|I|} |K_M(y_{\mathbf{h}_j}-y_{\mathbf{h}_l})| &= \sum_{l\in J_1} |K_M(y_{\mathbf{h}_j}-y_{\mathbf{h}_l} \bmod 1)| + \sum_{l\in J_2} |K_M(-y_{\mathbf{h}_j}+y_{\mathbf{h}_l} \bmod 1)| \\ &\le \frac{1}{2M}\sum_{l\in J_1}\frac{1}{y_{\mathbf{h}_j}-y_{\mathbf{h}_l} \bmod 1} + \frac{1}{2M}\sum_{l\in J_2}\frac{1}{y_{\mathbf{h}_l}-y_{\mathbf{h}_j} \bmod 1}.\end{aligned}$$

Now, we estimate the differences $y_{\mathbf{h}_j} - y_{\mathbf{h}_l} \bmod 1$. In principle, we interpret the index set J_1 as the indices of the left neighbors of $y_{\mathbf{h}_j}$. So, the distance of the nearest neighbor on the left hand side to $y_{\mathbf{h}_j}$ is at least $g_{\pi(1)}$. Clearly the second nearest neighbor at the left hand side brings a distance of at least $g_{\pi(1)} + g_{\pi(2)}$. In general the k-th nearest neighbor to the left of $y_{\mathbf{h}_j}$ has a distance not less than $\sum_{t=1}^{k} g_{\pi(t)}$ to $y_{\mathbf{h}_j}$. The index set J_2 can be interpreted as the index set of the right neighbors of $y_{\mathbf{h}_j}$ and we determine the lower bounds on the distances in the same way as done for the left neighbors. We obtain

$$\sum_{l=1;\, l\neq j}^{|I|} |K_M(y_{\mathbf{h}_j} - y_{\mathbf{h}_l})| \leq \frac{1}{2M} \sum_{k=1}^{|J_1|} \left(\sum_{t=1}^{k} g_{\pi(t)} \right)^{-1} + \frac{1}{2M} \sum_{k=1}^{|J_2|} \left(\sum_{t=1}^{k} g_{\pi(t)} \right)^{-1}.$$

Using $\sum_{t=1}^{k} g_{\pi(t)} \leq \sum_{t=1}^{r} g_{\pi(t)}$ for $k \leq r$ we balance the two sums and hence

$$\sum_{l=1;\, l\neq j}^{|I|} |K_M(y_{\mathbf{h}_j} - y_{\mathbf{h}_l})| \leq \frac{1}{M} \sum_{k=1}^{\left\lfloor \frac{|I|}{2} \right\rfloor} \left(\sum_{t=1}^{k} g_{\pi(t)} \right)^{-1},$$

which proves the theorem. □

Remark 1. In order to obtain the upper bound of the radii of all Gerschgorin circles in Theorem 1, we estimated the absolute value of the kernel K_M by a monotonically non-increasing upper bound $|2Mx|^{-1}$ in $\left[0, \frac{1}{2}\right]$. Due to $|K_M(\frac{t}{M})| = 0 < \frac{1}{|2t|} = |2M\frac{t}{M}|^{-1}$, for $t \in \mathbb{Z} \setminus M\mathbb{Z}$, the upper bound and the absolute value of the kernel K_M possibly differ widely. In addition, we sorted the pairwise distances of the sorted sequence $\left(y_{\mathbf{h}_j}\right)_{j=1,\ldots,|I|}$ in a worst case scenario. Thus, we also have to expect some differences between the estimation and the exact maximum Gerschgorin radius. Altogether, we obtain an estimation of the maximum Gerschgorin radius which eventually is much larger than the exact maximum Gerschgorin circle radius.

Corollary 1. *With the notation from Theorem 1, $\Delta(\mathbf{r}) < \infty$, and $C > 1$, we determine*

$$M^*(C) = \left\lceil \frac{C+1}{C-1} \Delta(\mathbf{r}) \right\rceil. \tag{5}$$

The condition number of the matrix $\mathbf{B}(\mathbf{r}, M^*(C))$ is bounded by

$$1 \leq \kappa(\mathbf{r}, M^*(C)) \leq C.$$

Proof. Fixing $M^*(C)$ in (5) ensures

$$C \geq \frac{1 + M^*(C)^{-1}\Delta(\mathbf{r})}{1 - M^*(C)^{-1}\Delta(\mathbf{r})} \geq \kappa(\mathbf{r}, M^*(C)) \geq 1,$$

as required. □

Algorithm 1 Computing $\Delta(\mathbf{r})$ from (3)

Input: I frequency index set
$\mathbf{r} \in \mathbb{R}^d$ generating vector

$\Delta(\mathbf{r}) = 0$
for $j = 1, \dots, |I|$ **do**
  $y_j = \mathbf{h}_j \cdot \mathbf{r} \bmod 1$
end for
in-place sort $\mathbf{y}$ in ascending order
$g_1 = 1 + y_1 - y_{|I|}$
for $j = 2, \dots, |I|$ **do**
  $g_j = y_j - y_{j-1}$
end for
in-place sort $\mathbf{g}$ in ascending order
for $j = 1, \dots, \left\lfloor \frac{|I|}{2} \right\rfloor$ **do**
  $\Delta(\mathbf{r}) = \Delta(\mathbf{r}) + \frac{1}{g_j}$
  $g_{j+1} = g_{j+1} + g_j$
end for
Output: $\Delta(\mathbf{r})$

Our approach is to find generated sets $\Lambda(\mathbf{r}, M)$ with small condition numbers $\kappa(\mathbf{r}, M)$. Obviously, the term $\Delta(\mathbf{r})$ should be of our main interest here. The functional Δ is the important term of the upper bound $M^{-1}\Delta(\mathbf{r})$ of the radii of all Gerschgorin circles of the matrix $\mathbf{B}(\mathbf{r}, M)$. Note that $\Delta(\mathbf{r})$ depends on the generating vector $\mathbf{r}$ of the generated set $\Lambda(\mathbf{r}, M)$ but not on M. On the contrary, knowing $\Delta(\mathbf{r})$ one can simply determine a suitable $M^*(C)$ guaranteeing the condition number $\kappa(\mathbf{r}, M^*(C)) \leq C$, see (5).

Algorithm 1 computes the value of $\Delta(\mathbf{r})$ for given I and $\mathbf{r}$ with a complexity of $\mathcal{O}(|I|(\log|I| + d))$.

Another point of view is described by our approach as follows: Let us assume, that we search for a generated set $\Lambda(\mathbf{r}, M)$ such that the condition number $\kappa(\mathbf{r}, M)$ of the matrix $\mathbf{B}(\mathbf{r}, M)$ does not exceed C. We call the generating vector $\mathbf{r}$ suitable in the sense of Theorem 1 if $\Delta(\mathbf{r}) < \infty$, i.e. $|\mathscr{Y}| = |I|$. For each suitable $\mathbf{r}$, Corollary 1 specifies an $M^*_{\mathbf{r}}(C)$ guaranteeing a condition number $\kappa(\mathbf{r}, M^*_{\mathbf{r}}(C))$ not larger than C. So, minimizing the functional Δ directly reduces the cardinality $M^*_{\mathbf{r}}(C)$ of the corresponding generated set for fixed C. This means that the theoretical number of sampling nodes needed for the fast and stable reconstruction of the Fourier coefficients $(f_{\mathbf{k}})_{\mathbf{k}\in I}$ decreases.

Note that a simple lower bound on the functional Δ is given by

$$\Delta(\mathbf{r}) \geq |I| \sum_{k=1}^{\left\lfloor \frac{|I|}{2} \right\rfloor} k^{-1}, \quad \text{for all } \mathbf{r} \in \mathbb{R}^d. \tag{6}$$

We obtain equality, iff the sequence of $(y_{\mathbf{h}})_{\mathbf{h}\in I}$ is an equispaced lattice on the one-dimensional torus. In that case we can translate $y_{\mathbf{h}}$ such that $y_{\mathbf{h}_1} = 0$ and apply an equispaced FFT of length $|I|$ to reconstruct all Fourier coefficients supported on I.

Example 2. Continuing Example 1, we obtain the following rounded results by minimizing Δ using a nonlinear simplex search method:

$$\mathbf{r}_{2,\Delta} = \begin{pmatrix} 0.14266632 \\ 0.40770614 \end{pmatrix} \qquad \text{and} \qquad \mathbf{r}_{5,\Delta} = \begin{pmatrix} 0.24342553 \\ 0.42933779 \\ 0.05122878 \\ 0.88917104 \\ 0.94691925 \end{pmatrix}$$

with

$$\Delta(\mathbf{r}_{2,\Delta}) \approx 113{,}324.3 \qquad \text{and} \qquad \Delta(\mathbf{r}_{5,\Delta}) \approx 161{,}500.5.$$

The corresponding cardinalities $M_{2,\Delta}(10)$ and $M_{5,\Delta}(10)$ of $\Lambda(\mathbf{r}_{s,\Delta}, M_{s,\Delta}(10))$ guaranteeing a condition number $\kappa(\mathbf{r}_{s,\Delta}, M_{s,\Delta}(10)) = \operatorname{cond}_2(\mathbf{B}(\mathbf{r}_{s,\Delta}, M_{s,\Delta}))$ of at most ten are determined by

$$M_{2,\Delta}(10) = 138{,}508 \qquad \text{and} \qquad M_{5,\Delta}(10) = 197{,}390,$$

cf. (5). Of course, these $M_{s,\Delta}(10)$ are simply based on an upper bound of the Gerschgorin radii. We also computed the exact Gerschgorin radii numerically for the generating vectors $\mathbf{r}_{2,\Delta}$ and $\mathbf{r}_{5,\Delta}$ and different M_s and obtain

$$M_2^* = 14{,}989 \qquad \text{and} \qquad M_5^* = 20{,}129$$

guaranteeing condition numbers of $\mathbf{B}(\mathbf{r}_{s,\Delta}, M_s^*)$ smaller or equal ten. In fact, we get condition numbers

$$\kappa(\mathbf{r}_{s,\Delta}, M_s^*) \approx \begin{cases} 2.1847, & \text{for } s = 2, \\ 2.1037, & \text{for } s = 5. \end{cases}$$

Finally, we give the condition numbers of the problem of Example 1. We simply took the generating vectors $\mathbf{r}_{s,\Delta}$ and computed the condition numbers of $\mathbf{B}(\mathbf{r}_{s,\Delta}, M)$ for $M = 16{,}381$ resulting in

$$\kappa(\mathbf{r}_{s,\Delta}, M) \approx \begin{cases} 1.7548, & \text{for } s = 2, \\ 2.9223, & \text{for } s = 5. \end{cases}$$

Obviously, these condition numbers are much smaller than those from Example 1, where we minimized the condition numbers directly. Note that the minimization of

the main term Δ of the upper bound of all Gerschgorin radii is much faster than the direct minimization of the condition number κ. □

4 Numerical Examples

The numerical minimization of $\Delta(\mathbf{r})$ returns minimizers $\mathbf{r}^*$, which are vectors of rational numbers. So one can find a possibly huge $\bar{M}$, such that the generated set $\Lambda(\mathbf{r}^*, \bar{M})$ is a rank-1 lattice. With $M < \bar{M}$, one can interpret the generated set $\Lambda(\mathbf{r}^*, M)$ as the first M elements of the rank-1 lattice $\Lambda(\mathbf{r}^*, \bar{M})$. In general we obtain $M \ll \bar{M}$.

Our numerical examples use generating vectors $\mathbf{r}^*$ found by minimizing Δ, cf. (3). We used the nonlinear simplex search method `fminsearch` of the Optimization Toolbox™ of MATLAB in version 7.14.0.739 (R2012a).

Using the `rand` function, we started the minimization at a randomly chosen vector. Providing a diameter of the actual simplex smaller than 10^{-10} and differences of the function values at the corners of the simplex smaller than $10^{-8}|I|$, we terminated the minimization. Alternatively, we stopped the minimization after a fixed number of function evaluations, even if these conditions are not fulfilled. In order to compute the minimizers of Tables 1 and 2 we limited the number of function evaluations to 3,000. In Table 3, we accepted at most 30,000 function evaluations. The applied simplex search method finds only local minimizers. For each index set I we computed 20 local minimizers of Δ and took as $\mathbf{r}^*$ the local minimizer that yields the smallest value $\Delta(\mathbf{r})$ in order to avoid obtaining minimizers of relatively large local minima.

Besides the computation of $M^*(C) = \left\lceil \frac{C+1}{C-1}\Delta(\mathbf{r}^*) \right\rceil$ from (5) guaranteeing a condition number smaller or equal C we computed exact maximum Gerschgorin circle radii defined by

$$\varrho(\mathbf{r}^*, M) = \max_{\mathbf{k}\in I} \sum_{\mathbf{h}\in I\setminus\{\mathbf{k}\}} |K_M(y_\mathbf{k} - y_\mathbf{h})|$$

for several M, where $K_M(y_\mathbf{k} - y_\mathbf{h})$ describes the elements of the matrix $\mathbf{B}(\mathbf{r}^*, M)$ as defined in (2). The Gerschgorin circle theorem ensures that the condition $\varrho(\mathbf{r}^*, M) \le \frac{C-1}{C+1}$ implies $\kappa(\mathbf{r}^*, M) \le C$.

For a fixed vector $\mathbf{r}^*$ we define $M^*_\mathrm{G}(C)$ as the smallest power of 2 such that the exact maximum Gerschgorin circle radius ensures a condition number not larger than C,

$$M^*_\mathrm{G}(C) = \min_{n\in\mathbb{N}} \left\{ 2^n : \varrho(\mathbf{r}^*, 2^n) \le \frac{C-1}{C+1} \right\}.$$

Moreover, we call $\kappa_{\mathsf{G}}^*(M) := \frac{1+\varrho(\mathbf{r}^*,M)}{1-\varrho(\mathbf{r}^*,M)}$ the estimation of the condition number of the matrix $\mathbf{B}(\mathbf{r}^*, M)$ based on the exact maximum Gerschgorin circle radius. Certainly, we have to assume $\varrho(\mathbf{r}^*, M) < 1$ to estimate the condition number $\kappa^*(M) := \kappa(\mathbf{r}^*, M) \leq \kappa_{\mathsf{G}}^*(M)$. Otherwise, i.e. $\varrho(\mathbf{r}^*, M) \geq 1$, we obtain $0 \in [1-\varrho(\mathbf{r}^*, M), 1+\varrho(\mathbf{r}^*, M)]$ and so zero is a candidate for the smallest eigenvalue of $\mathbf{B}(\mathbf{r}^*, M)$. Consequently, $\kappa_G^*(M) < 0$ does not bound the condition number of $\mathbf{B}(\mathbf{r}^*, M)$.

Applying (4), the computational costs for calculating $M_{\mathsf{G}}^*(C)$ is bounded by $c|I|^2 \log_2(M^*(C))$, where c is independent of I and $M^*(C)$. In general, this computation is not necessary in order to obtain stable spatial discretizations, but the costs for computing $M_{\mathsf{G}}^*(C)$ can be quickly compensated using the generated set $\Lambda(\mathbf{r}^*, M_{\mathsf{G}}^*(C))$ instead of $\Lambda(\mathbf{r}^*, M^*(C))$ in practical applications. In particular for a frequently used fixed index set I and generating vector $\mathbf{r}^*$, the generated set $\Lambda(\mathbf{r}^*, M_{\mathsf{G}}^*(C))$ with cardinality $M_{\mathsf{G}}^*(C) < M^*(C)$ saves sampling and computational costs.

4.1 Weighted Hyperbolic Crosses

Tables 1 and 2 show some numerical examples for weighted hyperbolic crosses $H_N^{d,\boldsymbol{\gamma}}$ as frequency index sets I. In Table 1 we consider weights $\boldsymbol{\gamma}_a = \left(2^{-1}\right)_{s\in\mathbb{N}}$, refinement $N = 32$, and dimensions d from 2 up to 12 as parameters and determine suitable generated sets for reconstructing trigonometric polynomials with frequencies supported on I. Table 2 contains similar results for weights $\boldsymbol{\gamma}_b = (3^{-1})_{s\in\mathbb{N}}$, refinement $N = 48$, and dimensions d up to 27. The parameters chosen ensure $H_{48}^{d,\boldsymbol{\gamma}_b} \subset H_{32}^{d,\boldsymbol{\gamma}_a}$. In detail, we obtain

$$H_{32}^{d,\boldsymbol{\gamma}_a} \setminus H_{48}^{d,\boldsymbol{\gamma}_b} \subset \left\{ \mathbf{k} \in H_{32}^{d,\boldsymbol{\gamma}_a} : \|\mathbf{k}\|_0 = \sum_{s=1}^{d} (1-\delta_0(k_s)) > 1 \right\},$$

i.e. the hyperbolic cross $H_{48}^{d,\boldsymbol{\gamma}_b}$ is sparser than $H_{32}^{d,\boldsymbol{\gamma}_a}$ in mixed indices only.

The first column of these tables shows the dimension d and the second column the cardinality of the considered frequency index set $H_N^{d,\boldsymbol{\gamma}}$. We minimize Δ like described above and obtain the resulting theoretical number of sampling points $M^*(10)$ needed to ensure a condition number of $\mathbf{B}(\mathbf{r}^*, M^*(10))$ not larger than ten. $M^*(10)$ is listed in column 3. Fixing I and $\mathbf{r}^*$, in column 4 we present the smallest power of 2 $M_{\mathsf{G}}^*(10)$ guaranteeing that the exact maximum Gerschgorin radius is not larger than $\frac{9}{11}$. This restriction ensures that even the condition number of $\mathbf{B}(\mathbf{r}^*, M_{\mathsf{G}}^*(10))$ is not larger than ten. In other words, sampling along the first $M_{\mathsf{G}}^*(10)$ multiples of the generating vector $\mathbf{r}^*$ already guarantees a stable reconstruction of all multivariate trigonometric polynomials with frequencies supported on $H_N^{d,\boldsymbol{\gamma}}$. We specify the corresponding estimations of the condition numbers

based on the maximum Gerschgorin radius labeled with $\kappa_{\mathrm{G}}^*(M_{\mathrm{G}}^*(10))$ in column 5. Column 6 shows the corresponding exact condition numbers $\kappa^*(M_{\mathrm{G}}^*(10))$.

Table 1 Cardinalities $M^*(10)$, $M_{\mathrm{G}}^*(10)$, and $M^{**}(10)$ of generated sets generated by vectors $\mathbf{r}^*$ that are found by minimizing Δ for index sets I which are weighted hyperbolic crosses $H_{32}^{d,\boldsymbol{\gamma}_a}$ with weights $\boldsymbol{\gamma}_a = (2^{-1})_{s\in\mathbb{N}}$ and dimensions $d = 2,\dots,12$; additionally, the condition numbers κ^* and upper bounds κ_{G}^* of the corresponding matrices $\mathbf{B}(\mathbf{r}^*, M_{\mathrm{G}}^*(10))$ and $\mathbf{B}(\mathbf{r}^*, M^{**}(10))$, respectively.

d	$\lvert H_{32}^{d,\boldsymbol{\gamma}_a}\rvert$	$M^*(10)$	$M_{\mathrm{G}}^*(10)$	$\kappa_{\mathrm{G}}^*(M_{\mathrm{G}}^*(10))$	$\kappa^*(M_{\mathrm{G}}^*(10))$	$M^{**}(10)$	$\kappa_{\mathrm{G}}^*(M^{**}(10))$	$\kappa^*(M^{**}(10))$
2	145	2,216	1,024	2.8466	1.4540	370	11.9876	2.0705
3	441	9,709	4,096	2.8441	1.2811	1,408	−7.1782	1.8418
4	1,105	48,328	8,192	8.1211	1.9993	6,291	−15.0278	2.1016
5	2,433	151,727	32,768	3.0456	1.4690	18,119	4.5528	1.4570
6	4,865	471,958	65,536	6.1468	1.8703	52,492	4.3597	1.6842
7	9,017	1,115,494	131,072	5.2118	1.9046	116,850	7.1495	1.8891
8	15,713	2,538,107	262,144	4.3721	1.7571	252,533	4.1645	1.7939
9	26,017	6,256,440	524,288	5.5048	2.1663	595,180	3.2571	1.7159
10	41,265	15,910,747	2,097,152	3.1616	1.7769	1,454,830	4.1586	1.8770
11	63,097	29,880,128	2,097,152	5.7801	2.5378	2,637,334	2.9012	1.8379
12	93,489	46,057,959	4,194,304	4.4024	1.7782	4,065,252	4.0095	1.7426

Regarding both tables, one observes that the values of $M_{\mathrm{G}}^*(10)$ behave like

$$M_{\mathrm{G}}^*(10) \sim M^{**}(10) := \left\lfloor M^*(10) \left(\sum_{k=1}^{\left\lfloor \frac{|I|}{2} \right\rfloor} k^{-1} \right)^{-1} \right\rfloor .$$

We listed the values of $M^{**}(10)$. The equispaced case discussed in the context of (6) illustrates that this observation is being caused by the construction of the functional Δ from (3). We also computed the exact maximum Gerschgorin circle radii $\varrho(\mathbf{r}^*, M^{**}(10))$, the estimator of the condition number $\kappa_{\mathrm{G}}^*(M^{**}(10))$, and the exact condition numbers $\kappa^*(M^{**}(10))$ of the corresponding matrices $\mathbf{B}(\mathbf{r}^*, M^{**}(10))$. One obtains a few exceptions only where the maximum Gerschgorin circle radii $\varrho(\mathbf{r}^*, M^{**}(10))$ strongly exceeds the bound $\frac{9}{11}$ that guarantees an upper bound $\kappa_{\mathrm{G}}^*(M^{**}(10))$ of the condition number $\kappa^*(M^{**}(10))$ smaller or equal ten. Nevertheless, all exact condition numbers $\kappa^*(M^{**}(10))$ of the matrices $\mathbf{B}(\mathbf{r}^*, M^{**}(10))$ do not exceed three in Table 1 and two in Table 2, evidently.

4.2 Randomly Chosen Index Sets

As described above, our approach finds stable spatial discretizations of trigonometric polynomials with frequencies supported on arbitrary known index sets I.

Table 2 Cardinalities $M^*(10)$, $M^*_{\mathrm{G}}(10)$, and $M^{**}(10)$ of generated sets generated by vectors $\mathbf{r}^*$ that are found by minimizing Δ for index sets I which are weighted hyperbolic crosses $H_{48}^{d,\boldsymbol{\gamma}_b}$ with weights $\boldsymbol{\gamma}_b = (3^{-1})_{s\in\mathbb{N}}$ and dimensions $d = 2, \ldots, 27$; additionally, the condition numbers κ^* and upper bounds κ^*_{G} of the corresponding matrices $\mathbf{B}(\mathbf{r}^*, M^*_{\mathrm{G}}(10))$ and $\mathbf{B}(\mathbf{r}^*, M^{**}(10))$, respectively.

d	$\lvert H_{48}^{d,\boldsymbol{\gamma}_b}\rvert$	$M^*(10)$	$M^*_{\mathrm{G}}(10)$	$\kappa^*_{\mathrm{G}}(M^*_{\mathrm{G}}(10))$	$\kappa^*(M^*_{\mathrm{G}}(10))$	$M^{**}(10)$	$\kappa^*_{\mathrm{G}}(M^{**}(10))$	$\kappa^*(M^{**}(10))$
2	105	1,878	512	5.6271	1.8467	354	8.9423	1.7548
3	225	5,916	2,048	2.6898	1.4053	1,006	31.1502	1.9347
4	401	16,432	4,096	5.4966	1.9237	2,588	5.0743	1.4437
5	641	34,464	8,192	3.8947	1.7342	5,110	2.6683	1.2588
6	953	91,526	16,384	4.5491	1.9292	12,912	3.1832	1.6643
7	1,345	120,893	16,384	5.3483	1.8287	16,351	5.3015	1.8288
8	1,825	244,266	32,768	3.1868	1.7918	31,856	3.6431	1.7963
9	2,401	400,917	65,536	2.1847	1.5143	50,639	3.1209	1.6695
10	3,081	595,978	65,536	3.3692	1.8932	73,163	3.0289	1.7584
11	3,873	960,647	131,072	2.3670	1.5973	114,946	2.3611	1.5942
12	4,785	1,265,910	131,072	3.2820	1.9410	147,990	2.3929	1.6242
13	5,825	1,875,694	262,144	2.6817	1.7231	214,662	2.4461	1.5622
14	7,001	2,135,009	262,144	2.9406	1.7604	239,603	2.9177	1.7633
15	8,321	3,310,334	262,144	5.2203	2.8743	364,835	2.7691	1.7001
16	9,793	4,831,312	524,288	2.6954	1.6401	523,570	2.6728	1.6412
17	11,425	6,156,192	1,048,576	3.0616	1.8571	656,735	2.4556	1.7802
18	13,225	7,735,764	1,048,576	2.9312	1.8649	813,162	2.5909	1.8368
19	15,201	9,885,874	1,048,576	2.0752	1.6079	1,024,862	2.1632	1.5777
20	17,361	11,784,210	1,048,576	2.8501	2.0048	1,205,779	2.2593	1.6075
21	19,713	16,342,704	2,097,152	2.8862	1.8002	1,651,639	2.0994	1.5800
22	22,265	18,916,637	2,097,152	1.9674	1.5404	1,889,453	2.0036	1.5755
23	25,025	27,027,375	2,097,152	4.1966	2.7862	2,669,617	2.1054	1.6669
24	28,001	30,693,609	4,194,304	2.8452	1.8359	2,999,686	1.8831	1.5614
25	31,201	37,040,314	4,194,304	1.8965	1.5780	3,583,403	1.9974	1.6094
26	34,633	41,051,986	4,194,304	2.7275	1.8536	3,933,160	2.8633	1.8562
27	38,305	46,404,278	4,194,304	2.0293	1.6813	4,328,544	2.0485	1.5950

So, we consider index sets I randomly chosen from the d-dimensional cube $[-128, 128]^d \subset \mathbb{Z}^d$ in Table 3. It presents results for several cardinalities of the index set I and dimensions d which are powers of 2. The content of each column is as described above. To achieve these results we increased the maximum number of the allowed function evaluations. In higher dimensions, this seems to be necessary to suitably decrease the diameter of the simplex in the used optimization method. For comparability we chose this parameter independent on the dimension d. So we allowed at most 30,000 function evaluations to minimize $\Delta(\mathbf{r})$.

We see that in principle the cardinalities $M^*(10)$, $M^*_{\mathrm{G}}(10)$, and $M^{**}(10)$ mildly decrease with growing dimensions. In other words, an increasing number of degrees of freedom of the functional Δ results in a lower minimal value.

Furthermore, we observe a growing oversampling with increasing cardinality of the index set I. For a doubled cardinality of I, the values of $M^*(10)$, $M^*_{\mathrm{G}}(10)$,

and $M^{**}(10)$ increase approximately fourfold. Thus, the cardinalities of the found generated sets grow nearly quadratical in the cardinality of the index set I. Taking into account some modifications of the results of Theorem 3.2 in [6], we also expect this behavior for rank-1 lattices which bring a full column rank of the corresponding Fourier matrix $\mathbf{A}$. Accordingly, we expect to evaluate and reconstruct the multivariate trigonometric polynomial with frequencies supported on I with a complexity of $\mathcal{O}\left(|I|^2 \log|I| + (|\log\varepsilon| + d)|I|\right)$. Precomputing the set $\mathscr{Y}$ and saving the necessarily bijective mapping $I \to [0,1) : \mathbf{h} \mapsto \mathbf{h}\cdot\mathbf{r} \pmod 1$ we reduce the complexity to $\mathcal{O}\left(|I|^2 \log|I| + |\log\varepsilon||I|\right)$, which is independent of the spatial dimension d.

Table 3 Cardinalities $M^*(10)$, $M^*_{\mathrm{G}}(10)$, and $M^{**}(10)$ of generated sets generated by vectors $\mathbf{r}^*$ that are found by minimizing Δ for index sets I of dimensions $d = 2^n$, $n = 1,\dots,8$; elements of I taken from $[-128, 128]^d \cap \mathbb{Z}^d$ uniformly at random; additionally, condition numbers κ^* and upper bounds κ^*_{G} of the corresponding matrices $\mathbf{B}(\mathbf{r}^*, M^*_{\mathrm{G}}(10))$ and $\mathbf{B}(\mathbf{r}^*, M^{**}(10))$, respectively.

d	$\lvert I\rvert$	$M^*(10)$	$M^*_{\mathrm{G}}(10)$	$\kappa^*_{\mathrm{G}}(M^*_{\mathrm{G}}(10))$	$\kappa^*(M^*_{\mathrm{G}}(10))$	$M^{**}(10)$	$\kappa^*_{\mathrm{G}}(M^{**}(10))$	$\kappa^*(M^{**}(10))$
2	750	195,091	32,768	1.8286	1.4154	29,988	2.1015	1.6151
4	750	196,928	32,768	2.8119	1.9440	30,271	3.9359	2.2560
8	750	166,797	32,768	1.9991	1.5679	25,639	2.2685	1.7329
16	750	144,334	16,384	7.0620	2.8097	22,186	2.5448	1.7345
32	750	104,756	16,384	2.8085	1.6086	16,102	2.6035	1.6000
64	750	61,856	8,192	7.7925	2.0178	9,508	5.9218	1.8894
128	750	62,873	8,192	8.3523	2.1612	9,664	6.1158	1.9719
256	750	54,611	8,192	5.3793	1.8811	8,394	5.3329	1.9067
2	1,500	412,137	65,536	1.7069	1.2490	57,257	1.8188	1.3153
4	1,500	851,612	131,072	2.3944	1.8503	118,313	3.6857	2.4721
8	1,500	647,439	65,536	5.0975	3.2850	89,947	3.4072	1.8601
16	1,500	619,395	65,536	6.2502	3.2448	86,051	2.2943	1.7509
32	1,500	411,622	65,536	2.0989	1.5788	57,185	2.6901	1.7571
64	1,500	324,658	32,768	6.7930	2.6120	45,104	2.5211	1.7025
128	1,500	254,375	32,768	3.2401	1.8037	35,339	2.6962	1.8009
256	1,500	226,842	32,768	3.5078	1.8600	31,514	3.4103	1.8070
2	3,000	547,361	65,536	1.9555	1.2360	69,367	2.0029	1.2113
4	3,000	3,265,505	262,144	8.4853	6.5363	413,838	2.6677	2.4452
8	3,000	3,078,366	262,144	8.7348	4.9750	390,122	2.4379	2.0957
16	3,000	2,434,510	262,144	3.0088	2.6453	308,526	2.0252	1.8461
32	3,000	1,675,774	262,144	1.9351	1.6033	212,371	2.2124	1.6160
64	3,000	1,213,198	131,072	5.4684	2.8743	153,749	2.5438	1.7883
128	3,000	1,034,141	131,072	2.3803	1.6471	131,057	2.3964	1.6471
256	3,000	805,916	131,072	3.1876	1.8898	102,134	2.7344	1.7359
2	6,000	674,627	65,536	6.0235	1.6905	78,593	6.9360	1.5907
4	6,000	13,250,802	1,048,576	7.2950	6.1781	1,543,707	2.9009	2.6842
8	6,000	11,501,870	2,097,152	1.8113	1.5978	1,339,958	4.3485	2.9201
16	6,000	10,191,192	1,048,576	3.2384	2.8720	1,187,265	2.3771	2.1829
32	6,000	7,573,185	1,048,576	2.3761	1.7520	882,269	2.1095	1.8821
64	6,000	5,661,152	524,288	6.4483	3.8637	659,519	2.2256	1.7379
128	6,000	3,777,565	524,288	2.2653	1.6630	440,083	2.1709	1.6788
256	6,000	3,311,017	524,288	2.6142	1.8378	385,730	2.3017	1.6661

5 Summary

The concept of generated sets provides mildly oversampled and stable spatial discretizations for multivariate trigonometric polynomials with frequencies supported on index sets I of reasonable cardinalities. In addition, the NFFT and some simple precomputations allow for the fast evaluation of multivariate trigonometric polynomials f at all sampling nodes of generated sets $\Lambda(\mathbf{r}, M)$. Assuming the condition number $\mathrm{cond}_2(\mathbf{B}(\mathbf{r}, M))$ equal or near one, the conjugate gradient method using the NFFT and its adjoint provide the fast, stable, and unique reconstruction of f from samples along the generated set $\Lambda(\mathbf{r}, M)$. Our approach imposes only one important condition on the generating vector $\mathbf{r} \in \mathbb{R}^d$: Successive elements of the one-dimensional sampling scheme $\mathscr{Y} = \{\mathbf{k} \cdot \mathbf{r} \bmod 1 : \mathbf{k} \in I\}$ should have relatively large distances.

Acknowledgements The author thanks Daniel Potts and Stefan Kunis for numerous valuable discussions. Moreover, he thanks the referees for their very useful suggestions for improvements and he gratefully acknowledges support by German Research Foundation within the project KU 2557/1-1.

References

1. Axelsson, O.: Iterative Solution Methods. Cambridge University Press, Cambridge (1996)
2. Cools, R., Kuo, F.Y., Nuyens, D.: Constructing lattice rules based on weighted degree of exactness and worst case error. Computing **87**, 63–89 (2010)
3. Cools, R., Sloan, I.H.: Minimal cubature formulae of trigonometric degree. Math. Comp. **65**, 1583–1600 (1996)
4. Gerschgorin, S.: Über die Abgrenzung der Eigenwerte einer Matrix. Izv. Akad. Nauk SSSR Otd. Fiz.-Mat. Nauk **1931**, 749–754 (1931)
5. Griebel, M., Hamaekers, J.: Fast discrete Fourier transform on generalized sparse grids. INS Preprint No. 1305 (2013)
6. Kämmerer, L.: Reconstructing Hyperbolic Cross Trigonometric Polynomials by Sampling along Rank-1 Lattices. SIAM J. Numer. Anal. **51**, 2773–2796 (2013)
7. Keiner, J., Kunis, S., Potts, D.: Using NFFT3 – a software library for various nonequispaced fast Fourier transforms. ACM Trans. Math. Software **36**, 1–30 (2009)
8. Knapek, S.: Hyperbolic cross approximation of integral operators with smooth kernel. Technical Report 665, SFB 256, University of Bonn (2000)
9. Kunis, S., Potts, D.: Stability results for scattered data interpolation by trigonometric polynomials. SIAM J. Sci. Comput. **29**, 1403–1419 (2007)
10. Li, D., Hickernell, F.J.: Trigonometric spectral collocation methods on lattices. In: Recent Advances in Scientific Computing and Partial Differential Equations. AMS Series in Contemporary Mathematics, vol. 330, 121–132 (2003)
11. Munthe-Kaas, H., Sørevik, T.: Multidimensional pseudo-spectral methods on lattice grids. Appl. Numer. Math. **62**, 155–165 (2012)
12. Nelder, J.A., Mead, R.: A simplex method for function minimization. Comput. J. **7**, 308–313 (1965)

13. Sickel, W., Ullrich, T.: The Smolyak algorithm, sampling on sparse grids and function spaces of dominating mixed smoothness. East J. Approx. **13**, 387–425 (2007)
14. Sloan, I.H., Joe, S.: Lattice Methods for Multiple Integration. Clarendon Press, Oxford (1994)
15. Sloan, I.H., Kachoyan, P.J.: Lattice methods for multiple integration: Theory, error analysis and examples. SIAM J. Numer. Anal. **24**, 116–128 (1987)

Bayesian Approaches to the Design of Markov Chain Monte Carlo Samplers

Jonathan M. Keith and Christian M. Davey

Abstract In the decades since Markov chain Monte Carlo methods were first introduced, they have revolutionised Bayesian approaches to statistical inference. Each new advance in MCMC methodology produces near immediate benefits for Bayesian practitioners, expanding the range of problems they can feasibly solve. In this paper, we explore ways in which Bayesian approaches can return something of the debt owed to MCMC, by using explicitly Bayesian concepts to aid in the design of MCMC samplers. The art of efficient MCMC sampling lies in designing a Markov process that (a) has the required limiting distribution, (b) has good convergence and mixing properties and (c) can be implemented in a computationally efficient manner. In this paper, we explore the idea that the selection of an appropriate process, and in particular the tuning of the parameters of the process to achieve the above goals, can be regarded as a problem of estimation. As such, it is amenable to a conventional Bayesian approach, in which a prior distribution for optimal parameters of the sampler is specified, data relevant to sampler performance is obtained and a posterior distribution for optimal parameters is formed. Sampling from this posterior distribution can then be incorporated into the MCMC sampler to produce an adaptive method. We present a new MCMC algorithm for Bayesian adaptive Metropolis-Hasting sampling (BAMS), using an explicitly Bayesian inference to update the proposal distribution. We show that author Keith's earlier Bayesian adaptive independence sampler (BAIS) and a new Bayesian adaptive random walk sampler (BARS) emerge as instances. More important than either of these instances, BAMS provides a general framework within which to explore adaptive schemes that are guaranteed to converge to the required limiting distribution.

J.M. Keith · C.M. Davey
School of Mathematical Sciences, Monash University, Clayton, VIC 3800, Australia
e-mail: jonathan.keith@monash.edu; christian.davey@monash.edu

J. Dick et al. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2012*,
Springer Proceedings in Mathematics & Statistics 65,
DOI 10.1007/978-3-642-41095-6_22, © Springer-Verlag Berlin Heidelberg 2013

1 Introduction

The goal of this paper is to initiate the study of a new class of adaptive Markov chain Monte Carlo (MCMC) techniques, which we refer to as *Bayesian adaptive Metropolis-Hastings sampling (BAMS)*. Like any adaptive MCMC technique, the goal is to progressively improve the sampling procedure, as one learns more about the target distribution $f(x)$ (which prior to sampling is considered a 'black box'). Reviews of the extensive literature on existing rigorously justified adaptive MCMC methods are available elsewhere [1, 2, 4, 11].

The novelty of the approach presented here is to cast the learning procedure as a problem of Bayesian inference. In other words, the approach considers the parameters of the adaptive procedure (and perhaps even the choice of adaptive procedure itself) as unknowns to be inferred from data consisting of a sample of points $(x_1, x_2, \ldots)$ and evaluations of the target density at those points $(f(x_1), f(x_2), \ldots)$.

In earlier work [8], author Keith presented the first algorithm of the general type to be developed here. This algorithm involved multiple parallel Markov chains, each implementing an independence sampler of Metropolis-Hastings type, with a common proposal distribution. The parameters of the proposal distribution were adaptively estimated using Bayesian inference. In this section, we review Metropolis-Hastings samplers and the Bayesian adaptive independence sampler (BAIS) before presenting a new generalisation.

The paper is structured as follows. Section 1.1 recapitulates the well known Metropolis-Hastings algorithm, using a non-standard framework that will be necessary for constructing the generalisation. Section 1.2 revises the Bayesian adaptive independence sampler. In Sect. 2.1, BAIS is generalised to produce the main result of this paper, a new Bayesian adaptive Metropolis-Hasting sampler. In Sect. 2.2, we briefly show that BAIS is an instance of BAMS, and in Sect. 2.3, we present a new instance of BAMS: a Bayesian adaptive random walk sampler (BARS). A simple example illustrating the use of BARS is presented in Sect. 3. Finally, in Sect. 4, some directions for further research are discussed.

1.1 Metropolis-Hastings Sampling

A major class of MCMC method is the Metropolis-Hastings algorithm [5, 9]. The algorithm is based on an arbitrary proposal density that is used to generate potential new values which are then either accepted or rejected. Let $\mathscr{X}$ be the *target space* from which elements are to be sampled and let $f(x)$ be the density of the *target distribution*, defined over $\mathscr{X}$. Let $g(y|x)$ be the *proposal density*, which determines how the proposed new element y is selected when the current element is x. In what follows, it will be convenient to describe the Metropolis-Hastings sampler in a non-standard way, using a construction described in Keith et al. [7]. A Markov chain is constructed on the space $\mathscr{X} \times \mathscr{X}$ with the limiting density

$$h(x, y) = f(x)g(y|x).$$

Note that the marginal density of x is the target distribution f, so the required sample is obtained by sampling from h and discarding the y values. The required Markov chain is generated by the following algorithm.

Algorithm 1 (Metropolis-Hastings). *Given initial values* $(x^{(0)}, y^{(0)})$*, set* $t := 1$ *and iterate the following:*

1. *Draw* y' *from* $h(y|x^{(t-1)}) = g(y|x^{(t-1)})$
2. *Draw* u *uniformly and randomly from the interval* $(0, 1)$ *and set*

$$(x^{(t)}, y^{(t)}) = \begin{cases} (y', x^{(t-1)}) \text{ if } u < \alpha(x^{(t-1)}, y') \\ (x^{(t-1)}, y') \text{ otherwise} \end{cases}$$

where

$$\alpha(x, y) = \min\left\{1, \frac{h(y, x)}{h(x, y)}\right\}.$$

The distribution $h(x, y)$ is stationary with respect to this process. To see this, note first that the transition in Step 1 leaves the marginal distribution of x unchanged, and trivially preserves the conditional distribution $h(y|x^{(t-1)})$. Moreover, h satisfies the detailed balance equations with respect to the transition function implied by Step 2. To see this, the transition function can be defined as follows:

$$R((x, y), (x', y')) = \begin{cases} \alpha(x, y) & \text{if } (x', y') = (y, x) \\ 1 - \alpha(x, y) & \text{if } (x', y') = (x, y) \\ 0 & \text{otherwise.} \end{cases}$$

Note that $h(x, y)\alpha(x, y) = h(y, x)\alpha(y, x)$ and hence

$$h(x, y)R((x, y), (x', y')) = h(x', y')R((x', y'), (x, y))$$

for all (x, y) and (x', y'). Stationarity of h follows by integrating both sides of the detailed balance equations over (x, y).

Two widely used special cases of the Metropolis-Hastings sampler are the *independence sampler* and the *random walk sampler*. In the independence sampler, the proposal density $h(y|x)$ is independent of x. The proposal density can thus be written $h(y)$ and can be described as a global proposal mechanism. In practice, h is chosen to be a distribution that is easy to sample, yet as similar as possible to the target distribution. In the random walk sampler, $h(y|x)$ is a distribution centred at x, and can therefore be described as a local proposal mechanism. If the difference $y - x$ is defined, h is a function of $y - x$, and is a density centred at 0.

1.2 BAIS

The first algorithm to employ a specifically Bayesian strategy for adaptive purposes in MCMC was the Bayesian adaptive independence sampler of Keith *et al.* [8], which uses a Bayesian approach to infer a parametric approximation to the target distribution $f(x)$ based on data obtained from N parallel chains. This approximation is then used as the new proposal distribution for each of the N chains, in the manner of Metropolis-Hastings independence samplers. The algorithm cycles between updating each individual chain and updating the parameter θ of the proposal distribution.

Consider a target density f defined on $\Re^d$. BAIS involves inferring a multivariate normal approximation to f, with mean μ and covariance matrix Σ. Put $\theta = (\mu, \Sigma)$. Given a sample $\mathbf{x} = (x_1, \ldots, x_N)$ and assuming a non-informative prior, the resulting posterior distribution for θ is the product of a multivariate normal density and an inverse Wishart density with $N - 1$ degrees of freedom:

$$p(\theta|\mathbf{x}) = \text{Norm}(\mu|\overline{x}, \Sigma/N) \cdot \text{InvW}_{N-1}(\Sigma|S) \tag{1}$$

where

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_n \tag{2}$$

and

$$S = \sum_{i=1}^{N}(x_n - \overline{x})(x_n - \overline{x})^T. \tag{3}$$

This estimate of θ is then used to construct a new proposal density $\text{Norm}(y|\theta)$. The BAIS algorithm can now be expressed as follows.

Algorithm 2 (BAIS). *Given initial values* $(x_1^{(0)}, \ldots, x_N^{(0)})$ *and* $\theta^{(0)} = (\mu^{(0)}, \Sigma^{(0)})$, *set* $t := 1$ *and iterate the following:*

1. *Draw* $\theta^{(t)}$ *from* $p(\theta|\mathbf{x}^{(t-1)})$.
2. *For each* $i = 1, \ldots, N$, *draw* y' *from* $\text{Norm}(y|\theta^{(t)})$ *and draw* u *uniformly and randomly from the interval* $(0, 1)$. *Set* $\mathbf{x}' = (x_1^{(t)}, \ldots, x_{i-1}^{(t)}, x_i^{(t-1)}, \ldots, x_N^{(t-1)})$ *and set:*

$$x_i^{(t)} = \begin{cases} y' & \text{if } u < \alpha_i(\mathbf{x}', y', \theta^{(t)}) \\ x_i^{(t-1)} & \text{otherwise} \end{cases}$$

where

$$\alpha_i(\mathbf{x}, y, \theta) = \min\left\{1, \frac{f(y)\text{Norm}(x_i|\theta)p(\theta|\mathbf{x}^*)}{f(x_i)\text{Norm}(y|\theta)p(\theta|\mathbf{x})}\right\}$$

and $\mathbf{x}^*$ *is* $\mathbf{x}$ *with* x_i *replaced by* y.

Note that $\alpha_i(\mathbf{x}, y, \theta)$ is not the standard Metropolis-Hastings acceptance ratio, because the parameter θ has been augmented to the search space, so that the augmented target distribution is

$$p(\theta|\mathbf{x}) \prod_{i=1}^{N} f(x_i).$$

BAIS can therefore be considered an auxiliary variable method [6] (as can the new BAMS and BARS algorithms described below). This augmentation is crucial, because it incorporates the adaptive parameter into the Markov chain, ensuring convergence to the target distribution. This solves a difficult problem for adaptive procedures: when the parameters of the process are constantly changing, there is no guarantee that there is a unique stationary distribution, or even that the chain will converge. Many adaptive techniques solve this problem by terminating the adaptive part of the algorithm once some objective has been achieved, so that the actual sampling of the target distribution is non-adaptive. The augmentation technique allows the adaptation to continue indefinitely.

Below we use a similar technique to adaptively estimate the scale parameter of a random walk Metropolis sampler. Although the justification of the adaptive part of the procedure is slightly more complex than for BAIS, the resulting algorithm is simpler in that the acceptance ratios do not contain an adjustment factor dependent on the sampling procedure.

2 Methods

In this section we introduce a general framework for Bayesian adaptive Metropolis-Hastings sampling, applicable to both global and local proposal mechanisms. We also show that both BAIS and a new Bayesian adaptive random walk sampler emerge as instances.

2.1 Bayesian Adaptive Metropolis-Hastings Sampling

First, we generalise BAIS to provide an adaptive scheme for a parallel implementation of *any* Metropolis-Hastings sampler, including both independence and random walk samplers. The generalisation requires that the following objects be defined:

1. A *performance vector* θ containing one or more parameters relevant to sampler performance,
2. A Bayesian model, encapsulated as a posterior distribution $p(\theta|\mathbf{x})$, for estimating the performance vector of the sampler given a sample $\mathbf{x} = (x_1, \ldots, x_N)$, and
3. A *proposal map* τ assigning to each performance vector θ a corresponding Metropolis-Hastings proposal distribution $\tau_\theta(y|x)$.

The term *performance vector* coined above should be interpreted broadly. In BAIS, the performance vector consists of the mean and covariance matrix of the multivariate normal model that is used as the proposal distribution. In general, the performance vector is merely a set of parameters to be estimated based on a sample, the values of which will be used to select the proposal distribution. The term is used to indicate that these parameters are estimated for the purpose of improving sampler performance, and may include some measure of sampler performance, or model relating proposal to performance.

As in BAIS, N parallel sampling chains will be constructed. Each of these will involve generating pairs $(x_i, y_i) \in \mathscr{X} \times \mathscr{X}$ in the manner of Sect. 1.1, so that the limiting distribution of x_i is the target distribution. A parameter chain will also be constructed, and will involve sampling θ from the parameter space Θ. The combined chain is thus constructed on the space:

$$(\mathscr{X} \times \mathscr{X})^N \times \Theta$$

such that the density of the limiting distribution is:

$$h(\mathbf{x}, \mathbf{y}, \theta) = f(\mathbf{x}) p(\theta|\mathbf{x}) \tau(\mathbf{y}|\mathbf{x}, \theta)$$

where $\mathbf{x} = (x_1, \ldots, x_N)$, $\mathbf{y} = (y_1, \ldots, y_N)$, $f(\mathbf{x}) = \prod_{i=1}^N f(x_i)$ and

$$\tau(\mathbf{y}|\mathbf{x}, \theta) = \prod_{i=1}^N \tau_\theta(y_i|x_i).$$

It is immediately clear that this is the density of a true distribution.

A Markov chain with the required limiting distribution $h(\mathbf{x}, \mathbf{y}, \theta)$ is generated by the following algorithm.

Algorithm 3 (BAMS). *Given initial values* $(\mathbf{x}^{(0)}, \mathbf{y}^{(0)}, \theta^{(0)})$*, set* $t := 1$ *and iterate the following:*

1. *Draw* $(\mathbf{y}', \theta^{(t)})$ *from* $h(\mathbf{y}, \theta|\mathbf{x}^{(t-1)})$ *by*
   (a) *Drawing* $\theta^{(t)}$ *from* $p(\theta|\mathbf{x}^{(t-1)})$
   (b) *Drawing* $\mathbf{y}'$ *from* $\tau(\mathbf{y}|\mathbf{x}^{(t-1)}, \theta^{(t)})$
2. *For each* $i = 1, \ldots, N$*, draw* u *uniformly and randomly from the interval* $(0, 1)$*, set* $\mathbf{x}' = (x_1^{(t)}, \ldots, x_{i-1}^{(t)}, x_i^{(t-1)}, \ldots, x_N^{(t-1)})$ *and set*

$$(x_i^{(t)}, y_i^{(t)}) = \begin{cases} (y_i', x_i^{(t-1)}) & \text{if } u < \alpha_i(\mathbf{x}', \mathbf{y}', \theta^{(t)}) \\ (x_i^{(t-1)}, y_i') & \text{otherwise} \end{cases}$$

where

$$\alpha_i(\mathbf{x}, \mathbf{y}, \theta) = \min\left\{1, \frac{h(\mathbf{x}^*, \mathbf{y}^*, \theta)}{h(\mathbf{x}, \mathbf{y}, \theta)}\right\}$$

and $\mathbf{x}^*$ *and* $\mathbf{y}^*$ *are* $\mathbf{x}$ *and* $\mathbf{y}$ *with* x_i *and* y_i *swapped.*

The distribution $h(\mathbf{x}, \mathbf{y}, \theta)$ is stationary with respect to this process. To see this, note first that the transition in Step 1 leaves the marginal distribution of $\mathbf{x}$ unchanged, and trivially preserves the conditional distribution $h(\mathbf{y}, \theta|\mathbf{x}^{(t-1)})$. Moreover, h satisfies the detailed balance equations with respect to the transitions in Step 2, that is

$$h(\mathbf{x}, \mathbf{y}, \theta)\alpha_i(\mathbf{x}, \mathbf{y}, \theta) = h(\mathbf{x}^*, \mathbf{y}^*, \theta)\alpha_i(\mathbf{x}^*, \mathbf{y}^*, \theta)$$

and hence h is stationary with respect to these transitions also.

Note also that since $\mathbf{x}^*$ and $\mathbf{y}^*$ differ from $\mathbf{x}$ and $\mathbf{y}$ in only one coordinate, $\alpha_i(\mathbf{x}, \mathbf{y}, \theta)$ simplifies to:

$$\alpha_i(\mathbf{x}, y_i, \theta) = \min\left\{1, \frac{f(y_i)\,p(\theta|\mathbf{x}^*)\tau_\theta(x_i|y_i)}{f(x_i)\,p(\theta|\mathbf{x})\tau_\theta(y_i|x_i)}\right\}. \tag{4}$$

2.2 BAIS as an Instance of BAMS

BAIS is a straightforward instance of the more general BAMS construction. As noted above, the parameter vector in BAIS consists of the mean vector and covariance matrix, that is $\theta = (\mu, \Sigma)$. The Bayesian model for estimating the performance vector is the multivariate normal model discussed in Sect. 1.2, and results in the posterior distribution $p(\theta|\mathbf{x})$ given by Eq. 1. The proposal map is given by:

$$\tau_\theta(y|x) = \text{Norm}(y|\mu, \Sigma).$$

Note that the proposal is independent of x, as it should be for an independence sampler.

The values y' drawn for each i in Step 2 of Algorithm BAIS can instead be drawn at Step 1 to give a vector $\mathbf{y}'$, as in Algorithm BAMS. The acceptance ratio in Step 2 of Algorithm BAIS is just Eq. 4 with the multivariate normal proposal map defined above. It remains only to note that accepting y' in Step 2 of Algorithm BAIS is equivalent to swapping $x_i^{(t-1)}$ with y_i' in Step 2 of Algorithm BAMS, whereas rejecting y' is equivalent to not swapping.

2.3 A Random Walk Instance of BAMS

It is now relatively easy to construct a random walk sampler with an adaptive proposal distribution within the BAMS framework. Consider a target distribution f defined on $\Re^d$. Let the random walk proposal distribution be the multivariate normal distribution $\text{Norm}(y|x, \Sigma')$ centred on the current sample $x \in \Re^d$ and

with covariance matrix Σ'. The covariance matrix Σ' can be adapted to improve sampler performance. Based on the investigations of Gelman et al. [3], we set $\Sigma' = (2.4/\sqrt{d})^2\Sigma$ where Σ is a covariance matrix for a multivariate normal model of the target distribution, to be estimated based on a sample. Thus $\theta = \Sigma$ is the performance vector, and the proposal map is defined by:

$$\tau_\theta(y|x) = \text{Norm}(y|x, (2.4/\sqrt{d})^2\Sigma).$$

The posterior distribution for Σ is obtained by integrating Eq. 1 over μ to obtain:

$$p(\Sigma|\mathbf{x}) = \text{InvW}_{N-1}(\Sigma|S)$$

where S is given by Equations 2 and 3. A Bayesian adaptive random walk sampler can thus be constructed as follows:

Algorithm 4 (BARS). *Given initial values* $(x_1^{(0)}, \ldots, x_N^{(0)})$ *and* $\Sigma^{(0)}$, *set* $t := 1$ *and iterate the following:*

1. *Draw* $\Sigma^{(t)}$ *from* $p(\Sigma|\mathbf{x}^{(t-1)})$ *and set* $\Sigma' = (2.4/\sqrt{d})^2\Sigma^{(t)}$.
2. *For each* $i = 1, \ldots, N$, *draw* y' *from* $\text{Norm}(y|x_i^{(t-1)}, \Sigma')$ *and draw* u *uniformly and randomly from the interval* $(0, 1)$. *Set* $\mathbf{x}' = (x_1^{(t)}, \ldots, x_{i-1}^{(t)}, x_i^{(t-1)}, \ldots, x_N^{(t-1)})$ *and set:*

$$x_i^{(t)} = \begin{cases} y' & \text{if } u < \alpha_i(\mathbf{x}', y', \Sigma^{(t)}) \\ x_i^{(t-1)} & \text{otherwise} \end{cases}$$

where

$$\alpha_i(\mathbf{x}, y, \Sigma) = \min\left\{1, \frac{f(y)p(\Sigma|\mathbf{x}^*)}{f(x_i)p(\Sigma|\mathbf{x})}\right\}$$

and $\mathbf{x}^*$ *is* $\mathbf{x}$ *with* x_i *replaced by* y.

Note that the proposal map cancels from the acceptance ratio, because in this case $\tau_\theta(y|x)$ is symmetric in x and y.

3 A Worked Example: Sampling from a Multimodal Distribution

To illustrate implementation of the BARS algorithm we include a simple example in which we sample from a distribution with multiple modes (for examples implementing BAIS see [8]).

The target density is given as follows:

$$f(x) \propto e^{-x^2/18}\sin^2(x). \tag{5}$$

This density is somewhat challenging for an adaptive random walk sampler, which has to avoid the pitfall of tuning to the scale of an individual mode. Ideally the algorithm should tune to the scale of the envelope function $e^{-x^2/18}$, which resembles a normal density with variance 9.

We assume a univariate normal proposal distribution, with variance $(2.4)^2\theta$, and we centre the distribution of the proposed value for each sampling chain, y_i, on the current value of that chain x_i, giving us the following form:

$$p(y_i \mid x_i, \theta) = \frac{1}{2.4\sqrt{2\pi\theta}} \exp\left[-\frac{1}{2}\frac{(y_i - x_i)^2}{(2.4)^2\theta}\right]$$

The factor of 2.4 is included to optimise the jumping kernel, as suggested by [3].

The parameter to be adapted in this example is the variance θ of a normal model fitted to the current vector of states of the chains, $\mathbf{x} = (x_1, \ldots, x_N)$. This is updated by sampling from an Inverse-Gamma distribution:

$$\theta \mid \mathbf{x} \sim \text{Inv}-\Gamma(\epsilon, \omega)$$

where the shape, ϵ, and scale, ω, are given by:

$$\epsilon = \alpha_0 + \frac{N}{2}$$

$$\omega = \frac{1}{2}\left[C - A\left(\frac{B}{2A}\right)^2\right]$$

Here

$$A = \frac{1}{\gamma} + N$$

$$B = -2\left(\frac{\delta}{\gamma} + \sum_{i=1}^{N} x_i\right)$$

$$C = \frac{\delta^2}{\gamma} + 2\beta_0 + \sum_{i=1}^{N} x_i^2$$

for N sampling chains and for prior shape α_0 and prior scale β_0 of the (gamma) distribution of θ. Here, δ is the mean and $\gamma\theta$ is the variance of the prior distribution for the fitted mean. In our simulation α_0, β_0 and γ were all taken to be 1 and δ was taken to be 0.

Using the BARS algorithm, 100 sampling chains were updated, followed by the parameter chain for θ. This was repeated for 10,000 iterations, with the first 5,000 discarded. Figure 1 shows a density plot for the remaining 5,000 samples from each of the sampling chains, fitted using the statistical computing package, R [10]. As

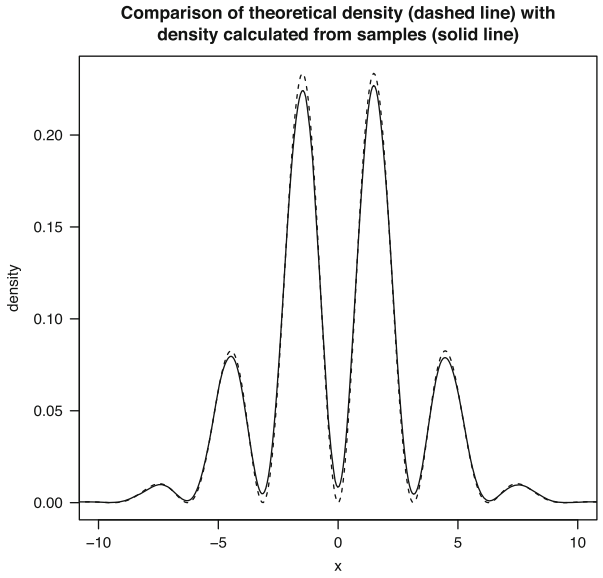


Fig. 1 Density plot of the samples produced using BARS for the target distribution given in Eq. (5). The *solid curve* is the estimated density plot of the samples and the *dashed curve* is the normalised theoretical density plot of the target distribution.

a density was calculated from the samples the theoretical curve was normalised (divided by the normalising constant 3.75994). This allows a more appropriate comparison between theory and observation. Convergence of the sampling chains was confirmed by a graphical check of the time series plots.

Figure 1 was plotted using 10,000 points per curve, with the same x values for both the density plot of the samples and the theoretical curve. The sum of squared differences between 10,000 points on the theoretical curve and the corresponding ones on the density plot of the samples was calculated to be 0.07568542. This is less than a 0.00001 average difference per pair of points, indicating a good agreement between theory and samples. Most of the difference occurs at peaks and troughs in the function, with larger differences occurring for sharper peaks and troughs. This may be due to the density fitting algorithm used by R rather than any defect in the sample. It should also be noted that the normalising constant used to scale the theoretical curve was a numerical approximation, calculated using R's built-in function integration, with an absolute error of less than 5×10^{-6}.

The posterior mean estimate of θ was 8.955, based on samples from the last 5,000 iterations. This is close to the variance of 9 relevant to the envelope function mentioned above, indicating that the algorithm has tuned to the appropriate scale.

This simple example is intended only to clarify the proposed algorithm. It is not our intention in this paper to present a state-of-the-art algorithm for real-world applications; rather, our focus is on developing a broad framework for

Bayesian adaptive sampling, within which many practical samplers can in future be investigated.

4 Future Directions

We have presented a new algorithm for Bayesian adaptive Metropolis-Hastings sampling. The algorithm involves three almost arbitrary objects: the performance vector, a Bayesian model relating the performance vector to a sample, and a proposal map. Convergence of the Markov chain implied by Algorithm BAMS to the required limiting distribution is guaranteed for any choice of these three objects, provided the chain is irreducible and aperiodic. However, it is not guaranteed that an arbitrary choice of these objects will result in an efficient sampler, or even in a sampler that adapts to improve performance. Identifying appropriate choices for these three objects is thus a significant direction for future research. What we have achieved here is to provide a broad framework that guarantees convergence and within which many adaptive schemes can be investigated.

It may also be difficult to sample the performance vector, so further work needs to be done in this area to determine when it is possible. In some cases approximations may be required to the underlying posterior distributions of the performance vector. However, complicated supports may make this difficult and further research into finding suitable approximations is required. Further work is also required to determine appropriate proposal distributions for specific cases.

As with all Markov chain Monte Carlo algorithms, it will be important to study how the variables of the method affect performance. In the case of algorithms that fit within the proposed framework, key variables are: (a) the number of sampling chains, (b) the length of each chain (run time) (c) the dimension of the target space and (d) the rate at which the algorithm adapts.

So far we have described only two instances of BAMS: Algorithms BAIS and BARS for independence and random walk samplers in $\Re^d$ respectively. It should be noted that these two instances scarcely begin to exploit the potential of the BAMS framework. For one thing, they are defined only on $\Re^d$, whereas BAMS is defined for an arbitrary target space $\mathscr{X}$, which might be combinatorial in nature. More importantly, both BAIS and BARS involve a very simple performance vector, estimated using a multivariate normal model of the target distribution. Other, more general methods of approximating the target distribution could be used, and in fact approximating the target distribution is not an essential element of BAMS. We envisage that a great many alternative methods for quantifying sampler performance and relating it to tuneable parameters are possible within the BAMS framework.

References

1. Andrieu, C., Moulines, E.: On the ergodicity properties of some adaptive MCMC algorithms. Ann. Appl. Probab. **16**, 1462–1505 (2006)
2. Andrieu, C., Thoms, J.: A tutorial on adaptive MCMC. Stat. Comput. **18**, 343–373 (2008)
3. Gelman, A., Roberts, G.O., Gilks, W.R.: Efficient Metropolis jumping rules. Bayesian Statist. **5**, 599–607 (1996)
4. Haario, H., Laine, M., Mira, A., Saksman, E.: DRAM: efficient adaptive MCMC. Stat. Comput. **16**, 339–354 (2006)
5. Hastings, W.K.: Monte Carlo sampling methods using Markov chains and their applications. Biometrika, **57**, 97–109 (1970)
6. Higdon, D.M.: Auxiliary variable methods for Markov chain Monte Carlo with applications. J. Amer. Statist. Stat. Assoc. **93**, 585–595 (1998)
7. Keith, J.M., Kroese, D.P., Bryant, D.: A generalized Markov sampler. Methodol. Comput. Appl. Probab. **6**, 29–53 (2004)
8. Keith, J.M., Kroese, D.P., Sofronov, G.Y.: Adaptive independence samplers. Stat. Comput. **18** 409–420 (2008)
9. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H.: Equations of state calculations by fast computing machines. J. Chem. Phys. **21**, 1087–1092 (1953)
10. R Development Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. ISBN 3-900051-07-0. http://www.R-project.org (2008)
11. Roberts, G.O., Rosenthal, J.S.: Examples of adaptive MCMC. J. Comput. Graph. Statist. **18**, 349–367 (2009)

Deterministic Consistent Density Estimation for Light Transport Simulation

Alexander Keller and Nikolaus Binder

Abstract Quasi-Monte Carlo methods often are more efficient than Monte Carlo methods, mainly, because deterministic low discrepancy sequences are more uniformly distributed than independent random numbers ever can be. So far, tensor product quasi-Monte Carlo techniques have been the only deterministic approach to consistent density estimation. By avoiding the repeated computation of identical information, which is intrinsic to the tensor product approach, a more efficient quasi-Monte Carlo method is derived. Its analysis relies on the properties of $(0, 1)$-sequences, provides new insights, and generalizes previous approaches to light transport simulation.

1 Introduction

Image synthesis comprises of computing functionals of the solution of a Fredholm integral equation of the second kind, which models light transport. Except for academic problems, no analytic solutions are available for any meaningful practical application. Therefore and due to high dimensionality and discontinuities of the functions involved, numerical algorithms average the contribution of transport paths sampled from the space of all possible light transport paths.

Monte Carlo methods [30] generate such samples from random numbers, which on a computer are simulated by pseudo-random number generators. Obviously, independence and unpredictability can only be mimicked, because pseudo-random number generators in fact are implemented as deterministic algorithms.

Quasi-Monte Carlo methods [22] are the deterministic counterpart of Monte Carlo methods. They often improve convergence speed as compared to Monte Carlo

A. Keller (✉) · N. Binder
NVIDIA, Fasanenstr. 81, 10623 Berlin, Germany
e-mail: keller.alexander@gmail.com; nikolaus.binder@googlemail.com

J. Dick et al. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2012*,
Springer Proceedings in Mathematics & Statistics 65,
DOI 10.1007/978-3-642-41095-6_23, © Springer-Verlag Berlin Heidelberg 2013

methods, because omitting the simulation of independence and unpredictability allows for a much better sampling. In addition, quasi-Monte Carlo methods have the desirable properties of simple parallelization [6, 15] and strict reproducibility.

In image synthesis, the aforementioned functionals correspond to measurements of transported light and can be formulated as integrals of a measurement contribution function $f(\mathbf{x}, \mathbf{y})$, which determines the light transported along a light transport path that is determined by $\mathbf{x}$ and $\mathbf{y}$. As illustrated in Fig. 3, such paths may result from connecting a path segment from a sink determined by $\mathbf{x}$ to a path segment originating from a source as determined by $\mathbf{y}$.

Often it is efficient to connect multiple path segments from a source to a single path segment from the sink and vice versa. For this setting, a new quasi-Monte Carlo algorithm is derived in Sect. 2 and applied to image synthesis in Sect. 3. The new algorithm is especially useful in difficult settings of light transport (see Fig. 4) and improves the efficiency of a first approach to consistent deterministic light transport simulation [16] by overcoming the necessity of a tensor product approach in order to guarantee consistency.

2 Consistent Blockwise Quasi-Monte Carlo Methods

Formulated in an abstract way, a measurement (as described above) is an integral

$$\int_{[0,1)^{s_1}} \int_{[0,1)^{s_2}} f(\mathbf{x}, \mathbf{y}) \, d\mathbf{y} \, d\mathbf{x} = \lim_{n \to \infty} \frac{1}{n} \sum_{i=0}^{n-1} \frac{1}{b^m} \sum_{k=0}^{b^m-1} f(\mathbf{x}_i, \mathbf{y}_{b^m \lfloor i/b^m \rfloor + k}), \quad (1)$$

where $\mathbf{x}$ represents the s_1 dimensions of a sink path, $\mathbf{y}$ specifies the trajectory of a source path in s_2 dimensions, and $f(\mathbf{x}, \mathbf{y})$ is the measurement contribution resulting from the connection of both path segments. Instead of considering the interaction of each source path with each sink path similar to the tensor product enumeration scheme in [16], the case $b^m < \infty$ results in a simpler and more efficient algorithm: As illustrated in Fig. 2, the algorithm averages the measurement contributions of contiguous blocks of b^m vectors $(\mathbf{x}_i, \mathbf{y}_i)$ from one low discrepancy sequence, where within each block with integer index $\lfloor i/b^m \rfloor$ each $\mathbf{x}_i$ is combined with all $\mathbf{y}_{b^m \lfloor i/b^m \rfloor + k}$ for $0 \leq k < b^m$.

The actual computation progresses in contiguous blocks, which allows one to increase efficiency by storing intermediate results of the path segments that are accessed b^m times in $\mathcal{O}(b^m)$ memory. The computation can be stopped and continued by only saving the current set of indices. For example, termination can be triggered by the user, by fixing a sample or time budget, or thresholding differences during the temporal progression of the algorithm [27].

Using the fact that quasi-Monte Carlo integration converges for bounded, square integrable functions [23], the consistency of Eq. 1 will be shown for quasi-Monte Carlo points that are based on radical inversion as reviewed in Sect. 2.1. The proof

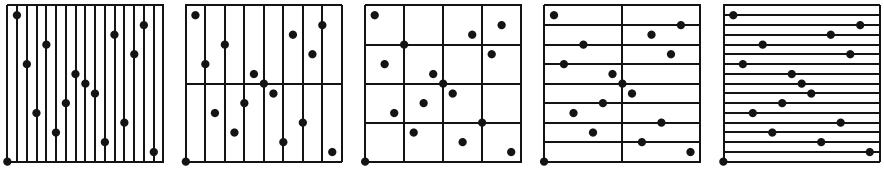


Fig. 1 All kinds of elementary intervals with area $\frac{1}{16}$ for $s = b = 2$. The first $2^4 = 16$ points of Sobol's $(0, 2)$-sequence, which form a $(0, 4, 2)$-net in base $b = 2$, are superimposed over each set of elementary intervals.

in Sect. 2.3 is based on the properties of such point sequences and a technique called scrambling (as reviewed in Sect. 2.2). The results can be generalized to include quasi-Monte Carlo integro-approximation [13] by adding a free variable $\mathbf{z}$ and considering $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$.

2.1 Low Discrepancy Sequences

Instead of sampling by using uniform (pseudo-) random numbers, quasi-Monte Carlo methods apply deterministic low discrepancy sequences, which are more uniformly distributed than random numbers can be. For a profound introduction including multiple constructions of low discrepancy sequences, we refer to the standard reference [22].

In order to prove Eq. 1, we focus on a framework that has been derived by Niederreiter [21] building on previous work on special cases by Sobol' [29] and Faure [4]. It is based on stratification properties as imposed by

Definition 1. For a fixed dimension $s \geq 1$ and an integer base $b \geq 2$ the subinterval

$$E(p_1, \ldots, p_s) := \prod_{j=1}^{s} \left[p_j \cdot b^{-d_j}, (p_j + 1) \cdot b^{-d_j} \right) \subseteq [0, 1)^s$$

with $0 \leq p_j < b^{d_j}$, $p_j, d_j \in \mathbb{N}_0$, is called an *elementary interval* (see [22, p. 48]). For integers $0 \leq t \leq m$, a (t, m, s)*-net in base* b (see [22, Definition 4.1]) is a point set of b^m points in $[0, 1)^s$ such that there are exactly b^t points in each elementary interval $E(p_1, \ldots, p_s)$ with volume b^{t-m}. For an integer $t \geq 0$, a sequence $\mathbf{x}_0, \mathbf{x}_1, \ldots$ of points in $[0, 1)^s$ is a (t, s)*-sequence in base* b (see [22, Definition 4.2]) if, for all integers $k \geq 0$ and $m > t$, the point set $\mathbf{x}_{kb^m}, \ldots, \mathbf{x}_{(k+1)b^m-1}$ is a (t, m, s)-net in base b.

One construction to enumerate the components of (t, s)-sequences in base b is to first represent the integer index

$$i =: \sum_{l=0}^{M-1} a_l(i) \cdot b^l$$

by its digits $a_l(i) \in \mathbb{Z}_b := \{0, \ldots, b-1\}$ in base b. Then the j-th component of the i-th point U_i is computed as

$$U_i^{(j)} = \sum_{k=1}^{M} u_{i,k}^{(j)} \cdot b^{-k} =_b 0.u_{i,1}^{(j)} u_{i,2}^{(j)} \ldots u_{i,M}^{(j)} \in [0,1) \cap \mathbb{Q}, \text{ where}$$

$$u_{i,k}^{(j)} := \eta_k^{(j)} \left(\sum_{l=0}^{M-1} c_{k,l}^{(j)} \cdot \psi_l\left(a_l(i)\right) \right) \tag{2}$$

and $\eta_k^{(j)} : R \to \mathbb{Z}_b$ and $\psi_l : \mathbb{Z}_b \to R$ are two families of bijections from and to a commutative ring $(R, +, \cdot)$ with $|R| = b$ elements. While in theory the generator matrices

$$C^{(j)} := \left(c_{k,l}^{(j)} \right)_{k=1,l=0}^{M,M-1} \in R^{M \times M}$$

could be infinite-dimensional, in practice they are finite due to the finite precision of computer arithmetic. Obviously, M digits allow for generating up to b^M points.

If the constructed point set $U = \{U_0, \ldots, U_{n-1}\}$ is a (t, m, s)-net in base b, then it is also called a *digital* (t, m, s)-net *constructed over* R. *Digital* (t, s)*-sequences* are defined analogously. The quality of a digital sequence is mainly determined by the choice of the ring R and the generator matrices $C^{(j)}$. Polynomial rings $R[X]$ over R are frequently used for the construction of the generator matrices $C^{(j)}$ (see [2, 4, 22, 28, 29, 31] and others).

The most popular implementation of digital sequences is the Sobol' sequence, which is a (t, s)-sequence in base $b = 2$. The perfect match of bit vector and computer arithmetic allows for the rapid generation of the Sobol' points at a speed comparable to high quality pseudo-random number generators. For example code and generator matrices see http://web.maths.unsw.edu.au/~fkuo/sobol/.

An important property of Sobol's constructions is that each component is a $(0, 1)$-sequence. Thus the matrices $C^{(j)}$ must be regular (in fact they are upper triangular matrices) and Eq. 2 is a bijection. The structure imposed by the elementary intervals is illustrated in Fig. 1 for the example of a $(0, m, 2)$-net resulting from the first two dimensions of the Sobol' sequence.

2.2 Scrambling

In order to scramble a point set on $H = [0, 1)^s$, the following steps are applied to each coordinate:

1. Slice H into b equal volumes $H_1, H_2, \ldots, H_b$ along the coordinate.
2. Permute these volumes.
3. For each volume H_h recursively repeat the procedure starting out with $H = H_h$.

While scrambling [24, 25] has been introduced to randomize point sets in order to allow for variance estimation, in fact, several classic improvements to low discrepancy sequence are deterministic scramblings [4, 13]. As the partitioning happens along elementary intervals (see Definition 1), the t parameter of (t, s)-sequences in base b is invariant with respect to scrambling.

Again, due to the finite precision of computer arithmetic, the scheme becomes a finite algorithm that can be formalized as follows: Given the j-th component $U_i^{(j)} =_b 0.u_{i,1}^{(j)} u_{i,2}^{(j)} u_{i,3}^{(j)} \cdots u_{i,M}^{(j)}$ of the i-th point of a point sequence, its scrambled version $V_i^{(j)} =_b 0.v_{i,1}^{(j)} v_{i,2}^{(j)} v_{i,3}^{(j)} \cdots v_{i,M}^{(j)}$ is determined by applying permutations to the digits

$$\begin{aligned} v_{i,1}^{(j)} &:= \pi^{(j)}\left(u_{i,1}^{(j)}\right) \\ v_{i,2}^{(j)} &:= \pi^{(j)}_{u_{i,1}^{(j)}}\left(u_{i,2}^{(j)}\right) \\ &\vdots \\ v_{i,M}^{(j)} &:= \pi^{(j)}_{u_{i,1}^{(j)}, u_{i,2}^{(j)}, \ldots, u_{i,M-1}^{(j)}}\left(u_{i,M}^{(j)}\right), \end{aligned} \tag{3}$$

where the k-th permutation from the symmetric group S_b of all permutations over the set $\{0, \ldots, b-1\}$ depends on the $k-1$ leading digits of $U_i^{(j)}$. The mapping is bijective on $[0, 1) \cap \mathbb{Q}$, because the inverse $u_{i,k}^{(j)}$ of any $v_{i,k}^{(j)}$ is found by recursively computing $u_{i,k}^{(j)} = \left(\pi^{(j)}_{u_{i,1}^{(j)}, \ldots, u_{i,k-1}^{(j)}}\right)^{-1}\left(v_{i,k}^{(j)}\right)$.

2.3 Replication by Partial Scrambling

With the properties of (t, s)-sequences and scrambling as reviewed in the two previous sections, we are ready to prove Eq. 1 formulated as

Theorem 1. *Given a deterministic digital (t, s)-sequence $(\mathbf{x}_i, \mathbf{y}_i)$ in base b, whose components in $\mathbf{y}_i$ are generated by regular upper triangular matrices,*

$$\int_{[0,1)^{s_1}} \int_{[0,1)^{s_2}} f(\mathbf{x}, \mathbf{y}) d\mathbf{y} d\mathbf{x} = \lim_{n \to \infty} \frac{1}{n} \sum_{i=0}^{n-1} \frac{1}{b^m} \sum_{k=0}^{b^m - 1} f(\mathbf{x}_i, \mathbf{y}_{b^m \lfloor i/b^m \rfloor + k})$$

is consistent.

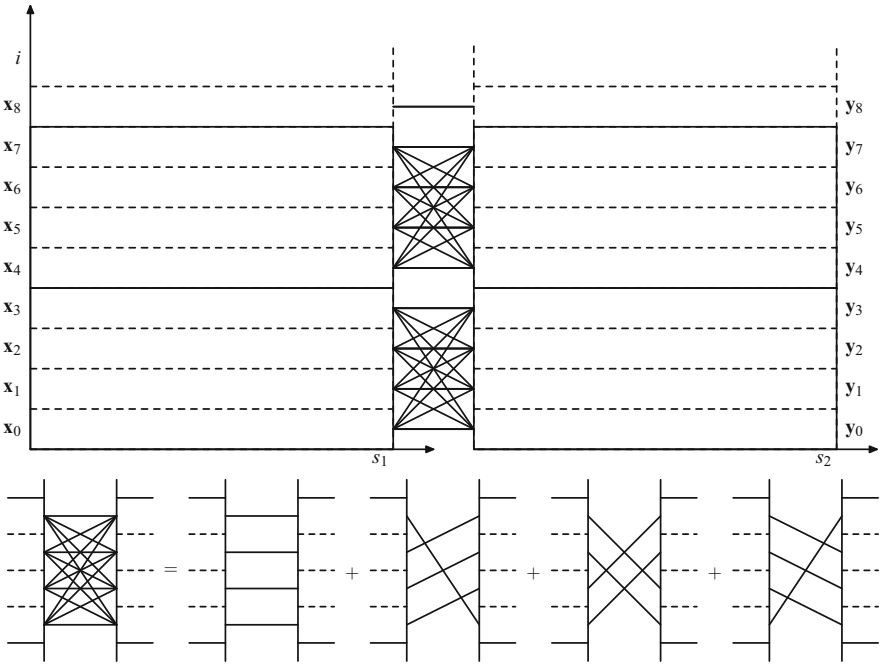


Fig. 2 *Top*: One $s_1 + s_2$-dimensional low discrepancy sequence is partitioned into contiguous blocks of b^m vectors. Simultaneously, all vectors are identically partitioned into s_1-dimensional $\mathbf{x}_i$- and s_2-dimensional $\mathbf{y}_i$-components. *Bottom*: Combining all vectors $\mathbf{x}_i$ of a block with all vectors $\mathbf{y}_i$ of the same block can be regarded as the sum of all tuples $(\mathbf{x}_i, \mathbf{y}_i)$, $(\mathbf{x}_i, \mathbf{y}_{b^m\lfloor i/b^m\rfloor+((i+1) \bmod b^m)})$, $(\mathbf{x}_i, \mathbf{y}_{b^m\lfloor i/b^m\rfloor+((i+2) \bmod b^m)})$, and so on, which is key to proving Eq. 1 as stated in Theorem 1.

Proof. Swapping the two sums in the algorithm yields

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \frac{1}{b^m} \sum_{k=0}^{b^m-1} f(\mathbf{x}_i, \mathbf{y}_{b^m\lfloor i/b^m\rfloor+k})$$

$$= \frac{1}{b^m} \sum_{k=0}^{b^m-1} \underbrace{\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(\mathbf{x}_i, \mathbf{y}_{b^m\lfloor i/b^m\rfloor+((i+k) \bmod b^m)})}_{(*)}$$

as illustrated at the bottom of Fig. 2.

For fixed $k \in \{0, \ldots, b^m - 1\}$, looking at the term (*), the j-th component

$$y^{(j)}_{b^m\lfloor i/b^m\rfloor+((i+k) \bmod b^m)}$$

$$= \left(b^{-1} \cdots b^{-M}\right) C^{(j)} \begin{pmatrix} a_0((i+k) \bmod b^m) \\ \vdots \\ a_{m-1}((i+k) \bmod b^m) \\ a_m(i) \\ \vdots \end{pmatrix}$$

$$= \left(b^{-1} \cdots b^{-M}\right) C^{(j)} \begin{pmatrix} \pi_k(a_0(i)) \\ \vdots \\ \pi_{k,a_0(i),\ldots,a_{m-2}(i)}(a_{m-1}(i)) \\ a_m(i) \\ \vdots \end{pmatrix}$$

of $\mathbf{y}_{b^m\lfloor i/b^m \rfloor + ((i+k) \bmod b^m)}$ is rewritten by inserting the index $b^m\lfloor i/b^m \rfloor + ((i + k) \bmod b^m)$ into the definition of the radical inverse as introduced in Eq. 2. The first step of the transformation reveals that only the m least significant digits of i are changed by the enumeration scheme. Consequently, the permutation $(i+k) \bmod b^m$ can only affect the m most significant digits of the j-th component, because by assumption the generator matrix $C^{(j)}$ is a regular, upper triangular matrix. The second step rewrites the extraction of a digit from the permutation $(i + k) \bmod b^m$ as a permutation applied to a digit of i, where the permutation depends on k and the previous least significant digits.

Obviously, changing the enumeration order of the partial sequence $\mathbf{y}_i$ results in the same points. However, the above derivation shows that this reordering can be understood as a scrambling, which is permuting elementary intervals. Hence, the sequence $(\mathbf{x}_i, \mathbf{y}_{b^m\lfloor i/b^m \rfloor + ((i+k) \bmod b^m)})$ remains a (t, s)-sequence with identical t parameter that results from applying a partial deterministic scrambling to the original (t, s)-sequence $(\mathbf{x}_i, \mathbf{y}_i)$ as explained in Sect. 2.2.

In combination with the assumption of a square integrable, bounded integrand, [23, Theorem 2] then guarantees convergence to the desired integral. Averaging the results of term (*) for all k yields the consistency of Eq. 1, which concludes the proof. □

As a corollary, the result extends to quasi-Mont Carlo integro-approximation by using the generalization of [23, Theorem 2] as derived in [13, Theorem 1].

In [13, 23] it has been shown that quasi-Monte Carlo integration and integro-approximation converge for bounded, square integrable functions, which is the setting in computer graphics. However, the error bounds do not provide a separation into a property of the function f and a rate of convergence depending on the number n of samples. In order to argue for the superiority of quasi-Monte Carlo methods over Monte Carlo methods, rates provided by theorems on discrete density approximation [10] can be consulted. Whenever the functions are more specific functions classes, like for example in the class of bounded variation in the sense of Hardy and Krause, the discrepancy of the point set provides a convergence

rate, like for example the Koksma-Hlawka inequality [22]. In conclusion, one can always argue for a superiority of deterministic low discrepancy sampling over (pseudo-) random sampling. While in some settings the advantage may be small, the deterministic method is simpler to parallelize and reproduce [6, 15].

2.3.1 Rank-1 Lattice Sequences

Rank-1 lattice sequences [9] are point sequences resulting from multiplying a generator vector $\mathbf{g}$ by a $(0, 1)$-sequence in base b on the unit torus. If now the base b is relatively prime to the components of the generator vector $\mathbf{g}$, each component of the rank-1 lattice sequence is a $(0, 1)$-sequence.

While scrambling (see Sect. 2.2) can destroy the lattice structure, it cannot change the order of uniformity: As only elementary intervals are permuted, scrambled low discrepancy rank-1 lattice sequences remain of low discrepancy. Therefore, Eq. 1 works with rank-1 lattice sequences for any block size b^m in analogy to Theorem 1.

2.3.2 Halton Sequence

Theorem 1 requires the components of $\mathbf{y}_i$ to be $(0, 1)$-sequences. While this is true for the Halton type sequences [8], the bases for each dimension are relatively prime, which requires the block size to be a product of powers of the bases in order to make Theorem 1 work. Then the approach may be considered impractical, because the block size would exponentially grow with dimension.

2.3.3 General Block Size and (0, 1)-Sequences

The theoretical restrictions of the previous section could be overcome if the algorithm in Theorem 1 worked for any block size $q \in \mathbb{N}$ instead of only b^m. In fact this conjecture is backed by numerical experiments with the specific algorithm described in Sect. 3.2 using the Halton sequence, whose convergence visually cannot be distinguished from sampling with the Sobol' sequence or rank-1 lattice sequences.

Although $y^{(j)}_{q\lfloor i/q\rfloor+((i+k) \bmod q)}$ and $y^{(j)}_i$ generate the identical set of numbers, the properties of the sequences may be different: For example for $q = 3$ and $k = 1$, $\Phi_2(3\lfloor i/3\rfloor + ((i + 1) \bmod 3))$ is not a $(0, 1)$-sequence, although $\Phi_2(i)$ is, which is easily verified by looking at the first four elements of either sequence. Therefore, enumerating a (t, s)-sequence in base b using the permuted enumeration index $q\lfloor i/q\rfloor + ((i + k) \bmod q)$ instead of i may increase the t parameter (see Definition 1) unless $q = b^m$, because the permutation does not match the structure of elementary intervals of (t, s)-sequences and (t, m, s)-nets in a selected base b.

Then, the sequence $(\mathbf{x}_i, \mathbf{y}_{q\lfloor i/q \rfloor + ((i+k) \bmod q)})$ cannot be written as a scrambling in the sense of [24,25] (see Sect. 2.2) and the assumptions of Theorem 1 are not fulfilled.

Still, since $q\lfloor i/q \rfloor + ((i + k) \bmod q)$ is a bijection on $\mathbb{N}_0$, there exists an automorphism on the s-dimensional unit cube that undoes the permutation. Consequently, for square integrable, bounded integrands the term (*) converges to the desired integral by Theorem 2 in [23]. The same argument allows for using any low discrepancy sequence, whose components $y^{(j)}_{q\lfloor i/q \rfloor + ((i+k) \bmod q)}$ are bijections, which especially includes (0, 1)-sequences generated by non-upper triangular matrices.

3 Application to Light Transport Simulation

In practice, the algorithm in Eq. 1 is evaluated in a progressive way: One low discrepancy sequence $(\mathbf{x}_i, \mathbf{y}_i)$ is enumerated in contiguous blocks of b^m elements. Then, the function f is evaluated for all possible pairs $(\mathbf{x}_i, \mathbf{y}_{b^m \lfloor i/b^m \rfloor + ((i+k) \bmod b^m)})$ of arguments within each block.

In light transport simulation (see Fig. 3), each tuple $(\mathbf{x}_i, \mathbf{y}_i)$ determines two end points $h(\mathbf{x}_i)$ and $h(\mathbf{y}_i)$ of path segments in space. While $h(\mathbf{y}_i)$ results from tracing the trajectory of a photon started on a light source, the $h(\mathbf{x}_i)$ results from following optical paths starting from the camera. In analogy to random walk simulation, decisions of how to scatter and/or terminate paths are controlled by the components of the low discrepancy points $(\mathbf{x}_i, \mathbf{y}_i)$ instead of using random numbers.

In order to determine the measurement contribution f, the path segments generated this way need to be connected by either testing their mutual visibility (see Sect. 3.1) or connecting end points of path segments that are sufficiently close (see Sect. 3.2). As discussed in Sect. 3.3, the efficiency of the methods depends on the parameter m, which controls the block size.

3.1 Consistent Bidirectional Path Tracing

The Monte Carlo method realized in [26] first generates a set of paths from the camera into the scene, then traces paths of photons from the light sources yielding a set of virtual point lights, and finally checks the visibility of each point light source from each end point of the camera paths in order to compute the radiance passing through the screen's pixels. Convergence is achieved by iterating the procedure.

Equation 1 is the deterministic and consistent quasi-Monte Carlo analog of that procedure. For $m = 0$, the methods coincides with bidirectional path tracing [20, 32,33], where each camera path segment is connected with its photon path segment. While for $m = 0$ the dominating transient artifact is noise, for $m > 0$ larger block sizes trade noise for coherence artifacts like for example discrete shadow boundaries [12].

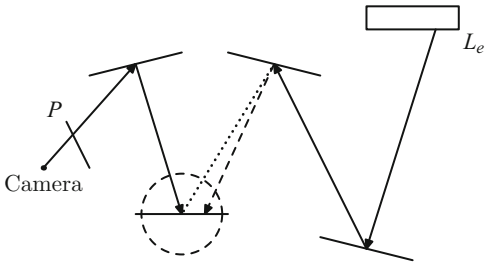


Fig. 3 Bidirectional generation of light transport paths: A path segment started from the camera and a path segment started from a light source L_e can be connected by a shadow ray (*dotted line*), which checks whether the vertices to connect are mutually visible. Alternatively, the basic idea of photon mapping is to relax the precise visibility check by allowing for a connection of both path segments if their end points are sufficiently close as indicated by the *dashed circle*. Both techniques are illustrated for identical path length, which is the reason for the dashed prolongation of the light path segment for photon mapping.

3.2 Consistent Photon Mapping

Connecting light path segments by checking the mutual visibility of their end points is not efficient in many common situations. In such situations (see [16] for an extensive background), connecting path segments if their end points are sufficiently close, i.e. if their difference vector lies within a ball $\mathscr{B}$ of radius $r(n)$ centered at the origin, can help overcome the problem of "insufficient techniques" [19, Fig. 2].

Using $\chi_{\mathscr{B}(r(n))}$ as the characteristic function of that ball, the radiance

$$L_P = \lim_{n\to\infty} \frac{|P|}{n} \sum_{i=0}^{n-1} \chi_P(\mathbf{x}_i) W(\mathbf{x}_i) \frac{1}{b^m} \sum_{k=0}^{b^m-1} \frac{\chi_{\mathscr{B}(r(n))}(h(\mathbf{x}_i) - h(\mathbf{y}_{b^m\lfloor i/b^m\rfloor+k}))}{\pi r^2(n)} \cdot f_s(\omega(\mathbf{x}_i), h(\mathbf{x}_i), \omega(\mathbf{y}_{b^m\lfloor i/b^m\rfloor+k}))\phi(\mathbf{y}_{b^m\lfloor i/b^m\rfloor+k}) \tag{4}$$

is determined by averaging the contributions to the query location $h(\mathbf{x}_i)$ attenuated by the throughput $W(\mathbf{x}_i)$ of the path, where χ_P selects the paths contributing to the pixel P. These contributions are the product of the flux ϕ of a photon attenuated by the bidirectional scattering distribution function (BSDF) f_s averaged over the disk with the radius $r(n)$. The BSDF determines the fraction of light incident from direction $\omega(\mathbf{y}_{b^m\lfloor i/b^m\rfloor+k})$ in the surface location $h(\mathbf{x}_i)$ into direction $\omega(\mathbf{x}_i)$, where the directions are determined by the respective last edge of the path segments.

The characteristic function $\chi_{\mathscr{B}(r)}$ selects a subset of path space. Since light transport is a linear problem, the number of paths in that set asymptotically must be linear in n. If now the radius

$$r^2(n) = \frac{r_0^2}{n^\alpha} \tag{5}$$

Fig. 4 The images compare three approaches to sampling in light transport simulation computed at an identical number of samples. Obviously the image quality achieved by consistent blockwise quasi-Monte Carlo method is superior to both the tensor product approach and the Monte Carlo method. The quasi-Monte Carlo methods were using the Sobol' sequence, while the pseudo-random numbers have been generated using the Mersenne Twister.

for some fixed $r_0 \in \mathbb{R}^+$, consistency can be proved similar to [16, Sect. 3.1]: For $\alpha = 1$, doubling the number n of paths results in half the squared radius, meaning half the area, while the number of paths connected in $\mathscr{B}(r(n))$ must remain the same due to linearity. For $0 < \alpha < 1$ the squared radius decreases slower than the increase in number of connected paths. As a consequence, more and more paths become connected with increasing n, which guarantees convergence. Note that for $\alpha = 0$ the radius r becomes independent of n and the computation converges to an average over the disk of fixed radius r.

In the practical computation of Eq. 4, for each block its smallest radius is used, which is $r(b^m(\lfloor i/b^m \rfloor + 1) - 1)$. If the number of blocks is finite and known beforehand, the overall smallest radius can be used right from the beginning.

As shown in [16], the radius vanishes arbitrarily slowly and the influence of the parameter α becomes negligible already after enumerating a few blocks. Consequently, the efficiency is controlled by the initial radius r_0 and the parameter m determining the block size b^m. The initial radius r_0 determines the ratio of how many photons can interact with a query location. Besides choosing r_0 constant, adaptive methods have been discussed in [16, 18].

3.2.1 Generalizations and Extensions

Similar to [11, 18], it is straightforward to extend the algorithm in Eq. 4 for light transport simulation in participating media: The BSDF is replaced by the product of phase function and scattering cross section, while for example Woodcock tracking is used to sample path segments in participating media.

With respect to efficiency, Eq. 4 is easily extended by a sum to account for multiple photons and/or query locations stored along path segments. Then, bounding memory in a conservative way results in a memory footprint proportional to the block size multiplied by the maximum path length. Therefore, it may be more efficient to fix an m, which is decreased during the course of computation whenever

the allocated memory block cannot store all path segments. It is obvious that in the limit this adaptive procedure converges as well. Optionally, Russian roulette can be used to adaptively control the average path length per path.

3.3 Selecting the Block Size

For consistent photon mapping (see Sect. 3.2), a larger block size increases the density of query locations and photons, which in turn results in more paths being connected. Since the memory footprint is proportional to the block size, m should be determined by the maximum amount of memory dedicated to storing query locations and photons.

With respect to path tracing (see Sect. 3.1), m blends between different algorithmic concepts. As proposed in [6], the $\mathbf{x}_i$ can be used to generate paths uniformly sampling the screen. If now b^m is about the number of pixels, each path through a pixel is connected to b^m paths from light sources. This approach very much resembles the method of dependent tests used in [12], however, now the scheme is deterministic and consistent. The application of smaller block sizes results in consistent version of interleaved sampling [17], where neighboring pixels do not share light path samples.

Although bidirectional path tracing and photon mapping, i.e. integration and density estimation, impose different constraints on m, both approaches can be efficiently combined: The b^m samples in a block used for density estimation can be partitioned into smaller blocks of the size of a power of b for integration purposes.

By using multiple importance sampling [1, 3, 20, 33], it is possible to further increase efficiency by weighting the contributions of both bidirectional path tracing and photon mapping [5, 7]. With respect to consistent quasi-Monte Carlo methods, this is subject to future investigation.

4 Conclusion

Compared to the tensor product quasi-Monte Carlo method introduced in [16], the new scheme is consistent without having to compute identical paths over and over again and thus can sample path space more densely in the same amount of time. For the case of photon mapping (see Sect. 3.2), the resulting improvement is shown in Fig. 4.

Other than the initial approach [16], the new algorithm does not allow for controlling the ratio of camera and photon paths. This, however, can be easily complemented using the orthogonal technique of enumerating low discrepancy sequences, whose components are $(0, 1)$-sequences, at different speeds as derived in [14, Sect. 3.2.4].

Acknowledgements The authors would like to thank Leonhard Grünschloß for the profound discussions.

References

1. van Antwerpen, D.: Unbiased physically based rendering on the GPU. Master's thesis, Computer Graphics Research Group, Department of Software Technology Faculty EEMCS, Delft University of Technology (2011)
2. Bratley, P., Fox, B., Niederreiter, H.: Implementation and tests of low-discrepancy sequences. ACM Trans. Model. Comput. Simul. **2**, 195–213 (1992)
3. Dahm, K.: A comparison of light transport algorithms on the GPU. Master's thesis, Computer Graphics Group, Saarland University (2011)
4. Faure, H.: Discrépance de suites associées à un système de numération (en dimension *s*). Acta Arith. **41**, 337–351 (1982)
5. Georgiev, I., Křivánek, J., Davidovič, T., Slusallek, P.: Light transport simulation with vertex connection and merging. ACM Trans. Graph. **31**, 192:1–192:10 (2012)
6. Grünschloß, L., Raab, M., Keller, A.: Enumerating quasi-Monte Carlo point sequences in elementary intervals. In: Plaskota, L., Woźniakowski, H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2010, pp. 399–408. Springer, Berlin/Heidelberg (2012)
7. Hachisuka, T., Pantaleoni, J., Jensen, H.: A path space extension for robust light transport simulation. ACM Trans. Graph. **31**, 191:1–191:10 (2012)
8. Halton, J.: On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. Numer. Math. **2**, 84–90 (1960)
9. Hickernell, F., Hong, H., L'Ecuyer, P., Lemieux, C.: Extensible lattice sequences for quasi-Monte Carlo quadrature. SIAM J. Sci. Comput. **22**, 1117–1138 (2001)
10. Hlawka, E., Mück, R.: Über eine Transformation von gleichverteilten Folgen II. Computing **9**, 127–138 (1972)
11. Jarosz, W., Nowrouzezahrai, D., Thomas, R., Sloan, P., Zwicker, M.: Progressive photon beams. ACM Trans. Graph. – Proc. ACM SIGGRAPH Asia **30**, 181:1–181:11 (2011)
12. Keller, A.: Instant radiosity. In: SIGGRAPH '97: Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques, Los Angeles, pp. 49–56 (1997)
13. Keller, A.: Myths of computer graphics. In: Niederreiter, H. (ed.) Monte Carlo and Quasi-Monte Carlo Methods 2004, pp. 217–243. Springer, Berlin/Heidelberg (2006)
14. Keller, A.: Quasi-Monte Carlo image synthesis in a nutshell. In: Dick, J., Kuo, F.Y., Peters, G.W., Sloan, I.H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2012, this volume 213–249. Springer, Berlin/Heidelberg (2013)
15. Keller, A., Grünschloß, L.: Parallel quasi-Monte Carlo integration by partitioning low discrepancy sequences. In: Plaskota, L., Woźniakowski, H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2010, pp. 487–498. Springer, Berlin/Heidelberg (2012)
16. Keller, A., Grünschloß, L., Droske, M.: Quasi-Monte Carlo progressive photon mapping. In: Plaskota, L., Woźniakowsi, H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2010, pp. 499–509. Springer, Berlin/Heidelberg (2012)
17. Keller, A., Heidrich, W.: Interleaved sampling. In: Myszkowski, K., Gortler, S. (eds.) Rendering Techniques 2001. Proceedings of the 12th Eurographics Workshop on Rendering, London, pp. 269–276. Springer (2001)
18. Knaus, C., Zwicker, M.: Progressive photon mapping: A probabilistic approach. ACM Trans. Graph. **30**, 25:1–25:13 (2011)
19. Kollig, T., Keller, A.: Efficient bidirectional path tracing by randomized quasi-Monte Carlo integration. In: Niederreiter, H., Fang, K., Hickernell, F. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2000, pp. 290–305. Springer, Berlin/New York (2002)

20. Lafortune, E.: Mathematical models and Monte Carlo algorithms for physically based rendering. Ph.D. thesis, KU Leuven (1996)
21. Niederreiter, H.: Point sets and sequences with small discrepancy. Monatsh. Math. **104**, 273–337 (1987)
22. Niederreiter, H.: Random Number Generation and Quasi-Monte Carlo Methods. SIAM, Philadelphia (1992)
23. Niederreiter, H.: Error bounds for quasi-Monte Carlo integration with uniform point sets. J. Comput. Appl. Math. **150**, 283–292 (2003)
24. Owen, A.: Randomly permuted (t, m, s)-nets and (t, s)-sequences. In: Niederreiter, H., Shiue, P.J.-S. (eds.) Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing. Lecture Notes in Statistics, vol. 106, pp. 299–315. Springer, New York (1995)
25. Owen, A.: Monte Carlo variance of scrambled net quadrature. SIAM J. Numer. Anal. **34**, 1884–1910 (1997)
26. Pajot, A., Barthe, L., Paulin, M., Poulin, P.: Combinatorial bidirectional path-tracing for efficient hybrid CPU/GPU rendering. Comput. Graph. Forum **30**, 315–324 (2011)
27. Paskov, S.: Termination criteria for linear problems. J. Complexity **11**, 105–137 (1995)
28. Schmid, W.: (t, m, s)-Nets: Digital construction and combinatorial aspects. Ph.D. thesis, Universität Salzburg (1995)
29. Sobol', I.: On the Distribution of points in a cube and the approximate evaluation of integrals. Zh. vychisl. Mat. mat. Fiz. **7**, 784–802 (1967)
30. Sobol', I.: Die Monte-Carlo-Methode. Deutscher Verlag der Wissenschaften, Berlin (1991)
31. Sobol', I., Asotsky, D., Kreinin, A., Kucherenko, S.: Construction and comparison of high-dimensional Sobol' generators. WILMOTT Mag. **56**, 64–79 (2011)
32. Veach, E.: Robust Monte Carlo methods for light transport simulation. Ph.D. thesis, Stanford University (1997)
33. Veach, E., Guibas, L.: Optimally combining sampling techniques for Monte Carlo rendering. In: SIGGRAPH '95 Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques, Los Angeles, pp. 419–428 (1995)

On Wavelet-Galerkin Methods for Semilinear Parabolic Equations with Additive Noise

Mihály Kovács, Stig Larsson, and Karsten Urban

Abstract We consider the semilinear stochastic heat equation perturbed by additive noise. After time-discretization by Euler's method the equation is split into a linear stochastic equation and a non-linear random evolution equation. The linear stochastic equation is discretized in space by a non-adaptive wavelet-Galerkin method. This equation is solved first and its solution is substituted into the nonlinear random evolution equation, which is solved by an adaptive wavelet method. We provide mean square estimates for the overall error.

1 Introduction

We consider the following semilinear parabolic problem with additive noise,

$$\begin{aligned} &\mathrm{d}u - \nabla \cdot (\kappa \nabla u)\,\mathrm{d}t = f(u)\,\mathrm{d}t + \mathrm{d}W, && x \in \mathscr{D},\ t \in (0,T), \\ &u = 0, && x \in \partial\mathscr{D},\ t \in (0,T), \\ &u(\cdot,0) = u_0, && x \in \mathscr{D}. \end{aligned} \tag{1}$$

M. Kovács (✉)
Department of Mathematics and Statistics, University of Otago, P.O. Box 56, Dunedin, New Zealand
e-mail: mkovacs@maths.otago.ac.nz

S. Larsson
Mathematical Sciences, Chalmers University of Technology, SE–412 96 Gothenburg, Sweden
e-mail: stig@chalmers.se

K. Urban
Institute for Numerical Mathematics, Ulm University, Helmholtzstr. 18, DE–89069 Ulm, Germany
e-mail: karsten.urban@uni-ulm.de

J. Dick et al. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2012*,
Springer Proceedings in Mathematics & Statistics 65,
DOI 10.1007/978-3-642-41095-6_24, © Springer-Verlag Berlin Heidelberg 2013

Here $T > 0$, $\mathscr{D} \subset \mathbb{R}^d$, $d = 1, 2, 3$, is a convex polygonal domain or a domain with smooth boundary $\partial\mathscr{D}$, and $\{W(t)\}_{t\geq 0}$ is an $L_2(\mathscr{D})$-valued Q-Wiener process on a filtered probability space $(\Omega, \mathscr{F}, \mathbb{P}, \{\mathscr{F}_t\}_{t\geq 0})$ with respect to the normal filtration $\{\mathscr{F}_t\}_{t\geq 0}$. We use the notation $H = L_2(\mathscr{D})$, $V = H_0^1(\mathscr{D})$ with $\|\cdot\| = \|\cdot\|_H$ and $(\cdot,\cdot) = (\cdot,\cdot)_H$. Moreover, $A\colon V \to V'$ denotes the linear elliptic operator $Au = -\nabla\cdot(\kappa\nabla u)$ for $u \in V$ where $\kappa(x) > \kappa_0 > 0$ is smooth. As usual we consider the bilinear form $a\colon V \times V \to \mathbb{R}$ defined by $a(u, v) = \langle Au, v\rangle$ for $u, v \in V$, and $\langle\cdot,\cdot\rangle$ denotes the duality pairing of V' and V. We denote by e^{-tA} the analytic semigroup in H generated by the realization of $-A$ in H with $D(A) = H^2(\mathscr{D}) \cap H_0^1(\mathscr{D})$. Finally, $f\colon H \to H$ is a nonlinear function, which is assumed to be globally Lipschitz continuous, i.e., there exists a constant L_f such that

$$\|f(u) - f(v)\| \leq L_f\|u - v\|, \quad u, v \in H. \tag{2}$$

It is well known that our assumptions on A and on the spatial domain $\mathscr{D}$ implies the existence of a sequence of nondecreasing positive real numbers $\{\lambda_k\}_{k\geq 1}$ and an orthonormal basis $\{e_k\}_{k\geq 1}$ of H such that

$$Ae_k = \lambda_k e_k, \quad \lim_{k\to+\infty} \lambda_k = +\infty.$$

Using the spectral functional calculus for A we introduce the fractional powers A^s, $s \in \mathbb{R}$, of A as

$$A^s v = \sum_{k=1}^{\infty} \lambda_k^s (v, e_k)e_k, \quad D(A^s) = \Big\{v \in H : \|A^s v\|^2 = \sum_{k=1}^{\infty} \lambda_k^{2s}(v, e_k)^2 < \infty\Big\}$$

and spaces $\dot{H}^\beta = D(A^{\beta/2})$ with norms $\|v\|_\beta = \|A^{\beta/2}v\|_\beta$. It is classical that if $0 \leq \beta < 1/2$, then $\dot{H}^\beta = H^\beta$ and if $1/2 < \beta \leq 2$, then $\dot{H}^\beta = \{u \in H^\beta : u|_{\partial\mathscr{D}} = 0\}$, where H^β denotes the standard Sobolev space of order β. We also use the spaces $L_2(\Omega, \dot{H}^\beta)$ with the mean square norms $\|v\|_{L_2(\Omega,\dot{H}^\beta)} = \big(\mathbb{E}[\|v\|_\beta^2]\big)^{\frac{1}{2}}$.

We assume for some $\beta \geq 0$ that

$$\|A^{\frac{\beta-1}{2}} Q^{\frac{1}{2}}\|_{\mathrm{HS}} < \infty, \quad u_0 \in L_2(\Omega, \dot{H}^\beta). \tag{3}$$

Here Q is the covariance operator of W and $\|\cdot\|_{\mathrm{HS}}$ denotes the Hilbert-Schmidt norm. The Hilbert-Schmidt condition in (3) can be viewed as a regularity assumption on the covariance operator Q. In particular, it holds with $\beta = 1$ if Q is a trace class operator and with $\beta < 1/2$ if $Q = I$ and $d = 1$. More generally, it holds if $\sum_{k=1}^{\infty} \lambda_k^{-\alpha} < \infty$ (thus $\alpha > d/2$) and $A^{\beta+\alpha-1}Q$ is a bounded linear operator on H (see, for example, [10, Theorem 2.1]).

It is known ([9], [11, Lemma 3.1]) that if (2) and (3) hold, then (1) has a unique mild solution, which is defined to be the solution of the fixed point equation

$$u(t) = \mathrm{e}^{-tA}u_0 + \int_0^t \mathrm{e}^{-(t-s)A} f(u(s))\,\mathrm{d}s + \int_0^t \mathrm{e}^{-(t-s)A}\,\mathrm{d}W(s). \tag{4}$$

This naturally splits the solution as $u = v + w$, where w is a stochastic convolution,

$$w(t) = \int_0^t \mathrm{e}^{-(t-s)A}\,\mathrm{d}W(s), \tag{5}$$

which is the solution of

$$\mathrm{d}w + Aw\,\mathrm{d}t = \mathrm{d}W,\ 0 < t \le T; \quad w(0) = 0, \tag{6}$$

and v is the solution of the random evolution equation

$$\dot{v} + Av = f(v + w),\ 0 < t \le T; \quad v(0) = u_0. \tag{7}$$

Our approach will be to first compute w and then to insert it into (7) which we then solve for v. Finally, $u = v + w$. For the numerical solution we use Rothe's method, where we first discretize with respect to time and then discretize the resulting elliptic problems with wavelet methods.

Thus, we fix a time step $\tau > 0$, set $t_n := n\tau$ with $t_N = T$, and consider a backward Euler discretization of (1). With $u^n \approx u(t_n)$ and increments $\Delta W^n = W(t_n) - W(t_{n-1})$ this reads

$$u^n + \tau A u^n = u^{n-1} + \tau f(u^n) + \Delta W^n,\ 1 \le n \le N; \quad u^0 = u_0. \tag{8}$$

Then we decompose $u^n = v^n + w^n$ to get time-discrete versions of (6) and (7):

$$w^n + \tau A w^n = w^{n-1} + \Delta W^n, \qquad 1 \le n \le N; \quad w^0 = 0, \tag{9a}$$

$$v^n + \tau A v^n = v^{n-1} + \tau f(v^n + w^n), \qquad 1 \le n \le N; \quad v^0 = u_0. \tag{9b}$$

This allows us to solve the linear problem (9a) first and use the result as an input for the nonlinear problem (9b). Moreover, the stochastic influence in (9b) is smoother than in (9a), which allows us to use fast nonlinear solvers.

We consider now the spatial discretization of (9). To this end, let S_J be a multiresolution space of order m (see (26) for the definition) and let $\{w_J^n\}_{n=0}^N \subset S_J$ be the corresponding Galerkin approximation of $\{w^n\}_{n=0}^N$, i.e.,

$$w_J^n + \tau A_J w_J^n = w_J^{n-1} + P_J \Delta W^n,\ 1 \le n \le N; \quad w_J^0 = 0. \tag{10}$$

We refer to Sect. 3 for further details. We enter this approximation instead of w^n into (9b). The corresponding equation reads

$$\bar{v}^n + \tau A \bar{v}^n = \bar{v}^{n-1} + \tau f(\bar{v}^n + w_J^n),\ 1 \le n \le N; \quad \bar{v}^0 = u_0. \tag{11}$$

For each $\omega \in \Omega$ and for each $n \geq 1$ the nonlinear equation in (11) is solved by an adaptive wavelet algorithm to yield an approximate solution v_ε^n with tolerance ε_n. Note that we use the same tolerance for each ω. More precisely, denoting $\bar{v}^n = E_n(\bar{v}^{n-1})$, where $E_n = (I + \tau A - \tau f(\cdot + w_J^n))^{-1}$ is the nonlinear one-step operator from (11), we assume that $v_\varepsilon^n = \tilde{E}_n(v_\varepsilon^{n-1})$, where $\tilde{E}_n$ is an approximation of E_n such that

$$\|E_n(v) - \tilde{E}_n(v)\| \leq \varepsilon_n, \quad 1 \leq n \leq N,\ v \in H. \tag{12}$$

The output of the computation will then be the sequence

$$u_\varepsilon^n = v_\varepsilon^n + w_J^n, \quad 0 \leq n \leq N. \tag{13}$$

The total error is $u_\varepsilon^n - u(t_n) = (v_\varepsilon^n - \bar{v}^n) + (\bar{v}^n - v^n) + (w_J^n - w^n) + (u^n - u(t_n))$. The contributions are bounded as follows, where the constants C depend on $\|u_0\|_{L_2(\Omega,\dot{H}^\beta)}$, $\|A^{\frac{\beta-1}{2}} Q^{\frac{1}{2}}\|_{\mathrm{HS}}$, and T, referring to assumption (3). We also assume $\tau L_f < \frac{1}{2}$.

First, in Sect. 2.1, an adaptive wavelet algorithm is described which realizes (12). In Theorem 4, we also analyze the computational effort of the algorithm applied to (11). We conclude the section by showing that

$$\max_{0 \leq t_n \leq T} \|v_\varepsilon^n - \bar{v}^n\|_{L_2(\Omega,H)} \leq C \sum_{n=1}^{N} \varepsilon_n. \tag{14}$$

The multiresolution approximation of the time-discrete stochastic convolution is studied in Sect. 3 and Theorem 5 shows that

$$\max_{0 \leq t_n \leq T} \|w_J^n - w^n\|_{L_2(\Omega,H)} \leq C\, 2^{-J \min(\beta,m)}. \tag{15}$$

In Sect. 4, Theorem 8, we study the time-discretization error and prove that

$$\max_{0 \leq t_n \leq T} \|u^n - u(t_n)\|_{L_2(\Omega,H)} \leq C\, \tau^{\frac{\beta}{2}}, \quad \text{if } 0 \leq \beta < 1. \tag{16}$$

Finally, in Sect. 5, we analyze the perturbation of the nonlinear term and obtain that

$$\max_{0 \leq t_n \leq T} \|\bar{v}^n - v^n\|_{L_2(\Omega,H)} \leq C \max_{0 \leq t_n \leq T} \|w_J^n - w^n\|_{L_2(\Omega,H)}. \tag{17}$$

Therefore, our main result is the following.

Theorem 1. *Assume (3) for some $\beta \geq 0$. Let $\{w_J^n\}_{n=0}^N \subset S_J$ be computed by a multiresolution Galerkin method of order m and $\{v_\varepsilon^n\}_{n=0}^N$ by an adaptive wavelet method with tolerances ε_n. Then for $0 \leq \beta < 1$, the total error in (13) is bounded by*

$$\max_{0\le t_n\le T} \|u_\varepsilon^n - u(t_n)\|_{L_2(\Omega,H)} \le C\,\tau^{\frac{\beta}{2}} + C\,2^{-J\min(\beta,m)} + C\sum_{n=1}^{N}\varepsilon_n,$$

for $\tau L_f < \frac{1}{2}$, *where* $C = C(\|u_0\|_{L_2(\Omega,\dot H^\beta)}, \|A^{\frac{\beta-1}{2}}Q^{\frac12}\|_{\mathrm{HS}}, T)$. *If* $\beta \ge 1$, *then first term is replaced by* $C_\delta\tau^{\frac12-\delta}$, *for any* $\delta > 0$.

The literature on numerics for nonlinear stochastic parabolic problems is now rather large. We mention, for example, [15] on pure time-discretization and [13,18] on complete discretization based on the method of lines, where the spatial discretization is first performed by finite elements and the resulting finite-dimensional evolution problem is then discretized. Wavelets have been used in [12] where the spatial approximation (without adaptivity) of stochastic convolutions were considered.

Our present paper is a first attempt towards spatial adaptivity by using Rothe's method together with known adaptive wavelet methods for solving the resulting nonlinear elliptic problems.

The spatial Besov regularity of solutions of stochastic PDEs is investigated in [2, 3]. The comparison of the Sobolev and Besov regularity is indicative of whether adaptivity is advantageous. For problems on domains with smooth or convex polygonal boundary with boundary adapted additive noise (that is, (3) holds for β high enough), where the solution can be split as $u = v + w$, we expect that the adaptivity is not needed for the stochastic convolution w, which then has sufficient Sobolev regularity. We therefore apply it only to the random evolution problem (7). Once the domain is not convex, or the boundary is not regular, or the noise is not boundary adapted, adaptivity might pay off also for the solution of the linear problem (9a).

The recent paper [1] is a first attempt for a more complete error analysis of Rothe's method for both deterministic and stochastic evolution problems. The overlap with our present work is not too large, since we take advantage of special features of equations with additive noise.

2 Wavelet Approximation

In this section, we collect the notation and the main properties of wavelets that will be needed in the sequel. We refer to [4, 8, 17] for more details on wavelet methods for PDEs. For the space discretization, let

$$\Psi = \{\psi_\lambda : \lambda \in \mathscr{J}^\Psi\}, \quad \tilde\Psi = \{\tilde\psi_\lambda : \lambda \in \mathscr{J}^\Psi\}$$

be a biorthogonal basis of H, i.e., in particular $(\psi_\lambda, \tilde\psi_\mu)_H = \delta_{\delta,\mu}$. Here, λ typically is an index vector $\lambda = (j,k)$ containing both the information on the level $j = |\lambda|$ and the location in space k (e.g., the center of the support of ψ_λ). Note that Ψ also

contains the scaling functions on the coarsest level that are not wavelets. We will refer to $|\lambda| = 0$ as the level of the *scaling functions*.

In addition, we assume that $\psi_\lambda \in V$, which is an assumption on the regularity (and boundary conditions) of the primal wavelets. To be precise, we pose the following assumptions on the wavelet bases:

1. Regularity: $\psi_\lambda \in H^t(\mathscr{D})$, $\lambda \in \mathscr{J}^\Psi$ for all $0 \le t < s_\Psi$;
2. Vanishing moments: $((\cdot)^r, \psi_\lambda)_{0;\mathscr{D}} = 0$, $0 \le r < m_\Psi$, $|\lambda| > 0$.
3. Locality: $\operatorname{diam}(\operatorname{supp} \psi_\Lambda) \sim 2^{-|\lambda|}$.

We assume the same properties for the dual wavelet basis with s_Ψ and m_Ψ replaced by $\tilde{s}_\Psi$ and $\tilde{m}_\Psi$. Note that the dual wavelet $\tilde{\psi}_\lambda$ does not need to be in V, typically one expects $\tilde{\psi}_\lambda \in V'$.

We will consider (often finite-dimensional) subspaces generated by (adaptively generated finite) sets of indices $\Lambda \subset \mathscr{J}^\Psi$ and

$$\Psi_\Lambda := \{\psi_\lambda : \lambda \in \Lambda\}, \quad S_\Lambda := \operatorname{clos} \operatorname{span}(\Psi_\Lambda),$$

where the closure is of course not needed if Λ is a finite set. If $\Lambda = \Lambda_J := \{\lambda \in \mathscr{J}^\Psi : |\lambda| \le J - 1\}$, then $S_J := S_{\Lambda_J}$ contains all wavelets up to level $J - 1$ so that S_J coincides with the multiresolution space generated by all scaling functions on level J, i.e.,

$$S_J = \operatorname{span} \Phi_J, \quad \Phi_J = \{\varphi_{J,k} : k \in \mathscr{I}_J\}, \tag{18}$$

where $\mathscr{I}_J$ is an appropriate index set.

2.1 Adaptive Wavelet Methods for Nonlinear Variational Problems

In this section, we quote from [7] the main facts on adaptive wavelet methods for solving stationary nonlinear variational problems. Note, that all what is said in this section is taken from [7]. However, we abandon further reference for easier reading.

Let $F: V \to V'$ be a nonlinear map. We consider the problem of finding $u \in V$ such that

$$\langle v, R(u)\rangle := \langle v, F(u) - g\rangle = 0, \quad v \in V, \tag{19}$$

where $g \in V'$ is given. As an example, let F be given as $\langle v, F(u)\rangle := a(v,u) + \langle v, f(u)\rangle$ which covers (11). The main idea is to consider an equivalent formulation of (19) in terms of the wavelet coefficients $\mathbf{u}$ of the unknown solution $u = \mathbf{u}^T \Psi$. Setting

$$\mathbf{R}(\mathbf{v}) := (\langle \psi_\lambda, R(v)\rangle)_{\lambda \in \mathscr{J}^\Psi}, \quad v = \mathbf{v}^T \Psi,$$

the equivalent formulation amounts to finding $\mathbf{u} \in \ell_2(\mathcal{J}^\Psi)$ such that

$$\mathbf{R}(\mathbf{v}) = \mathbf{0}. \tag{20}$$

The next ingredient is a basic iteration in the (infinite-dimensional) space $\ell_2(\mathcal{J}^\Psi)$ and replacing the infinite operator applications in an adaptive way by finite approximations in order to obtain a computable version. Starting by some finite $\mathbf{u}^{(0)}$, the iteration reads

$$\mathbf{u}^{(n+1)} = \mathbf{u}^{(n)} - \Delta\mathbf{u}^{(n)}, \quad \Delta\mathbf{u}^{(n)} := \mathbf{B}^{(n)}\mathbf{R}(\mathbf{u}^{(n)}) \tag{21}$$

where the operator $\mathbf{B}^{(n)}$ is to be chosen and determines the nonlinear solution method (such as Richardson or Newton). The sequence $\Delta\mathbf{u}^{(n)} = \mathbf{B}^{(n)}\mathbf{R}(\mathbf{u}^{(n)})$ (possibly infinite even for finite input $\mathbf{u}^{(n)}$) is then replaced by some finite sequence $\mathbf{w}_\eta^{(n)} := \mathbf{RES}[\eta_n, \mathbf{B}^{(n)}, \mathbf{R}, \mathbf{u}^{(n)}]$ such that

$$\|\Delta\mathbf{u}^{(n)} - \mathbf{w}_\eta^{(n)}\| \le \eta_n.$$

Replacing $\Delta\mathbf{u}^{(n)}$ by $\mathbf{w}_\eta^{(n)}$ in (21) and choosing the sequence of tolerances $(\eta_n)_{n\in\mathbb{N}_0}$ appropriately results in a convergent algorithm such that any tolerance ε is reached after finitely many steps. We set $\bar{\mathbf{u}}(\varepsilon) := \mathbf{SOLVE}[\varepsilon, \mathbf{R}, \mathbf{B}^{(n)}, \mathbf{u}^{(0)}]$ such that we get $\|\mathbf{u} - \bar{\mathbf{u}}(\varepsilon)\| \le \varepsilon$.

In terms of optimality, there are several issues to be considered:

- How many iterations $n(\varepsilon)$ are required in order to achieve ε-accuracy?
- How many "active" coefficients are needed to represent the numerical approximation and how does that compare with a "best" approximation?
- How many operations (arithmetic, storage) and how much storage is required?

In order to quantify that, one considers so-called *approximation classes*

$$\mathcal{A}^s := \{\mathbf{v} \in \ell_2(\mathcal{J}^\Psi) : \sigma_N(\mathbf{v}) \lesssim N^{-s}\}$$

of all those sequences whose *error of best N-term approximation*

$$\sigma_N(\mathbf{v}) := \min\{\|\mathbf{v} - \mathbf{w}\|_{\ell_2} : \#\operatorname{supp}\mathbf{w} \le N\}$$

decays at a certain rate ($\operatorname{supp}\mathbf{v} := \{\lambda \in \mathcal{J}^\Psi : v_\lambda \neq 0\}$, $\mathbf{v} = (v_\lambda)_{\lambda\in\mathcal{J}^\Psi}$).

Let us first consider the case where $F = A$ is a linear elliptic partial differential operator, i.e., $Au = g \in V'$, where $A: V \to V'$, $g \in V'$ is given and $u \in V$ is to be determined. For the discretization we use a wavelet basis Ψ in H where rescaled versions admit Riesz bases in V and V', respectively. Then, the operator equation can equivalently be written as

$$\mathbf{Au} = \mathbf{g} \in \ell_2(\mathcal{J}^\Psi),$$

where $\mathbf{A} := \mathbf{D}^{-1}a(\Psi,\Psi)\mathbf{D}^{-1}$, $\mathbf{g} := \mathbf{D}^{-1}(g,\Psi)$ and $\mathbf{u} := \mathbf{D}(u_\lambda)_{\lambda\in\mathcal{J}^\Psi}$, with u_λ being the wavelet coefficients of the unknown function $u \in V$, $\|u\|_V \sim \|\mathbf{u}\|_{\ell_2(\mathcal{J}^\Psi)}$. Wavelet preconditioning results in the fact that $\kappa_2(\mathbf{A}) < \infty$, [5].

The (biinfinite) matrix $\mathbf{A}$ is said to be s^*-*compressible*, $\mathbf{A} \in \mathcal{C}_{s^*}$, if for any $0 < s < s^*$ and every $j \in \mathbb{N}$ there exists a matrix $\mathbf{A}_j$ with the following properties: For some summable sequence $(\alpha_j)_{j\in\mathbb{N}}$, the matrix $\mathbf{A}_j$ is obtained by replacing all but the order of $\alpha_j 2^j$ entries per row and column in $\mathbf{A}$ by zero and satisfies

$$\|\mathbf{A} - \mathbf{A}_j\| \le C\alpha_j 2^{-js}, \quad j \in \mathbb{N}.$$

Wavelet representations of differential (and certain integral) operators fall into this category. Typically, s^* depends on the regularity and the order of vanishing moments of the wavelets. Then, one can construct a linear counterpart $\mathbf{RES}_{\text{lin}}$ of $\mathbf{RES}$ such that $\mathbf{w}_\eta := \mathbf{RES}_{\text{lin}}[\eta, \mathbf{A}, \mathbf{g}, \mathbf{v}]$ for finite input $\mathbf{v}$ satisfies

$$\|\mathbf{w}_\eta - (\mathbf{A}\mathbf{v} - \mathbf{g})\|_{\ell_2} \le \eta, \tag{22a}$$

$$\|\mathbf{w}_\eta\|_{\mathcal{A}^s} \lesssim (\|\mathbf{v}\|_{\mathcal{A}^s} + \|\mathbf{u}\|_{\mathcal{A}^s}), \tag{22b}$$

$$\#\operatorname{supp}\mathbf{w}_\eta \lesssim \eta^{-1/s}(\|\mathbf{v}\|_{\mathcal{A}^s}^{1/s} + \|\mathbf{u}\|_{\mathcal{A}^s}^{1/s}), \tag{22c}$$

where the constants in (22b), (22c) depend only on s. Here, we have used the quasi-norm

$$\|\mathbf{v}\|_{\mathcal{A}^s} := \sup_{N\in\mathbb{N}} N^s \sigma_N(\mathbf{v}).$$

This is the main ingredient for proving optimality of the scheme in the following sense.

Theorem 2 ([5, 7]). *If* $\mathbf{A} \in \mathcal{C}_{s^*}$ *and if the exact solution* $\mathbf{u}$ *of* $\mathbf{A}\mathbf{u} = \mathbf{g}$ *satisfies* $\mathbf{u} \in \mathcal{A}^s$, $s < s^*$, *then* $\bar{\mathbf{u}}(\varepsilon) = \mathbf{SOLVE}_{\text{lin}}[\varepsilon]$ *satisfies*

$$\|\mathbf{u} - \bar{\mathbf{u}}(\varepsilon)\| \le \varepsilon, \tag{23a}$$

$$\#\operatorname{supp}\bar{\mathbf{u}}(\varepsilon) \lesssim \varepsilon^{-1/s}, \tag{23b}$$

$$\textit{computational complexity} \sim \#\operatorname{supp}\bar{\mathbf{u}}(\varepsilon). \tag{23c}$$

It turns out that most of what is said before also holds for the nonlinear case except that the analysis of the approximate evaluation of nonlinear expressions $\mathbf{R}(\mathbf{v})$ poses a constraint on the structure of the active coefficients, namely that it has *tree structure*. In order to define this, one uses the notation $\mu \prec \lambda$, $\lambda, \mu \in \mathcal{J}^\Psi$ to express that μ is a *descendant* of λ. We explain this in the univariate case with $\psi_\lambda = \psi_{j,k} = 2^{j/2}\psi(2^j \cdot -k)$. Then, the *children* of $\lambda = (j,k)$ are, as one would also intuitively define, $\mu = (j+1, 2k)$ and $\nu = (j+1, 2k+1)$. The descendants of λ are its children, the children of its children and so on. In higher dimensions and even on more complex domains this can also be defined – with some more technical effort, however.

Then, a set $\mathscr{T} \subset \mathscr{J}^\Psi$ is called a *tree* if $\lambda \in \mathscr{T}$ implies $\mu \in \mathscr{T}$ for all $\mu \in \mathscr{J}^\Psi$ with $\lambda \prec \mu$. Given this, the error of the *best N-term tree approximation* is given as

$$\sigma_N^{\text{tree}}(\mathbf{v}) := \min\{\|\mathbf{v} - \mathbf{w}\|_{\ell_2} : \mathscr{T} := \#\operatorname{supp}\mathbf{w} \text{ is a tree and } \#\mathscr{T} \le N\}$$

and define the *tree approximation space* as

$$\mathscr{A}^s_{\text{tree}} := \{\mathbf{v} \in \ell_2(\mathscr{J}^\Psi) : \sigma_N^{\text{tree}}(\mathbf{v}) \lesssim N^{-s}\}$$

which is a quasi-normed space under the quasi-norm

$$\|\mathbf{v}\|_{\mathscr{A}^s_{\text{tree}}} := \sup_{N\in\mathbb{N}} N^s \sigma_N^{\text{tree}}(\mathbf{v}).$$

Remark 1. For the case $V = H^t$ (or, a closed subspace of H^t) it is known that the solution being in some Besov space $u \in B_q^{t+ds}(L_q)$, $q = (s + \frac{1}{2})^{-1}$, implies that $\mathbf{u} \in \mathscr{A}^r_{\text{tree}}$, for $r < s$, see [6, Remark 2.3].

The extension of the s^*-compressibility $\mathscr{C}_{s^*}$ is the s^*-*sparsity* of the scheme **RES** which is defined by the following property: *If the exact solution* $\mathbf{u}$ *of* (20) *is in* $\mathscr{A}^s_{\text{tree}}$ *for some* $s < s^*$, *then* $\mathbf{w}_\eta := \mathbf{RES}[\eta, \mathbf{B}, \mathbf{R}, \mathbf{v}]$ *for finite* $\mathbf{v}$ *satisfies*

$$\|\mathbf{w}_\eta\|_{\mathscr{A}^s_{\text{tree}}} \le C(\|\mathbf{v}\|_{\mathscr{A}^s_{\text{tree}}} + \|\mathbf{u}\|_{\mathscr{A}^s_{\text{tree}}} + 1),$$

$$\#\text{supp}\mathbf{w}_\eta \le C\,\eta^{-1/s}(\|\mathbf{v}\|^{1/s}_{\mathscr{A}^s_{\text{tree}}} + \|\mathbf{u}\|^{1/s}_{\mathscr{A}^s_{\text{tree}}} + 1),$$

$$\textit{comp. complexity} \sim C(\eta^{-1/s}(\|\mathbf{v}\|^{1/s}_{\mathscr{A}^s_{\text{tree}}} + \|\mathbf{u}\|^{1/s}_{\mathscr{A}^s_{\text{tree}}} + 1) + \#\mathscr{T}(\text{supp}\mathbf{v})),$$

where C depends only on s when $s \to s^*$ and $\mathscr{T}(\text{supp}\mathbf{v})$ denotes the smallest tree containing supp$\mathbf{v}$. Now, we are ready to collect the main result.

Theorem 3 ([7, Theorem 6.1]). *If* **RES** *is s^*-sparse, $s^* > 0$ and if $\mathbf{u} \in \mathscr{A}^s_{\text{tree}}$ for some $s < s^*$, then the approximations $\bar{\mathbf{u}}(\varepsilon)$ satisfy $\|\mathbf{u} - \mathbf{u}(\varepsilon)\| \le \varepsilon$ with*

$$\#\operatorname{supp}\bar{\mathbf{u}}(\varepsilon) \le C\,\varepsilon^{-1/s}\|\mathbf{u}\|^{1/s}_{\mathscr{A}^s_{\text{tree}}}, \quad \|\bar{\mathbf{u}}(\varepsilon)\|_{\mathscr{A}^s_{\text{tree}}} \le C\,\|\mathbf{u}\|_{\mathscr{A}^s_{\text{tree}}},$$

where C depends only on s when $s \to s^*$. The number of operations is bounded by $C\,\varepsilon^{-1/s}\|\mathbf{u}\|^{1/s}_{\mathscr{A}^s_{\text{tree}}}$.

We remark that since the wavelet transform is of linear complexity the overall number of operations needed is the one mentioned in Theorem 3.

Next we show that the wavelet coefficients $\bar{\mathbf{v}}^n$ of the solution of (11) belong to a certain approximation class $\mathscr{A}^s_{\text{tree}}$ and hence, in view of Theorem 3, we obtain an estimate on the support of $\bar{\mathbf{v}}^n_\varepsilon$ and the number of operations required to compute it.

Theorem 4. *The wavelet coefficients $\bar{\mathbf{v}}^n$ of the solution of* (11) *belong to $\mathscr{A}^s_{\text{tree}}$ for all $s < \frac{1}{2d-2}$, where $d \ge 2$ is the spatial dimension of $\mathscr{D}$.*

Proof. It follows from [1, Lemma 5.15] that $r(\tau A) \in \mathscr{L}(L_2(\mathscr{D}), B_q^r(L_q))$ for $r = \frac{3d-2+4\varepsilon}{2d-2+4\varepsilon}$, where $1/q = (r-1)/d + 1/2$ and $\varepsilon > 0$. Thus, the statement follows from Remark 1 noting that $t = 1$ and hence $r = 1 + ds$. □

We end this section by showing (14); that is, the overall error after n steps, when in every step (11) is solved approximately up to an error tolerance ε_n using the adaptive wavelet algorithm described above. Define

$$E_j^n = E_n \circ \cdots \circ E_{j+1}, \; E_n^n = I; \quad 0 \le j < n \le N,$$

and similarly $\tilde{E}_j^n$. Then we have

$$\begin{aligned} v_\varepsilon^n - \bar{v}^n &= \tilde{E}_0^n(u_0) - E_0^n(u_0) \\ &= \sum_{j=0}^{n-1} \big(E_{j+1}^n(\tilde{E}_0^{j+1}(u_0)) - E_j^n(\tilde{E}_0^j(u_0))\big) \\ &= \sum_{j=0}^{n-1} \big(E_{j+1}^n(\tilde{E}_j^{j+1}(\tilde{E}_0^j(u_0))) - E_{j+1}^n(E_j^{j+1}(\tilde{E}_0^j(u_0)))\big) \\ &= \sum_{j=0}^{n-1} \big(E_{j+1}^n(\tilde{E}_{j+1}(v_\varepsilon^j)) - E_{j+1}^n(E_{j+1}(v_\varepsilon^j))\big). \end{aligned}$$

A simple argument shows that the Lipschitz constant of E_n is bounded by $(1 - \tau L_f)^{-1} \le \mathrm{e}^{c\tau L_f}$ for some $c > 0$, if $\tau L_f \le \frac{1}{2}$, cf. the proof of Lemma 5. Hence E_{j+1}^n has a Lipschitz constant bounded by $\mathrm{e}^{c(t_n - t_{j+1})} \le \mathrm{e}^{ct_N}$. Thus, using (12), we obtain

$$\begin{aligned} \|v_\varepsilon^n - \bar{v}^n\|_{L_2(\Omega,H)} &\le \sum_{j=0}^{n-1} \mathrm{e}^{c(t_n - t_{j+1})} \|\tilde{E}_{j+1}(v_\varepsilon^j) - E_{j+1}(v_\varepsilon^j)\|_{L_2(\Omega,H)} \\ &\le \sum_{j=1}^{n} \mathrm{e}^{c(t_n - t_j)} \varepsilon_j \le \mathrm{e}^{ct_N} \sum_{j=1}^{n} \varepsilon_j. \end{aligned}$$

After taking a mean square we obtain (14).

3 Error Analysis for the Stochastic Convolution

Let $S_J = S_{\Lambda_J}$ be a multiresolution space (18). The multiresolution Galerkin approximation of the equation $Au = f$ in V' is to find $u_J \in S_J$ such that

$$a(u_J, v_J) = (f, v_J) \quad \forall v \in S_J. \tag{24}$$

Define the orthogonal projector $P_J : H \to S_J$ by

$$(P_J v, w_J) = (v, w_J), \quad v \in H,\ w_J \in S_J. \tag{25}$$

Note that P_J can be extended to V' by (25) since $S_J \subset V$. Next, we define the operator $A_J : S_J \to S_J$ by

$$a(A_J v_J, w_J) = a(v_J, w_J), \quad u_J, v_J \in S_J.$$

Then (24) reads $A_J u_J = P_J f$ in S_J. Alternatively we may write $u_J = R_J u$, where $R_J : V \to S_J$ is the *Ritz projector*, defined by

$$a(R_J v, w_J) = a(v, w_J), \quad v \in V,\ w_J \in S_J.$$

The multiresolution space is of order m if

$$\inf_{w_J \in S_J} \|v - w_J\|_{L_2(\Omega,H)} \lesssim 2^{-mJ} \|v\|_{m;\mathscr{D}}, \quad v \in H^m(\mathscr{D}) \cap V. \tag{26}$$

Standard arguments then show, using elliptic regularity thanks to our assumptions on $\mathscr{D}$, that $\|u_J - u\|_{m;\mathscr{D}} \lesssim 2^{-mJ} \|u\|_{m;\mathscr{D}}$, or in other words

$$\|v - R_J v\|_{m;\mathscr{D}} \lesssim 2^{-mJ} \|v\|_{m;\mathscr{D}}, \quad v \in H^m(\mathscr{D}) \cap V. \tag{27}$$

The next lemma is of independent interest and we state it in a general form.

Lemma 1. *Let $-A$ and $-B$ generate strongly continuous semigroups e^{-tA} and e^{-tB} on a Banach space X and let $r(s) = (1+s)^{-1}$. Then, for all $x, y \in X$, $N \in \mathbb{N}$, $\tau > 0$,*

$$\tau \sum_{n=1}^{N} \|r^n(\tau B)y - r^n(\tau A)x\|^p \le \int_0^\infty \|\mathrm{e}^{-tB}y - \mathrm{e}^{-tA}x\|^p \,\mathrm{d}t, \quad 1 \le p < \infty, \tag{28}$$

$$\|r^n(\tau B)y - r^n(\tau A)x\| \le \sup_{t \ge 0} \|\mathrm{e}^{-tB}y - \mathrm{e}^{-tA}x\|. \tag{29}$$

Proof. By the Hille-Phillips functional calculus, we have

$$r^n(\tau B)y - r^n(\tau A)x = \int_0^\infty (\mathrm{e}^{-t\tau B}y - \mathrm{e}^{-t\tau A}x)\, f_n(t)\,\mathrm{d}t, \tag{30}$$

where f_n denotes the nth convolution power of $f(t) = \mathrm{e}^{-t}$. Since $\|f_n\|_{L_1(\mathbb{R}_+)} = 1$ inequality (29) follows immediately by Hölder's inequality. To see (28) we note that f_n is a probability density and hence by Jensen's inequality and (30),

$$\tau\sum_{n=1}^{N}\|r^n(\tau B)y-r^n(\tau A)x\|^p=\tau\sum_{n=1}^{N}\Big\|\int_0^\infty(\mathrm{e}^{-t\tau B}y-\mathrm{e}^{-t\tau A}x)\,f_n(t)\,\mathrm{d}t\Big\|^p$$
$$\le\tau\sum_{n=1}^{N}\int_0^\infty\|\mathrm{e}^{-t\tau B}y-\mathrm{e}^{-t\tau A}x\|^p\,f_n(t)\,\mathrm{d}t$$
$$=\int_0^\infty\|\mathrm{e}^{-tB}y-\mathrm{e}^{-tA}x\|^p\,\mathrm{d}t\,\sup_{t>0}\sum_{n=1}^{\infty}f_n(t).$$

Finally, by monotone convergence, the Laplace transform of $\sum_{n=1}^{\infty}f_n$ is given by

$$\Big(\sum_{n=1}^{\infty}f_n\Big)\hat{}\,(\lambda)=\sum_{n=1}^{\infty}\hat{f}_n(\lambda)=\sum_{n=1}^{\infty}\Big(\frac{1}{1+\lambda}\Big)^n=\frac{1}{\lambda},\quad\lambda>0.$$

Thus, $\sum_{n=1}^{\infty}f_n\equiv 1$ and the proof is complete. □

Next we derive an error estimate for the multiresolution approximation of the semigroup e^{-tA} and its Euler approximation $r^n(\tau A)$.

Lemma 2. *Let S_J be a multiresolution space of order m and let A, A_J, and P_J be as above. Then, for $T\ge 0$, $N\ge 1$, τ, we have*

$$\Big(\int_0^T\|\mathrm{e}^{-tA_J}P_Jv-\mathrm{e}^{-tA}v\|^2\,\mathrm{d}t\Big)^{\frac12}\le C2^{-J\beta}\|v\|_{\beta-1},\quad 0\le\beta\le m,\tag{31}$$

and

$$\Big(\tau\sum_{n=1}^{N}\|r^n(\tau A_J)P_Jv-r^n(\tau A)v\|^2\Big)^{\frac12}\le C2^{-J\beta}\|v\|_{\beta-1},\quad 0\le\beta\le m.\tag{32}$$

Proof. Estimate (31) is known in the finite element context, see for example [16, Theorem 2.5], and may be proved in a completely analogous fashion for using the approximation property (27) of the Ritz projection R_J, the parabolic smoothing (35), and interpolation. Finally, (32) follows from (31) by using Lemma 1 with $x=v$, $y=P_Jv$, and $B=A_J$. (Note that C is independent of T.) □

Now we are ready to consider the multiresolution approximation of w^n in (9a).

Theorem 5. *Let S_J be a multiresolution space of order m and w and w_J^n the solutions of (9a) and (10). If $\|A^{\frac{\beta-1}{2}}Q^{\frac12}\|_{\mathrm{HS}}<\infty$ for some $0\le\beta\le m$, then*

$$(\mathbb{E}[\|w_J^n-w^n\|^2])^{\frac12}\le C2^{-J\beta}\|A^{\frac{\beta-1}{2}}Q^{\frac12}\|_{\mathrm{HS}}.$$

Proof. Let $t_k = k\tau$, $k = 0, \dots, n$. By (10), (9a), and induction,

$$w_J^n - w^n = \sum_{k=1}^{n} \int_{t_{k-1}}^{t_k} \left[r^{n-k+1}(\tau A_J) P_J - r^{n-k+1}(\tau A) \right] \mathrm{d}W(s),$$

whence, by Itô's isometry, we get

$$\begin{aligned} \mathbb{E}[\|w_J^n - w^n\|^2] &= \sum_{k=1}^{n} \int_{t_{k-1}}^{t_k} \left\| \left[r^{n-k+1}(\tau A_J) P_J - r^{n-k+1}(\tau A) \right] Q^{\frac{1}{2}} \right\|_{\mathrm{HS}}^2 \mathrm{d}s \\ &= \sum_{k=1}^{n} \tau \left\| \left[r^k(\tau A_J) P_J - r^k(\tau A) \right] Q^{\frac{1}{2}} \right\|_{\mathrm{HS}}^2. \end{aligned}$$

Let $\{e_l\}_{l=1}^{\infty}$ be an orthonormal basis of H. Then, using Lemma 2, we obtain

$$\begin{aligned} \mathbb{E}[\|w_J^n - w^n\|^2] &= \sum_{l=1}^{\infty} \sum_{k=1}^{n} \tau \| [r^k(\tau A_J) P_J - r^k(\tau A)] Q^{\frac{1}{2}} e_l \|^2 \\ &\le C \sum_{l=1}^{\infty} 2^{-2J\beta} \| Q^{\frac{1}{2}} e_l \|_{\beta-1}^2 = C 2^{-2J\beta} \| A^{\frac{\beta-1}{2}} Q^{\frac{1}{2}} \|_{\mathrm{HS}}^2. \end{aligned}$$

This completes the proof. □

4 Pure Time Discretization

In the proofs below we will often make use of the following well-known facts about the analytic semigroup e^{-tA}, namely

$$\| A^{\alpha} \mathrm{e}^{-tA} \| \le C t^{-\alpha}, \quad \alpha \ge 0,\ t > 0, \tag{33}$$

$$\| (\mathrm{e}^{-tA} - I) A^{-\alpha} \| \le C t^{\alpha}, \quad 0 \le \alpha \le 1,\ t \ge 0, \tag{34}$$

for some $C = C(\alpha)$, see, for example, [14, Chap. II, Theorem 6.4]. Also, by a simple energy argument we may prove

$$\int_0^t \| A^{\frac{1}{2}} \mathrm{e}^{-sA} v \|^2 \, \mathrm{d}s \le \frac{1}{2} \| v \|^2, \quad v \in H,\ t \ge 0. \tag{35}$$

We quote the following existence, uniqueness and stability result from [11, Lemma 3.1]. For the mild, and other solution concepts we refer to [9, Chaps. 6 and 7].

Lemma 3. *If* $\|A^{\frac{\beta-1}{2}}Q^{\frac{1}{2}}\|_{\mathrm{HS}} < \infty$ *for some* $\beta \geq 0$, $u_0 \in L_2(\Omega, H)$, *and* (2) *holds, then there is a unique mild solution* $\{u(t)\}_{t\geq 0}$ *of* (1) *with* $\sup_{t\in[0,T]} \mathbb{E}\|u(t)\|^2 \leq K$, *where* $K = K(u_0, T, L_f)$.

Concerning the temporal regularity of the stochastic convolution we have the following theorem.

Theorem 6. *Let* $\|A^{-\eta}Q^{\frac{1}{2}}\|_{\mathrm{HS}} < \infty$ *for some* $\eta \in [0, \frac{1}{2}]$. *Then the stochastic convolution* $w(t) := \int_0^t \mathrm{e}^{-(t-\sigma)A}\,\mathrm{d}W(\sigma)$ *is mean square Hölder continuous on* $[0,\infty)$ *with Hölder constant* $C = C(\eta)$ *and Hölder exponent* $\frac{1}{2} - \eta$, *i.e.,*

$$\left(\mathbb{E}\|w(t) - w(s)\|^2\right)^{\frac{1}{2}} \leq C|t-s|^{\frac{1}{2}-\eta}, \quad t, s \geq 0.$$

Proof. For $\eta = \frac{1}{2}$ the result follows from Lemma 3. Let $\eta \in [0, \frac{1}{2})$ and, without loss of generality, let $s < t$. By independence of the increments of W,

$$\begin{aligned}\mathbb{E}\|w(t) - w(s)\|^2 &= \mathbb{E}\Big\| \int_s^t \mathrm{e}^{-(t-\sigma)A}\,\mathrm{d}W(\sigma)\Big\|^2 \\ &\quad + \mathbb{E}\Big\| \int_0^s \mathrm{e}^{-(t-\sigma)A} - \mathrm{e}^{-(s-\sigma)A}\,\mathrm{d}W(\sigma)\Big\|^2 = I_1 + I_2.\end{aligned}$$

From Itô's isometry and (33) it follows that

$$\begin{aligned}I_1 &= \mathbb{E}\Big\| \int_s^t A^{\eta}\mathrm{e}^{-(t-\sigma)A}A^{-\eta}\,\mathrm{d}W(\sigma)\Big\|^2 = \int_s^t \|A^{\eta}\mathrm{e}^{-(t-\sigma)A}A^{-\eta}Q^{\frac{1}{2}}\|_{\mathrm{HS}}^2\,\mathrm{d}\sigma \\ &\leq C\int_s^t (t-\sigma)^{-2\eta}\|A^{-\eta}Q^{\frac{1}{2}}\|_{\mathrm{HS}}^2\,\mathrm{d}\sigma \leq \frac{C}{1-2\eta}(t-s)^{1-2\eta}\|A^{-\eta}Q^{\frac{1}{2}}\|_{\mathrm{HS}}^2.\end{aligned}$$

Finally, let $\{e_k\}_{k=1}^{\infty}$ be an orthonormal basis of H. Then, by (34) and (35),

$$\begin{aligned}I_2 &= \int_0^s \|(\mathrm{e}^{-(t-\sigma)A} - \mathrm{e}^{-(s-\sigma)A})Q^{\frac{1}{2}}\|_{\mathrm{HS}}^2\,\mathrm{d}\sigma \\ &= \sum_{k=1}^{\infty}\int_0^s \|(\mathrm{e}^{-(t-s)A} - I)A^{-(\frac{1}{2}-\eta)}A^{\frac{1}{2}-\eta}\mathrm{e}^{-(s-\sigma)A}Q^{\frac{1}{2}}e_k\|^2\,\mathrm{d}\sigma \\ &\leq C(t-s)^{1-2\eta}\sum_{k=1}^{\infty}\int_0^s \|A^{\frac{1}{2}}\mathrm{e}^{-(s-\sigma)A}A^{-\eta}Q^{\frac{1}{2}}e_k\|^2\,\mathrm{d}\sigma \\ &\leq C(t-s)^{1-2\eta}\sum_{k=1}^{\infty}\|A^{-\eta}Q^{\frac{1}{2}}e_k\| = C(t-s)^{1-2\eta}\|A^{-\eta}Q^{\frac{1}{2}}\|_{\mathrm{HS}}^2.\end{aligned} \tag{36}$$

This completes the proof. □

The next result shows that the time regularity of w transfers to the solution of the semilinear problem.

Theorem 7. *If $u_0 \in L_2(\Omega, \dot{H}^\beta)$ and $\|A^{\frac{\beta-1}{2}} Q^{\frac{1}{2}}\|_{\mathrm{HS}} < \infty$ for some $0 \le \beta < 1$, then there is $C = C(T, u_0, \beta)$ such that the mild solution u of (1) satisfies*

$$\left(\mathbb{E}\|u(t) - u(s)\|^2\right)^{\frac{1}{2}} \le C\,|t-s|^{\frac{\beta}{2}}, \quad t, s \in [0, T].$$

Proof. Let $T > 0$ and $0 \le s < t \le T$. Then, by (4),

$$u(t) - u(s) = (\mathrm{e}^{-tA} - \mathrm{e}^{-sA})u_0 + \int_s^t \mathrm{e}^{-(t-r)A} f(u(r))\,\mathrm{d}r + \int_0^s (\mathrm{e}^{-(t-r)A} - \mathrm{e}^{-(s-r)A}) f(u(r))\,\mathrm{d}r + w(t) - w(s).$$

In a standard way, for $0 \le \beta \le 2$, we have $\mathbb{E}\|(\mathrm{e}^{-tA} - \mathrm{e}^{-sA})u_0\|^2 \le C\,|t-s|^\beta \mathbb{E}\|u_0\|_\beta^2$. Using that f is Lipschitz and hence $\|f(u)\| \le C(1 + \|u\|)$, the norm boundedness of the semigroup e^{-tA}, and Lemma 3, we have that

$$\mathbb{E}\left\| \int_s^t \mathrm{e}^{-(t-r)A} f(u(r))\,\mathrm{d}r \right\|^2 \le C\,|t-s|^2 \left(1 + \sup_{r\in[0,T]} \mathbb{E}\|u(r)\|^2\right) \le C\,|t-s|^2.$$

For $0 \le \beta < 1$, by Lemma 3, (33) and (34), it follows that

$$\begin{aligned}
&\mathbb{E}\left\| \int_0^s (\mathrm{e}^{-(t-r)A} - \mathrm{e}^{-(s-r)A}) f(u(r))\,\mathrm{d}r \right\|^2 \\
&\quad \le s\mathbb{E} \int_0^s \|(\mathrm{e}^{-(t-r)A} - \mathrm{e}^{-(s-r)A}) f(u(r))\|^2\,\mathrm{d}r \\
&\quad \le Cs\left(1 + \sup_{r\in[0,T]} \mathbb{E}\|u(r)\|^2\right) \int_0^s \|\mathrm{e}^{-(t-r)A} - \mathrm{e}^{-(s-r)A}\|^2\,\mathrm{d}r \\
&\quad \le Cs \int_0^s \|A^{\frac{\beta}{2}} \mathrm{e}^{-(s-r)A} (\mathrm{e}^{-(t-s)A} - I) A^{-\frac{\beta}{2}}\|^2\,\mathrm{d}r \le C\,|t-s|^\beta s^{2-\beta} \le C\,|t-s|^\beta.
\end{aligned}$$

Finally, by Theorem 6 with $\eta = -\frac{\beta-1}{2}$, we have $\mathbb{E}\|w(t) - w(s)\|^2 \le C\,|t-s|^\beta$, which finishes the proof. □

In order to analyze the order of the backward Euler time-stepping (8) we quote the following deterministic error estimates, where $r(\tau A) = (I + \tau A)^{-1}$.

Lemma 4. *The following error estimates hold for $t_n = n\tau > 0$.*

$$\|[\mathrm{e}^{-n\tau A} - r^n(\tau A)]v\| \le C\tau^{\frac{\beta}{2}} \|v\|_\beta, \quad 0 \le \beta \le 2, \tag{37}$$

$$\|[\mathrm{e}^{-n\tau A} - r^n(\tau A)]v\| \le C\tau t_n^{-1} \|v\|, \tag{38}$$

$$\sum_{k=1}^{n} \tau \left\| [r^k(\tau A) - \mathrm{e}^{-k\tau A}]v \right\|^2 \le C\tau^\beta \|v\|_{\beta-1}^2, \quad 0 \le \beta \le 2. \tag{39}$$

Proof. Estimates (37) and (38) are shown in, for example, [16, Chap. 7]. Estimate (39) can be proved in a similar way as (2.17) in [18, Lemma 2.8]. □

Theorem 8. *If $u_0 \in L_2(\Omega, \dot{H}^\beta)$ and $\|A^{\frac{\beta-1}{2}} Q^{\frac{1}{2}}\|_{\mathrm{HS}} < \infty$ for some $0 \le \beta < 1$, then there is $C = C(T, u_0, \beta)$ such that for $0 < \tau < \frac{1}{2L_f}$, the solutions u of (4) and u^n of (8) satisfy*

$$(\mathbb{E}\|u(t_n) - u^n\|^2)^{\frac{1}{2}} \le C\tau^{\beta/2}, \quad t_n = n\tau \in [0, T].$$

Proof. We have, with $e^n := u(t_n) - u^n$,

$$e^n = [\mathrm{e}^{-t_n A} - r^n(\tau A)]u_0 + \sum_{k=1}^{n} \int_{t_{k-1}}^{t_k} [\mathrm{e}^{-(t_n - s)A} - r^{n-k+1}(\tau A)]\, \mathrm{d}W(s)$$

$$+ \sum_{k=1}^{n} \int_{t_{k-1}}^{t_k} \mathrm{e}^{-(t_n - s)A} f(u(s)) - r^{n-k+1}(\tau A) f(u_k)\, \mathrm{d}s = e_1 + e_2 + e_3.$$

The error e_1 is easily bounded, using (37), as

$$\mathbb{E}\|e_1\|^2 \le C\tau^\beta \mathbb{E}\|u_0\|_\beta^2, \quad 0 \le \beta \le 2.$$

The contribution of e_2 is the linear stochastic error. First, we decompose e_2 as

$$e_2 = \sum_{k=1}^{n} \int_{t_{k-1}}^{t_k} [\mathrm{e}^{-t_{n-k+1}A} - r^{n-k+1}(\tau A)]\, \mathrm{d}W(s)$$

$$+ \sum_{k=1}^{n} \int_{t_{k-1}}^{t_k} [\mathrm{e}^{-(t_n - s)A} - \mathrm{e}^{-t_{n-k+1}A}]\, \mathrm{d}W(s) = e_{21} + e_{22}.$$

Let $\{f_l\}_{l=1}^\infty$ be an ONB of H. By Itô's isometry, the independence of the increments of W and (39),

$$\mathbb{E}\|e_{21}\|^2 = \sum_{k=1}^{n} \tau \|[r^k(\tau A) - \mathrm{e}^{-k\tau A}]Q^{\frac{1}{2}}\|_{\mathrm{HS}}^2 \le = \sum_{l=1}^{\infty} \sum_{k=1}^{n} \tau \|[r^k(\tau A) - \mathrm{e}^{-k\tau A}]Q^{\frac{1}{2}} f_l\|^2$$

$$\le C \sum_{l=1}^{\infty} \tau^\beta \|Q^{\frac{1}{2}} f_l\|_{\beta-1}^2 = C\tau^\beta \|A^{\frac{\beta-1}{2}} Q^{\frac{1}{2}}\|_{\mathrm{HS}}^2, \quad 0 \le \beta \le 2.$$

The term e_{22} can be bounded using a similar argument as in (36) by

$$\mathbb{E}\|e_{22}\|^2 \le C\tau^\beta \|A^{\frac{\beta-1}{2}} Q^{\frac{1}{2}}\|_{\mathrm{HS}}^2, \quad 0 \le \beta \le 2.$$

Next, we can further decompose e_3 as

$$\begin{aligned} e_3 &= \sum_{k=1}^{n} \int_{t_{k-1}}^{t_k} r^{n-k+1}(\tau A)[f(u(t_k)) - f(u_k)]\,\mathrm{d}s \\ &+ \sum_{k=1}^{n} \int_{t_{k-1}}^{t_k} [\mathrm{e}^{-t_{n-k+1}A} - r^{n-k+1}(\tau A)] f(u(t_k))\,\mathrm{d}s \\ &+ \sum_{k=1}^{n} \int_{t_{k-1}}^{t_k} \mathrm{e}^{-t_{n-k+1}A}[f(u(s)) - f(u(t_k))]\,\mathrm{d}s \\ &+ \sum_{k=1}^{n} \int_{t_{k-1}}^{t_k} [\mathrm{e}^{-(t_n-s)A} - \mathrm{e}^{-t_{n-k+1}A}] f(u(s))\,\mathrm{d}s = e_{31} + e_{32} + e_{33} + e_{34}. \end{aligned}$$

By the stability of $r^n(\tau A)$ and the Lipschitz condition on f, we have

$$\mathbb{E}\|e_{31}\|^2 \le 2L_f^2\tau^2\mathbb{E}\|e^n\|^2 + 2L_f^2\tau^2 n \sum_{k=1}^{n-1} \mathbb{E}\|e^k\|^2 \le 2L_f^2\tau^2\mathbb{E}\|e^n\|^2 + C\tau \sum_{k=1}^{n-1} \mathbb{E}\|e^k\|^2.$$

By (38) and Lemma 3, with $\tau t_{n-k+1}^{-1} = (n-k+1)^{-1} = l^{-1}$,

$$\begin{aligned} \mathbb{E}\|e_{32}\|^2 &\le C\mathbb{E}\Big(\sum_{k=1}^{n} \tau\tau t_{n-k+1}^{-1} \|f(u(t_k))\|\Big)^2 \le C\tau^2 \sum_{l=1}^{n} \frac{1}{l^2} \sum_{k=1}^{n} \mathbb{E}\|f(u(t_k))\|^2 \\ &\le C\tau^2 \sum_{k=1}^{n} (1 + \mathbb{E}\|u(t_k)\|^2) \le C\tau t_n \le C\tau. \end{aligned}$$

Furthermore, by Theorem 7,

$$\mathbb{E}\|e_{33}\|^2 \le t_n \sum_{k=1}^{n} \int_{t_{k-1}}^{t_k} \mathbb{E}\|f(u(s)) - f(u(t_k))\|^2\,\mathrm{d}s \le Ct_n^2\tau^\beta \le C\tau^\beta,\ 0 \le \beta < 1.$$

To estimate e_{34} we have, using again that $t_{n-k+1} = t_n - t_{k-1}$ and Lemma 3,

$$\begin{aligned} \mathbb{E}\|e_{34}\|^2 &= \mathbb{E}\Big(\sum_{k=1}^{n} \int_{t_{k-1}}^{t_k} \|[A^{\frac{\beta}{2}} \mathrm{e}^{-(t_n-s)A}(I - \mathrm{e}^{-(s-t_{k-1})A})]A^{-\frac{\beta}{2}} f(u(s))\|\,\mathrm{d}s\Big)^2 \\ &\le Ct_n \sum_{k=1}^{n} \int_{t_{k-1}}^{t_k} (t_n - s)^{-\beta}\tau^\beta \mathbb{E}\|f(u(s))\|^2\,\mathrm{d}s \le C\tau^\beta, \quad 0 \le \beta < 1. \end{aligned}$$

Putting the pieces together, we have

$$\mathbb{E}\|e^n\|^2 \le C\tau^\beta + 2L_f^2\tau^2\mathbb{E}\|e^n\|^2 + C\tau\sum_{k=1}^{n-1}\mathbb{E}\|e^k\|^2, \quad 0 \le \beta < 1.$$

Finally, if $\tau < \frac{1}{2L_f}$, then by the discrete Gronwall lemma,

$$\mathbb{E}\|e^n\|^2 \le C\tau^\beta \mathrm{e}^{Ct_n} \le C\tau^\beta,\ 0 \le \beta < 1,$$

and the theorem is established. □

5 Error Analysis for the Nonlinear Random Problem

In this section we bound the term $\mathbb{E}\big[\|\bar{v}^n - v^n\|^2\big]$ in (17). We use the global Lipschitz condition (2).

Lemma 5. *Assume that* $\tau L_f \le \frac{1}{2}$. *Then, with* $C = 2L_f T e^{2L_f T}$,

$$\max_{1\le n\le N}\Big(\mathbb{E}\big[\|\bar{v}^n - v^n\|^2\big]\Big)^{\frac{1}{2}} \le C\max_{1\le n\le N}\Big(\mathbb{E}\big[\|w_J^n - w^n\|^2\big]\Big)^{\frac{1}{2}}.$$

Proof. Let $e^n := \bar{v}^n - v^n$. Then, we have by (9b) and (11)

$$e^n + \tau A e^n = \tau\big(f(\bar{v}^n + w_J^n) - f(v^n + w^n)\big) + e^{n-1}.$$

Since $e^0 = 0$, we get by induction

$$e^n = \tau\sum_{j=1}^{n}(I + \tau A)^{-(n+1-j)}\big(f(\bar{v}^j + w_J^j) - f(v^j + w^j)\big).$$

In view of the global Lipschitz condition (2), this results in the estimate

$$\begin{aligned}\|e^n\| &\le L_f\tau\sum_{j=1}^{n}\|(I + \tau A)^{-(n+1-j)}\|\,\|\bar{v}^j + w_J^j - v^j - w^j\| \\ &\le L_f\tau\sum_{j=1}^{n}\big(\|w_J^j - w^j\| + \|e^j\|\big),\end{aligned}$$

since $\|(I + \tau A)^{-1}\| \le 1$. Thus, we obtain

$$\|e^n\| \le (1 - L_f\tau)^{-1} L_f \tau \Big(\sum_{j=1}^{n} \|w_J^j - w^j\| + \sum_{j=1}^{n-1} \|e^j\| \Big).$$

With $L_f\tau \le \frac{1}{2}$ we complete the proof by the standard discrete Gronwall lemma. □

References

1. Cioca, P.A., Dahlke, S., Döhring, N., Friedrich, U., Kinzel, S., Lindner, F., Raasch, T., Ritter, K., Schilling, R.: On the convergence analysis of Rothe's method. Preprint Nr. 124, DFG-Schwerpunktprogramm 1324 "Extraktion Quantifizierbarer Information aus Komplexen Systemen" (2012)
2. Cioca, P.A., Dahlke, S., Döhring, N., Kinzel, S., Lindner, F., Raasch, T., Ritter, K., Schilling, R.: Adaptive wavelet methods for the stochastic poisson equation. BIT **52**, 589–614 (2012)
3. Cioca, P.A., Dahlke, S., Kinzel, S., Lindner, F., Raasch, T., Ritter, K., Schilling, R.: Spatial Besov regularity for stochastic partial differential equations on Lipschitz domains. Studia Mathematica **207**, 197–234 (2011)
4. Cohen, A.: Wavelet methods in numerical analysis. Handb. Numer. Anal. **7**, 417–711 (2000)
5. Cohen, A., Dahmen, W., DeVore, R.A.: Adaptive wavelet schemes for elliptic operator equations—convergence rates. Math. Comp. **70**, 27–75 (2001)
6. Cohen, A., Dahmen, W., DeVore, R.A.: Sparse evaluation of compositions of functions using multiscale expansions. SIAM J. Math. Anal. **35**, 279–303 (2003)
7. Cohen, A., Dahmen, W., DeVore, R.A.: Adaptive wavelet schemes for nonlinear variational problems. SIAM J. Numer. Anal. **41**, 1785–1823 (2003)
8. Dahmen, W.: Wavelet and multiscale methods for operator equations. Acta Numerica **6**, 55–228 (1997)
9. Da Prato, G., Zabczyk, J.: Stochastic Equations in Infinite Dimensions. Cambridge University Press, Cambridge (1992)
10. Kovács, M., Larsson, S., Lindgren, F.: Weak convergence of finite element approximations of stochastic evolution equations with additive noise. BIT **52**, 85–108 (2012)
11. Kovács, M., Lindgren, F., Larsson, S.: Strong convergence of the finite element method with truncated noise for semilinear parabolic stochastic equations with additive noise. Numer. Algorithms **53**, 309–320 (2010)
12. Kovács, M., Lindgren, F., Larsson, S.: Spatial approximation of stochastic convolutions. J. Comput. Appl. Math. **235**, 3554–3570 (2011)
13. Kruse, R.: Optimal error estimates of Galerkin finite element methods for stochastic partial differential equations with multiplicative noise. IMA J. Numer. Anal. (2013). doi:10.1093/imanum/drs055
14. Pazy, A.: Semigroups of Linear Operators and Applications to Partial Differential Equations. Springer, New York (1983)
15. Printems, J.: On the discretization in time of parabolic stochastic partial differential equations. Math. Model. Numer. Anal. **35**, 1055–1078 (2001)
16. Thomée, V.: Galerkin Finite Element Methods for Parabolic Problems, 2nd edn. Springer, Berlin (2006)
17. Urban, K.: Wavelet Methods for Elliptic Partial Differential Equations. Oxford University Press, Oxford (2009)
18. Yan, Y.: Galerkin finite element methods for stochastic parabolic partial differential equations. SIAM J. Numer. Anal. **43**, 1363–1384 (2005)

Component-by-Component Construction of Hybrid Point Sets Based on Hammersley and Lattice Point Sets

Peter Kritzer, Gunther Leobacher, and Friedrich Pillichshammer

Abstract In a series of recent articles, such as, e.g., (Hellekalek (2012) Hybrid function systems in the theory of uniform distribution of sequences. Monte Carlo and quasi-Monte Carlo methods 2010. Springer, Berlin, pp. 435–450; Hofer, Kritzer, Larcher, Pillichshammer (Int J Number Theory 5:719–746, 2009); Kritzer (Monatsh Math 168:443–459, 2012); Niederreiter (Acta Arith 138:373–398, 2009)), point sets mixed from integration node sets in different sorts of quasi-Monte Carlo rules have been studied. In particular, a finite version, based on Hammersley and lattice point sets, was introduced in Kritzer (Monatsh Math 168:443–459, 2012), where the existence of such hybrid point sets with low star discrepancy was shown. However, up to now it has remained an open problem whether such low discrepancy hybrid point sets can be explicitly constructed. In this paper, we solve this problem and discuss component-by-component constructions of the desired point sets.

1 Introduction

In many applications of mathematics (most notably in numerical integration) one is in need of having at hand finite or infinite sequences of points which are evenly spread in a certain domain, which, for the sake of simplicity, will be assumed to be the unit cube $[0,1)^s$. It is well-known that point sets which are evenly distributed in the unit cube yield a low integration error when applying a quasi-Monte Carlo (QMC) algorithm using the point set as integration nodes (see, e.g., [2, 3, 17, 19, 21, 29] for comprehensive introductions to this topic). Whenever we speak of a point set in the following, we mean a multi-set, i.e., points may occur repeatedly, and, if so,

P. Kritzer (✉) · G. Leobacher · F. Pillichshammer
Institute of Financial Mathematics, Johannes Kepler University Linz, Altenbergerstrasse 69, 4040 Linz, Austria
e-mail: peter.kritzer@jku.at; gunther.leobacher@jku.at; friedrich.pillichshammer@jku.at

J. Dick et al. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2012*, Springer Proceedings in Mathematics & Statistics 65,
DOI 10.1007/978-3-642-41095-6_25,

are counted repeatedly. There are several well-known classes of point sets that are traditionally considered when one speaks of point sets with excellent distribution properties, such as, e.g., (t, m, s)-nets (see [2,21]) or lattice point sets (see [21,29]).

As one is interested in point sets that are evenly distributed in the unit cube, one is in need of a way of measuring uniformity of distribution. One way of actually assessing the quality of distribution of a point set is to consider its star discrepancy, which is defined as follows.

Definition 1. Let $\mathscr{P}_N$ be a point set of N points in $[0,1)^s$. The *star discrepancy* of $\mathscr{P}_N$ is defined as

$$D_N^*(\mathscr{P}_N) := \sup_{I \subseteq [0,1)^s} \left| \frac{A_N(\mathscr{P}_N, I)}{N} - \lambda_s(I) \right|,$$

where the supremum is extended over all half-open, rectangular subintervals I of $[0,1)^s$ with their lower left corner in the origin, where $A_N(\mathscr{P}_N, I)$ is the number of points of $\mathscr{P}_N$ contained in I, and where λ_s denotes the s-dimensional Lebesgue measure.

If we consider an infinite sequence $\mathscr{S}$ of points in $[0,1)^s$, then $D_N^*(\mathscr{S})$ denotes the star discrepancy of the first N elements of $\mathscr{S}$.

Remark 1. From the definition, it is easy to see that the star discrepancy always takes on values in $[0,1]$. The star discrepancy is a measure of uniformity of distribution, i.e., point sets that are evenly distributed in the unit cube have a lower star discrepancy than point sets that are not evenly distributed. An infinite sequence $\mathscr{S}$ of points in $[0,1)^s$ is called uniformly distributed if and only if its star discrepancy tends to zero for growing N.

Remark 2. It is well-known in the theory of low discrepancy point sets that computing exact values of the star discrepancy is, at least for large values of s and/or N, infeasible (in fact, as shown in [4], calculating the star discrepancy is NP-hard). Therefore, one usually has to resort to considering bounds on the discrepancy of a given point set (see, e.g., [2,21]).

Remark 3. The star discrepancy of a point set $\mathscr{P}_N$ is related to the integration error of a QMC algorithm based on the points $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1}$ of $\mathscr{P}_N$ via the Koksma-Hlawka inequality

$$\left| \int_{[0,1]^s} f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} - \frac{1}{N} \sum_{k=0}^{N-1} f(\boldsymbol{x}_k) \right| \leq D_N^*(\mathscr{P}_N) V(f),$$

where $V(f)$ denotes the variation of f in the sense of Hardy and Krause; see [2,3, 17,21] for further information.

In what follows, we are going to study a combination of two well-known classes of finite point sets with low discrepancy. The first are Hammersley point sets, which

are based on the so-called radical inverse function. Here and in the following, we denote by $\mathbb{N}_0$ the set of non-negative integers and by $\mathbb{N}$ the set of positive integers.

Definition 2. Let $b \geq 2$ be an integer. For $n \in \mathbb{N}_0$, let $n = n_0 + n_1 b + n_2 b^2 + \cdots$ be the base b expansion of n with digits $n_i \in \{0, 1, \ldots, b-1\}$ for $i \geq 0$. The *radical inverse function* $\varphi_b : \mathbb{N}_0 \to [0, 1)$ is defined by

$$\varphi_b(n) := \sum_{k=1}^{\infty} n_{k-1} b^{-k}.$$

Based on the radical inverse function, we can now define the s-dimensional Hammersley point set.

Definition 3. Let $s \geq 2$ and let $b_1, \ldots, b_{s-1}$ be integers, $b_i \geq 2$ for all indices $1 \leq i \leq s-1$. Let $N \in \mathbb{N}$. Then the s-dimensional *Hammersley point set* in bases $b_1, \ldots, b_{s-1}$ is defined to be the point set $\mathscr{H}_N = (\boldsymbol{x}_n)_{n=0}^{N-1} \subseteq [0, 1)^s$, where

$$\boldsymbol{x}_n = \left(\frac{n}{N}, \varphi_{b_1}(n), \varphi_{b_2}(n), \ldots, \varphi_{b_{s-1}}(n)\right) \quad \text{for all } 0 \leq n \leq N-1.$$

Remark 4. It is well-known (see, e.g., [1,2,21]) that the Hammersley point set has good distribution properties in $[0, 1)^s$ if the bases $b_1, \ldots, b_{s-1}$ are pairwise co-prime. To be more precise, the s-dimensional Hammersley point set in pairwise co-prime bases $b_1, \ldots, b_{s-1}$ of N points satisfies

$$D_N^*(\mathscr{H}_N) = O((\log N)^{s-1}/N),$$

where the implied constant depends on s and $b_1, \ldots, b_{s-1}$. For this reason, we are frequently going to choose the bases $b_1, \ldots, b_{s-1}$ as distinct prime numbers $p_1, \ldots, p_{s-1}$ in the following.

Another class of finite point sets that are known to have low discrepancy is that of lattice point sets, going back to Korobov [14] and Hlawka [7].

Definition 4. Let N be an integer and let $\boldsymbol{g} = (g_1, \ldots, g_d)$ be a d-dimensional vector of integers. The *lattice point set* $\mathscr{L}_N = (\boldsymbol{y}_n)_{n=0}^{N-1}$ with *generating vector* $\boldsymbol{g}$, consisting of N points in $[0, 1)^d$, is defined by

$$\boldsymbol{y}_n = \left(\left\{\frac{ng_1}{N}\right\}, \ldots, \left\{\frac{ng_d}{N}\right\}\right) \quad \text{for all } \ 0 \leq n \leq N-1,$$

where $\{\cdot\}$ denotes the fractional part of a number. For short, we write

$$\boldsymbol{y}_n = \left(\left\{\frac{n\boldsymbol{g}}{N}\right\}\right) \quad \text{for all } \ 0 \leq n \leq N-1.$$

Remark 5. Note that it is sufficient to consider only the generating vectors $\boldsymbol{g} \in \{0, 1, \ldots, N-1\}^d$ in the definition of lattice point sets.

Remark 6. Regarding the discrepancy of lattice point sets, it was shown in [20] that, given N, there always exists a d-dimensional generating vector $\boldsymbol{g}$, such that the d-dimensional lattice point set $\mathscr{L}_N$ of N points, generated by $\boldsymbol{g}$, satisfies

$$D_N^*(\mathscr{L}_N) = O((\log N)^d/N),$$

where the implied constant depends on d. It was shown later by Joe [11] (see also [2]) that, at least for prime N, generating vectors $\boldsymbol{g}$ yielding lattice points with a discrepancy as above can be constructed by means of a component-by-component algorithm which chooses one component of the generating vector at a time. Component-by-component (CBC) constructions are nowadays quite standard in the theory of lattice point sets and they apply to many quality measures, such as the worst-case error criteria in Korobov and Sobolev spaces (see, e.g., [18,30]). This approach was first introduced by Korobov [15] and later it was re-invented by Sloan and Reztsov [31].

Considerable progress has been made during the last years in the analysis of QMC algorithms where low discrepancy point sets served as the integration nodes. Traditionally, these approaches only use one fixed sort of point set in the algorithm. However, Spanier [32] considered a mixture of quasi-Monte Carlo and Monte Carlo methods, where he suggested concatenating vectors of QMC point sets and (pseudo-) random vectors, i.e., one obtains $(s + d)$-dimensional "hybrid" point sets or sequences. Recent results regarding hybrid sequences built from QMC point sets and pseudo-random points can be found in [22–26] and the references therein. Moreover, the paper [5] contains a collection of general results regarding these point sets.

Here, we do not consider the concatenation of (pseudo-) random and quasi-random point sets, but we deal with a slightly different concept of a hybrid sequence, namely one that is obtained by using different QMC point sets. Therefore, whenever we speak of a hybrid point set or sequence in the following, we mean a set of points that is obtained by concatenating the components of two different quasi-random point sets (see also Owen [28], where a randomized variant is considered). As pointed out by Keller in [13], such hybrid point sets can be used in computer graphics.

The rest of the paper is structured as follows. In Sect. 2 we shall define a special type of finite hybrid point sets that will be studied in the rest of the paper. These are a mixture of Hammersley point sets on the one hand, and lattice point sets on the other hand. In Sect. 3 we are going to show a CBC construction of such point sets with particularly low discrepancy. We would like to stress that the aim of the current article is not to find general point sets with a discrepancy as low as possible, but the aim is to find hybrid point sets with low star discrepancy.

2 Finite Hybrid Quasi-Random Point Sets

In several papers, such as, e.g., [8–10], hybrid point sets with infinitely many points were considered. The drawback of these considerations is that, at least in many cases, it is not known how to explicitly find hybrid point sets with good distribution properties. For this reason, in the paper [16] a finite version of hybrid point sets was introduced.

To be more precise, it was shown in [16] that there exist finite hybrid point sets, consisting of N points (N a prime) which are obtained from Hammersley and lattice point sets, that have low discrepancy. Indeed, let $p_1, \ldots, p_{s-1}$ be $s-1$ distinct prime numbers and, for technical reasons, let N be an odd prime number that is different from $p_1, \ldots, p_{s-1}$. Furthermore, we write $G_N = \{0, 1, \ldots, N-1\}$ in the following. Let $(\boldsymbol{x}_n)_{n=0}^{N-1}$ be the s-dimensional Hammersley point set in bases $p_1, \ldots, p_{s-1}$. Let $(\boldsymbol{y}_n)_{n=0}^{N-1}$ be the d-dimensional lattice point set generated from the vector $\boldsymbol{g} \in G_N^d$, i.e., $\boldsymbol{y}_n = \left\{\frac{n\boldsymbol{g}}{N}\right\}$ for $0 \le n \le N-1$. Define now

$$\mathscr{P}_N := ((\boldsymbol{x}_n, \boldsymbol{y}_n))_{n=0}^{N-1} \in [0,1)^{s+d}.$$

By an averaging argument, it was shown in [16] that for any prime N different from $p_1, \ldots, p_{s-1}$ there exists a generating vector $\boldsymbol{g} \in G_N^d$ for which

$$D_N^*(\mathscr{P}_N) = O((\log N)^{s+d}/N) \tag{1}$$

holds, where the implied constant depends on s, d and $p_1, \ldots, p_{s-1}$. Theoretically speaking, one could now search for a "good" generating vector $\boldsymbol{g} \in G_N^d$ to explicitly construct low discrepancy hybrid point sets, which is a step forward in comparison to many existence results on infinite hybrid sequences, where it is not clear how one could search for examples with low discrepancy. However, a search over G_N^d is practically infeasible if d and N are not very small. A similar observation is true if one considers only Korobov type generating vectors $\boldsymbol{g}$, as it was also discussed in [16]. Therefore, we are, as outlined in the conclusion of [16], in need of an efficient construction algorithm for a generating vector $\boldsymbol{g} \in G_N^d$ such that a bound as in (1) holds.

As we already noted above, if one considers "pure" lattice point sets, it is known due to earlier papers that there exist CBC constructions yielding low discrepancy. Therefore, it is near at hand to also use a CBC construction for the lattice part of the hybrid point sets under consideration. We introduce such an algorithm in the following section.

3 A CBC Construction

As before, let $p_1, \ldots, p_{s-1}$ be $s-1$ distinct prime numbers and let N be an odd prime number that is different from $p_1, \ldots, p_{s-1}$. Let $(\boldsymbol{x}_n)_{n=0}^{N-1}$ be the s-dimensional Hammersley point set in bases $p_1, \ldots, p_{s-1}$. Let $(\boldsymbol{y}_n)_{n=0}^{N-1}$ be the d-dimensional lattice point set generated from the vector $\boldsymbol{g} \in G_N^d$, i.e., $\boldsymbol{y}_n = \left\{\frac{n\boldsymbol{g}}{N}\right\}$ for $0 \le n \le N-1$. Let now

$$\mathscr{P}_N = ((\boldsymbol{x}_n, \boldsymbol{y}_n))_{n=0}^{N-1} \in [0,1)^{s+d}.$$

In this section, we are going to show how we can use a CBC algorithm to construct a generating vector $\boldsymbol{g} \in G_N^d$ such that $D_N^*(\mathscr{P}_N)$ is low.

For $1 \le i \le s-1$ let m_i be the minimal integer such that $N \le p_i^{m_i}$, i.e., $m_i = \lceil \log_{p_i} N \rceil \le \frac{\log N}{\log p_i} + 1$.

In the derivation of (1) in [16], the constants not depending on the cardinality N of $\mathscr{P}_N$ were not given in their explicit form. However, going through exactly the same steps as in the proof of [16, Eq. (5)], one can also give a bound on $D_N^*(\mathscr{P}_N)$ where all constants can be stated explicitly. This discrepancy bound is given in Eq. (2) below. At this stage we need to introduce some further notation: for $N \in \mathbb{N}$, we define $C(N)$ as the set $(-N/2, N/2] \cap \mathbb{Z}$, $C_d(N)$ as $C(N)^d$, and $C_d^*(N)$ as $C_d(N) \setminus \{\boldsymbol{0}\}$. For $h \in \mathbb{Z}$, we define $r(h) := \max(1, |h|)$ and we put $r(\boldsymbol{h}) = \prod_{j=1}^d r(h_j)$ for $\boldsymbol{h} = (h_1, \ldots, h_d)$. Let further $S_N := \sum_{h \in C^*(N)} |h|^{-1}$ and let $\|x\|$ to denote the distance of a real x to the nearest integer.

Now we have

$$ND_N^*(\mathscr{P}_N) \le 1 + (2s + (1+S_N)^d) \prod_{i=1}^{s-1} m_i p_i + \mathscr{T}_N(\boldsymbol{g}) \prod_{i=1}^{s-1} p_i, \tag{2}$$

with

$$\mathscr{T}_N(\boldsymbol{g}) = \sum_{j_1=1}^{m_1} \cdots \sum_{j_{s-1}=1}^{m_{s-1}} \sum_{\boldsymbol{h} \in C_d^*(N)} \frac{1}{r(\boldsymbol{h})} H_{N,Q}(\boldsymbol{h} \cdot \boldsymbol{g}),$$

where $Q = p_1^{j_1} \cdots p_{s-1}^{j_{s-1}}$ and where

$$H_{N,Q}(\boldsymbol{h} \cdot \boldsymbol{g}) = \begin{cases} N/Q & \text{if } \boldsymbol{h} \cdot \boldsymbol{g} \equiv 0 \bmod N, \\ \|Q\boldsymbol{h} \cdot \boldsymbol{g}/N\|^{-1} & \text{if } \boldsymbol{h} \cdot \boldsymbol{g} \not\equiv 0 \bmod N. \end{cases}$$

In [16], an averaging argument was used to show the existence of a generating vector $\boldsymbol{g} \in G_N^d$ for which (1) holds. Here we would like to use (2) as a starting point for a CBC algorithm for the construction of a good generating vector $\boldsymbol{g}$, i.e., one component of $\boldsymbol{g}$ will be constructed at a time. To begin with, we analyze how the quantity $\mathscr{T}_N$ can be computed effectively.

3.1 Computing the Quantity $\mathscr{T}_N$

Throughout this section, we let $\hat{f}$ denote the discrete Fourier transform of a function $f : C(N) \to \mathbb{C}$, that is

$$\hat{f}(k) = \sum_{h \in C(N)} f(h) \exp(-2\pi \mathtt{i} hk/N), \quad k \in C(N).$$

For $n \in G_N$, let

$$\begin{aligned} \mathscr{R}_N(\boldsymbol{g}, n) &:= -1 + \prod_{i=1}^{d} \left(\sum_{h \in C(N)} \frac{1}{r(h)} \exp(-2\pi \mathtt{i} h g_i n/N) \right) \\ &= -1 + \prod_{i=1}^{d} \widehat{\left(\frac{1}{r} \right)}(g_i n). \end{aligned} \tag{3}$$

We can now write

$$\begin{aligned} \mathscr{T}_N(\boldsymbol{g}) &= \sum_{j_1=1}^{m_1} \cdots \sum_{j_{s-1}=1}^{m_{s-1}} \sum_{k=0}^{N-1} H_{N,Q}(k) \sum_{\substack{\boldsymbol{h} \in C_d^*(N) \\ \boldsymbol{h} \cdot \boldsymbol{g} \equiv k \bmod N}} \frac{1}{r(\boldsymbol{h})} \\ &= \sum_{j_1=1}^{m_1} \cdots \sum_{j_{s-1}=1}^{m_{s-1}} \sum_{k=0}^{N-1} H_{N,Q}(k) \sum_{\boldsymbol{h} \in C_d^*(N)} \frac{1}{r(\boldsymbol{h})} \\ &\quad \times \frac{1}{N} \sum_{n=0}^{N-1} \exp\left(2\pi \mathtt{i} \frac{\boldsymbol{h} \cdot \boldsymbol{g} - k}{N} n \right) \\ &= \sum_{j_1=1}^{m_1} \cdots \sum_{j_{s-1}=1}^{m_{s-1}} \frac{1}{N} \sum_{n=0}^{N-1} \sum_{k=0}^{N-1} H_{N,Q}(k) \exp\left(-2\pi \mathtt{i} \frac{k}{N} n \right) \\ &\quad \times \sum_{\boldsymbol{h} \in C_d^*(N)} \frac{1}{r(\boldsymbol{h})} \exp\left(2\pi \mathtt{i} \frac{\boldsymbol{h} \cdot \boldsymbol{g}}{N} n \right) \\ &= \sum_{n=0}^{N-1} \mathscr{R}_N(\boldsymbol{g}, n) \sum_{j_1=1}^{m_1} \cdots \sum_{j_{s-1}=1}^{m_{s-1}} \frac{1}{N} \sum_{k=0}^{N-1} H_{N,Q}(k) \exp\left(-2\pi \mathtt{i} \frac{k}{N} n \right). \end{aligned}$$

Note that

$$\sum_{k=0}^{N-1} H_{N,Q}(k) \exp\left(-2\pi \mathrm{i} \frac{k}{N} n\right) = \sum_{k \in C(N)} H_{N,Q}(k) \exp\left(-2\pi \mathrm{i} \frac{k}{N} n\right) = \widehat{H_{N,Q}}(n).$$

Therefore the vector F_N with entries

$$F_N(n) := \sum_{j_1=1}^{m_1} \cdots \sum_{j_{s-1}=1}^{m_{s-1}} \frac{1}{N} \sum_{k=0}^{N-1} H_{N,Q}(k) \exp\left(-2\pi \mathrm{i} \frac{k}{N} n\right), \quad 0 \le n \le N-1, \tag{4}$$

can be computed using $O(N \log(N)^s)$ operations by fast Fourier transform. We state this as a lemma for later reference.

Lemma 1. *The vector F_N defined in (4) can be computed using $O(N(\log N)^s)$ operations.*

Now, using (3),

$$\begin{aligned}
\mathscr{T}_N(\boldsymbol{g}) &= \sum_{n=0}^{N-1} F_N(n) \mathscr{R}_N(\boldsymbol{g}, n) \\
&= -\sum_{n=0}^{N-1} F_N(n) + \sum_{n=0}^{N-1} F_N(n) \prod_{i=1}^{d} \widehat{\left(\frac{1}{r}\right)}(g_i n) \\
&= -\sum_{n=0}^{N-1} F_N(n) + \sum_{n=0}^{N-1} \widehat{\left(\frac{1}{r}\right)}(g_d n) \left(F_N(n) \prod_{i=1}^{d-1} \widehat{\left(\frac{1}{r}\right)}(g_i n) \right).
\end{aligned} \tag{5}$$

For a given $\boldsymbol{g} = (g_1, \ldots, g_d)$, $\mathscr{T}_N(\boldsymbol{g})$ can therefore be computed recursively using $O(N(\log(N)^s + d))$ operations. Note that F_N and $\widehat{\left(\frac{1}{r}\right)}$ have to be computed only once. Thus, having pre-computed F_N and $\widehat{\left(\frac{1}{r}\right)}$, a straightforward CBC construction of a vector $\boldsymbol{g}$ with a low value of $\mathscr{T}_N(\boldsymbol{g})$ would require $O(N \log(N)^s + N^2 d)$ operations. However, we are going to present an algorithm with a better run-time in the next section (see Algorithm 1 and the subsequent comments).

3.2 The CBC Construction

In this section, we would like to show a CBC algorithm for generating a vector $\boldsymbol{g} \in G_N^d$ such that (1) is satisfied. Before we state the algorithm and show that it indeed generates a vector of the desired quality, we prove three lemmas.

Lemma 2. *Let N be an odd prime and let Q be an integer such that* $\gcd(N, Q) = 1$. *Then we have* $\sum_{k=1}^{N-1} \|Qk/N\|^{-1} = NS_N$.

Proof. If k runs through all elements of $\{1, \ldots, N-1\}$, then the residues of Qk modulo N run through all elements of $\{1, \ldots, N-1\}$ as well. Hence we have

$$\sum_{k=1}^{N-1} \left\| \frac{Qk}{N} \right\|^{-1} = \sum_{h=1}^{N-1} \left\| \frac{h}{N} \right\|^{-1} = N \sum_{h=1}^{(N-1)/2} \frac{1}{h} + N \sum_{h=(N+1)/2}^{N-1} \frac{1}{N-h}$$

$$= 2N \sum_{h=1}^{(N-1)/2} \frac{1}{h} = NS_N.$$

This completes the proof. □

Lemma 3. *There exists a $g \in G_N$ such that*

$$\mathscr{T}_N(g) \le S_N \left[\prod_{i=1}^{s-1} \frac{1}{p_i - 1} + S_N \prod_{i=1}^{s-1} m_i \right].$$

Proof. We have

$$\frac{1}{N} \sum_{g \in G_N} \mathscr{T}_N(g) = \sum_{j_1=1}^{m_1} \cdots \sum_{j_{s-1}=1}^{m_{s-1}} \frac{1}{N} \sum_{g \in G_N} \sum_{h \in C^*(N)} \frac{1}{r(h)} H_{N,Q}(hg).$$

Now, using Lemma 2,

$$\frac{1}{N} \sum_{g \in G_N} \sum_{h \in C^*(N)} \frac{1}{r(h)} H_{N,Q}(hg)$$

$$= \frac{1}{N} \sum_{g \in G_N} \sum_{k=0}^{N-1} H_{N,Q}(k) \sum_{\substack{h \in C^*(N) \\ hg \equiv k \bmod N}} \frac{1}{r(h)}$$

$$= \frac{1}{Q} \sum_{h \in C^*(N)} \frac{1}{r(h)} \sum_{\substack{g \in G_N \\ hg \equiv 0 \bmod N}} 1$$

$$+ \frac{1}{N} \sum_{k=1}^{N-1} \|Qk/N\|^{-1} \sum_{h \in C^*(N)} \frac{1}{r(h)} \sum_{\substack{g \in G_N \\ hg \equiv k \bmod N}} 1$$

$$= \frac{1}{Q} S_N + S_N^2.$$

Since $Q = p_1^{j_1} \cdots p_{s-1}^{j_{s-1}}$ we obtain

$$\frac{1}{N} \sum_{g \in G_N} \mathscr{T}_N(g) = \sum_{j_1=1}^{m_1} \cdots \sum_{j_{s-1}=1}^{m_{s-1}} \left[\frac{1}{p_1^{j_1} \cdots p_{s-1}^{j_{s-1}}} S_N + S_N^2 \right]$$

$$\le S_N \left[\prod_{i=1}^{s-1} \frac{1}{p_i - 1} + S_N \prod_{i=1}^{s-1} m_i \right],$$

and the result follows. □

For $\boldsymbol{g} = (g_1, \ldots, g_d) \in G_N^d$ and $g_{d+1} \in G_N$ we write in the following $(\boldsymbol{g}, g_{d+1}) := (g_1, \ldots, g_d, g_{d+1})$.

Lemma 4. *Let N be an odd prime number. Assume that there exists $\boldsymbol{g} \in G_N^d$ such that*

$$\mathscr{T}_N(\boldsymbol{g}) \le [(1 + S_N)^d - 1] \left[\prod_{i=1}^{s-1} \frac{1}{p_i - 1} + S_N \prod_{i=1}^{s-1} m_i \right].$$

Then there exists $g_{d+1} \in G_N$, such that

$$\mathscr{T}_N((\boldsymbol{g}, g_{d+1})) \le [(1 + S_N)^{d+1} - 1] \left[\prod_{i=1}^{s-1} \frac{1}{p_i - 1} + S_N \prod_{i=1}^{s-1} m_i \right].$$

Proof. Let

$$\mathscr{T}_N(\boldsymbol{g}, Q) := \sum_{\boldsymbol{h} \in C_d^*(N)} \frac{1}{r(\boldsymbol{h})} H_{N,Q}(\boldsymbol{h} \cdot \boldsymbol{g}).$$

We show the result by an averaging argument. We have

$$\frac{1}{N} \sum_{g_{d+1} \in G_N} \mathscr{T}_N((\boldsymbol{g}, g_{d+1}), Q)$$

$$= \frac{1}{N} \sum_{g_{d+1} \in G_N} \sum_{(\boldsymbol{h}, h_{d+1}) \in C_{d+1}^*(N)} \frac{1}{r((\boldsymbol{h}, h_{d+1}))} H_{N,Q}((\boldsymbol{h}, h_{d+1}) \cdot (\boldsymbol{g}, g_{d+1}))$$

$$= \mathscr{T}_N(\boldsymbol{g}, Q)$$

$$+ \frac{1}{N} \sum_{g_{d+1} \in G_N} \sum_{k=0}^{N-1} H_{N,Q}(k) \sum_{\substack{(\boldsymbol{h}, h_{d+1}) \in C_d(N) \times C^*(N) \\ g_{d+1} h_{d+1} \equiv k - \boldsymbol{h} \cdot \boldsymbol{g} \bmod N}} \frac{1}{r((\boldsymbol{h}, h_{d+1}))}$$

$$
\begin{aligned}
&= \mathscr{T}_N(\boldsymbol{g}, Q) \\
&\quad + \frac{1}{Q} \sum_{(\boldsymbol{h},h_{d+1})\in C_d(N)\times C^*(N)} \frac{1}{r((\boldsymbol{h},h_{d+1}))} \sum_{\substack{g_{d+1}\in G_N \\ g_{d+1}h_{d+1}\equiv -\boldsymbol{h}\cdot\boldsymbol{g} \bmod N}} 1 \\
&\quad + \frac{1}{N} \sum_{k=1}^{N-1} \|Qk/N\|^{-1} \sum_{(\boldsymbol{h},h_{d+1})\in C_d(N)\times C^*(N)} \frac{1}{r((\boldsymbol{h},h_{d+1}))} \\
&\qquad \times \sum_{\substack{g_{d+1}\in G_N \\ g_{d+1}h_{d+1}\equiv k-\boldsymbol{h}\cdot\boldsymbol{g} \bmod N}} 1 \\
&\le \mathscr{T}_N(\boldsymbol{g}, Q) + \frac{1}{Q} \sum_{(\boldsymbol{h},h_{d+1})\in C_d(N)\times C^*(N)} \frac{1}{r((\boldsymbol{h},h_{d+1}))} \\
&\quad + \frac{1}{N} \sum_{k=1}^{N-1} \|Qk/N\|^{-1} \sum_{(\boldsymbol{h},h_{d+1})\in C_d(N)\times C^*(N)} \frac{1}{r((\boldsymbol{h},h_{d+1}))} \\
&= \mathscr{T}_N(\boldsymbol{g}, Q) + \frac{1}{Q}(1+S_N)^d S_N + (1+S_N)^d S_N^2,
\end{aligned}
$$

where we used Lemma 2. Therefore

$$
\begin{aligned}
&\frac{1}{N} \sum_{g_{d+1}\in G_N} \mathscr{T}_N((\boldsymbol{g}, g_{d+1})) \\
&\le \mathscr{T}_N(\boldsymbol{g}) + (1+S_N)^d S_N \left[\prod_{i=1}^{s-1} \frac{1}{p_i - 1} + S_N \prod_{i=1}^{s-1} m_i \right] \\
&\le [(1+S_N)^d - 1 + (1+S_N)^d S_N] \left[\prod_{i=1}^{s-1} \frac{1}{p_i - 1} + S_N \prod_{i=1}^{s-1} m_i \right] \\
&= [(1+S_N)^{d+1} - 1] \left[\prod_{i=1}^{s-1} \frac{1}{p_i - 1} + S_N \prod_{i=1}^{s-1} m_i \right].
\end{aligned}
$$

Hence there exists a $g_{d+1} \in G_N$ such that

$$
\mathscr{T}_N((\boldsymbol{g}, g_{d+1})) \le [(1+S_N)^{d+1} - 1] \left[\prod_{i=1}^{s-1} \frac{1}{p_i - 1} + S_N \prod_{i=1}^{s-1} m_i \right].
$$

This completes the proof. □

We are now ready to state our CBC algorithm for generating the vector $\boldsymbol{g} \in G_N^d$.

Algorithm 1. *Let $N, d_* \in \mathbb{N}$. Set $\boldsymbol{g}_0 := ()$. For $d \in \{0, \dots, d_* - 1\}$, assume we have already constructed $\boldsymbol{g}_d \in G_N^d$. Then choose $g_{d+1} \in G_N$ which minimizes $\mathscr{T}_N((\boldsymbol{g}_d, \cdot))$.*

We now assess the complexity of the CBC construction. In the d-th step of the CBC construction, $g_1, \dots, g_{d-1}$ are already fixed and we need to compute $\mathscr{T}_N((g_1, \dots, g_{d-1}, g_d))$ for all g_d simultaneously. Referring to Eq. (5), we therefore need a quick way to compute the sum

$$\sum_{n=0}^{N-1} \widehat{\left(\frac{1}{r}\right)}(g_d n) \left(F_N(n) \prod_{i=1}^{d-1} \widehat{\left(\frac{1}{r}\right)}(g_i n) \right) \tag{6}$$

for all g_d. This is best done by viewing (6) as a multiplication of the matrix $\Omega := \left(\widehat{\left(\frac{1}{r}\right)}(gn)\right)_{g,n=0}^{N-1}$ with the transpose of the vector $x = \left(F_N(n) \prod_{i=1}^{d-1} \widehat{\left(\frac{1}{r}\right)}(g_i n)\right)_{n=0}^{N-1}$. This matrix-vector multiplication can be computed using $N \log N$ operations, by transforming the matrix into a circulant one so that multiplication with a vector can be done using fast Fourier transform. See the by now classical article by Nuyens and Cools [27] on fast CBC constructions for details.

We summarize in the following lemma.

Lemma 5. *The vector $F_N^{(d)}$ defined by*

$$F_N^{(d)}(g) := \sum_{n=0}^{N-1} \widehat{\left(\frac{1}{r}\right)}(gn) \left(F_N(n) \prod_{i=1}^{d-1} \widehat{\left(\frac{1}{r}\right)}(g_i n) \right)$$

can by computed recursively in d using $O(N((\log N)^{s-1} + d) \log N)$ operations.

Corollary 1. *The number of required operations in Algorithm 1 is of order*

$$O(N((\log N)^{s-1} + d) \log N).$$

The following theorem shows that Algorithm 1 yields $\boldsymbol{g}$ such that, if we use $\boldsymbol{g}$ as the generating vector of the lattice of our hybrid point set, we obtain low discrepancy.

Theorem 1. *Let $p_1, \dots, p_{s-1}$ be $s-1$ distinct prime numbers and let N be an odd prime number that is different from $p_1, \dots, p_{s-1}$. Let $(\boldsymbol{x}_n)_{n=0}^{N-1}$ be the s-dimensional Hammersley point set in bases $p_1, \dots, p_{s-1}$. Let $d_* \in \mathbb{N}$ and assume that $\boldsymbol{g} = (g_1, \dots, g_{d_*}) \in G_N^{d_*}$ is constructed according to Algorithm 1. Then for all $d \in \{1, \dots, d_*\}$ the point set*

$$\mathscr{P}_N = ((\boldsymbol{x}_n, \boldsymbol{y}_n))_{n=0}^{N-1} \in [0,1)^{s+d}$$

with $\boldsymbol{y}_n = \left\{\frac{n\boldsymbol{g}}{N}\right\}$ *for* $0 \le n \le N-1$ *where* $\boldsymbol{g} = (g_1, \ldots, g_d)$*, satisfies*

$$ND_N^*(\mathscr{P}_N) \le 1 + [2s + (1+S_N)^{d+1} - S_N] \prod_{i=1}^{s-1} m_i\, p_i + [(1+S_N)^d - 1] \prod_{i=1}^{s-1} \frac{p_i}{p_i - 1}.$$

Proof. The result can be shown by combining (2) and an inductive argument. Indeed, starting from (2), one only needs to show the existence of $\boldsymbol{g} \in G_N^{d_*}$ with small $\mathscr{T}_N(\boldsymbol{g})$. The latter step is done by induction on d_*, with Lemma 3 covering the one-dimensional case, and Lemma 4 covering the induction step. □

Remark 7. Recall that $m_i \le \frac{\log N}{\log p_i} + 1$ for all $1 \le i \le s-1$. Furthermore we have $S_N \le 2 \log N$. Therefore the discrepancy bound in Theorem 1 is of the form $D_N^*(\mathscr{P}_N) = O((\log N)^{s+d}/N)$. This corresponds to the existence result in [16, Theorem 1].

We report some concrete computations. Table 1 lists some results for $s = d = 2$. The number N is chosen as the N_0-th prime number and the bound on the discrepancy, labeled by "Bound (Theorem 1)" is that from Theorem 1. However, we also state the values of the preliminary bound in (2), labeled "Bound (2)" for comparison. Computing time is in seconds on an icore7 CPU.

Remark 8. It is not surprising that the bounds on the discrepancy reported in Table 1 cannot compete with the ones for non-hybrid sequences, which can be found in [21, Theorem 3.8] (Hammersley sequence) and [12, Eq. (20)] (lattice points). But to our best knowledge our bounds are the first non-trivial ones given for this kind of hybrid sequence. Note further that the order of convergence is as good as that of non-hybrid sequences.

Table 1 Discrepancy bounds and good generating vectors for $s = d = 2$.

N_0	$N = \pi(N_0)$	Bound (Theorem 1)	g_1	g_2	Bound (2)	Time
25,000	287,117	2.302	39,578	893	2.065	14
50,000	611,953	1.348	222,724	283,785	1.211	32
75,000	951,161	0.954	202,628	38,908	0.856	51
100,000	1,299,709	0.782	1,196,035	448,030	0.703	71
125,000	1,655,131	0.646	531,187	81,506	0.579	91
150,000	2,015,177	0.552	1,421,738	994,192	0.496	112
175,000	2,381,147	0.506	1,847,221	90,990	0.454	136
200,000	2,750,159	0.451	1,665,684	312,176	0.404	160
225,000	3,122,321	0.407	673,590	1,982,517	0.364	191
250,000	3,497,861	0.371	2,430,895	1,686,792	0.332	219

4 Concluding Remarks

In this paper, we have discussed the explicit construction of hybrid point sets obtained by mixing Hammersley point sets and lattice point sets. Algorithm 1 is a common CBC approach for generating the lattice part of our hybrid set. This algorithm yields point sets satisfying a discrepancy bound as good as the bound in the existence result of [16]. We have therefore solved the open problem outlined in the final section of [16] and taken a further step towards obtaining explicit versions of hybrid point sets of high quality.

Finally, we would like to remark that the point sets under consideration in this paper have recently also been analyzed with respect to a different measure of uniformity, namely the diaphony (cf. [6]).

Acknowledgements P. Kritzer gratefully acknowledges the support of the Austrian Science Fund (FWF), Project P23389-N18. G. Leobacher gratefully acknowledges the support of the Austrian Science Fund (FWF), Project P21196. F. Pillichshammer is partially supported by the Austrian Science Fund (FWF), Project S9609, that is part of the Austrian National Research Network "Analytic Combinatorics and Probabilistic Number Theory".

The authors would like to thank Josef Dick, Frances Y. Kuo and Ian H. Sloan for their hospitality during the authors' stay at the School of Mathematics and Statistics at the University of New South Wales in February 2012, where parts of this paper were written. Furthermore, the authors gratefully acknowledge the support of the Australian Research Council.

References

1. Atanassov, E.I.: On the discrepancy of the Halton sequences. Math. Balkanica (New Series) **18**, 15–32 (2004)
2. Dick, J., Pillichshammer, F.: Digital Nets and Sequences. Discrepancy Theory and Quasi-Monte Carlo Integration. Cambridge University Press, Cambridge (2010)
3. Drmota, M., Tichy, R.F.: Sequences, Discrepancies and Applications. Lecture Notes in Mathematics 1651. Springer, Berlin (1997)
4. Gnewuch, M., Srivastav, A., Winzen, C.: Finding optimal volume subintervals with k points and calculating the star discrepancy are NP-hard problems. J. Complexity **25**, 115–127 (2009)
5. Hellekalek, P.: Hybrid function systems in the theory of uniform distribution of sequences. In: Plaskota, L., Woźniakowski, H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2010, pp. 435–450. Springer, Berlin/Heidelberg (2012)
6. Hellekalek, P., Kritzer, P.: The diaphony of some finite hybrid point sets. Acta Arith. **156**, 257–282 (2012)
7. Hlawka, E.: Zur angenäherten Berechnung mehrfacher Integrale. Monatsh. Math. **66**, 140–151 (1962) (in German)
8. Hofer, R., Kritzer, P.: On hybrid sequences built of Niederreiter-Halton sequences and Kronecker sequences. Bull. Aust. Math. Soc. **84**, 238–254 (2011)
9. Hofer, R., Kritzer, P., Larcher, G., Pillichshammer, F.: Distribution properties of generalized van der Corput-Halton sequences and their subsequences. Int. J. Number Theory **5**, 719–746 (2009)
10. Hofer, R., Larcher, G.: Metrical results on the discrepancy of Halton-Kronecker sequences. Mathematische Zeitschrift **271**, 1–11 (2012)

11. Joe, S.: Component by component construction of rank-1 lattice rules having $\mathcal{O}(n^{-1}(\ln(n))^d)$ star discrepancy. In: Niederreiter, H. (ed.) Monte Carlo and Quasi-Monte Carlo Methods 2002, pp. 293–298. Springer, Berlin/Heidelberg (2004)
12. Joe, S.: An intermediate bound on the star discrepancy. In: Plaskota, L., Woźniakowski, H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2010, pp. 451–470. Springer, Berlin/Heidelberg (2012)
13. Keller, A.: Quasi-Monte Carlo image synthesis in a Nutshell. In: Dick, J., Kuo, F.Y., Peters, G.W., Sloan, I.H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2012, this volume 213–249. Springer, Berlin/Heidelberg (2013)
14. Korobov, N.M.: Approximate evaluation of repeated integrals. Doklady Akademii Nauk, SSSR **124**, 1207–1210 (1959) (in Russian)
15. Korobov, N.M.: Number-Theoretic Methods in Approximate Analysis. Fizmatgiz, Moscow (1963) (in Russian)
16. Kritzer, P.: On an example of finite hybrid quasi-Monte Carlo Point Sets. Monatsh. Math. **168**, 443–459 (2012)
17. Kuipers, L., Niederreiter, H.: Uniform Distribution of Sequences. Wiley, New York (1974) (Reprint, Dover Publications, Mineola 2006)
18. Kuo, F.Y.: Component-by-component constructions achieve the optimal rate of convergence for multivariate integration in weighted Korobov and Sobolev spaces. J. Complexity **19**, 301–320 (2003)
19. Lemieux, C.: Monte Carlo and Quasi-Monte Carlo Sampling. Springer Series in Statistics. Springer, New York (2009)
20. Niederreiter, H.: Existence of good lattice points in the sense of Hlawka. Monatsh. Math. **86**, 203–219 (1978)
21. Niederreiter, H.: Random Number Generation and Quasi-Monte Carlo Methods. No. 63 in CBMS-NSF Series in Applied Mathematics. SIAM, Philadelphia (1992)
22. Niederreiter, H.: On the discrepancy of some hybrid sequences. Acta Arith. **138**, 373–398 (2009)
23. Niederreiter, H.: Further discrepancy bounds and an Erdős-Turán-Koksma inequality for hybrid sequences. Monatsh. Math. **161**, 193–222 (2010)
24. Niederreiter, H.: Discrepancy bounds for hybrid sequences involving matrix-method pseudo-random vectors. Publ. Math. Debrecen **79**, 589–603 (2011)
25. Niederreiter, H.: Improved discrepancy bounds for hybrid sequences involving Halton sequences. Acta Arith. **55**, 71–84 (2012)
26. Niederreiter, H., Winterhof, A.: Discrepancy bounds for hybrid sequences involving digital explicit inversive pseudorandom numbers. Unif. Distrib. Theory **6**, 33–56 (2011)
27. Nuyens, D., Cools, R.: Fast algorithms for component-by-component construction of rank-1 lattice rules in shift-invariant reproducing Kernel Hilbert spaces. Math. Comp. **75**, 903–920 (2006)
28. Owen, A.B.: Latin supercube sampling for very high dimensional simulations. ACM Trans. Model. Comput. Simul. **8**, 71–102 (1998)
29. Sloan, I.H., Joe, S.: Lattice Methods for Multiple Integration. Oxford University Press, New York/Oxford (1994)
30. Sloan, I.H., Kuo, F.Y., Joe, S.: On the step-by-step construction of quasi-Monte Carlo integration rules that achieve strong tractability error bounds in weighted Sobolev spaces. Math. Comp. **71**, 1609–1640 (2002)
31. Sloan, I.H., Reztsov, A.V.: Component-by-component construction of good lattice rules. Math. Comp. **71**, 263–273 (2002)
32. Spanier, J.: Quasi-Monte Carlo methods for particle transport problems. In: Niederreiter, H., Shiue, P.J.-S. (eds.) Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing. Lecture Notes in Statistics, vol. 106, pp. 121–148. Springer, New York (1995)

A QMC-Spectral Method for Elliptic PDEs with Random Coefficients on the Unit Sphere

Quoc Thong Le Gia

Abstract We present a quasi-Monte Carlo spectral method for a class of elliptic partial differential equations (PDEs) with random coefficients defined on the unit sphere. The random coefficients are parametrised by the Karhunen-Loève expansion, while the exact solution is approximated by the spherical harmonics. The expectation of the solution is approximated by a quasi-Monte Carlo integration rule. A method for obtaining error estimates between the exact and the approximate solution is also proposed. Some numerical experiments are provided in the last section.

1 Introduction

Let S be the unit sphere in $\mathbb{R}^3$, i.e. $S = \{\mathbf{x} \in \mathbb{R}^3 : |\mathbf{x}| = 1\}$. Let (Ω, Σ, P) be a probability space and assume that $a(\cdot, \omega) : \Omega \to L^\infty(S)$ is a P-measurable map. We assume

$$a \in L^2(\Omega, dP; L^\infty(S)), \tag{1}$$

which renders the mean and variance of the random field a as elements of $L^\infty(S)$ and respectively of $L^\infty(S \times S)$, finite.

Given a random diffusion coefficient $a(\mathbf{x}, \omega)$, a prediction of the concentration $u(\mathbf{x}, \omega)$ for $\mathbf{x} \in S$ requires a solution of a stochastic differential equation such as

$$\begin{cases} -\text{Div}\,(a(\mathbf{x}, \omega)\text{Grad}\, u(\mathbf{x}, \omega)) &= f(\mathbf{x}) \quad \text{on } S, \\ \int_S u(\mathbf{x}, \omega) dS(\mathbf{x}) &= 0, \quad \omega \in \Omega, \end{cases} \tag{2}$$

Q.T. Le Gia (✉)
School of Mathematics and Statistics, University of New South Wales, Sydney, NSW, Australia
e-mail: qlegia@unsw.edu.au

J. Dick et al. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2012*,
Springer Proceedings in Mathematics & Statistics 65,
DOI 10.1007/978-3-642-41095-6_26, © Springer-Verlag Berlin Heidelberg 2013

where Div and Grad are the surface divergence and surface gradient on the sphere respectively.

These equations can be used to describe a diffusion on the sphere in which the diffusivity contains random noises coming from measurements. This situation can occur when turbulent diffusivity of the atmosphere is inferred from radar measurements [11].

To ensure (2) has a unique solution, we assume further that $a \in L^\infty(S \times \Omega)$ is strictly positive, with lower and upper bound $a_{\min} > 0$ and $a_{\max} < \infty$ respectively, i.e.

$$a_{\min} \leq \text{ess inf } a(\mathbf{x}, \omega) \text{ and ess sup } a(\mathbf{x}, \omega) \leq a_{\max} \quad \text{P-a.s,} \tag{3}$$

where the essential infimum and supremum are taken with respect to the Lebesgue measure in S.

In this work, we propose an approximation scheme for (2) using the Karhunen-Loève expansion on the unit sphere of the random coefficient. A similar approximation model for elliptic PDEs with random coefficients on bounded domains in $\mathbb{R}^n$ has been proposed recently [4].

The paper is organised as follows. In Sect. 2, we review background materials on spherical harmonics, Karhunen-Loève expansion on the unit sphere, quasi Monte Carlo method using lattice rules. In Sect. 3, we describe the parametric variational formulation of the PDE and discuss the regularity of the solution, QMC integration for the exact solution of the PDE. The spectral method on the sphere and the combined error estimates are presented in the last two sections of the paper.

2 Preliminaries

2.1 Spherical Harmonics

Spherical harmonics are the restriction to S of homogeneous polynomials Y in $\mathbb{R}^3$ which satisfy $\Delta Y = 0$, where Δ is the Laplacian operator in $\mathbb{R}^3$. The space of all spherical harmonics of degree ℓ on S, denoted by $\mathscr{H}_\ell$, has an orthonormal basis

$$\{Y_{\ell,m} : m = -\ell, \ldots, \ell\}.$$

The space of spherical harmonics of degree $\leq L$ will be denoted by $\mathscr{P}_L := \bigoplus_{\ell=0}^{L} \mathscr{H}_\ell$; it has dimension $(\ell+1)^2$. Every function $f \in L^2(S)$ can be expanded in terms of spherical harmonics,

$$f = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} \hat{f}_{\ell,m} Y_{\ell,m}, \quad \hat{f}_{\ell,m} = \int_S f \overline{Y_{\ell,m}} dS,$$

where dS is the surface measure of the unit sphere. The following formulas are the addition theorems for spherical harmonics [10, p. 223].

$$\sum_{m=-\ell}^{\ell} Y_{\ell,m}(\mathbf{x})\overline{Y_{\ell,m}}(\mathbf{y}) = \frac{2\ell+1}{4\pi} P_\ell(\mathbf{x}\cdot\mathbf{y}), \tag{4}$$

$$\sum_{m=-\ell}^{\ell} \text{Grad}\, Y_{\ell,m}(\mathbf{x}) \cdot \overline{\text{Grad}\, Y_{\ell,m}}(\mathbf{y}) = \frac{(2\ell+1)\ell(\ell+1)}{4\pi} P_\ell(\mathbf{x}\cdot\mathbf{y}), \tag{5}$$

where P_ℓ is the Legendre polynomial of degree ℓ normalised so that $P_\ell(1) = 1$.

Lemma 1. *Let Y_ℓ be a spherical harmonic of degree ℓ. Then*

$$|Y_\ell(\mathbf{x})| \le \sqrt{\frac{2\ell+1}{4\pi}} \left(\int_S |Y_\ell(\mathbf{x})|^2 dS \right)^{1/2} \tag{6}$$

and

$$|\text{Grad}\, Y_\ell(\mathbf{x})| \le \sqrt{\frac{2\ell+1}{4\pi}} \left(\int_S |\text{Grad}\, Y_\ell(\mathbf{x})|^2 dS \right)^{1/2} \tag{7}$$

Proof. Inequality (6) is a result of [6, Lemma 8, p. 14]. In order to prove inequality (7), suppose $Y_\ell(\mathbf{x}) = \sum_{m=-\ell}^{\ell} d_m Y_{\ell,m}(\mathbf{x})$, where $d_m = (Y_\ell, Y_{\ell,m})_{L^2(S)}$. We use the orthogonality [10, p. 227] $\int_S \text{Grad}\, Y_{\ell,m} \cdot \overline{\text{Grad}\, Y_{\ell',m'}} dS = \ell(\ell+1)\delta_{\ell,\ell'}\delta_{m,m'}$, to obtain

$$\int_S |\text{Grad}\, Y_\ell|^2 dS = \ell(\ell+1) \sum_{m=-\ell}^{\ell} (d_m)^2. \tag{8}$$

Applying Cauchy-Schwarz's inequality and (5) we have

$$|\text{Grad}\, Y_\ell(\mathbf{x})|^2 \le \sum_{m=-\ell}^{\ell} (d_m)^2 \sum_{m=-\ell}^{\ell} |\text{Grad}\, Y_{\ell,m}(\mathbf{x})|^2 = \frac{(2\ell+1)\ell(\ell+1)}{4\pi} \sum_{m=-\ell}^{\ell} (d_m)^2.$$

Combining this with (8), we obtain (7). □

2.2 Karhunen-Loève Expansion on the Unit Sphere

To define the Karhunen-Loève (KL) expansion of $a(\mathbf{x}, \omega)$, we assume the mean field and two-point correlation of $a(\mathbf{x}, \omega)$ are known, i.e. that

$$\bar{a}(\mathbf{x}) := \int_\Omega a(\mathbf{x}, \omega) dP(\omega) \text{ and } C_a(\mathbf{x}, \mathbf{y}) := \int_\Omega a(\mathbf{x}, \omega) a(\mathbf{y}, \omega) dP(\omega) \tag{9}$$

are known. An equivalent assumption is that the mean field $\bar{a}$ and its covariance V_a are known, since by definition,

$$V_a(\mathbf{x}, \mathbf{y}) = C_a(\mathbf{x}, \mathbf{y}) - \bar{a}(\mathbf{x})\bar{a}(\mathbf{y}). \tag{10}$$

The 2-point correlation of $a(\mathbf{x}, \omega)$ is well-defined and belongs to $L^\infty(S \times S)$ due to (1). Associated with V_a we can define a compact, self-adjoint operator $\mathscr{V}_a : L^2(S) \to L^2(S)$ by

$$(\mathscr{V}_a u)(\mathbf{x}) = \int_S V_a(\mathbf{x}, \mathbf{y}) u(\mathbf{y}) dS(\mathbf{y}), \tag{11}$$

where dS is the surface measure of the unit sphere S.

A covariance kernel $V_a(\mathbf{x}, \mathbf{y}) \in L^2(S \times S)$ given by (10) is said to be *admissible* if it is symmetric and positive definite in the sense that

$$\sum_{k=1}^{n} \sum_{j=1}^{n} a_k V_a(\mathbf{x}_k, \mathbf{x}_j) \overline{a}_j \geq 0, \quad \forall \mathbf{x}_j, \mathbf{x}_k \in S, \quad a_k, a_j \in \mathbb{C}.$$

Using a characterisation of positive definite functions on the unit sphere by Schoenberg [8, Theorem 1], we conclude that an admissible covariance kernel $V_a(\mathbf{x}, \mathbf{y})$ admits the following expansion into spherical harmonics

$$V_a(\mathbf{x}, \mathbf{y}) = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} \hat{v}_\ell Y_{\ell,m}(\mathbf{x}) \overline{Y_{\ell,m}}(\mathbf{y}), \quad \mathbf{x}, \mathbf{y} \in S. \tag{12}$$

where

$$\sum_{\ell=0}^{\infty} (2\ell + 1)\hat{v}_\ell < \infty, \quad \hat{v}_\ell > 0, \quad \ell = 0, 1, \ldots \tag{13}$$

From the addition theorem for spherical harmonics (4),

$$V_a(\mathbf{x}, \mathbf{y}) = \sum_{\ell=0}^{\infty} \frac{2\ell + 1}{4\pi} \hat{v}_\ell P_\ell(\mathbf{x} \cdot \mathbf{y}).$$

Therefore, in view of condition (13), the series (12) converge uniformly by Weierstrass M-test.

Using the orthogonality of the spherical harmonics, we have

$$\int_S V_a(\mathbf{x}, \mathbf{y}) Y_{\ell,m}(\mathbf{y}) dS(\mathbf{y}) = \hat{v}_\ell Y_{\ell,m}(\mathbf{x}).$$

Therefore $\{(\hat{v}_\ell, Y_{\ell,m}) : \ell = 0, 1, \ldots; m = -\ell, \ldots, \ell\}$ is the sequence of eigenpairs of the integral operator $\mathscr{V}_a$.

Using the Loève representation theorem, since S is a compact set and the spherical harmonics form an orthonormal basis of $L^2(S)$, the random field (1) takes the form

$$a(\mathbf{x},\omega) = \bar{a}(\mathbf{x}) + \sum_{\ell=0}^{\infty}\sum_{m=-\ell}^{\ell} \sqrt{\hat{v}_\ell}\, Y_{\ell,m}(\mathbf{x}) X_{\ell,m}(\omega), \tag{14}$$

where $X_{\ell,m}(\omega)$ are centred at 0, pairwise uncorrelated random variables on probability spaces $(\Omega_{\ell,m}, \Sigma_{\ell,m}, P_{\ell,m})$ for $\ell = 0, 1, 2, \ldots; m = -\ell, \ldots, \ell$.

We now assume that

$$\bar{a} \in W^{1,\infty}(S), \quad \sum_{\ell=1}^{\infty} \sqrt{\ell(\ell+1)(2\ell+1)}\sqrt{\hat{v}_\ell} < \infty, \tag{15}$$

where $\|v\|_{W^{1,\infty}(S)} = \max\{\|v\|_{L^\infty(S)}, \|\mathrm{Grad}\, v\|_{L^\infty(S)}\}$.

Since $Y_{\ell,m}$ is an element of an orthonormal basis, we deduce from (6),

$$|Y_{\ell,m}(\mathbf{x})| \le \sqrt{\frac{2\ell+1}{4\pi}} \quad \forall \mathbf{x} \in S. \tag{16}$$

From assumption (15) and estimate (16) we obtain

$$\sum_{\ell=0}^{\infty}\sum_{m=-\ell}^{\ell} \sqrt{\hat{v}_\ell}\, \|Y_{\ell,m}(\mathbf{x})\|_{L^\infty(S)} < \infty. \tag{17}$$

We sometimes make a stronger assumption that

$$\sum_{\ell=0}^{\infty}\sum_{m=-\ell}^{\ell} \sqrt{(\hat{v}_\ell)^p}\, \|Y_{\ell,m}(\mathbf{x})\|^p_{L^\infty(S)} < \infty, \quad 0 < p < 1. \tag{18}$$

Using the orthogonality of $\mathrm{Grad}\, Y_{\ell,m}$, we deduce from (7) that

$$|\mathrm{Grad}\, Y_{\ell,m}(\mathbf{x})| \le \sqrt{\frac{\ell(2\ell+1)(\ell+1)}{4\pi}}. \tag{19}$$

So from (19) and (15) we deduce that

$$\sum_{\ell=0}^{\infty}\sum_{m=-\ell}^{\ell} \sqrt{\hat{v}_\ell}\, \|Y_{\ell,m}(\mathbf{x})\|_{W^{1,\infty}(S)} < \infty, \tag{20}$$

which will guarantee the convergence of the approximate solution.

Following [1], we assume that the diffusion coefficient $a(\mathbf{x},\omega)$ satisfies (3), the covariance kernel V_a is admissible and $a(\mathbf{x},\omega)$ admits a Karhunen-Loève expansion

(14). Furthermore, we make the following assumptions on the random variables $X_{\ell,m}$ in the KL-expansion (14) of the random coefficient.

Assumption 1. *(i) The family* $\{X_{\ell,m} : \ell = 0,1,2,\ldots; m = -\ell,\ldots,\ell\}$ *is independent.*

(ii) The KL-expansion (14) *of the random coefficient is finite, i.e. there exists* $\overline{M} < \infty$ *such that* $X_{\ell,m} = 0$ *for all* $\ell > \overline{M}$.

(iii) Each $X_{\ell,m}(\omega)$ *in* (14) *is associated with a probability space* $(\Omega_{\ell,m}, \Sigma_{\ell,m}, P_{\ell,m})$ *with the following properties:*

(a) The range of $X_{\ell,m}$, $U_{\ell,m} := Range(X_{\ell,m}) \subset \mathbb{R}$, *is assumed to be compact,*

(b) The probability measure $P_{\ell,m}$ *admits a probability density function* $\rho_{\ell,m} : U_{\ell,m} \to [0,\infty)$ *such that* $dP_{\ell,m}(\omega) = \rho_{\ell,m}(y_{\ell,m})dy_{\ell,m}$, $y_{\ell,m} \in U_{\ell,m}$, *and*

(c) The sigma algebras $\Sigma_{\ell,m}$ *are subsets of the Borel sets of the interval* $U_{\ell,m}$, *i.e.* $\Sigma_{\ell,m} \subset \mathscr{B}(U_{\ell,m})$

In the sequel, we let $\Lambda := \{\ell = 0,1,2,\ldots,\overline{M}; m = -\ell,\ldots,\ell\}$.

Assumption 1 (ii) is made so that we can represent the measure P on the space of input data as an $\overline{M}-$fold product measure and to avoid technical issues related to countable product measures on the space $L^\infty(S)$ of input data. We have

$$\Sigma = \bigotimes_{(\ell,m)\in\Lambda} \Sigma_{\ell,m}, \quad dP = \bigotimes_{(\ell,m)\in\Lambda} dP_{\ell,m}, \quad U = \bigotimes_{(\ell,m)\in\Lambda} U_{\ell,m}.$$

While the double index (ℓ,m) follows the convention of spherical harmonics, it is inconvenient in subsequent analysis. We introduce a single index via the map

$$j(\ell,m) = \ell(\ell+1) + 1 + m, \quad \ell = 0,1,2,\ldots; m = -\ell,\ldots,\ell.$$

In subsequent sections, we assume that $X_{\ell,m}(\omega) = \omega_{\ell,m}$ uniformly distributed and $U = \left[-\frac{1}{2},\frac{1}{2}\right]^s$, for some $s \leq \overline{M}$.

2.3 Quasi-Monte Carlo Integration in Weighted Spaces

Let s be a positive integer, we consider integrals over the s-dimensional cube $\left[-\frac{1}{2},\frac{1}{2}\right]^s$ of the form

$$\mathscr{I}_s(F) := \int_{\left[-\frac{1}{2},\frac{1}{2}\right]^s} F(\mathbf{y})d\mathbf{y}.$$

An N-point QMC approximation to this integral is an equal weight quadrature of the form

$$Q_{s,N}(F) := \frac{1}{N}\sum_{i=1}^{N} F(\mathbf{y}^{(i)}),$$

with carefully chosen points $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(N)} \in [-\frac{1}{2}, \frac{1}{2}]^s$. We shall assume that the integrand F belongs to weighted and anchored Sobolev space $\mathscr{W}_{s,\gamma}$, which is a Hilbert space equipped with the norm

$$\|F\|_{\mathscr{W}_{s,\gamma}} := \left(\sum_{\mathfrak{u} \subseteq \{1:s\}} \gamma_{\mathfrak{u}}^{-1} \int_{[-\frac{1}{2},\frac{1}{2}]^{|\mathfrak{u}|}} \left| \int_{[-\frac{1}{2},\frac{1}{2}]^{s-|\mathfrak{u}|}} \frac{\partial^{|\mathfrak{u}|} F}{\partial \mathbf{y}_{\mathfrak{u}}} (\mathbf{y}_{\mathfrak{u}}; \mathbf{y}_{\{1:s\}\setminus\mathfrak{u}}; 0) d\mathbf{y}_{\{1:s\}\setminus\mathfrak{u}} \right|^2 d\mathbf{y}_{\mathfrak{u}} \right)^{1/2} \tag{21}$$

The norm of $\mathscr{I}_s$ as a linear functional on the function space $\mathscr{W}_{s,\gamma}$ is, from [9],

$$\|\mathscr{I}_s\| := \sup_{\|F\|_{\mathscr{W}_{s,\gamma}} \le 1} |\mathscr{I}_s(F)| = \left(\sum_{\mathfrak{u} \subseteq \{1:s\}} \gamma_{\mathfrak{u}} \left(\frac{1}{12} \right)^{|\mathfrak{u}|} \right)^{1/2}.$$

We shall assume that we have a sequence of positive weights $\gamma = (\gamma_{\mathfrak{u}})_{|\mathfrak{u}|<\infty}$ satisfying

$$\sum_{|\mathfrak{u}| \le \infty} \gamma_{\mathfrak{u}} \left(\frac{1}{12} \right)^{|\mathfrak{u}|} < \infty. \tag{22}$$

For the moment, we focus on a family of QMC rules known as "shifted rank-1 lattice rules", whose quadrature points are given by the following formula

$$\mathbf{y}^{(i)} = \mathrm{frac}\left(\frac{i\mathbf{z}}{N} + \boldsymbol{\Delta} \right) - \left(\frac{1}{2}, \ldots, \frac{1}{2} \right), \quad i = 1, \ldots, N,$$

where $\mathbf{z} \in \mathbb{Z}^s$ is known as the *generating vector*, $\boldsymbol{\Delta} \in [0,1]^s$ is the *shift*, and $\mathrm{frac}(\cdot)$ means to take the fractional part of each component in the vector.

An application of $Q_{s,N}$ with a realization for a draw of the shift $\boldsymbol{\Delta}$ will be denoted by $Q_{s,N}(\cdot; \boldsymbol{\Delta})$.

Theorem 1 ([4, Theorem 2.1]). *Let $s, N \in \mathbb{N}$ be given, and assume $F \in \mathscr{W}_{s,\gamma}$ for a particular choice of weights γ. Then a randomly shifted lattice rule can be constructed using a component-by-component algorithm such that the root-mean-square error for approximating the s-dimensional integral $\mathscr{I}_s(F)$ satisfies, for all $\lambda \in (1/2, 1]$,*

$$\sqrt{\mathbb{E}[|\mathscr{I}_s(F) - Q_{s,N}(F;\cdot)|^2]} \tag{23}$$

$$\le \left(\sum_{|\mathfrak{u}|<\infty} \gamma_{\mathfrak{u}}^{\lambda} \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^{\lambda}} + \frac{1}{12^{\lambda}} \right)^{|\mathfrak{u}|} \right)^{1/(2\lambda)} (N-1)^{-1/(2\lambda)} \|F\|_{\mathscr{W}_{s,\gamma}}, \tag{24}$$

where $\mathbb{E}[\cdot]$ *denotes the expectation with respect to the random shift* $\boldsymbol{\Delta}$ *which is uniformly distributed over* $[0,1]^s$.

3 Parametric Variational Formulation

3.1 Function Spaces

Define the following Sobolev space

$$H^1(S) := \{u \in L^2(S) : \operatorname{Grad} u \in \mathbb{L}^2(S)\}.$$

That is, $H^1(S)$ consists of functions $u \in L^2(S)$, whose weak gradient $\operatorname{Grad} u$ exists and is in the space $\mathbb{L}^2(S)$, which contains all vector fields $\mathbf{u}$ so that $\int_S \mathbf{u} \cdot \mathbf{u} < \infty$.

Let V be a subspace of $H^1(S)$ which contains all functions with zero mean over S, i.e.

$$V := \left\{ u \in H^1(S) : \int_S u = 0 \right\}.$$

It can be shown that V is a Hilbert space with the following inner product and norm

$$(u,v)_V = (\operatorname{Grad} u, \operatorname{Grad} v)_{\mathbb{L}^2(S)}, \quad \|u\|_V = \|\operatorname{Grad} u\|_{\mathbb{L}^2(S)}.$$

Let V^* be the dual space of V with respect to the $L^2(S)$ inner product $(\cdot,\cdot)$, i.e., the space of all continuous linear functionals defined on V. Since $V \subset H^1(S)$, we have $H^{-1}(S) \subset V^*$. We also consider the following function space

$$Z := \{v \in V : \Delta^* v \in L^2(S)\},$$

where Δ^* is the Laplace-Beltrami operator on S. The space $Z \subset V$ is a closed subspace which, when equipped with the norm

$$\|v\|_Z := \left(\|v\|^2_{L^2(S)} + \|\Delta^* v\|^2_{L^2(S)} \right)^{1/2},$$

is a Hilbert space. The space Z is a subspace of $H^2(S)$, see [5] for definitions of Sobolev spaces on manifolds based on the Laplace–Beltrami operator.

We also define the weighted spaces $\mathscr{W}_{s,\gamma}(U;V)$, which are the Bochner versions of the weighted spaces $\mathscr{W}_{s,\gamma}$, with the norm

$$\|u\|_{\mathscr{W}_{s,\gamma}(U;V)} :=$$

$$\left(\sum_{|\mathrm{u}|\subseteq\{1:s\}} \gamma_{\mathrm{u}}^{-1} \int_{[-\frac{1}{2},\frac{1}{2}]^{|\mathrm{u}|}} \left\| \int_{[-\frac{1}{2},\frac{1}{2}]^{s-|\mathrm{u}|}} \frac{\partial^{|\mathrm{u}|} u}{\partial \mathbf{y}_{\mathrm{u}}}(\cdot, (\mathbf{y}_{\mathrm{u}}; \mathbf{y}_{\{1:s\}\setminus\mathrm{u}}; 0)) d\mathbf{y}_{\{1:s\}\setminus\mathrm{u}} \right\|_V^2 d\mathbf{y}_{\mathrm{u}} \right)^{1/2}. \tag{25}$$

In this paper, we wish to compute

$$\int_U F(\mathbf{y}) d\mathbf{y}, \quad \text{with} \quad F(\mathbf{y}) = G(u(\cdot, \mathbf{y})), \quad G \in V^*. \tag{26}$$

Then for $s \geq 1$ and for $u \in \mathscr{W}_{s,\gamma}(U; V)$, using (21) and

$$\frac{\partial^{|\mathrm{u}|} F}{\partial \mathbf{y}_{\mathrm{u}}}(\mathbf{y}) = G\left(\frac{\partial^{|\mathrm{u}|} u}{\partial \mathbf{y}_{\mathrm{u}}}(\cdot, \mathbf{y}) \right),$$

we have

$$\|F\|_{\mathscr{W}_\gamma} \leq \|G(\cdot)\|_{V^*} \|u\|_{\mathscr{W}_{s,\gamma}(U;V)} < \infty. \tag{27}$$

3.2 The Parametric Variational Formulation

The weak formulation of (2) is given by

$$E\left[\int_S a(\mathbf{x}, \omega) \mathrm{Grad}\, u(\mathbf{x}, \omega) \cdot \mathrm{Grad}\, v(\mathbf{x}, \omega) dS(\mathbf{x}) \right] = E\left[\int_S f(\mathbf{x}) v(\mathbf{x}, \omega) dS(\mathbf{x}) \right],$$

where E denotes the expectation with respect to the random variable ω.

As a consequence of the independence in Assumption 1, the multivariate probability density on U is given by

$$\rho(\mathbf{y}) := \Pi_{(\ell,m)\in\Lambda} \rho_{\ell,m}(y_{\ell,m}).$$

We substitute $X_{\ell,m}(\omega)$ by $y_{\ell,m}$ and equip the range U of the vector $\mathbf{y}$ with the product measure $dP(\omega) = \bigotimes \rho_{\ell,m}(y_{\ell,m}) dy_{\ell,m}$. Here we assume that the random variable $X_{\ell,m}$ and the random numbers $y_{\ell,m}$ have the same probability distribution. Changing measure from $dP(\omega)$ to $\Pi_{(\ell,m)\in\Lambda} \rho_{\ell,m}(y_{\ell,m}) dy_{\ell,m}$, problem (2) is equivalent to the parametric, deterministic problem:

$$\text{Find } u(\cdot, \mathbf{y}) \in V \text{ such that } -\mathrm{Div}\, a(\mathbf{x}, \mathbf{y}) \mathrm{Grad}\, u(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}), \quad \mathbf{x} \in S, \quad \mathbf{y} \in U. \tag{28}$$

For a fixed $\mathbf{y} \in U$, the parametric variational formulation of the PDE (28) is obtained by multiplying it with a test function and integrating by part: find $u(\cdot, \mathbf{y}) \in V$ such that

$$\int_S a(\mathbf{x},\mathbf{y})\text{Grad}\, u(\mathbf{x},\mathbf{y})\cdot \text{Grad}\, v(\mathbf{x},\mathbf{y})dS(\mathbf{x}) = \int_S f(\mathbf{x})v(\mathbf{x})dS \quad \forall v \in V.$$

Let us define the parametric bilinear form $b(\mathbf{y}; v, w)$ by

$$b(\mathbf{y}; v, w) := \int_S a(\cdot,\mathbf{y})\text{Grad}\, v \cdot \text{Grad}\, w dS \quad v, w \in V. \tag{29}$$

In view of (3) the bilinear form $b(\cdot,\cdot)$ is continuous and coercive on $V \times V$, i.e., for all $\mathbf{y} \in U$ and all $v, w \in V$ we have

$$b(\mathbf{y}; v, v) \geq a_{\min}\|v\|_V^2 \quad \text{and} \quad |b(\mathbf{y}; v, w)| \leq a_{\max}\|v\|_V\|w\|_V.$$

The variational form now reads:

$$\text{Find } u(\cdot,\mathbf{y}) \in V \text{ such that } b(\mathbf{y}; u(\cdot,\mathbf{y}), v) = \langle f, v\rangle \quad \forall v \in V. \tag{30}$$

By the Lax-Milgram Lemma we conclude that for every $f \in V^*$, there exists a unique solution to the parametric weak problem (30).

Theorem 2. *Under Assumptions (3) and (17), for every $f \in V^*$ and every $\mathbf{y} \in U$, there exists a unique solution $u(\cdot,\mathbf{y})$ of the parametric weak problem (30), which satisfies*

$$\|u(\cdot,\mathbf{y})\|_V \leq \frac{1}{a_{\min}}\|f\|_{V^*}.$$

3.3 Regularity of the PDE Solution

Assume that $f \in L^2(S)$, we want to obtain a bound on the Z norm of $u(\cdot,\mathbf{y})$ for each value of the parameter $\mathbf{y}$.

Theorem 3. *Under Assumptions (3) and (15), there exists a constant $C > 0$ such that for every $f \in L^2(S)$ and every $\mathbf{y} \in U$, the solution $u(\cdot,\mathbf{y}) \in V$ of the parametric weak problem (30) satisfies*

$$\|u(\cdot,\mathbf{y})\|_Z \leq C\|f\|_{L^2(S)}. \tag{31}$$

Proof. Using Assumption (15), which implies (20), for every $\mathbf{y} \in U$, we have,

$$\|a(\cdot,\mathbf{y})\|_{W^{1,\infty}(S)} \leq \|\bar{a}\|_{W^{1,\infty}(S)} + \frac{1}{2}\sum_{\ell=0}^{\infty}\sum_{m=-\ell}^{\ell}\sqrt{\hat{v}_\ell}\|Y_{\ell,m}\|_{W^{1,\infty}(S)} < \infty.$$

We apply the identity

$$\mathrm{Div}\,(\alpha(\mathbf{x})\mathrm{Grad}\,w(\mathbf{x})) = \alpha(\mathbf{x})\Delta^* w(\mathbf{x}) + \mathrm{Grad}\,\alpha(\mathbf{x}) \cdot \mathrm{Grad}\,w(\mathbf{x}),$$

to (2) in order to obtain

$$-a(\cdot,\mathbf{y})\Delta^* u(\cdot,\mathbf{y}) = \mathrm{Grad}\,a(\cdot,\mathbf{y}) \cdot \mathrm{Grad}\,u(\cdot,\mathbf{y}) + f(\cdot) \quad \text{on S},$$
$$\int_S u(\mathbf{x},\mathbf{y})dS(\mathbf{x}) = 0.$$

This implies that for every $\mathbf{y} \in U$ there holds

$$a_{\min}\|\Delta^* u(\cdot,\mathbf{y})\|_{L^2(S)} \le \|a(\cdot,\mathbf{y})\|_{W^{1,\infty}(S)}\|u(\cdot,\mathbf{y})\|_V + \|f\|_{L^2(S)},$$

and this yields

$$\|u(\cdot,\mathbf{y})\|_Z^2 \le \|u(\cdot,\mathbf{y})\|_{L^2(S)}^2 + \frac{1}{a_{\min}^2}(\|a(\cdot,\mathbf{y})\|_{W^{1,\infty}(S)}\|u(\cdot,\mathbf{y})\|_V + \|f\|_{L^2(S)})^2.$$

Using the Poincaré inequality $\|u\|_{L^2(S)} \le C_P\|u\|_V$ for all $u \in V$, we obtain

$$\|u(\cdot,\mathbf{y})\|_Z^2 \le \left(\frac{1}{C_P^2} + \frac{2}{a_{\min}^2}\sup_{\mathbf{z}\in U}\|a(\cdot,\mathbf{z})\|_{W^{1,\infty}(S)}^2\right)\|u(\cdot,\mathbf{y})\|_V^2 + \frac{2}{a_{\min}^2}\|f\|_{L^2(S)}^2.$$

The proof is completed by using Theorem 2. □

In the following, we will discuss the regularity of $u(\mathbf{x},\mathbf{y})$ with respect to the $\mathbf{y}$ variable. Firstly, we introduce a multi-index notation. For $\boldsymbol{\mu} = (\mu_j)_{j\ge1} \in \mathbb{N}_0^{\mathbb{N}}$, where $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$, we define $|\boldsymbol{\mu}| := \mu_1 + \mu_2 + \cdots$, and we refer to $\boldsymbol{\mu}$ as a "multi-index" and $|\boldsymbol{\mu}|$ as the length of $\boldsymbol{\mu}$. For $\mu \subset \mathbb{N}$ of finite cardinality, we denote by

$$\partial_{\mathbf{y}}^{\boldsymbol{\mu}} u := \frac{\partial^{|\boldsymbol{\mu}|}}{\partial_{y_1}^{\mu_1}\partial_{y_2}^{\mu_2}\cdots}u$$

the partial derivative of order $|\boldsymbol{\mu}|$ of u with respect to $\mathbf{y}$.

Theorem 4. *Under Assumptions (3) and (17), for every $f \in V^*$, every $\mathbf{y} \in U$ and every $\boldsymbol{\mu} \subset \mathbb{N}$ of finite cardinality, the solution $u(\cdot,\mathbf{y})$ of the parametric weak problem (30) satisfies*

$$\|\partial_{\mathbf{y}}^{\boldsymbol{\mu}} u(\cdot,\mathbf{y})\|_V \le |\boldsymbol{\mu}|!\frac{\|f\|_{V^*}}{a_{\min}}\prod_{j\in\boldsymbol{\mu}} b_j, \tag{32}$$

where the sequence $\mathbf{b} = (b_j)_{j\ge1} \in \ell^1(\mathbb{N})$ is defined by

$$b_j := \frac{1}{a_{\min}}\sqrt{\tilde{v}_j}\,\|Y_j\|_{L^\infty(S)}. \tag{33}$$

Remark 1. As mentioned earlier, the relationship between j, ℓ, m is given by

$$j(\ell, m) = \ell(\ell+1) + 1 + m, \quad \ell = 0, 1, 2, \ldots; m = -\ell, \ldots, \ell.$$

The coefficients $\tilde{v}_j$ are defined by

$$\tilde{v}_j = \hat{v}_\ell, \quad \text{for} \quad \ell(\ell+1) + 1 - \ell \leq j \leq \ell(\ell+1) + 1 + \ell.$$

Proof. The proof of this theorem can be adapted from the proof of [4, Theorem 4.2] with a slight modification, namely the ∇ is replaced by the surface gradient Grad on the sphere. The key idea is the following recurrence relation deduced from (30)

$$\int_S a(\mathbf{x}, \mathbf{y}) \text{Grad}\,(\partial_{\mathbf{y}}^{\boldsymbol{\mu}} u(\mathbf{x}, \mathbf{y})) \cdot \text{Grad}\, v(\mathbf{x}) dS(\mathbf{x}) + \sum_{j \in \text{supp}(\boldsymbol{\mu})} \mu_j \int_S \sqrt{\tilde{v}_j} Y_j(\mathbf{x}) \text{Grad}\,(\partial_{\mathbf{y}}^{\boldsymbol{\mu} - \mathbf{e}_j} u(\mathbf{x}, \mathbf{y})) \cdot \text{Grad}\, v(\mathbf{x}) dS(\mathbf{x}) = 0, \tag{34}$$

for every $v \in V$, $\mathbf{y} \in U$ and $\boldsymbol{\mu} \subset \mathbb{N}$ with $0 \neq |\boldsymbol{\mu}| < \infty$. Here $\mathbf{e}_j \in \mathbb{N}^{\mathbb{N}}$ denotes the multi-index with entry 1 in the jth position and zeros elsewhere, and where $\text{supp}(\boldsymbol{\mu}) = \{j \in \mathbb{N} : \nu_j \neq 0\}$ denotes the "support" of $\boldsymbol{\mu}$. We now select in (34) the function $v(\mathbf{x}) = \partial_{\mathbf{y}}^{\boldsymbol{\mu}} u(\mathbf{x}, \mathbf{y}) \in V$ to estimate $\partial_{\mathbf{y}}^{\boldsymbol{\mu}} u(\mathbf{x}, \mathbf{y})$ in appropriate norms. □

3.4 Dimensional Truncation

Theorem 5. *Under Assumptions (3) and (17), for every $f \in V^*$, every $\mathbf{y} \in U$ and every $s \in \mathbb{N}$, the solution $u(\cdot, (\mathbf{y}_{\{1:s\}}; 0))$ of the truncated parametric weak problem (30) satisfies*

$$\|u(\cdot, \mathbf{y}) - u(\cdot, (\mathbf{y}_{\{1:s\}}; 0))\|_V \leq \frac{\|f\|_{V^*}}{2a_{\min}^2} \sum_{j \geq s+1} \sqrt{\tilde{v}_j} \|Y_j\|_{L^\infty(S)}.$$

Proof. The proof of this theorem can also be adapted from the proof of [4, Theorem 5.1] when the ∇ is replaced by the surface gradient Grad and the eigenfunctions $\psi_j = \sqrt{\tilde{v}_j} Y_j$. □

4 Spectral Method on the Sphere

Let $\mathscr{P}_L^*$ be the space of all spherical harmonics of degree $\leq L$ excluding the constant function, i.e.

$$\mathscr{P}_L^* = \text{span}\{Y_{\ell,m} : \ell = 1, \ldots, L; \;\; m = -\ell, \ldots, \ell\}.$$

The space $\mathscr{P}_L^*$ is a subspace of V since the orthogonality of the spherical harmonics implies

$$\int_S p dS = 0 \quad \forall p \in \mathscr{P}_L^*.$$

For a given function $u \in H^m(S)$ with $m \geq 1$, there is a constant $C > 0$ such that [3]

$$\inf_{p \in \mathscr{P}_L^*} \|u - p\|_{H^1(S)} \leq C L^{1-m} \|u\|_{H^m(S)}. \tag{35}$$

Consequently, for a given function $u \in Z \subset H^2(S)$, there is a positive constant C such that

$$\inf_{p \in \mathscr{P}_L^*} \|u - p\|_V \leq C L^{-1} \|u\|_Z. \tag{36}$$

For any $\mathbf{y} \in U$, we define the parametric spectral approximation $u_L(\cdot, \mathbf{y})$ as the spectral solution of the parametric deterministic problem: for $f \in V^*$ and $\mathbf{y} \in U$, find

$$u_L(\cdot, \mathbf{y}) \in \mathscr{P}_L^* : \quad b(\mathbf{y}; u_L(\cdot, \mathbf{y}), p) = \langle f, p \rangle \quad \forall p \in \mathscr{P}_L^*. \tag{37}$$

Theorem 6. *Under Assumptions (3) and (15), for every $f \in V^*$ and every $\mathbf{y} \in U$, the spectral approximations $u_L(\cdot, \mathbf{y})$ are stable in the sense that*

$$\|u_L(\cdot, \mathbf{y})\|_V \leq \frac{\|f\|_{V^*}}{a_{\min}}. \tag{38}$$

Moreover, for every $f \in L^2(S)$, as $L \to \infty$, there exists a constant $C > 0$ independent of L such that

$$\|u(\cdot, \mathbf{y}) - u_L(\cdot, \mathbf{y})\|_V \leq C L^{-1} \|f\|_{L^2(S)}. \tag{39}$$

Proof. The result follows from Cea's lemma together with the approximation property (36). □

Since we are interested in estimating the error in approximating functionals (26), we also assume that $G(\cdot) \in L^2(S)$.

Theorem 7. *Under Assumptions (3) and (15), for every $f \in L^2(S)$, every $G(\cdot) \in L^2(S)$, and every $\mathbf{y} \in U$, the spectral approximation $G(u_L(\cdot, \mathbf{y}))$ satisfy the asymptotic estimate*

$$|G(u(\cdot, \mathbf{y})) - G(u_L(\cdot, \mathbf{y}))| \leq C L^{-2} \|f\|_{L^2(S)} \|G(\cdot)\|_{L^2(S)},$$

where the constant $C > 0$ is independent of $\mathbf{y} \in U$.

Proof. For $G(\cdot) \in L^2(S)$ and any $\mathbf{y} \in U$, we define $v_G(\cdot, \mathbf{y}) \in V$ as the unique solution to the adjoint problem

$$b(\mathbf{y}; w, v_G(\cdot, \mathbf{y})) = G(w) \quad \forall w \in V. \tag{40}$$

Since b is symmetric, $b(\mathbf{y}; w, v) = b(\mathbf{y}; v, w)$ for all $v, w \in V$, we also have

$$b(\mathbf{y}; v_G(\cdot, \mathbf{y}), w) = G(w) \quad \forall w \in V.$$

So, by the regularity estimate (31), there is a constant $C > 0$, which is independent of $\mathbf{y}$ such that

$$\|v_G(\cdot, \mathbf{y})\|_Z \leq C \|G(\cdot)\|_{L^2(S)}. \tag{41}$$

Using the orthogonality property and (40), we may write for every $\mathbf{y} \in U$ and every $v_L \in \mathscr{P}_L^*$,

$$\begin{aligned} |G(u(\cdot, \mathbf{y})) - G(u_L(\cdot, \mathbf{y}))| &= |G(u(\cdot, \mathbf{y}) - u_L(\cdot, \mathbf{y}))| \\ &= |b(\mathbf{y}; u(\cdot, \mathbf{y}) - u_L(\cdot, \mathbf{y}), v_G(\cdot, \mathbf{y})| \\ &= |b(\mathbf{y}; u(\cdot, \mathbf{y}) - u_L(\cdot, \mathbf{y}), v_G(\cdot, \mathbf{y}) - p)| \\ &\leq C \|u(\cdot, \mathbf{y}) - u_L(\cdot, \mathbf{y})\|_V \|v_G(\cdot, \mathbf{y}) - p\|_V. \end{aligned}$$

Finally, we apply (36), (38) and (41) to obtain

$$|G(u(\cdot, \mathbf{y}) - G(u_L(\cdot, \mathbf{y}))| \leq C L^{-2} \|f\|_{L^2(S)} \|v_G\|_Z \leq C L^{-2} \|f\|_{L^2(S)} \|G(\cdot)\|_{L^2(S)}.$$

This completes the proof. □

5 Combined Error Estimates

We now present the error analysis for the combined QMC spectral approximation for the integral (26) using a randomly shifted lattice rule with N points in s dimensions. A realization for a draw of the shift $\boldsymbol{\Delta}$ will be denoted by $Q_{s,N}(\cdot; \boldsymbol{\Delta})$ and for each evaluation of the integrand, the exact solution $u(\cdot, \mathbf{y})$ of the parametric weak problem (30) is replaced by its spectral approximation $u_L \in \mathscr{P}_L^*$.

Theorem 8. *Under the same assumptions and definitions as in Theorems 5–7 and (18), if we approximate the integral over U by the randomly shifted lattice rule from Theorem 1 with N points in s dimensions, and for each shifted lattice point we solve the approximate elliptic problem (38) by a spectral method then we have the root-mean-square error bound*

$$\sqrt{\mathbb{E}[|\mathscr{I}(G(u)) - Q_{s,N}(G(u_L); \cdot)|^2]}$$

$$\leq C(\kappa(s,N)\|f\|_{V^*}\|G(\cdot)\|_{V^*} + CL^{-2}\|f\|_{L^2(S)}\|G\|_{L^2(S)}),$$

where

$$\kappa(s,N) = \begin{cases} s^{-2(1/p-1)} + N^{-(1-\delta)} & \text{when } p \in (0,2/3], \\ s^{-2(1/p-1)} + N^{-(1/p-1/2)} & \text{when } p \in (2/3,1), \\ (\sum_{j\geq s+1} b_j)^2 + N^{-1/2} & \text{when } p = 1, \end{cases}$$

and $\mathbb{E}[\cdot]$ *denotes the expectation with respect to the random shift* $\boldsymbol{\Delta}$ *which is uniformly distributed over* $[0,1]^s$.

Proof. We express the overall error as the sum of a dimensional truncation error, a QMC quadrature error, and a spectral approximation error:

$$\begin{aligned} &\mathscr{I}(G(u)) - Q_{s,N}(G(u_L);\boldsymbol{\Delta}) \\ &= (\mathscr{I} - \mathscr{I}_s)(G(u)) + (\mathscr{I}_s(G(u)) - Q_{s,N}(G(u);\boldsymbol{\Delta})) + Q_{s,N}(G(u-u_L);\boldsymbol{\Delta}). \end{aligned}$$

The mean-square error with respect to the random shift can then be bounded by

$$\begin{aligned} \mathbb{E}[|\mathscr{I}(G(u)) - Q_{s,N}(G(u_L);\cdot)|^2] \leq\ & 3|(\mathscr{I} - \mathscr{I}_s)(G(u))|^2 + \\ & 3\mathbb{E}\,[|I_s(G(u)) - Q_{s,N}(G(u);\cdot)|^2] + \\ & 3\mathbb{E}\,[|Q_{s,N}(G(u-u_L);\cdot)|^2]. \end{aligned} \tag{42}$$

For the truncation error, i.e., the first term in (42), we use the estimate

$$\begin{aligned} |(\mathscr{I} - \mathscr{I}_s)(G(u))| &= \left| \int_U G(u(\cdot,\mathbf{y}) - u(\cdot,(\mathbf{y}_{\{1:s\}};0)))d\mathbf{y} \right| \\ &\leq \sup_{\mathbf{y}\in U} |G(u(\cdot,\mathbf{y}) - u(\cdot,(\mathbf{y}_{\{1:s\}};0)))| \\ &\leq \|G(\cdot)\|_{V^*} \sup_{\mathbf{y}\in U} \|u(\cdot,\mathbf{y}) - u(\cdot,(\mathbf{y}_{\{1:s\}};0))\|_V, \end{aligned}$$

and then apply Theorem 5. The QMC error is already analysed in [4, Sect. 6]. Finally, for the spectral error, i.e., the third term in (42), using the property that the QMC quadrature weights $1/N$ are positive and sum to 1, we obtain

$$\mathbb{E}[|Q_{s,N}(G(u-u_L);\cdot)|^2] \leq \sup_{\mathbf{y}\in U} |G(u(\cdot,\mathbf{y}) - u_L(\cdot,\mathbf{y}))|^2,$$

and then apply Theorem 7. □

6 Numerical Experiments

Let the unit sphere S by parametrised by

$$\mathbf{x} = (x_1, x_2, x_3) = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta), \quad \theta \in [0, \pi], \phi \in [0, 2\pi).$$

Let $U = \left[-\frac{1}{2}, \frac{1}{2}\right]^2$, we consider the following simplified problem

$$\begin{cases} -\text{Div}\,(a(\mathbf{x}, \mathbf{y})\text{Grad}\, u(\mathbf{x}, \mathbf{y})) & = f(\mathbf{x}, \mathbf{y}), \\ \int_S u(\mathbf{x}, \mathbf{y}) dS(\mathbf{x}) & = 0, \quad \forall \mathbf{y} = (y_1, y_2) \in \left[-\frac{1}{2}, \frac{1}{2}\right]^2 \end{cases} \tag{43}$$

with

$$a(\mathbf{x}, \mathbf{y}) = 3 + y_1 + y_2 x_3,$$

and

$$f(\mathbf{x}, \mathbf{y}) = 2y_1 x_1 (8y_2 x_3^2 + 18x_3 + 6x_3 y_1 - y_2).$$

It can be shown that the exact solution is given by

$$u(\mathbf{x}, \mathbf{y}) = y_1 \cos(\phi)\sin(2\theta) = 2y_1 x_1 x_3.$$

The spherical harmonics and their gradients are computed explicitly using formulas in Varshalovich's book [10], see also [2]. Integration of a function f on the sphere is approximated by a quadrature of the form

$$\int_S f dS \approx \frac{2\pi}{M} \sum_{p=1}^{M/2} w_p \sum_{q=0}^{M-1} f(\sin\theta_p \cos\phi_q, \sin\theta_p \sin\phi_q, \cos\theta_p),$$

for an even number $M \geq 2$, where $\int_{-1}^{1} g(z)dz \approx \sum_{p=1}^{M/2}$ is a Gauss-Legendre rule and $\phi_q = 2\pi q/R$.

For fixed value $\mathbf{y} = (1/2, \ -1/4)^T$, Table 1 shows the values of the quantities

$$e_{\max} = \max_{\mathbf{x} \in \mathscr{Q}} |u(\mathbf{x}, \mathbf{y}) - u_L(\mathbf{x}, \mathbf{y})| \text{ and } e_2 = \left(\sum_{\mathbf{x} \in \mathscr{Q}} w_{\mathbf{x}} |u(\mathbf{x}, \mathbf{y}) - u_L(\mathbf{x}, \mathbf{y})|^2 \right)^{1/2},$$

where $\mathscr{Q}$ is the set of quadrature points.

Table 1 Errors of the PDE solvers for fixed $\mathbf{y} = (1/2, \ -1/4)^T$.

L	1	2	3	5
$e_{\max}$	0.2581	3.3307e−16	3.2613e−16	4.1633e−16
e_2	0.4583	3.3729e−16	3.4730e−16	3.8967e−16

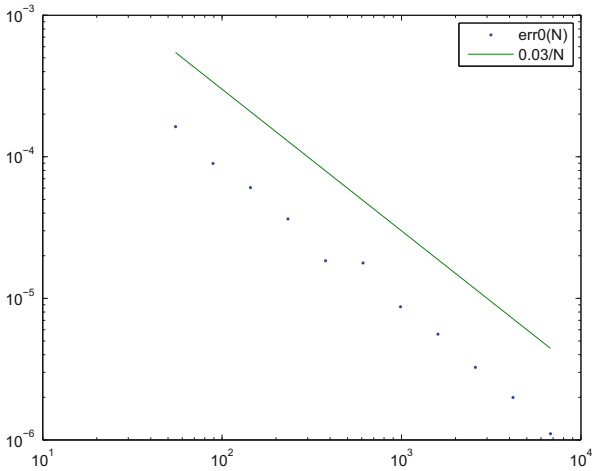


Fig. 1 Errors plot of err0(N).

Table 1 shows spectral approximation errors are neglectable for $L \geq 2$ in this example.

We used Fibonacci lattice point sets for our QMC rule since these are optimal for any choice of weights [7, Chap. 5]. The Fibonacci point set $\{\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(N)}\}$ of size $N = F_n$, where F_n is the nth Fibonacci number, has the generating vector $(1, \quad F_{n-1})$, so that $\mathbf{y}^{(k)} = \left(\frac{k}{F_n}, \quad \frac{kF_{n-1}}{F_n}\right) \mod 1$. The random shifts Δ are drawn uniformly from $[0,1]^2$.

We let $n = 10, 11, \ldots, 20$ and hence $N = 55{,}89, \ldots, 6{,}765$. Figure 1 shows the plot of quantities

$$\text{err0}(N) = \left(\mathbb{E} \left| \frac{1}{N} \sum_{k=1}^{N} u_5(\mathbf{x}_0, \mathbf{y}_k + \Delta) - \int_{\left[\frac{-1}{2}, \frac{1}{2}\right]^2} u(\mathbf{x}_0, \mathbf{y}) d\mathbf{y} \right|^2 \right)^{1/2},$$

where $\mathbf{x}_0 = (-0.9994, \ 0.0314, \ -0.0156) \in S$ and $\mathbb{E}$ is taken over 10 random shifts Δ's. The plot is in the log-log scale compared with $0.03/N$. Since the truncation error is also neglectable in this example, the numerical results are consistent with Theorem 8.

Acknowledgements The author acknowledges the support of the Australian Research Council under its Centre of Excellence program and helpful conversations with Prof. Ian Sloan, Prof. Christoph Schwab, Dr. Josef Dick and Dr. Frances Kuo.

References

1. Bieri, M., Schwab, C.: Sparse high order FEM for elliptic sPDEs. Comput. Methods Appl. Mech. Engrg. **198**, 1149–1170 (2009)
2. Ganesh, M., Le Gia, Q.T., Sloan, I.H.: A pseudospectral quadrature method for Navier-Stokes equations on rotating spheres. Math. Comp. **80**, 1397–1430 (2011)
3. Kamzolov, A.I.: The best approximation of classes of functions $W_p^\alpha(S^n)$ by polynomials in spherical harmonics. Mat. Zametki **32**, 285–293 (1982)
4. Kuo, F.Y., Schwab, C., Sloan, I.H.: Quasi-Monte Carlo finite element methods for a class of elliptic partial differential equations with random coefficients. SIAM J. Numer. Anal. **50**, 3351–3374 (2012)
5. Lions, J.L., Magenes, E.: Non-Homogeneous Boundary Value Problems and Applications I. Springer, New York (1972)
6. Müller, C.: Spherical harmonics. Lecture Notes in Mathematics, vol. 17. Springer, Berlin (1966)
7. Niederreiter, H.: Random Number Generation and Quasi-Monte Carlo Methods. SIAM, Philadelphia (1992)
8. Schoenberg, I.J.: Positive definite function on spheres. Duke Math. J. **9**, 96–108 (1942)
9. Sloan, I.H., Wang, X., Woźniakowski, H.: Finite-order weights imply tractability of multivariate integration. J. Complexity **20**, 46–74 (2004)
10. Varshalovich, D.A., Moskalev, A.N., Khersonskii, V.K.: Quantum Theory of Angular Momentum. World Scientific, Singapore (1988)
11. Wilson: Turbulent diffusivity in the free atmosphere inferred from MST radar measurements: a review. Annales Geophysicae **22**, 3869–3887 (2004)

Sampling and Low-Rank Tensor Approximation of the Response Surface

Alexander Litvinenko, Hermann G. Matthies, and Tarek A. El-Moselhy

Abstract Most (quasi)-Monte Carlo procedures can be seen as computing some integral over an often high-dimensional domain. If the integrand is expensive to evaluate—we are thinking of a stochastic PDE (SPDE) where the coefficients are random fields and the integrand is some functional of the PDE-solution—there is the desire to keep all the samples for possible later computations of similar integrals. This obviously means a lot of data. To keep the storage demands low, and to allow evaluation of the integrand at points which were not sampled, we construct a low-rank tensor approximation of the integrand over the whole integration domain. This can also be viewed as a representation in some problem-dependent basis which allows a sparse representation. What one obtains is sometimes called a "surrogate" or "proxy" model, or a "response surface". This representation is built step by step or sample by sample, and can already be used for each new sample. In case we are sampling a solution of an SPDE, this allows us to reduce the number of necessary samples, namely in case the solution is already well-represented by the low-rank tensor approximation. This can be easily checked by evaluating the residuum of

A. Litvinenko (✉)
Institute for Scientific Computing, Technische Universität Braunschweig, Hans-Sommerstr. 65, Brunswick, Germany

Division of Mathematics and Computational Sciences and Engineering (MCSE), 4700 King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia
e-mail: alexander.litvinenko@kaust.edu.sa

H.G. Matthies
Institute for Scientific Computing, Technische Universität Braunschweig, Hans-Sommerstr. 65, Brunswick, Germany
e-mail: wire@tu-bs.de

T.A. El-Moselhy
MIT, Cambridge, MA, USA
e-mail: tmoselhy@mit.edu

J. Dick et al. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2012*,
Springer Proceedings in Mathematics & Statistics 65,
DOI 10.1007/978-3-642-41095-6_27, © Springer-Verlag Berlin Heidelberg 2013

the PDE with the approximate solution. The procedure will be demonstrated in the computation of a compressible transonic Reynolds-averaged Navier-Strokes flow around an airfoil with random/uncertain data.

1 Motivation

Situations where one is concerned with uncertainty quantification often come in the following guise: we are investigating some physical system which is modelled by an evolution equation for its state:

$$\frac{\partial}{\partial t}u(t) = F(p; u(t)) + f(p; t), \tag{1}$$

where $u(t)$ describes the state of the system at time $t \in [0, T]$ lying in a Hilbert space $\mathcal{V}$ (for the sake of simplicity), F is an operator modelling the physics of the system, and f is some external influence (action/excitation/loading). In particular $F(p; u(t))$ could be some parameter-dependent differential operator, for example

$$\frac{\partial}{\partial t}u(x,t) = \nabla \cdot (\kappa(x,\omega)\nabla u(x,t)) + f(x,t), \quad x \in \mathcal{G} \subset \mathbb{R}^d,\ t \in [0,T],$$

where $\kappa(x,\omega)$ is a random field dependent on a random parameter ω in some probability space Ω, and one may take $\mathcal{V} = L_2(\mathcal{G})$.

The model depends on some parameter $p \in \mathcal{P}$; in the context of uncertainty quantification the actual value of p is uncertain. Often this uncertainty is modelled by giving the set $\mathcal{P}$ a probability measure. Evaluation and quantification of the uncertainty will often involve functionals of the state $\Psi(u(p;t))$, and the functional dependence of u on p becomes important. Similar situations arise in design, where p may be a design parameter still to be chosen, and one may seek a design such that a functional $\Psi(u(p;t))$ is e.g. maximised.

The situation just sketched involves a number of objects which are functions of the parameter values. While evaluating $F(p)$ or $f(p)$ for a certain p may be straightforward, one may easily envisage situations where evaluating $u(p)$ or $\Psi(u(p))$ may be very costly as it may involve some very time consuming simulation or computation, like for example running a climate model.

As will be shown in the following Sect. 2, any such parametric object like $u(p)$, $F(p)$, or $f(p)$ may be seen as an element of a tensor product space [4, 5]. This in turn can be used to find very sparse approximations to those objects, and in turn much cheaper ways to evaluate other parameter values. In particular, this may be used in the uncertainty quantification to large advantage, like computing means, covariances, exceedance probabilities, etc. For this, the dependence of $F(p)$ and $f(p)$ on p has to be propagated to the solution or state vector $u(p)$, see e.g. [18].

2 Parametric Problems

Let $r : \mathscr{P} \to \mathscr{V}$ be a parametric description of one of the objects alluded to in the introduction, where $\mathscr{P}$ is some set, and $\mathscr{V}$ for the sake of simplicity is assumed as a separable Hilbert space with inner product $\langle\cdot|\cdot\rangle_{\mathscr{U}}$ (the meaning of the index $\mathscr{U}$ will soon become clear). What we desire is a simple representation/approximation of that function, which avoids solving Eq. 1 every time one wants to know $r(p)$ for a new $p \in \mathscr{P}$, i.e. a *response surface* (RS) or surrogate model, sometimes also called an *emulator*, whereas the solver for Eq. 1 is termed a *simulator*.

One relatively well-known way, particularly in statistical estimation [10], turns the problem into one of approximation of a linear mapping: let $\mathscr{U} = \overline{\text{span}}\, r(\mathscr{P}) = \overline{\text{span}}\, \text{im}\, r \subseteq \mathscr{V}$ be the smallest closed subspace of $\mathscr{V}$ which is spanned by all the vectors $\{r(p)|\ p \in \mathscr{P}\}$. Then to each such function $r : \mathscr{P} \to \mathscr{U}$ one may associate a linear map

$$R : \mathscr{U} \ni u \mapsto \langle r(\cdot)|u\rangle_{\mathscr{U}} \in \mathbb{R}^{\mathscr{P}}. \tag{2}$$

By construction, R is injective. This may be used to define an inner product on $\text{im}\, R$ as

$$\forall \phi, \psi \in \text{im}\, R : \quad \langle\phi|\psi\rangle_{\mathscr{R}} := \langle R^{-1}\phi|R^{-1}\psi\rangle_{\mathscr{U}},$$

and let $\mathscr{R}$ be the completion of $\text{im}\, R$ with that inner product. It is obvious that R is a unitary map between the Hilbert spaces $\mathscr{U}$ and $\mathscr{R}$.

Up to now, no structure on the set $\mathscr{P}$ has been assumed, whereas on $\mathscr{U}$ the inner product is assumed to measure what is important for the state $r(p) \in \mathscr{U}$. This is carried via the map R in Eq. 2 onto the space of scalar functions $\mathscr{R}$ on the set $\mathscr{P}$, and the inner product there measures essentially the same thing as the one on $\mathscr{U}$.

2.1 Reproducing Kernel Hilbert Space

This is a first representation, and $\mathscr{R}$ is called a reproducing kernel Hilbert space (RKHS) [9] with *reproducing kernel* $\varkappa \in \mathbb{R}^{\mathscr{P}} \times \mathbb{R}^{\mathscr{P}}$

$$\varkappa(p_1, p_2) := \langle r(p_1)|r(p_2)\rangle_{\mathscr{U}}. \tag{3}$$

It is straightforward to verify that it defines an obviously continuous (on $\mathscr{R}$) point-evaluation functional

$$\delta_p : \mathscr{R} \ni \phi \mapsto \langle\delta_p, \phi\rangle_{\mathscr{R}^* \times \mathscr{R}} := \phi(p) = \langle\varkappa(p, \cdot)|\phi\rangle_{\mathscr{R}} \in \mathbb{R},$$

hence the name.

In other settings like classification or machine learning, e.g. with support vector machines, where $p \in \mathscr{P}$ has to be classified as belonging to certain subsets of $\mathscr{P}$, the space $\mathscr{V}$ and the map $r : \mathscr{P} \to \mathscr{V}$ may often be freely chosen. This is then referred to as the "kernel trick", and classification may be achieved by mapping these subsets with r into $\mathscr{U}$ and separating them with hyperplanes—a linear classifier.

In terms of representation, one may now choose a basis $\{\varphi_m\}_{m\in\mathbb{N}}$ in $\mathscr{R}$, which may be assumed to be a complete orthonormal system (CONS). With the CONS $\{y_m \mid y_m = R^{-1}\varphi_m\}_{m\in\mathbb{N}}$ in $\mathscr{U}$, the operator R, its inverse R^{-1}, and the parametric element $r(p)$ become

$$R = \sum_m \varphi_m \otimes y_m; \quad R^{-1} = \sum_m y_m \otimes \varphi_m; \quad r(p) = \sum_m y_m \varphi_m(p), \tag{4}$$

exhibiting the tensorial nature of the representation mapping. With such a basis one may define a unitary map from ℓ_2 to $\mathscr{R}$ and via R^{-1} further to $\mathscr{U}$:

$$\ell_2 \ni \boldsymbol{a} = (a_1, a_2, \ldots) \mapsto \sum_m a_m \varphi_m \mapsto \sum_m a_m y_m \in \mathscr{U}. \tag{5}$$

Note that this representation is linear in the new 'parameters' $(a_1, a_2, \ldots) \in \ell_2$. Model reductions may be achieved by choosing only subspaces of $\mathscr{R}$ or ℓ_2, or by approximating the map R^{-1}. This pattern of Eqs. 4 or 5 repeats itself for all representations to follow.

2.2 Spectral Decomposition

As a way of measuring of what is important on the set $\mathscr{P}$, assume that there is another inner product $\langle\cdot|\cdot\rangle_{\mathscr{W}}$ for scalar functions $\phi \in \mathbb{R}^{\mathscr{P}}$, and denote the Hilbert space of functions with that inner product by $\mathscr{W}$. With this, one may define [10] a densely defined map C in $\mathscr{U}$ through the bilinear form

$$\langle Cu|v\rangle_{\mathscr{U}} := \langle Ru|Rv\rangle_{\mathscr{W}}, \quad \forall u, v \in \mathscr{U}$$

The map $C = R^*R$ (the adjoint is taken w.r.t. the $\mathscr{W}$-inner product, by abuse of notation we shall still call the map R) may be called the 'correlation' operator. By construction it is injective, positive, and self-adjoint.

Often the inner product $\langle\cdot|\cdot\rangle_{\mathscr{W}}$ comes from a measure ϖ on $\mathscr{P}$, so that $\mathscr{W}$ may be taken as $L_2(\mathscr{P}, \varpi)$. One important class of problems is when ϖ is a probability measure on $\mathscr{P}$, i.e. $\varpi(\mathscr{P}) = 1$. Often the set has more structure, like being in a topological space, differentiable (Riemann) manifold, or a Lie group, which then may induce the choice of σ-algebra or measure. In all such cases one has $C = R^*R = \int_{\mathscr{P}} r(p) \otimes r(p)\ \varpi(\mathrm{d}p)$. It is the factorisation of $C = R^*R$ which paves the

way for further possibilities of representation. Most common is to use the spectral decomposition (e.g. [25]) of C:

$$Cu = \int_0^\infty \lambda \, \mathrm{d}E_\lambda(u), \tag{6}$$

where E_λ is the corresponding projection valued spectral measure, with the spectrum $\sigma(C) \subseteq \mathbb{R}_+$. For the sake of simplicity assume that C has a pure point spectrum $\sigma_p(C) = \sigma(C)$—the important case where C has also a continuous spectrum requires too many technical tools such as Gel'fand triplets (rigged Hilbert spaces) and generalised eigenvectors to be treated in this short note—such that Eq. 6 may be written with the CONS of unit-$\mathscr{U}$-norm eigenvectors v_m:

$$Cu = \sum_m \lambda_m \langle v_m | u \rangle_{\mathscr{U}} v_m = \sum_m \lambda_m (v_m \otimes v_m) u.$$

From this follows the singular value decomposition of R, with $\lambda_m^{1/2} s_m := R v_m$:

$$R = \sum_m \lambda_m^{\frac{1}{2}} (s_m \otimes v_m); \quad R^* = \sum_m \lambda_m^{\frac{1}{2}} (v_m \otimes s_m); \quad r(p) = \sum_m \lambda_m^{\frac{1}{2}} s_m(p) v_m,$$

where the last relation is the so-called Karhunen-Loève or proper orthogonal decomposition (POD). Observe that r—as well as R^*—is linear in the s_m. Similarly to Eq. 5, we have the—linear in $\boldsymbol{a}$—representation:

$$\ell_2 \ni \boldsymbol{a} = (a_1, a_2, \ldots) \mapsto \sum_m a_m s_m \mapsto \sum_m \lambda_m^{1/2} a_m v_m \in \mathscr{U}.$$

An alternative formulation of the spectral decomposition Eq. 6 is [25] that C is unitarily equivalent with a multiplication operator:

$$C = V M_k V^* = (V M_k^{1/2})(V M_k^{1/2})^*,$$

where V is unitary between some $L_2(\mathscr{T})$ and $\mathscr{U}$, M_k is a multiplication operator on the measure space $\mathscr{T}$ with a positive function $k(s) > 0$, and $M_k^{1/2} = M_{\sqrt{k}}$. The essential range of k is the spectrum of C. This gives in the now familiar manner a representation on $L_2(\mathscr{T})$ through the choice of a CONS $\{\varsigma_m\}$. Setting $u_m := V M_{\sqrt{k}} \varsigma_m$, one obtains

$$(V M_k^{1/2}) = (V M_{\sqrt{k}}) = \sum_m u_m \otimes \varsigma_m$$

as tensorial representation. A representation on $\mathscr{U}$ is given by the factorisation $C = C^{1/2} C^{1/2}$, where the positive square root of C is $C^{1/2} = V M_{\sqrt{k}} V^*$.

2.3 Other Factorisations of C

Other factorisations $C = B^*B$—which are all unitarily equivalent—lead to analogous representations. Let $B : \mathscr{U} \to \mathscr{H}$ be an injective mapping into another Hilbert space $\mathscr{H}$. Pick a CONS $\{e_m\}$ in $\mathscr{H}$ and set $f_m := B^*e_m$, then

$$B^* = \sum_m f_m \otimes e_m,$$

again a tensorial representation. All the representations considered so far are of this type. Similarly to Eq. 5, we have the—linear in $\boldsymbol{a}$—representation:

$$\ell_2 \ni \boldsymbol{a} = (a_1, a_2, \ldots) \mapsto \sum_m a_m e_m \mapsto \sum_m a_m f_m \in \mathscr{U}.$$

For finite dimensional spaces, a favourite choice for such a decomposition of C is the Cholesky factorisation $C = LL^T$.

3 Discretisation and Computation

For brevity we follow [17], where more references may be found, cf. also the recent monograph [13]. For the sake of simplicity, let us concentrate on the time-independent or stationary version of Eq. 1, namely $F(p;u) = f(p)$. Usually this is some partial differential equation and has to be discretised, approximated, or somehow projected onto some finite dimensional subspace $\mathscr{V}_N \subset \mathscr{V}$, with $\dim \mathscr{V}_N = N$. The entities of Eq. 1 which are projected or induced on the corresponding $\mathbb{R}^N$ will be denoted by boldface, such that the stationary, projected equation reads as

$$\boldsymbol{F}(p;\boldsymbol{u}) = \boldsymbol{f}(p). \tag{7}$$

To propagate the parametric dependence, choose a finite dimensional subspace of the Hilbert spaces, say $\mathscr{S}_M \subset \mathscr{S}$ for the solution $\boldsymbol{u}(p)$ in Eq. 7. Via Galerkin projection or collocation, or other such techniques, the still parametric model Eq. 7 is thereby formulated on the tensor product $\mathscr{V}_N \otimes \mathscr{S}_M$, denoted as

$$\mathbf{F}(\mathbf{u}) = \mathbf{f}. \tag{8}$$

The results of Sect. 2 (particularly Sect. 2.2) show that there are multiple possibilities for the choice of $\mathscr{S}$, and hence finite dimensional subspaces $\mathscr{S}_M$. The solution of Eq. 8 is often computationally challenging, as $\dim \mathscr{V}_N \otimes \mathscr{S}_M = N \times M$ may be very large. One possibility for such high-dimensional problems are the low-rank approximations, by representing the entities in Eq. 7 such as $\boldsymbol{F}$, $\boldsymbol{u}$, and $\boldsymbol{f}$ in a low-rank format. Several numerical techniques [3, 12, 19, 21] have been developed

recently to obtain an approximation to the solution $\mathbf{u} \approx \sum_j \boldsymbol{u}_j \otimes \boldsymbol{z}_j$ to Eq. 8 in this format by only ever operating on the data-sparse low-rank representation, thus allowing for an efficient resolution of the high-dimensional problem [5,8,11,22–24].

Once this has been computed, any other functional such as $\Psi(u(p))$ may be computed with relative ease. In case there is a probability measure on $\mathscr{P}$, for example to quantify some uncertainty in the parameters, the functionals usually take the form of expectations, a variance, an exceedance probability, or other such quantity needed in an *uncertainty quantification*.

3.1 Sampling by Simulation

The probability space is given as a triplet $(\Omega, \mathscr{A}, \mathbb{P})$, where Ω is the set of elementary events $\omega \in \Omega$, $\mathscr{A}$ a σ-algebra, and $\mathbb{P}$ the probability measure. Assume we want to compute

$$\overline{\Psi} = \mathbb{E}\left(\Psi\left(\cdot, u_e(\cdot)\right)\right) = \int_{\Omega} \Psi\left(\omega, u_e(\omega)\right) \mathbb{P}(d\omega), \tag{9}$$

where $\mathbb{P}$ is a probability measure on Ω, and u_e is the exact solution of a PDE depending on the parameter $\omega \in \Omega$:

$$\boldsymbol{F}[\omega]\left(u_e(\omega)\right) = f(\omega), \quad \text{a.s. in } \omega \in \Omega,$$

$u_e(\omega)$ is a $\mathscr{U}$-valued random variable (RV), where e.g. $\mathscr{U} := \mathring{H}^1(\mathscr{G}) = \{u \in H^1(\mathscr{G}) \mid u = 0 \text{ on } \partial\mathscr{G}\}$ and $\mathscr{G}$ computational domain.

To compute an approximation $u_a(\omega)$ to $u_e(\omega)$ via MC simulation is expensive, even for one value of ω. Put $u_a(\omega)$ into Eq. 9, obtain

$$\overline{\Psi} \approx \sum_{i=1}^{N_s} \Psi\left(\omega_i, u_a(\omega_i)\right) w_i,$$

where N_s is number of quadrature points (or MC simulations), w_i weights and ω_i sample points.

To simulate u_a one may need samples of the random field (RF) which depends on infinitely many random variables (RVs). This has to be reduced/transformed $\Theta : \Omega \to [0,1]^M$ to a finite number M of RVs $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_M)$, with $\mu = \Theta_* \mathbb{P}$ the push-forward measure:

$$\overline{\Psi} = \int_{\Omega} \Psi(\omega, u_e(\omega))\, \mathbb{P}(\mathrm{d}\omega) = \int_{[0,1]^M} \hat{\Psi}(\boldsymbol{\theta}, u_a(\boldsymbol{\theta}))\, \mu(\mathrm{d}\boldsymbol{\theta}).$$

This is a product measure for independent RVs $(\theta_1, \ldots, \theta_M)$. The next aim is to approximate an expensive simulation $u_a(\boldsymbol{\theta})$ by cheaper emulation.

4 Constructing the Emulation

The idea is to approximate the solution by a (multivariate) polynomial (see Eqs. 10 and 13) and then to use it for sampling.

Let $\boldsymbol{u}(\omega) = (u(x_1,\omega),\dots,u(x_N,\omega))^T$ be the solution (or a functional of the solution), which depends on a random parameter ω. $\boldsymbol{u}(\omega)$ can be approximated in a set of new independent Gaussian random variables (Wiener's [26] polynomial chaos expansions (PCE)):

$$\boldsymbol{u}(\boldsymbol{\theta}(\omega)) = \sum_{\beta\in\mathscr{J}} \boldsymbol{u}_\beta H_\beta(\boldsymbol{\theta}(\omega)) \approx \sum_{\beta\in\mathscr{J}_{M,p}} \boldsymbol{u}_\beta H_\beta(\boldsymbol{\theta}) := \mathbf{U}\boldsymbol{H}, \tag{10}$$

where $\mathbf{U} := (\dots \boldsymbol{u}_\beta \dots) \in \mathbb{R}^{N\times L}$, $\boldsymbol{H} := (\dots H_\beta(\boldsymbol{\theta})\dots)^T \in \mathbb{R}^L$, $L := |\mathscr{J}_{M,p}| = \frac{(M+p)!}{M!p!}$, $\boldsymbol{\theta}(\omega) = (\theta_1(\omega),\dots,\theta_M(\omega))$, $\boldsymbol{u}_\beta = (u_\beta(x_1),\dots,u_\beta(x_N))^T$ are coefficients, $H_\beta(\boldsymbol{\theta})$ the multivariate Hermite polynomials

$$H_\beta(\boldsymbol{\theta}) := \prod_{j=1}^{M} h_{\beta_j}(\theta_j) \quad \forall\boldsymbol{\theta}\in\mathbb{R}^M,\ \beta\in\mathscr{J}_{M,p} := \{\beta\in\mathscr{J} \mid \gamma(\beta)\le M, \quad |\beta|\le p\},$$

where $\gamma(\beta) := \max\{j\in\mathbb{N}\,|\,\beta_j > 0\}$, h_{β_j} are univariate Hermite polynomials, $\beta\in\mathscr{J}$ a multiindex. For the purpose of actual computation, truncate PCE after finitely many terms (Eq. 10) and obtain the finite multiindex subset $\mathscr{J}_{M,p}$ of the infinite set $\mathscr{J} := \{\beta|\beta = (\beta_1,\dots,\beta_j,\dots),\ \beta_j\in\mathbb{N}_0\}$. Since Hermite polynomials are orthogonal, the coefficients $\boldsymbol{u}_\beta$ can be computed by projection:

$$\boldsymbol{u}_\beta = \frac{1}{\beta!}\int_\Theta H_\beta(\boldsymbol{\theta})\boldsymbol{u}(\boldsymbol{\theta})\,\mathbb{P}(d\boldsymbol{\theta}) \approx \frac{1}{\beta!}\sum_{i=1}^{Z} H_\beta(\boldsymbol{\theta}_i)\boldsymbol{u}(\boldsymbol{\theta}_i)w_i =: \mathbf{W}\Lambda_\beta, \tag{11}$$

where $\mathbf{W} := (\boldsymbol{u}(\boldsymbol{\theta}_1),\dots,\boldsymbol{u}(\boldsymbol{\theta}_Z)) \in \mathbb{R}^{N\times Z}$, $\Lambda_\beta := \frac{1}{\beta!}(H_\beta(\boldsymbol{\theta}_1)w_1,\dots,H_\beta(\boldsymbol{\theta}_Z)w_Z)^T \in \mathbb{R}^Z$. The multidimensional integral over Θ is computed approximately, for example, on a sparse Gauss-Hermite grid where $\boldsymbol{\theta}_i$ are quadrature points, w_i corresponding weights and Z is the number of quadrature points [2, 6]. The matrix $\mathbf{W}$ is represented as $\boldsymbol{A}\boldsymbol{B}^T$ via a rank k approximation (see Sect. 5). The matrix $\mathbf{U}$ of all PCE coefficients will be

$$\mathbf{U} := (\dots\boldsymbol{u}_\beta\dots) = \mathbf{W}\boldsymbol{\Lambda} \approx \boldsymbol{A}\boldsymbol{B}^T\boldsymbol{\Lambda}, \quad \beta\in\mathscr{J}_{M,p}, \tag{12}$$

where $\boldsymbol{\Lambda} = (\dots\Lambda_\beta\dots)\in\mathbb{R}^{Z\times L}$. The final low-rank response surface will be

$$\boldsymbol{u}(\boldsymbol{\theta}) \approx \boldsymbol{A}\boldsymbol{B}^T\boldsymbol{\Lambda}\boldsymbol{H}. \tag{13}$$

Now, having low-rank response surface (Eq. 13), we can easily generate a large sample ($\boldsymbol{u}(\boldsymbol{\theta}_1)$, $\boldsymbol{u}(\boldsymbol{\theta}_2)$,...) by generating the random vector $\boldsymbol{\theta}$ and evaluating the expression in Eq. 13. After each m new samples we perform the update of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ as shown in the Algorithm in Sect. 5.

4.1 Emulation Instead of Simulation

In applications it is often very time-consuming (or/and expensive) to generate a sample which is large enough to compute statistical functionals, probability density functions (pdf), cumulative distribution functions (cdf), etc. Our idea [15] is to construct an approximation of the response surface (RS) from few samples and then to use the residual for its improvement. A motivation for this idea comes from the fact that in many software packages for solving engineering and physical problems it is impossible or very difficult to change the code, but it is possible to access the residual.

Assume that our response surface is an approximation via multivariate Hermite polynomials as in Eq. 10, where coefficients are computed like in Eqs. 11 and 12 with quadrature points $\boldsymbol{\theta}_j$, $j = 1..Z$.

Assume now that we get the new random point $\boldsymbol{\theta}_{Z+j}$, $j = 1, \ldots m$, where the solution $u(x, \boldsymbol{\theta}_{Z+j})$ should be computed. The following algorithm computes this solution efficiently and updates the given response surface.

Algorithm 1.

1. *Evaluate the RS from Eq. 13 in* $\boldsymbol{\theta}_{Z+j}$. *Let* $u_a(x, \boldsymbol{\theta}_{Z+j})$ *be the obtained value.*
2. *Compute the norm of residual* $\|r\| := \|F(u_a(x, \boldsymbol{\theta}_{Z+j})) - f(\boldsymbol{\theta}_{Z+j})\|$ *of the deterministic problem (e.g. evaluate one iteration). If* $\|r\|$ *is small then there is no need to solve the expensive deterministic problem in* $\boldsymbol{\theta}_{Z+j}$, *otherwise solve the deterministic problem.*
3. *Extend matrix* $\mathbf{W}$ *and vector* Λ_β *and recompute* $\boldsymbol{A}$, $\boldsymbol{B}^T$ *and* $\boldsymbol{\Lambda}$ *in Eq. 13. Go to Step 1.*

In the best case we never have to solve the deterministic problem again. In the worst case we must solve the deterministic problem for each $\boldsymbol{\theta}_{Z+j}$, $j = 1, 2, ...$ The numerical results (Fig. 4, left) in Sect. 6 (solution is smooth, no shock) show that with this algorithm one can reduce the number of needed TAU iterations from 10,000 to 1,000. If the solution is discontinuous (e.g. with shock) then our response surface is a very poor approximation and the produced value can not be used as a good starting point (see Fig. 6).

5 Low-Rank Data Compression

A large number of stochastic realisations of random fields requires a large amount of memory and powerful computational resources. To decrease memory requirements and computing time, we use a low-rank approximation [5, 8, 11, 22–24] for all realisations of the solution. This low-rank approximation allows an effective post-processing with drastically reduced memory requirements. For each new realisation only a corresponding low-rank update is computed (see e.g. [1]). This can be practical when, e.g., the results of many thousands of Monte Carlo simulations should be computed and stored. Let $\boldsymbol{u}_i \in \mathbb{R}^N$, $i = 1..N_s$ (N_s can be e.g. equal to Z), be the solution vectors (snapshots), where N_s is a number of stochastic realisations of the solution. Let us build from all these vectors the matrix $\mathbf{W} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{N_s}) \in \mathbb{R}^{N \times N_s}$ and consider the approximation

$$\mathbf{W} \approx \mathbf{W}_k = \boldsymbol{A}\boldsymbol{B}^T, \text{ where } \boldsymbol{A} \in \mathbb{R}^{N \times k}, \ \boldsymbol{B} \in \mathbb{R}^{N_s \times k} \text{ and } \|\mathbf{W} - \mathbf{W}_k\| < \varepsilon, \ k \ll \min\{N, N_s\}. \tag{14}$$

To compute factors $\boldsymbol{A}$ and $\boldsymbol{B}$ in Eq. 14 we omit all singular values which are smaller than a given level ε or, an alternative variant, we leave a fixed number ($= k$) of largest singular values. After truncation we speak about *reduced singular value decomposition (denoted by rSVD)* $\mathbf{W}_k = \boldsymbol{U}_k \Sigma_k \boldsymbol{V}_k{}^T$, where $\boldsymbol{U}_k \in \mathbb{R}^{N \times k}$ contains the first k columns of $\boldsymbol{U}$, $\boldsymbol{V}_k \in \mathbb{R}^{N_s \times k}$ contains the first k columns of $\boldsymbol{V}$ and $\Sigma_k \in \mathbb{R}^{k \times k}$ contains the k-biggest singular values of Σ. We denote $\boldsymbol{A} = \boldsymbol{U}_k \Sigma_k$ and $\boldsymbol{B} = \boldsymbol{V}_k$. The Eckart-Young-Mirsky theorem (see more in [20] or [7]) states that matrix $\mathbf{W}_k$ is the best approximation of $\mathbf{W}$ in the set of all rank-k matrices w.r.t. the Frobenius norm.

Suppose $\mathbf{W}_k$ is given. Suppose also that the matrix $\mathbf{W}' \in \mathbb{R}^{N \times m}$ contains new m solution vectors. For a small m, computing the factors $\boldsymbol{E} \in \mathbb{R}^{N \times k}$ and $\boldsymbol{D} \in \mathbb{R}^{m \times k}$ such that $\mathbf{W}' \approx \boldsymbol{E}\boldsymbol{D}^T$ is not expensive. Now our purpose is to compute with linear complexity the rank-k approximation of $\mathbf{W}_{\text{new}} := [\mathbf{W}\,\mathbf{W}'] \in \mathbb{R}^{N \times (N_s + m)}$. To do this, we build two concatenated matrices $\boldsymbol{A}_{\text{new}} := [\boldsymbol{A}\ \boldsymbol{E}] \in \mathbb{R}^{N \times 2k}$ and $\boldsymbol{B}^T_{\text{new}} = \text{blockdiag}[\boldsymbol{B}^T\ \boldsymbol{D}^T] \in \mathbb{R}^{2k \times (N_s + m)}$. Note that the difficulty now is that matrices $\boldsymbol{A}_{\text{new}}$ and $\boldsymbol{B}_{\text{new}}$ have rank $2k$. To truncate the rank from $2k$ to k we use the QR-algorithm below with complexity $\mathcal{O}((N + N_s)k^2 + k^3)$ [1, 7], linear in $(N + N_s)$.

Algorithm 2 (Rank truncation operation $\tau_{2k \to k}(\cdot)$).

1. *Compute (reduced) QR-factorization of* $\boldsymbol{A}_{new} = \boldsymbol{Q}_A \boldsymbol{R}_A$ *and* $\boldsymbol{B}_{new} = \boldsymbol{Q}_B \boldsymbol{R}_B$, *where* $\boldsymbol{Q}_A \in \mathbb{R}^{N \times 2k}$, $\boldsymbol{Q}_B \in \mathbb{R}^{N_s \times 2k}$, *and upper triangular matrices* $\boldsymbol{R}_A$, $\boldsymbol{R}_B \in \mathbb{R}^{2k \times 2k}$.
2. *Compute rSVD with k largest eigenvalues of* $\boldsymbol{R}_A \boldsymbol{R}_B^T = \boldsymbol{U} \Sigma \boldsymbol{V}^T$.
3. *Compute* $\boldsymbol{U}_k := \boldsymbol{Q}_A \boldsymbol{U}$, $\boldsymbol{V}_k := \boldsymbol{Q}_A \boldsymbol{V}^T$.

Finally, obtain $(\boldsymbol{U}_k \Sigma)\boldsymbol{V}_k^T = \tau_{2k \to k}(\boldsymbol{A}_{\text{new}} \boldsymbol{B}^T_{\text{new}})$.

5.1 Mean and Variance in the Rank-k Format

Denote the j-th row of matrix $\boldsymbol{A}$ by $\boldsymbol{a}_j \in \mathbb{R}^k$ and the i-th row of matrix $\boldsymbol{B}$ by $\boldsymbol{b}_i \in \mathbb{R}^k$. Then one can estimate the mean solution $\overline{\boldsymbol{u}} \in \mathbb{R}^N$ as follows

$$\overline{\boldsymbol{u}} := \frac{1}{N_s} \sum_{i=1}^{N_s} \boldsymbol{u}_i = \frac{1}{N_s} \sum_{i=1}^{N_s} \boldsymbol{A} \cdot \boldsymbol{b}_i = \boldsymbol{A}\overline{\boldsymbol{b}}. \tag{15}$$

The computational complexity is $\mathcal{O}(k(N_s + N))$, in contrast to $\mathcal{O}(N \cdot N_s)$ for usual dense data format. One can compute the variance of the solution $\mathrm{var}(\boldsymbol{u}) \in \mathbb{R}^N$ by the computing the covariance matrix and taking its diagonal. For this, first, one computes the centred matrix $\boldsymbol{W}_c := \boldsymbol{W} - \overline{\boldsymbol{W}}\mathbb{1}^T$, where $\overline{\boldsymbol{W}} = \boldsymbol{W} \cdot \mathbb{1}/N_s$ and $\mathbb{1} = (1, \ldots, 1)^T$.
Computing $\boldsymbol{W}_c$ costs $\mathcal{O}(k^2(N + N_s))$ (addition and truncation of rank-k matrices). By definition

$$\boldsymbol{C} = \frac{1}{N_s - 1} \boldsymbol{W}_c \boldsymbol{W}_c^T \approx \frac{1}{N_s - 1} \boldsymbol{U}_k \Sigma_k \Sigma_k^T \boldsymbol{U}_k^T. \tag{16}$$

The variance of the solution vector (i.e. the diagonal of the covariance matrix in (16)) can be computed with the complexity $\mathcal{O}(k^2(N_s + N))$.

Lemma 1. *Let $\|\boldsymbol{W} - \boldsymbol{W}_k\|_2 \le \varepsilon$, and $\overline{\boldsymbol{u}}_k$ be a rank-k approximation of the mean $\overline{\boldsymbol{u}}$. Then (a) $\|\overline{\boldsymbol{u}} - \overline{\boldsymbol{u}}_k\| \le \frac{\varepsilon}{\sqrt{N_s}}$, (b) $\|\boldsymbol{W}_c - (\boldsymbol{W}_c)_k\| \le \varepsilon$, c) $\|\boldsymbol{C} - \boldsymbol{C}_k\| \le \frac{1}{N_s - 1}\varepsilon^2$.*

Proof. Since $\overline{\boldsymbol{u}} = \frac{1}{N_s}\boldsymbol{W}\,\mathbb{1}$ and $\overline{\boldsymbol{u}}_k = \frac{1}{N_s}\boldsymbol{W}_k\,\mathbb{1}$, then

$$\|\overline{\boldsymbol{u}} - \overline{\boldsymbol{u}}_k\|_2 = \frac{1}{N_s}\|(\boldsymbol{W} - \boldsymbol{W}_k)\mathbb{1}\|_2 \le \frac{1}{N_s}\|(\boldsymbol{W} - \boldsymbol{W}_k)\|_2 \cdot \|\mathbb{1}\|_2 \le \frac{\varepsilon}{\sqrt{N_s}}.$$

Let $\boldsymbol{I} \in \mathbb{R}^{N_s \times N_s}$ be the identity matrix, then

$$\|\boldsymbol{W}_c - (\boldsymbol{W}_c)_k\|_2 \le \|\boldsymbol{W} - \boldsymbol{W}_k\|_2 \cdot \|\boldsymbol{I} - \frac{1}{N_s} \cdot \mathbb{1} \cdot \mathbb{1}^T\|_2 \le \varepsilon, \quad \text{and}$$

$$\begin{aligned} \|\boldsymbol{C} - \boldsymbol{C}_k\|_2 &\le \frac{1}{N_s - 1}\|\boldsymbol{W}_c \boldsymbol{W}_c^T - (\boldsymbol{W}_c)_k (\boldsymbol{W}_c)_k^T\|_2 \\ &= \frac{1}{N_s - 1}\|\boldsymbol{U}\,\Sigma\,\Sigma^T \boldsymbol{U}^T - \boldsymbol{U}_k \Sigma_k \Sigma_k^T \boldsymbol{U}_k^T\| \le \frac{1}{N_s - 1}\varepsilon^2. \end{aligned}$$

6 Numerical Tests

In this work we consider an example from aerodynamic, described by a system of Navier-Stokes equations with a k-ω turbulence model. Uncertainties in parameters

such as the angle of attack α and Mach number Ma are modelled by random variables (see details in [14, 16]):

$$\alpha(\theta_1, \theta_2) = \overline{\alpha} + \tilde{\alpha}(\theta_1, \theta_2), \quad \mathrm{Ma}(\theta_1, \theta_2) = \overline{\mathrm{Ma}} + \widetilde{\mathrm{Ma}}(\theta_1, \theta_2), \tag{17}$$

where θ_1 and θ_2 are Gaussian random variables, which model the components of the random fluctuations of the velocity vector, $\overline{\alpha}$ and $\overline{\mathrm{Ma}}$ are the mean values, and $\tilde{\alpha}(\theta_1, \theta_2)$ and $\widetilde{\mathrm{Ma}}(\theta_1, \theta_2)$ are the random fluctuations.

Uncertain output fields such as pressure, density, velocity, turbulence kinetic energy are modelled by random fields as well. The lift, drag and moments will be random variables. We will consider two cases: Case 1 ($\overline{\alpha} = 1.93$ and $\overline{\mathrm{Ma}} = 0.676$, no shock) and Case 9 ($\overline{\alpha} = 2.79$ and $\overline{\mathrm{Ma}} = 0.73$ with shock).

As the deterministic solver for the Navier-Stokes problem we use the DLR (German Aerospace Agency) TAU code. Our aim is the appropriate modelling of uncertainties and developing stochastic/statistical numerical techniques for further quantification of uncertainties. See, for instance, Fig. 1 which shows 5 % and 95 % quantiles of the (left) pressure (`cp`) and (right) the skin friction (`cf`) coefficients for RAE-2822 airfoil. With probability 90 % (`cp`) and (`cf`) belong to the interval between the lower and upper curves.

Figure 2 demonstrates ranks 5 (left) and rank 30 (right) approximation errors of the variance of the density. Both errors are smaller than, e.g. discretisation error or MC error.

To compute the mean value we use sparse Gauss-Hermite two-dimensional grids with 281 (Fig. 3 on the left) and 201 (Fig. 3 on the right) nodes. Figure 3 compares the mean pressure, obtained from 2,600 Monte-Carlo simulations with the mean pressure obtained from the sparse grid. We do this comparison for both cases—with shock (Case 9, right) and without (Case 1, left). Both errors are very small. Thus the response surface in Eq. 13 and Monte Carlo produce similar results.

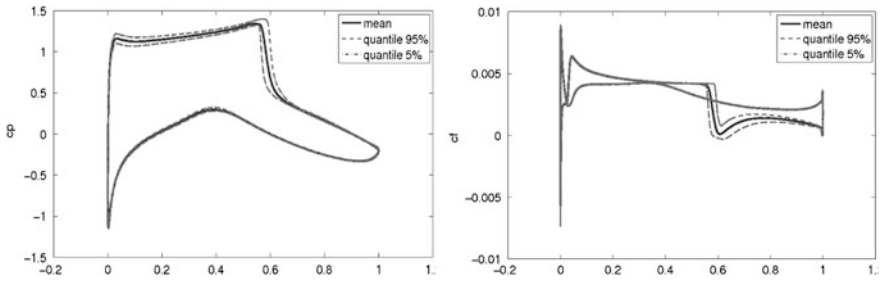


Fig. 1 5 % and 95 % quantiles for the (*left*) pressure coefficient cp; and (*right*) friction coefficients cf. Computations are done for the Case 9, RAE-2822 airfoil.

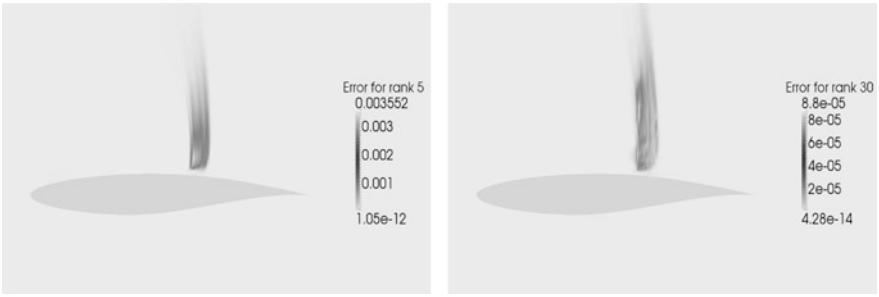


Fig. 2 (*Left*) Rank-k approximation errors $|\mathrm{var}(\rho) - \mathrm{var}(\rho)_5|$ and (*right*) $|\mathrm{var}(\rho) - \mathrm{var}(\rho)_{30}|$. The exact density ρ and its variance $\mathrm{var}(\rho)$ are computed from $\mathbf{W} \in \mathbb{R}^{65568\times 2600}$ (see Eq. 14), whereas $\mathrm{var}(\rho)_k$ is computed from $\mathbf{W}_k$. All computations are done for the Case 9 (with shock), RAE-2822 airfoil. The ranges of $\overline{\rho}$ and $\mathrm{var}(\rho)$ are shown in Fig. 5.

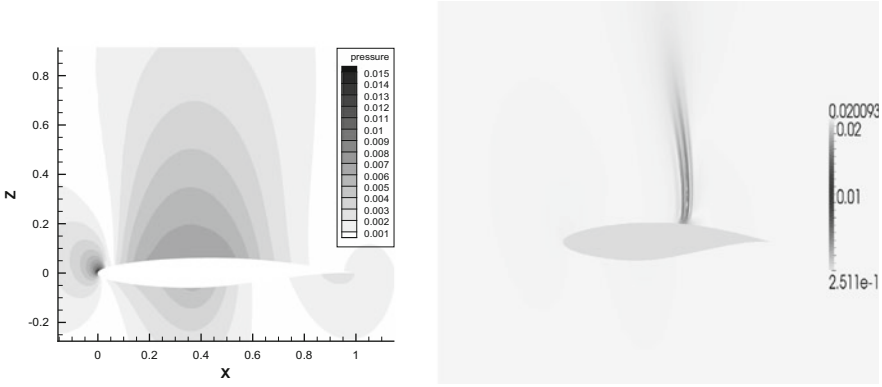


Fig. 3 Comparison of the mean pressures computed with PCE and with MC. (*Left*) $\Delta\bar{p} := |\bar{p}_{PCE281} - \bar{p}_{MC}|$, Case 1 without shock, (*right*) $\Delta\bar{p} := |\bar{p}_{PCE201} - \bar{p}_{MC}|$, Case 9 with shock.

6.1 Numerical Experiment with Sampling from the Response Surface

In this experiment, a sparse Gauss-Hermite two-dimensional grid [2, 6] is used for building the response surface. The response surface is the multivariate Hermite polynomial of order $p = 2$ with two random variables θ_1 and θ_2.

At first, we take 20 grid points (shown in Fig. 4, right) and compute TAU solutions in these points. This can be expensive, since one evaluation of the TAU code in one points requires already at least 10,000 TAU iterations.

In the next step, we evaluate the just constructed response surface in the 21st grid point and take the obtained value as the start value for TAU iterations. If this start value is good (meaning that the response surface gives a good initial approximation) then only few (may be none) TAU iterations are needed. If the start value is not good, we need again around 10,000 TAU iterations as above. After obtaining the

solution we update our response surface (see Sect. 4.1). The same procedure is repeated for all remaining sparse grid points. Figure 4 (left) shows relative errors in the Frobenius and the maximum norms for pressure and density computed in 10 points from a neigbourhood of $\overline{\alpha} = 1.93$ and $\overline{\text{Ma}} = 0.676$ (shown on the right). These relative errors are computed between solutions which we obtained after 10,000 TAU iterations without any starting and solutions which we obtained after only 1,000 TAU iterations with start values taken from the response surface (multivariate Hermite polynomials with $M = 2$ variables and of order $p = 2$). One can see that the errors are very small, i.e. the response surface produces a good approximation.

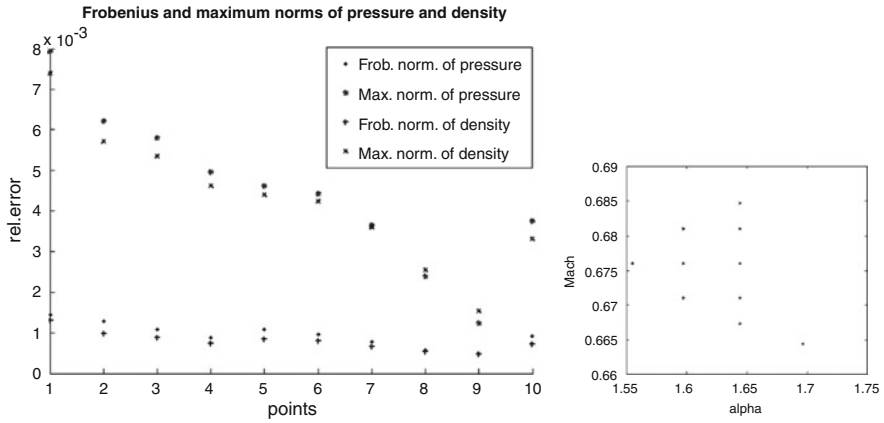


Fig. 4 (*Left*) Relative errors in the Frobenius and the maximum norms for pressure and density. (*Right*) 10 points (α, Ma) were chosen in the neigbourhood of $\overline{\alpha} = 1.93$ and $\overline{\text{Ma}} = 0.676$.

Figure 5 demonstrates the mean and the variance of the density for Case 9.

Figure 6 demonstrates the density evaluated from the TAU code (6,000 iterations) and from two different response surfaces ($p = 2$ and $p = 4$). Both response surfaces fail to produce a good result. Please note the increasing range of density—it does not reflect the physics (compare with Fig. 5). This effect is similar to the effect when one tries to approximate a step function by a polynomial—the amplitude of oscillations increases. Another negative effect which we observed during further iterations on the approximation obtained from the response surface is that the deterministic solver (TAU) may fail due to non-physical values after a few iterations. A possible reason is that some important solution values, obtained from the response surface, are out of the physical range and are non-realistic, or just that the approximation from the response surface is not a valid starting point for the iteration.

We also observed that RS produces a very poor solution in the discontinuous case (with shock). For instance, we obtained the wrong (and non-physical) solution range $(-6; 5)$, in contrast to the correct one $(0.6, 1.35)$. A possible remedy would be to compute the response surface for the log of the pressure.

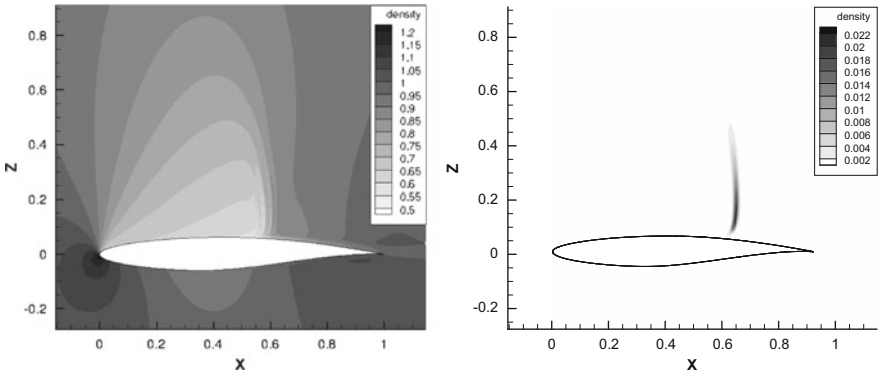


Fig. 5 (*Left*) The mean density and (*right*) variance of the density. Case 9, RAE-2822 airfoil.

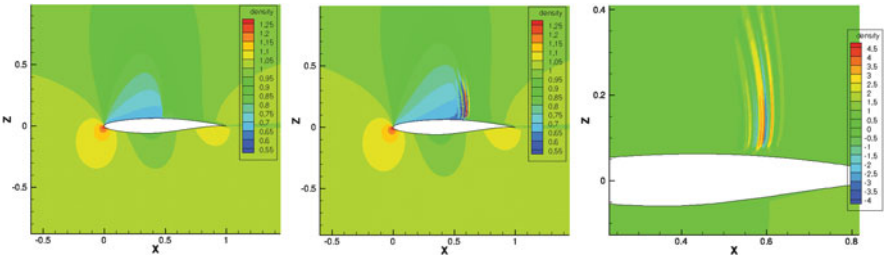


Fig. 6 Density computed from the TAU code (*left*) and from the response surfaces, Case 9, $Z = 201$. (*Centered*) $p = 2$ and (*right*) $p = 4$.

Thus, if the solution is smooth (e.g. as in Case 1) then the response surface produces a good starting point. In the case when the solution has a shock, the response surface produces a very poor starting point (Fig. 6) and further iterations do not help.

7 Conclusion

Stochastic calculations produce a huge amount of data which require a low-rank approximation (representation). The ansatz in low-rank tensor products reduces the numerical work, as well as the amount of storage for the solution and residuum. Low-rank approximation works in sampling and emulation as well as for non-linear problems and solvers. The set of random realisations (snapshots) can be approximated with a low rank k ($k = 5$ and $k = 30$ in Fig. 2). The numerical complexity is almost linear, i.e. $\mathcal{O}(k^2(N + N_s))$, where N is the number of degrees of freedom and N_s the number of realisations (simulations). One can successfully use PCE for the quantification of uncertainties (Fig. 3 compares MC with PCE).

PCE produces results which are similar to MC and requires a smaller number of deterministic computations. Using the response surface approximation as starting point for further iterations requires some care and possibly transformation to ensure physically valid values.

Acknowledgements We thank the German Ministry of Economics for the financial support of the project "MUNA—Management and Minimisation of Uncertainties and Errors in Numerical Aerodynamic" within the Luftfahrtforschungsprogramm IV under contract number 20A0604A. We thank also Nathalie Rauschmayr for Fig. 2.

References

1. Brand, M.: Fast low-rank modifications of the thin singular value decomposition. Linear Algebra Appl. **415**, 20–30 (2006).
2. Bungartz, H.J., Griebel, M.: Sparse grids. Acta Numer. **13**, 147–269 (2004)
3. Doostan, A., Iaccarino, G.: A least-squares approximation of partial differential equations with high-dimensional random inputs. J. Comput. Phys. **228**, 4332–4345 (2009)
4. Espig, M., Hackbusch, W., Litvinenko, A., Matthies, H.G., Wähnert, P.: Efficient low-rank approximation of the stochastic galerkin matrix in tensor formats. Comput. Math. Appl. (2012). doi:10.1016/j.camwa.2012.10.008
5. Espig, M., Hackbusch, W., Litvinenko, A., Matthies, H.G., Zander, E.: Efficient analysis of high dimensional data in tensor formats. In: Garcke, J., Griebel, M. (eds.) Sparse Grids and Applications. Lecture Notes in Computational Science and Engineering, vol. 88, pp. 31–56. Springer, Berlin/Heidelberg (2013)
6. Gerstner, T., Griebel, M.: Numerical integration using sparse grids. Numer. Algorithms **18**, 209–232 (1998)
7. Golub, G.H., Van Loan, C.F.: Matrix Computations, 3rd edn. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore (1996)
8. Hackbusch, W., Khoromskij, B.N.: Tensor-product approximation to operators and functions in high dimensions. J. Complexity **23**, 697–714 (2007)
9. Janson, S.: Gaussian Hilbert spaces. In: Cambridge Tracts in Mathematics, vol. 129. Cambridge University Press, Cambridge (1997)
10. Krée, P., Soize, C.: Mathematics of random phenomena. D. Reidel Publishing Co., Dordrecht (1986)
11. Kressner, D., Tobler, C.: Low-rank tensor Krylov subspace methods for parametrized linear systems. SIAM J. Matrix Anal. Appl. **32**, 1288–1316 (2011)
12. Krosche, M., Niekamp, R.: Low rank approximation in spectral stochastic finite element method with solution space adaption. Informatikbericht 2010–03, Technische Universität Braunschweig, Brunswick (2010). http://www.digibib.tu-bs.de/?docid=00036351
13. Le Maître, O.P., Knio, O.M.: Spectral methods for uncertainty quantification. In: Scientific Computation. Springer, New York (2010)
14. Litvinenko, A., Matthies, H.G.: Sparse data representation of random fields. In: Proceedings in Applied Mathematics and Mechanics. PAMM, vol. 9, pp. 587–588. Wiley-InterScience, Hoboken (2009)
15. Litvinenko, A., Matthies, H.G.: Low-rank data format for uncertainty quantification. In: Skiadas, Chr.H. (ed.) International Conference on Stochastic Modeling Techniques and Data Analysis Proceedings, pp. 477–484, Chania Crete (2010)
16. Litvinenko, A., Matthies, H.G.: Uncertainties quantification and data compression in numerical aerodynamics. Proc. Appl. Math. Mech. **11**, 877–878 (2011)

17. Matthies, H.G.: Uncertainty Quantification with Stochastic Finite Elements. Part 1. Fundamentals. Encyclopedia of Computational Mechanics. Wiley, Chichester (2007)
18. Matthies, H.G., Keese, A.: Galerkin methods for linear and nonlinear elliptic stochastic partial differential equations. Comput. Methods Appl. Mech. Engrg. **194**, 1295–1331 (2005)
19. Matthies, H.G., Zander, E.: Solving stochastic systems with low-rank tensor compression. Linear Algebra Appl. **436**, 3819–3838 (2012)
20. Mirsky, L.: Symmetric gauge functions and unitarily invariant norms. Q. J. Math. Oxf. Second Series(2) **11**, 50–59 (1960)
21. Nouy, A., Maître, O.L.: Generalized spectral decomposition for stochastic nonlinear problems. J. Comput. Phys. **228**, 202–235 (2009)
22. Oseledets, I.V., Savostyanov, D.V., Tyrtyshnikov, E.E.: Linear algebra for tensor problems. Computing **85**, 169–188 (2009)
23. Oseledets, I.V., Tyrtyshnikov, E.E.: Breaking the curse of dimensionality, or how to use SVD in many dimensions. SIAM J. Sci. Comput. **31**, 3744–3759 (2009)
24. Schwab, C., Gittelson, C.J.: Sparse tensor discretizations of high-dimensional parametric and stochastic PDEs. Acta Numer. **20**, 291–467 (2011)
25. Segal, I.E., Kunze, R.A.: Integrals and Operators. Springer, Berlin (1978)
26. Wiener, N.: The homogeneous chaos. Amer. J. Math. **60**, 897–936 (1938)

The Stochastic EM Algorithm for Censored Mixed Models

Ian C. Marschner

Abstract The Stochastic EM algorithm is a Monte Carlo method for approximating the regular EM algorithm in missing data situations where the E step is intractable. It produces a stationary Markov chain iterative sequence where each iteration is the result of applying complete data maximum likelihood estimation to a single simulation of the complete data conditional on the observed data. Analogously to other Markov chain Monte Carlo methods, the final estimate is the mean of the iterative sequence after a burn-in period. We consider a case study of the application of the Stochastic EM algorithm for censored mixed models, a computationally challenging context. The Stochastic EM algorithm is particularly simple to apply to either linear or non-linear mixed models with censoring. All that is required is a routine to simulate censored multivariate normal observations, and a routine to fit the desired uncensored mixed model. An application is presented involving repeated measures of HIV viral load subject to censoring caused by a lower detection limit of the assay. It is found that crude methods ignoring the censoring are biased compared to results from the Stochastic EM algorithm.

1 Introduction

The purpose of this paper is to provide a case study of the application of a straightforward Monte Carlo iterative technique for undertaking analyses in a computationally challenging context, mixed models with censoring.

Since mixed model analysis is standard when no censoring is present, it is natural to accommodate censoring by formulating the problem as a missing data problem and making use of the EM algorithm [5]. However, despite the applicability and

I.C. Marschner (✉)
Department of Statistics, Macquarie University, Sydney, NSW 2109, Australia
e-mail: ian.marschner@mq.edu.au

J. Dick et al. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2012*,
Springer Proceedings in Mathematics & Statistics 65,
DOI 10.1007/978-3-642-41095-6_28, © Springer-Verlag Berlin Heidelberg 2013

convenience of the EM algorithm in many missing data contexts, its usefulness is limited in contexts where either the conditional expectations required by the E step, or the maximisation required by the M step, are complicated or intractable. In the context of mixed models with censoring, the E step requires multi-dimensional numerical integration which can render the EM algorithm highly intensive and even infeasible in the presence of a reasonable level of censoring. Thus, while the EM approach is a natural one, modifications of the EM algorithm are required to make it generally applicable.

A general approach to dealing with an intractable E step in the EM algorithm is to carry out the so-called Monte Carlo EM algorithm (MCEM), which involves calculating the conditional expectations at the E step via simulation [19]. Thus, at each iteration of the EM algorithm, a large number of simulations of the missing data must be carried out, conditional on the observed data and the parameter value from the previous iteration. This approach has been suggested for censored linear mixed models by Hughes [9]. A drawback of the MCEM algorithm is its computational requirements which mandate, at each iteration, sufficiently many simulations to estimate the E step conditional expectations with high precision. In the present paper we make use of an approach to mixed model analysis of censored HIV viral load data which is computationally less intensive than MCEM, and is trivial to implement with the aid of standard software for fitting uncensored mixed models. An additional advantage of the approach is that, unlike MCEM, it is similarly convenient in non-linear contexts, although here we make use of it only in the context of linear models.

The approach we make use of is based on a modification of the EM algorithm known as the Stochastic EM (SEM) algorithm, which was introduced by Celeux and Diebolt [3] and described in detail by Diebolt and Ip [7]. Over the years this method has been used in a range of applications, including finite mixture and censoring problems [2,4], as well as various genetics and bioinformatics applications [10,18]. In the spirit of the MCEM algorithm, the SEM algorithm substitutes an intractable E step with a single simulation of the missing data conditional on the observed data and the parameter value from the previous iteration. The result of this scheme is a Markov chain, indexed by the iteration number, which can be averaged to produce a point estimate of the parameter. In contexts where the MCEM algorithm might be employed, the SEM algorithm will generally be computationally less burdensome because it requires only one simulation at each iteration. This, combined with the asymptotic equivalence of the SEM and EM estimates makes the SEM approach attractive in some contexts [7].

In the next section we consider the details of a laboratory assay data set which provides motivation for the use of mixed models with censored data. Subsequently we discuss the SEM algorithm and its implementation for fitting mixed models with censoring. A detailed analysis of the data set will then be presented in which the results of the SEM algorithm are assessed and compared with other approaches.

2 Motivating Application

HIV viral load, typically defined as the concentration of HIV RNA in plasma, is a useful prognostic tool for the management of HIV infection. In this context it is of some interest to assess patterns of change in HIV viral load subsequent to interventions such as the initiation or cessation of antiviral therapy. Since statistical models for repeated measures of HIV viral load over time need to accommodate substantial differences between individuals in rates of change and other parameters, mixed models are a natural tool for the analysis of repeated measures of HIV viral load. However, as with many laboratory assay contexts, a complicating feature of viral load data is that some measurements may be censored because the assay may have limits of detection beyond which measurement of the viral load is not possible. This motivates the need for mixed model analysis of censored data.

The data we consider in this paper are from a study of HIV-infected individuals who were initially treated with antiviral therapy so that their HIV viral load was suppressed below a predetermined threshold [8]. Subsequently these individuals experienced a loss of viral load suppression when part of the therapy was ceased. The goal of the analysis is to address the question of how the rate of increase in HIV viral load is related to certain subject-specific covariates, in order to obtain information about the manner in which HIV replicates in the body.

The present analysis involves 14 subjects for whom it could be verified that there was complete compliance with the demanding drug regimens over the course of the study. A plot of four subjects' data is given in Fig. 1; in this figure, and throughout the paper, HIV viral load is analysed on a $\log_{10}$ scale since changes tend to be linear and approximately normally distributed on this scale. In the data to be analysed, the average number of measurements per subject was 6.2, with a range of 4–8 measurements per subject, and 87 measurements in total. A discussion of the crude versus adjusted slope estimates in Fig. 1 will be deferred to Sect. 3.

A feature of the data presented in Fig. 1 is that some observations are known only to be less than a particular value, or left censored. In the full dataset to be analysed, the number of censored observations ranges from 1 to 5 per subject, with 40 % of all measurements being censored. Censoring arises because the virologic assays used to measure HIV viral load are subject to measurement limitations which impose a lower limit on the detectable concentration of HIV RNA. Thus, if the actual HIV viral load is less than the lower limit, it is left censored. The lower limits vary from assay run to assay run and are associated with factors other than the underlying HIV viral load value (assay reagents, technologist performance etc.). This last point implies that it is reasonable to assume non-informative censoring, and this will be done in the analysis presented later in the paper. As an aside we note that it is also possible that HIV viral load measurements are subject to upper limits of detection, leading to right censoring. The methods discussed in this paper can be applied in essentially the same way when the data are right censored (or even interval censored), however, the data analysed here were subject only to left censoring so we restrict our discussion to that context.

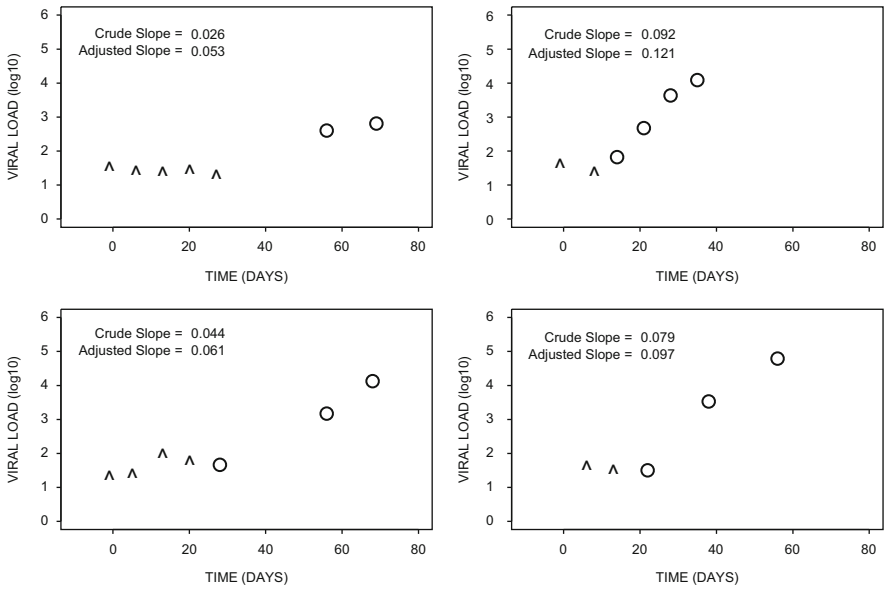


Fig. 1 Plots of HIV viral load measurements ($\log_{10}$ scale) for four subjects. *Circles* denote observed values, *arrows* denote left censoring. For each plot, the crude and adjusted slopes correspond to the slope estimates from a linear regression ignoring the censoring, and incorporating the censoring, respectively.

The association between HIV viral load changes and subject-specific factors will be addressed using the following covariates: (i) baseline HIV viral load prior to beginning therapy; (ii) CD4 cell count (a measure of the state of the immune system) prior to cessation of therapy; and (iii) the change in CD4 cell count during the initial period of therapy. In the original analysis of the data from this study it was observed that the probability of retaining suppression of HIV viral load is lower for those subjects having greater increases in CD4 cell count [8]; however, the result would be stronger if an association could be identified with the rate of increase in HIV viral load subject to loss of suppression. Thus, from a scientific point of view, the main goal of this analysis is to investigate whether such an association exists with change in CD4 count, as well as whether there is a lack of an association with the other subject-specific covariates. Although it is somewhat paradoxical that improving the immune system could lead to greater production of virus, the hypothesis is motivated by the presence of more targets for HIV infection (that is, more CD4 cells) in subjects having greater increases in CD4 cell count. An association between HIV viral replication and increases in CD4 cell count would be important because it would stress the need for early treatment prior to deterioration of the immune system, and may argue for caution in the application of treatment regimens which cause dramatic rises in CD4 count.

3 Models and Exploratory Analyses

Before describing a full analysis of the data, we consider some exploratory analyses based on separate models for each subject's data. As can be seen in Fig. 1, taking account of the fact that some observations are censored, linear increases in viral load are generally consistent with the data. Thus, a censored linear regression model can be used to estimate subject-specific rates of viral load increase and intercepts, for example using the LIFEREG procedure in the SAS software package [17]. The adjusted estimates in Fig. 1 were calculated in this manner, making an assumption of normal errors. An important point arising from Fig. 1 is that the censoring mechanism leads to censoring primarily among earlier observations, because these observations tend to be the smallest. Consequently, the censored observations tend to have high leverage in determining estimates of the rate of viral load increase. By comparing the crude estimates and the estimates adjusted for censoring in Fig. 1, it can be seen that ignoring the censoring tends to underestimate the slopes. This obviously leads to the potential for bias in the assessment of factors associated with the rate of viral load increase. The potential for bias when the censoring indicators are ignored will be further illustrated in the full analysis of Sect. 5.

Figure 2 contains a plot, versus various subject-specific quantities, of the estimated rates of increase in viral load for each of the 14 subjects, based on separate censored linear regression analyses with normal errors. The analyses displayed in Fig. 2 will be used below in motivating a linear mixed model for the data.

To describe an appropriate model for these data we use the following notation: let N be the total number of subjects and n_i be the number of observations taken on subject i; let t_{ij} be the jth time at which subject i is observed ($i = 1, \ldots, N$; $j = 1, \ldots, n_i$), and V_{ij} be the value of subject i's ($\log_{10}$-transformed) HIV viral load at time t_{ij}; let A_i and B_i be the values, for subject i, of the baseline HIV viral load and CD4 cell count prior to drug cessation, respectively; and finally, let C_i be the CD4 change ratio, defined as the CD4 cell count just prior to partial cessation of antiviral therapy divided by the baseline CD4 cell count. The quantity C_i measures the change in the CD4 cell count during the initial period of therapy. The exploratory analyses of Fig. 2a–c suggest that this covariate has the strongest association with rate of viral load increase.

A linear dependence of HIV viral load on time of measurement is suggested by the exploratory analyses in Fig. 1. In addition, a linear dependence of the subject-specific rates of change on the covariates A, B and C, is consistent with the exploratory analyses in Fig. 2a–c. Furthermore, subject-specific differences in both initial viral load and rates of change in viral load are evident in Fig. 2d. These considerations suggest the following model, which allows for subject-specific intercepts and slopes, and rates of change that depend, in linear fashion, on the three covariates of interest:

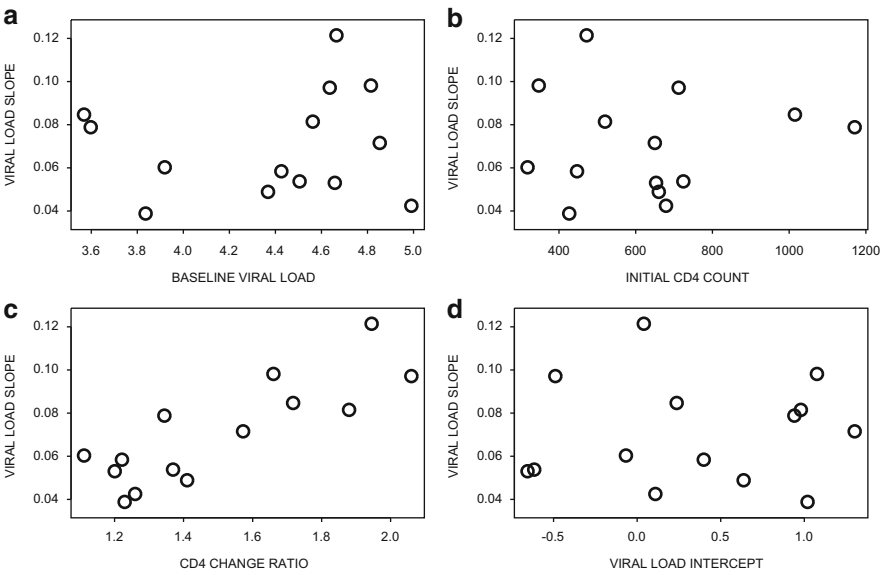


Fig. 2 Rate of viral load increase for each patient, obtained by fitting a censored linear regression model to each subject's data separately. Plotted against: (**a**) baseline viral load (A_i); (**b**) CD4 count prior to cessation of therapy (B_i); (**c**) CD4 change ratio (C_i); and (**d**) estimated viral load intercept obtained from the subject-specific censored regression analyses.

$$V_{ij} = \mu_i + g_i(A_i, B_i, C_i)t_{ij} + \epsilon_{ij} \tag{1}$$

where

$$g_i(A, B, C) = \alpha_i + \beta A + \gamma B + \delta C. \tag{2}$$

Random effects and mixed models have been used for some time in other contexts to model variation over subjects in HIV viral load parameters [12, 22]. In the present context, μ_i, α_i and ϵ_{ij} will be taken to be normally distributed random effects, which implies that (1) and (2) are equivalent to a linear mixed model. Thus, it is assumed that $\{\mu_i\}$ are i.i.d. $N(\mu, \sigma_\mu^2)$, $\{\alpha_i\}$ are i.i.d. $N(\alpha, \sigma_\alpha^2)$, and $\{\epsilon_{ij}\}$ are i.i.d. $N(0, \sigma_\epsilon^2)$. Because the SEM analysis in Sect. 5 will incorporate standard software for fitting uncensored mixed models, it is possible to include a correlation structure between the random effects. However, in view of the exploratory analyses in Fig. 2d, which show a clear lack of correlation between the subject-specific intercepts and slopes, our analysis will not include such a correlation structure; that is, the random effects μ_i, α_i and ϵ_{ij} will be assumed to be independent.

Observe that under the model specification in (1) and (2), the covariate effects are treated as fixed effects; that is, β, γ and δ do not vary over subjects. It is natural to consider whether to allow variation over subjects in the way viral load changes depend on the covariates, and thus to include them as random effects. In fact, such a model is not identifiable for these data, which can be seen in an intuitive way as follows, assuming we have just a single covariate C. It is possible to estimate a rate of change $\hat{g}_i$ for each subject, and hence to fit the model $\alpha_i + \delta C_i$ to the $\hat{g}_i$'s; β here is analogous to the gradient of the plot in Fig. 2c. However, it is not possible to fit the model $\alpha_i + \delta_i C_i$ to the $\hat{g}_i$'s; this would be analogous to fitting a separate gradient for each observation in Fig. 2c, which is obviously not identifiable. Thus, only a single baseline random effect will be used to model the subject-specific gradients, and this baseline gradient will be modified in fixed fashion by the covariate effects, as in Eq. (2).

For completely observed data $\{V_{ij}\}$, the mixed model specified by Eqs. (1) and (2) can be conveniently accommodated in a number of standard statistical software packages. However, in view of the discussion in Sect. 2, the observed data in the present context are potentially left censored, corresponding to $\{V_{ij}^*\}$ where $V_{ij}^* = \max(V_{ij}, L_{ij})$, and L_{ij} is the lower limit of detection for the assay run corresponding to the jth measurement on subject i. As discussed in Sect. 2, L_{ij} can be assumed to be independent of V_{ij}, so that the censoring is non-informative. In the next section, treating $\{V_{ij}\}$ as a complete data set, we discuss the SEM algorithm in order to conveniently fit the mixed model subject to censoring, making use of standard mixed model software at the M step, combined with a simulated E step to fill in the censored data.

4 Stochastic EM Algorithm

The EM algorithm is useful for maximising a log-likelihood $L(\theta; Y)$ to obtain the maximum likelihood estimate (MLE) of the parameter θ, when the observed data Y can be considered to be an incomplete version of an unobserved complete data set X [5, 14]. Given a current estimate $\hat{\theta}^{\text{old}}$ of θ, an iteration of the algorithm proceeds in two steps, the expectation (E) and maximisation (M) steps. The E step corresponds to calculation of

$$Q(\theta|\hat{\theta}^{\text{old}}) = E\left[L_c(\theta; X) \middle| Y; \hat{\theta}^{\text{old}} \right], \tag{3}$$

where L_c is the log-likelihood corresponding to the complete data X, and the M step corresponds to maximisation of $Q(\theta|\hat{\theta}^{\text{old}})$ over θ to produce the updated estimate $\hat{\theta}^{\text{new}}$. This process is continued until convergence, that is, until $\hat{\theta}^{\text{new}} = \hat{\theta}^{\text{old}}$.

The SEM algorithm, described in detail by Diebolt and Ip [7], replaces the E step above with a single simulation X^* of the complete data X, conditional on the

observed data Y and the current estimate $\hat{\theta}^{\text{old}}$. Thus, an SEM iteration consists of maximising $L_c(\theta; X^*)$. Let $\Theta(X)$ be the MLE for θ based on the complete data X, that is, the maximiser of $L_c(\theta; X)$. If we let $F_c(z|Y; \theta)$ be the distribution function for the random variable $\Theta(X)$, conditional on the observed data Y, then an SEM iteration is equivalent to

$$\hat{\theta}^{\text{new}} = \text{random draw from } F_c(z|Y; \hat{\theta}^{\text{old}}). \tag{4}$$

The updated estimate $\hat{\theta}^{\text{new}}$ is therefore an observation from a distribution that depends on the current estimate $\hat{\theta}^{\text{old}}$, meaning that the sequence of iterates arising from an SEM algorithm is a Markov chain. This sequence converges to a stationary distribution and, analogously to other Markov chain Monte Carlo (MCMC) techniques, after a sufficiently long burn-in period during which iterates are discarded, a point estimate of θ can be obtained by averaging the remaining sequence of SEM iterates. In many contexts it can be shown that the resulting SEM estimate is asymptotically equivalent to the EM estimate [7].

The SEM algorithm described above is useful when the Monte Carlo update (4) is easy and the deterministic E step (3) is hard. This is indeed the case for fitting mixed models with censoring. In this case (4) requires two straightforward steps at each iteration: (i) construction of a complete data set by simulating the censored measurements conditional on the observed data; and (ii) analysis of the complete data set by fitting an uncensored mixed model. Step (ii) is straightforward using existing software for mixed models, and requires a trivial amount of programming using standard packages. Step (i) is also straightforward, requiring simulation from a truncated multivariate normal distribution, which can be carried out using simple rejection/acceptance sampling, or using more efficient Gibbs sampling. Such simulation is straightforward to implement directly, but is also conveniently available in standard software [20].

The ease with which the SEM algorithm can be implemented contrasts with the deterministic EM algorithm. In this case, (3) involves multi-dimensional numerical integration over the censoring region $(-\infty, L_{i1}] \times \cdots \times (-\infty, L_{in_i}]$ for each individual i. When individuals have more than one censored observation, which is common in the application considered here, this can involve a large number of multi-dimensional integrations at each iteration which can be computationally prohibitive.

Implementation of the SEM algorithm requires choices for the initial parameter estimates, the burn-in period and the length of the MCMC sequence that will be averaged to produce the final estimates. Good initial estimates can speed up convergence of the MCMC sequence to a stationary distribution, and a simple method for providing good initial estimates is described below, which led to almost immediate stationarity. Using these initial estimates, a burn-in period of 50 iterations was used followed by a further 500 iterations which were averaged to give the final estimates. In view of the almost immediate stationarity, these were conservative but satisfactory choices. In view of the asymptotic equivalence of the SEM and EM algorithms, approximate standard errors for the fixed effects components of

the model can be obtained using a natural application of the method of Louis [11], formulas for which were given by Hughes [9] in the present context.

In order to provide good initial estimates for the SEM algorithm, a crude multiple imputation procedure was used. In Sect. 2 we described subject-specific estimates of viral load intercept and slope, calculated using censored linear regression analysis. These subject-specific estimates were used to simulate, or impute, each subject's censored measurements, with simulation of each censored measurement being carried out based on a univariate truncated normal distribution. The mixed model was then fitted to these uncensored data using a standard routine, leading to an imputed parameter estimate. One hundred such imputed estimates were obtained and averaged to give the multiple imputation estimate, which was then used as the initial estimate for the SEM algorithm. In practice, an even cruder calculation of the initial estimates is likely to be satisfactory, for example based on ignoring the censoring indicators altogether; however, it is of some interest to compare estimates from the stochastic algorithms with those from the crude multiple imputation approach, so the latter is presented below.

In the next section we discuss the results of the analysis produced by applying the above SEM procedure to the data described in Sect. 2.

5 Analysis Results

Before discussing the results in detail, we consider some initial SEM analyses assessing the importance of the various covariates. The full model specified in (1) and (2) allows the rate of viral load increase to depend on three covariates. When this full model was fitted, little evidence was found that the rate of viral load increase depends on the baseline viral load (A_i) or the CD4 count prior to cessation of drugs (B_i). In particular, the SEM estimates of the covariate effects in (2) were: for baseline viral load, $\hat{\beta} = -0.011$ and s.e.($\hat{\beta}$) $= 0.0090$ ($P = 0.22$); and for CD4 count prior to cessation of drugs, $\hat{\gamma} = -1.85 \times 10^{-5}$ and s.e.($\hat{\gamma}$) $= 1.84 \times 10^{-5}$ ($P = 0.32$). In contrast, consistent with the exploratory analyses in Fig. 2a–c, there was strong evidence that the rate of viral load increase depends on the CD4 change ratio (C_i), with $\hat{\delta} = 0.066$ and s.e.($\hat{\delta}$) $= 0.014$ ($P < 0.0001$). While it is prudent to interpret the asymptotic standard errors with caution, these results, taken together with the exploratory analyses in Fig. 2, suggest that it is only the CD4 change ratio which has a significant association with viral load change. Thus, the detailed results quoted below will make use of the model without the two insignificant covariates (A_i and B_i).

Table 1 contains estimates of the six parameters obtained using various methods. When the censoring indicators are ignored there is clearly great potential for bias, as shown by a comparison of the uncensored estimates with the SEM estimates. Understandably the between-subject variation parameter (σ_μ) is strongly underestimated because ignoring the censoring indicators has the effect of reducing

Table 1 Estimates for the parameters of the viral load mixed model specified in Eqs. (1) and (2), comparing the stochastic EM (SEM) procedure with two other methods of estimation. The uncensored analysis corresponds to ignoring the censoring indicators, while the imputed analysis corresponds to a multiple imputation procedure described in Sect. 4.

Method	$\hat{\mu}$	$\hat{\alpha}$	$\hat{\delta}$	$\hat{\sigma}_\mu$	$\hat{\sigma}_\alpha$	$\hat{\sigma}_\epsilon$
Uncensored	0.96	–0.026	0.054	0.21	0.012	0.35
Imputed	0.36	–0.024	0.061	0.62	0.0097	0.36
SEM	0.46	–0.025	0.061	0.57	0.0091	0.40

the variability of smaller observations, taken soon after cessation of therapy. The bias is substantially improved by carrying out the crude imputation procedure outlined in Sect. 4, however, there is still some bias evident, relative to the SEM procedure, for all parameters except α and δ.

For the three fixed effects (μ, α, δ), 95 % confidence intervals based on the SEM estimates are $[0.098, 0.81]$, $[-0.067, 0.018]$ and $[0.032, 0.090]$, respectively. The fact that δ is significantly greater than zero reflects a positive relationship between the rate of viral load increase and the increase in CD4 count prior to cessation of therapy. This result is consistent with the exploratory analyses summarized in Fig. 2c. The interpretation of this finding is that, paradoxically, the rate of viral load increase is greatest for individuals who had a better immune response to therapy. This is likely explained by the fact that a better immune response on therapy means that there are more targets for the virus if treatment is ceased. Such a finding could have implications for the way treatment is administered, particularly in relation to highlighting the need for a high level of treatment compliance.

It is of interest to investigate the extent of between subject variation in the rate of viral load increase that is explained by the differences in immune response. The variation between subjects in the rate of viral load increase is illustrated in Fig. 1, using the adjusted subject-specific slopes. When taken across all individuals in the sample, these adjusted slopes have a standard deviation of 0.023, compared to the variance component estimate $\hat{\sigma}_\alpha = 0.0091$ from Table 1. That is, adjustment for differences in the immune response explains approximately 86 % of the variation between subjects in the rate of viral load increase. When the analogous computations are carried out for the crude analysis which ignores the censoring, the variance component estimate is $\hat{\sigma}_\alpha = 0.012$ from Table 1. Compared with a standard deviation of 0.024 for the crude subject-specific slopes, this suggests that approximately 73 % of the variation between subjects in viral load increase is explained by the immune response. Thus, compared with the SEM analysis, the crude analysis substantially underestimates the importance of immune response in explaining the variability of viral load rebound rate.

Since the SEM estimates require averaging of stationary Markov chains, it is of interest to inspect the behaviour of these sequences. Figure 3 shows the SEM iterates for each of the six parameters, together with the final estimates as listed in Table 1. Beginning with the good initial estimates provided by the crude imputation

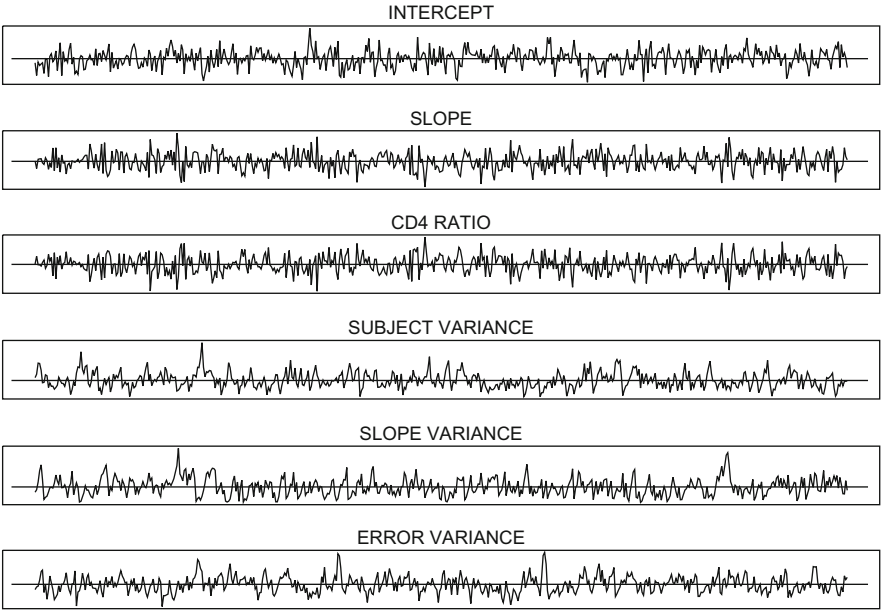


Fig. 3 Plots of SEM iterates for HIV viral load data set, showing stationarity of the Monte Carlo Markov chain sequences about the final estimates shown in Table 1 (*horizontal lines*). The 550 iterates are provided for each of the six parameters in the mixed effects model: intercept (μ), slope (α), CD4 ratio coefficient (δ), between-subject variance (σ^2_μ), between-subject slope variance (σ^2_α) and error variance (σ^2_ϵ).

procedure, it is seen that stationarity occurs almost immediately and that subsequent iterates vary stably about the final estimates.

In view of its stochastic nature, it is important to investigate the extent of stochastic variation across different runs of the SEM algorithm. The results of 50 replications of the SEM algorithm are displayed in Table 2, where each replication was implemented identically to the main analysis displayed in Table 1. It can be seen that the means of the SEM replications are virtually identical to the SEM estimates in Table 1. Furthermore, the standard deviations of the SEM replications are extremely small relative to the magnitude of the estimates. Indeed, compared to the sampling variation summarised in the confidence intervals discussed earlier in this section, the stochastic variation in the SEM algorithm is negligible and inconsequential. It can also be seen that the difference between the SEM estimates and the crude estimates displayed in Table 1 are greater than could be explained by the negligible stochastic variation in the SEM estimates. This indicates that these differences are genuine, rather than simply reflecting stochastic variation in the SEM algorithm. Additional analyses conducted with other choices for the burn-in period and the length of the averaged MCMC iterative sequence were also highly consistent with the analyses displayed in Table 1.

Table 2 Means and standard deviations of the parameter estimates from 50 identical replications of the stochastic EM (SEM) algorithm applied to the viral load data. Each replication of the SEM algorithm was implemented identically to the primary analysis displayed in Table 1.

Summary	$\hat{\mu}$	$\hat{\alpha}$	$\hat{\delta}$	$\hat{\sigma}_\mu$	$\hat{\sigma}_\alpha$	$\hat{\sigma}_\epsilon$
Mean	0.46	–0.025	0.061	0.56	0.0092	0.40
Standard deviation	0.0049	0.00026	0.00015	0.0077	0.00012	0.0019

Although it is preferable to start with good initial estimates, in this application the SEM algorithm is insensitive to the initialisation and always converges rapidly to a stationary Markov chain centred around the estimates in Table 1. This behaviour is illustrated in Fig. 4, which shows the burn-in period of the SEM iterative sequences beginning with initial estimates that correspond to an implausible special case of the linear mixed model, namely, independent and identically distributed viral load measurements with unit mean and variance. It is seen that the sequences reach stability about the final estimates within ten iterations, even for the parameters that have very poor initial estimates. Similar behaviour arises for other choices of the initial estimates.

6 Discussion

Adjustment for censoring when fitting mixed models is not a trivial problem but is important to avoid bias resulting from more crude methods of analysis. The present paper illustrates a straightforward approach which can be implemented easily in routine data analyses, as illustrated using a case study on changes in HIV viral load subsequent to partial cessation of antiviral therapy. In this application, the analyses allowed the rate of increase in viral load to depend on a number of subject-specific covariates and pointed to a strong association between the rate of viral load increase and the increase in CD4 count prior to cessation of therapy, a finding that would potentially have implications for the way treatment is administered. Furthermore, the models fitted provide information about parameters governing the dynamics of viral load increases over time, and the mixed effects approach allows quantification of the extent of variation between subjects with respect to these parameters. Importantly, in the analyses presented there were substantial biases in both the parameter estimates and the extent of variation explained by covariates when the analysis method did not take account of the censoring.

The SEM algorithm that we have used here to carry out the adjustment for censoring is a convenient tool in any context where a Monte Carlo E step might be employed. A particularly useful feature of the SEM algorithm for censored mixed models is that it can be straightforwardly applied in the context of non-linear models. All that is required is a standard routine for fitting non-linear mixed effects models for uncensored data, and then the same approach can be used as that described in Sect. 4. On the other hand, accommodation of non-linear models

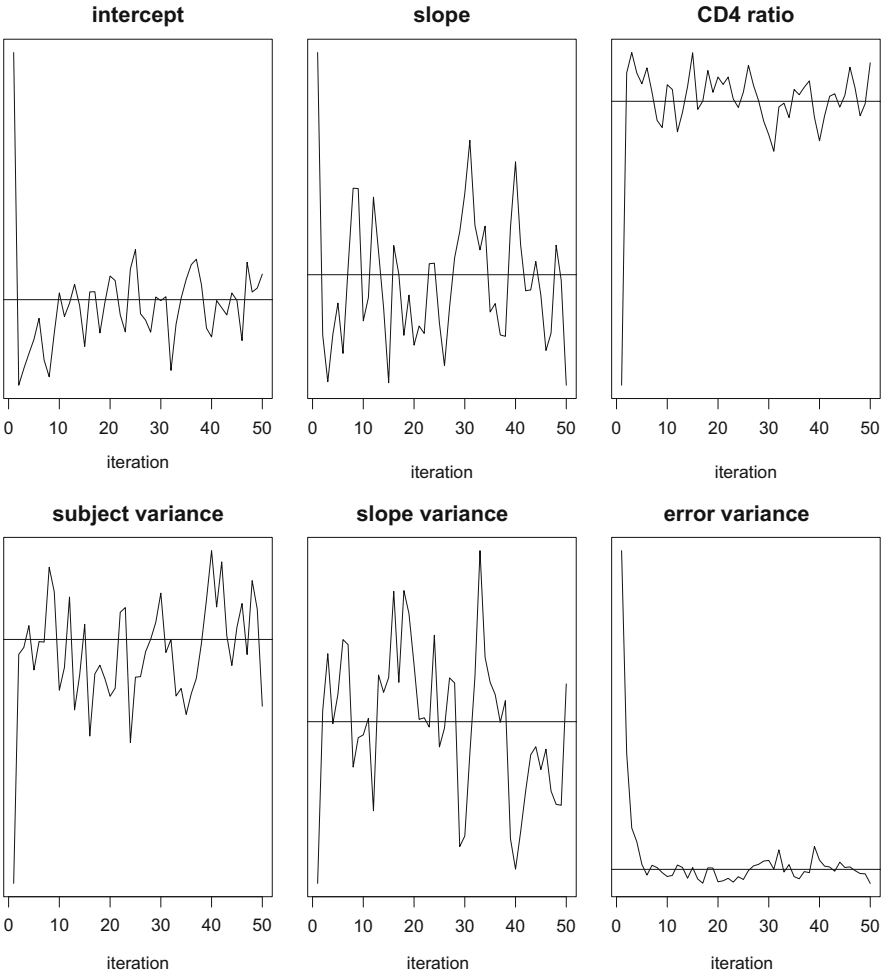


Fig. 4 Burn-in periods for SEM iterates with poor initial estimates showing rapid convergence of the Monte Carlo Markov chain sequences to stationarity about the final estimates shown in Table 1 (*horizontal lines*). The poor initial estimates correspond to independent and identically distributed viral load observations with unit mean and unit variance, $(\mu, \alpha, \delta, \sigma_\mu^2, \sigma_\alpha^2, \sigma_\epsilon^2) = (1, 0, 0, 0, 0, 1)$.

in an EM or MCEM framework is less straightforward [21]. Other Monte Carlo based algorithms have also been useful in this context to overcome the complexities of full MCEM, including the so-called stochastic approximation EM algorithm [6, 16], which makes use of the technique of stochastic approximation to deal with intractable expectations during the E step.

Censoring in both linear and non-linear mixed effects models can also be handled under a Bayesian formulation of the model with conventional MCMC methods used for model fitting [1]. While this may be useful in that it would provide full posterior

distributions for the parameters of interest, it would also be subject to considerably increased computational burden, particularly given convergence in Bayesian models is typically much slower than the very rapid stationarity achieved by the SEM sequence.

Enhancement to the SEM algorithm implemented here may also be useful. For example Nielsen [15] studied the use of more than one simulation during each iteration, an approach that is intermediate to the SEM and MCEM algorithms, while Marschner [13] proposed a general correction to the SEM algorithm that can improve its performance in small samples. While these and other enhancements of the basic SEM algorithm may be worthy of further study in the context of censored mixed models, overall the approach illustrated here offers a straightforward way to accommodate censoring in mixed models, and is amenable to routine data analysis using only standard software and straightforward Monte Carlo methods.

References

1. Bennett, J.E., Racine-Poon, A., Wakefield, J.C.: MCMC for non-linear hierarchical models. In: Gilks, W.R., Richardson, S., Spiegelhalter, D.J. (eds.) Markov Chain Monte Carlo in Practice, pp. 339–357. Chapman and Hall, London (1996)
2. Bordes, L., Chaveau, D., Vandekerkhove, P.: A stochastic EM algorithm for a semiparametric mixture model. Comput. Stat. Data Anal. **51**, 5429–5443 (2007)
3. Celeux, G., Diebolt, J.: The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. Comput. Stat. Q. **2**, 73–82 (1985)
4. Chaveau, D.: A stochastic EM algorithm for mixtures with censored data. J. Stat. Plan. Inference **46**, 1–25 (1995)
5. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. R. Stat. Soc. Ser. B **39**, 1–38 (1977)
6. Deylon, B., Lavielle, M., Moulines, E.: Convergence of a stochastic approximation version of the EM algorithm. Ann. Statist. **27**, 94–128 (1999)
7. Diebolt, J., Ip, E.H.S.: Stochastic EM: methods and applications. In: Gilks, W.R., Richardson, S., Spiegelhalter, D.J. (eds.) Markov Chain Monte Carlo in Practice, pp. 259–273. Chapman and Hall, London (1996)
8. Havlir, D.V., Marschner, I.C., Hirsch, M.S., Collier, A.C., Tebas, P., Bassett, R.L., Ioannidis, J.P.A., Holohan, M.K., Leavitt, R., Boone, G., Richman, D.D.: Maintenance antiretroviral therapy in HIV infected patients who have achieved undetectable plasma HIV RNA with triple combination therapy. New England Journal of Medicine **339**, 1261–1268 (1998)
9. Hughes, J.P.: Mixed effects models with censored data with application to HIV RNA levels. Biometrics **55**, 625–629 (1999)
10. Liu, F.: A Bayesian hierarchical model for high-dimensional meta-analysis. In: Bang, H., Zhou, X.K., Van Epps, H.L., Mazumdar, M. (eds.) Statistical Methods in Molecular Biology, pp. 531–539. Springer, New York (2010)
11. Louis, T.A.: Finding the observed information matrix when using the EM algorithm. J. R. Stat. Soc. Ser. B **44**, 226–233 (1982)
12. Marschner, I.C.: Design of HIV viral dynamics studies. Stat. Med. **17**, 2421–2434 (1998)
13. Marschner, I.C.: On stochastic versions of the EM algorithm. Biometrika **88**, 281–286 (2001)
14. McLachlan, G.J., Krishnan, T.: The EM Algorithm and Extensions, 2nd edn. Wiley, Hoboken (2008)

15. Nielsen, S.F.: The stochastic EM algorithm: estimation and asymptotic results. Bernoulli **6**, 457–489 (2000)
16. Samson, A., Lavielle, M., Mentre, F.: Extension of the SAEM algorithm to left-censored data in non-linear mixed-effects model: application to HIV dynamics model. Comput. Stat. Data Anal. **51**, 1562–1574 (2006)
17. SAS Institute Inc.: SAS/STAT 9.2 User's Guide. Chapter 48, The LIFEREG Procedure. SAS Institute Inc., Cary (2008)
18. Tregouet, D.A., Escolano, S., Tiret, L., Mallett, A., Golmard, J.L.: A new algorithm for haplotype-based association analysis: the stochastic EM algorithm. Ann. Hum. Genet. **68**, 165–177 (2004)
19. Wei, G.C.G., Tanner, M.A.: A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. J. Amer. Statist. Assoc. **85**, 699–704 (1990)
20. Wilhelm, S., Manjunath, B.G.: tmvtnorm: truncated multivariate normal and t distributions. R package version 1.4–4. http://CRAN.R-project.org/package=tmvtnorm (2012)
21. Wu, L.: A joint model for non-linear mixed-effects models with censoring and covariates measured with error, with application to AIDS studies. J. Amer. Statist. Assoc. **97**, 955–964 (2002)
22. Wu, H., Ding, A.A., DeGruttola, V.: Estimation of HIV dynamic parameters. Stat. Med. **17**, 2463–2485 (1998)

Existence of Higher Order Convergent Quasi-Monte Carlo Rules via Walsh Figure of Merit

Makoto Matsumoto and Takehito Yoshiki

Abstract The Walsh figure of merit WAFOM(P) is a quality measure of point sets $P \subset [0,1)^S$ in the S-dimensional unit cube for quasi-Monte Carlo integration constructed by a digital net method with n-bit precision over the two element field. We prove that there are explicit constants E, C, D such that for any $d \geq 9S$ and n, there is a point set P of size $N := 2^d$ with WAFOM(P) $\leq E \cdot 2^{-Cd^2/S+Dd} = E \cdot N^{-C(\log_2 N)/S+D}$, by bounding WAFOM($P$) by the minimum Dick-weight of $P^\perp$, and by proving the existence of point sets with large minimum Dick-weight by a probabilistic argument.

1 WAFOM and Its Background

Let S, n be positive integers. Let $\mathbb{F}_2 = \{0,1\}$ denote the two element field. Let $V := V_{S,n}$ denote the set of $S \times n$ matrices with coefficients in $\mathbb{F}_2$. For $A = (a_{ij}) \in V$, we define a non-negative integer

$$\mu(A) := \sum_{1 \leq T \leq S, 1 \leq j \leq n} j \times a_{T,j}, \tag{1}$$

where each $a_{T,j} \in \{0,1\}$ is considered as an integer, not an element of $\mathbb{F}_2$. If $n = 1$, then A is a column vector, and $\mu(A)$ is the Hamming weight of A. From this view

M. Matsumoto (✉)
Department of Mathematics, Graduate School of Science, Hiroshima University, Higashi-Hiroshima, Japan
e-mail: m-mat@math.sci.hiroshima-u.ac.jp

T. Yoshiki
Graduate School of Mathematical Sciences, University of Tokyo, Tokyo 153-8914, Japan
e-mail: yosiki@ms.u-tokyo.ac.jp

J. Dick et al. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2012*,
Springer Proceedings in Mathematics & Statistics 65,
DOI 10.1007/978-3-642-41095-6_29, © Springer-Verlag Berlin Heidelberg 2013

point, we may call $\mu(A)$ a *Dick weight* of A since it is a special case (where $n = \alpha$) of μ_α introduced by Dick in [2] (see also a comprehensive book [5], to which we refer for all the basic terminology), or, we may call it simply the weight of A in this article.

Let $P \subset V$ be an $\mathbb{F}_2$-linear subspace of V. We may consider P as a point set in $[0,1)^S$ by the digital net method (see Niederreiter's book [8]), but we will choose a slightly different formulation in the following. We define the *Walsh figure of merit* [7] of P by

Definition 1 (WAFOM).

$$\mathrm{WAFOM}(P) := \sum_{A \in P^\perp \setminus \{0\}} 2^{-\mu(A)},$$

where $P^\perp$ denotes the orthogonal space of P in V with respect to the standard inner product on V, namely, $A \cdot B = (a_{i,j}) \cdot (b_{i,j}) = \sum_{i,j} a_{i,j} b_{i,j} \in \mathbb{F}_2$.

Remark 1. This figure of merit is based on the decay of Walsh coefficients of smooth functions proved by Dick (see a survey [3, Sects. 4.1 and 4.2]), and hence *Walsh* in its name.

Theorem 1. *There are positive constants C, D and E such that for any positive integers S, n and $d \geq 9S$, there exists a linear subspace $P \subset V_{S,n}$ of dimension d satisfying*

$$\mathrm{WAFOM}(P) \leq E \cdot 2^{-Cd^2/S + Dd}.$$

We explain the background briefly (see [7, Sect. 2] for detail). A matrix $B \in V_{S,n}$ has S rows. Each row is n-digit of bits, and by identifying with the corresponding 2-adic fraction, each row gives a real number in $[0,1)$, and an elementary 2-adic interval of length 2^{-n}. More precisely, we identify an element $\mathbf{b} = (b_1, \ldots, b_n) \in \mathbb{F}_2^n$ with a real number $\mathbf{b} := \sum_{i=1}^n b_i 2^{-i}$ and then with the interval $[\mathbf{b}, \mathbf{b} + 2^{-n})$ (these intervals are called 2-adic elementary interval of length 2^{-n}). Thus, $\mathbb{F}_2^n$ is identified with the set of 2^n such intervals. Then, $(\mathbb{F}_2^n)^S$ is identified with the set of $(2^n)^S$ of 2-adic elementary cubes in $[0,1)^S$ represented as the S-dimensional Cartesian products of these intervals. Thus, each $B \in V$ represents a 2-adic elementary *cube*, which we shall denote $\mathbf{I}_B \subset [0,1)^S$.

Let $f : [0,1]^S \to \mathbb{R}$ be a Riemann integrable function. We define its *n-digit discretization* $f_n : V \to \mathbb{R}$ by

$$f_n(B) := \frac{1}{\mathrm{Vol}(\mathbf{I}_B)} \int_{\mathbf{I}_B} f(\mathbf{x}) d\mathbf{x},$$

namely, $f_n(B)$ is the average of the value of f over $\mathbf{I}_B$. Then, the integration equals to the discrete sum, i.e.,

$$\int_{[0,1)^S} f(\mathbf{x})d\mathbf{x} = \frac{1}{\#(V)} \sum_{B \in V} f_n(B).$$

(Here $\#(F)$ denotes the cardinality of finite set F.) We may approximate $f_n(B)$ by the value of f at the center of the cube $\mathbf{I}_B$, for example. Then, under Lipschitz continuity of f, as $n \to \infty$ and the cube collapses to its center, the approximation error of the integration when $f_n(B)$ in the right sum is replaced with the value of f sampled from the cube B is $O(C_f \sqrt{S} 2^{-n})$ in n, where C_f is the Lipschitz constant (see [7], this is the discretization error). We assume that n is taken to be large enough, so that the precision required in the particular computation is less than or equal to n binary digits. Practically, we may take n large so that the above discretization error is small enough for our particular purposes.

Let $P \subset V$. The discretized quasi-Monte Carlo integration of a function $f : [0,1)^S \to \mathbb{R}$ by the point set P is an approximation of $\int_{[0,1)^S} f(\mathbf{x})d\mathbf{x}$ by the average $\frac{1}{\#(P)} \sum_{B \in P} f_n(B)$ (see [7, (2.3)]).

We have a bound on the integration error [7, (3.7)]

$$\left| \int_{[0,1)^S} f(\mathbf{x})d\mathbf{x} - \frac{1}{\#(P)} \sum_{B \in P} f_n(B) \right| \le c_{S,n} ||f||_n \times \mathrm{WAFOM}(P),$$

where $||f||_n$ is a norm of f defined in [2,5, Chap. 14.6] and $c_{S,n}$ is a constant independent of f and P. Thus, finding point sets with the smaller value of $\mathrm{WAFOM}(P)$ implies the smaller error bound. Dick [2] and Dick and Pillichshammer [5] showed that for arbitrary $\alpha > 0$, there is a construction of a family of point sets with error bound $O(N^{-\alpha}(\log N)^{S\alpha})$, where N is the size of the point set under the assumption of α-smoothness (in the classical, non-digital sense) on the function class.

Our result below can be written as $\mathrm{WAFOM}(P) = O(N^{-C(\log_2 N)/S+D})$. These results go in a similar direction, but there is no implication between them; Dick fixed the smoothness α and gave a construction of series of point sets with error bound $C_\alpha N^{-\alpha}(\log N)^{S\alpha}$, where the constant C_α depends on α. While our method requires n-smoothness on the function for n being as above. Thus, in our case, the function class is getting smaller for n being increased. On the other hand, we do not get an implied constant depending on n. Although our method is not constructive, the existence of point sets P for every $d \ge 9S$ has been shown using probabilistic arguments. Further, a random search has successfully been performed in [7, Sect. 5].

2 Proof

2.1 Geometry of Weight and Enumeration

Recall that with the weight function $\mu : V \to \mathbb{N}_0$ defined in (1), V is a metric space, by defining

$$d(A, B) := \mu(A - B) \text{ for } A, B \in V.$$

It is easy to show that this is a metric. We use this metric from now on. We remark that V is not distance-transitive, differently from the case of the Hamming weight. Namely, for two given pairs $(A, B), (A', B')$ with $d(A, B) = d(A', B')$, there may be no $\mathbb{F}_2$-affine isometry $f : V \to V$ that maps A to A' and B to B'.

Let us define

$$\mathscr{S}_{S,n}(m) := \{A \in V_{S,n} \mid \mu(A) = m\}, \tag{2}$$

$$s_{S,n}(m) := \#(\mathscr{S}_{S,n}(m)). \tag{3}$$

Note that $\mathscr{S}_{S,n}(m)$ is the sphere in V with center 0 and radius m, and $s_{S,n}(m)$ is its cardinality. A combinatorial interpretation of $s_{S,n}(m)$ is as follows. One has coins with values $1, 2, \ldots, n$. For each value i, one has exactly S labeled coins with value i. Then, $s_{S,n}(m)$ is the number of ways to pay m by using these coins.

Another equivalent definition is in the following identity:

$$\prod_{k=1}^{n}(1 + x^k)^S = \sum_{m=0}^{\infty} s_{S,n}(m)x^m. \tag{4}$$

Note that the right hand side is a finite sum. It is easy to see that $s_{S,n}(m)$ is monotonically increasing with respect to S and n, and $s_{S,m}(m) = s_{S,m+1}(m) = s_{S,m+2}(m) = \cdots$ holds.

Definition 2. $s_S(m) := s_{S,m}(m)$.

We have the following identity between formal power series:

$$\prod_{k=1}^{\infty}(1 + x^k)^S = \sum_{m=0}^{\infty} s_S(m)x^m. \tag{5}$$

For any positive integer M, we define

$$\mathscr{B}_{S,n}(M) := \cup_{m=0}^{M}\mathscr{S}_{S,n}(m), \quad \mathrm{vol}_{S,n}(M) := \#(\mathscr{B}_{S,n}(M)). \tag{6}$$

Note that $\mathscr{B}_{S,n}(M)$ is the ball in V with center 0 and radius M, and $\mathrm{vol}_{S,n}(M)$ is its cardinality. Since $\mathrm{vol}_{S,n}(M) = \sum_{m=0}^{M} s_{S,n}(m)$, $\mathrm{vol}_{S,n}(M)$ inherits properties of $s_{S,n}(m)$, namely, it is monotonically increasing with respect to S and n, and has the stability property

$$\mathrm{vol}_S(M) := \mathrm{vol}_{S,M}(M) = \mathrm{vol}_{S,M+1}(M) = \cdots.$$

2.2 Combinatorial Inequalities on Volumes of Spheres

Lemma 1.

$$s_{S,n}(m) \le s_S(m) < e^{2\sqrt{Sm}}$$

Proof. We have already seen the first inequality. We prove the next inequality along [6, Exercise 3(b), p. 332], which treats only $S = 1$ case. Define a polynomial with non-negative integer coefficients by

$$f_{S,m}(x) := \prod_{k=1}^{m} (1 + x^k)^S.$$

From (4) and Definition 2, $s_S(m)$ is the coefficient of x^m in $f_{S,m}(x)$.

Since $f_{S,m}(x)$ has only non-negative coefficients, by dropping all terms except the last in (4)

$$s_S(m) \le f_{S,m}(x)/x^m \quad (x \in (0,1)) \tag{7}$$

holds. Taking $\log := \log_e$ of the both sides, well-known bounds $\log(1 + x^k) < x^k$ for $0 < x$ give

$$\begin{aligned} \log s_S(m) &\le S \sum_{k=1}^{m} \log(1 + x^k) + m \log(1/x) \\ &< S \sum_{k=1}^{m} x^k + m \log(1 + (1/x - 1)) \\ &< Sx/(1-x) + m(1-x)/x. \end{aligned}$$

Here, we can take any value of $x \in (0,1)$. By comparison of the arithmetic mean and the geometric mean, the last expression attains the minimum value $2\sqrt{Sm}$ when $Sx/(1-x) = m(1-x)/x$, namely, $(1-x)/x = \sqrt{S/m}$. A direct computation shows that $x = (\sqrt{S/m} + 1)^{-1} \in (0,1)$ attains this equality. □

Remark 2. The function $\prod_{k=1}^{\infty}(1+x^k)^S$ is an automorphic form, and the magnitude of the coefficients $s_S(m) \sim (\frac{1}{\sqrt{2}})^{S+3}(\frac{S}{3})^{\frac{1}{4}} m^{-\frac{3}{4}} e^{\frac{\pi}{\sqrt{3}}\sqrt{Sm}}$ $(m \to \infty)$ can be proved by using a method in [1, Chap. 6, Theorem 6.2], which may improve the following estimation. Here, $f(m) \sim g(m)$ means that $f(m)/g(m) \to 1$ when $m \to \infty$.

Lemma 2 (Volume of balls). *For a positive integer M, we have*

$$\mathrm{vol}_{S,n}(M-1) \le \mathrm{vol}_S(M-1) \le \sqrt{M/S}\, e^{2\sqrt{SM}}.$$

Proof. It is an easy exercise that $s_S(m)$ is monotonically increasing with respect to m. Thus, we have

$$\begin{aligned}
\mathrm{vol}_S(M-1) = \sum_{m=0}^{M-1} s_S(m) &\le \sum_{m=0}^{M-1} e^{2\sqrt{Sm}} \\
&\le \int_{m=0}^{M} e^{2\sqrt{Sm}} dm \\
&= \int_{x=0}^{4SM} e^{\sqrt{x}} \frac{dx}{4S} \\
&= \frac{1}{4S}\left[2\sqrt{x}e^{\sqrt{x}} - 2e^{\sqrt{x}}\right]_{x=0}^{x=4SM} \\
&= \frac{1}{4S}\left[4\sqrt{SM}e^{2\sqrt{SM}} - 2e^{2\sqrt{SM}} + 2\right] \\
&< \frac{1}{4S}\left[4\sqrt{SM}e^{2\sqrt{SM}}\right] \\
&= \sqrt{M/S}e^{2\sqrt{SM}},
\end{aligned}$$

which completes the proof. □

2.3 Bounding WAFOM by the Minimum Weight

Let $P \subset V$ be a d-dimensional subspace.

Definition 3. The minimum weight of $P^\perp$ is defined by

$$\delta_{P^\perp} := \min_{A \in P^\perp \setminus \{0\}} \mu(A).$$

Thus, $\delta_{\{0\}} = +\infty$.

We have the following equivalence:

$$\delta_{P^\perp} \ge M \Leftrightarrow P^\perp \cap \mathscr{B}_{S,n}(M-1) = \{0\}.$$

The next easy lemma bounds WAFOM(P) by the minimum weight of $P^\perp$.

Lemma 3. *For an integer M, define*

$$C_{S,n}(M) := \sum_{A \in V_{S,n} \setminus \mathscr{B}_{S,n}(M-1)} 2^{-\mu(A)} = \sum_{m=M}^{\infty} s_{S,n}(m) 2^{-m}.$$

Then we have

$$\mathrm{WAFOM}(P) = \sum_{A \in P^{\perp} \setminus \{0\}} 2^{-\mu(A)} < \quad C_{S,n}(\delta_{P^{\perp}}).$$

Proof. The result holds because $\delta_{P^{\perp}}$ is the maximum M satisfying $P^{\perp} \setminus \{0\} \subset V_{S,n} \setminus \mathscr{B}_{S,n}(M-1)$. □

We shall estimate $C_{S,n}(M)$ (C for the Complement of the ball) for rather general M: from Lemma 1 it follows that

$$C_{S,n}(M) = \sum_{m=M}^{\infty} s_{S,n}(m) 2^{-m} \le \sum_{m=M}^{\infty} s_S(m) 2^{-m} \le \sum_{m=M}^{\infty} e^{2\sqrt{Sm}} 2^{-m}. \tag{8}$$

The sequence $e^{2\sqrt{Sm}} 2^{-m}$ ($m \ge M$) is decreasing if

$$e^{2\sqrt{S(M+1)} - 2\sqrt{SM}}/2 \le 1.$$

The left hand side is

$$e^{2\sqrt{S}(\sqrt{M+1}-\sqrt{M})}/2 = e^{2\sqrt{S}\left(\frac{1}{\sqrt{M+1}+\sqrt{M}}\right)}/2 < e^{\sqrt{\frac{S}{M}}}/2,$$

and hence if $\log 2 > \sqrt{S/M}$ or equivalently $M > (\log 2)^{-2} S$. Under this condition, we have

$$\begin{aligned} C_{S,n}(M+1) &\le \sum_{m=M+1}^{\infty} e^{2\sqrt{Sm}} 2^{-m} \le \int_{m=M}^{\infty} e^{2\sqrt{Sm}} e^{-m \log 2} dm \\ &= \int_{m=M}^{\infty} e^{-(\log 2)(\sqrt{m} - \frac{\sqrt{S}}{\log 2})^2 + \frac{S}{\log 2}} dm \\ &= \int_{x=\sqrt{M}}^{\infty} e^{-(\log 2)(x - \frac{\sqrt{S}}{\log 2})^2 + \frac{S}{\log 2}} 2x dx. \end{aligned}$$

For a positive number c, we assume $\sqrt{M} \ge (1+c)\frac{\sqrt{S}}{\log 2}$ or equivalently $M \ge (1+c)^2 (\log 2)^{-2} S$, which is stronger than the previous assumption $M \ge (\log 2)^{-2} S$.

Then, since on the domain of integration $x \ge \sqrt{M} \ge (1+c)\frac{\sqrt{S}}{\log 2}$, we have $cx \le (1+c)(x - \frac{\sqrt{S}}{\log 2})$, and the estimation continues:

$$C_{S,n}(M+1) \le \int_{x=\sqrt{M}}^{\infty} e^{-(\log 2)(x-\frac{\sqrt{S}}{\log 2})^2+\frac{S}{\log 2}} 2 \cdot \frac{1+c}{c}(x-\frac{\sqrt{S}}{\log 2})dx$$
$$= \frac{1+c}{c}\frac{1}{\log 2} e^{-(\log 2)(\sqrt{M}-\frac{\sqrt{S}}{\log 2})^2+\frac{S}{\log 2}}$$
$$= \frac{1+c}{c}\frac{1}{\log 2} e^{-(\log 2)M+2\sqrt{SM}} = \frac{1+c}{c}\frac{1}{\log 2} 2^{-M} e^{2\sqrt{SM}}.$$

Since $C_{S,n}(M) \le C_{S,n}(M+1) + 2^{-M} e^{2\sqrt{SM}}$, we proved:

Proposition 1. *Let M be a positive integer, c a positive real number. Assume $M \ge (1+c)^2(\log 2)^{-2}S$. Then we have the following bound*

$$C_{S,n}(M) := \sum_{A \in V_{S,n} \setminus \mathscr{B}_{S,n}(M-1)} 2^{-\mu(A)} < \left(1 + \frac{1+c}{c}\frac{1}{\log 2}\right) 2^{-M} e^{2\sqrt{SM}}.$$

Remark 3. Definition 3 is a special case where $\alpha = n$ holds in: [4, Definition 3], where their Theorem 3 describes the condition of higher-order digital nets in terms of $\delta_{P^\perp}$.

Recently, Kousuke Suzuki [10] proved that the construction of higher order digital nets given in [5, Theorem 15.7] combined with some Niederreiter-Xing point sets [9] yields an explicit construction of low-WAFOM point sets, whose order of WAFOM is almost same with this paper.

2.4 Existence of Large Minimum-Weight Point Sets by Probabilistic Argument

Let d be a positive integer. Choose d matrices $B_1, \dots, B_d \in V_{S,n}$ uniformly randomly. Let $P \subset V$ be the $\mathbb{F}_2$-linear span of $B_1, \dots, B_d$. For any given nonzero matrix $L \in V_{S,n}$, let Perp_L be the event that $B_1, \dots, B_d$ are all perpendicular to L, which occurs with probability 2^{-d}. Consider the event that $\delta_{P^\perp} \ge M$, namely

$$P^\perp \cap \mathscr{B}_{S,n}(M-1) = \{0\}. \tag{9}$$

This is the complement of the union $\cup_{L \in \mathscr{B}_{S,n}(M-1)} \mathrm{Perp}_L$. The probability of this union is bounded from above by $\#(\mathscr{B}_{S,n}(M-1)) \times 2^{-d} = \mathrm{vol}_{S,n}(M-1)2^{-d}$. Thus, the probability that (9) holds is larger than $1 - \mathrm{vol}_{S,n}(M-1)2^{-d}$. This shows:

Proposition 2. *If $\mathrm{vol}_{S,n}(M-1) < 2^d$, then there exists $P \subset V$ of dimension at most d such that $\delta_{P^\perp} \ge M$. By Lemma 2, the condition is satisfied if*

$$\sqrt{M/S} e^{2\sqrt{SM}} \le 2^d. \tag{10}$$

Remark 4. Suppose that $B_1, \dots, B_d$ are linearly dependent. Still, if n is sufficiently large, we may choose any P of dimension d containing $B_1, \dots, B_d$, then $\delta_{P^\perp} \geq M$.

Corollary 1. *If the inequality (10) on d, M is satisfied, then there exists $P \subset V$ with dimension at most d with* $\mathrm{WAFOM}(P) < \left(1 + \frac{1+c}{c}\frac{1}{\log 2}\right) 2^{-M} e^{2\sqrt{SM}}$, *if M and c satisfy the condition in Proposition 1.*

For a given d, we want to estimate a large M satisfying (10). For this, we put $M = x^2 d^2/S$ for a positive indetermined x. Then, the inequality (10) is rewritten as

$$\frac{xd}{S} e^{2xd} \leq 2^d.$$

Assume $x \leq 1$ or equivalently $M \leq d^2/S$. Then a sufficient condition is

$$\frac{d}{S} e^{2xd} \leq 2^d,$$

and by taking log this is equivalent to

$$2xd \leq d(\log 2) - \log d + \log S. \tag{11}$$

This condition is tighter when S is smaller, so we may assume $S = 1$. Then, the above inequality is $x \leq (\log 2)/2 - (\log d)/(2d)$. Since $(\log d)/d$ is monotonically decreasing for $d \geq e$, by assuming $d \geq 4$ we have $(\log d)/d \geq (\log 4)/4 = (\log 2)/2$ and obtain a sufficient condition for (10):

$$x \leq (\log 2)/2 - (\log 2)/4 = (\log 2)/4 =: \alpha. \tag{12}$$

(We remark that by taking a larger constant $C \geq 4$ and assuming that $d \geq C$, we may replace α with $\frac{\log 2}{2} - \frac{\log C}{2C}$.)

Proposition 3. *Let $\alpha := (\log 2)/4$ and assume $d \geq 4$. If $M \leq \alpha^2 d^2/S$, then the inequality (10) is satisfied, and hence a subspace P of dimension at most d with $\delta_{P^\perp} \geq M$ exists (for all n) by Proposition 2.*

From now on, we take M to be $\lfloor \alpha^2 d^2/S \rfloor$ so that P with $\dim_{\mathbb{F}_2}(P) \leq d$ and $\delta_{P^\perp} \geq M$ exists. To make the estimation easier, we assume $M = \alpha^2 d^2/S$, which is not precise but has no effect on the order of estimation. Then the condition $M \geq (1+c)^2 (\log 2)^{-2} S$ in Proposition 1 is equivalent to $d \geq \frac{(1+c)S}{\alpha \log 2}$.

By plugging $M = \alpha^2 d^2/S$ in Corollary 1, we obtain Theorem 1; more precisely:

Theorem 2. *Let $\alpha := (\log 2)/4$, and take an arbitrary number $c > 0$. Then for any n and $d \geq \frac{(1+c)S}{\alpha \log 2}$, there is a subspace $P \subset V_{S,n}$ of $\mathbb{F}_2$-dimension at most d such that*

$$\mathrm{WAFOM}(P) \leq \left(1 + \frac{1+c}{c}\frac{1}{\log 2}\right) 2^{-\alpha^2 \frac{d^2}{S} + 2(\log 2)^{-1} \alpha d}.$$

If we write $N := 2^d$*, the right hand side is*

$$\left(1 + \frac{1+c}{c}\frac{1}{\log 2}\right) N^{-\alpha^2 \frac{\log_2 N}{S} + 2(\log 2)^{-1}\alpha}.$$

The constant α is approximately $0.173286\cdots$, and $d \geq \frac{(1+c)S}{\alpha \log 2} = (1+c)S \times 8.3254\cdots$. Hence, we may take c so that the latter condition is $d \geq 9S$.

Remark 5. Approximation. When d is large, then in the inequality (11), the effect of $\log d$ is negligible compared to $d(\log 2)$. Then, we may approximately replace α with $(\log 2)/2 = 0.34657\cdots$, and obtain approximately

$$\begin{aligned}\mathrm{WAFOM}(P) &\leq \left(1 + \frac{1+c}{c}\frac{1}{\log 2}\right) 2^{-((\log 2)/2)^2 d^2/S + d} \\ &= \left(1 + \frac{1+c}{c}\frac{1}{\log 2}\right) N^{-((\log 2)/2)^2 (\log_2 N)/S + 1}.\end{aligned}$$

Thus, the convergence ratio $N^{-\beta}$ would be realized when $d \geq S(\beta+1)(2/\log 2)^2 = S(\beta+1) \times 8.3254\cdots$. The condition $d \geq \frac{(1+c)S}{\alpha \log 2}$ becomes, for $\alpha = (\log 2)/2$, $d \geq S(1+c)/2 \cdot (2/\log 2)^2$ and in this case we may take c such that $(1+c)/2 = \beta + 1$, so that the both conditions coincide. For example, for $\beta \geq 1$, it suffices to take $d \geq S \times 16.6509\cdots$. This seems not practical for $S \geq 3$ in quasi-Monte Carlo integration, since the number of points 2^d is too large.

A finer estimation may lead to a better bound, but it would be a future work.

Acknowledgements The authors would like to thank Professor Harald Niederreiter, Professor Art Owen, and Mr. Kyle Matoba for helpful discussions and comments on the manuscript, and thank the anonymous referees for invaluable suggestions.

The first author is supported by JSPS/MEXT Grant-in-Aid for Scientific Research No.24654019, No.23244002, No.21654017. The second author is supported by Leading Graduate Course of Frontiers of Mathematical Sciences and Physics.

References

1. Andrews, G.E.: The Theory of Partitions. Cambridge University Press, Cambridge (1984)
2. Dick, J.: Walsh spaces containing smooth functions and quasi-Monte Carlo rules of arbitrary high order. SIAM J. Numer. Anal. **46**, 1519–1553 (2008)
3. Dick, J.: On quasi-Monte Carlo rules achieving higher order convergence. In: L'Ecuyer P., Owen, A.B. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2008, pp. 73–96. Springer, Berlin/Heidelberg (2010)
4. Dick, J., Kritzer, P.: Duality theory and propagation rules for generalized digital nets. Math. Comput. **79**, 993–1017 (2010)
5. Dick, J., Pillichshammer, F.: Digital Nets and Sequences: Discrepancy Theory and Quasi-Monte Carlo Integration. Cambridge University Press, Cambridge (2010)

6. Matousek, J., Nesetril, J.: Invitation to Discrete Mathematics. Oxford University Press, New York (2008)
7. Matsumoto, M., Saito, M., Matoba, K.: A computable figure of merit for quasi-Monte Carlo point sets. Math. Comp. (2013, to appear)
8. Niederreiter, H.: Random Number Generation and Quasi-Monte Carlo Methods. CBMS-NSF. Society for Industrial and Applied Mathematics, Philadelphia (1992)
9. Niederreiter, H., Xing, C.P.: Low-discrepancy sequences and global function fields with many rational places. Finite Fields Appl. **2**, 241–273 (1996)
10. Suzuki, K.: An explicit construction of point sets with large minimum Dick weight (Submitted)

ANOVA Decomposition of Convex Piecewise Linear Functions

Werner Römisch

Abstract Piecewise linear convex functions arise as integrands in stochastic programs. They are Lipschitz continuous on their domain, but do not belong to tensor product Sobolev spaces. Motivated by applying Quasi-Monte Carlo methods we show that all terms of their ANOVA decomposition, except the one of highest order, are smooth if the underlying densities are smooth and a certain geometric condition is satisfied. The latter condition is generically satisfied in the normal case.

1 Introduction

During the last decade much progress has been achieved in *Quasi-Monte Carlo (QMC) theory* for computing multidimensional integrals. Appropriate function spaces of integrands were discovered that allowed to improve the classical convergence rates. It is referred to the monographs [17, 27] for providing an overview of the earlier work and to [2, 12, 15] for presenting much of the more recent achievements.

In particular, certain reproducing kernel Hilbert spaces $\mathbb{F}_d$ of functions $f : [0,1]^d \to \mathbb{R}$ became important for estimating the quadrature error (see [7]). If the integral $I_d(f) = \int_{[0,1]^d} f(x)dx$ defines a linear continuous functional on $\mathbb{F}_d$ and $Q_{n,d}(f)$ denotes a Quasi-Monte Carlo method for computing $I_d(f)$, i.e.,

$$Q_{n,d}(f) = \frac{1}{n}\sum_{j=1}^{n} f(x_j) \quad (n \in \mathbb{N})$$

for some sequence $x_i \in [0,1)^d$, $i \in \mathbb{N}$, the quadrature error $e_n(\mathbb{F}_d)$ allows the representation

W. Römisch (✉)
Institute of Mathematics, Humboldt-University Berlin, Berlin, Germany
e-mail: romisch@math.hu-berlin.de

J. Dick et al. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2012*,
Springer Proceedings in Mathematics & Statistics 65,
DOI 10.1007/978-3-642-41095-6_30, © Springer-Verlag Berlin Heidelberg 2013

$$e_n(\mathbb{F}_d) = \sup_{f \in \mathbb{F}_d, \|f\| \le 1} \left| I_d(f) - Q_{n,d}(f) \right| = \sup_{\|f\| \le 1} |\langle f, h_n \rangle| = \|h_n\| \qquad (1)$$

according to Riesz' theorem for linear bounded functionals. The *representer* $h_n \in \mathbb{F}_d$ of the quadrature error is of the form

$$h_n(x) = \int_{[0,1]^d} K(x,y)dy - \frac{1}{n}\sum_{i=1}^{n} K(x,x_i) \quad (\forall x \in [0,1]^d),$$

where $K : [0,1]^d \times [0,1]^d \to \mathbb{R}$ denotes the kernel of $\mathbb{F}_d$. It satisfies the conditions $K(\cdot, y) \in \mathbb{F}_d$ and $\langle f, K(\cdot, y)\rangle = f(y)$ for each $y \in [0,1]^d$ and $f \in \mathbb{F}_d$, where $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ denote inner product and norm in $\mathbb{F}_d$.

In particular, the weighted tensor product Sobolev space [26]

$$\mathbb{F}_d = \mathscr{W}_{2,\mathrm{mix}}^{(1,\dots,1)}([0,1]^d) = \bigotimes_{i=1}^{d} W_2^1([0,1]) \qquad (2)$$

equipped with the weighted norm $\|f\|_\gamma^2 = \langle f, f \rangle_\gamma$ and inner product (see Sect. 2 for the notation)

$$\langle f, g \rangle_\gamma = \sum_{u \subseteq \{1,\dots,d\}} \prod_{j \in u} \gamma_j^{-1} \int_{[0,1]^{|u|}} \frac{\partial^{|u|}}{\partial x^u} f(x^u, 1^{-u}) \frac{\partial^{|u|}}{\partial x^u} g(x^u, 1^{-u}) dx^u, \qquad (3)$$

and a weighted Walsh space consisting of Walsh series (see [2, Example 2.8] and [1]) are reproducing kernel Hilbert spaces.

They became important for analyzing the recently developed randomized lattice rules (see [11, 13, 25] and [1, 2]) and allowed for deriving optimal error estimates of the form

$$e_n(\mathbb{F}_d) \le C(\delta) n^{-1+\delta} \quad (n \in \mathbb{N}, \delta \in (0, \tfrac{1}{2}]), \qquad (4)$$

where the constant $C(\delta)$ does not depend on the dimension d if the nonnegative weights γ_j satisfy

$$\sum_{j=1}^{\infty} \gamma_j^{\frac{1}{2(1-\delta)}} < \infty.$$

Unfortunately, a number of integrands do not belong to such tensor product Sobolev or Walsh spaces and are even not of bounded variation in the sense of Hardy and Krause. The latter condition represents the standard requirement on an integrand f to justify Quasi-Monte Carlo algorithms via the Koksma-Hlawka theorem [17, Theorem 2.11].

Often integrands are non-differentiable like those in option pricing models [31] or max-type functions in general. It has been discovered in [4, 5] that the so-called ANOVA decomposition (see Sect. 2) of such integrands may have a smoothing effect in the sense that many ANOVA terms are smooth if the underlying densities are sufficiently smooth.

In this paper we show that such a smoothing effect occurs also in case of piecewise linear convex functions f. More precisely, we show that all ANOVA terms except the one of highest order of such functions are infinitely differentiable if the densities are sufficiently smooth and a geometric property is satisfied. This geometric property is generic if the underlying densities are normal. The results pave the way to extensions for composite functions $f(g(\cdot))$ with a smooth mapping g. Since piecewise linear convex functions appear as the result of linear optimization processes, our results apply to linear two-stage stochastic programs and (slightly) extend the main result of [6]. Hence, the results justify earlier studies of QMC methods in stochastic programming [3, 9, 21] and motivate that the recently developed randomized lattice rules [2, 25] may be efficient for stochastic programming models if their superposition dimension is small. The computational experience reported in [6] confirms the efficiency of randomly shifted lattice rules.

The paper starts by recalling the ANOVA decomposition in Sect. 2 and convex piecewise linear functions in Sect. 3. Section 4 contains the main results on the smoothing effect of the ANOVA decomposition of convex piecewise linear functions, followed by discussing the generic character of the geometric property (Sect. 5) and dimension reduction (Sect. 6) both in the normal case.

2 ANOVA Decomposition and Effective Dimension

The analysis of variance (ANOVA) decomposition of a function was first proposed as a tool in statistical analysis (see [8] and the survey [29]). Later it was often used for the analysis of quadrature methods mainly on $[0, 1]^d$. Here, we will use it on $\mathbb{R}^d$ equipped with a probability measure given by a density function ρ of the form

$$\rho(\xi) = \prod_{j=1}^{d} \rho_j(\xi_j) \quad (\forall \xi = (\xi_1, \ldots, \xi_d) \in \mathbb{R}^d) \tag{5}$$

with continuous one-dimensional marginal densities ρ_j on $\mathbb{R}$. As in [5] we consider the weighted $\mathscr{L}_p$ space over $\mathbb{R}^d$, i.e., $\mathscr{L}_{p,\rho}(\mathbb{R}^d)$, with the norm

$$\|f\|_{p,\rho} = \begin{cases} \left(\int\limits_{\mathbb{R}^d} |f(\xi)|^p \rho(\xi) d\xi \right)^{\frac{1}{p}} & \text{if } 1 \le p < +\infty, \\ \operatorname{ess\,sup}\limits_{\xi \in \mathbb{R}^d} \rho(\xi) |f(\xi)| & \text{if } p = +\infty. \end{cases}$$

Let $I = \{1, \dots, d\}$ and $f \in \mathscr{L}_{1,\rho}(\mathbb{R}^d)$. The projection P_k, $k \in I$, is defined by

$$(P_k f)(\xi) := \int_{-\infty}^{\infty} f(\xi_1, \dots, \xi_{k-1}, s, \xi_{k+1}, \dots, \xi_d)\rho_k(s)ds \quad (\xi \in \mathbb{R}^d).$$

Clearly, the function $P_k f$ is constant with respect to ξ_k. For $u \subseteq I$ we use $|u|$ for its cardinality, $-u$ for $I \setminus u$ and write

$$P_u f = \Big(\prod_{k \in u} P_k\Big)(f),$$

where the product sign means composition. Due to Fubini's theorem the ordering within the product is not important and $P_u f$ is constant with respect to all $\xi_k, k \in u$.

The ANOVA decomposition of $f \in \mathscr{L}_{1,\rho}(\mathbb{R}^d)$ is of the form [14, 30]

$$f = \sum_{u \subseteq I} f_u \tag{6}$$

with f_u depending only on ξ^u, i.e., on the variables ξ_j with indices $j \in u$. It satisfies the property $P_j f_u = 0$ for all $j \in u$ and the recurrence relation

$$f_\emptyset = P_I(f) \quad \text{and} \quad f_u = P_{-u}(f) - \sum_{v \subseteq u} f_v\,.$$

It is known from [14] that the ANOVA terms are given explicitly in terms of the projections by

$$f_u = \sum_{v \subseteq u} (-1)^{|u|-|v|} P_{-v} f = P_{-u}(f) + \sum_{v \subset u} (-1)^{|u|-|v|} P_{u-v}(P_{-u}(f)), \tag{7}$$

where P_{-u} and P_{u-v} mean integration with respect to ξ_j, $j \in I \setminus u$ and $j \in u \setminus v$, respectively. The second representation motivates that f_u is essentially as smooth as $P_{-u}(f)$ due to the Inheritance Theorem [5, Theorem 2].

If f belongs to $\mathscr{L}_{2,\rho}(\mathbb{R}^d)$, the ANOVA functions $\{f_u\}_{u \subseteq I}$ are orthogonal in the Hilbert space $\mathscr{L}_{2,\rho}(\mathbb{R}^d)$ (see e.g. [30]).

Let the variance of f be defined by $\sigma^2(f) = \|f - P_I(f)\|^2_{L_2}$. Then it holds

$$\sigma^2(f) = \|f\|^2_{2,\rho} - (P_I(f))^2 = \sum_{\emptyset \neq u \subseteq I} \|f_u\|^2_{2,\rho} =: \sum_{\emptyset \neq u \subseteq I} \sigma^2_u(f).$$

To avoid trivial cases we assume $\sigma(f) > 0$ in the following. The normalized ratios $\frac{\sigma^2_u(f)}{\sigma^2(f)}$ serve as indicators for the importance of the variable ξ^u in f. They are used to define sensitivity indices of a set $u \subseteq I$ for f in [28] and the dimension distribution of f in [16, 18].

For small $\varepsilon \in (0,1)$ ($\varepsilon = 0.01$ is suggested in a number of papers), the *effective superposition (truncation) dimension* $d_S(\varepsilon)$ ($d_T(\varepsilon)$) is defined by

$$d_S(\varepsilon) = \min\left\{s \in I : \sum_{|u|\le s} \frac{\sigma_u^2(f)}{\sigma^2(f)} \ge 1-\varepsilon\right\}$$

$$d_T(\varepsilon) = \min\left\{s \in I : \sum_{u\subseteq\{1,\dots,s\}} \frac{\sigma_u^2(f)}{\sigma^2(f)} \ge 1-\varepsilon\right\}$$

and it holds $d_S(\varepsilon) \le d_T(\varepsilon)$ and (see [30])

$$\left\| f - \sum_{|u|\le d_S(\varepsilon)} f_u \right\|_{2,\rho} \le \sqrt{\varepsilon}\sigma(f). \tag{8}$$

For linear functions f one has $\sigma_u(f) = 0$ for $|u| > 1$, $d_S(\varepsilon) = 1$, but $d_T(\varepsilon)$ may be close to d [18, 30]. For the simple convex piecewise linear function $f(\xi_1,\xi_2) = \max\{\xi_1,\xi_2\}$ on $[0,1]^2$ with the uniform distribution it holds $f_\emptyset = \frac{1}{3}$, $\sigma^2(f) = \frac{1}{18}$,

$$f_{\{i\}}(\xi_i) = -\frac{1}{2}\xi_i^2 + \xi_i - \frac{1}{3},\ \ \sigma^2_{\{i\}}(f) = \frac{2}{45},\ (i=1,2),\ \ \sigma^2_{\{1,2\}}(f) = \frac{1}{90}.$$

Hence, we obtain $d_S(\varepsilon) = 2$ for $\varepsilon \in (0, \frac{1}{5})$ and the situation is entirely different for convex piecewise linear functions.

3 Convex Piecewise Linear Functions

Convex piecewise linear functions appear as optimal value functions of linear programs depending on parameters in right-hand sides of linear constraints or in the objective function. In general, they are nondifferentiable and not of bounded variation in the sense of Hardy and Krause (for the latter see [19]). On the other hand, such functions enjoy structural properties which make them attractive for variational problems.

As in [22, Sect. 2.I] a function f from $\mathbb{R}^d$ to the extended reals $\bar{\mathbb{R}}$ is called *piecewise linear* on $D = \operatorname{dom} f = \{\xi \in \mathbb{R}^d : f(\xi) < \infty\}$ if D can be represented as the union of finitely many polyhedral sets relative to each of which $f(\xi)$ is given by $f(\xi) = a^\top\xi + \alpha$ for some $a \in \mathbb{R}^d$ and $\alpha \in \mathbb{R}$.

Proposition 1. *Let $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ be proper, i.e., $f(\xi) > -\infty$ and $D = \operatorname{dom} f$ be nonempty. Then the function f is convex and piecewise linear if and only if it has a representation of the form*

$$f(\xi) = \begin{cases} \max\{a_1^\top \xi + \alpha_1, \ldots, a_\ell^\top \xi + \alpha_\ell\} & , \ \xi \in D, \\ \infty & , \ \xi \notin D, \end{cases} \tag{9}$$

for some $\ell \in \mathbb{N}$, $a_j \in \mathbb{R}^d$ *and* $\alpha_j \in \mathbb{R}$, $j = 1, \ldots, \ell$. *Moreover,* D *is polyhedral and, if* int D *is nonempty,* D *may represented as the union of a finite collection of polyhedral sets* D_j, $j = 1, \ldots, \ell$, *such that* int $D_j \neq \emptyset$ *and* int $D_j \cap$ int $D_{j'} = \emptyset$ *when* $j \neq j'$.

Proof. The two parts of the results are proved as Theorem 2.49 and Lemma 2.50 in [22, Sect. 2.I]. □

Example 1 (Linear two-stage stochastic programs). We consider the linear optimization problem

$$\min\left\{c^\top x + \mathbb{E}_P[q^\top y(\xi)] : Wy(\xi) + T(\xi)x = h(\xi),\ x \in X,\ y(\xi) \geq 0, \forall \xi \in \mathbb{R}^d\right\},$$

where $c \in \mathbb{R}^m$, $q \in \mathbb{R}^{\bar{m}}$, W is a $r \times \bar{m}$ matrix, $T(\xi)$ a $r \times m$ matrix, $h(\xi) \in \mathbb{R}^r$ for each $\xi \in \mathbb{R}^d$, X is convex and polyhedral in $\mathbb{R}^m$, P is a probability measure on $\mathbb{R}^d$ and $\mathbb{E}_P$ denotes expectation with respect to P. We assume that $T(\cdot)$ and $h(\cdot)$ are affine functions of ξ. The above problem may be reformulated as minimizing a convex integral functional with respect to x, namely,

$$\min\left\{c^\top x + \int_{\mathbb{R}^d} \Phi(h(\xi) - T(\xi)x)P(d\xi) : x \in X\right\}, \tag{10}$$

where Φ is the optimal value function assigning to each parameter $t \in \mathbb{R}^r$ an extended real number by $\Phi(t) = \inf\{q^\top y : Wy = t, y \geq 0\}$. The value $\Phi(t) = -\infty$ appears if there exists $y \in \mathbb{R}^{\bar{m}}_+$, $y \neq 0$ such that $Wy = 0$ and $\Phi(t) = +\infty$ means infeasibility, i.e., $\{y \in \mathbb{R}^{\bar{m}} : Wy = t, y \geq 0\}$ is empty. The integrand in (10) is $f(\xi) = c^\top x + \Phi(h(\xi) - T(\xi)x)$ for every $x \in X$.

Now, we assume that both dom $\Phi = \{t \in \mathbb{R}^r : \Phi(t) < +\infty\}$ and the dual polyhedron $\mathscr{D} = \{z \in \mathbb{R}^r : W^\top z \leq q\}$ are nonempty. Then $\Phi(t) > -\infty$ holds for all $t \in \mathbb{R}^r$ and the original primal as well as the dual linear program $\max\{t^\top z : z \in \mathscr{D}\}$ are solvable due to the duality theorem. If v^j, $j = 1, \ldots, l$, denote the vertices of $\mathscr{D}$, it holds

$$\Phi(t) = \max_{j=1,\ldots,l} t^\top v^j \quad (t \in \text{dom}\,\Phi = \mathbb{R}^r),$$

i.e., the integrand $f(\cdot)$ is convex and piecewise linear on $D = \mathbb{R}^d$ for every $x \in X$. For more information on stochastic programming see [23, 24].

4 ANOVA Decomposition of Convex Piecewise Linear Functions

We consider a piecewise linear convex function f and assume that its polyhedral domain $D = \operatorname{dom} f$ has nonempty interior. Let D_j, $j = 1, \ldots, \ell$, be the polyhedral subsets of D according to Proposition 1 such that

$$f(\xi) = a_j^\top \xi + \alpha_j \quad (\forall \xi \in D_j)$$

holds for some $a_j \in \mathbb{R}^d$, $\alpha_j \in \mathbb{R}$, $j = 1, \ldots, \ell$. For each $i \in I = \{1, \ldots, d\}$ there exist finitely many $(d-1)$-dimensional intersections H_{ij}, $j = 1, \ldots, J(i)$, of D_i with adjacent polyhedral sets D_j, $j \in \{1, \ldots, d\} \setminus \{i\}$. These polyhedral sets are subsets of finitely many $(d-1)$-dimensional affine subspaces of $\mathbb{R}^d$ which are renumbered by H_i, $i = 1, \ldots, \theta(f)$.

Furthermore, we assume that the support Ξ of the probability measure is contained in D and its density ρ is of the form (5). For any $k \in I$ we denote the kth coordinate projection of D by $\pi_k(D)$, i.e.,

$$\pi_k(D) = \{\xi_k \in \mathbb{R} : \exists \xi_j,\ j \in I, j \neq k, \text{ such that } \xi = (\xi_1, \ldots, \xi_d) \in D\}.$$

Next we intend to compute projections $P_k(f)$ for $k \in I$. For $\xi \in D$ we set $\bar{\xi}^k = (\xi_1, \ldots, \xi_{k-1}, \xi_{k+1}, \ldots, \xi_d)$, and $\bar{\xi}^k_s = (\xi_1, \ldots, \xi_{k-1}, s, \xi_{k+1}, \ldots, \xi_d)$ for $s \in \pi_k(D)$. We know that

$$\bar{\xi}^k_s \in \bigcup_{j=1}^{\ell} D_j = D \tag{11}$$

for every $s \in \pi_k(D)$ and assume $\rho_k(s) = 0$ for every $s \in \mathbb{R} \setminus \pi_k(D)$. Hence, we obtain by definition of the projection

$$(P_k f)(\bar{\xi}^k) = \int_{-\infty}^{\infty} f(\bar{\xi}^k_s)\rho_k(s)ds = \int_{-\infty}^{\infty} f(\xi_1, \ldots, \xi_{k-1}, s, \xi_{k+1}, \ldots, \xi_d)\rho_k(s)ds.$$

Due to (11) the one-dimensional affine subspace $\{\bar{\xi}^k_s : s \in \mathbb{R}\}$ intersects a finite number of the polyhedral sets D_j. Hence, there exist $p = p(k) \in \mathbb{N} \cup \{0\}$, $s_i = s_i^k \in \mathbb{R}$, $i = 1, \ldots, p$, and $j_i = j_i^k \in \{1, \ldots, \ell\}$, $i = 1, \ldots, p+1$, such that $s_i < s_{i+1}$ and

$$\begin{aligned}
&\bar{\xi}^k_s \in D_{j_1} && \forall s \in (-\infty, s_1] \cap \pi_k(D)\\
&\bar{\xi}^k_s \in D_{j_i} && \forall s \in [s_{i-1}, s_i] \quad (i = 2, \ldots, p)\\
&\bar{\xi}^k_s \in D_{j_{p+1}} && \forall s \in [s_p, +\infty) \cap \pi_k(D).
\end{aligned}$$

By setting $s_0 := -\infty$, $s_{p+1} := \infty$, we obtain the following explicit representation of $P_k f$.

$$(P_k f)(\bar{\xi}^k) = \sum_{i=1}^{p+1} \int_{s_{i-1}}^{s_i} (a_{j_i}^\top \bar{\xi}_s^k + \alpha_{j_i}) \rho_k(s) ds \tag{12}$$

$$= \sum_{i=1}^{p+1} \Big(\Big(\sum_{\substack{l=1 \\ l \neq k}}^{d} a_{j_i l} \xi_l + \alpha_{j_i} \Big) \int_{s_{i-1}}^{s_i} \rho_k(s) ds + a_{j_i k} \int_{s_{i-1}}^{s_i} s \rho_k(s) ds \Big)$$

$$= \sum_{i=1}^{p+1} \Big(\Big(\sum_{\substack{l=1 \\ l \neq k}}^{d} a_{j_i l} \xi_l + \alpha_{j_i} \Big) (\varphi_k(s_i) - \varphi_k(s_{i-1})) + a_{j_i k} (\psi_k(s_i) - \psi_k(s_{i-1})) \Big) \tag{13}$$

Here, φ_k is the one-dimensional distribution function with density ρ_k, ψ_k the corresponding mean value function and μ_k the mean value, i.e.,

$$\varphi_k(u) = \int_{-\infty}^{u} \rho_k(s) ds, \quad \psi_k(u) = \int_{-\infty}^{u} s \rho_k(s) ds, \quad \mu_k = \int_{-\infty}^{+\infty} s \rho_k(s) ds.$$

Next we reorder the outer sum to collect the factors of $\varphi_k(s_i)$ and $\psi_k(s_i)$, and a remainder.

$$(P_k f)(\bar{\xi}^k) = \sum_{i=1}^{p} \Big(\Big(\sum_{\substack{l=1 \\ l \neq k}}^{d} (a_{j_i l} - a_{j_{i+1} l}) \xi_l + (\alpha_{j_i} - \alpha_{j_{i+1}}) \Big) \varphi_k(s_i) + (a_{j_i k} - a_{j_{i+1} k}) \psi_k(s_i) \Big) + \sum_{\substack{l=1 \\ l \neq k}}^{d} a_{j_{p+1} l} \xi_l + \alpha_{j_{p+1}} + a_{j_{p+1} k} \mu_k. \tag{14}$$

As the convex function f is continuous on int D, it holds

$$a_{j_i}^\top \bar{\xi}_s^k + \alpha_{j_i} = a_{j_{i+1}}^\top \bar{\xi}_s^k + \alpha_{j_{i+1}}$$

and, thus, the points s_i, $i = 1, \ldots, p$, satisfy the equations

$$\sum_{\substack{l=1 \\ l \neq k}}^{d} \xi_l (a_{j_{i+1} l} - a_{j_i l}) + s_i (a_{j_{i+1} k} - a_{j_i k}) + \alpha_{j_{i+1}} - \alpha_{j_i} = 0 \quad (i = 1, \ldots, p).$$

This leads to the explicit formula

$$s_i = \frac{1}{a_{j_i k} - a_{j_{i+1} k}} \Big(\sum_{\substack{l=1 \\ l \neq k}}^{d} \xi_l (a_{j_{i+1} l} - a_{j_i l}) + \alpha_{j_{i+1}} - \alpha_{j_i} \Big) \text{ if } a_{j_i k} \neq a_{j_{i+1} k}. \quad (15)$$

for $i = 1, \ldots, p$. Hence, all s_i, $i = 1, \ldots, p$, are linear combinations of the remaining components ξ_j, $j \neq k$, of ξ if the following *geometric condition* is satisfied: All kth components of adjacent vectors a_j are different from each other, i.e., all polyhedral sets H_j are subsets of $(d-1)$-dimensional subspaces that are not parallel to the kth coordinate axis in $\mathbb{R}^r$ or, with other words, not parallel to the canonical basis element e_k (whose components are equal to δ_{ik}, $i = 1, \ldots, d$).

To simplify notation we set $w_i = a_{j_i} - a_{j_{i+1}}$ and $v_i = \alpha_{j_i} - \alpha_{j_{i+1}}$. If the above geometric condition is satisfied, we obtain the following representation of $P_k f$:

$$(P_k f)(\bar{\xi}^k) = \sum_{i=1}^{p} w_{ik} \Big(- s_i(\bar{\xi}^k) \varphi_k(s_i(\bar{\xi}^k)) + \psi_k(s_i(\bar{\xi}^k)) \Big) + \sum_{\substack{l=1 \\ l \neq k}}^{d} a_{j_{p+1} l} \xi_l + \alpha_{j_{p+1}} + a_{j_{p+1} k} \mu_k \quad (16)$$

$$s_i = s_i(\bar{\xi}^k) = -\frac{1}{w_{ik}} \Big(\sum_{\substack{l=1 \\ l \neq k}}^{d} w_{il} \xi_l + v_i \Big). \quad (17)$$

Hence, the projection represents a sum of products of differentiable functions and of affine functions of ξ^k.

Proposition 2. *Let f be piecewise linear convex having the form*

$$f(\xi) = a_j^\top \xi + \alpha_j \quad (\forall \xi \in D_j). \quad (18)$$

Let $k \in I$ and assume that vectors a_j belonging to adjacent polyhedral sets D_j have different kth components. Then the kth projection $P_k f$ is twice continuously differentiable. The projection $P_k f$ belongs to $C^{s+1}(\mathbb{R}^d)$ if the density ρ_k is in $C^{s-1}(\mathbb{R})$ ($s \in \mathbb{N}$). $P_k f$ is infinitely differentiable if the density ρ_k is in $C^\infty(\mathbb{R})$.

Proof. Let $l \in I$, $l \neq k$. The projection $P_k f$ is partially differentiable with respect to ξ_l and it holds

$$\begin{aligned} \frac{\partial P_k f}{\partial \xi_l}(\bar{\xi}^k) &= \sum_{i=1}^{p} w_{ik} \frac{\partial}{\partial \xi_l} \Big(- s_i(\bar{\xi}^k) \varphi_k(s_i(\bar{\xi}^k)) + \psi_k(s_i(\bar{\xi}^k)) \Big) + a_{j_{p+1} l} \\ &= \sum_{i=1}^{p} w_{il} \big(\varphi_k(s_i(\bar{\xi}^k)) + s_i(\bar{\xi}^k) \varphi_k'(s_i(\bar{\xi}^k)) - \psi_k'(s_i(\bar{\xi}^k)) \big) + a_{j_{p+1} l} \\ &= \sum_{i=1}^{p} w_{il} \varphi_k(s_i(\bar{\xi}^k)) + a_{j_{p+1} l} \end{aligned}$$

due to (16)–(17) and $\varphi_k'(s) = \rho_k(s)$ and $\psi_k'(s) = s\rho_k(s)$. Hence, the behavior of all first order partial derivatives of $P_k f$ only depends on the kth marginal distribution functions. The first order partial derivatives are continuous and again partially differentiable. The second order partial derivatives are of the form

$$\frac{\partial^2 P_k f}{\partial \xi_l \partial \xi_r}(\bar{\xi}^k) = \sum_{i=1}^{p} \frac{-w_{il} w_{ir}}{w_{ik}} \rho_k(s_i(\bar{\xi}^k))$$

and, thus, only depend on the marginal density ρ_k. Hence, $P_k f$ is twice continuously differentiable as ρ_k is continuous. If $\rho_k \in C^{s-1}(\mathbb{R})$ for some $s \in \mathbb{N}$, $P_k f$ belongs to $C^{s+1}(\mathbb{R}^d)$. If $\rho_k \in C^\infty(\mathbb{R})$, $P_k f$ is in $C^\infty(\mathbb{R}^d)$. □

Our next example shows that the geometric condition imposed in Proposition 2 is not superfluous.

Example 2. Let us consider the function

$$f(\xi) = \max\{\xi_1, -\xi_1, \xi_2\} \quad (\forall \xi = (\xi_1, \xi_2) \in \mathbb{R}^2)$$

on $D = \mathbb{R}^2$, i.e., we have $\alpha_1 = \alpha_2 = \alpha_3 = 0$ and $a_1 = (1, 0)^\top$, $a_2 = (-1, 0)^\top$ and $a_3 = (0, 1)^\top$. The decomposition of D according to Proposition 1 consists of

$$D_1 = \{\xi \in \mathbb{R}^2 : \xi_1 \geq 0, \xi_2 \leq \xi_1\}, \quad D_2 = \{\xi \in \mathbb{R}^2 : \xi_1 \leq 0, \xi_2 \leq -\xi_1\},$$
$$D_3 = \{\xi \in \mathbb{R}^2 : \xi_2 \geq \xi_1, \xi_2 \geq -\xi_1\}.$$

All polyhedral sets are adjacent and the second component of two of the vectors a_j, $j = 1, 2, 3$, coincides. Hence, the geometric condition in Proposition 2 is violated. Indeed, the projection $P_2 f$ is of the form

$$(P_2 f)(\xi_1) = |\xi_1| \int_{-\infty}^{|\xi_1|} \rho(\xi_2) d\xi_2 + \int_{|\xi_1|}^{+\infty} \xi_2 \rho(\xi_2) d\xi_2$$

and, thus, nondifferentiable on $\mathbb{R}$ (see also [6, Example 3]).

The previous result extends to more general projections P_u.

Proposition 3. *Let $\emptyset \neq u \subseteq I$, f be given by (18) and the vectors a_j belonging to adjacent polyhedral sets D_j have kth components which are all different for some $k \in u$. Then the projection $P_u f$ is continuously differentiable. The projection $P_u f$ is infinitely differentiable if $\rho_k \in C_b^\infty(\mathbb{R})$. Here, the subscript b at $C_b^\infty(\mathbb{R})$ indicates that all derivatives of functions in that space are bounded on $\mathbb{R}$.*

Proof. If $|u| = 1$ the result follows from Proposition 2. For $u = \{k, r\}$ with $k, r \in I$, $k \neq r$, we obtain from the Leibniz theorem [5, Theorem 1] for $l \notin u$

$$D_l P_u f(\xi^u) = \frac{\partial}{\partial \xi_l} P_u f(\xi^u) = P_r \frac{\partial}{\partial \xi_l} P_k f(\xi^u)$$

and from the proof of Proposition 2

$$D_l P_u f(\xi^u) = \sum_{i=1}^{p} w_{il} \int_{\mathbb{R}} \varphi_k(s_i(\bar{\xi}^k))\rho_r(\xi_r) d\xi_r + a_{j_{p+1}l} .$$

If u contains more than two elements, the integral on the right-hand side becomes a multiple integral. In all cases, however, such an integral is a function of the remaining variables ξ_j, $j \in I \setminus u$, whose continuity and differentiability properties correspond to those of φ_k and ρ_k. This follows from Lebesgue's dominated convergence theorem as φ_k and all densities ρ_j, $j \in u$, and their derivatives are bounded on $\mathbb{R}$. □

The following is the main result of this section.

Theorem 1. *Let $u \subset I$, f given by (18) and the vectors a_j belonging to adjacent polyhedral sets D_j have kth components which are all different for some $k \in -u = I \setminus u$. Then the ANOVA term f_u is infinitely differentiable if $\rho_k \in C_b^\infty(\mathbb{R})$.*

Proof. According to formula (7) it holds

$$f_u = P_{-u}(f) + \sum_{v \subset u} (-1)^{|u|-|v|} P_{u-v}(P_{-u}(f))$$

and Proposition 3 implies that $P_{-u} f$ is infinitely differentiable. The result follows from the Inheritance Theorem [5, Theorem 2] applied to $P_{u-v}(P_{-u}(f))$ for each $v \subset u$. □

Corollary 1. *Let f be given by (18) and the following geometric condition (GC) be satisfied: All $(d-1)$-dimensional subspaces containing $(d-1)$-dimensional intersections of adjacent polyhedral sets D_j are not parallel to any coordinate axis. Then the ANOVA approximation*

$$f_{d-1} := \sum_{u \subset I} f_u \tag{19}$$

of f is infinitely differentiable if all densities ρ_k, $k \in I$, belong to $C_b^\infty(\mathbb{R})$.

Proof. The result follows immediately from Theorem 1 when applying it to all nonempty strict subsets of I. □

Remark 1. Under the assumptions of Corollary 1 all ANOVA terms f_u are at least continuously differentiable if ρ is continuous and $|u| \leq d-1$. Hence, the function f_{d-1} is in $C^1(\mathbb{R}^d)$ ($C^\infty(\mathbb{R}^d)$) if each ρ_k, $k \in I$, belongs to $C(\mathbb{R})$ ($C_b^\infty(\mathbb{R})$). On the other hand, it holds

$$f = f_{d-1} + f_I \quad \text{and} \quad \|f - f_{d-1}\|_{L_2}^2 = \|f_I\|_{L_2}^2 = \sigma_I^2(f)$$

according to (6). Hence, the question arises: For which convex piecewise linear functions f is $\sigma_I^2(f)$ small or, in terms of the effective superposition dimension $d_S(\varepsilon)$ of f, is $d_S(\varepsilon)$ smaller than d (see also (8))?

5 Generic Smoothness in the Normal Case

We consider the convex, piecewise linear function

$$f(\xi) = \max\{a_1^\top \xi + \alpha_1, \ldots, a_\ell^\top \xi + \alpha_\ell\} \quad (\forall \xi \in \mathbb{R}^d)$$

on $\operatorname{dom} f = \mathbb{R}^d$ and assume that ξ is normal with mean μ and nonsingular covariance matrix Σ. Then there exists an orthogonal matrix Q such that $\Delta = Q\,\Sigma\,Q^\top$ is a diagonal matrix. Then the d-dimensional random vector η given by

$$\xi = Q\eta + \mu \quad \text{or} \quad \eta = Q^\top(\xi - \mu)$$

is normal with zero mean and covariance matrix Δ, i.e., has independent components. The transformed function $\hat{f}$

$$\hat{f}(\eta) = f(Q\eta+\mu) = \max_{j=1,\ldots,\ell}\{a_j^\top(Q\eta+\mu)+\alpha_j\} = \max_{j=1,\ldots,\ell}\{(Q^\top a_j)^\top\eta+a_j^\top\mu+\alpha_j\}$$

is defined on the polyhedral set $Q^\top D - Q^\top\mu$ and it holds

$$\hat{f}(\eta) = (Q^\top a_j)^\top\eta + a_j^\top\mu + \alpha_j \quad \text{for each } \eta \in Q^\top(D_j - \mu).$$

We consider now $(d-1)$-dimensional intersections H_{ij} of two adjacent polyhedral sets D_i and $D_j, i, j = 1, \ldots, \ell$. They are polyhedral subsets of $(d-1)$-dimensional affine subspaces H_i. The orthogonal matrix $Q^\top$ causes a rotation of the sets H_{ij} and the corresponding affine subspaces H_i. However, there are only countably many orthogonal matrices Q such that the geometric condition (GC) (see Corollary 1) on the subspaces is *not satisfied*. When equipping the linear space of all orthogonal $d \times d$ matrices with the standard norm topology, the set of all orthogonal matrices Q that satisfy the geometric condition, is a *residual set*, i.e., the countable intersection of open dense subsets. A property for elements of a topological space is called *generic* if it holds in a residual set. This proves

Corollary 2. *Let f be a piecewise linear convex function on* $\operatorname{dom} f = \mathbb{R}^d$ *and let ξ be normally distributed with nonsingular covariance matrix. Then the infinite differentiability of the ANOVA approximation f_{d-1} of f (given by (19)) is a generic property.*

Proof. Let μ be the mean vector and Σ be the nonsingular covariance matrix of ξ. Let Q be the orthogonal matrix satisfying $Q\,\Sigma\,Q^\top = \Delta = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ and

ρ be the normal density with mean μ and covariance matrix Δ. Then $\sigma_j > 0$, $j = 1, \ldots, d$, and ρ is equal to the product of all one-dimensional marginal densities ρ_k, where

$$\rho_k(t) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\Big(-\frac{(t-\mu_k)^2}{2\sigma_k^2}\Big) \quad (k = 1, \ldots, d),$$

and all ρ_k belong to $C_b^\infty(\mathbb{R})$. Hence, the result follows from Corollary 1. □

6 Dimension Reduction of Piecewise Linear Convex Functions

In order to replace a piecewise linear convex function f by the sum f_{d-1} of ANOVA terms until order $d-1$ (see Corollary 1), we need that the effective superposition dimension d_S of f is smaller than d. Hence, one is usually interested in determining and reducing the effective dimension. This topic is discussed in a number of papers, e.g., [3, 16, 18, 28, 30, 32].

In the *normal* or *lognormal case* there exist universal (i.e., independent on the structure of f) and problem dependent principles for dimension reduction.

A universal principle for dimension reduction is *principal component analysis* (PCA). In PCA one uses the decomposition $\Sigma = U_P U_P^\top$ of Σ with the matrix $U_P = (\sqrt{\lambda_1}u_1, \ldots, \sqrt{\lambda_d}u_d)$, the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d > 0$ of Σ in decreasing order and the corresponding orthonormal eigenvectors u_i, $i = 1, \ldots, d$, of Σ. Several authors report an enormous reduction of the effective truncation dimension in financial models if PCA is used (see, for example, [30, 31]). However, PCA may become expensive for large d and the reduction effect depends on the eigenvalue distribution.

Several *dimension reduction techniques* exploit the fact that a normal random vector ξ with mean μ and covariance matrix Σ can be transformed by $\xi = B\eta + \mu$ and *any* matrix B satisfying $\Sigma = B\,B^\top$ into a standard normal random vector η with independent components. The following equivalence principle is [32, Lemma 1] and already mentioned in [20, p. 182].

Proposition 4. *Let Σ be a $d \times d$ nonsingular covariance matrix and A be a fixed $d \times d$ matrix such that $A\,A^\top = \Sigma$. Then it holds $\Sigma = B\,B^\top$ if and only if B is of the form $B = A\,Q$ with some orthogonal $d \times d$ matrix Q.*

To apply the proposition, one may choose $A = L_C$, where L_C is the standard Cholesky matrix, or $A = U_P$. Then any other decomposition matrix B with $\Sigma = B\,B^\top$ is of the form $B = A\,Q$ with some orthogonal matrix Q.

A dimension reduction approach now consists in determining a *good* orthogonal matrix Q such that the truncation dimension is minimized by exploiting the structure of the underlying integrand f. Such an approach is proposed in [10] for linear functions and refined and extended in [32].

Piecewise linear convex functions are of the form

$$f(\xi) = G(a_1^\top \xi + \alpha_1, \ldots, a_\ell^\top \xi + \alpha_\ell), \tag{20}$$

where $G(t_1, \ldots, t_\ell) = \max\{t_1, \ldots, t_\ell\}$. Hence, f is of the form as considered in [32] shortly after Theorem 3. The transformed function is

$$\hat{f}(\eta) = f(B\eta + \mu) = G((B^\top a_1)^\top \eta_1 + a_1^\top \mu + \alpha_1, \ldots, (B^\top a_\ell)^\top \eta_d + a_\ell^\top \mu + \alpha_\ell). \tag{21}$$

In order to minimize the truncation dimension of $\hat{f}$ in (21), the following result is recorded from [32, Theorem 2] (see also Proposition 1 in [10]).

Proposition 5. *Let $\ell = 1$. If the matrix $Q = (q_1, \ldots, q_d)$ is determined such that*

$$q_1 = \pm \frac{A^\top a_1}{\|A^\top a_1\|} \quad \textit{and} \quad Q \textit{ is orthogonal,} \tag{22}$$

the transformed function is

$$\hat{f}(\eta) = G(\|A^\top a_1\| \eta_1 + a_1^\top \mu + \alpha_1)$$

and has effective truncation dimension $d_T = 1$.

The orthogonal columns $q_2, \ldots, q_d$ may be computed by the Householder transformation. In case $1 < \ell \leq d$ it is proposed in [32] to determine the orthogonal matrix $Q = (q_1, \ldots, q_d)$ by applying an orthogonalization technique to the matrix

$$M = (A^\top a_1, \ldots, A^\top a_\ell, b_{\ell+1}, \ldots, b_d), \tag{23}$$

where we assume that the $a_1, \ldots, a_\ell$ are linearly independent and $b_{\ell+1}, \ldots, b_d$ are selected such that M has rank d. It is shown in [32, Theorem 3] that then the function $\hat{f}$ depends only on $\eta_1, \ldots, \eta_\ell$. The practical computation may again be done by the Householder transformation applied to M in (23).

Acknowledgements The author wishes to express his gratitude to Prof. Ian Sloan (University of New South Wales, Sydney) for inspiring conversations during his visit to the Humboldt-University Berlin in 2011 and the referee for his/her constructive criticism. The research of the author is supported by the DFG Research Center MATHEON at Berlin.

References

1. Dick, J.: Walsh spaces containing smooth functions and Quasi-Monte Carlo rules of arbitrary high order. SIAM J. Numer. Anal. **46**, 1519–1553 (2008)
2. Dick, J., Pillichshammer, F.: Digital Nets and Sequences. Cambridge University Press, Cambridge (2010)

3. Drew, S.S., Homem-de-Mello, T.: Quasi-Monte Carlo strategies for stochastic optimization. In: Proceedings of the 2006 Winter Simulation Conference, Monterey, pp. 774–782. IEEE (2006)
4. Griebel, M., Kuo, F.Y., Sloan, I.H.: The smoothing effect of the ANOVA decomposition. J. Complexity **26**, 523–551 (2010)
5. Griebel, M., Kuo, F.Y., Sloan, I.H.: The smoothing effect of integration in $\mathbb{R}^d$ and the ANOVA decomposition. Math. Comp. **82**, 383–400 (2013)
6. Heitsch, H., Leövey, H., Römisch, W.: Are Quasi-Monte Carlo algorithms efficient for two-stage stochastic programs? Stochastic Programming E-Print Series 5-2012. www.speps.org
7. Hickernell, F.J.: A generalized discrepancy and quadrature error bound. Math. Comp. **67**, 299–322 (1998)
8. Hoeffding, W.: A class of statistics with asymptotically normal distribution. Ann. Math. Stat. **19**, 293–325 (1948)
9. Homem-de-Mello, T.: On rates of convergence for stochastic optimization problems under non-i.i.d. sampling. SIAM J. Optim. **19**, 524–551 (2008)
10. Imai, J., Tan, K.S.: Minimizing effective dimension using linear transformation. In: Niederreiter, H. (ed.) Monte Carlo and Quasi-Monte Carlo Methods 2002, pp. 275–292. Springer, Berlin/Heidelberg (2004)
11. Kuo, F.Y.: Component-by-component constructions achieve the optimal rate of convergence in weighted Korobov and Sobolev spaces. J. Complexity **19**, 301–320 (2003)
12. Kuo, F.Y., Schwab, Ch., Sloan, I.H.: Quasi-Monte Carlo methods for high-dimensional integration – the standard (weighted Hilbert space) setting and beyond. ANZIAM J. **53**, 1–37 (2011)
13. Kuo, F.Y., Sloan, I.H., Wasilkowski, G.W., Waterhouse, B.J.: Randomly shifted lattice rules with the optimal rate of convergence for unbounded integrands. J. Complexity **26**, 135–160 (2010)
14. Kuo, F.Y., Sloan, I.H., Wasilkowski, G.W., Woźniakowski, H.: On decomposition of multivariate functions. Math. Comp. **79**, 953–966 (2010)
15. Lemieux, C.: Monte Carlo and Quasi-Monte Carlo Sampling. Springer, New York (2009)
16. Liu, R., Owen, A.B.: Estimating mean dimensionality of analysis of variance decompositions. J. Amer. Statist. Assoc. **101**, 712–721 (2006)
17. Niederreiter, H.: Random Number Generation and Quasi-Monte Carlo Methods. SIAM, Philadelphia (1992)
18. Owen, A.B.: The dimension distribution and quadrature test functions. Stat. Sin. **13**, 1–17 (2003)
19. Owen, A.B.: Multidimensional variation for Quasi-Monte Carlo. In: Fan, J., Li, G. (eds.) Contemporary Multivariate Analysis and Design of Experiments (in celebration of K. T. Fang's 65th birthday), pp. 49–74. World Scientific Publishing (2005)
20. Papageorgiou, A.: The Brownian bridge does not offer a consistent advantage in Quasi-Monte Carlo integration. J. Complexity **18**, 171–186 (2002)
21. Pennanen, T., Koivu, M.: Epi-convergent discretizations of stochastic programs via integration quadratures. Numer. Math. **100**, 141–163 (2005)
22. Rockafellar, R.T., Wets, R.J.-B.: Variational Analysis. Springer, Berlin (1998)
23. Ruszczyński, A., Shapiro, A. (eds.): Stochastic Programming. Handbooks in Operations Research and Management Science, vol. 10. Elsevier, Amsterdam (2003)
24. Shapiro, A., Dentcheva, D., Ruszczyński, A.: Lectures on Stochastic Programming. MPS-SIAM Series on Optimization. SIAM, Philadelphia (2009)
25. Sloan, I.H., Kuo, F.Y., Joe, S.: On the step-by-step construction of Quasi-Monte Carlo integration rules that achieve strong tractability error bounds in weighted Sobolev spaces. Math. Comp. **71**, 1609–1640 (2002)
26. Sloan, I.H., Woźniakowski, H.: When are Quasi Monte Carlo algorithms efficient for high-dimensional integration. J. Complexity **14**, 1–33 (1998)
27. Sobol', I.M.: Multidimensional Quadrature Formulas and Haar Functions. Nauka, Moscow (1969, in Russian)

28. Sobol', I.M.: Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Math. Comput. Simulation **55**, 271–280 (2001)
29. Takemura, A.: Tensor analysis of ANOVA decomposition. J. Amer. Statist. Assoc. **78**, 894–900 (1983)
30. Wang, X., Fang, K.-T.: The effective dimension and Quasi-Monte Carlo integration. J. Complexity **19**, 101–124 (2003)
31. Wang, X., Sloan, I.H.: Why are high-dimensional finance problems often of low effective dimension. SIAM J. Sci. Comput. **27**, 159–183 (2005)
32. Wang, X., Sloan, I.H.: Quasi-Monte Carlo methods in financial engineering: an equivalence principle and dimension reduction. Oper. Res. **59**, 80–95 (2011)

Hit-and-Run for Numerical Integration

Daniel Rudolf

Abstract We study the numerical computation of an expectation of a bounded function f with respect to a measure given by a non-normalized density on a convex body $K \subset \mathbb{R}^d$. We assume that the density is log-concave, satisfies a variability condition and is not too narrow. In [19, 25, 26] it is required that K is the Euclidean unit ball. We consider general convex bodies or even the whole $\mathbb{R}^d$ and show that the integration problem satisfies a refined form of tractability. The main tools are the hit-and-run algorithm and an error bound of a multi run Markov chain Monte Carlo method.

1 Introduction and Results

In many applications, for example in Bayesian inference, see [5, 8], or in statistical physics, see [18, 27], it is desirable to compute an expectation of the form

$$\int_K f(x)\, \pi_\rho(\mathrm{d}x) = \int_K f(x)\, c\, \rho(x)\, \mathrm{d}x,$$

where the probability measure π_ρ is given by the density $c\,\rho$ with $c > 0$. The normalizing constant of the density

$$\frac{1}{c} = \int_K \rho(x)\, \mathrm{d}x$$

D. Rudolf (✉)
Institute of Mathematics, Friedrich-Schiller-University Jena, Ernst-Abbe-Platz 2, 07743 Jena, Germany

School of Mathematics and Statistics, University of New South Wales, Sydney, NSW, Australia
e-mail: daniel.rudolf@uni-jena.de

J. Dick et al. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2012*,
Springer Proceedings in Mathematics & Statistics 65,
DOI 10.1007/978-3-642-41095-6_31,

is not known and hard to compute. We want to have algorithms that are able to compute the expectation without any precomputation of c.

More precisely, let $\rho\colon \mathbb{R}^d \to \mathbb{R}_+$ be a possibly non-normalized density function, let $K = \operatorname{supp}(\rho) \subset \mathbb{R}^d$ be a convex body and let $f\colon K \to \mathbb{R}$ be integrable with respect to π_ρ. For a tuple (f, ρ) we define the desired quantity

$$A(f,\rho) = \frac{\int_K f(x)\,\rho(x)\,\mathrm{d}x}{\int_K \rho(x)\,\mathrm{d}x}. \tag{1}$$

In [19] a simple Monte Carlo method is considered which evaluates the numerator and denominater of $A(f,\rho)$ on a common independent, uniformly distributed sample in K. There it must be assumed that one can sample the uniform distribution in K. The authors show that this algorithm is not able to use any additional structure, such as log-concavity, of the density function. But they show that such structure can be used by Markov chain Monte Carlo which then outperforms the simple Monte Carlo method.

Markov chain Monte Carlo algorithms for the integration problem of the form (1) are considered in [19, 21, 25, 26]. Basically it is always assumed that K is the Euclidean unit ball rather than a general convex body. We extend the results to the case where K might even be the whole $\mathbb{R}^d$ if the density satisfies some further properties. We do not assume that we can sample with respect to π_ρ. The idea is to compute $A(f,\rho)$ by using a Markov chain which approximates π_ρ. We prove that the integration problem (1) satisfies an extended type of tractability. Now let us introduce the error criterion and the new notion of tractability.

Error criterion and algorithms. Let $t\colon \mathbb{N}\times\mathbb{N} \to \mathbb{N}$ be a function and let A_{n,n_0} be a generic algorithm which uses $t(n,n_0)$ Markov chain steps. Intuitively, the number n_0 determines the number of steps to approximate π_ρ. The number n determines the number of pieces of information of f used by the algorithm. The error is measured in mean square sense, for a tuple (f,ρ) it is given by

$$e(A_{n,n_0}(f,\rho)) = \left(\mathrm{E}\,|A_{n,n_0}(f,\rho) - A(f,\rho)|^2\right)^{1/2},$$

where E denotes the expectation with respect to the joint distribution of the used sequence of random variables determined by the Markov chain.

For example the algorithm might be a single or multi run Markov chain Monte Carlo. More precisely, assume that we have a Markov chain with limit distribution π_ρ and let $X_1, \dots, X_{n+n_0}$ be the first $n+n_0$ steps. Then

$$S_{n,n_0}(f,\rho) = \frac{1}{n}\sum_{j=1}^{n} f(X_{j+n_0})$$

is an approximation of $A(f,\rho)$ and the function $t(n,n_0) = n + n_0$. In contrast to the single run Markov chain Monte Carlo S_{n,n_0} one might consider a multi run

Markov chain Monte Carlo, say M_{n,n_0}, given as follows. Assume that we have n independent Markov chains with the same transition kernel, the same initial distribution and limit distribution π_ρ. Let $X_{n_0}^1, \ldots, X_{n_0}^n$ be the sequence of the n_0th steps of the Markov chains, then

$$M_{n,n_0}(f,\rho) = \frac{1}{n}\sum_{j=1}^{n} f(X_{n_0}^j)$$

is an approximation of $A(f,\rho)$. In this setting the function $t(n,n_0) = n \cdot n_0$.

Tractability. In [19,21] a notion of tractability for the integration problem (1) is introduced. It is assumed that $\|f\|_\infty \leq 1$ and that the density function satisfies

$$\frac{\sup_{x\in K}\rho(x)}{\inf_{x\in K}\rho(x)} \leq \gamma,$$

for some $\gamma \geq 3$. Let $s_{\varepsilon,\gamma}(n,n_0)$ be the minimal number of function values of (f,ρ) to guarantee an ε-approximation with respect to the error above. Then the integration problem is called tractable with respect to γ if $s_{\varepsilon,\gamma}(n,n_0)$ depends polylogarithmically on γ and depends polynomially on ε^{-1}, d. We extend this notion of tractability. We study a class of tuples (f,ρ) which satisfy $\|f\|_\infty \leq 1$ and we assume that for any ρ there exists a set $G \subset K$ such that for $\kappa \geq 3$ holds

$$\frac{\int_K \rho(x)\,\mathrm{d}x}{\mathrm{vol}_d(G)\ \inf_{x\in G}\rho(x)} \leq \kappa, \tag{2}$$

where $\mathrm{vol}_d(G)$ denotes the d-dimensional volume of G. Then we call the integration problem tractable with respect to κ if the minimal number of function values $t_{\varepsilon,\kappa}(n,n_0)$ of (f,ρ) to guarantee an ε-approximation satisfies for some non-negative numbers p_1, p_2 and p_3 that

$$t_{\varepsilon,\kappa}(n,n_0) = \mathcal{O}(\varepsilon^{-p_1} d^{\,p_2} [\log\kappa]^{p_3}), \quad \varepsilon > 0,\ d \in \mathbb{N},\ \kappa \geq 3.$$

Hence we permit only polylogarithmical dependence on the number κ, since it might be very large (e.g. 10^{30} or 10^{40}). The extended notion of tractability allows us to consider $K = \mathrm{supp}(\rho) = \mathbb{R}^d$.

The structure of the work and the main results are as follows. We use the hit-and-run algorithm to approximate π_ρ. An explicit estimate of the total variation distance of the hit-and-run algorithm, proven by Lovász and Vempala in [15,16], and an error bound of the mean square error of M_{n,n_0} are essential. In Sect. 2 we provide the basics on Markov chains and prove an error bound of M_{n,n_0}. In Sect. 3 we define the class of density functions. Roughly we assume that the densities are log-concave, that for any ρ there exists a set $G \subset K$ such that condition (2) holds for $\kappa \geq 3$ and that the densities are not too narrow. Namely, we assume that level sets of ρ of measure larger than $1/8$ contain a ball with radius r. We distinguish

two settings which guarantee that the densities are not too spread out. Either the convex body $K = \mathrm{supp}(\rho)$ is bounded by a ball with radius R around 0, then we say $\rho \in \mathscr{U}_{r,R,\kappa}$, or the support of ρ is bounded in average sense,

$$\int_K \left|x - x_\rho\right|^2 \pi_\rho(\mathrm{d}x) \leq 4R^2,$$

where $x_\rho = \int_K x\, \pi_\rho(\mathrm{d}x) \in \mathbb{R}^d$ is the centroid. Then we say $\rho \in \mathscr{V}_{r,R,\kappa}$. For precise definitions see Sect. 3. In Sect. 4 we provide the hit-and-run algorithm and state convergence properties of the algorithm for densities from $\mathscr{U}_{r,R,\kappa}$ and $\mathscr{V}_{r,R,\kappa}$. Then we show that the integration problem (1) is tractable with respect to κ, see Sect. 5. For $\rho \in \mathscr{U}_{r,R,\kappa}$ we obtain in Theorem 4 that

$$t_{\varepsilon,\kappa}(n, n_0) = \mathscr{O}(d^2\,[\log d]^2\,\varepsilon^{-2}\,[\log \varepsilon^{-1}]^3\,[\log \kappa]^3). \tag{3}$$

For $\rho \in \mathscr{V}_{r,R,\kappa}$ we find in Theorem 5 a slightly worse bound of the form

$$t_{\varepsilon,\kappa}(n, n_0) = \mathscr{O}(d^2\,[\log d]^2\,\varepsilon^{-2}\,[\log \varepsilon^{-1}]^5\,[\log \kappa]^5). \tag{4}$$

Here the $\mathscr{O}$ notation hides the polynomial dependence on r and R.

In [19, 21, 25, 26] it is proven that the problem (1) is tractable with respect to γ for $K = B_d$, where B_d denotes the Euclidean unit ball. Note that for $G = B_d$ we have

$$\frac{\int_K \rho(x)\,\mathrm{d}x}{\mathrm{vol}_d(G)\ \inf_{x\in G} \rho(x)} \leq \frac{\sup_{x\in K} \rho(x)}{\inf_{x\in K} \rho(x)} \leq \gamma.$$

Furthermore it is assumed that $\rho\colon B_d \to \mathbb{R}_+$ is log-concave and $\log \rho$ is Lipschitz. Then the Metropolis algorithm with a ball walk proposal is used to approximate π_ρ. For $\|f\|_p \leq 1$ with $p > 2$ the algorithm S_{n,n_0} is considered for the approximation of $A(f, \rho)$. It is proven that

$$s_{\varepsilon,\gamma}(n, n_0) = \mathscr{O}(d\,\max\{\log^2(\gamma), d\}(\varepsilon^{-2} + \log \gamma)). \tag{5}$$

In open problem 84 of [21] it is asked whether one can extend this result to other families of convex sets. The complexity bound of (5) is better than the results of (3) and (4) in terms of the dimension, the precision and γ. On the one hand the assumption that $K = B_d$ is very restrictive but on the other hand the estimates of (3) and (4) seem to be pessimistic. However, with our results we contribute to problem 84 in the sense that tractability with respect to γ can be shown for arbitrary convex bodies or even the whole $\mathbb{R}^d$ if the density functions satisfy certain properties.

2 Markov Chains and an Error Bound

Let $(X_n)_{n\in\mathbb{N}}$ be a Markov chain with transition kernel $P(\cdot,\cdot)$ and initial distribution ν on a measurable space $(K,\mathscr{B}(K))$, where $K\subset\mathbb{R}^d$ and $\mathscr{B}(K)$ is the Borel σ-algebra. We assume that the transition kernel $P(\cdot,\cdot)$ is reversible with respect to π_ρ. For $p\in[1,\infty]$ we denote by $L_p=L_p(\pi_\rho)$ the class of functions $f\colon K\to\mathbb{R}$ with

$$\|f\|_p=\left(\int_K |f(x)|^p\,\pi_\rho(\mathrm{d}x)\right)^{1/p}<\infty.$$

Similarly we denote by $\mathscr{M}_p$ the class of measures ν which are absolutely continuous with respect to π_ρ and where the density $\frac{d\nu}{d\pi_\rho}\in L_p$. The transition kernel induces an operator $P\colon L_p\to L_p$ given by

$$Pf(x)=\int_K f(y)\,P(x,\mathrm{d}y),\quad x\in K,$$

and it induces an operator $P\colon\mathscr{M}_p\to\mathscr{M}_p$ given by

$$\mu P(C)=\int_K P(x,C)\,\mu(\mathrm{d}x),\quad C\in\mathscr{B}(K).$$

For $n\in\mathbb{N}$ and a probability measure ν note that $\Pr(X_n\in C)=\nu P^n(C)$, where $C\in\mathscr{B}(K)$. We define the total variation distance between νP^n and π_ρ as

$$\left\|\nu P^n-\pi_\rho\right\|_{\mathrm{tv}}=\sup_{C\in\mathscr{B}(K)}\left|\nu P^n(C)-\pi_\rho(C)\right|.$$

Under suitable assumptions on the Markov chain one obtains that $\left\|\nu P^n-\pi_\rho\right\|_{\mathrm{tv}}\to 0$ as $n\to\infty$.

Now we consider the multi run Markov chain Monte Carlo method and prove an error bound. This bound is not new, see for example [4].

Theorem 1. *Assume that we have n_0 independent Markov chains with transition kernel $P(\cdot,\cdot)$ and initial distribution $\nu\in\mathscr{M}_1$. Let π_ρ be a stationary distribution of $P(\cdot,\cdot)$. Let $X_{n_0}^1,\ldots,X_{n_0}^n$ be the sequence of the n_0th steps of the Markov chains and let*

$$M_{n,n_0}(f,\rho)=\frac{1}{n}\sum_{j=1}^{n} f(X_{n_0}^j).$$

Then

$$e(M_{n,n_0}(f,\rho))^2\le\frac{1}{n}\|f\|_\infty^2+2\|f\|_\infty^2\left\|\nu P^n-\pi_\rho\right\|_{tv}.$$

Proof. With an abuse of notation let us denote

$$A(f) = \int_K f(x)\,\pi_\rho(\mathrm{d}x) \quad \text{and} \quad \nu P^{n_0}(f) = \int_K f(x)\,\nu P^{n_0}(\mathrm{d}x).$$

We decompose the error into variance and bias. Then

$$\begin{aligned} e(M_{n,n_0}(f,\rho))^2 &= \frac{1}{n}\int_K |f(x) - \nu P^{n_0}(f)|^2\,\nu P^{n_0}(\mathrm{d}x) + |\nu P^{n_0}(f) - A(f)|^2 \\ &= \frac{1}{n}\left(\nu P^{n_0}(f^2) - \nu P^{n_0}(f)^2\right) + |\nu P^{n_0}(f) - A(f)|^2 \\ &\le \frac{1}{n}\,\|f\|_\infty^2 + \int_K f(x)^2\left|\nu P^{n_0}(\mathrm{d}x) - \pi_\rho(\mathrm{d}x)\right| \\ &\le \frac{1}{n}\,\|f\|_\infty^2 + 2\,\|f\|_\infty^2\left\|\nu P^{n_0} - \pi_\rho\right\|_{\mathrm{tv}}. \end{aligned}$$

The last inequality follows by a well known characterization of the total variation distance, see for example [24, Proposition 3]. □

Very often there exists a number $\beta \in [0,1)$ and a number $C_\nu < \infty$ such that

$$\left\|\nu P^n - \pi_\rho\right\|_{\mathrm{tv}} \le C_\nu \beta^n.$$

For example, if $\beta = \|P - A\|_{L_2\to L_2} < 1$ and $C_\nu = \frac{1}{2}\left\|\nu - \pi_\rho\right\|_2$, see [23] for more details. Let us define the L_2-spectral gap as

$$\mathrm{gap}(P) = 1 - \|P - A\|_{L_2\to L_2}.$$

This is a significant quantity, see for instance [2, 26–29]. In [26] it is shown that

$$e(S_{n,n_0}(f,\rho))^2 \le \frac{4\,\|f\|_4}{n\,\mathrm{gap}(P)} \qquad \text{for} \qquad n_0 \ge \frac{\log\left(64\left\|\frac{d\nu}{d\pi_\rho} - 1\right\|_2\right)}{\mathrm{gap}(P)}.$$

There are several Markov chains where it is possible to provide, for certain classes of density functions, a lower bound of $\mathrm{gap}(P)$ which grows polynomially with respect to the dimension, see for example [16, 19]. Then, the error bound of the single run Markov chain Monte Carlo method might imply that the integration problem (1) is tractable with respect to some κ.

Note that there are also other possible approximation schemes and other bounds of the error of S_{n,n_0} which depend on different assumptions to the Markov chain (e.g. Ricci curvature condition, drift condition, small set), see for instance [9, 11–13]. For example one might consider a multi run Markov chain Monte Carlo method where function values of a trajectory of each Markov chain after a sufficiently large n_0 are used. But all known error bounds of such methods include quantities such as the L_2-spectral gap or the conductance.

It is not an easy task to prove that a Markov chain satisfies the different assumptions stated above and it is also not an easy task to prove a lower bound of the L_2-spectral gap. It might be easier to estimate the total variation distance of νP^{n_0} and π_ρ directly. Then one can use Theorem 1 to show that the integration problem (1) is tractable with respect to some κ.

3 Densities with Additional Structure

Let us assume that the densities have some additional structure. For $0 < r \leq R$ and $\kappa \geq 3$ a density function $\rho\colon K \to \mathbb{R}_+$ is in $\mathscr{U}_{r,R,\kappa}$ if the following properties are satisfied:

(a) ρ is log-concave, i.e. for all $x, y \in K$ and $\lambda \in [0,1]$ one has

$$\rho(\lambda x + (1-\lambda)y) \geq \rho(x)^\lambda \rho(y)^{1-\lambda}.$$

(b) ρ is strictly positive, i.e. $K = \operatorname{supp}(\rho)$ and we assume that $K \subset RB_d$, where RB_d is the Euclidean ball with radius R around 0.

(c) There exists a set $G \subset K$ such that

$$\frac{\int_K \rho(x)\,\mathrm{d}x}{\operatorname{vol}_d(G)\ \inf_{x\in G}\rho(x)} \leq \kappa,$$

and we can sample the uniform distribution on G.

(d) For $s > 0$ let $K(s) = \{x \in K \mid \rho(x) \geq t\}$ be the level set of ρ and let $B(z,r)$ be the Euclidean ball with radius r around z. Then

$$\pi_\rho(K(s)) \geq \frac{1}{8} \quad \Longrightarrow \quad \exists z \in K \quad B(z,r) \subset K(s).$$

The log-concavity of ρ implies that the maximal value is attained on a convex set, that the function is continuous and that one has an isoperimetric inequality, see [16]. Assumption (b) gives that K is bounded.

By (c) we can sample the uniform distribution on G. We can choose it as initial distribution for a Markov chain, where the number κ provides an estimate of the influence of this initial distribution.

The condition on the level set $K(s)$ guarantees that the peak is not too narrow. Roughly speaking, if the π_ρ measure of a level set is not too small, then the Lebesgue measure is also not too small. Note that K is bounded from below, since condition (d) implies that $B(z,r) \subset K$.

Now we enlarge the class of densities. Let us define the following property:

(b′) ρ is strictly positive, i.e. $K = \operatorname{supp}(\rho)$ and $x_\rho = \int_K x\ \pi_\rho(\mathrm{d}x) \in \mathbb{R}^d$ is the centroid of π_ρ. Then

$$\int_K \left|x - x_\rho\right|^2 \pi_\rho(\mathrm{d}x) \le 4\, R^2.$$

We have $\rho \in \mathscr{V}_{r,R,\kappa}$ if the density ρ satisfies (a), (b′), (c) and (d). We substituted the boundedness condition (b) by (b′). Note that (b) implies (b′). Hence $\mathscr{U}_{r,R,\kappa} \subset \mathscr{V}_{r,R,\kappa}$. Condition (b′) provides a boundedness criterion in average sense. Namely, it implies that

$$\int_K \int_K |x - y|^2 \, \pi_\rho(\mathrm{d}x)\, \pi_\rho(\mathrm{d}y) \le 8R^2.$$

Example of a Gaussian function in $\mathscr{V}_{r,R,\kappa}$. Let Σ be a symmetric and positive definite $d \times d$ matrix. We consider the non-normalized density

$$\varphi(x) = \exp(-\frac{1}{2}\, x^T \Sigma^{-1} x), \quad x \in \mathbb{R}^d.$$

The target distribution π_φ is a normal distribution with mean $x_\varphi = 0 \in \mathbb{R}^d$ and covariance matrix Σ. There exists an orthogonal matrix $V = (v_1, \dots, v_d)$, where $v_1, \dots, v_d$ are the eigenvectors of Σ. Then

$$V^{-1} \Sigma V = \Lambda,$$

where $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_d)$ and $\lambda_1, \dots, \lambda_d$ with $\lambda_i > 0$ for $i \in \{1, \dots, d\}$ are the corresponding eigenvalues of Σ. Recall that the trace and the determinant of Σ are

$$\mathrm{tr}(\Sigma) = \sum_{i=1}^d \lambda_i \quad \text{and} \quad \det(\Sigma) = \prod_{i=1}^d \lambda_i.$$

We show that if r, R and κ are appropriately chosen, then $\varphi \in \mathscr{V}_{r,R,\kappa}$.

To (a): The density φ is obviously log-concave.

To (b′): Since $x_\varphi = 0$ we obtain

$$\int_K \left|x - x_\varphi\right|^2 \pi_\varphi(\mathrm{d}x) = \frac{1}{(2\pi)^{d/2}\sqrt{\det(\Sigma)}} \int_{\mathbb{R}^d} |x|^2 \, \varphi(x)\, \mathrm{d}x = \mathrm{tr}(\Sigma).$$

Hence we set $R = \frac{1}{2}\sqrt{\mathrm{tr}(\Sigma)}$.

To (c): Let $\lambda_{\min} = \min_{i=1,\dots,d} \lambda_i$ and let $v_{\min}$ be the corresponding eigenvector. Note that $x^T \Sigma^{-1} x \le \lambda_{\min}^{-1} |x|^2$ and that equality holds for $x = v_{\min}$. With $G = B_d$ we obtain

$$\frac{\int_{\mathbb{R}^d} \varphi(x)\, \mathrm{d}x}{\mathrm{vol}_d(B_d)\ \inf_{x \in B_d} \varphi(x)} = \exp(\frac{1}{2}\, \lambda_{\min}^{-1})\ \Gamma(d/2 + 1)\ 2^{d/2} \sqrt{\det(\Sigma)},$$

where $\Gamma(d) = \int_0^\infty t^{d-1} \exp(-t)\, \mathrm{d}t$ is the gamma function. Hence we set

$$\kappa = \exp(\frac{1}{2}\,\lambda_{\min}^{-1})\;\Gamma(d/2+1)\;2^{d/2}\sqrt{\det(\Sigma)}.$$

To (d): The level sets of φ are ellipsoids

$$K(s) = \{x \in \mathbb{R}^d \mid x^T\,\Sigma^{-1}x \le 2\log(s^{-1})\}, \quad s \in [0,1].$$

In general one has

$$\pi_\varphi(K(s)) = \frac{\int_0^\infty \mathrm{vol}_d(K(s)\cap K(t))\;\mathrm{d}t}{\int_0^\infty \mathrm{vol}_d(K(t))\;\mathrm{d}t} = \frac{s\;\mathrm{vol}_d(K(s)) + \int_s^\infty \mathrm{vol}_d(K(t))\,\mathrm{d}t}{\int_0^\infty \mathrm{vol}_d(K(t))\;\mathrm{d}t}.$$

By the well known formula of the volume of an ellipsoid we obtain

$$\mathrm{vol}_d(K(t)) = 2^{d/2}\;\log^{d/2}(t^{-1})\sqrt{\det(\Sigma)}\;\mathrm{vol}_d(B_d), \quad t \in [0,1]$$

and

$$\pi_\varphi(K(s)) = \frac{s\;\log^{d/2}(s^{-1}) + \int_s^1 \log^{d/2}(t^{-1})\;\mathrm{d}t}{\int_0^1 \log^{d/2}(t^{-1})\;\mathrm{d}t}, \quad s \in [0,1].$$

Hence

$$\pi_\varphi(K(s)) = \frac{\gamma(\log s^{-1}, d/2)}{\Gamma(d/2)}, \quad s \in [0,1],$$

where $\gamma(r,d) = \int_0^r t^{d-1}\exp(-t)\,\mathrm{d}t$ is the lower incomplete gamma function. Let us define a function $r^* : \mathbb{N} \to \mathbb{R}$ by

$$r^*(d) = \inf\{r \in [0,\infty):\ \gamma(r, d/2) \ge \frac{1}{8}\,\Gamma(d/2)\}.$$

If we substitute $1/8$ by $1/2$ in the definition of $r^*(d)$ we have the median of the gamma distribution with parameter $d/2$ and 1. It is known that the median is in $\Theta(d)$, see [1]. Figure 1 suggests that $r^*(d)$ behaves also linearly in d.
Let $\log(s^*(d)^{-1}) = r^*(d)$, such that $s^*(d) = \exp(-r^*(d))$. Then

$$\pi_\varphi(K(s^*(d))) = \frac{1}{8} \quad \text{and} \quad B(0, (\lambda_{\min} r^*(d))^{1/2}) \subset K(s^*(d)).$$

Hence we set $r = (\lambda_{\min} r^*(d))^{1/2}$.

Let us summarize. For $r = (\lambda_{\min} r^*(d))^{1/2}$, $R = \frac{1}{2}\sqrt{\mathrm{tr}(\Sigma)}$ and

$$\kappa = \exp(\frac{1}{2}\,\lambda_{\min}^{-1})\;\Gamma(d/2+1)\;2^{d/2}\sqrt{\det(\Sigma)}$$

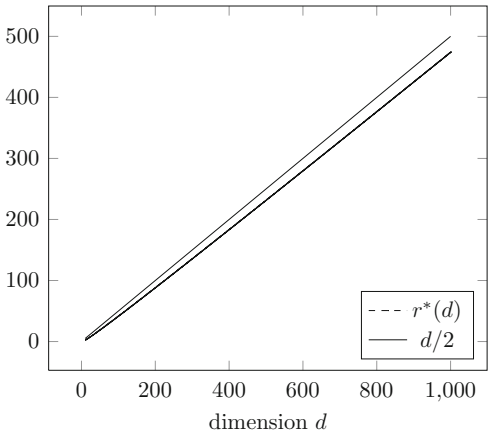


Fig. 1 Plot of an approximation of $r^*(d)$ with a Newton method and an appropriately chosen initial value.

we obtain that $\varphi \in \mathcal{V}_{r,R,\kappa}$. Note that κ depends exponentially on the dimension d. However, if one has tractability with respect to κ, then the error depends polynomially on the dimension.

4 Hit-and-Run Algorithm

For $\rho\colon K \to \mathbb{R}_+$ the hit-and-run algorithm is as follows. Let ν be a probability measure on $(K, \mathscr{B}(K))$ and let $x_1 \in K$ be chosen by ν. For $k \in \mathbb{N}$ suppose that the states $x_1, \dots, x_k$ are already computed. Then

1. Choose a direction u uniformly distributed on ∂B_d;
2. Set $x_{k+1} = x_k + \alpha\, u$, where $\alpha \in I_k = \{\alpha \in \mathbb{R} \mid x_k + \alpha u \in K\}$ is chosen with respect to the distribution determined by the density

$$\ell_k(s) = \frac{\rho(x_k + s\, u)}{\int_{I_k} \rho(x_k + t\, u)\, \mathrm{d}t}, \quad s \in I_k.$$

The second step might cause implementation issues. However, if we have a log-concave density ρ then ℓ_k is also log-concave. In this setting one can use different acceptance/rejection methods. For more details see for example [6, Sect. 2.4.2] or [17]. In the following we assume that we can sample the distribution determined by ℓ_k.

Other algorithms for the approximation of π_ρ would be a Metropolis algorithm with suitable proposal [19] or a combination of a hit-and-run algorithm with uniform stationary distribution and a Ratio-of-uniforms method [10]. Also hybrid samplers are promising methods, especially when ρ decreases exponentially in the tails [7].

Now let us state the transition kernel, say H_ρ, of the hit-and-run algorithm

$$H_\rho(x, C) \;=\; \frac{2}{\mathrm{vol}_{d-1}(\partial B^d)} \int_C \frac{\rho(y)\,\mathrm{d}y}{\ell_\rho(x, y)\,|x-y|^{d-1}}, \quad x \in K,\, C \in \mathscr{B}(K),$$

where

$$\ell_\rho(x, y) \;=\; \int_{-\infty}^{\infty} \rho(\lambda x + (1-\lambda)y)\mathbf{1}_K(\lambda x + (1-\lambda)y)\,\mathrm{d}\lambda.$$

The transition kernel H_ρ is reversible with respect to π_ρ, let us refer to [3] for further details.

In the following we state several results from Lovaśz and Vempala. This part is based on [15]. We start with a special case of [15, Theorem 1.1] and sketch the proof of this theorem.

Theorem 2. *Let $\varepsilon \in (0, 1/2)$ and $\rho \in \mathscr{U}_{r,R,\kappa}$. Let ν be an initial distribution with the following property. There exists a set $S_{\varepsilon \subset K}$ and a number $D \geq 1$ such that*

$$\frac{d\nu}{d\pi_\rho}(x) \leq D, \quad x \in K \setminus S_\varepsilon,$$

where $\nu(S_\varepsilon) \leq \varepsilon$. Then for

$$n_0 > 10^{27}(dr^{-1}\,R)^2 \log^2(8\,D\,dr^{-1}\,R\varepsilon^{-1}) \log(4\,D\,\varepsilon^{-1})$$

the total variation distance between $\nu H_\rho^{n_0}$ and π_ρ is less than 2ε.

Proof (Sketch).

1. Let us assume that $S_\varepsilon = \emptyset$:
   Then it follows $\left\| \frac{d\nu}{d\pi_\rho} \right\|_\infty \leq D$, so that $\nu \in \mathscr{M}_\infty$. We use [14, Corollary 1.6] with $s = \frac{\varepsilon}{2D}$ and obtain

$$\left\| \nu H_\rho^n - \pi_\rho \right\|_{\mathrm{tv}} \leq \varepsilon/2 + D \exp(-\frac{1}{2}\, n\, \Phi^2_{\frac{\varepsilon}{2D}}),$$

   where $\Phi_{\frac{\varepsilon}{2D}}$ is the $\frac{\varepsilon}{2D}$-conductance of H_ρ. By Theorem 3.7 of [15] and the scaling invariance of the hit-and-run algorithm we find a lower bound of $\Phi_{\frac{\varepsilon}{2D}}$. It is given by

$$\Phi_{\frac{\varepsilon}{2D}} \geq \frac{10^{-13}}{2\,dr^{-1}\,R \log(4\,dr^{-1}\,R\,D\,\varepsilon^{-1})}. \tag{6}$$

   This leads to

$$\left\| \nu H_\rho^n - \pi_\rho \right\|_{\mathrm{tv}} \leq \varepsilon/2 + D\, \exp\left(\frac{-10^{-26}\, n}{8\,(dr^{-1}\,R)^2 \log^2(4\,dr^{-1}\,R\,D\,\varepsilon^{-1})} \right). \tag{7}$$

2. Now let us assume that $S_\varepsilon \neq \emptyset$:
Let $\tilde{\varepsilon} := \nu(S_\varepsilon)$, so that $0 < \tilde{\varepsilon} \leq \varepsilon \leq 1/2$ and for $C \in \mathscr{B}(K)$ let

$$\mu_1(C) = \frac{\nu(C \cap S_\varepsilon^c)}{\nu(S_\varepsilon^c)} \quad \text{and} \quad \mu_2(C) = \frac{\nu(C \cap S_\varepsilon)}{\nu(S_\varepsilon)}.$$

Then

$$\nu = (1 - \tilde{\varepsilon})\mu_1 + \tilde{\varepsilon}\mu_2$$

and $\left\| \frac{d\mu_1}{d\pi_\rho} \right\|_\infty \leq 2D$. Furthermore for any $C \in \mathscr{B}(K)$ we find

$$\left| \nu H_\rho^n(C) - \pi_\rho(C) \right| \leq (1 - \tilde{\varepsilon}) \left| \mu_1 H_\rho^n(C) - \pi_\rho(C) \right| + \tilde{\varepsilon}.$$

By using (7) we get

$$\left\| \mu_1 H_\rho^n - \pi_\rho \right\|_{\text{tv}} \leq \varepsilon/2 + 2\, D \exp\left(\frac{-10^{-26}\, n}{8\, (dr^{-1}\, R)^2 \log^2(8\, dr^{-1}\, R\, D\, \varepsilon^{-1})} \right),$$

and altogether

$$\left\| \nu H_\rho^n - \pi_\rho \right\|_{\text{tv}} \leq 3\, \varepsilon/2 + 2\, D \exp\left(\frac{-10^{-26}\, n}{8\, (dr^{-1}\, R)^2 \log^2(8\, dr^{-1}\, R\, D\, \varepsilon^{-1})} \right). \tag{8}$$

Choosing n so that the right hand side of the previous equation is less than or equal to 2ε completes the proof. □

The next Corollary provides an explicit upper bound of the total variation distance.

Corollary 1. *Under the assumptions of Theorem 2 with*

$$\beta = \exp\left(\frac{-10^{-9}}{(dr^{-1}\, R)^{2/3}} \right) \quad \textit{and} \quad C = 12\, dr^{-1}\, R\, D$$

one obtains

$$\left\| \nu H_\rho^{n_0} - \pi_\rho \right\|_{tv} \leq C\ \beta^{\sqrt[3]{n_0}}, \quad n \in \mathbb{N}.$$

Proof. Set $\varepsilon = 8\, dr^{-1}\, R\, D \exp\left(\frac{-10^{-9}\, n^{1/3}}{(dr^{-1}\, R)^{2/3}} \right)$ and use (8) to complete the proof. □

Note that the result of Theorem 2 is better than the result of Corollary 1. However, Corollary 1 provides an explicit estimate of the total variation distance. One can see that there is an almost exponential decay, namely the total variation distance goes to zero at least as $\beta^{\sqrt[3]{n_0}}$ goes to zero for increasing n_0.

In the previous results we assumed that $\rho \in \mathscr{U}_{r,R,\kappa}$. It is essentially used that (b) holds. Now let us assume that $\rho \in \mathscr{V}_{r,R,\kappa}$. The next statement is proven in [15, Theorem 1.1].

Theorem 3. *Let $\varepsilon \in (0, 1/2)$, $\rho \in \mathscr{V}_{r,R,\kappa}$. Let ν be an initial distribution with the following property. There exists a set $S_{\varepsilon \subset K}$ and a number $D \geq 1$ such that*

$$\frac{d\nu}{d\pi_\rho}(x) \leq D, \quad x \in K \setminus S_\varepsilon,$$

where $\nu(S_\varepsilon) \leq \varepsilon$. Then for

$$n_0 \geq 4 \cdot 10^{30} (dr^{-1} R)^2 \log^2(2\, D\, dr^{-1}\, R\, \varepsilon^{-1}) \log^3(D\, \varepsilon^{-1})$$

the total variation distance between $\nu H_\rho^{n_0}$ and π_ρ is less than 2ε.

Note that Theorems 2 and 3 can be applied if the initial distribution is bounded, i.e. we can set $D = \left\| \frac{d\nu}{d\pi_\rho} \right\|_\infty$ and $S_\varepsilon = \emptyset$. Furthermore if $\nu \in \mathscr{M}_2$, i.e. $\left\| \frac{d\nu}{d\pi_\rho} \right\|_2$ is bounded, then we can also apply Theorems 2 and 3 with $D = \left\| \frac{d\nu}{d\pi_\rho} \right\|_2^2 \varepsilon^{-1}$ and

$$S_\varepsilon = \left\{ x \in K \mid \frac{d\nu}{d\pi_\rho}(x) > \left\| \frac{d\nu}{d\pi_\rho} \right\|_2^2 \varepsilon^{-1} \right\}.$$

5 Main Results

Now we are able to state and to prove the main results. To avoid any pathologies we assume that $r^{-1} Rd \geq 3$.

Theorem 4. *Let $\varepsilon \in (0, 1/2)$ and*

$$\mathscr{F}_{r,R,\kappa} = \{(f, \rho) \mid \rho \in \mathscr{U}_{r,R,\kappa},\ \|f\|_\infty \leq 1\}.$$

For $(f, \rho) \in \mathscr{F}_{r,R,\kappa}$ let ν be the uniform distribution on $G \subset \mathbb{R}^d$ from (c). Let $X_{n_0}^1, \ldots, X_{n_0}^n$ be a sequence of the n_0th steps of n independent hit-and-run Markov chains with stationary distribution π_ρ and initial distribution ν. Recall that

$$M_{n,n_0}(f, \rho) = \frac{1}{n} \sum_{j=1}^{n} f(X_{n_0}^j).$$

Then for $n \geq \varepsilon^{-2}$ and

$$n_0 \geq 10^{27} (dr^{-1} R)^2 \log^2(8\, dr^{-1}\, R\, \kappa\, \varepsilon^{-2}) \log(4\kappa\, \varepsilon^{-2})$$

we obtain

$$\sup_{(f,\rho)\in\mathscr{F}_{r,R,\kappa}} e(M_{n,n_0}(f,\rho)) \le 3\varepsilon.$$

Hence

$$t_{\varepsilon,\kappa}(n,n_0) = \mathscr{O}(d^2\,(r^{-1}\,R)^2\,\log^2(dr^{-1}\,R)\,\varepsilon^{-2}\,[\log\varepsilon^{-1}]^3\,[\log\kappa]^3).$$

Proof. For $C \in \mathscr{B}(K)$ we have

$$\nu(C) = \int_C \frac{\mathbf{1}_G(y)\int_K \rho(x)\,\mathrm{d}x}{\mathrm{vol}_d(G)\rho(y)}\,\pi_\rho(\mathrm{d}y).$$

It implies that $\frac{d\nu}{d\pi_\rho}(x) \le \kappa$ for all $x \in K$. Then the assertion follows by Theorems 1 and 2. □

Now let us consider densities which belong to $\mathscr{V}_{r,R,\kappa}$.

Theorem 5. *Let* $\varepsilon \in (0,1/2)$ *and*

$$\mathscr{G}_{r,R,\kappa} = \{(f,\rho) \mid \rho \in \mathscr{V}_{r,R,\kappa},\ \|f\|_\infty \le 1\}.$$

Let M_{n,n_0} *be given as in Theorem 4. Then for* $n \ge \varepsilon^{-2}$ *and*

$$n_0 \ge 4\cdot 10^{30}(dr^{-1}\,R)^2\log^2(2\,dr^{-1}\,R\,\kappa\,\varepsilon^{-2})\log^3(\kappa\,\varepsilon^{-2})$$

we obtain

$$\sup_{(f,\rho)\in\mathscr{G}_{r,R,\kappa}} e(M_{n,n_0}(f,\rho)) \le 3\varepsilon.$$

Hence

$$t_{\varepsilon,\kappa}(n,n_0) = \mathscr{O}(d^2\,(r^{-1}\,R)^2\,\log^2(dr^{-1}\,R)\,\varepsilon^{-2}\,[\log\varepsilon^{-1}]^5\,[\log\kappa]^5).$$

Proof. The assertion follows by the same steps as the proof of Theorem 4. Note that we use Theorem 3 instead of Theorem 2. □

Note that in both theorems there is no hidden dependence on further parameters in the $\mathscr{O}$ notation. However, the explicit constant might be very large, of the magnitude of 10^{30}. The theorems imply that the problem of integration (1) is tractable with respect to κ on the classes $\mathscr{F}_{r,R,\kappa}$ and $\mathscr{G}_{r,R,\kappa}$.

Example of a Gaussian function revisited. In the Gaussian example of Sect. 3 we obtained

$$R/r = (2\,r^*(d)^{1/2})^{-1} \cdot \sqrt{\mathrm{tr}(\Sigma)/\lambda_{\min}},$$
$$\kappa = \exp(\frac{1}{2}\,\lambda_{\min}^{-1})\;\Gamma(d/2+1)\;2^{d/2}\sqrt{\det(\Sigma)}.$$

If we assume that $r^*(d)$ increases linearly in d (Fig. 1), that $\sqrt{\mathrm{tr}(\Sigma)/\lambda_{\min}}$ and $\log(\exp(\frac{1}{2}\;\lambda_{\min}^{-1})\sqrt{\det(\Sigma)})$ grows polynomially in the dimension, then $t_{\varepsilon,\kappa}(n, n_0)$ grows also polynomially in the dimension. This implies that the integration problem with respect to the Gaussian function is polynomially tractable in the sense of Novak and Woźniakowski [20–22].

Acknowledgements The author gratefully acknowledges the comments of the referees and wants to express his thanks to the local organizers of the Tenth International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing for their hospitality. The research was supported by the DFG Priority Program 1324, the DFG Research Training Group 1523, and an Australian Research Council Discovery Project.

References

1. Adell, J., Jodrá, P.: Sharp estimates for the median of the $\Gamma(n+1, 1)$ distribution. Stat. Probab. Lett. **71**, 185–191 (2005)
2. Aldous, D.: On the Markov chain simulation method for uniform combinatorial distributions and simulated annealing. Probab. Engrg. Inform. Sci. **1**, 33–46 (1987)
3. Bélisle, C., Romeijn, E., Smith, R.: Hit-and-run algorithms for generating multivariate distributions. Math. Oper. Res. **18**, 255–266 (1993)
4. Belloni, A., Chernozhukov, V.: On the computational complexity of MCMC-based estimators in large samples. Ann. Statist. **37**, 2011–2055 (2009)
5. Brooks, S., Gelman, A., Jones, G., Meng, X.: Handbook of Markov Chain Monte Carlo. Chapman & Hall, Boca Raton (2011)
6. Casella, G., Robert, C.: Monte Carlo Statistical Methods, 2nd edn. Springer Texts in Statistics. Springer, New York (2004)
7. Fort, G., Moulines, E., Roberts, G., Rosenthal, J.: On the geometric ergodicity of hybrid samplers. J. Appl. Probab. **40**, 123–146 (2003)
8. Gilks, W., Richardson, S., Spiegelhalter, D.: Markov Chain Monte Carlo in Practice. Chapman & Hall, Boca Raton (1996)
9. Joulin, A., Ollivier, Y.: Curvature, concentration and error estimates for Markov chain Monte Carlo. Ann. Probab. **38**, 2418–2442 (2010)
10. Karawatzki, R., Leydold, J., Potzelberger, K.: Automatic Markov chain Monte Carlo procedures for sampling from multivariate distributions. Tech. Rep. 27, Department of Statistics and Mathematics, WU Wien (2005)
11. Łatuszyński, K., Miasojedow, B., Niemiro, W.: Nonasymptotic bounds on the estimation error of MCMC algorithms. ArXiv e-prints (2011)
12. Łatuszyński, K., Miasojedow, B., Niemiro, W.: Nonasymptotic bounds on the mean square error for MCMC estimates via renewal techniques. ArXiv e-prints (2011)
13. Łatuszyński, K., Niemiro, W.: Rigorous confidence bounds for MCMC under a geometric drift condition. J. Complexity **27**, 23–38 (2011)
14. Lovász, L., Simonovits, M.: Random walks in a convex body and an improved volume algorithm. Random Structures Algorithms **4**, 359–412 (1993)

15. Lovász, L., Vempala, S.: Fast algorithms for logconcave functions: sampling, rounding, integration and optimization. In: Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science, FOCS '06, Berkeley, pp. 57–68. IEEE Computer Society, Washington, DC (2006)
16. Lovász, L., Vempala, S.: Hit-and-run from a corner. SIAM J. Comput. **35**, 985–1005 (2006)
17. Lovász, L., Vempala, S.: The geometry of logconcave functions and sampling algorithms. Random Structures Algorithms **30**, 307–358 (2007)
18. Martinelli, F.: Relaxation times of Markov chains in statistical mechanics and combinatorial structures. In: Probability on Discrete Structures. Encyclopaedia Mathematical Sciences, vol. 110, pp. 175–262. Springer, Berlin (2004)
19. Mathé, P., Novak, E.: Simple Monte Carlo and the Metropolis algorithm. J. Complexity **23**, 673–696 (2007)
20. Novak, E., Woźniakowski, H.: Tractability of Multivariate Problems. Vol. 1: Linear Information. EMS Tracts in Mathematics, vol. 6. European Mathematical Society (EMS), Zürich (2008)
21. Novak, E., Woźniakowski, H.: Tractability of Multivariate Problems. Vol. 2: Standard Information for Functionals. EMS Tracts in Mathematics, vol. 12. European Mathematical Society (EMS), Zürich (2010)
22. Novak, E., Woźniakowski, H.: Tractability of Multivariate Problems. Vol. 3: Standard Information for Operators. EMS Tracts in Mathematics, vol. 12. European Mathematical Society (EMS), Zürich (2012)
23. Roberts, G., Rosenthal, J.: Geometric ergodicity and hybrid Markov chains. Electron. Commun. Probab. **2**, 13–25 (1997)
24. Roberts, G., Rosenthal, J.: General state space Markov chains and MCMC algorithms. Probab. Surv. **1**, 20–71 (2004)
25. Rudolf, D.: Explicit error bounds for lazy reversible Markov chain Monte Carlo. J. Complexity **25**, 11–24 (2009)
26. Rudolf, D.: Explicit error bounds for Markov chain Monte Carlo. Dissertationes Math. **485**, 93 (2012)
27. Sokal, A.: Monte Carlo methods in statistical mechanics: foundations and new algorithms. In: Functional Integration (Cargèse, 1996). NATO Advanced Study Institutes Series B Physics, vol. 361, pp. 131–192. Plenum, New York (1997)
28. Ullrich, M.: Comparison of Swendsen-Wang and heat-bath dynamics. ArXiv e-prints (2011)
29. Ullrich, M.: Swendsen-wang is faster than single-bond dynamics. ArXiv e-prints (2012)

QMC Galerkin Discretization of Parametric Operator Equations

Christoph Schwab

Abstract We extend recent results from Kuo et al. (SIAM J Numer Anal 50:3351–3374, 2012) of QMC quadrature and Finite Element discretization for parametric, scalar second order elliptic partial differential equations to general QMC-Galerkin discretizations of parametric operator equations, which depend on possibly countably many parameters. Such problems typically arise in the numerical solution of differential and integral equations with random field inputs. The present setting covers general second order elliptic equations which are possibly indefinite (Helmholtz equation), or which are given in saddle point variational form (such as mixed formulations). They also cover nonsymmetric variational formulations which appear in space-time Galerkin discretizations of parabolic problems or countably parametric nonlinear initial value problems (Hansen and Schwab, Vietnam J. Math 2013, to appear).

1 Introduction

The efficient numerical computation of statistical quantities for solutions of partial differential and of integral equations with random inputs is a key task in uncertainty quantification in engineering and in the sciences. The quantity of interest being expressed as a mathematical expectation, the efficient computation of these quantities involves two basic steps: (i) approximate (numerical) solution of the operator equation, and (ii) numerical integration. In the present note, we outline a general strategy towards these two aims which is based on (i) stable Galerkin discretization and (ii) Quasi Monte-Carlo (QMC) integration by a randomly shifted, first order lattice rule following [6, 17, 22].

C. Schwab (✉)
Seminar für Angewandte Mathematik, ETH Zürich, Zürich, Switzerland
e-mail: schwab@math.ethz.ch

J. Dick et al. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2012*,
Springer Proceedings in Mathematics & Statistics 65,
DOI 10.1007/978-3-642-41095-6_32, © Springer-Verlag Berlin Heidelberg 2013

QMC (and other) quadrature methods require the *introduction of coordinates of integration* prior to numerical quadrature. In the context of random field inputs with nondegenerate covariance operators, a *countable number of coordinates* is required to describe the random input data, e.g. by a Karhunen-Loève expansion. Therefore, in the present note, we consider in particular that the operator equation contains not only a finite number of random input parameters, but rather depends on *random field inputs*, i.e. it contains random functions of space and, in evolution problems, of time which describe uncertainty in the problem under consideration. Combined QMC – Finite Element error analysis for scalar diffusion problems with random coefficients was obtained recently in [9, 17]. In the present note, we indicate how the main conclusions in [17] extend to larger classes of problems.

2 Parametric Operator Equations

2.1 Abstract Saddle Point Problems

Throughout, we denote by $\mathscr{X}$ and $\mathscr{Y}$ two reflexive Banach spaces over $\mathbb{R}$ (all results will hold with the obvious modifications also for spaces over $\mathbb{C}$) with (topological) duals $\mathscr{X}'$ and $\mathscr{Y}'$, respectively. By $\mathscr{L}(\mathscr{X}, \mathscr{Y}')$, we denote the set of bounded linear operators $A : \mathscr{X} \to \mathscr{Y}'$. The Riesz representation theorem associates each $A \in \mathscr{L}(\mathscr{X}, \mathscr{Y}')$ in a one-to-one correspondence with a bilinear form $\mathbf{b}(\cdot,\cdot) : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}$ by means of

$$\mathbf{b}(v, w) = \langle w, Av \rangle_{\mathscr{Y} \times \mathscr{Y}'} \quad \text{for all } v \in \mathscr{X}, w \in \mathscr{Y} \ . \tag{1}$$

Here and in what follows, we indicate spaces in duality pairings $\langle \cdot, \cdot \rangle$ by subscripts.

We shall be interested in the solution of linear operator equations $Au = f$ and make use of the following solvability result which is a straightforward consequence of the closed graph theorem, see, e.g., [1] or [8, Chap. 4].

Proposition 1. *A bounded, linear operator $A \in \mathscr{L}(\mathscr{X}, \mathscr{Y}')$ is boundedly invertible* if and only if *its bilinear form satisfies* inf-sup conditions*: ex. $\alpha > 0$ s.t.*

$$\inf_{0 \neq v \in \mathscr{X}} \sup_{0 \neq w \in \mathscr{Y}} \frac{\mathbf{b}(v, w)}{\|v\|_{\mathscr{X}} \|w\|_{\mathscr{Y}}} \geq \alpha \ , \quad \inf_{0 \neq w \in \mathscr{Y}} \sup_{0 \neq v \in \mathscr{X}} \frac{\mathbf{b}(v, w)}{\|v\|_{\mathscr{X}} \|w\|_{\mathscr{Y}}} \geq \alpha \ . \tag{2}$$

If (2) holds then for every $f \in \mathscr{Y}'$ the operator equation

$$\text{find } u \in \mathscr{X} : \quad \mathbf{b}(u, v) = \langle f, v \rangle_{\mathscr{Y}' \times \mathscr{Y}} \quad \forall v \in \mathscr{Y} \tag{3}$$

admits a unique solution $u \in \mathscr{X}$ and there holds $\|u\|_{\mathscr{X}} = \|A^{-1} f\|_{\mathscr{X}} \leq \alpha^{-1} \|f\|_{\mathscr{Y}'}$.

2.2 Parametric Operator Families

We shall be interested in QMC quadratures applied to solutions of *parametric families of operators A*. From partial differential equations with random field input (see, e.g. [27]), we consider, in particular, operator families which depend on infinitely many parameters (obtained, for example, by Karhunen-Loève expansion of random input functions). To this end, we denote by $\boldsymbol{y} := (y_j)_{j\geq 1} \in \mathscr{U}$ the possibly (for random field inputs with nondegenerate covariance kernels) countable set of parameters. We assume the parameters to take values in a bounded parameter domain $\mathscr{U} \subseteq \mathbb{R}^{\mathbb{N}}$. Then, in particular, each realization of $\boldsymbol{y}$ is a sequence of real numbers. Two main cases arise in practice: first, the *"uniform case"*: the parameter domain $\mathscr{U} = [-1/2, 1/2]^{\mathbb{N}}$ and, second, the *"truncated lognormal case"*: the parameter domain $\mathscr{U} \subset \mathbb{R}^{\mathbb{N}}$. In both cases, we account for randomness in inputs by equipping these parameter domains with countable product probability measures (thereby stipulating *mathematical independence* of the random variables y_j). Specifically,

$$\varrho(dy) = \bigotimes_{j\geq 1} \varrho_j(y_j)dy_j \ , \quad \boldsymbol{y} \in \mathscr{U} \tag{4}$$

where, for $j \in \mathbb{N}$, $\varrho_j(y_j) \geq 0$ denotes a probability density on $(-1/2, 1/2)$; for example, $\varrho_j(y_j) = 1$ denotes the uniform density, and in the truncated lognormal case, $\varrho_j = \gamma_1$, the Gaussian measure truncated to the bounded parameter domain $(-1/2, 1/2) \subset \mathbb{R}$, normalized so that $\gamma_1([-1/2, 1/2]) = 1$.

Often, mathematical expectations w.r. to the probability measure ϱ of (functionals of) the solutions $u(\boldsymbol{y})$ of operator equations depending on the parameter vector $\boldsymbol{y}$ are of interest. One object of this note is to address error analysis of QMC evaluation of such, possibly infinite dimensional, integrals. A key role in QMC convergence analysis is played by *parametric regularity* of integrand functions, in terms of weighted (reproducing kernel) Hilbert spaces which were identified in recent years as pivotal for QMC error analysis (see, e.g., [20, 21, 30, 30, 33]) and QMC rule construction (see, e.g., [4, 5, 26]). By $\mathbb{N}_0^{\mathbb{N}}$ we denote the set of all sequences of nonnegative integers, and by $\mathfrak{F} = \{\nu \in \mathbb{N}_0^{\mathbb{N}} : |\nu| < \infty\}$ the set of "finitely supported" such sequences, i.e., sequences of nonnegative integers which have only a finite number of nonzero entries. For $\nu \in \mathfrak{F}$, we denote by $\mathfrak{n} \subset \mathbb{N}$ the set of coordinates j such that $\nu_j \neq 0$, *with j repeated* $\nu_j \geq 1$ *many times*. Analogously, $\mathfrak{m} \subset \mathbb{N}$ denotes the supporting coordinate set for $\mu \in \mathfrak{F}$.

We consider *parametric* families of continuous, linear operators which we denote as $A(\boldsymbol{y}) \in \mathscr{L}(\mathscr{X}, \mathscr{Y}')$. We now make precise the dependence of $A(\boldsymbol{y})$ on the parameter sequence $\boldsymbol{y}$ which is required for our regularity and approximation results.

Assumption 1. *The parametric operator family* $\{A(\boldsymbol{y}) \in \mathscr{L}(\mathscr{X}, \mathscr{Y}') : \boldsymbol{y} \in \mathscr{U}\}$ *is a regular p-analytic operator family for some* $0 < p \leq 1$*, i.e.,*

1. *$A(\boldsymbol{y}) \in \mathscr{L}(\mathscr{X}, \mathscr{Y}')$ is boundedly invertible for every $\boldsymbol{y} \in \mathscr{U}$ with uniformly bounded inverses $A(\boldsymbol{y})^{-1} \in \mathscr{L}(\mathscr{Y}', \mathscr{X})$, i.e., there exists $C_0 > 0$ such that*

$$\sup_{\boldsymbol{y}\in\mathscr{U}} \|A(\boldsymbol{y})^{-1}\|_{\mathscr{L}(\mathscr{Y}',\mathscr{X})} \leq C_0 \tag{5}$$

and

2. *For any fixed $\boldsymbol{y} \in \mathscr{U}$, the operators $A(\boldsymbol{y})$ are analytic with respect to each y_j such that there exists a nonnegative sequence $b = (b_j)_{j\geq 1} \in \ell^p(\mathbb{N})$ such that*

$$\forall \nu \in \mathfrak{F}\backslash\{0\}: \quad \sup_{\boldsymbol{y}\in\mathscr{U}} \left\| (A(0))^{-1}(\partial_{\boldsymbol{y}}^{\nu} A(\boldsymbol{y})) \right\|_{\mathscr{L}(\mathscr{X},\mathscr{X})} \leq C_0 b^{\nu} . \tag{6}$$

Here $\partial_{\boldsymbol{y}}^{\nu} A(\boldsymbol{y}) := \partial_{y_1}^{\nu_1} \partial_{y_2}^{\nu_2} \cdots A(\boldsymbol{y})$; the notation b^{ν} signifies the (finite due to $\nu \in \mathfrak{F}$) product $b_1^{\nu_1} b_2^{\nu_2} \ldots$ where we use the convention $0^0 := 1$.

We verify the abstract assumptions in the particular setting of *affine parameter dependence*; this case arises, for example, in diffusion problems where the diffusion coefficients are given in terms of a Karhunen-Loève expansion (see, e.g. [28] for such Karhunen-Loève expansions and their numerical analysis, in the context of elliptic PDEs with random coefficients). Then, there exists a family $\{A_j\}_{j\geq 0} \subset \mathscr{L}(\mathscr{X}, \mathscr{Y}')$ such that $A(\boldsymbol{y})$ can be written in the form

$$\forall \boldsymbol{y} \in \mathscr{U}: \quad A(\boldsymbol{y}) = A_0 + \sum_{j\geq 1} y_j A_j . \tag{7}$$

We shall refer to $A_0 = A(0)$ as "nominal" operator, and to the operators A_j, $j \geq 1$ as "fluctuation" operators. In order for the sum in (7) to converge, we impose the following assumptions on the sequence $\{A_j\}_{j\geq 0} \subset \mathscr{L}(\mathscr{X}, \mathscr{Y}')$. In doing so, we associate with the operator A_j the bilinear forms $\mathbf{b}_j(\cdot,\cdot) : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}$ via

$$\forall v \in \mathscr{X},\ w \in \mathscr{Y}: \quad \mathbf{b}_j(v,w) =_{\mathscr{Y}} \langle w, A_j v\rangle_{\mathscr{Y}'} , \quad j = 0, 1, 2 \ldots .$$

Assumption 2. *The family $\{A_j\}_{j\geq 0}$ in (7) satisfies the following conditions:*

1. *The "nominal" or "mean field" operator $A_0 \in \mathscr{L}(\mathscr{X}, \mathscr{Y}')$ is boundedly invertible, i.e. (cf. Proposition 1) there exists $\alpha_0 > 0$ such that*

$$\inf_{0\neq v\in\mathscr{X}} \sup_{0\neq w\in\mathscr{Y}} \frac{\mathbf{b}_0(v,w)}{\|v\|_{\mathscr{X}}\|w\|_{\mathscr{Y}}} \geq \alpha_0 , \ \inf_{0\neq w\in\mathscr{Y}} \sup_{0\neq v\in\mathscr{X}} \frac{\mathbf{b}_0(v,w)}{\|v\|_{\mathscr{X}}\|w\|_{\mathscr{Y}}} \geq \alpha_0 . \tag{A1}$$

2. *The "fluctuation" operators $\{A_j\}_{j\geq 1}$ are small with respect to A_0 in the following sense: there exists a constant $0 < \kappa < 2$ such that for α_0 as in (A1) holds*

$$\sum_{j\geq 1} b_j \leq \kappa < 2 , \quad \text{where} \quad b_j := \|A_0^{-1} A_j\|_{\mathscr{L}(\mathscr{X},\mathscr{X})} , \quad j = 1, 2, \ldots . \tag{A2}$$

Condition (**A2**) (and, hence, Assumption 2) is sufficient for the bounded invertibility of $A(\boldsymbol{y})$, uniformly w.r. to the parameter vector $\boldsymbol{y} \in \mathscr{U}$.

Theorem 1. *Under Assumption 2, for every realization $\boldsymbol{y} \in \mathscr{U} = [-1/2, 1/2]^{\mathbb{N}}$ of the parameter vector, the parametric operator $A(\boldsymbol{y})$ is boundedly invertible. Specifically, for the bilinear form $\mathbf{b}(\boldsymbol{y}; \cdot, \cdot) : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}$ associated with $A(\boldsymbol{y}) \in \mathscr{L}(\mathscr{X}, \mathscr{Y}')$ via*

$$\mathbf{b}(\boldsymbol{y}; w, v) := {}_{\mathscr{Y}}\langle v, A(\boldsymbol{y})w\rangle_{\mathscr{Y}'} \tag{8}$$

there hold uniform (w.r. to $\boldsymbol{y} \in \mathscr{U}$) inf-sup conditions (2) with $\alpha = (1-\kappa/2)\alpha_0 > 0$,

$$\forall \boldsymbol{y} \in \mathscr{U} : \quad \inf_{0\neq v\in\mathscr{X}} \sup_{0\neq w\in\mathscr{Y}} \frac{\mathbf{b}(\boldsymbol{y}; v, w)}{\|v\|_{\mathscr{X}}\|w\|_{\mathscr{Y}}} \geq \alpha \ , \quad \inf_{0\neq w\in\mathscr{Y}} \sup_{0\neq v\in\mathscr{X}} \frac{\mathbf{b}(\boldsymbol{y}; v, w)}{\|v\|_{\mathscr{X}}\|w\|_{\mathscr{Y}}} \geq \alpha \ . \tag{9}$$

In particular, for every $f \in \mathscr{Y}'$ and for every $\boldsymbol{y} \in \mathscr{U}$, the parametric operator equation

$$\text{find } u(\boldsymbol{y}) \in \mathscr{X} : \quad \mathbf{b}(\boldsymbol{y}; u(\boldsymbol{y}), v) = \langle f, v\rangle_{\mathscr{Y}'\times\mathscr{Y}} \quad \forall v \in \mathscr{Y} \tag{10}$$

admits a unique solution $u(\boldsymbol{y})$ which satisfies the a-priori estimate

$$\sup_{\boldsymbol{y}\in\mathscr{U}} \|u(\boldsymbol{y})\|_{\mathscr{X}} \leq C\|f\|_{\mathscr{Y}'} \ . \tag{11}$$

Proof. We use Proposition 1, which gives necessary and sufficient conditions for bounded invertibility; also, $1/\alpha$ is a bound for the inverse. By Assumption 2, the nominal part A_0 of $A(\boldsymbol{y})$ in (7) is boundedly invertible, and we write for every $\boldsymbol{y} \in \mathscr{U}$: $A(\boldsymbol{y}) = A_0\left(I + \sum_{j\geq 1} y_j A_0^{-1} A_j\right)$. We see that $A(\boldsymbol{y})$ is boundedly invertible iff the Neumann Series in the second factor is. Since $|y_j| \leq 1/2$, a sufficient condition for this is (**A2**) which implies, with Proposition 1, the assertion with $\alpha = \alpha_0(1-\kappa/2)$. □

From the preceding considerations, the following is readily verified.

Corollary 1. *The affine parametric operator family (7) satisfies Assumption 1 with*

$$C_0 = \frac{1}{(1-\kappa/2)\alpha_0} \quad \text{and} \quad b_j := \|A_0^{-1}A_j\|_{\mathscr{L}(\mathscr{X},\mathscr{X})} \ , \quad \text{for all } j \geq 1 \ .$$

Examples for families of parametric operator equation include certain linear and parabolic evolution equations [12], linear second order wave equations [13], nonlinear elliptic equations [11], elliptic problems in random media with multiple scales [14], and elliptic and parabolic control problems [15].

2.3 Analytic Parameter Dependence of Solutions

The dependence of the solution $u(\boldsymbol{y})$ of the parametric, variational problem (10) on the parameter vector $\boldsymbol{y}$ is analytic, with precise bounds on the growth of the partial derivatives. The following bounds of the parametric solution's dependence on the parameter vector $\boldsymbol{y}$ will, as in [17], allow us to prove dimension independent rates of convergence of QMC quadratures.

Theorem 2. *Under Assumption 1, for every $f \in \mathscr{Y}'$ and for every $\boldsymbol{y} \in \mathscr{U}$, the unique solution $u(\boldsymbol{y}) \in \mathscr{X}$ of the parametric operator equation*

$$A(\boldsymbol{y})\, u(\boldsymbol{y}) = f \qquad \text{in } \mathscr{Y}' \tag{12}$$

depends analytically on the parameters, and the partial derivatives of the parametric solution family $u(\boldsymbol{y})$ satisfy the bounds

$$\sup_{\boldsymbol{y}\in\mathscr{U}} \|(\partial_{\boldsymbol{y}}^{\nu} u)(\boldsymbol{y})\|_{\mathscr{X}} \le C_0 |\nu|!\, \tilde{b}^{\nu} \|f\|_{\mathscr{Y}'} \quad \text{for all } \nu \in \mathfrak{F}, \tag{13}$$

*where $0! := 1$ and where the sequence $\tilde{b} = (\tilde{b}_j)_{j\ge 1} \in \ell^p(\mathbb{N})$ is defined by $\tilde{b}_j = b_j$ with b_j as in (**A2**) in the affine case (7), and with $\tilde{b}_j = b_j / \ln 2$ for all $j \in \mathbb{N}$ in the truncated lognormal case (6).*

For a proof in the case of stationary diffusion problems we refer, for example, to [3], for control problems to [15]. The regularity estimates (13) (and, therefore, also sparsity and, as shown ahead, QMC convergence) results are available for linear parabolic and hyperbolic PDE problems [12, 13], and for solutions of nonlinear, parametric initial value problems on possibly infinite dimensional state spaces [10].

2.4 Spatial Regularity of Solutions

Convergence rates of Galerkin discretizations will require regularity of the parametric solution $u(\boldsymbol{y})$. To state it, we assume given *scales of smoothness spaces* $\{\mathscr{X}_t\}_{t\ge 0}$ and $\{\mathscr{Y}_t'\}_{t\ge 0}$, with

$$\mathscr{X} = \mathscr{X}_0 \supset \mathscr{X}_1 \supset \mathscr{X}_2 \supset \dots, \qquad \mathscr{Y}' = \mathscr{Y}_0' \supset \mathscr{Y}_1' \supset \mathscr{Y}_2' \supset \dots. \tag{14}$$

The scales $\{\mathscr{X}_t\}_{t\ge 0}$ and $\{\mathscr{Y}_t'\}_{t\ge 0}$ (and analogously $\{\mathscr{X}_t'\}_{t\ge 0}$, $\{\mathscr{Y}_t\}_{t\ge 0}$) are defined for noninteger values of $t \ge 0$ by interpolation.

Instances of smoothness scales (14) in the context of the diffusion problem considered in [3, 17] are, in a *convex domain* D, the choices $\mathscr{X} = H_0^1(D)$, $\mathscr{X}_1 = (H^2 \cap H_0^1)(D)$, $\mathscr{Y}' = H^{-1}(D)$, $\mathscr{Y}_1' = L^2(D)$. In a nonconvex polygon (or polyhedron), analogous smoothness scales are available, but involve Sobolev spaces with weights (see, e.g., [25]). In the ensuing convergence analysis of QMC – Galerkin discretizations of (12), we assume $f \in \mathscr{Y}_t'$ for some $t > 0$ implies that

$$\sup_{y\in\mathscr{U}} \|u(y)\|_{\mathscr{X}_t} = \sup_{y\in\mathscr{U}} \|A(y)^{-1} f\|_{\mathscr{X}_t} \le C_t \|f\|_{\mathscr{Y}_t'} . \tag{15}$$

Such regularity is available for a wide range of parametric differential equations (see [10, 15, 27] and the references there). For the analysis of Multi-Level QMC Galerkin discretizations, however, stronger bounds which combined (15) and (13) are necessary (see [19]).

2.5 Discretization

As the inverse $A(y)^{-1}$ is not available explicitly, we will have to compute, for given QMC quadrature points $\boldsymbol{y} \in \mathscr{U}$, an approximate inverse. We consider the case when it is obtained by *Galerkin discretization*: we assume given two one-parameter families $\{\mathscr{X}^h\}_{h>0} \subset \mathscr{X}$ and $\{\mathscr{Y}^h\}_{h>0} \subset \mathscr{Y}$ of subspaces of equal, finite dimension N_h, which are dense in $\mathscr{X}$ resp. in $\mathscr{Y}$, i.e.

$$\forall u \in \mathscr{X} : \quad \limsup_{h\to 0} \inf_{0\neq u^h\in\mathscr{X}^h} \|u - u^h\|_{\mathscr{X}} = 0 \tag{16}$$

and likewise for $\{\mathscr{Y}^h\}_{h>0} \subset \mathscr{Y}$. We also assume the *approximation property*:

$$\forall 0 < t \le \bar{t} : \exists C_t > 0 : \forall u \in \mathscr{X}_t \ \forall 0 < h \le h_0 : \quad \inf_{w^h\in\mathscr{X}^h} \|u - w^h\|_{\mathscr{X}} \le C_t h^t \|u\|_{\mathscr{X}_t} . \tag{17}$$

The maximum amount of smoothness in the scale $\mathscr{X}_t$, denoted by $\bar{t}$, depends of the problem class under consideration and on the Sobolev scale: e.g. for elliptic problems in polygonal domains, it is well known that choosing for $\mathscr{X}_t$ the usual Sobolev spaces will allow (15) with t only in a rather small interval $0 < t \le \bar{t}$, whereas choosing $\mathscr{X}_t$ as *weighted Sobolev spaces* will allows large values of $\bar{t}$ (see [25]).

Proposition 2. *Assume that the subspace sequences* $\{\mathscr{X}^h\}_{h>0} \subset \mathscr{X}$ *and* $\{\mathscr{Y}^h\}_{h>0} \subset \mathscr{Y}$ *are stable, i.e. that there exists* $\bar{\alpha} > 0$ *and* $h_0 > 0$ *such that for every* $0 < h \le h_0$, *there hold the* uniform (w.r. to $\boldsymbol{y} \in \mathscr{U}$) discrete inf-sup conditions

$$\forall \boldsymbol{y} \in \mathscr{U} : \quad \inf_{0\neq v^h\in\mathscr{X}^h} \sup_{0\neq w^h\in\mathscr{Y}^h} \frac{\mathbf{b}(\boldsymbol{y}; v^h, w^h)}{\|v^h\|_{\mathscr{X}} \|w^h\|_{\mathscr{Y}}} \ge \bar{\alpha} > 0 \tag{18}$$

and

$$\forall \boldsymbol{y} \in \mathscr{U} : \quad \inf_{0\neq w^h\in\mathscr{Y}^h} \sup_{0\neq v^h\in\mathscr{X}^h} \frac{\mathbf{b}(\boldsymbol{y}; v^h, w^h)}{\|v^h\|_{\mathscr{X}} \|w^h\|_{\mathscr{Y}}} \ge \bar{\alpha} > 0 . \tag{19}$$

Then, for every $0 < h \leq h_0$, *and for every* $\boldsymbol{y} \in \mathscr{U}$, *the Galerkin approximation* $u^h \in \mathscr{X}^h$, *given by*

$$\text{find } u^h(\boldsymbol{y}) \in \mathscr{X}^h : \quad \mathbf{b}(\boldsymbol{y}; u^h(\boldsymbol{y}), v^h) = \langle f, v^h \rangle_{\mathscr{Y}' \times \mathscr{Y}} \quad \forall v^h \in \mathscr{Y}^h \tag{20}$$

admits a unique solution $u^h(\boldsymbol{y})$ *which satisfies the a-priori estimate*

$$\sup_{\boldsymbol{y} \in \mathscr{U}} \|u^h(\boldsymbol{y})\|_{\mathscr{X}} \leq \bar{\alpha}^{-1} \|f\|_{\mathscr{Y}'} \,. \tag{21}$$

Moreover, there exists a constant $C > 0$ *such that for all* $\boldsymbol{y} \in \mathscr{U}$ *holds* quasioptimality

$$\|u(\boldsymbol{y}) - u^h(\boldsymbol{y})\|_{\mathscr{X}} \leq C \bar{\alpha}^{-1} \inf_{0 \neq w^h \in \mathscr{X}^h} \|u(\boldsymbol{y}) - w^h\|_{\mathscr{X}} \,. \tag{22}$$

We remark that under Assumption 2, the validity of the discrete inf-sup conditions (18), (19) for the "nominal" bilinear forms $\mathbf{b}_0(\boldsymbol{y}; \cdot, \cdot)$ with constant $\bar{\alpha}_0 > 0$ independent of h implies (18), (19) for the form $\mathbf{b}(\boldsymbol{y}; \cdot, \cdot)$ with constant $\bar{\alpha} = (1 - \kappa/2)\bar{\alpha}_0 > 0$.

3 QMC Integration

For a given bounded, linear functional $G(\cdot) : \mathscr{X} \to \mathbb{R}$, we are interested in computing expected values of

$$F(\boldsymbol{y}) := G(u(\cdot, \boldsymbol{y})) \,, \qquad \boldsymbol{y} \in \mathscr{U} \,, \tag{23}$$

(respectively of its parametric Galerkin approximation $u^h(\boldsymbol{y}) \in \mathscr{X}_h \subset \mathscr{X}$ defined in (20)). The expected value of F is an infinite-dimensional, iterated integral of the functional $G(\cdot)$ of the parametric solution:

$$\int_{\mathscr{U}} F(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} = \int_{\mathscr{U}} G(u(\cdot, \boldsymbol{y})) \, \mathrm{d}\boldsymbol{y} = G\left(\int_{\mathscr{U}} u(\cdot, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} \right) . \tag{24}$$

The issue is thus the numerical evaluation of Bochner integrals of $\mathscr{X}$-valued functions over the infinite dimensional domain of integration $\mathscr{U}$. We also observe that for the parametric operator equation (12), to evaluate F at a single QMC point $\boldsymbol{y} \in \mathscr{U}$ requires the approximate (Galerkin) solution of one instance of the operator equation for $u(\cdot, \boldsymbol{y}) \in \mathscr{X}$. This introduces an additional *Galerkin discretization error*, and can be accounted for as in [17] in the present, more general, setting with analogous proofs.

In [3] and the present paper, the summability of the fluctuation operators A_j, $j \geq 1$, plays an important role for proving dimension-independent convergence

rates of approximations of the parametric solution maps. Accordingly, we will make the assumption, stronger than Assumption (**A2**) that there exists $0 < p < 1$ such that

$$\sum_{j\geq 1} \|A_j\|^p_{\mathscr{L}(\mathscr{X},\mathscr{Y}')} < \infty\,. \tag{**A3**}$$

Notice that this condition is, by (**A1**), equivalent to $(b_j)_{j\geq 1} \in \ell^p(\mathbb{N})$, and implies decay of the fluctuation coefficients A_j, with stronger decay as the value of p becomes smaller. In both [3, 17] and the present paper, the rate of convergence $\mathscr{O}(N^{-1+\delta})$ is attained if (**A3**) is satisfied with $p = 2/3$. Here and throughout what follows, N denotes the number of points used in QMC integration. For values of p between $2/3$ and 1, the rate of convergence in both cases is $\mathscr{O}(N^{-(1/p-1/2)})$.

Recall that the purpose of the present paper is to analyze accuracy and complexity of QMC methods in connection with the Galerkin approximation (20) of (10). To obtain convergence rates, we strengthen Assumption (**A2**) to the requirement

$$\sup_{\boldsymbol{y}\in\mathscr{U}} \|A(\boldsymbol{y})^{-1}\|_{\mathscr{L}(\mathscr{Y}'_t,\mathscr{X}_t)} < \infty\,, \quad 0 \leq t \leq 1\,. \tag{**A4**}$$

For application of QMC quadrature rules, the infinite sum in (7) must be truncated to a finite sum of, say, s terms. Below, the parameter s shall be referred to as "QMC-truncation dimension". In order for the dimension truncation to be meaningful, we will assume additionally that the A_j are decreasingly, i.e. the sequence of bounds b_j in (**A2**) is nonincreasing:

$$b_1 \geq b_2 \geq \ldots \geq b_j \geq \cdots\,. \tag{**A5**}$$

The overall error for the QMC-Galerkin approximation is then a sum of three terms: a *truncation error*, a *QMC error*, and the *Galerkin discretization error*. We bound the three errors and finally combine them to arrive at an overall QMC-Galerkin error bound.

3.1 Finite Dimensional Setting

In this subsection we review QMC integration when the truncation dimension (i.e. the number of integration variables), denoted by s, is assumed to be finite and fixed. The domain of integration is taken to be the s-dimensional unit cube $[-\frac{1}{2},\frac{1}{2}]^s$ centered at the origin so that QMC integration methods formulated for $[0,1]^s$ may require a coordinate translation. We thus consider integrals of the form

$$I_s(F) := \int_{[-\frac{1}{2},\frac{1}{2}]^s} F(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}\,. \tag{25}$$

In our later applications F will be of the form (23), but for the present it is general and depends only on s variables. An N-point QMC approximation to this integral is an equal-weight rule of the form

$$Q_{s,N}(F) := \frac{1}{N}\sum_{i=1}^{N} F(\boldsymbol{y}^{(i)}),$$

with carefully chosen points $\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(N)} \in [-\frac{1}{2}, \frac{1}{2}]^s$. For classical results on QMC methods, see, e.g. [24, 29].

We shall assume that our integrand F belongs to a *weighted* and *anchored* Sobolev space $\mathscr{W}^a_{s,\boldsymbol{\gamma}}$. This is a Hilbert space over the unit cube $[-\frac{1}{2}, \frac{1}{2}]^s$ with norm given by

$$\|F\|^2_{\mathscr{W}^a_{s,\boldsymbol{\gamma}}} := \sum_{\mathfrak{u}\subseteq\{1:s\}} \frac{1}{\gamma_{\mathfrak{u}}} \int_{[-\frac{1}{2},\frac{1}{2}]^{|\mathfrak{u}|}} \left| \frac{\partial^{|\mathfrak{u}|} F}{\partial \boldsymbol{y}_{\mathfrak{u}}}(\boldsymbol{y}_{\mathfrak{u}}; 0) \right|^2 \mathrm{d}\boldsymbol{y}_{\mathfrak{u}}\,, \tag{26}$$

where $\{1:s\}$ is a shorthand notation for the set of indices $\{1, 2, \ldots, s\}$, $\frac{\partial^{|\mathfrak{u}|} F}{\partial \boldsymbol{y}_{\mathfrak{u}}}$ denotes the mixed first derivative with respect to the variables y_j with $j \in \mathfrak{u}$, and $(\boldsymbol{y}_{\mathfrak{u}}; 0)$ denotes the vector whose jth component is y_j if $j \in \mathfrak{u}$ and 0 if $j \notin \mathfrak{u}$.

A closely related family of weighted spaces are the so-called *unanchored spaces* denoted by $\mathscr{W}^u_{s,\boldsymbol{\gamma}}$. Here, "inactive" arguments of integrands are averaged, rather than fixed at the origin as in (26). Accordingly, the *un*anchored norm $\|\circ\|_{\mathscr{W}^u_{s,\boldsymbol{\gamma}}}$ is given by

$$\|F\|^2_{\mathscr{W}^u_{s,\boldsymbol{\gamma}}} := \sum_{\mathfrak{u}\subseteq\{1:s\}} \frac{1}{\gamma_{\mathfrak{u}}} \int_{[-\frac{1}{2},\frac{1}{2}]^{|\mathfrak{u}|}} \left(\int_{[-\frac{1}{2},\frac{1}{2}]^{s-|\mathfrak{u}|}} \frac{\partial^{|\mathfrak{u}|} F}{\partial \boldsymbol{y}_{\mathfrak{u}}}(\boldsymbol{y}_{\mathfrak{u}}; \boldsymbol{y}_{\{1:s\}\setminus\mathfrak{u}}) \mathrm{d}\boldsymbol{y}_{\{1:s\}\setminus\mathfrak{u}} \right)^2 \mathrm{d}\boldsymbol{y}_{\mathfrak{u}}\,. \tag{27}$$

We omit the superscripts a and u in statements which apply for either choice of space; we will also require $u \in \mathscr{W}^a_{s,\boldsymbol{\gamma}}(U; \mathscr{X})$ which is defined as the Bochner space of strongly measurable, $\mathscr{X}$-valued functions for which the (26) (with the $\|\circ\|_{\mathscr{X}}$ norm in place of the absolute value) is finite.

Weighted, anchored spaces $\mathscr{W}^a_{s,\boldsymbol{\gamma}}$ were first introduced by Sloan and Woźniakowski in [32]. By now there are many variants and generalizations, see e.g. [7, 31] and the references there. In (26) the "anchor" is $(0, \ldots, 0)$, the center of the unit cube $[\frac{1}{2}, \frac{1}{2}]^s$, corresponding to the anchor $(\frac{1}{2}, \ldots, \frac{1}{2})$ in the standard unit cube $[0, 1]^s$. For parametric operator equations (12) anchoring at the origin is preferable, since *the parametric solution of* (12) *with anchored operators corresponds to the anchored parametric solution*.

Regarding the choice of weights, from derivative bounds (13), in [17] *product and order dependent ("POD" for short) weights* were derived which are given by

$$\gamma_{\mathfrak{u}} = \Gamma_{|\mathfrak{u}|} \prod_{j\in\mathfrak{u}} \gamma_j > 0\,. \tag{28}$$

Here $|\mathfrak{u}|$ denotes the cardinality (or the "order") of $\mathfrak{u}$. The weights are therefore determined by a specific choice of the sequences $\Gamma_0 = \Gamma_1 = 1, \Gamma_2, \Gamma_3, \ldots$ and $\gamma_1, \gamma_2, \gamma_3, \ldots$. (a precise choice of $\gamma_{\mathfrak{u}}$ will be given in (38) ahead).

QMC error analysis is based on the *worst case error* of a QMC rule (or a family of QMC rules). It is defined as supremum of the (bounded, linear) QMC error functional over all functions in the unit ball of $\mathscr{W}_{s,\boldsymbol{\gamma}}$:

$$e^{\mathrm{wor}}(Q_{s,N}; \mathscr{W}_{s,\boldsymbol{\gamma}}) := \sup_{\|F\|_{\mathscr{W}_{s,\boldsymbol{\gamma}}} \le 1} |I_s(F) - Q_{s,N}(F)| \,. \tag{29}$$

Due to linearity of the functionals $I_s(\cdot)$ and $Q_{s,N}(\cdot)$, we have

$$|I_s(F) - Q_{s,N}(F)| \le e^{\mathrm{wor}}(Q_{s,N}; \mathscr{W}_{s,\boldsymbol{\gamma}}) \, \|F\|_{\mathscr{W}_{s,\boldsymbol{\gamma}}} \qquad \text{for all} \quad F \in \mathscr{W}_{s,\boldsymbol{\gamma}} \,. \tag{30}$$

In shifted rank-1 lattice rules, quadrature points in $\mathscr{U}$ are given by

$$\boldsymbol{y}^{(i)} = \mathrm{frac}\left(\frac{i\boldsymbol{z}}{N} + \boldsymbol{\Delta}\right) - \left(\tfrac{1}{2}, \ldots, \tfrac{1}{2}\right) \,, \qquad i = 1, \ldots, N \,,$$

where $\boldsymbol{z} \in \mathbb{Z}^s$ is the *generating vector*, $\boldsymbol{\Delta} \in [0,1]^s$ is the *shift*, and $\mathrm{frac}(\cdot)$ indicates the fractional part of each component in the vector. Subtraction by the vector $(\frac{1}{2}, \ldots, \frac{1}{2})$ translates the rule from $[0,1]^s$ to $[-\frac{1}{2}, \frac{1}{2}]^s$. In *randomly shifted lattice rules* the shift $\boldsymbol{\Delta}$ is a vector with independent, uniformly in $[0,1)$ distributed components; we denote the application of the QMC rule to the integrand function F for one draw of the shift $\boldsymbol{\Delta}$ by $Q_{s,N}(\boldsymbol{\Delta}; F)$.

Theorem 3 ([16, Theorem 5]). *Let $s, N \in \mathbb{N}$ be given, and assume that $F \in \mathscr{W}_{s,\boldsymbol{\gamma}}$ for a particular choice of weights $\boldsymbol{\gamma}$, with $\mathscr{W}_{s,\boldsymbol{\gamma}}$ denoting either the anchored space with norm (26) or the unanchored space with norm (27).*

In each case, there exists a randomly shifted lattice such that its root-mean-square error (with respect to averages over all shifts) satisfies, for all $\lambda \in (1/2, 1]$,

$$\sqrt{\mathbb{E}\left[|I_s(F) - Q_{s,N}(\cdot; F)|^2\right]} \le \left(\sum_{\emptyset \ne \mathfrak{u} \subseteq \{1:s\}} \gamma_{\mathfrak{u}}^{\lambda} \rho(\lambda)^{|\mathfrak{u}|}\right)^{1/(2\lambda)} [\varphi(N)]^{-1/(2\lambda)} \, \|F\|_{\mathscr{W}_{s,\boldsymbol{\gamma}}} \,, \tag{31}$$

where $\mathbb{E}[\cdot]$ denotes the expectation with respect to the random shift which is uniformly distributed over $[0,1]^s$. In (31), with $\zeta(x)$ denotes the Riemann zeta function, and $\varphi(N)$ the Euler totient function which satisfies $\varphi(N) \le 9N$ for all $N \le 10^{30}$,

$$\rho(\lambda) := \begin{cases} \dfrac{2\zeta(2\lambda)}{(2\pi^2)^\lambda} + \dfrac{1}{12^\lambda} & \text{if} \quad \mathscr{W}_{s,\boldsymbol{\gamma}} = \mathscr{W}^a_{s,\boldsymbol{\gamma}} \,, \\ \dfrac{2\zeta(2\lambda)}{(2\pi^2)^\lambda} & \text{if} \quad \mathscr{W}_{s,\boldsymbol{\gamma}} = \mathscr{W}^u_{s,\boldsymbol{\gamma}} \,. \end{cases} \tag{32}$$

The result with general weights, restricted to prime N in the anchored case was first obtained in [31, Theorem 3(A)], for general N and unanchored spaces in [16, Theorem 4.1] (with the choice $m = 0$ in the statement of that theorem), and for general N and anchored spaces in [16, Theorem 4.1],

The question of *efficient construction of lattice rules* has received much attention in recent years [30]. Algorithms which obtain the generating vector with favourable (w.r. to N and s) scaling have been obtained for integrands in unanchored spaces in [26], where the first algorithm for fast CBC construction using FFT at a cost of $O(sN \log N)$ was given. Efficient algorithms for construction of so-called embedded families of lattice rules where proposed in [4]. We refer to [16, 18] for a discussion.

3.2 Dimensional Truncation

Given $s \in \mathbb{N}$ and $\boldsymbol{y} \in \mathscr{U}$, we observe that truncating the sum in (7) at s terms amounts to setting $y_j = 0$ for $j > s$. We thus denote by $u^s(\boldsymbol{x}, \boldsymbol{y}) := u(\boldsymbol{x}, (\boldsymbol{y}_{\{1:s\}}; 0))$ the solution of the parametric weak problem (10) corresponding to the parametric operator $A((\boldsymbol{y}_{\{1:s\}}; 0))$ in which the sum (7) is truncated at s terms. Then Theorem 1 remains valid *with constants independent of* s when $u(\cdot, \boldsymbol{y})$ is replaced by its dimensionally truncated approximation $u^s(\cdot, \boldsymbol{y})$.

Theorem 4. *Under Assumptions* (**A2**), (**A3**), (**A5**), *for every* $f \in \mathscr{Y}'$ *and for every* $\boldsymbol{y} \in \mathscr{U}$ *and for every* $s \in \mathbb{N}$, *the dimensionally truncated, parametric solution* $u^s(\cdot, \boldsymbol{y}) = u(\cdot, (\boldsymbol{y}_{\{1:s\}}; 0))$ *of the* s*-term truncated parametric weak problem* (10) *satisfies, with* b_j *as defined in* (**A2**),

$$\|u(\cdot, \boldsymbol{y}) - u^s(\cdot, \boldsymbol{y})\|_{\mathscr{X}} \leq C\alpha^{-1} \|f\|_{\mathscr{Y}'} \sum_{j \geq s+1} b_j \tag{33}$$

for some constant $C > 0$ *independent of* s, $\boldsymbol{y}$ *and* f. *For every* $G(\cdot) \in \mathscr{X}'$

$$|I(G(u)) - I_s(G(u))| \leq \tilde{C}\, \alpha^{-1} \|f\|_{\mathscr{Y}'} \|G(\cdot)\|_{\mathscr{X}'} \left(\sum_{j \geq s+1} b_j \right)^2 \tag{34}$$

for some constant $\tilde{C} > 0$ *independent of* s, f *and* $G(\cdot)$. *In addition, if Assumptions* (**A3**) *and* (**A5**) *hold, then*

$$\sum_{j\geq s+1} b_j \;\leq\; \min\left(\frac{1}{1/p-1},1\right)\left(\sum_{j\geq 1} b_j^p\right)^{1/p} s^{-(1/p-1)}\,.$$

This result is proved in the affine case (7) in [17, Theorem 5.1], and for operators depending lognormally on $\boldsymbol{y}$ in [2]. It will hold for general probability densities $\varrho(y)$ in (4) whenever the factor measures $\varrho_j(dy_j)$ are centered.

4 Analysis of QMC and Galerkin Disretization

We apply QMC quadrature $Q_{s,N}$ to the dimensionally truncated approximaion $I_s(G(u))$ of the integral (24), where the integrand $F(\boldsymbol{y}) = G(u(\cdot,\boldsymbol{y}))$ is a continuous, linear functional $G(\cdot)$ of the parametric solution $u(\cdot,\boldsymbol{y})$ of the operator equation (10).

As proposed in [7,23], choices of QMC weights can be based on minimizing the product of worst case error and of (upper bounds for) the weighted norms $\|F\|_{\mathscr{W}_{s,\boldsymbol{\gamma}}}$ in the error bound (30). This idea was combined with the bounds (13) in [17] to identify POD QMC weights (28) as sufficient to ensure a QMC convergence rate of $O(N^{-1+\delta})$ with $O()$ being independent of the truncation dimension s. Another issue raised by the infinite dimensional nature of the problem is to choose the value of s and estimate the truncation error $I(G(u)) - I_s(G(u))$, which was estimated in Theorem 4. The following QMC quadrature error bound is proved in [17, Theorem 5.1] for scalar, parametric diffusion problems; its statement and proof generalize to the parametric operator equations (12) with solution regularity (13).

Theorem 5 (Root-mean-square error bound). *Under Assumptions* (**A2**) *and* (9) *let b_j be defined as in* (**A2**). *For every $f \in \mathscr{Y}'$ and for every $G(\cdot) \in \mathscr{X}'$, let $u(\cdot,\boldsymbol{y})$ denote the solution of the parametric variational problem* (10).

Then for $s, N \in \mathbb{N}$ and weights $\boldsymbol{\gamma} = (\gamma_{\mathfrak{u}})$, randomly shifted lattice rules $Q_{s,N}(\cdot;\cdot)$ with N points in s dimensions can be constructed by a component-by-component algorithm such that the root-mean-square error for approximating the finite dimensional integral $I_s(G(u))$ satisfies, for all $\lambda \in (1/2,1]$, and all $N \leq 10^{30}$

$$\sqrt{\mathbb{E}\left[|I_s(G(u)) - Q_{s,N}(\cdot;G(u))|^2\right]} \;\leq\; \frac{C_{\boldsymbol{\gamma}}(\lambda)}{\alpha} N^{-1/(2\lambda)} \|f\|_{\mathscr{Y}'} \|G(\cdot)\|_{\mathscr{X}'}\,, \qquad (35)$$

where $\mathbb{E}[\cdot]$ denotes the expectation with respect to the random shift $\boldsymbol{\Delta}$ (uniformly distributed over $[0,1]^s$) and $C_{\boldsymbol{\gamma}}(\lambda)$ is independent of s as in [17, Eq. (6.2)].

In [17, Theorem 6.1], a choice of weights which minimizes the upper bound was derived. As the derivation in [17, Theorem 6.1] generalizes verbatim to the presently considered setting we only state the result. Under the assumptions of Theorem 5, for b_j as in (**A2**) suppose that (**A3**) holds, i.e.

$$\sum_{j\geq 1} b_j^p \;<\; \infty \qquad \text{for some} \quad 0<p<1\,, \tag{36}$$

For the choice

$$\lambda \;:=\; \begin{cases} \frac{1}{2-2\delta} & \text{for some} \quad \delta\in(0,1/2) \quad \text{when } p\in(0,2/3]\,, \\ \frac{p}{2-p} & \qquad\qquad\qquad\quad \text{when } p\in(2/3,1)\,, \end{cases} \tag{37}$$

the choice of weights

$$\gamma_{\mathfrak{u}} \;=\; \gamma_{\mathfrak{u}}^* \;:=\; \left(|\mathfrak{u}|!\prod_{j\in\mathfrak{u}}\frac{b_j}{\sqrt{\rho(\lambda)}}\right)^{2/(1+\lambda)} \tag{38}$$

with $\rho(\lambda)$ in (32) minimizes the constant $C_{\boldsymbol{\gamma}}(\lambda)$ in the bound (35). To account for the impact of Galerkin discretization of the operator equation, recall Sect. 2.5. For any $\boldsymbol{y}\in\mathscr{U}$, the *parametric FE approximation* $u^h(\cdot,\boldsymbol{y})\in\mathscr{X}^h$ is defined as in (20). Here, $\boldsymbol{b}(\boldsymbol{y};\cdot,\cdot)$ denotes the parametric bilinear form (8). In particular the FE approximation (20) is defined *pointwise* with respect to the parameter $\boldsymbol{y}\in\mathscr{U}$.

Theorem 6. *Under Assumptions* (**A2**), (9) *and* (15) *for every* $f\in\mathscr{Y}'$ *and for every* $\boldsymbol{y}\in\mathscr{U}$, *the approximations* $u^h(\cdot,\boldsymbol{y})$ *are stable, i.e.* (21) *holds. For every* $f\in\mathscr{Y}_t'$ *with* $0<t\leq 1$ *exists a constant* $C>0$ *such that for all* $\boldsymbol{y}\in\mathscr{U}$ *as* $h\to 0$ *holds*

$$\sup_{\boldsymbol{y}\in\mathscr{U}}\|u(\cdot,\boldsymbol{y})-u^h(\cdot,\boldsymbol{y})\|_{\mathscr{X}} \;\leq\; C\,h^t\,\|f\|_{\mathscr{Y}_t'}\,. \tag{39}$$

Proof. Since $f\in\mathscr{Y}_t'$ for some $t>0$ implies with (15) that $u(\boldsymbol{y})\in\mathscr{X}_t$ and, with the approximation property (22),

$$\|u(\cdot,\boldsymbol{y})-u^h(\cdot,\boldsymbol{y})\|_{\mathscr{X}} \;\leq\; C\,h^t\,\|u(\cdot,\boldsymbol{y})\|_{\mathscr{X}_t}$$

where the constant C is independent h and of $\boldsymbol{y}$. This proves (39). □

Since we are interested in estimating the error in approximating functionals (24), we will also impose a regularity assumption on the functional $G(\cdot)\in\mathscr{X}'$:

$$\exists\, 0<t'\leq 1:\quad G(\cdot)\in\mathscr{X}_{t'}' \tag{40}$$

and the *adjoint regularity*: for t' as in (40), and for every $\boldsymbol{y}\in\mathscr{U}$,

$$w(\boldsymbol{y})=(A^*(\boldsymbol{y}))^{-1}G\in\mathscr{Y}_{t'}\,,\quad \sup_{\boldsymbol{y}\in\mathscr{U}}\|w(\boldsymbol{y})\|_{\mathscr{Y}_{t'}}\leq C\|G\|_{\mathscr{X}_{t'}'}\,. \tag{41}$$

Moreover, since in the expression (23) only a bounded linear functional $G(\cdot)$ of u rather than the parametric solution u itself enters, the discretization error of $G(u)$ is

of main interest in QMC error analysis. An Aubin-Nitsche duality argument shows that $|G(u(\cdot,\boldsymbol{y})) - G(u^h(\cdot,\boldsymbol{y}))|$ converges faster than $\|u(\cdot,\boldsymbol{y}) - u^h(\cdot,\boldsymbol{y})\|_{\mathscr{X}}$: under Assumptions (**A2**), (9), (**A4**), and (15), (41) there exists a constant $C > 0$ such that for every $f \in \mathscr{Y}_t'$ with $0 < t \leq 1$, for every $G(\cdot) \in \mathscr{X}_{t'}'$ with $0 < t' \leq 1$ and for every $\boldsymbol{y} \in \mathscr{U}$, as $h \to 0$, the Galerkin approximations $G(u^h(\cdot,\boldsymbol{y}))$ satisfy

$$\left|G(u(\cdot,\boldsymbol{y})) - G(u^h(\cdot,\boldsymbol{y}))\right| \;\leq\; C\, h^{\tau}\, \|f\|_{\mathscr{Y}_t'}\, \|G(\cdot)\|_{\mathscr{X}_{t'}'}\,, \tag{42}$$

where $0 < \tau := t + t'$ and where the constant $C > 0$ is independent of $\boldsymbol{y} \in \mathscr{U}$.

We conclude with bounds for the combined QMC FE approximation of the integral (24). To define the approximation of (24), we approximate the infinite dimensional integral using a randomly shifted lattice rule with N points in s dimensions. The QMC rule with N points for integration over $(-1/2, 1/2)^s$ for one single draw $\boldsymbol{\Delta}$ of the shift will be denoted by $Q_{s,N}(\cdot;\boldsymbol{\Delta})$. For each evaluation of the integrand F, we replace the exact solution $u(\cdot,\boldsymbol{y})$ of the parametric weak problem (10) by the Galerkin approximation $u^h(\cdot,\boldsymbol{y})$ in the subspace $\mathscr{X}^h \subset \mathscr{X}$ of dimension $M^h := \dim \mathscr{X}^h < \infty$.

Thus we may express the overall error as a sum of a *dimension truncation error* (which is implicit when a finite dimensional QMC method is used for an infinite dimensional integral), a *QMC quadrature error*, and a *FE discretization error*:

$$\begin{aligned} &I(G(u)) - Q_{s,N}(G(u^h);\boldsymbol{\Delta}) \\ &\quad = (I - I_s)(G(u)) \;+\; (I_s(G(u)) - Q_{s,N}(G(u);\boldsymbol{\Delta})) \;+\; Q_{s,N}(G(u-u^h);\boldsymbol{\Delta})\,. \end{aligned}$$

We bound the mean-square error with respect to the random shift by

$$\begin{aligned} \mathbb{E}\left[|I(G(u)) - Q_{s,N}(G(u^h);\cdot)|^2\right] \leq\; & 3\,|(I - I_s)(G(u))|^2 \\ & + 3\,\mathbb{E}\left[|I_s(G(u)) - Q_{s,N}(G(u);\cdot)|^2\right] + 3\,\mathbb{E}\left[|Q_{s,N}(G(u-u^h);\cdot)|^2\right]\,. \end{aligned} \tag{43}$$

The dimension truncation error, i.e., the first term in (43), was estimated in Theorem 4. The QMC error, i.e., the second term in (43), is already analyzed in Theorem 5. Finally, for the Galerkin projection error, i.e., for the third term in (43), we apply the property that the QMC quadrature weights $1/N$ are positive and sum to 1, to obtain

$$\mathbb{E}\left[|Q_{s,N}(G(u-u^h);\cdot)|^2\right] \;\leq\; \sup_{\boldsymbol{y}\in\mathscr{U}} |G(u(\cdot,\boldsymbol{y}) - u^h(\cdot,\boldsymbol{y}))|^2\,,$$

and apply (42) Then, under the assumptions in Theorems 4, 5 and in (42), we approximate the dimensionally truncated approximation (25) of the integral (24) over $\mathscr{U}$ by the randomly shifted lattice rule from Theorem 5 with N points in s dimensions. For each lattice point we solve the approximate problem (20) with *one common subspace* $\mathscr{X}^h \subset \mathscr{X}$ with $M_h = \dim(\mathscr{X}^h)$ degrees of freedom and with the approximation property (17). Then, there holds the root-mean-square error bound

$$\sqrt{\mathbb{E}\left[|I(G(u)) - Q_{s,N}(\cdot; G(u^h))|^2\right]}$$
$$\leq C\left(\kappa(s,N)\,\|f\|_{\mathscr{Y}'}\,\|G(\cdot)\|_{\mathscr{X}'} + h^\tau\,\|f\|_{\mathscr{Y}'_t}\|G(\cdot)\|_{\mathscr{X}'_{t'}}\right)\,,$$

where $\tau = t + t'$, and, assuming $\varphi(N) \leq CN$, for fixed $\delta > 0$ arbitrary small,

$$\kappa(s,N) = \begin{cases} s^{-2(1/p-1)} & + N^{-(1-\delta)} \quad \text{when } p \in (0, 2/3]\,, \\ s^{-2(1/p-1)} & + N^{-(1/p-1/2)} \quad \text{when } p \in (2/3, 1)\,. \end{cases}$$

Acknowledgements The author is supported by ERC under Grant AdG 247277.

References

1. Brezzi, F., Fortin, M.: Mixed and Hybrid Finite Element Methods. Springer, Berlin (1991)
2. Charrier, J.: Strong and weak error estimates for elliptic partial differential equations with random coefficients. SIAM J. Numer. Anal. **50**, 216–246 (2012)
3. Cohen, A., DeVore, R., Schwab, Ch.: Convergence rates of best N-term Galerkin approximation for a class of elliptic sPDEs., Found. Comput. Math. **10**, 615–646 (2010).
4. Cools, R., Kuo, F.Y., Nuyens, D.: Constructing embedded lattice rules for multivariate integration. SIAM J. Sci. Comput. **28**, 2162–2188 (2006)
5. Dick, J.: On the convergence rate of the component-by-component construction of good lattice rules. J. Complexity **20**, 493–522 (2004)
6. Dick, J., Pillichshammer, F.: Digital Nets and Sequences. Cambridge University Press, Cambridge, UK (2010)
7. Dick, J., Sloan, I.H., Wang, X., Woźniakowski, H.: Liberating the weights. J. Complexity **20**, 593–623 (2004)
8. Ern, A., Guermond, J.-L.: Theory and Practice of Finite Elements. Applied Mathematical Sciences, vol. 159. Springer, Heidelberg (2004)
9. Graham, I.G., Kuo, F.Y., Nichols, J.A., Scheichl, R., Schwab, Ch., Sloan, I.H.: Quasi-Monte Carlo finite element methods for elliptic PDEs with log-normal random coefficients. Report 2013–14, Seminar for Applied Mathematics, ETH Zurich, http://www.sam.math.ethz.ch/sam_reports/index.php?id=2013-14 (in review)
10. Hansen, M., Schwab, Ch.: Analytic regularity and best N-term approximation of high dimensional parametric initial value problems. Vietnam J. Math. **41/2**, 181–215 (2013)
11. Hansen, M., Schwab, Ch.: Analytic regularity and nonlinear approximation of a class of parametric semilinear elliptic PDEs. Math. Nachr. (2013, to appear)
12. Hoang, V.H., Schwab, Ch.: Sparse tensor Galerkin discretizations for parametric and random parabolic PDEs: analytic regularity and gpc-approximation. Report 2010-11, Seminar for Applied Mathematics, ETH Zürich. SIAM J. Math. Anal. **45/5**, 3050–3083 (2013)
13. Hoang, V.H., Schwab, Ch.: Analytic regularity and gpc approximation for parametric and random 2nd order hyperbolic PDEs. Report 2010-19, Seminar for Applied Mathematics, ETH Zürich. Anal. Appl. (Singapore) **10**(3) (2012)
14. Hoang, V.H., Schwab, Ch.: Analytic regularity and polynomial approximation of stochastic, parametric elliptic multiscale PDEs. Anal. Appl. (Singapore) (2013, to appear)
15. Kunoth, A., Schwab, Ch.: Analytic regularity and GPC approximation for stochastic control problems constrained by Linear parametric elliptic and parabolic PDEs. SIAM J. Control Optim. **51**, 2442–2471 (2013)

16. Kuo, F.Y., Schwab, Ch., Sloan, I.H.: Quasi-Monte Carlo methods for very high dimensional integration: the standard weighted-space setting and beyond. ANZIAM J. **53**, 1–37 (2011)
17. Kuo, F.Y., Schwab, Ch., Sloan, I.H.: Quasi-Monte Carlo finite element methods for a class of elliptic partial differential equations with random coefficient. SIAM J. Numer. Anal. **50**, 3351–3374 (2012)
18. Kuo, F.Y., Schwab, Ch., Sloan, I.H.: Corrigendum for Quasi-Monte Carlo methods for very high dimensional integration: the standard weighted-space setting and beyond. ANZIAM J. (2013, to appear)
19. Kuo, F.Y., Schwab, Ch., Sloan, I.H.: Multi-level Quasi-Monte Carlo finite element methods for a class of elliptic partial differential equations with random coefficient. Report 2012-25, Seminar für Angewandte Mathematik, ETH Zürich (in review)
20. Kuo, F.Y., Sloan, I.H., Wasilkowski, G.W., Waterhouse, B.J.: Randomly shifted lattice rules with the optimal rate of convergence for unbounded integrands. J. Complexity **26**, 135–160 (2010)
21. Kuo, F.Y., Sloan, I.H., Wasilkowski, G.W., Woźniakowski, H.: Liberating the dimension. J. Complexity **26**, 422–454 (2010)
22. Kuo, F.Y., Wasilkowski, G.W., Waterhouse, B.J.: Randomly shifted lattice rules for unbounded integrals. J. Complexity **22**, 630–651 (2006)
23. Larcher, G., Leobacher, G., Scheicher, K.: On the tractability of the Brownian bridge algorithm. J. Complexity **19**, 511–528 (2004)
24. Niederreiter, H.: Random Number Generation and Quasi-Monte Carlo Methods. SIAM, Philadelphia (1992)
25. Nistor, V., Schwab, Ch.: High order Galerkin approximations for parametric second order elliptic partial differential equations. Report 2012-21 Seminar for Applied Mathematics, ETH Zürich. Math. Method Model Appl. Sci. (2013, to appear)
26. Nuyens, D., Cools, R.: Fast algorithms for component-by-component construction of rank-1 lattice rules in shift-invariant reproducing kernel Hilbert spaces. Math. Comp. **75**, 903–920 (2006)
27. Schwab, Ch., Gittelson, C.J.: Sparse tensor discretizations of high-dimensional parametric and stochastic PDEs. Acta Numer. **20**, 291–467 (2011). Cambridge University Press
28. Schwab, Ch., Todor, R.A.: Karhunen-Loève approximation of random fields by generalized fast multipole methods. J. Comput. Phys. **217**, 100–122 (2006)
29. Sloan, I.H., Joe, S.: Lattice Methods for Multiple Integration. Oxford University Press, Oxford (1994)
30. Sloan, I.H., Kuo, F.Y., Joe, S.: Constructing randomly shifted lattice rules in weighted Sobolev spaces. SIAM J. Numer. Anal. **40**, 1650–1665 (2002)
31. Sloan, I.H., Wang, X., Woźniakowski, H.: Finite-order weights imply tractability of multivariate integration. J. Complexity **20**, 46–74 (2004)
32. Sloan, I.H., Woźniakowski, H.: When are Quasi-Monte Carlo algorithms efficient for high-dimensional integrals? J. Complexity **14**, 1–33 (1998)
33. Wang, X.: Strong tractability of multivariate integration using Quasi-Monte Carlo algorithms. Math. Comp. **72**, 823–838 (2002)

On the Choice of Weights in a Function Space for Quasi-Monte Carlo Methods for a Class of Generalised Response Models in Statistics

Vasile Sinescu, Frances Y. Kuo, and Ian H. Sloan

Abstract Evaluation of the likelihood of generalised response models in statistics leads to integrals over unbounded regions in high dimensions. In order to apply a quasi-Monte Carlo (QMC) method to approximate such integrals, one has to transform the original integral into an equivalent integral over the unit cube. From the point of view of QMC, this leads to a known (but non-standard) space of functions for the transformed problem. The "weights" in this function space describe the relative importance of variables or groups of variables. The quadrature error produced via a QMC method is bounded by the product of the worst-case error and the norm of the transformed integrand. This paper is mainly concerned with finding a suitable error bound for the integrand arising from a particular generalised linear model for time series regression, and then determining the choice of weights that minimises this error bound. We obtained "POD weights" ("product and order dependent weights") which are of a simple enough form to permit the construction of randomly shifted lattice rules with the optimal rate of convergence in the given function space setting.

1 Introduction

In this paper we use *quasi-Monte Carlo* (QMC) methods for the approximation of high-dimensional integrals arising from a class of *generalised response models* in statistics. In particular, we study the log-likelihood function considered in [6]

V. Sinescu (✉) · F.Y. Kuo · I.H. Sloan
School of Mathematics and Statistics, University of New South Wales, Sydney, NSW 2052, Australia
e-mail: v.sinescu@unsw.edu.au; f.kuo@unsw.edu.au; i.sloan@unsw.edu.au

J. Dick et al. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2012*,
Springer Proceedings in Mathematics & Statistics 65,
DOI 10.1007/978-3-642-41095-6_33,

$$\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \log \int_{\mathbb{R}^d} \exp\left\{ \boldsymbol{y}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{W}\boldsymbol{w}) - \mathbf{1}^{\mathrm{T}} b(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{W}\boldsymbol{w}) - \frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{w} \right\} \mathrm{d}\boldsymbol{w} - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{d}{2}\log(2\pi) + \mathbf{1}^{\mathrm{T}} c(\boldsymbol{y}), \tag{1}$$

where $\boldsymbol{y}$ is a response vector containing counts, $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are model parameters, $\boldsymbol{X}$ is a design matrix corresponding to the effects in $\boldsymbol{\beta}$, $\boldsymbol{W}$ is a design matrix sometimes assumed to be the identity matrix, and d is the number of counts. The functions b and c are scalar functions which are applied component-wise to a vector. For such models a conditional Poisson likelihood is often assumed, hence we take $b(x) = e^x$ and $c(x) = \log(1/x!)$. The log-likelihood (1) represents the logarithm of an expectation with respect to the random variable $\boldsymbol{w}$ from the multivariate normal density with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. See [2, 3, 6, 10, 13] for further background information on the problem.

QMC methods are equal-weight quadrature rules defined over the unit cube. The usual approach to approximating integrals over unbounded regions via QMC is to transform the original integral into one over the unit cube. In this case, however, the typical approach of diagonalising $\boldsymbol{\Sigma}$ and using its factors to map to the unit cube yields very poor results, see [6]. The main reason is that integrals arising from the log-likelihood have a narrow peak located far away from the origin, attributable to the exponential function $b(x) = e^x$ in the exponent. Therefore, in [6] the authors recentred and rescaled the exponent in the integrand before mapping to the unit cube; the details of the transformation will be given in Sect. 3.

After recentering around the stationary point of the exponent and then rescaling, we divide and multiply the resulting integrand by the product of a suitable univariate probability density ϕ, so obtaining the integral in the form

$$I_d(f, \phi) = \int_{\mathbb{R}^d} f(\boldsymbol{x}) \prod_{j=1}^{d} \phi(x_j)\, \mathrm{d}\boldsymbol{x}. \tag{2}$$

The integral (2) is then transformed into an equivalent integral over the unit cube by using the mapping

$$u = \Phi(x) = \int_{-\infty}^{x} \phi(t)\, \mathrm{d}t, \quad \forall x \in \mathbb{R}, \tag{3}$$

for each coordinate direction. The inverse mapping is $\Phi^{-1} : (0, 1) \to \mathbb{R}$, $\Phi^{-1}(u) = x$. In dimension d, we have $\Phi^{-1}(\boldsymbol{u}) := (\Phi^{-1}(u_1), \ldots, \phi^{-1}(u_d))$, and the transformed integral in the unit cube is

$$I_d(\tilde{f}) = \int_{[0,1]^d} \tilde{f}(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u}, \quad \text{where} \quad \tilde{f} = f \circ \Phi^{-1}. \tag{4}$$

We stress that a different choice of the density ϕ will lead to a different integrand f in (2), and a different transformed integrand $\tilde{f}$ in (4).

This transformed integrand $\tilde{f}$ does *not* lie in a function space where QMC methods are typically analysed (e.g., the standard Sobolev space in [17]). Therefore standard QMC results do not apply. Fortunately, the integrand $\tilde{f}$ lies in a function space studied in [9, 11] and therefore the results in those papers can be used for the problem considered here. Further details will be provided in Sect. 2.

To construct the particular QMC methods known as *randomly shifted lattice rules*, we need to choose some nonnegative parameters known as *weights*, which are used to describe the relative importance of variables or groups of variables. Weighted spaces of functions were first proposed in [17] to explain the success of QMC. By now the literature on finding good QMC methods in weighted spaces of functions, once the weights are known, is very rich, see for instance the recent reviews [4, 7].

In [6] the weights were chosen in an ad hoc fashion. In this paper, in contrast, our goal is to obtain weights suitable for approximating the log-likelihood integral arising from a particular time series model (see Sect. 3), using the following approach. The integration error is bounded by the product of the *worst-case error* (depending only on the QMC point set) and the norm of the function (depending only on the integrand). We choose weights that minimise a certain upper bound on this product. In order to do that we focus first on bounding the norm of f under different choices of ϕ, a difficult task for the specific integrands arising from the log-likelihood (1). The strategy of then finding the weights will be adapted from [8], where the authors were concerned with QMC methods suitable for a class of elliptic partial differential equations. As there, the resulting weights are "POD weights", see further details in Sect. 2.

2 Background on Lattice Rules and the Function Space Setting

The function space we use is the same as the function space in [9, 11], see also [18, 19]. The norm in this ("anchored") space is given by

$$\|f\|_{\boldsymbol{\gamma}} := \left(\sum_{\mathfrak{u} \subseteq \mathscr{D}} \gamma_{\mathfrak{u}}^{-1} \int_{\mathbb{R}^{|\mathfrak{u}|}} \left(\frac{\partial^{|\mathfrak{u}|}}{\partial \boldsymbol{x}_{\mathfrak{u}}} f(\boldsymbol{x}_{\mathfrak{u}}, \mathbf{0}) \right)^2 \prod_{j \in \mathfrak{u}} \psi^2(x_j) \, \mathrm{d}\boldsymbol{x}_{\mathfrak{u}} \right)^{1/2}. \tag{5}$$

In (5), the notation $(\boldsymbol{x}_{\mathfrak{u}}, \mathbf{0})$ means that we anchor to 0 all the components of $\boldsymbol{x}$ that do not belong to $\mathfrak{u}$, and $\mathscr{D} := \{1, 2, \ldots, d\}$. The space is "non-standard" in the sense that the integral in (5) is over the unbounded domain (rather than the unit cube) and is weighted by a positive and continuous function $\psi : \mathbb{R} \to \mathbb{R}^+$. The role of ψ is

to control the growth of the mixed first partial derivatives of the functions as any component x_j goes to $\pm\infty$, and hence to control the size of the function space.

For each subset $\mathfrak{u} \subseteq \mathscr{D}$, the nonnegative parameter $\gamma_\mathfrak{u}$ is the *weight* associated with the variables $\boldsymbol{x}_\mathfrak{u} = \{x_j : j \in \mathfrak{u}\}$. Popular classes of weights include *product weights* (first introduced in [17]) and *order dependent weights* (see e.g., [16]). The recent paper [8] led us to consider a hybrid between these two classes of weights called *POD weights* ("*product and order dependent*" weights): they take the form

$$\gamma_\mathfrak{u} := \Gamma_{|\mathfrak{u}|} \prod_{j \in \mathfrak{u}} \gamma_j,$$

where $\Gamma_{|\mathfrak{u}|}$ depends only on the cardinality of the set $\mathfrak{u}$ (the order dependent part), while γ_j is a weight associated with the individual coordinate x_j. We will see that POD weights are also suitable for our problem here.

Note that via the transformation (3) which links (2) to (4), the norm of f defined as above in the Euclidean space will be equal to the norm of the transformed integrand $\tilde{f} = f \circ \Phi^{-1}$ in the corresponding isometric space over the unit cube. In what follows it will be more convenient, however, to analyse the norm of f in $\mathbb{R}^d$.

The integral (2) may be approximated by *randomly shifted lattice rules*

$$Q_{n,d}(f; \boldsymbol{\Delta}) = \frac{1}{n} \sum_{k=0}^{n-1} f\left(\Phi^{-1}\left(\left\{\frac{k\boldsymbol{z}}{n} + \boldsymbol{\Delta}\right\}\right)\right),$$

where $\boldsymbol{\Delta} \in [0,1]^d$ is a *random shift* drawn from the uniform distribution over the unit cube and $\boldsymbol{z} \in \mathbb{Z}^d$ is a (deterministic) *generating vector*. (More accurately, the randomly shifted lattice rule is applied to the transformed integral (4).) The braces around a vector indicate that we take the fractional part of each vector component. The generating vector can be restricted to the set $\mathscr{Z}_n^d$, where $\mathscr{Z}_n := \{z \in \{1, 2, \ldots, n-1\} : \gcd(z, n) = 1\}$. We have $|\mathscr{Z}_n| = \varphi(n)$, where $\varphi(n)$ is Euler's totient function, that is, the number of positive integers smaller than n and co-prime with n.

We define the *worst-case error* of a shifted lattice rule with generating vector $\boldsymbol{z}$ and shift $\boldsymbol{\Delta}$ by

$$e_{n,d,\boldsymbol{\gamma}}(\boldsymbol{z}, \boldsymbol{\Delta}) := \sup_{\|f\|_{\boldsymbol{\gamma}} \le 1} |I_d(f, \phi) - Q_{n,d}(f; \boldsymbol{\Delta})|.$$

It follows that for all f with finite norm (5) we have

$$|I_d(f, \phi) - Q_{n,d}(f; \boldsymbol{\Delta})| \le e_{n,d,\boldsymbol{\gamma}}(\boldsymbol{z}, \boldsymbol{\Delta}) \cdot \|f\|_{\boldsymbol{\gamma}}.$$

Squaring both sides of this inequality, integrating over $\boldsymbol{\Delta} \in [0,1]^d$, and then taking the square root on both sides, we obtain

$$\sqrt{\mathbb{E}|I_d(f,\phi) - Q_{n,d}(f;\cdot)|^2} \le \hat{e}_{n,d,\boldsymbol{\gamma}}(\boldsymbol{z}) \cdot \|f\|_{\boldsymbol{\gamma}}, \tag{6}$$

where the expectation is with respect to the random shift, and $\hat{e}_{n,d,\boldsymbol{\gamma}}(\boldsymbol{z})$ is the "*shift-averaged*" *worst-case error* for randomly shifted lattice rules, defined by

$$\hat{e}_{n,d,\boldsymbol{\gamma}}(\boldsymbol{z}) := \left(\int_{[0,1]^d} e^2_{n,d,\boldsymbol{\gamma}}(\boldsymbol{z},\boldsymbol{\Delta})\,\mathrm{d}\boldsymbol{\Delta}\right)^{1/2}.$$

A lattice rule can be constructed via the component-by-component (CBC) algorithm, which is a greedy algorithm that constructs the generating vector one component at a time. The CBC construction has been extensively used in many research papers, see among others [5,12,14,15]. Specific (fast) CBC constructions of randomly shifted lattice rules for POD weights were considered in [7,11]. Note that the CBC construction and the error bound for this specific function space, see [11], depend explicitly on the choices of ϕ and ψ as well as on the weights $\gamma_{\mathfrak{u}}$. Further analysis on the worst-case error will be given in Sect. 6.

3 The Integrand for the Likelihood Problem

In this paper we focus on the particular time series model in [6, Example 1]. The integral of interest is (leaving out a scaling constant)

$$\int_{\mathbb{R}^d} \exp(F(\boldsymbol{w}))\,\mathrm{d}\boldsymbol{w}, \tag{7}$$

where the function F is defined by

$$F(\boldsymbol{w}) := \sum_{j=1}^{d} \left(y_j(\beta + w_j) - e^{\beta+w_j}\right) - \frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{w}.$$

Here $\beta \in \mathbb{R}$ is a parameter, $y_1, \ldots, y_d \in \{0, 1, \ldots\}$ are the count data, and $\boldsymbol{\Sigma}$ is a (Toeplitz) covariance matrix with the (j,i)-th entry given by $\sigma^2\kappa^{|i-j|}/(1-\kappa^2)$, where σ^2 is the variance and $\kappa \in (-1,1)$ is the autoregression coefficient.

As mentioned earlier, we follow the idea from [6] of first recentering and rescaling the exponent of the integrand, as follows:

1. Find the unique stationary point $\boldsymbol{w}^*$ satisfying $\nabla F(\boldsymbol{w}^*) = 0$.
2. Determine the matrix $\boldsymbol{\Sigma}^* = (-\nabla^2 F(\boldsymbol{w}^*))^{-1}$, which describes the convexity of F around the stationary point and factorise it as $\boldsymbol{\Sigma}^* = A^* A^{*\mathrm{T}}$.
3. Introduce the transformation $\boldsymbol{w} = A^*\boldsymbol{x} + \boldsymbol{w}^*$. Then $\mathrm{d}\boldsymbol{w} = |\det A^*|\,\mathrm{d}\boldsymbol{x}$, and the integral (7) becomes $|\det A^*| \int_{\mathbb{R}^d} \exp(F(A^*\boldsymbol{x} + \boldsymbol{w}^*))\,\mathrm{d}\boldsymbol{x}$.

We remark that Steps 1–2 use the idea from the "Laplace method" (see e.g., [1]) which involves a multivariate normal approximation to the integrand via a second order Taylor series approximation of the exponent about its stationary point. The matrix $\boldsymbol{\Sigma}^*$ will most likely be positive definite due to our underlying model assumptions.

We then multiply and divide the new integrand $\exp(F(A^*\boldsymbol{x} + \boldsymbol{w}^*))$ by $\prod_{j=1}^d \phi(x_j)$ for a suitable univariate probability density function $\phi : \mathbb{R} \to \mathbb{R}$. Thus we have transformed the integral (7) into an integral of the form (2), where f is given by (omitting here the factor $|\det A^*|$)

$$f(\boldsymbol{x}) := \exp(g(\boldsymbol{x})) \prod_{j=1}^{d} h(x_j), \quad \text{with} \quad h(x) = 1/\phi(x), \tag{8}$$

and

$$g(\boldsymbol{x}) := \sum_{j=1}^{d} \left(y_j((A^*\boldsymbol{x})_j + \beta + w_j^*) - e^{(A^*\boldsymbol{x})_j + \beta + w_j^*} \right) - \frac{1}{2}(A^*\boldsymbol{x} + \boldsymbol{w}^*)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (A^*\boldsymbol{x} + \boldsymbol{w}^*).$$

By construction, $\nabla g(\boldsymbol{x})$ vanishes at $\boldsymbol{x} = \boldsymbol{0}$, while near the origin $g(\boldsymbol{x})$ behaves like $g(\boldsymbol{0}) - \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}$.

Let us remark that f is determined after we have chosen ϕ. Later we will consider logistic, normal and Student densities for ϕ; these have been analysed in [9]. We also need to choose a weight function ψ to ensure that the norm (5) of f is finite. We will show that for our three choices of ϕ it suffices to take $\psi \equiv 1$. (We stress that even with $\psi \equiv 1$ the corresponding isometric space over the unit cube is *not* the standard Sobolev space considered in e.g., [17].)

4 Bounding the Mixed First Partial Derivatives of the Integrand

Due to the error bound (6), we are interested in finding an upper bound on the norm (5) of f in (8). In turn, we need to estimate the mixed first partial derivatives of f.

Lemma 1. *Let $(a_{ji}^*)_{1\le i,j\le d}$ denote the (j,i)-entry in the matrix A^*, and let $(A^*)_i$ denote the i-th column of A^*. Let $\Pi(\mathfrak{v})$ denote the set of all partitions of a finite set $\mathfrak{v}$. For any probability density ϕ and any $\mathfrak{u} \subseteq \mathscr{D}$, the mixed first partial derivatives of f defined by (8) are given by*

$$\frac{\partial^{|\mathfrak{u}|} f(\boldsymbol{x})}{\partial \boldsymbol{x}_{\mathfrak{u}}} = \sum_{\mathfrak{v}\subseteq\mathfrak{u}} \left[\frac{\partial^{|\mathfrak{v}|} \exp(g(\boldsymbol{x}))}{\partial \boldsymbol{x}_{\mathfrak{v}}} \left(\prod_{\substack{i=1 \\ i\notin \mathfrak{u}\setminus\mathfrak{v}}}^{d} h(x_i) \right) \left(\prod_{\substack{i=1 \\ i\in \mathfrak{u}\setminus\mathfrak{v}}}^{d} h'(x_i) \right) \right], \tag{9}$$

where

$$\frac{\partial^{|\mathfrak{v}|}\exp(g(\boldsymbol{x}))}{\partial \boldsymbol{x}_{\mathfrak{v}}} = \exp(g(\boldsymbol{x})) \sum_{\pi\in\Pi(\mathfrak{v})} \prod_{\mathfrak{t}\in\pi} \left(-\sum_{j=1}^{d} \left(\prod_{i\in\mathfrak{t}} a_{ji}^* \right) e^{(A^*\boldsymbol{x})_j+\beta+w_j^*} + T_{\mathfrak{t}}(\boldsymbol{x}) \right), \tag{10}$$

and for any subset $\mathfrak{t} \subseteq \mathfrak{v}$

$$T_{\mathfrak{t}}(\boldsymbol{x}) := \begin{cases} \sum_{j=1}^{d} y_j a_{ji}^* - (A^*)_i^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(A^*\boldsymbol{x}+\boldsymbol{w}^*) & \text{if} \quad |\mathfrak{t}|=1, \quad \mathfrak{t}=\{i\}, \\ -(A^*)_i^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(A^*)_k & \text{if} \quad |\mathfrak{t}|=2, \quad \mathfrak{t}=\{i,k\}, \\ 0 & \text{if} \quad |\mathfrak{t}|\geq 3. \end{cases} \tag{11}$$

Proof. From (8), we see that we can write

$$\frac{\partial^{|\mathfrak{u}|} f(\boldsymbol{x})}{\partial \boldsymbol{x}_{\mathfrak{u}}} = \sum_{\mathfrak{v}\subseteq\mathfrak{u}} \left[\frac{\partial^{|\mathfrak{v}|}\exp(g(\boldsymbol{x}))}{\partial \boldsymbol{x}_{\mathfrak{v}}} \cdot \frac{\partial^{|\mathfrak{u}|-|\mathfrak{v}|}}{\partial \boldsymbol{x}_{\mathfrak{u}\setminus\mathfrak{v}}} \left(\prod_{i=1}^{d} h(x_i) \right) \right].$$

For the derivative of the product function on the right-hand side, we obtain

$$\frac{\partial^{|\mathfrak{u}|-|\mathfrak{v}|}}{\partial \boldsymbol{x}_{\mathfrak{u}\setminus\mathfrak{v}}} \left(\prod_{i=1}^{d} h(x_i) \right) = \left(\prod_{\substack{i=1 \\ i\notin\mathfrak{u}\setminus\mathfrak{v}}}^{d} h(x_i) \right) \left(\prod_{\substack{i=1 \\ i\in\mathfrak{u}\setminus\mathfrak{v}}}^{d} h'(x_i) \right).$$

For the partial derivatives of $\exp(g(\boldsymbol{x}))$, we use the well-known Faà di Bruno's formula for differentiating compositions of multivariate functions, to obtain

$$\frac{\partial^{|\mathfrak{v}|}\exp(g(\boldsymbol{x}))}{\partial \boldsymbol{x}_{\mathfrak{v}}} = \exp(g(\boldsymbol{x})) \sum_{\pi\in\Pi(\mathfrak{v})} \prod_{\mathfrak{t}\in\pi} \frac{\partial^{|\mathfrak{t}|} g(\boldsymbol{x})}{\partial \boldsymbol{x}_{\mathfrak{t}}}.$$

We turn our attention to finding the partial derivatives of g. We have

$$\frac{\partial g(\boldsymbol{x})}{\partial x_i} = \sum_{j=1}^{d} \left(y_j a_{ji}^* - a_{ji}^* e^{(A^*\boldsymbol{x})_j+\beta+w_j^*} \right) - (A^*)_i^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(A^*\boldsymbol{x}+\boldsymbol{w}^*),$$

$$\frac{\partial^2 g(\boldsymbol{x})}{\partial x_i \partial x_k} = -\sum_{j=1}^{d} a_{ji}^* a_{jk}^* e^{(A^*\boldsymbol{x})_j+\beta+w_j^*} - (A^*)_i^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(A^*)_k \qquad \text{for } i \neq k,$$

$$\frac{\partial^{|\mathfrak{t}|} g(\boldsymbol{x})}{\prod_{i\in\mathfrak{t}} \partial x_i} = -\sum_{j=1}^{d} \left(\prod_{i\in\mathfrak{t}} a_{ji}^* \right) e^{(A^*\boldsymbol{x})_j+\beta+w_j^*} \qquad \text{for any } \mathfrak{t}\subseteq\mathfrak{v} \text{ with } |\mathfrak{t}|\geq 3.$$

Combining all formulas in this proof leads to the stated result. □

To proceed further from (9), we need a bound on the mixed first partial derivatives of $\exp(g(\boldsymbol{x}))$ given by (10), and a bound on those products involving $h = 1/\phi$ and its derivative. The former is done in Lemma 2 below; the latter is treated in Assumption 1 and Lemma 3 in the next section.

We need the following definitions:

$$\alpha_i := \max_{j=1,\dots,d} |a^*_{ji}|, \quad \omega := 1 + \sum_{i=1}^{d}\sum_{j=1}^{d} |\varpi_{ij}| \text{ with } \Sigma^{-1} = (\varpi_{ij})_{i,j=1}^{d}, \tag{12}$$

$$\tau_1 := \sum_{i=1}^{d} \frac{|\sum_{j=1}^{d} y_j a^*_{ji} - (A^*)_i^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{w}^*|}{\alpha_i}, \quad \theta_k := \sum_{i=1}^{d} \frac{\left|((A^*)_i^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} A^*)_k\right|}{\alpha_i}, \tag{13}$$

$$\tau_2 := \sum_{j=1}^{d} y_j(\beta + w^*_j) - \frac{1}{2} \boldsymbol{w}^{*\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{w}^*, \quad \mu_k := \sum_{j=1}^{d} y_j a^*_{jk} - (\boldsymbol{w}^{*\mathrm{T}} \boldsymbol{\Sigma}^{-1} A^*)_k. \tag{14}$$

Furthermore, it is easy to see that the matrix $A^{*\mathrm{T}} \boldsymbol{\Sigma}^{-1} A^*$ is real, symmetric and positive definite, and consequently all the eigenvalues are positive. We define

$$\lambda_{\min} := \text{the smallest eigenvalue of the matrix } A^{*\mathrm{T}} \boldsymbol{\Sigma}^{-1} A^*. \tag{15}$$

Also we denote by B_m the *Bell number* of order m, that is, the number of partitions of a set with m elements.

Lemma 2. *For any* $\mathfrak{v} \subseteq \mathscr{D}$, *the mixed first partial derivative of* $\exp(g(\boldsymbol{x}))$ *from* (10) *is bounded by*

$$\left|\frac{\partial^{|\mathfrak{v}|} \exp(g(\boldsymbol{x}))}{\partial \boldsymbol{x}_{\mathfrak{v}}}\right| \le e^{\tau_1+\tau_2} B_{|\mathfrak{v}|} |\mathfrak{v}|! \omega^{|\mathfrak{v}|/2} \left(\prod_{i\in\mathfrak{v}} \alpha_i\right) \prod_{k=1}^{d} \exp\left(-\frac{\lambda_{\min}}{2} x_k^2 + \mu_k x_k + \theta_k |x_k|\right).$$

Proof. For convenience, we define for the scope of this proof

$$K(\boldsymbol{x}) := \sum_{j=1}^{d} e^{(A^*\boldsymbol{x})_j + \beta + w^*_j} \quad \text{and} \quad J(\boldsymbol{x}) := g(\boldsymbol{x}) + K(\boldsymbol{x}).$$

These, together with the definition of α_i in (12), yield the following bound on (10)

$$\begin{aligned}\left|\frac{\partial^{|\mathfrak{v}|} \exp(g(\boldsymbol{x}))}{\partial \boldsymbol{x}_{\mathfrak{v}}}\right| &\le \exp(J(\boldsymbol{x}) - K(\boldsymbol{x})) \sum_{\pi\in\Pi(\mathfrak{v})} \prod_{\mathfrak{t}\in\pi} \left(K(\boldsymbol{x}) \prod_{i\in\mathfrak{t}} \alpha_i + |T_{\mathfrak{t}}(\boldsymbol{x})|\right) \\ &= \left(\prod_{i\in\mathfrak{v}} \alpha_i\right) \exp(J(\boldsymbol{x}) - K(\boldsymbol{x})) \sum_{\pi\in\Pi(\mathfrak{v})} \prod_{\mathfrak{t}\in\pi} \left(K(\boldsymbol{x}) + \frac{|T_{\mathfrak{t}}(\boldsymbol{x})|}{\prod_{i\in\mathfrak{t}} \alpha_i}\right), \end{aligned} \tag{16}$$

where we used the property that for any partition π of $\mathfrak{v}$ we have

$$\prod_{\mathfrak{t}\in\pi}\prod_{i\in\mathfrak{t}}\alpha_i = \prod_{i\in\mathfrak{v}}\alpha_i.$$

For (16) we now consider $K(\boldsymbol{x}) \le 1$ and $K(\boldsymbol{x}) > 1$ separately. Note that $K(\boldsymbol{x}) \ge 0$ for any $\boldsymbol{x}$. If $K(\boldsymbol{x}) \le 1$ then

$$\left|\frac{\partial^{|\mathfrak{v}|}\exp(g(\boldsymbol{x}))}{\partial \boldsymbol{x}_{\mathfrak{v}}}\right| \le \left(\prod_{i\in\mathfrak{v}}\alpha_i\right)\exp(J(\boldsymbol{x}))\sum_{\pi\in\Pi(\mathfrak{v})}\prod_{\mathfrak{t}\in\pi}\left(1+\frac{|T_{\mathfrak{t}}(\boldsymbol{x})|}{\prod_{i\in\mathfrak{t}}\alpha_i}\right). \tag{17}$$

If $K(\boldsymbol{x}) > 1$ then

$$\begin{aligned}\left|\frac{\partial^{|\mathfrak{v}|}\exp(g(\boldsymbol{x}))}{\partial \boldsymbol{x}_{\mathfrak{v}}}\right| &\le \left(\prod_{i\in\mathfrak{v}}\alpha_i\right)\exp(J(\boldsymbol{x})-K(\boldsymbol{x}))\sum_{\pi\in\Pi(\mathfrak{v})}\prod_{\mathfrak{t}\in\pi}\left(K(\boldsymbol{x})\left(1+\frac{|T_{\mathfrak{t}}(\boldsymbol{x})|}{\prod_{i\in\mathfrak{t}}\alpha_i}\right)\right)\\ &\le \left(\prod_{i\in\mathfrak{v}}\alpha_i\right)\exp(J(\boldsymbol{x}))\exp(-K(\boldsymbol{x}))(K(\boldsymbol{x}))^{|\mathfrak{v}|}\sum_{\pi\in\Pi(\mathfrak{v})}\prod_{\mathfrak{t}\in\pi}\left(1+\frac{|T_{\mathfrak{t}}(\boldsymbol{x})|}{\prod_{i\in\mathfrak{t}}\alpha_i}\right). \end{aligned} \tag{18}$$

Using the elementary inequality $e^{-x}x^{\eta} \le (\frac{\eta}{e})^{\eta}$ for all $x \ge 0$ and $\eta \ge 1$ together with Stirling's formula, we obtain

$$\exp(-K(\boldsymbol{x}))(K(\boldsymbol{x}))^{|\mathfrak{v}|} \le \left(\frac{|\mathfrak{v}|}{e}\right)^{|\mathfrak{v}|} \le |\mathfrak{v}|!.$$

Since $|\mathfrak{v}|! \ge 1$ for any subset $\mathfrak{v}$ with $|\mathfrak{v}| \ge 1$, it follows from (17) and (18) that

$$\left|\frac{\partial^{|\mathfrak{v}|}\exp(g(\boldsymbol{x}))}{\partial \boldsymbol{x}_{\mathfrak{v}}}\right| \le |\mathfrak{v}|!\left(\prod_{i\in\mathfrak{v}}\alpha_i\right)\exp(J(\boldsymbol{x}))\sum_{\pi\in\Pi(\mathfrak{v})}\prod_{\mathfrak{t}\in\pi}\left(1+\frac{|T_{\mathfrak{t}}(\boldsymbol{x})|}{\prod_{i\in\mathfrak{t}}\alpha_i}\right). \tag{19}$$

Next we use the definition of $T_{\mathfrak{t}}(\boldsymbol{x})$ in (11) which depends on the cardinality of the block $\mathfrak{t}$ in the partition π. For the product over partition blocks of cardinality 1, we have

$$\prod_{\substack{\mathfrak{t}\in\pi\\|\mathfrak{t}|=1}}\left(1+\frac{|T_{\mathfrak{t}}(\boldsymbol{x})|}{\prod_{i\in\mathfrak{t}}\alpha_i}\right) \le \exp\left(\sum_{\substack{\mathfrak{t}\in\pi\\|\mathfrak{t}|=1}}\frac{|T_{\mathfrak{t}}(\boldsymbol{x})|}{\prod_{i\in\mathfrak{t}}\alpha_i}\right) \le \exp\left(\sum_{i=1}^{d}\frac{|T_{\{i\}}(\boldsymbol{x})|}{\alpha_i}\right), \tag{20}$$

where the first inequality was obtained by using $1 + x \leq e^x$ for $x \in \mathbb{R}$. The second inequality above is conservative, but allows us to obtain a bound independent of the partition. It follows from the first case in (11) and the definitions of τ_1 and θ_k in (13) that

$$\sum_{i=1}^{d} \frac{|T_{\mathfrak{t}i\}}(\boldsymbol{x})|}{\alpha_i} \leq \sum_{i=1}^{d} \frac{\left|\sum_{j=1}^{d} y_j a_{ji}^* - (A^*)_i^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{w}^*\right| + \left|(A^*)_i^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} A^* \boldsymbol{x}\right|}{\alpha_i} \leq \tau_1 + \sum_{k=1}^{d} \theta_k |x_k|. \tag{21}$$

For the product over partition blocks of cardinality 2, we use the definitions of α_i and ω in (12) to obtain for any pair of distinct indices $\{i, k\} \subseteq \mathscr{D}$

$$1 + \frac{|T_{\{i,k\}}(\boldsymbol{x})|}{\alpha_i \alpha_k} = 1 + \frac{|(A^*)_i^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (A^*)_k|}{\alpha_i \alpha_k} \leq \omega.$$

Then since for any partition π of $\mathfrak{v}$ we have at most $\lfloor |\mathfrak{v}|/2 \rfloor$ blocks of cardinality 2, we conclude that

$$\prod_{\substack{\mathfrak{t} \in \pi \\ |\mathfrak{t}|=2}} \left(1 + \frac{|T_{\mathfrak{t}}(\boldsymbol{x})|}{\prod_{i \in \mathfrak{t}} \alpha_i}\right) \leq \omega^{|\mathfrak{v}|/2}. \tag{22}$$

Substituting (20), (21), and (22) into (19), and noting that the product over partition blocks of cardinality greater than 2 always equals 1, we obtain

$$\left|\frac{\partial^{|\mathfrak{v}|} \exp(g(\boldsymbol{x}))}{\partial \boldsymbol{x}_{\mathfrak{v}}}\right| \leq B_{|\mathfrak{v}|} |\mathfrak{v}|! \, \omega^{|\mathfrak{v}|/2} \left(\prod_{i \in \mathfrak{v}} \alpha_i\right) \exp\left(J(\boldsymbol{x}) + \tau_1 + \sum_{k=1}^{d} \theta_k |x_k|\right), \tag{23}$$

where $B_{|\mathfrak{v}|}$ is the number of partitions of the set $\mathfrak{v}$.

Finally we use the definitions of τ_2 and μ_k in (14) and $\lambda_{\min}$ in (15) to write

$$\begin{aligned} J(\boldsymbol{x}) = g(\boldsymbol{x}) + K(\boldsymbol{x}) &= \tau_2 - \frac{1}{2} \boldsymbol{x}^{\mathrm{T}} A^{*\mathrm{T}} \boldsymbol{\Sigma}^{-1} A^* \boldsymbol{x} + \sum_{j=1}^{d} y_j (A^* \boldsymbol{x})_j - \boldsymbol{w}^{*\mathrm{T}} \boldsymbol{\Sigma}^{-1} A^* \boldsymbol{x} \\ &\leq \tau_2 + \sum_{k=1}^{d} \left(-\frac{\lambda_{\min}}{2} x_k^2 + \mu_k x_k\right), \end{aligned} \tag{24}$$

where we used the property that $\boldsymbol{x}^{\mathrm{T}} A^{*\mathrm{T}} \boldsymbol{\Sigma}^{-1} A^* \boldsymbol{x} \geq \lambda_{\min} \sum_{k=1}^{d} x_k^2$ for any $\boldsymbol{x} \in \mathbb{R}^d$. Substituting (24) into (23) completes the proof. □

Note that in Lemma 2 we intentionally obtained a bound that depends on the variables x_k in a product manner, so that the integral in the norm (5) can be easily calculated.

5 Bounding the Norm of the Integrand for Specific Densities

Specific bounds on the norm (5) of the integrand f given in (8) will depend on the chosen density ϕ. We first state an assumed bound for the products in (9) involving $h = 1/\phi$ and its derivative, and verify that this generic bound holds for three specific choices of ϕ. We then proceed to obtain a generic bound on the norm of f.

Assumption 1. *Suppose ϕ is a probability density such that for all $\mathfrak{v} \subseteq \mathfrak{u} \subseteq \mathscr{D}$ we have, with $h = 1/\phi$,*

$$\left(\prod_{\substack{i=1\\ i\notin \mathfrak{u}\setminus\mathfrak{v}}}^{d} h(x_i)\right)\left(\prod_{\substack{i=1\\ i\in \mathfrak{u}\setminus\mathfrak{v}}}^{d} |h'(x_i)|\right) \le p^d q^{|\mathfrak{u}|-|\mathfrak{v}|} \prod_{k=1}^{d} \exp\left(\lambda_0 x_k^2 + \theta_0 |x_k|\right), \tag{25}$$

where p, q, λ_0 and θ_0 are constants independent of $\mathfrak{v}$, $\mathfrak{u}$ and d.

Lemma 3. *Assumption 1 holds for the logistic, normal, and Student densities with parameter $\nu > 0$ as follows:*

	$\phi(x)$	p	q	λ_0	θ_0
Logistic	$\dfrac{e^{x/\nu}}{\nu(1+e^{x/\nu})^2}$	4ν	$\dfrac{1}{4\nu}$	0	$\dfrac{1}{\nu}$
Normal	$\dfrac{1}{\sqrt{2\pi\nu}}\exp\left(-\dfrac{x^2}{2\nu}\right)$	$\sqrt{2\pi\nu}$	$\dfrac{1}{\nu}$	$\dfrac{1}{2\nu}$	1
Student	$T_\nu\left(1+\dfrac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$	$\dfrac{1}{T_\nu}$	$1+\dfrac{1}{\nu}$	0	$\dfrac{\nu+\max(1,\sqrt{2\nu}-1)}{\sqrt{2\nu}}$

Here $T_\nu = (\pi\nu)^{-1/2}\,\Gamma(\frac{\nu+1}{2})/\Gamma(\frac{\nu}{2})$, where $\Gamma(\cdot)$ denotes Euler's gamma function.

Proof. Recall that $h(x) = 1/\phi(x)$. For the logistic density we use the estimates

$$h(x) = \nu(e^{x/\nu} + 2 + e^{-x/\nu}) \le 4\nu e^{|x|/\nu} \quad \text{and} \quad |h'(x)| = |e^{x/\nu} - e^{-x/\nu}| \le e^{|x|/\nu}.$$

For the normal density we use $|x| \le e^{|x|}$ for all $x \in \mathbb{R}$ to obtain

$$h(x) = \sqrt{2\pi\nu}\exp\left(\frac{x^2}{2\nu}\right),$$

$$|h'(x)| = \left|\sqrt{\frac{2\pi}{\nu}}x\,\exp\left(\frac{x^2}{2\nu}\right)\right| \le \frac{\sqrt{2\pi\nu}}{\nu}\exp\left(\frac{x^2}{2\nu} + |x|\right).$$

For the Student density we use $1 + x^2/\nu \le \exp(|x|\sqrt{2/\nu})$ and $|x| \le e^{|x|}$ for all $x \in \mathbb{R}$ to obtain

$$h(x) = \frac{1}{T_\nu}\left(1+\frac{x^2}{\nu}\right)^{\frac{\nu+1}{2}} \le \frac{1}{T_\nu}\exp\left(\frac{\nu+1}{\sqrt{2\nu}}|x|\right),$$

$$|h'(x)| = \left|\frac{1}{T_\nu}\frac{\nu+1}{\nu}x\left(1+\frac{x^2}{\nu}\right)^{\frac{\nu-1}{2}}\right| \le \frac{1}{T_\nu}\left(1+\frac{1}{\nu}\right)\exp\left(\left(\frac{\nu-1}{\sqrt{2\nu}}+1\right)|x|\right).$$

These estimates lead to the parameters in the lemma. □

Theorem 1. *If the weight function is $\psi \equiv 1$ and the probability density ϕ satisfies Assumption 1 with*

$$\lambda_0 < \frac{\lambda_{\min}}{2},$$

then the norm (5) of f defined by (8) is bounded by

$$\|f\|_{\boldsymbol{\gamma}}^2 \le e^{2(\tau_1+\tau_2)}p^{2d}\sum_{\mathfrak{u}\subseteq\mathscr{D}}\left[\gamma_{\mathfrak{u}}^{-1}B_{|\mathfrak{u}|}^2(|\mathfrak{u}|!)^2\left(\frac{\pi}{\lambda_{\min}-2\lambda_0}\right)^{|\mathfrak{u}|/2}\prod_{i\in\mathfrak{u}}(q+\omega^{1/2}\alpha_i)^2\right.$$
$$\left.\cdot\prod_{k\in\mathfrak{u}}\left(\exp\left(\frac{(\theta_k+\theta_0-\mu_k)^2}{\lambda_{\min}-2\lambda_0}\right)+\exp\left(\frac{(\theta_k+\theta_0+\mu_k)^2}{\lambda_{\min}-2\lambda_0}\right)\right)\right].$$

Proof. Combining Lemmas 1 and 2 and Assumption 1, we obtain

$$\left|\frac{\partial^{|\mathfrak{u}|}f(\boldsymbol{x})}{\partial\boldsymbol{x}_{\mathfrak{u}}}\right| \le e^{\tau_1+\tau_2}p^d\sum_{\mathfrak{v}\subseteq\mathfrak{u}}\left(B_{|\mathfrak{v}|}|\mathfrak{v}|!\omega^{|\mathfrak{v}|/2}q^{|\mathfrak{u}|-|\mathfrak{v}|}\prod_{i\in\mathfrak{v}}\alpha_i\right)$$
$$\cdot\prod_{k=1}^{d}\exp\left(-\left(\frac{\lambda_{\min}}{2}-\lambda_0\right)x_k^2+\mu_k x_k+(\theta_k+\theta_0)|x_k|\right),$$

where

$$\sum_{\mathfrak{v}\subseteq\mathfrak{u}}\left(B_{|\mathfrak{v}|}|\mathfrak{v}|!\omega^{|\mathfrak{v}|/2}q^{|\mathfrak{u}|-|\mathfrak{v}|}\prod_{i\in\mathfrak{v}}\alpha_i\right) \le B_{|\mathfrak{u}|}|\mathfrak{u}|!\prod_{i\in\mathfrak{u}}(q+\omega^{1/2}\alpha_i).$$

Thus

$$\|f\|_{\boldsymbol{\gamma}}^2 \le e^{2(\tau_1+\tau_2)}p^{2d}\sum_{\mathfrak{u}\subseteq\mathscr{D}}\left[\gamma_{\mathfrak{u}}^{-1}B_{|\mathfrak{u}|}^2(|\mathfrak{u}|!)^2\prod_{i\in\mathfrak{u}}(q+\omega^{1/2}\alpha_i)^2\right.$$
$$\left.\cdot\prod_{k\in\mathfrak{u}}\int_{-\infty}^{\infty}\psi^2(x_k)\exp\left(-(\lambda_{\min}-2\lambda_0)x_k^2+2\mu_k x_k+2(\theta_k+\theta_0)|x_k|\right)\mathrm{d}x_k\right]. \tag{26}$$

Now with $\psi \equiv 1$ and $\lambda_0 < \lambda_{\min}/2$, the desired result follows from the estimate

$$\begin{aligned}
&\int_{-\infty}^{\infty} \exp(-\lambda x^2 + 2\mu x + 2\theta|x|)\,\mathrm{d}x \\
&\le \int_{-\infty}^{\infty} \exp(-\lambda x^2 + 2(\mu+\theta)x)\,\mathrm{d}x + \int_{-\infty}^{\infty} \exp(-\lambda x^2 + 2(\mu-\theta)x)\,\mathrm{d}x \\
&= \sqrt{\frac{\pi}{\lambda}} \left(\exp\left(\frac{(\theta-\mu)^2}{\lambda} \right) + \exp\left(\frac{(\theta+\mu)^2}{\lambda} \right) \right) \quad \text{for all } \lambda > 0.
\end{aligned}$$

This completes the proof. □

We conclude from Lemma 3 that Theorem 1 holds for the logistic density and the Student density with $\nu > 0$, and for the normal density with

$$\nu > \frac{1}{\lambda_{\min}}.$$

Alternatively, it is also possible to use a normal density ϕ with $\nu \le 1/\lambda_{\min}$, but we would need to consider a weight function $\psi \not\equiv 1$ that decays fast enough to make the integral in (26) finite.

6 Choosing the Weights

From Theorem 1 and Lemma 3, we see that for the three choices of the density ϕ we can bound the norm of f by an expression of the form

$$\|f\|_{\boldsymbol{\gamma}} \le \left(c \sum_{\mathfrak{u} \subseteq \mathscr{D}} \gamma_{\mathfrak{u}}^{-1} \Lambda_{|\mathfrak{u}|} \prod_{k \in \mathfrak{u}} b_k \right)^{1/2}. \tag{27}$$

It now remains to obtain a bound on the integration error, and then determine the most appropriate weights $\gamma_{\mathfrak{u}}$ for our problem.

Theorem 2. *Let $\psi \equiv 1$. Let ϕ be one of the three probability densities in Lemma 3 with parameter $\nu > 0$, and let Φ denote the corresponding cumulative distribution function. Assume additionally that $\nu > 1/\lambda_{\min}$ when ϕ is the normal density. Define*

$$\begin{cases}
C := \dfrac{2^{2-\delta}\nu}{\pi^{2-\delta}\delta e^{1-\delta/2}}, \quad r := 1-\delta/2, \quad \delta \in (0,1), & \text{if } \phi \text{ is logistic,} \\
C := \dfrac{2^{3/2-\delta}\nu^{1/2}}{\pi^{3/2-\delta}\delta e^{1-\delta/2}}, \quad r := 1-\delta/2, \quad \delta \in (0,1), & \text{if } \phi \text{ is normal,} \\
C := \dfrac{4T_\nu^{1/\nu}\nu^{1-1/\nu}(\nu+1)^{(\nu+1)^2/(2\nu)}}{\pi^{2-1/\nu}(\min(1,\nu))^{(\nu+1)/2}(2\nu-1)}, \quad r := 1-1/(2\nu), & \text{if } \phi \text{ is Student,}
\end{cases}$$

and

$$\varrho(\lambda) := \left(2\int_{-\infty}^{0} \Phi^2(x)\,\mathrm{d}x\right)^{\lambda} + 2C^{\lambda}\zeta(2r\lambda), \quad \lambda \in (1/(2r), 1], \tag{28}$$

where $\zeta(x) = \sum_{k=1}^{\infty} k^{-x}$ *is the Riemann zeta function. For* f *defined by* (8) *with norm bounded as in* (27)*, the following statements hold:*

- *A randomly shifted lattice rule can be constructed via the CBC algorithm such that the root-mean-square error satisfies for all* $\lambda \in (1/(2r), 1]$

$$\sqrt{\mathbb{E}|I_d(f) - Q_{n,d}(f;\cdot)|^2} \le \left(\frac{1}{\varphi(n)} \sum_{\mathfrak{u}\subseteq\mathscr{D}} \gamma_{\mathfrak{u}}^{\lambda} (\varrho(\lambda))^{|\mathfrak{u}|}\right)^{1/(2\lambda)} \left(c \sum_{\mathfrak{u}\subseteq\mathscr{D}} \gamma_{\mathfrak{u}}^{-1} \Lambda_{|\mathfrak{u}|} \prod_{k\in\mathfrak{u}} b_k\right)^{1/2}, \tag{29}$$

  where the expectation is with respect to the random shift which is uniformly distributed over $[0,1]^d$*, and* $\varphi(n)$ *is Euler's totient function.*
- *For a given* $\lambda \in (1/(2r), 1]$*, the weights* $\gamma_{\mathfrak{u}}$ *that minimise the error bound* (29) *are given by*

$$\gamma_{\mathfrak{u}} = \gamma_{\mathfrak{u}}(\lambda) := \left(\Lambda_{|\mathfrak{u}|} \prod_{k\in\mathfrak{u}} \frac{b_k}{\sqrt{\varrho(\lambda)}}\right)^{2/(1+\lambda)} \quad \textit{for each set} \quad \mathfrak{u} \subseteq \mathscr{D}. \tag{30}$$

Proof. Setting $\psi \equiv 1$ and using the symmetry of ϕ, it follows from the results in [11] (see also [9] for product weights) that the generating vector $\boldsymbol{z}$ for a randomly shifted lattice rule can be constructed by the CBC algorithm such that the shift-averaged worst-case error satisfies

$$\hat{e}_{n,d,\boldsymbol{\gamma}}(\boldsymbol{z}) \le \left(\frac{1}{\varphi(n)} \sum_{\mathfrak{u}\subseteq\mathscr{D}} \gamma_{\mathfrak{u}}^{\lambda} (\varrho(\lambda))^{|\mathfrak{u}|}\right)^{1/(2\lambda)} \quad \text{for all } \lambda \in (1/(2r), 1], \tag{31}$$

where $\rho(\lambda)$ is defined by (28), with the values of $C > 0$ and $r > 1/2$ chosen to satisfy

$$\hat{\theta}(h) := \frac{2}{\pi^2 h^2} \int_0^{1/2} \frac{\sin^2(\pi h y)}{\phi(\Phi^{-1}(y))}\,\mathrm{d}y \le \frac{C}{h^r} \quad \text{for all positive integers } h.$$

Below we obtain C and r for our three choices of the density ϕ. (We remark that a few combinations of ϕ and ψ were analyzed in [9], but the case $\psi \equiv 1$ was not included.)

For the logistic density we have $\Phi^{-1}(y) = \nu \ln(y/(1-y))$ and $\phi(\Phi^{-1}(y)) = y(1-y)/\nu$. Since $1 - y \ge 1/2$ for $y \in [0, 1/2]$, we have

$$\hat{\theta}(h) \le \frac{4\nu}{\pi^2 h^2} \int_0^{1/2} \frac{\sin^2(\pi h y)}{y}\,\mathrm{d}y = \frac{4\nu}{\pi^2 h^2} \left(\int_0^{1/\pi} \frac{\sin^2(\pi t)}{t}\,\mathrm{d}t + \int_{1/\pi}^{h/2} \frac{\sin^2(\pi t)}{t}\,\mathrm{d}t \right),$$

where the last equality was obtained by the substitution $t = hy$. We now use $\sin^2(\pi t) \le \pi^2 t^2$ for $t \in [0, 1/\pi]$ and $\sin^2(\pi t) \le 1$ for $t \ge 1/\pi$ to obtain

$$\hat{\theta}(h) \le \frac{4\nu}{\pi^2 h^2} \left(\frac{1}{2} + \ln\left(\frac{\pi h}{2}\right) \right) \le \frac{2^{2-\delta}\nu}{\pi^{2-\delta}\delta e^{1-\delta/2}}\, h^{-2+\delta} \quad \text{for all } \delta > 0,$$

where the last inequality follows from $1/2 + \ln(\pi h/2) \le h^\delta \pi^\delta / (2^\delta \delta e^{1-\delta/2})$ for all $\delta > 0$. Indeed, it can be easily checked that the function $\xi(h) := (1/2 + \ln(\pi h/2))h^{-\delta}$ attains its maximum at $h_0 = (2/\pi)e^{1/\delta - 1/2}$. Consequently, $\xi(h) \le \xi(h_0)$ which is equivalent to the stated inequality.

For the normal density we have from [9, Eq. (25)] that $\exp((\Phi^{-1}(y))^2/(2\nu)) \le 1/y$ for $y \in (0, 1/2)$, which leads to

$$\hat{\theta}(h) \le \frac{2\sqrt{2\pi\nu}}{\pi^2 h^2} \int_0^{1/2} \frac{\sin^2(\pi h y)}{y}\,\mathrm{d}y.$$

The remaining derivation of C and r follows the argument for the logistic density.

For the Student density the values of C and r are obtained by taking $\alpha = 0$ in [9, Example 3].

The estimates (6), (27) and (31) together yield (29). The choice of weights which minimizes the right-hand side of (29) follows from [8, Lemma 6.2]. □

7 Discussion

In summary, we obtained POD weights (30) for the likelihood integrand from a time series model. This is no coincidence: it is the consequence of our deliberate aim to obtain a bound in Lemma 2 that depends on the set $\mathfrak{v}$ only through its cardinality and depends on the variables x_k in a multiplicative way.

POD weights provide enough flexibility to model more complicated structure between variables, and yet they are of a simple enough form to allow the practical implementation of CBC construction for lattice rules.

Since r can be arbitrarily close to 1 in Theorem 2, we can take λ close to $1/2$ to yield nearly order n^{-1} convergence, which is optimal in this function space setting.

In our analysis we have considered the "anchored" version of the function space setting. However, as recently highlighted in [4, Sect. 5] and [11], the CBC construction in the anchored setting is more costly due to the need to switch to some "auxiliary weights" for the implementation. One way to avoid this issue is to consider instead the "unanchored" function space setting in the analysis, meaning that we need to modify our estimates to bound a different norm for the integrand.

Alternatively, we can continue to work with the anchored setting, but modify the weights following the strategy in [4, pp. 208–209], at the expense of enlarging the overall error estimate. We leave these possible refinements for future work.

Acknowledgements The authors acknowledge support from the Australian Research Council.

References

1. Breslow, N.E., Clayton, D.G.: Approximate inference in generalized linear mixed models. JASA **88**, 9–25 (1993)
2. Davis, R.A., Dunsmuir, W.T.M., Wang, Y.: On autocorrelation in a Poisson regression model. Biometrika **87**, 491–505 (2000)
3. Davis, R.A., Wang, Y., Dunsmuir, W.T.M.: Modeling time series of count data. In: Ghosh, S. (eds.) Asymptotics, Nonparametrics, and Time Series. Statistics, Textbooks and Monographs, vol. 158, pp. 63–113. Dekker, New York (1999)
4. Dick, J., Kuo, F.Y., Sloan, I.H.: High-dimensional integration – the quasi-Monte Carlo way. Acta Numer. **22**, 133–288 (2013)
5. Kuo, F.Y.: Component-by-component constructions achieve the optimal rate of convergence for multivariate integration in weighted Korobov and Sobolev spaces. J. Complexity **19**, 301–320 (2003)
6. Kuo, F.Y., Dunsmuir, W.T.M., Sloan, I.H., Wand, M.P., Womersley, R.S.: Quasi-Monte Carlo for highly structured generalised response models. Methodol. Comput. Appl. Probab. **10**, 239–275 (2008)
7. Kuo, F.Y., Schwab, Ch., Sloan, I.H.: Quasi-Monte Carlo methods for high dimensional integration: the standard (weighted Hilbert space) setting and beyond. ANZIAM J. **53**, 1–37 (2011)
8. Kuo, F.Y., Schwab, Ch., Sloan, I.H.: Quasi-Monte Carlo finite element methods for a class of elliptic partial differential equations with random coefficients. SIAM J. Numer. Anal. **50**, 3351–3374 (2012)
9. Kuo, F.Y., Sloan, I.H., Wasilkowski, G.W., Waterhouse, B.J.: Randomly shifted lattice rules with the optimal rate of convergence for unbounded integrands. J. Complexity **26**, 135–160 (2010)
10. McCulloch, C.E., Searle, S.R.: Generalized, Linear, and Mixed Models. Wiley Series in Probability and Statistics: Texts, References, and Pocketbooks Section. Wiley-Interscience, New York (2001)
11. Nichols, J.A., Kuo, F.Y.: Fast CBC construction of randomly shifted lattice rules achieving $\mathcal{O}(N^{-1+\delta})$ convergence for unbounded integrands in $\mathbb{R}^d$ in weighted spaces with POD weights (2013, submitted)
12. Nuyens, D., Cools, R.: Fast algorithms for component-by-component construction of rank-1 lattice rules in shift-invariant reproducing kernel Hilbert spaces. Math. Comp. **75**, 903–920 (2006)
13. Ruppert, D., Wand, M.P., Carroll, R.J.: Semiparametric Regression. Cambridge Series in Statistical and Probabilistic Mathematics, vol. 12. Cambridge University Press, Cambridge (2003)
14. Sinescu, V., L'Ecuyer, P.: Existence and construction of shifted lattice rules with an arbitrary number of points and bounded weighted star discrepancy for general decreasing weights. J. Complexity **27**, 449–465 (2011)
15. Sloan, I.H., Kuo, F.Y., Joe, S.: Constructing randomly shifted lattice rules in weighted Sobolev spaces. SIAM J. Numer. Anal. **40**, 1650–1665 (2002)

16. Sloan, I.H., Wang, X., Woźniakowski, H.: Finite-order weights imply tractability of multivariate integration. J. Complexity **20**, 46–74 (2004)
17. Sloan, I.H., Woźniakowski, H.: When are quasi-Monte Carlo algorithms efficient for high dimensional integrals? J. Complexity **14**, 1–33 (1998)
18. Wasilkowski, G.W., Woźniakowski, H.: Complexity of weighted approximation over $\mathbb{R}^1$. J. Approx. Theory **103**, 223–251 (2000)
19. Wasilkowski, G.W., Woźniakowski, H.: Tractability of approximation and integration for weighted tensor product problems over unbounded domains. In: Fang, K.-T., Hickernell, F.J., Niederreiter, H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2000, pp. 497–522. Springer, Berlin/Heidelberg (2002)

Multi-level Monte Carlo Finite Difference and Finite Volume Methods for Stochastic Linear Hyperbolic Systems

Jonas Šukys, Siddhartha Mishra, and Christoph Schwab

Abstract We consider stochastic linear hyperbolic systems of conservation laws in several space dimensions. We prove existence and uniqueness of a random weak solution and provide estimates for the space-time as well as statistical regularity of the solution in terms of the corresponding estimates for the random input data. Multi-Level Monte Carlo Finite Difference and Finite Volume algorithms are used to approximate such statistical moments in an efficient manner. We present novel probabilistic computational complexity analysis which takes into account the *sample path dependent* complexity of the underlying FDM/FVM solver, due to the *random CFL-restricted time step size* on account of the wave speed in a random medium. Error bounds for mean square error vs. *expected* computational work are obtained. We present numerical experiments with uncertain uniformly as well as log-normally distributed wave speeds that illustrate the theoretical results.

1 Introduction

Linear hyperbolic systems of conservation laws arise in a very large number of models in physics and engineering such as the acoustic wave equation, equations of linear elasticity and linearized shallow water and Euler equations. For a given bounded domain $\mathbf{D} \subset \mathbb{R}^d$, the general form of a linear hyperbolic system is given by

J. Šukys (✉) · S. Mishra · C. Schwab
SAM, ETH Zürich, Rämistrasse 101, Zurich, Switzerland
e-mail: jonas.sukys@sam.math.ethz.ch; smishra@sam.math.ethz.ch;
christoph.schwab@sam.math.ethz.ch

J. Dick et al. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2012*,
Springer Proceedings in Mathematics & Statistics 65,
DOI 10.1007/978-3-642-41095-6_34,

$$\begin{cases} \mathbf{U}_t(\mathbf{x},t) + \sum_{r=1}^{d} (\mathbf{A}_r(\mathbf{x})\mathbf{U}(\mathbf{x},t))_{\mathbf{x}_r} = \mathbf{S}(\mathbf{x}), & \forall (\mathbf{x},t) \in \mathbf{D} \times \mathbb{R}_+. \\ \mathbf{U}(\mathbf{x},0) = \mathbf{U}_0(\mathbf{x}), \end{cases} \quad (1)$$

Here, $\mathbf{U} : \mathbf{D} \times \mathbb{R}_+ \to \mathbb{R}^m$ denotes the vector of conserved variables, $\mathbf{A}_r : \mathbb{R}^m \to \mathbb{R}^m$ denote linear maps (linear fluxes), and $\mathbf{S} : \mathbf{D} \to \mathbb{R}^m$ denotes the source term. Equation (1) is augmented with initial data $\mathbf{U}_0 \in \mathbf{D} \to \mathbb{R}^m$ and suitable boundary conditions.

Given the lack of explicit solution formulas (particularly for variable coefficients and in several space dimensions), numerical methods are widely used to approximate (1). Popular discretization methods include finite difference, finite volume and discontinuous Galerkin methods, see [6,9,16] and references therein.

These numerical methods require the specification of the coefficient matrices, initial data, source terms and boundary data as input. However, these quantities are often determined by measurements. Measurements are typically uncertain and only provide statistical information about the input data. As an example, consider the propagation of acoustic waves in the subsurface. The wave speeds, being dependent on the material properties of medium, are generally determined up to some statistical moments. This uncertainty in the input data results in the solution being uncertain. The efficient computation of the solution uncertainty, given the input uncertainty, is the central theme of uncertainty quantification (UQ).

A necessary prerequisite in UQ is to formulate an appropriate mathematical notion of random solutions for linear hyperbolic systems. *The first aim* of this paper is to provide an appropriate framework of random solutions of (1) and prove existence, uniqueness as well as (spatio-temporal and statistical) regularity of these solutions.

The second aim is to present *efficient* numerical methods for approximation of *random* version of linear hyperbolic systems (1). Examples of such methods include the stochastic Galerkin and stochastic collocation, see references in [11]. Currently these methods are not able to handle large number of uncertainty sources, are *intrusive* (existing deterministic solvers need to be reconfigured) and hard to parallelize.

Another class of methods are the *Monte Carlo* (MC) methods where the underlying deterministic PDE is solved for each statistical *sample* and the samples are combined to ascertain statistical information. However, MC methods are inefficient due to the error convergence rate of $1/2$: a large number of numerical solves of (1) is required. Such slow convergence has inspired the development of *Multi-Level Monte Carlo* (MLMC) methods. They were introduced by S. Heinrich for numerical quadrature [8], developed by M. Giles for Itô SPDE [4], and applied to various SPDEs [2,3,13]. In particular, recent papers [10–12] extended the MLMC algorithm to nonlinear conservation laws. Massively parallel simulations of the random multi-dimensional Euler, magnetohydrodynamics (MHD) and shallow water equations were conducted using novel static load balancing [15].

In this paper, we extend the MLMC methods for computing uncertainty in the solutions of random linear hyperbolic systems. It is essential to point out the following two novel features of this paper in the context of MC and MLMC methods:

- Due to the linearity of the underlying evolution equations, we are able to prove rigorous error estimates for linear hyperbolic systems, both in terms of the spatio-temporal discretization parameter as well as in terms of the number of samples, for the MC as well as MLMC discretizations. This should be contrasted with the results in a recent paper [11], where we postulated error estimates for the MC and MLMC discretizations for nonlinear hyperbolic systems.
- A new feature that emerges in the treatment of linear equations with random coefficients is the *probabilistic nature of the computational complexity estimates*. This is due to the fact that the wave speed (depending on the eigenvalues of the coefficient matrices in (1)) is random and can have a large statistical spread, for instance, if the wave speed is log-normally distributed. Consequently, for the popular explicit time stepping schemes (such as those employed in the current paper), the time step size, being specified in terms of the wave speed due to the CFL condition, is *random*. Hence, the number of time steps as well as the total computational work for the whole simulation is a random quantity. Thus, we device novel probabilistic computational complexity estimates to account for this randomness in computational work. *A crucial result of this paper is to derive such probabilistic work estimates* for the MLMC methods and to show that the expected work of the MLMC methods is asymptotically the same as that of a single deterministic run of the underlying finite difference or finite volume scheme. Hence, the MLMC method is considerably superior to the MC methods. To the best of our knowledge, this is the first time that such probabilistic complexity estimates have been obtained for MC discretizations of random PDEs.

To illustrate theoretical results, we present numerical experiments for the acoustic wave equation in uniformly as well as log-normally distributed random medium.

2 Linear Systems of Stochastic Hyperbolic Conservation Laws

Definition 1 (Strong hyperbolicity). In the case $d = 1$, the linear system of conservation laws (1) is called strongly hyperbolic [6] if $\forall \mathbf{x} \in \mathbf{D}, \exists \mathbf{Q}_{\mathbf{x}} : \mathbb{R}^m \to \mathbb{R}^m$:

$$\sup_{\mathbf{x}\in\mathbf{D}} \|\mathbf{Q}_{\mathbf{x}}^{-1}\| \|\mathbf{Q}_{\mathbf{x}}\| \leq K < \infty, \qquad \mathbf{Q}_{\mathbf{x}}^{-1}\mathbf{A}_1(\mathbf{x})\mathbf{Q}_{\mathbf{x}} \quad \text{is diagonal.} \tag{2}$$

For extension of strong hyperbolicity to $d > 1$ spatial variables we refer to [6].

Let V denote an arbitrary Banach space. The following notation will be used:

$$\|\mathbf{U}, \mathbf{S}, t\|_V = \|\mathbf{U}\|_V + t\|\mathbf{S}\|_V, \quad \mathbf{U}, \mathbf{S} \in V,\ t \geq 0. \tag{3}$$

The following result recapitulates some of the classical existence and uniqueness results [6, 9, 16] for weak solutions of linear hyperbolic systems (1).

Theorem 1. *Denote* $\mathbf{L}^p(\mathbf{D}) = L^p(\mathbf{D})^m$, $\mathbf{W}^{r,\infty}(\mathbf{D}) = W^{r,\infty}(\mathbf{D})^m$, *and assume that*

1. *The linear system* (1) *is strongly hyperbolic with* $K < \infty$ *in* (2),
2. *There exist* $r_{\mathbf{0}}, r_{\mathbf{S}}, r_{\mathbf{A}} \in \mathbb{N} \cup \{0, \infty\}$ *such that:*

$$\mathbf{U}_0 \in \mathbf{W}^{r_{\mathbf{0}},\infty}(\mathbf{D}), \quad \mathbf{S} \in \mathbf{W}^{r_{\mathbf{S}},\infty}(\mathbf{D}), \quad \mathbf{A}_r \in \mathbf{W}^{r_{\mathbf{A}},\infty}(\mathbf{D})^m. \tag{4}$$

Then, for every finite time horizon $T < \infty$, (1) *admits a unique weak solution* $\mathbf{U} \in L^\infty(\mathbf{D} \times [0, T])^m$. *Furthermore, for every* $0 \leq t \leq T$, *the a priori estimates hold:*

$$\|\mathbf{U}(\cdot, t)\|_{\mathbf{L}^2(\mathbf{D})} \leq K\|\mathbf{U}_0, \mathbf{S}, t\|_{\mathbf{L}^2(\mathbf{D})}, \tag{5}$$

$$\|\mathbf{U}(\cdot, t) - \mathbf{V}(\cdot, t)\|_{\mathbf{L}^2(\mathbf{D})} \leq K\|\mathbf{U}_0 - \mathbf{V}_0,\ \mathbf{S}_\mathbf{U} - \mathbf{S}_\mathbf{V},\ t\|_{\mathbf{L}^2(\mathbf{D})}, \tag{6}$$

$$\mathbf{U} \in C([0, T], \mathbf{W}^{\bar{r},\infty}(\mathbf{D})), \quad \textit{with} \quad \bar{r} = \min\{r_{\mathbf{0}}, r_{\mathbf{S}}, r_{\mathbf{A}}\}. \tag{7}$$

Let $(\Omega, \mathscr{F}, \mathbb{P})$ denote a complete probability space and $\mathscr{B}(V)$ a Borel σ-algebra.

Definition 2 (Random field). A V-valued random field is a measurable mapping

$$\mathbf{U} : (\Omega, \mathscr{F}) \to (V,\ \mathscr{B}(V)), \quad \omega \mapsto \mathbf{U}(\mathbf{x}, t, \omega).$$

The stochastic version of the linear system of hyperbolic conservation laws (1) is

$$\begin{cases} \mathbf{U}_t(\mathbf{x}, t, \omega) + \displaystyle\sum_{r=1}^{d} \mathbf{A}_r(\mathbf{x}, \omega)\mathbf{U}_{\mathbf{x}_r} = \mathbf{S}(\mathbf{x}, \omega), & \forall (\mathbf{x}, t) \in \mathbf{D} \times \mathbb{R}_+, \quad \forall \omega \in \Omega. \\ \mathbf{U}(\mathbf{x}, 0, \omega) = \mathbf{U}_0(\mathbf{x}, \omega), \end{cases} \tag{8}$$

Here, $\mathbf{U}_0$ and $\mathbf{S}$ are $\mathbf{L}^2(\mathbf{D})$-valued random fields $(\Omega, \mathscr{F}) \to (\mathbf{L}^2(\mathbf{D}), \mathscr{B}(\mathbf{L}^2(\mathbf{D})))$. The fluxes $\mathbf{A}_r$ are $\mathbf{L}^\infty(\mathbf{D})^m$-valued random fields $(\Omega, \mathscr{F}) \to (\mathbf{L}^\infty(\mathbf{D})^m, \mathscr{B}(\mathbf{L}^\infty(\mathbf{D})^m))$. We define the following notion of solutions of (8):

Definition 3 (Random weak solution). A $C([0, T], \mathbf{L}^2(\mathbf{D}))$-valued random field $\mathbf{U} : \Omega \ni \omega \mapsto \mathbf{U}(\mathbf{x}, t, \omega)$ is a random weak solution to the stochastic linear hyperbolic system of conservation laws (8) if it is a weak solution of (1) for $\mathbb{P}$-a.e. $\omega \in \Omega$.

Based on Theorem 1, we obtain the following well-posedness result for (8).

Theorem 2. *In* (8), *assume that the following holds for some* $k \in \mathbb{N} \cup \{0, \infty\}$:

1. *Equation (8) is strongly hyperbolic, with* $\bar{K}_k = \|K(\omega)\|_{L^k(\Omega,\mathbb{R})} < \infty$,
2. *There exists non-negative integers* $r_{\mathbf{0}}, r_{\mathbf{S}}, r_{\mathbf{A}} \in \mathbb{N} \cup \{0, \infty\}$ *such that:*

$$\mathbf{U}_0 \in L^k(\Omega, \mathbf{W}^{r_{\mathbf{0}},\infty}(\mathbf{D})), \quad \mathbf{S} \in L^k(\Omega, \mathbf{W}^{r_{\mathbf{S}},\infty}(\mathbf{D})), \quad \mathbf{A}_r \in L^0(\Omega, \mathbf{W}^{r_{\mathbf{A}},\infty}(\mathbf{D})^m), \tag{9}$$

3. *Each random field* $\mathbf{A}_r$, $r = 1, \ldots, d$, *is independent of* $\mathbf{U}_0$ *and* $\mathbf{S}$ *on* $(\Omega, \mathscr{F}, \mathbb{P})$.

Then, for $T < \infty$, (8) *admits a unique random weak solution*

$$\mathbf{U} : \Omega \to C([0,T], \mathbf{L}^2(\mathbf{D})), \qquad \omega \mapsto \mathbf{U}^\omega(\cdot,\cdot), \quad \forall \omega \in \Omega, \tag{10}$$

where $\mathbf{U}^\omega(\cdot,\cdot)$ *is the solution to the* deterministic *system* (1)*. Moreover,* $\forall t \in [0,T]$,

$$\|\mathbf{U}(\cdot,t,\omega)\|_{\mathbf{L}^2(\mathbf{D})} \le K(\omega)\|\mathbf{U}_0(\cdot,\omega), \mathbf{S}(\cdot,\omega), t\|_{\mathbf{L}^2(\mathbf{D})}, \quad \mathbb{P}\text{-}a.s., \tag{11}$$

$$\|\mathbf{U}\|_{L^k(\Omega,C([0,T],\mathbf{L}^2(\mathbf{D})))} \le \bar{K}_k \|\mathbf{U}_0, \mathbf{S}, t\|_{L^k(\Omega,\mathbf{L}^2(\mathbf{D}))}, \tag{12}$$

with $\|\mathbf{U}, \mathbf{S}, t\|_{L^k(\Omega,V)} = \|\|\mathbf{U}, \mathbf{S}, t\|_V\|_{L^k(\Omega,\mathbb{R})}$.

We will use the following lemma in the proof.

Lemma 1. *Let* E *be a separable Banach space and* $X : \Omega \to E$ *be an* E*-valued random variable on* $(\Omega, \mathscr{F})$*. Then, mapping* $\Omega \ni \omega \mapsto \|X(\omega)\|_E \in \mathbb{R}$ *is measurable.*

Proof (of Theorem 2). To retain brevity of exposition, we only outline the key steps, following [10].

1. By Theorem 1, the random field in (10) is well defined for $\mathbb{P}$-a.e. $\omega \in \Omega$. Furthermore, for $\mathbb{P}$-a.e. $\omega \in \Omega$, $\mathbf{U}(\cdot,\cdot,\omega)$ is a weak solution of (1).
2. $\forall t \in [0,T]$, $\forall j = 1, \ldots, m$, we verify the measurability of the component map $\Omega \ni \omega \mapsto \mathbf{U}_j(\cdot,t,\omega) \in L^2(\mathbf{D})$. Since $L^2(\mathbf{D})$ is a separable Hilbert space, the $\mathscr{B}(L^2(\mathbf{D}))$ is the smallest σ-algebra containing all subsets

$$\{v \in L^2(\mathbf{D}) : \varphi(v) \le \alpha\} \quad : \quad \varphi \in L^2(\mathbf{D}), \ \alpha \in \mathbb{R}.$$

   For a fixed $\alpha \in \mathbb{R}$, $\varphi \in L^2(\mathbf{D})$, consider the set $\{\mathbf{U}_j(\cdot,t,\omega) : \varphi(\mathbf{U}_j(\cdot,t,\omega)) \le \alpha\}$. By continuity (5) in $L^2(\mathbf{D})$, since $\mathbf{U}_0, \mathbf{S} \in L^0(\Omega, \mathbf{L}^2(\mathbf{D}))$ and $\mathbf{A}_r \in L^0(\Omega, \mathbf{L}^2(\mathbf{D})^m)$, we obtain $\mathbf{U}_j(\cdot,t,\cdot) \in L^0(\Omega, L^2(\mathbf{D}))$, for every $0 \le t \le T$.
3. Equation (11) follows from (5) and Lemma 1; (12) follows from (11) and hypothesis 3,

$$\begin{aligned}\|\mathbf{U}\|^k_{L^k(\Omega,C([0,T],\mathbf{L}^2(\mathbf{D})))} &= \mathbb{E}\left[\max_{0\le t\le T} \|\mathbf{U}(\cdot,t,\omega)\|^k_{\mathbf{L}^2(\mathbf{D})}\right] \\ &\le \mathbb{E}\left[K^k(\omega)\|\mathbf{U}_0, \mathbf{S}, t\|^k_{\mathbf{L}^2(\mathbf{D})}\right] = \bar{K}^k_k \|\mathbf{U}_0, \mathbf{S}, t\|^k_{L^k(\Omega,\mathbf{L}^2(\mathbf{D}))}.\end{aligned}$$

This theorem ensures the existence of the k-th moments $\mathscr{M}^k(\mathbf{U}) \in (\mathbf{L}^2(\mathbf{D}))^k$ [10] of the random weak solution, provided $\mathbf{U}_0, \mathbf{S} \in L^k(\Omega, \mathbf{L}^2(\mathbf{D}))$ and $K \in L^k(\Omega, \mathbb{R})$. □

3 Multi-level Monte Carlo FVM and FDM Methods

3.1 Monte Carlo Method

Under the hypotheses of Theorem 2, the unique random weak solution exists and has bounded k-th moments [10]. We are interested in the computational estimation of the "mean field" or "ensemble average", i.e. of $\mathscr{M}^1(\mathbf{U}) = \mathbb{E}[\mathbf{U}]$. To this end, we use the *Monte Carlo (MC) method* to approximate $\mathbb{E}[\mathbf{U}]$: fix $M \in \mathbb{N}$ and let $\hat{\mathbf{I}}^i := \{\hat{\mathbf{U}}_0^i, \hat{\mathbf{S}}^i, \hat{\mathbf{A}}_1^i, \ldots, \hat{\mathbf{A}}_d^i\}$ be independent, identically distributed (i.i.d.) samples of input data $\hat{\mathbf{I}}(\omega) := \{\hat{\mathbf{U}}_0(\omega), \hat{\mathbf{S}}(\omega), \hat{\mathbf{A}}_1(\omega), \ldots, \hat{\mathbf{A}}_d(\omega)\}$. The *Monte Carlo (MC)* estimate of the expectation $\mathscr{M}^1(\mathbf{U}) = \mathbb{E}[\mathbf{U}(\cdot, t, \cdot)]$ at fixed time t is given by the sample average

$$E_M[\mathbf{U}(\cdot, t, \cdot)] := \frac{1}{M} \sum_{i=1}^{M} \hat{\mathbf{U}}^i(\cdot, t), \tag{13}$$

where $\hat{\mathbf{U}}^i(\cdot, t)$ denotes the M unique random weak solutions of the deterministic linear system of conservation laws (1) with the input data $\hat{\mathbf{I}}^i$. By (11), we have

$$\|E_M[\mathbf{U}]\|_{\mathbf{L}^2(\mathbf{D})} = \frac{1}{M} \left\| \sum_{i=1}^{M} \hat{\mathbf{U}}^i \right\|_{\mathbf{L}^2(\mathbf{D})} \leq \frac{1}{M} \sum_{i=1}^{M} \|\hat{\mathbf{U}}^i\|_{\mathbf{L}^2(\mathbf{D})} \overset{(11)}{\leq} \frac{1}{M} \sum_{i=1}^{M} K^i \|\hat{\mathbf{U}}_0^i, \hat{\mathbf{S}}^i, t\|_{\mathbf{L}^2(\mathbf{D})}.$$

Using the i.i.d. property of the samples $\{\hat{\mathbf{I}}^i\}_{i=1}^M$ of the random input data $\mathbf{I}(\omega)$, Lemma 1, the linearity of $\mathbb{E}[\cdot]$ and hypothesis 3 in Theorem 2, we obtain

$$\mathbb{E}\left[\|E_M[\mathbf{U}(\cdot, t, \omega)]\|_{\mathbf{L}^2(\mathbf{D})}\right] = \bar{K}_1 \|\hat{\mathbf{U}}_0(\cdot, \omega), \hat{\mathbf{S}}(\cdot, \omega), t\|_{L^1(\Omega, \mathbf{L}^2(\mathbf{D}))} < \infty. \tag{14}$$

The following result states that MC estimates (13) converge as $M \to \infty$.

Theorem 3. *Assume the hypothesis of Theorem 2 is satisfied with $k \geq 2$, i.e. the second moments of the random initial data $\mathbf{U}_0$, source $\mathbf{S}$ and K exist. Then, the MC estimates $E_M[\mathbf{U}(\cdot, t, \omega)]$ in (13) converge to $\mathscr{M}^1(\mathbf{U}) = \mathbb{E}[\mathbf{u}]$ as $M \to \infty$. Furthermore,*

$$\|\mathbb{E}[\mathbf{U}(t)] - E_M[\mathbf{U}(t)](\omega)\|_{L^2(\Omega, \mathbf{L}^2(\mathbf{D}))} \leq M^{-\frac{1}{2}} \bar{K}_2 \|\mathbf{U}_0, \mathbf{S}, t\|_{L^2(\Omega, \mathbf{L}^2(\mathbf{D}))}. \tag{15}$$

Proof. We follow the structure of the analogous proofs in [2]. The M samples $\{\hat{\mathbf{I}}^i\}_{i=1}^M$ are interpreted as realizations of M independent "copies" of $\mathbf{I}(\omega)$ on the probability space $(\Omega, \mathscr{F}, \mathbb{P})$, i.e. $\hat{\mathbf{I}}^i = \hat{\mathbf{I}}^i(\omega)$. By $\mathbf{L}^2(\mathbf{D})$ contractivity (6), $\forall 0 \le t \le T$, solutions $\hat{\mathbf{U}}(\cdot, t, \omega)$ of any two i.i.d. realizations of $\mathbf{I}(\omega)$ are strongly measurable as $\mathbf{L}^2(\mathbf{D})$-valued functions, hence are independent random fields. By Lemma 1 and by continuity (11), the mapping $\omega \mapsto \|\mathbf{U}(\cdot, t, \omega)\|_{\mathbf{L}^2(\mathbf{D})}$ is measurable. Hence,

$$\mathbb{E}\left[\|\mathbb{E}[\mathbf{U}] - E_M[\mathbf{U}](\omega)\|^2_{\mathbf{L}^2(\mathbf{D})}\right] = \frac{1}{M^2}\mathbb{E}\left[\sum_{i=1}^{M}\|\mathbb{E}[\mathbf{U}] - \hat{\mathbf{U}}^i(\omega)\|^2_{\mathbf{L}^2(\mathbf{D})}\right]$$
$$= \frac{1}{M}\mathbb{E}\left[\|\mathbb{E}[\mathbf{U}] - \mathbf{U}\|^2_{\mathbf{L}^2(\mathbf{D})}\right] = \frac{1}{M}\left(\mathbb{E}\|\mathbf{U}\|^2_{\mathbf{L}^2(\mathbf{D})} - \|\mathbb{E}[\mathbf{U}]\|^2_{\mathbf{L}^2(\mathbf{D})}\right) \le \frac{1}{M}\mathbb{E}\|\mathbf{U}\|^2_{\mathbf{L}^2(\mathbf{D})}.$$

Using (11), hypothesis 3 of Theorem 2, and notation $\mathbf{U}(t) = \mathbf{U}(\cdot, t, \omega)$, we deduce

$$\mathbb{E}\left[\|\mathbb{E}[\mathbf{U}(t)] - E_M[\mathbf{U}(t)](\omega)\|^2_{\mathbf{L}^2(\mathbf{D})}\right] \le M^{-1}\bar{K}_2^2\mathbb{E}\left[\|\mathbf{U}_0, \mathbf{S}, t\|^2_{\mathbf{L}^2(\mathbf{D})}\right],$$

which implies (15) upon taking square roots. □

3.2 Finite Difference and Finite Volume Methods

In the derivation of (15), we have assumed that the *exact* random weak solutions $\hat{\mathbf{U}}^i(\mathbf{x}, t, \omega)$ of (1) are available. In MC-FDM/FVM and MLMC-FDM/FVM, solutions are approximated by Finite Difference [6] and Finite Volume [9] methods.

If $\mathbf{U}_0$ and $\mathbf{S}$ are continuous (then solution $\mathbf{U}$ is also continuous), conventional Finite Difference methods [6,16] can be used where spatial and temporal derivatives in (1) are approximated by upwinded difference quotients. For discontinuous $\mathbf{U}_0$ and $\mathbf{S}$, (then solution $\mathbf{U}$ is also discontinuous) we present Finite Volume Method.

For simplicity of exposition, we consider here *periodic* Cartesian physical domains $\mathbf{D} = I_1 \times \cdots \times I_d \subset \mathbb{R}^d$. However, all results of the present paper also extend to systems (1) in general polyhedral domains with suitable boundary conditions.

Let $\mathscr{T} = \mathscr{T}^1 \times \cdots \times \mathscr{T}^d$ denote a uniform axiparallel quadrilateral mesh of the domain $\mathbf{D}$, consisting of identical cells $C_{\mathbf{j}} = C_{\mathbf{j}_1} \times \cdots \times C_{\mathbf{j}_d}, \mathbf{j}_r = 1, \ldots, \#\mathscr{T}^r$.

Assume *mesh widths* are equal in each dimension, i.e. $\Delta x := \frac{|I_1|}{\#\mathscr{T}_1} = \cdots = \frac{|I_d|}{\#\mathscr{T}_d}$. Define the approximations to cell averages of the solution $\mathbf{U}$ and source term $\mathbf{S}$ by

$$\mathbf{U}_{\Delta x}(\mathbf{x}, t) = \mathbf{U}_{\mathbf{j}}(t) \approx \frac{1}{|C_{\mathbf{j}}|}\int_{C_{\mathbf{j}}}\mathbf{U}(\mathbf{x}, t)d\mathbf{x}, \quad \forall \mathbf{x} \in C_{\mathbf{j}}, \quad \mathbf{S}_{\mathbf{j}} \approx \frac{1}{|C_{\mathbf{j}}|}\int_{C_{\mathbf{j}}}\mathbf{S}(\mathbf{x})d\mathbf{x}.$$

Then, a semi-discrete finite volume scheme [9] for approximating (1) is given by

$$\partial_t \mathbf{U}_{\mathbf{j}}(t) = -\sum_{r=1}^{d} \frac{1}{\Delta x}\left(\mathbf{F}^r_{\mathbf{j}+\frac{1}{2}} - \mathbf{F}^r_{\mathbf{j}-\frac{1}{2}}\right) - \mathbf{S}_{\mathbf{j}}, \tag{16}$$

where *numerical fluxes* $\mathbf{F}^r_{\cdot}$ are defined by using (approximate) solutions of local Riemann problems (in direction r) at each cell interface. High order accuracy is achieved by using non-oscillatory TVD, ENO, WENO methods [5,7]. Approximations $\mathbf{U}^n_{\Delta x} = \mathbf{U}_{\Delta x}(\cdot, t^n)$ at time steps t^n are obtained by SSP Runge-Kutta methods.

Assumption 1. *We assume that the abstract FDM or FVM scheme* (16) *satisfies*

$$\|\mathbf{U}^n_{\Delta x}\|_{\mathbf{L}^2(\mathbf{D})} \le K \|\mathbf{U}^0_{\Delta x}\|_{\mathbf{L}^2(\mathbf{D})}, \tag{17}$$

and the approximation error converges (as $\Delta x \to 0$*) with rate* $s > 0$*, i.e.*

$$\|\mathbf{U}_0 - \mathbf{U}^0_{\Delta x}\|_{\mathbf{L}^2(\mathbf{D})} \le C\,\Delta x^s \|\mathbf{U}_0\|_{\mathbf{H}^s(\mathbf{D})}, \quad \|\mathbf{S} - \mathbf{S}_{\Delta x}\|_{\mathbf{L}^2(\mathbf{D})} \le C\,\Delta x^s \|\mathbf{S}\|_{\mathbf{H}^s(\mathbf{D})}, \tag{18}$$

$$\|\mathbf{U}(t^n) - \mathbf{U}^n_{\Delta x}\|_{\mathbf{L}^2(\mathbf{D})} \le C\,\Delta x^s \left(\|\mathbf{U}_0, \mathbf{S}, t^n\|_{\mathbf{H}^s(\mathbf{D})} + t^n \|\mathbf{U}_0, \mathbf{S}, t^n\|_{\mathbf{L}^2(\mathbf{D})}\right), \tag{19}$$

provided $\Delta t = \Delta x/\lambda$. $C, \lambda > 0$ *are independent of* Δx. $\mathbf{H}^s(\mathbf{D})$ *denotes* $W^{s,2}(\mathbf{D})^m$.

Assumption 1 is satisfied by many standard FDM and FVM (for small s) schemes, we refer to [6,9,16] and the references therein. For q-th order (formally) accurate schemes, $q \in \mathbb{N}$, the convergence estimate (19) holds [6,9] with

$$s = \min\{q, \bar{r}\} \quad \text{(FDM)}, \qquad s = \min\{q, \max\{\min\{2, q\}/2, \bar{r}\}\} \quad \text{(FVM)}. \tag{20}$$

We assume the computational work of FDM/FVM for a time step and for a complete run to behave as

$$\mathrm{Work}^{\mathrm{step}}_{\Delta x} = B\,\Delta x^{-d}, \quad \mathrm{Work}_{\Delta x} = \mathrm{Work}^{\mathrm{step}}_{\Delta x} \frac{T}{\Delta t} = \lambda T B\,\Delta x^{-(d+1)}, \tag{21}$$

where $B > 0$ is independent of Δx and Δt. However, in the *random* case (8), the computational work (21) of FDM/FVM for one complete run depends on the particular realization of the coefficient $c(\cdot, \omega)$: due to the CFL condition ensuring the numerical stability of the explicit time stepping, the number of time steps $N(\Delta x, \omega)$ depends on the speed λ of the fastest moving wave, where $\lambda(\omega) = \sum_{r=1}^{d} \|\sigma^r_{\max}(\cdot, \omega)\|_{L^\infty(\mathbf{D})}$,

$$N(\Delta x, \omega) = \lambda(\omega) T/\Delta x = T/\Delta x \max_{1 \le r \le d} \|\sigma^r_{\max}(\cdot, \omega)\|_{L^\infty(\mathbf{D})}, \quad \forall \omega \in \Omega. \tag{22}$$

Here, $\sigma^r_{\max} = \max\{\sigma^r_1, \ldots, \sigma^r_m\}$, where $\sigma^r_1(\mathbf{x}, \omega), \ldots, \sigma^r_m(\mathbf{x}, \omega)$ are the eigenvalues of $\mathbf{A}_r(\mathbf{x}, \omega)$ and correspond to the directional speeds of the wave propagation at $\mathbf{x} \in \mathbf{D}$.

3.3 MC-FDM and MC-FVM Schemes

The MC-FDM or MC-FVM algorithm consists of the following three steps:

1. **Sample:** We draw M independent identically distributed (i.i.d.) input data samples $\mathbf{I}^i$ with $i = 1, 2, \ldots, M$ from the random fields $\mathbf{I}(\cdot, \omega)$ and approximate these by piece-wise constant functions obtained from cell averages.
2. **Solve:** For each realization $\mathbf{I}^i$, the underlying balance law (1) is solved numerically by the Finite Volume/Difference Method (16). Denote the solutions by $\mathbf{U}^{i,n}_{\Delta x}$.
3. **Estimate Statistics:** We estimate the expectation of the random solution field with the sample mean (ensemble average) of the approximate solution:

$$E_M[\mathbf{U}^n_{\Delta x}] := \frac{1}{M} \sum_{i=1}^{M} \mathbf{U}^{i,n}_{\Delta x}. \tag{23}$$

Higher statistical moments can be approximated analogously under suitable statistical regularity assumptions on the underlying random entropy solutions [10].

Theorem 4. *Assume the hypothesis of Theorem 2 is satisfied with $k \geq 2$, i.e. second moments of the random initial data $\mathbf{U}_0$, source $\mathbf{S}$ and K exist. Under Assumption 1, the MC-FDM/FVM estimate (23) satisfies the following error bound,*

$$\begin{aligned} \|\mathbb{E}[\mathbf{U}(t^n)] - E_M[\mathbf{U}^n_{\Delta x}](\omega)\|_{L^2(\Omega,\mathbf{L}^2(\mathbf{D}))} &\leq M^{-\frac{1}{2}} \bar{K}_2 \|\mathbf{U}_0, \mathbf{S}, t\|_{L^2(\Omega,\mathbf{L}^2(\mathbf{D}))} \\ &+ C\,\Delta x^s \left(\|\mathbf{U}_0, \mathbf{S}, t^n\|_{L^2(\Omega,\mathbf{H}^s(\mathbf{D}))} + t^n \|\mathbf{U}_0, \mathbf{S}, t^n\|_{L^2(\Omega,\mathbf{L}^2(\mathbf{D}))} \right), \end{aligned} \tag{24}$$

where $C > 0$ is independent of M, K and Δx.

Proof. Firstly, we bound the left hand side of (24) using the triangle inequality,

$$\begin{aligned} \|\mathbb{E}[\mathbf{U}(t^n)] - E_M[\mathbf{U}^n_{\Delta x}](\omega)\|_{L^2(\Omega,\mathbf{L}^2(\mathbf{D}))} &\leq \|\mathbb{E}[\mathbf{U}(t^n)] - E_M[\mathbf{U}(t^n)](\omega)\|_{L^2(\Omega,\mathbf{L}^2(\mathbf{D}))} \\ &+ \|E_M[\mathbf{U}(t^n)](\omega) - E_M[\mathbf{U}^n_{\Delta x}](\omega)\|_{L^2(\Omega,\mathbf{L}^2(\mathbf{D}))} = \mathrm{I} + \mathrm{II}. \end{aligned}$$

Term I is bounded by (15). Next, by the triangle inequality, and by (5) and (19),

$$\begin{aligned} \mathrm{II} &= \|E_M[\mathbf{U}(t^n) - \mathbf{U}^n_{\Delta x}](\omega)\|_{L^2(\Omega,\mathbf{L}^2(\mathbf{D}))} \\ &\leq \frac{1}{M} \sum_{i=1}^{M} \|\mathbf{U}^i(\cdot, t^n, \omega) - \mathbf{U}^{i,n}_{\Delta x}(\omega)\|_{L^2(\Omega,\mathbf{L}^2(\mathbf{D}))} = \|\mathbf{U}(\cdot, t^n, \omega) - \mathbf{U}^n_{\Delta x}(\omega)\|_{L^2(\Omega,\mathbf{L}^2(\mathbf{D}))} \\ &\leq C\,\Delta x^s \left\| \|\mathbf{U}_0, \mathbf{S}, t^n\|_{\mathbf{H}^s(\mathbf{D})} + t^n \|\mathbf{U}_0, \mathbf{S}, t^n\|_{\mathbf{L}^2(\mathbf{D})} \right\|_{L^2(\Omega,\mathbb{R})}. \end{aligned}$$

Finally, (24) is obtained by applying the triangle inequality on the last term. □

To equilibrate statistical and spatio-temporal discretization errors in (24), we require $M = O(\Delta x^{-2s})$. Next, we are interested in the asymptotic behavior of the error (24) vs. the *expected* computational work. We choose the algorithm parameters in order to maximize the convergence rate $\alpha > 0$, ie.

$$\|\mathbb{E}[\mathbf{U}(t^n)] - E_M[\mathbf{U}^n_{\Delta x}](\omega)\|_{L^2(\Omega,\mathbf{L}^2(\mathbf{D}))} \leq C \left(\mathbb{E}\left[\text{Work}^{\text{MC}}_{\Delta x}\right]\right)^{-\alpha}.$$

Assuming that the expected fastest wave speed $\bar{\lambda} = \mathbb{E}[\lambda(\omega)]$ in (22) is finite,

$$\mathbb{E}[\text{Work}^{\text{MC}}_{\Delta x}] = \mathbb{E}\left[M\,\text{Work}_{\Delta x}(\omega)\right] = \Delta x^{-(d+1+2s)} TB\bar{\lambda} < \infty. \tag{25}$$

Consequently, the asymptotic error bound (29) is satisfied with $\alpha = s/(d+1+2s)$, which is considerably lower than the deterministic rate $\alpha = s/(d+1)$.

3.4 MLMC-FDM and MLMC-FVM Schemes

Given the slow convergence of MC-FDM/FVM, we propose the Multi-Level Monte Carlo methods: MLMC-FDM and MLMC-FVM. The key idea is to simultaneously draw MC samples on a hierarchy of nested grids [10]. There are four steps:

0. **Nested meshes:** Consider *nested* meshes $\{\mathscr{T}_\ell\}_{\ell=0}^{\infty}$ of the domain $\mathbf{D}$ with corresponding mesh widths $\Delta x_\ell = 2^{-\ell}\Delta x_0$, where Δx_0 is the mesh width for the coarsest resolution and corresponds to the lowest level $\ell = 0$.
1. **Sample:** For each level of resolution $\ell \in \mathbb{N}_0$, we draw M_ℓ independent identically distributed (i.i.d) samples $\mathbf{I}^i_\ell$ with $i = 1, 2, \ldots, M_\ell$ from the random input data $\mathbf{I}(\omega)$ and approximate these by cell averages.
2. **Solve:** For each level ℓ and each realization $\mathbf{I}^i_\ell$, the balance law (1) is solved for $\mathbf{U}^{i,n}_{\Delta x_\ell}$ and $\mathbf{U}^{i,n}_{\Delta x_{\ell-1}}$ by the FDM/FVM method (16) with mesh widths Δx_ℓ and $\Delta x_{\ell-1}$.
3. **Estimate solution statistics:** Fix some positive integer $L < \infty$ corresponding to the highest level. Denoting MC estimator (23) with $M = M_\ell$ by E_{M_ℓ}, the expectation of the random solution field $\mathbf{U}$ is estimated by

$$E^L[\mathbf{U}^n_{\Delta x_L}] := \sum_{\ell=0}^{L} E_{M_\ell}[\mathbf{U}^n_{\Delta x_\ell} - \mathbf{U}^n_{\Delta x_{\ell-1}}]. \tag{26}$$

As MLMC-FDM/FVM is *non-intrusive*, any standard FDM/FVM codes can be used in step 2. Furthermore, MLMC-FDM/FVM is amenable to *efficient parallelization* [11, 15] as data from different grid resolutions and samples only interacts in step 3.

Theorem 5. *Assume the hypothesis of Theorem 2 is satisfied with $k \geq 2$, i.e. second moments of the random initial data $\mathbf{U}_0$, source $\mathbf{S}$ and K exist. Under Assumption 1, the MLMC-FDM/FVM estimate (26) satisfies the following error bound,*

$$\left\|\mathbb{E}[\mathbf{U}(t^n)] - E^L[\mathbf{U}^n_{\Delta x_L}](\omega)\right\|_{L^2(\Omega,\mathbf{L}^2(\mathbf{D}))} \leq C_1 \Delta x_L^s + C_2 \sum_{\ell=1}^{L} M_\ell^{-\frac{1}{2}} \Delta x_\ell^s + C_3 M_0^{-\frac{1}{2}}, \tag{27}$$

$$\begin{aligned} C_1 &= C\left(\|\mathbf{U}_0, \mathbf{S}, t^n\|_{L^1(\Omega,\mathbf{H}^s(\mathbf{D}))} + t^n \|\mathbf{U}_0, \mathbf{S}, t^n\|_{L^1(\Omega,\mathbf{L}^2(\mathbf{D}))}\right), \\ C_2 &= C\left(\|\mathbf{U}_0, \mathbf{S}, t^n\|_{L^2(\Omega,\mathbf{H}^s(\mathbf{D}))} + t^n \|\mathbf{U}_0, \mathbf{S}, t^n\|_{L^2(\Omega,\mathbf{L}^2(\mathbf{D}))}\right), \\ C_3 &= \bar{K}_2 \|\mathbf{U}_0, \mathbf{S}, t^n\|_{L^2(\Omega,\mathbf{L}^2(\mathbf{D}))}. \end{aligned}$$

Proof. Using the triangle inequality, the left hand side of (27) is bounded by

$$\|\mathbb{E}[\mathbf{U}(t^n)] - \mathbb{E}[\mathbf{U}^n_{\Delta x_L}]\|_{L^2(\Omega,\mathbf{L}^2(\mathbf{D}))} + \|\mathbb{E}[\mathbf{U}^n_{\Delta x_L}] - E^L[\mathbf{U}^n_{\Delta x_L}](\omega)\|_{L^2(\Omega,\mathbf{L}^2(\mathbf{D}))} = \mathrm{I} + \mathrm{II}.$$

We estimate term I and II separately. By linearity of the expectation, term I equals

$$\mathrm{I} = \|\mathbb{E}[\mathbf{U}(\cdot, t^n, \omega) - \mathbf{U}^n_{\Delta x_L}(\cdot, \omega)]\|_{\mathbf{L}^2(\mathbf{D})} = \|\mathbf{U}(\cdot, t^n, \omega) - \mathbf{U}^n_{\Delta x_L}(\cdot, \omega)\|_{L^1(\Omega,\mathbf{L}^2(\mathbf{D}))},$$

which can be bounded by (19). Using MLMC definition (26) and, again, linearity of the expectation, and the MC bound (15), term II is bounded by

$$\begin{aligned} \mathrm{II} &= \left\| \sum_{\ell=0}^{L} \mathbb{E}[\mathbf{U}^n_{\Delta x_\ell}(\cdot,\omega) - \mathbf{U}^n_{\Delta x_{\ell-1}}(\cdot,\omega)] - E_{M_\ell}[\mathbf{U}^n_{\Delta x_\ell}(\cdot,\omega) - \mathbf{U}^n_{\Delta x_{\ell-1}}(\cdot,\omega)] \right\|_{L^2(\Omega,\mathbf{L}^2(\mathbf{D}))} \\ &\leq \sum_{\ell=0}^{L} \|\mathbb{E}[\mathbf{U}^n_{\Delta x_\ell}(\cdot,\omega) - \mathbf{U}^n_{\Delta x_{\ell-1}}(\cdot,\omega)] - E_{M_\ell}[\mathbf{U}^n_{\Delta x_\ell}(\cdot,\omega) - \mathbf{U}^n_{\Delta x_{\ell-1}}(\cdot,\omega)]\|_{L^2(\Omega,\mathbf{L}^2(\mathbf{D}))} \\ &\leq M_0^{-\frac{1}{2}} \|\mathbf{U}^n_{\Delta x_0}(\cdot,\omega)\|_{L^2(\Omega,\mathbf{L}^2(\mathbf{D}))} + \sum_{\ell=1}^{L} M_\ell^{-\frac{1}{2}} \|\mathbf{U}^n_{\Delta x_\ell}(\cdot,\omega) - \mathbf{U}^n_{\Delta x_{\ell-1}}(\cdot,\omega)\|_{L^2(\Omega,\mathbf{L}^2(\mathbf{D}))}. \end{aligned}$$

The first term is bounded by (17); the detail terms $\mathbf{U}^n_{\Delta x_\ell} - \mathbf{U}^n_{\Delta x_{\ell-1}}$ are bounded by

$$\|\mathbf{U}^n_{\Delta x_\ell} - \mathbf{U}^n_{\Delta x_{\ell-1}}\|_{L^2(\Omega,\mathbf{L}^2(\mathbf{D}))} \leq \|\mathbf{U} - \mathbf{U}^n_{\Delta x_\ell}\|_{L^2(\Omega,\mathbf{L}^2(\mathbf{D}))} + \|\mathbf{U} - \mathbf{U}^n_{\Delta x_{\ell-1}}\|_{L^2(\Omega,\mathbf{L}^2(\mathbf{D}))}.$$

Using (19), detail terms can be further bounded by

$$\|\mathbf{U}^n_{\Delta x_\ell} - \mathbf{U}^n_{\Delta x_{\ell-1}}\|_{L^2(\Omega,\mathbf{L}^2(\mathbf{D}))} \leq C\, \Delta x_\ell^s \left\| \|\mathbf{U}_0, \mathbf{S}, t^n\|_{\mathbf{H}^s(\mathbf{D})} + t^n \|\mathbf{U}_0, \mathbf{S}, t^n\|_{\mathbf{L}^2(\mathbf{D})} \right\|_{L^2(\Omega,\mathbb{R})}.$$

Using triangle inequality and summing over all levels $\ell > 0$, bound (27) follows. □

To equilibrate the statistical and the spatio-temporal errors in (27), we require

$$M_\ell = O(2^{2(L-\ell)s}), \quad 0 \leq \ell \leq L. \tag{28}$$

Notice that (28) implies that the largest number of MC samples is required on the coarsest mesh level $\ell = 0$, whereas only a few MC samples are needed for $\ell = L$. Next, we are interested in the largest $\alpha > 0$ and smallest $\beta > 0$, such that:

$$\|\mathbb{E}[\mathbf{U}(t^n)] - E^L[\mathbf{U}^n_{\Delta x_L}](\omega)\|_{L^2(\Omega,\mathbf{L}^2(\mathbb{R}))} \leq C\,(\mathbb{E}\,[\mathrm{Work}_L])^{-\alpha}\log(\mathbb{E}[\mathrm{Work}_L])^{\beta}. \quad (29)$$

Assuming that $\bar{\lambda} = \mathbb{E}[\lambda(\omega)]$ in (22) is finite and using (21) with (28),

$$\begin{aligned}\mathbb{E}\,[\mathrm{Work}_L] &= \mathbb{E}\left[\sum_{\ell=0}^{L} M_\ell \mathrm{Work}_{\Delta x_\ell}(\omega)\right] = \sum_{\ell=0}^{L} M_\ell \mathbb{E}\left[\mathrm{Work}_{\Delta x_\ell}(\omega)\right] \\ &= \sum_{\ell=0}^{L} M_\ell T B \bar{\lambda} \Delta x_\ell^{-(d+1)} = T B \bar{\lambda} \sum_{\ell=0}^{L} M_\ell \Delta x_\ell^{-(d+1)}.\end{aligned} \quad (30)$$

The last term in (30) was already estimated in [11]. Since the *expectation* of computational work is obtain from the *deterministic* computational work by scaling with a problem dependent constant $\bar{\lambda}$, the *asymptotic* error vs. *expected* (in the mean) computation work estimate (29) remains analogous to the estimates derived in [11],

$$(\alpha, \beta) = \begin{cases} \left(\min\left\{\frac{1}{2}, \frac{s}{d+1}\right\}, 1\right) & \text{if } s \neq (d+1)/2, \\ \left(\frac{1}{2}, \frac{3}{2}\right) & \text{if } s = (d+1)/2. \end{cases} \quad (31)$$

Finally, we would like to note that bounds (15), (24) and (27) can be easily generalized (all steps in proofs are analogous) for higher statistical moments, i.e. $k > 1$.

4 Acoustic Isotropic Wave Equation as a Linear Hyperbolic System

The *stochastic isotropic* linear acoustic wave equation, modeling the propagation of acoustic pressure p in a random medium, is given by

$$\begin{cases} p_{tt}(\mathbf{x}, t, \omega) - \nabla \cdot (c(\mathbf{x}, \omega) \nabla p(\mathbf{x}, t, \omega)) = f(\mathbf{x}, \omega), \\ p(\mathbf{x}, 0, \omega) = p_0(\mathbf{x}, \omega), \quad \mathbf{x} \in \mathbf{D},\ t > 0,\ \omega \in \Omega. \\ p_t(\mathbf{x}, 0, \omega) = p_1(\mathbf{x}, \omega), \end{cases} \quad (32)$$

In many cases, the initial data p_0, p_1, the coefficient c and the source f are *not* known exactly. We propose to model them as random fields $p_0, p_1 \in L^k(\Omega, \mathbf{W}^{r_0,\infty}(\mathbf{D}))$, $f \in L^k(\Omega, \mathbf{W}^{r_f,\infty}(\mathbf{D}))$ and $c \in L^0(\Omega, W^{r_c,\infty}(\mathbf{D}))$ with $\mathbb{P}[c(\mathbf{x}, \omega) > 0, \forall \mathbf{x} \in \mathbf{D}] = 1$. For implementation, we rewrite the stochastic linear

acoustic wave equation (32) as a *linear system* of $d+1$ *first order* conservation laws. One possibility (out of many) is the following,

$$\begin{cases} p_t(\mathbf{x},t,\omega) - \nabla\cdot(c(\mathbf{x},\omega)\mathbf{u}(\mathbf{x},t,\omega)) = tf(\mathbf{x},\omega), \\ \mathbf{u}_t(\mathbf{x},\omega) - \nabla p(\mathbf{x},\omega) = 0, \\ p(\mathbf{x},0,\omega) = p_0(\mathbf{x},\omega), \\ \mathbf{u}(\mathbf{x},0,\omega) = \mathbf{u}_0(\mathbf{x},\omega), \end{cases} \quad \mathbf{x}\in\mathbf{D},\ t>0,\ \omega\in\Omega, \tag{33}$$

To verify equivalence of (33) and (32), differentiate the first equation of (33) in time:

$$f = p_{tt} - \nabla\cdot(c(\mathbf{x},\omega)\mathbf{u}_t) = p_{tt} - \nabla\cdot(c(\mathbf{x},\omega)\nabla p).$$

The linear hyperbolic system (33) is a system of conservation laws (8) for $m = d+1$,

$$\mathbf{U} = \begin{bmatrix} p \\ \mathbf{u} \end{bmatrix}, \quad \mathbf{U}_0 = \begin{bmatrix} p_0 \\ \mathbf{u}_0 \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} tf \\ 0 \end{bmatrix}, \quad \mathbf{A}_r(\mathbf{x},\omega) \in \mathbb{R}^{(d+1)\times(d+1)}. \tag{34}$$

All elements of $\mathbf{A}_r$ are zero, except $(\mathbf{A}_r(\mathbf{x},\omega))_{1,r+1} = -c(\mathbf{x},\omega)$ and $(\mathbf{A}_r)_{r+1,1} = -1$. Note, that $\mathbf{A}_r$ defines a strongly hyperbolic linear system of conservation laws. This is easily verifiable for $d=1$; there exists an invertible $\mathbf{Q}_\mathbf{x}(\omega)$ diagonalizing $\mathbf{A}$:

$$\mathbf{Q}_\mathbf{x}(\omega) = \frac{1}{\sqrt{2}}\begin{bmatrix} \frac{1}{\sqrt{c}} & -\frac{1}{\sqrt{c}} \\ 1 & 1 \end{bmatrix} \quad\Longrightarrow\quad \mathbf{Q}_\mathbf{x}(\omega)\mathbf{A}_1(\mathbf{x},\omega)\mathbf{Q}_\mathbf{x}(\omega)^{-1} = \begin{bmatrix} -\sqrt{c} & 0 \\ 0 & \sqrt{c} \end{bmatrix}.$$

Since $\|\mathbf{Q}_\mathbf{x}(\omega)\|\|\mathbf{Q}_\mathbf{x}^{-1}(\omega)\| = \max\{c^{\frac{1}{2}}, c^{-\frac{1}{2}}\} \le c^{\frac{1}{2}} + c^{-\frac{1}{2}}$, the uniform boundedness $c, c^{-1} \in L^\infty(\Omega, L^\infty(\mathbf{D}))$ ensures $\bar{K}_\infty < \infty$. For $k < \infty$: $c, c^{-1} \in L^{k/2}(\Omega, L^\infty(\mathbf{D}))$ implies

$$\bar{K}_k^k = \mathbb{E}[K^k(\omega)] \le \|c(\cdot,\omega)\|^{\frac{k}{2}}_{L^{k/2}(\Omega,L^\infty(\mathbf{D}))} + \|c^{-1}(\cdot,\omega)\|^{\frac{k}{2}}_{L^{k/2}(\Omega,L^\infty(\mathbf{D}))} < \infty. \tag{35}$$

Since the non-zero eigenvalues of $\mathbf{A}_r \in \mathbb{R}^{m\times m}$ are $\pm\sqrt{c(\mathbf{x},\omega)}$, the expected maximum wave speed $\bar{\lambda}$ required in (25) and (30) is *finite* provided $c \in L^{1/2}(\Omega, L^\infty(\mathbf{D}))$,

$$\bar{\lambda} = \|c\|_{L^{1/2}(\Omega,L^\infty(\mathbf{D}))} < \infty. \tag{36}$$

Finally, hypothesis 2 of the Theorem 1 holds with $r_\mathbf{0} = r_0$, $r_\mathbf{S} = r_f$, $r_\mathbf{A} = r_c$.

5 Numerical Experiments for Acoustic Isotropic Wave Equation

All simulations reported below were performed on Cray XE6 in CSCS [14] with the recently developed massively parallel code ALSVID-UQ [1, 11, 15].

We assume that random wave speed c is given by its Karhunen-Loève expansion

$$\log c(\mathbf{x},\omega) = \log \bar{c}(\mathbf{x}) + \sum_{m=1}^{\infty} \sqrt{\lambda_m}\Psi_m(\mathbf{x})Y_m(\omega), \tag{37}$$

with eigenvalues $\{\lambda_m\}_{m=1}^{\infty} \in \ell^{\frac{1}{2}}(\mathbb{N})$, eigenfunctions $\Psi_m, \bar{c} \in L^2(\mathbf{D})$, $\|\Psi_m\|_{L^2(\mathbf{D})} = 1$, and *independent* random variables Y_m with zero mean and finite variance.

5.1 Smooth Wave with Uniformly Distributed Coefficient

For physical domain $\mathbf{D} = [0, 2]$, consider *deterministic*, smooth ($r_0 = \infty$) initial data

$$p_0(x,\omega) := \sin(\pi x), \quad p_1(x,\omega) \equiv 0, \tag{38}$$

and *random* coefficient $c(\mathbf{x},\omega)$ that is given in terms of its KL expansion (37) with identical, *uniformly* distributed $Y_m \sim \mathscr{U}[-1, 1]$. We choose eigenvalues $\lambda_m = m^{-2.5}$, eigenfunctions $\Psi_m(\mathbf{x}) = \sin(\pi m x)$ and the mean field $\bar{c}(\mathbf{x}) \equiv 0.1$. Then both c and c^{-1} are uniformly bounded in Ω: $c(\mathbf{x},\omega), c^{-1}(\mathbf{x},\omega) \in L^{\infty}(\Omega, L^{\infty}(\mathbf{D}))$. Hence (35) and (36) holds with any $k \in \mathbb{N}_0 \cup \{\infty\}$. For simulations, KL expansion is truncated up to the first 10 terms: $\lambda_m = 0$, $\forall m > 10$. Since $r_0 = \infty$, $r_c \geq 1$, by Theorem 1 the solution $\mathbb{P}$-a.s. has bounded weak derivatives of first order, i.e. $\mathbf{U}(\cdot,\cdot,\omega) \in \mathbf{W}^{\bar{r},\infty}(\mathbf{D})$ with $\bar{r} = 1$. First order accurate FVM scheme ($q = 1$, HLL Rusanov flux [9], FE time stepping) will be used, hence, in (20), $s = \min\{1, \max\{1/2, 1\}\} = 1$. Higher order schemes ($s > 1$) for case $d = 1$ are inefficient since $s/(d+1) > 1/2$ in (31). Results of the MLMC-FVM simulation at $t = 2.0$ are presented in Fig. 1.

Using MLMC-FVM approximation from Fig. 1 (computed on 12 levels of resolution with the finest resolution having 16,384 cells) as a reference solution $\mathbf{U}_{\text{ref}}$, we run MC-FVM and MLMC-FVM (with $\Delta x_0 = 1/4$) on a series of mesh resolutions from 32 cells up to 1,024 cells and monitor the convergence behavior. For $L^2(\Omega,\cdot)$ norms in (24) and (27), the $L^2(\Omega; L^2(\mathbf{D}))$-based relative error estimator from [10] was used. $K = 5$ delivered sufficiently small relative standard deviation σ_K.

In Fig. 2, we compare the MC-FVM scheme with $M = O(\Delta x^{-2s})$ and the MLMC-FVM scheme with $M_\ell = M_L 2^{2s(L-\ell)}$, where $M_L = 16$ is chosen as suggested in [10]. Dashed lines indicate convergence rate slopes proved in Theorems 4 and 5. Theoretical and numerically observed convergence rates coincide, confirming

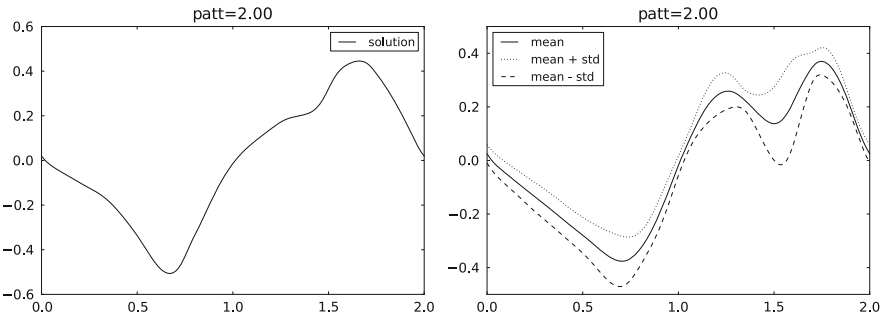


Fig. 1 One sample (*left*) and mean and variance (*right*) acoustic pressure $p(x, \omega)$ as in (33).

the robustness of our implementation. MLMC method is observed to be three orders of magnitude faster than MC method. This numerical experiment clearly illustrates the superiority of the MLMC algorithm over the MC algorithm (for $q = 1, s = 1$).

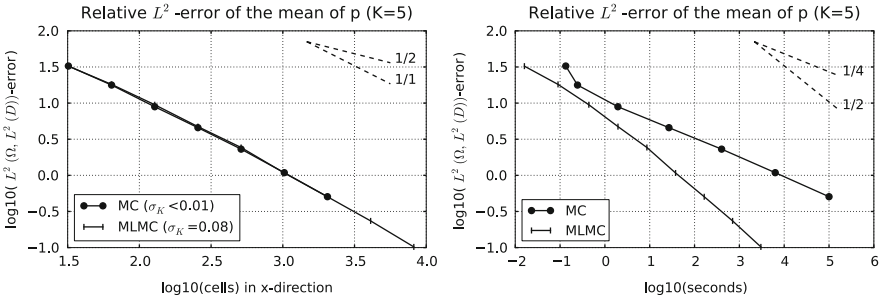


Fig. 2 Convergence of estimated mean for (38). Both MLMC and MC give similar errors for the same spatial resolution. However, MLMC method is 3 orders of magnitude faster than MC.

5.2 Discontinuous Wave with Normally Distributed Coefficient

For domain $\mathbf{D} = [0, 2]$, consider *deterministic*, discontinuous ($r_0 = 0$) initial data

$$p_0(x, \omega) := 2\chi_{(0.5,1.5)}(x) - 1.0, \quad p_1(x, \omega) \equiv 0, \tag{39}$$

and *stochastic* coefficient $c(\mathbf{x}, \omega)$ that is given by KL expansion (37) with identical, *normally* distributed $Y_m \sim \mathcal{N}[0, 1]$. We choose eigenvalues $\lambda_m = m^{-2.5}$, eigenfunctions $\Psi_m(\mathbf{x}) = \sin(\pi m x)$ and the mean field $\bar{c}(\mathbf{x}) \equiv 0.1$. Then, unlike in the *uniform* case before, $c, c^{-1} \notin L^\infty(\Omega, L^\infty(\mathbf{D}))$. However, (35) and (36) holds by

Proposition 1. *Assume* $\{\lambda_m\} \in \ell^{\frac{1}{2}}(\mathbb{N})$. *Then* $c, c^{-1} \in L^k(\Omega, L^\infty(\mathbf{D}))$, $\forall k \in \mathbb{N} \cup \{0\}$.

Proof. Using triangle inequality and $\|\Psi_m\|_{L^\infty(\mathbf{D})} = 1$, we obtain the following bound,

$$\frac{\|c(\cdot,\omega)\|_{L^\infty(\mathbf{D})}}{\|\bar{c}\|_{L^\infty(\mathbf{D})}} \leq \exp\left(\sum_{m=1}^{\infty} \sqrt{\lambda_m}|Y_m(\omega)|\right) =: \tilde{c}(\omega).$$

Next, we bound $\mathbb{E}[\tilde{c}(\omega)^k]$. Since Y_m are independent and normally distributed,

$$\mathbb{E}[\tilde{c}(\omega)^k] = \prod_{m=1}^{\infty} \mathbb{E}\left[\exp\left(k\sqrt{\lambda_m}|Y_m(\omega)|\right)\right] = \prod_{m=1}^{\infty} \exp\left(\frac{k^2\lambda_m}{2}\right)\left(1 + \operatorname{erf}\left(\frac{k\sqrt{\lambda_m}}{\sqrt{2}}\right)\right),$$

where the error function is defined by $\operatorname{erf}(a) = \dfrac{2}{\sqrt{(\pi)}} \displaystyle\int_0^a \exp(-t^2)dt$.

Using inequalities $\operatorname{erf}(a) \leq \frac{2}{\sqrt{\pi}}a$ and $1 + a \leq \exp(a)$, for any real $a \geq 0$,

$$\prod_{m=1}^{\infty}\left(1 + \operatorname{erf}\left(\frac{k\sqrt{\lambda_m}}{\sqrt{2}}\right)\right) \leq \prod_{m=1}^{\infty}\left(1 + \frac{2}{\sqrt{\pi}}\frac{k\sqrt{\lambda_m}}{\sqrt{2}}\right) \leq \exp\left(\sum_{m=1}^{\infty}\frac{2}{\sqrt{\pi}}\frac{k\sqrt{\lambda_m}}{\sqrt{2}}\right).$$

Hence, $\|c\|_{L^k(\Omega,L^\infty(\mathbf{D}))} \leq \|\bar{c}\|_{L^\infty(\mathbf{D})}\mathbb{E}[\tilde{c}(\omega)^k]^{\frac{1}{k}}$ is bounded further using

$$\mathbb{E}[\tilde{c}(\omega)^k]^{\frac{1}{k}} \leq \exp\left(\frac{k}{2}\|\{\lambda_m\}\|_{\ell^1(\mathbb{N})} + \sqrt{\frac{2}{\pi}}\left\|\left\{\sqrt{\lambda_m}\right\}\right\|_{\ell^1(\mathbb{N})}\right) < \infty.$$

Proof of $c^{-1} \in L^k(\Omega, L^\infty(\mathbf{D}))$ is analogous due to symmetry of Y_m. □

Since $r_0 = 0$, by Theorem 1, the solution $\mathbf{U}(\omega) \in \mathbf{W}^{\bar{r},\infty}(\mathbf{D})$ is $\mathbb{P}$-a.s. discontinuous ($\bar{r} = 0$). First order accurate ($q_1 = 1$, HLL Rusanov flux [9], FE time stepping) and second order accurate ($q_2 = 2$, HLL Rusanov flux, WENO reconstruction, SSP-RK2 time stepping [9]) FVM schemes will be used; hence, in (20), $s_1 = 1/2$ and $s_2 = 1$. For simulations, KL expansion is truncated up to first 10 terms: $\lambda_m = 0$, $\forall m > 10$. Results of the MLMC-FVM simulation at $t = 2.0$ are presented in Fig. 3.

MLMC-FVM approximation from Fig. 3 (computed on 12 levels of resolution with the finest resolution being on a mesh of 16,384 cells) is used as a reference solution $\mathbf{U}_{\text{ref}}$. Additionally to MC and MLMC schemes with $s = s_1$, we consider MC2 and MLMC2 schemes with $s = s_2$. In Fig. 4, we show convergence plots for *variance*; MLMC methods appear to be two orders of magnitude faster than MC methods.

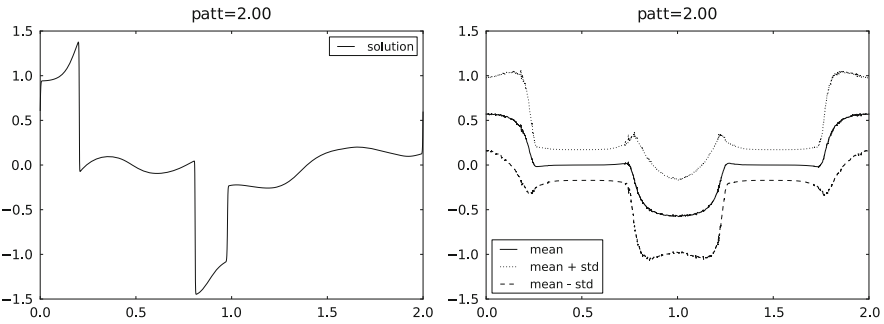


Fig. 3 One sample (*left*) and mean and variance (*right*) acoustic pressure $p(x, \omega)$ as in (33).

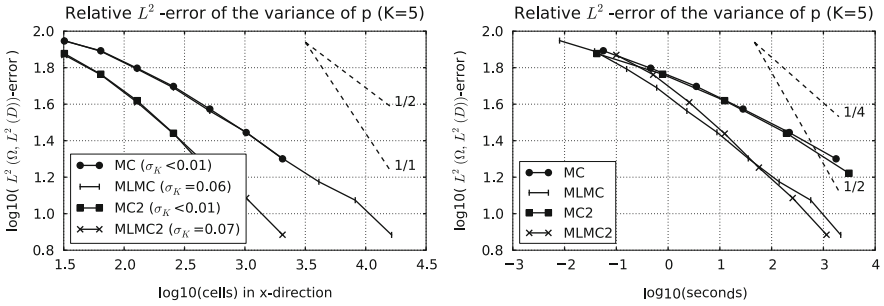


Fig. 4 Convergence of estimated variance for (39). Both MLMC(2) and MC(2) give similar errors for the same spatial resolution. However, MLMC methods are 2 orders of magnitude faster.

6 Conclusion

We formulate the proper notion of random solution for stochastic linear hyperbolic systems of conservation laws with random fluxes and show the resulting problem to be well-posed. We propose Multi-Level Monte Carlo FDM/FVM methods and prove their convergence. The complexity of the underlying FDM/FVM solver is *sample path dependent* due to the *random CFL-restricted time step size* in the *explicit* time-stepping scheme. To this end, a novel *probabilistic complexity analysis* of the error vs. *expected computational cost* is introduced. MLMC-FDM/FVM are proven to be much faster than MC methods and have the *same* accuracy vs. *expected* work ratio as *one deterministic solve*; they are also non-intrusive (existing FDM/FVM solvers can be used) and easily parallelizable [15]. As an example, we consider acoustic wave propagation in random medium with time independent statistical properties. The acoustic wave equation is rewritten as a linear hyperbolic system. In particular, the case of log-normal, isotropic Gaussian wave speed was considered, where the FVM with *explicit* time stepping incurs a *random CFL stability bound*. Numerical experiments are presented which are consistent with the findings from the theory.

Acknowledgements Jonas Šukys was supported in part by ETH CHIRP1-03 10-1. Siddhartha Mishra was supported by ERC StG No. 306279 SPARCCLE. Christoph Schwab was supported by ERC AdG No. 247277.

References

1. ALSVID-UQ. Version 2.0 (2013). Available from http://www.sam.math.ethz.ch/alsvid-uq
2. Barth, A., Schwab, Ch., Zollinger, N.: Multilevel MC method for elliptic PDEs with stochastic coefficients. Numer. Math. **119**, 123–161 (2011)
3. Cliffe, K.A., Giles, M.B., Scheichl, R., Teckentrup, A.L.: Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients. Comput. Vis. Sci. **14**, 3–15 (2011)
4. Giles, M.: Multilevel Monte Carlo path simulation. Oper. Res. **56**, 607–617 (2008)
5. Gottlieb, S., Shu, C.W., Tadmor, E.: High order time discretizations with strong stability property. SIAM Rev. **43**, 89–112 (2001)
6. Gustafsson, B., Kreiss, H.O., Oliger, J.: Time dependent problems and difference methods. Wiley, New York (1995)
7. Harten, A., Engquist, B., Osher S., Chakravarty, S.R.: Uniformly high order accurate essentially non-oscillatory schemes. J. Comput. Phys. **7**, 231–303, (1987)
8. Heinrich, S.: Multilevel Monte Carlo methods. In: Margenov, S., Waśniewski, J., Yalamov, P. (eds.) Large-Scale Scientific Computing, Sozopol, 2001. Lecture Notes in Computer Science, vol. 2170, pp. 58–67. Springer, Berlin/London (2001)
9. LeVeque, R.: Numerical Solution of Hyperbolic Conservation Laws. Cambridge University Press, Cambridge (2002)
10. Mishra, S., Schwab, Ch.: Sparse tensor multi-level Monte Carlo finite volume methods for hyperbolic conservation laws with random initial data. Math. Comp. **280**, 1979–2018 (2012)
11. Mishra, S., Schwab, Ch., Šukys, J.: Multi-level Monte Carlo finite volume methods for nonlinear systems of conservation laws. J. Comput. Phys. **231**, 3365–3388 (2012)
12. Mishra, S., Schwab, Ch., Šukys, J.: Multilevel Monte Carlo finite volume methods for shallow water equations with uncertain topography. SIAM J. Sci. Comput. **34**, B761–B784 (2012)
13. Müller, F., Jenny, P., Meyer, D.W.: Multilevel Monte Carlo for two phase flow and Buckley-Leverett transport in random heterogeneous porous media. J. Comput. Phys. **250**, 685–702 (2013)
14. Rosa (Cray XE6), Swiss National Supercomputing Center (CSCS), Lugano (2013), www.cscs.ch
15. Šukys, J., Mishra, S., Schwab, Ch.: Static load balancing for multi-level Monte Carlo finite volume solvers. In: PPAM 2011, Torun, Part I. Lecture Notes in Computer Science, vol. 7203, pp. 245–254. Springer, Heidelberg (2012)
16. Wloka, J.: Partial Differential Equations. Cambridge University Press, Cambridge/New York (1987)

Conference Participants

Nico Achtsis
Computer Science, KU Leuven, Belgium
e-mail: nico.achtsis@cs.kuleuven.be

Chengshi Ai
MSI, ANU, Australia
e-mail: chengshi.ai@anu.edu.au

Christoph Aistleitner
TU Graz, Austria
e-mail: aistleitner@math.tugraz.at

Abdulrahman Sulaiman Alangari
Statistics and Operation Research, King Saud University, Saudi Arabia
e-mail: asalangari@gmail.com

Adil Alharthi
IT and Computer Science, RMIT, AUSTRALIA
e-mail: s3125087@student.rmit.edu.au

Hamza Alkhatib
Geodetic Institute Hanover, Leibniz University Hanover, Germany
e-mail: alkhatib@gih.uni-hannover.de

James Allison
Statistics, Nort-West University, Hoffmanstreet 1, Potchefstroom, South Africa
e-mail: james.allison@nwu.ac.za

Abdulmohsen Almalawi
Distributed Systems & Networking, RMIT, Australia
e-mail: S3125089@student.rmit.edu.au

J. Dick et al. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2012*,
Springer Proceedings in Mathematics & Statistics 65,
DOI 10.1007/978-3-642-41095-6, © Springer-Verlag Berlin Heidelberg 2013

Abdullah A. Alshiha
Observatory on Higher Education, Ministry of Higher Education, Riyadh, Saudi Arabia, and Statistics & O.R., King Saud University, College of Science, P.O.Box 2455, Riyadh, 11451, Saudi Arabia
e-mail: abdullahalshiha@yahoo.com

Martin Altmayer
FB Mathematik, TU Kaiserslautern, Germany
e-mail: altmayer@mathematik.uni-kl.de

David Anderson
Mathematics, University of Wisconsin, USA
e-mail: anderson@math.wisc.edu

Andrea N. Arnold
Mathematics, Case Western Reserve University, USA
e-mail: ana33@case.edu

Yves Atchade
Statistics, University of Michigan, USA
e-mail: yvesa@umich.edu

Frank Aurzada
Institut für Mathematik, Technische Universität Berlin, Straße des 17. Juni 136, MA7-4, 10623 Berlin, Germany
e-mail: aurzada@math.tu-berlin.de

Jan Baldeaux
Quant Models & Development, Danske Bank, Denmark
e-mail: JanBaldeaux@gmail.com

Serge Barbeau
France
e-mail: serge.barbeau@thomsonreuters.com

Bjoern Baumeier
Max Planck Institute for Polymer Research, Mainz, Germany
e-mail: baumeier@mpip-mainz.mpg.de

Francisco Bernal
Mathematics, Lisbon Technical University, Portugal
e-mail: francisco.bernal@ist.utl.pt

Dmitriy Bilyk
Mathematics, University of South Carolina, USA
e-mail: bilyk.dmitriy@gmail.com

Benjamin Börschinger
Computing, Macquarie University, Sydney NSW 2109, Australia
e-mail: benjamin.borschinger@mq.edu.au

Zdravko Botev
School of Mathematics and Statistics, University of New South Wales, Sydney NSW 2052, Australia
e-mail: botev@unsw.edu.au

Richard J. Boys
School of Mathematics & Statistics, Newcastle University, UK
e-mail: richard.boys@ncl.ac.uk

Johann Brauchart
School of Mathematics and Statistics, University of New South Wales, Sydney NSW 2052, Australia
e-mail: j.brauchart@unsw.edu.au

Tim Brereton
Mathematics and Physics, University of Queensland, Brisbane QLD 4072, Australia
e-mail: tim.brereton@uqconnect.edu.au

Daniela Calvetti
Mathematics, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH, USA
e-mail: dxc57@case.edu

Christopher Carter
Economics, University of New South Wales, Australia
e-mail: chris.carter@unsw.edu.au

Jiating Chen
School of Software, Tsinghua University, China
e-mail: chenjt04@gmail.com

Nan Chen
Systems Engineering and Engineering Management, The Chinese University of Hong Kong, 709A William Mong Engineering Building, Hong Kong
e-mail: nchen@se.cuhk.edu.hk

William Chen
Mathematics, Macquarie University, Australia
e-mail: william.chen@mq.edu.au

Andrew Chernih
School of Mathematics and Statistics, University of New South Wales, Sydney NSW 2052, Australia
e-mail: andrew@andrewch.com

Boris Choy
Business Analytics, University of Sydney, Australia
e-mail: boris.choy@sydney.edu.au

Ronald Cools
Dept. of Computer Science, KU Leuven, Celestijnenlaan 200A, B-3001 Heverlee, Belgium
e-mail: ronald.cools@cs.kuleuven.be

Edward Cripps
Mathematics and Statistics, University of Western Australia, Australia
e-mail: edward.cripps@uwa.edu.au

Bernie Daigle, Jr.
Computer Science, University of California Santa Barbara, USA
e-mail: bdaigle@gmail.com

Fred Daum
Mathematics, Raytheon, 225 Presidential Way, Woburn, MA 01801, USA
e-mail: frederick_e_daum@raytheon.com

Pierre Del Moral
Institut de Mathematiques, Universite Bordeaux I, France
e-mail: Pierre.Del_Moral@inria.fr

Gary Delaney
CMIS, CSIRO, Australia
e-mail: gary.delaney@csiro.au

Josef Dick
School of Mathematics and Statistics, University of New South Wales, Sydney NSW 2052, Australia
e-mail: josef.dick@unsw.edu.au

Benjamin Doerr
Max Planck Institute for Computer Science, Campus E 1.4, 66123 Saarbrücken, Germany
e-mail: doerr@mpi-inf.mpg.de

Margaret Donald
Australia
e-mail: merricks.merricks@gmail.com

Jean-Luc Dortet-Bernadet
IRMA, Université de Strasbourg, 7 Rue René Descartes, 67084 Strasbourg, France
e-mail: jean-luc.dortet-bernadet@math.unistra.fr

Christopher Drovandi
Mathematical Sciences, Queensland University of Technology, 2 George St, Brisbane QLD 4000, Australia
e-mail: c.drovandi@qut.edu.au

Haining Fan
School of Software, Tsinghua University, China
e-mail: fhn@tsinghua.edu.cn

Yanan Fan
School of Mathematics and Statistics, University of New South Wales, Sydney NSW 2052, Australia
e-mail: y.fan@unsw.edu.au

Henri Faure
Institut de Mathématiques de Luminy, Aix-marseille, France
e-mail: faure@iml.univ-mrs.fr

Sarah Filippi
Department of Biological Sciences, Imperial College London, SW7 2AZ London, UK
e-mail: s.filippi@imperial.ac.uk

James M. Flegal
Department of Statistics, University of California, Riverside, 900 University Ave, Riverside, CA 92521, USA
e-mail: jflegal@ucr.edu

Colin Fox
Department of Physics, University of Otago, PO Box 56, Dunedin 9054, New Zealand
e-mail: fox@physics.otago.ac.nz

Mike Giles
Mathematical Institute, University of Oxford, Oxford OX2 6GG, UK
e-mail: mike.giles@maths.ox.ac.uk

Francois Giraud
Institut mathematique de Bordeaux, Bordeaux 1, 351 cours de la Liberation, 33405 Talence, CEDEX, France, and DSGA/SSPP/LOR, Centre CEA/CESTA, 15 Avenue des Sablieres, 33114 Le Barp, France
e-mail: fr.giraud@laposte.net

Michael Gnewuch
Department of Computer Science, Christian-Albrechts-Universität Kiel, Christian-Albrechts-Platz 4, 24098 Kiel, Germany
e-mail: mig@numerik.uni-kiel.de

Andrew Golightly
School of Mathematics & Statistics, Newcastle University, UK
e-mail: a.golightly@ncl.ac.uk

Domingo Gomez
Department of Mathematics and Statistics, University of Cantabria, Spain
e-mail: gomezd@unican.es

Venkiteswaran Gopalakrishnan
Mathematics, Birla Institute of Technology and Science, Vidya Vihar Campus, Pilani, India
e-mail: gvenki@bits-pilani.ac.in

Leonhard Grünschloß
Rendering Research, Weta Digital, New Zealand
e-mail: leonhard.gruenschloss@googlemail.com

Hiroshi Haramoto
Faculty of Education, Ehime University, Address 3, Bunkyo-cho, Matsuyama, Ehime, 790-8577, Japan
e-mail: haramoto@ehime-u.ac.jp

Shin Harase
Graduate School of Mathematical Sciences, The University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo, 153-8914, Japan
e-mail: harase@ms.u-tokyo.ac.jp

Keith Hayes
CSIRO, Australia
e-mail: keith.hayes@csiro.au

Stefan Heinrich
Department of Computer Science, TU Kaiserslautern, 67653 Kaiserslautern, Germany
e-mail: heinrich@informatik.uni-kl.de

Peter Hellekalek
Department of Mathematics, University of Salzburg, Hellbrunner Strasse 34, 5020 Salzburg, Austria
e-mail: peter.hellekalek@sbg.ac.at

Gillian Heller
Statistics, Macquarie University, Sydney NSW 2109, Australia
e-mail: gillian.heller@mq.edu.au

Radu Herbei
Statistics, The Ohio State University, USA
e-mail: herbei@stat.osu.edu

Samuel Herrmann
University of Burgundi, France
e-mail: Samuel.Herrmann@u-bourgogne.fr

Fred J. Hickernell
Department of Applied Mathematics, Illinois Institute of Technology, E1-208, 10 W. 32nd St., Chicago, IL 60616, USA
e-mail: hickernell@iit.edu

Aicke Hinrichs
Institut für Mathematik, Universität Rostock, Ulmenstraße 69, Haus 3, D-18051 Rostock, Germany
e-mail: aicke.hinrichs@uni-rostock.de

Roswitha Hofer
Institute of Financial Mathematics, Johannes Kepler University Linz, Altenbergerstrasse 69, 4040 Linz, Austria
e-mail: roswitha.hofer@jku.at

Geoff Hosack
Mathematics, Informatics and Statistics, CSIRO, Castray Esplanade, Hobart, TAS 7000, Australia
e-mail: geoff.hosack@csiro.au

Pierre Jacob
CEREMADE, Université Paris-Dauphine, France, and Statistics laboratory, CREST
e-mail: pierre.jacob@ensae.fr

Stephen Joe
Department of Mathematics, University of Waikato, Private Bag 3105, Hamilton 3240, New Zealand
e-mail: stephenj@waikato.ac.nz

Chaitanya Joshi
Statistics, University of Waikato, Private Bag 3105, Hamilton 3240, New Zealand
e-mail: cjoshi@waikato.ac.nz

Paul Kabaila
Department of Mathematics and Statistics, La Trobe University, Australia
e-mail: P.Kabaila@latrobe.edu.au

Lutz Kämmerer
Department of Mathematics, Chemnitz University of Technology, Germany
e-mail: kaemmerer@mathematik.tu-chemnitz.de

Roman Kapuscinski
Ross School of Business, University of Michigan, USA
e-mail: kapuscin@umich.edu

Jonathan M. Keith
School of Mathematical Sciences, Monash University, Clayton, VIC 3800, Australia
e-mail: jonathan.keith@monash.edu

Alexander Keller
NVIDIA, Fasanenstrasse 81, 10623 Berlin, Germany
e-mail: keller.alexander@gmail.com

Sunghwan Kim
Computing, Macquarie University, Australia
e-mail: sunghwan.kim@students.mq.edu.au

Jonas Knape
Environmental Sience, Policy and Management, University of California, Berkeley, USA
e-mail: jknape@berkeley.edu

Robert Kohn
ASB, UNSW, Australia
e-mail: r.kohn@unsw.edu.au

Mihaly Kovacs
Mathematics and Statistics, University of Otago, New Zealand
e-mail: mkovacs@maths.otago.ac.nz

Peter Kritzer
Institute of Financial Mathematics, Johannes Kepler University Linz, Altenbergerstrasse 69, 4040 Linz, Austria
e-mail: peter.kritzer@jku.at

Dirk Kroese
Mathematics, University of Queensland, Brisbane QLD 4072, Australia
e-mail: kroese@maths.uq.edu.au

Sergei Kucherenko
Centre for Process Systems Engineering, Imperial College London, Exhibition road, SW7 2AZ, UK
e-mail: s.kucherenko@broda.co.uk

Thomas Kühn
Mathematisches Institut, Universität Leipzig, Johannisgasse 26, 04103 Leipzig, Germany
e-mail: kuehn@mathematik.uni-leipzig.de

Frances Y. Kuo
School of Mathematics and Statistics, University of New South Wales, Sydney NSW 2052, Australia
e-mail: f.kuo@unsw.edu.au

Pierre L'Ecuyer
DIRO, University of Montreal, C.P.6128, succ. centre-ville,, Montreal, H3C 3J7, Canada
e-mail: lecuyer@iro.umontreal.ca

Michael Lacey
Mathematics, Georgia Institute of Technology, USA
e-mail: lacey@math.gatech.edu

Stig Larsson
Mathematical Sciences, Chalmers University of Technology, SE-41296 Gothenburg, Sweden
e-mail: stig@chalmers.se

Krzysztof Łatuszyński
Department of Statistics, Univeristy of Warwick, CV4 7AL, Coventry, UK
e-mail: k.g.latuszynski@warwick.ac.uk

Quoc Thong Le Gia
School of Mathematics and Statistics, University of New South Wales, Sydney NSW 2052, Australia
e-mail: qlegia@unsw.edu.au

Kate Lee
School of Computing & Mathematical Sciences, Auckland University of Technology, New Zealand
e-mail: jelee@aut.ac.nz

Sebastien Lemaire
Commissariat a l'Energie Atomique, CEA, France
e-mail: sebastien.lemaire@gmail.com

Gunther Leobacher
Institute of Financial Mathematics, Johannes Kepler University Linz, Altenbergerstrasse 69, 4040 Linz, Austria
e-mail: gunther.leobacher@jku.at

Paul Leopardi
Mathematical Sciences Insttitute, Australian National University, Building 27, ANU ACT 0200, Australia
e-mail: paul.leopardi@anu.edu.au

Hernan Eugenio Leövey
Mathematics, Humboldt Universität Berlin, Rudower Chausee 25, 12489, Berlin, Germany
e-mail: leovey@math.hu-berlin.de

Josef Leydold
Institute for Statistics and Mathematics, WU Vienna, Austria
e-mail: josef.leydold@wu.ac.at

Faming Liang
Department of Statistics, Texas A&M University, TAMU3143, USA
e-mail: fliang@stat.tamu.edu

Timothy Ling
Mathematical Sciences, University of Technology, Sydney, Australia
e-mail: Timothy.G.Ling@student.uts.edu.au

Fang Lou
Department of Mathematical Science, Tsinghua University, P.R. China
e-mail: louf@163.com

Maurizio Manuguerra
Statistics, Macquarie University, Australia
e-mail: maurizio.manuguerra@mq.edu.au

Lev Markhasin
Department of Mathematics, Friedrich-Schiller-University Jena, Ernst-Abbe-Platz 2, 07743 Jena, Germany
e-mail: lev.markhasin@uni-jena.de

Ian C. Marschner
Department of Statistics, Macquarie University, Building E4A, NSW, 2109, Australia
e-mail: ian.marschner@mq.edu.au

Kyle Matoba
Finance, University of California, Los Angeles, 110 Westwood Plaza, Ste. C-402, Los Angeles, CA 90095, USA
e-mail: kyle.matoba.2014@anderson.ucla.edu

Makoto Matsumoto
Department of Mathematics, Graduate School of Science, Hiroshima University, Japan
e-mail: m-mat@math.sci.hiroshima-u.ac.jp

Hermann G. Matthies
Institute of Scientific Computing, Technische Universitaet Braunschweig, Germany
e-mail: wire@tu-bs.de

Parviz Mehdi
Computing, Macquarie University, E6A, Room 347, Sydney NSW 2109, Australia
e-mail: mehdi.parviz@students.mq.edu.au

Kerrie Mengersen
Science, Queensland University of Technology, Australia
e-mail: k.mengersen@qut.edu.au

Blazej Miasojedow
Institute of Applied Mathematics, University of Warsaw, ul. Banacha 2, 02-097 Warszawa, Poland, and LTCI, TELECOM ParisTech, 46 rue Barrault, 75634 Paris Cedex 13, France
e-mail: bmia@mimuw.edu.pl

Kenichi Miura
Center for Grid Research and Development, National Institute of Informatics, Japan
e-mail: kenmiura@nii.ac.jp

David Munger
Informatique et recherche opérationnelle, University of Montreal, Canada
e-mail: david.munger@umontreal.ca

Lawrence Murray
Mathematics, Informatics and Statistics, CSIRO, Private Bag 5, Wembley WA 6913, Australia
e-mail: lawrence.murray@csiro.au

Jesper Møller
Dept. of Mathematical Sciences, Aalborg University, Denmark
e-mail: jm@math.aau.dk

Kenji Nagata
Graduate School of Frontier Science, The University of Tokyo, Japan
e-mail: nagata@mns.k.u-tokyo.ac.jp

Andreas Neuenkirch
Department of Mathematics, TU Kaiserslautern, Postfach 3049, 67653 Kaiserslautern, Germany
e-mail: neuenkirch@mathematik.uni-kl.de

Ido Nevat
CSIRO, Australia
e-mail: idonevat@gmail.com

Ken Newman
US Fish and Wildlife Service, 4001 N. Wilson Way, Stockton, CA 95204, USA
e-mail: ken_newman@fws.gov

James Nichols
School of Mathematics and Statistics, University of New South Wales, Sydney NSW 2052, Australia
e-mail: james.ashton.nichols@gmail.com

Harald Niederreiter
RICAM, Austrian Academy of Sciences, Altenbergerstrasse 69, Linz, Austria
e-mail: ghnied@gmail.com

Wojciech Niemiro
Faculty of Mathematics and Computer Science, Nicolaus Copernicus University, Torun, Gagarina 11, 87-100 Torun, Poland, and Mathematics, Informatics and Mechanics, University of Warsaw, Krakowskie Przedmiescie 26/28, 00-927 Warszawa, Poland
e-mail: wniemiro@gmail.com

Erich Novak
Department of Mathematics, Friedrich-Schiller-University Jena, Ernst-Abbe-Platz 2, 07743 Jena, Germany
e-mail: erich.novak@uni-jena.de

Dirk Nuyens
Department of Computer Science, KU Leuven, Celestijnenlaan 200A - bus 2402, BE-3001 Heverlee, Belgium
e-mail: dirk.nuyens@cs.kuleuven.be

John Ormerod
School of Mathematics and Statistics, University of Sydney, Australia
e-mail: john.ormerod@sydney.edu.au

Alina Ostafe
Computing, Macquarie University, Sydney NSW 2109, Australia
e-mail: alina.ostafe@mq.edu.au

Art B. Owen
Statistics, Stanford University, Sequoia Hall, Stanford, CA 94305, USA
e-mail: owen@stanford.edu

Minho Park
Mathematical Sciece, University of Nottingham, UK
e-mail: min.park@nottingham.ac.uk

Gareth W. Peters
School of Mathematics and Statistics, University of New South Wales, Sydney NSW 2052, Australia, and Mathematical and Information Sciences, CSIRO, Macquarie Park, Sydney NSW 2113, Australia
e-mail: garethpeters@unsw.edu.au

Martin Peters
Math.Editorial IV, Springer-Verlag, Germany
e-mail: Martin.Peters@springer.com

Anthony Pettitt
Mathematics, Queensland University of Technology, Australia
e-mail: a.pettitt@qut.edu.au

Friedrich Pillichshammer
Institute of Financial Mathematics, Johannes Kepler University Linz, Altenbergerstrasse 69, 4040 Linz, Austria
e-mail: friedrich.pillichshammer@jku.at

Jill Pipher
Mathematics, Brown University, USA
e-mail: jill_pipher@brown.edu

Gottlieb Pirsic
Institute of Financial Mathematics, Johannes Kepler University Linz, Altenbergerstrasse 69, 4040 Linz, Austria
e-mail: gottlieb.pirsic@jku.at

Leszek Plaskota
Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw, ul. Banacha 2, 02-097, Warsaw, Poland
e-mail: leszekp@mimuw.edu.pl

Eckhard Platen
Mathematical Sciences, University of Technology, Sydney, Australia
e-mail: eckhard.platen@uts.edu.au

Koen Poppe
Department of Computer Science, KU Leuven, Celestijnenlaan 200a - bus 2402, 3001 Heverlee, Belgium
e-mail: koen.poppe@cs.kuleuven.be

James Propp
Department of Mathematical Sciences, University of Massachusetts - Lowell, 1 University Avenue, Lowell, MA 01854, USA, and Department of Mathematics, University of California - Berkeley, 970 Evans Hall, Berkeley, CA 94720-3840, USA
e-mail: JamesPropp@gmail.com

Antonija Pršlja
Statistics, University of Ljubljana, Kongresni trg 12, 1000, Ljubljana, Slovenia
e-mail: antonija.prslja@arctur.si

Paweł Przybyłowicz
Department of Applied Mathematics, AGH University of Science and Technology, Al. Mickiewicza 30, Krakow, 30-059, Poland
e-mail: Przybyl83@gmail.com

Klaus Ritter
Department of Mathematics, TU Kaiserslautern, Postfach 3049, 67653 Kaiserslautern, Germany
e-mail: ritter@mathematik.uni-kl.de

Dale Roberts
MSI & School of Finance and Applied Statistics, Australian National University, Australia
e-mail: dale.roberts@anu.edu.au

Gareth Roberts
Univeristy of Warwick, UK
e-mail: gareth.o.roberts@warwick.ac.uk

Leonardo Rojas-Nandayapa
School of Mathematics and Physics, University of Queensland, Australia
e-mail: l.rojasnandayapa@uq.edu.au

Werner Römisch
Department of Mathematics, Humboldt-University Berlin, Rudower Chaussee 25, 12489 Berlin, Germany
e-mail: romisch@math.hu-berlin.de

Andreas Rößler
Institut für Mathematik, Universität zu Lübeck, Wallstr. 40, D-23560 Lübeck, Germany
e-mail: roessler@mathematik.tu-darmstadt.de

Daniel Rudolf
Department of Mathematics, Friedrich-Schiller-University Jena, Ernst-Abbe-Platz 2, 07743 Jena, Germany
e-mail: daniel.rudolf@uni-jena.de

Donna Mary Salopek
School of Mathematics and Statistics, University of New South Wales, Sydney NSW 2052, Australia
e-mail: dm.salopek@unsw.edu.au

Christian Schäfer
Statistics, Université Paris Dauphine, France
e-mail: christian.schafer@ensae.fr

Scott C. Schmidler
Statistical Science, Duke University, Box 90251, Durham, NC 27708-0251, USA
e-mail: schmidler@stat.duke.edu

Volker Schmidt
Institute of Stochastics, Ulm University, Helmholtzstr. 18, D-89069 Ulm, Germany
e-mail: volker.schmidt@uni-ulm.de

Christoph Schwab
Seminar of Applied Math, Dept of Mathematics, ETH Zürich, ETH Zentrum, HG G57.1, CH 8092 Zürich, Switzerland
e-mail: schwab@math.ethz.ch

David Scollnik
Mathematics and Statistics, Calgary, Canada
e-mail: scollnik@ucalgary.ca

Jungsuk Shim
Statistics, University of Seoul, 601, Miraegwan, University of Seoul, Siripdae-gil, jeonnong-dong 90, Dongdaemun-gu, South Korea
e-mail: js-beloved@hanmail.net

Alla Shymanska
School of Computing and Mathematical Sciences, Auckland University of Technology, New Zealand
e-mail: alla.shymanska@aut.ac.nz

Vasile Sinescu
School of Mathematics and Statistics, University of New South Wales, Sydney NSW 2052, Australia
e-mail: v.sinescu@unsw.edu.au

Scott Sisson
School of Mathematics and Statistics, University of New South Wales, Sydney NSW 2052, Australia
e-mail: scott.sisson@unsw.edu.au

Ian H. Sloan
School of Mathematics and Statistics, University of New South Wales, Sydney NSW 2052, Australia
e-mail: i.sloan@unsw.edu.au

Georgy Sofronov
Department of Statistics, Macquarie University, Sydney NSW 2109, Australia
e-mail: georgy.sofronov@mq.edu.au

Erkki Somersalo
Department of Mathematics, Case Western Reserve University, USA
e-mail: ejs49@case.edu

Eric Song
School of Mathematics and Statistics, University of New South Wales, Sydney NSW 2052, Australia
e-mail: ericlson@gmail.com

Živa Stepančič
Statistics, University of Ljubljana, Kongresni trg 12, 1000 Ljubljana, Slovenia
e-mail: ziva.stepancic@arctur.si

Megan Stephens
MetService, New Zealand
e-mail: meghan.stephens@metservice.com

Osnat Stramer
Statistics & Actuarial Science, University of Iowa, USA
e-mail: osnat-stramer@uiowa.edu

Christopher Strickland
Mathematics, Queensland University of Technology, Australia
e-mail: christopher.strickland@qut.edu.au

Jonas Šukys
Seminar of Applied Math, Dept of Mathematics, ETH Zürich, ETH Zentrum, HG G57.1, CH 8092 Zürich, Switzerland
e-mail: jonas.sukys@sam.math.ethz.ch

Hidemaro Suwa
Department of Applied Physics, The University of Tokyo, Hongo 7-3-1, Bunkyo, Tokyo, Japan
e-mail: suwamaro@looper.t.u-tokyo.ac.jp

Lukasz Szpruch
Mathematical Institute, University of Oxford, UK
e-mail: szpruch@maths.ox.ac.uk

Tor Sørevik
Dept. of Mathematics, University of Bergen, Norway
e-mail: tor.sorevik@math.uib.no

Aretha L. Teckentrup
Mathematical Sciences, University of Bath, Claverton Down, Bath, BA2 7AY, UK
e-mail: alt24@bath.ac.uk

Alexandre H. Thiéry
Department of Statistics, University of Warwick, UK
e-mail: a.h.thiery@warwick.ac.uk

Tomáš Tichy
Department of Finance, Technical University Ostrava, Sokolska 33, Ostrava, 701 21, Czech Republic
e-mail: tomas.tichy@vsb.cz

Henghsiu Tsai
Institute of Statistical Science, Academia Sinica, Taiwan
e-mail: htsai@stat.sinica.edu.tw

Elisabeth Ullmann
Mathematical Sciences, University of Bath, UK
e-mail: E.Ullmann@bath.ac.uk

Mario Ullrich
Department of Mathematics, Friedrich-Schiller-University Jena, Ernst-Abbe-Platz 2, 07743 Jena, Germany
e-mail: mario.ullrich@uni-jena.de

Sebastian Vollmer
Mathematics, University of Warwick, UK
e-mail: sjvollmer@gmail.com

Yuguang Wang
School of Mathematics and Statistics, University of New South Wales, Sydney NSW 2052, Australia
e-mail: yuguang.wang@student.unsw.edu.au

Markus Weimar
Department of Mathematics, Friedrich-Schiller-University Jena, Ernst-Abbe-Platz 2, 07743 Jena, Germany
e-mail: markus.weimar@uni-jena.de

Heidi Weyhausen
Department of Mathematics, Friedrich-Schiller-University Jena, Ernst-Abbe-Platz 2, 07737 Jena, Germany
e-mail: heidi.weyhausen@gmail.com

Richard Wilson
School of Mathematics and Physics, The University of Queensland, Australia
e-mail: richard.wilson@uq.edu.au

Arne Winterhof
RICAM, Austrian Academy of Sciences, Altenbergerstrasse 69, 4040 Linz, Austria
e-mail: arne.winterhof@oeaw.ac.at

Carola Winzen
Department 1: Algorithms and Complexity, Max-Planck-Institut für Informatik, Campus E1 4, 66123 Saarbrücken, Germany
e-mail: winzen@mpi-inf.mpg.de

Robert Womersley
School of Mathematics and Statistics, University of New South Wales, Sydney NSW 2052, Australia
e-mail: R.Womersley@unsw.edu.au

Henryk Woźniakowski
Computer Science, Columbia University, 520 Amsterdam Ave., New York, NY 10027, USA, and Applied Mathematics, University of Warsaw, Banacha 2, 02-097 Warsaw, Poland
e-mail: henryk@cs.columbia.edu

Will Wright
MRQS, NAB, Australia
e-mail: will.wright@nab.com.au

Larisa Yaroslavtseva
Fakultät für Informatik und Mathematik, Universität Passau, Innstrasse 33, 94032 Passau, Germany
e-mail: larisa.yaroslavtseva@uni-passau.de

Qi Ye
Applied Mathematics, Illinois Institute of Technology, 10 West 32nd Street, Chicago, IL 60616, USA
e-mail: qye3@hawk.iit.edu

Index

J. Dick et al. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2012*,
Springer Proceedings in Mathematics & Statistics 65,
DOI 10.1007/978-3-642-41095-6, © Springer-Verlag Berlin Heidelberg 2013