Springer INdAM Series 29

Maurizio Falcone
Roberto Ferretti
Lars Grüne
William M. McEneaney *Editors*

Numerical Methods for Optimal Control Problems

Springer

Springer INdAM Series

Volume 29

More information about this series at http://www.springer.com/series/10283

Maurizio Falcone • Roberto Ferretti •
Lars Grüne • William M. McEneaney
Editors

Numerical Methods for Optimal Control Problems

Springer

Editors
Maurizio Falcone
Department of Mathematics
Sapienza University of Rome
Roma, Italy

Roberto Ferretti
Department of Mathematics & Physics
Roma Tre University
Rome, Italy

Lars Grüne
Mathematical Institut
Universität Bayreuth
Bayreuth
Bayern, Germany

William M. McEneaney
Department of Mechanical and Aerospace
Engineering
University of California, San Diego
La Jolla
CA, USA

ISSN 2281-518X ISSN 2281-5198 (electronic)
Springer INdAM Series
ISBN 978-3-030-01958-7 ISBN 978-3-030-01959-4 (eBook)
https://doi.org/10.1007/978-3-030-01959-4

Library of Congress Control Number: 2018966854

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The contributions included in this volume were presented at the workshop "Numerical methods for optimal control problems: algorithms, analysis and applications", held in Rome on 19–23 June 2017, at the Istituto Nazionale di Alta Matematica—INdAM (for more information on the workshop see the website http://www1.mat.uniroma1.it/ricerca/convegni/2017/numoc2017/). The goal of the workshop was to compare and (possibly) integrate a number of techniques of heterogeneous origin, such as open-loop optimization, nonlinear and dynamic programming, and model predictive control. Specific tools for high dimensional problems, such as state space reduction, adaptive and sparse grids, max-plus algebra and radial basis functions, were also addressed. The idea was to gather experts from different communities in numerical analysis and control engineering, and the varied nature of the topics covered in the volume reflect this goal.

Optimal control theory concerns the determination of control strategies for complex dynamical systems, in order to optimize measures of their performance. Nowadays, the field embraces a variety of areas ranging from process control to traffic flow optimization, renewable-resource exploitation and management of financial markets. In addition, since many of these systems are described by stochastic models, optimal control theory has also been adapted to stochastic dynamical systems, as well as to multi-agent systems. In this generality, the theory is also of increasing interest for large-scale economic planning.

Although a widely studied topic, the numerical computation of optimal strategies remains a problem of high complexity. Consequently, along with advances in the theory, the development of new numerical methods is a crucial issue for many industrial applications as well as for social and economic planning. With the rapid growth in computational power in recent years, problems of larger and larger scale are becoming computationally feasible in many application areas. However, depending on the techniques used, the bottleneck for numerical algorithms may be either, on the one hand, a lack of robustness, or, on the other, "curse of dimensionality"—the tendency for problem complexity to grow exponentially with respect to the number of state variables. This is particularly true for large-scale systems, in which the number of variables and agents is usually very high.

While optimal control techniques are expected to provide some of the theoretical foundations for new, much-needed technological developments, advances in real-life applications will require corresponding improvements in the efficiency and reliability of such numerical methods.

The field of numerical analysis for optimal control problems is experiencing rapid growth: new algorithms have been proposed in recent years, problems of increasing complexity have been approached, various state-of-the-art numerical recipes have been applied, and new ways of coupling the various techniques have been suggested. For all these reasons, we have decided to place the focus of this volume on the comparison of a broad range of numerical tools for this class of problems, and on their respective advantages, drawbacks and possible interactions.

We would like to take this opportunity to express our gratitude to INdAM for its financial support, for its continuous help with the organizational tasks, and for the warm hospitality during the workshop. We also want to thank all the speakers who contributed to the scientific program and this volume, and the institutions which sponsored the workshop: the University of Bayreuth, the Sapienza University of Rome, and Roma Tre University.

Rome, Italy — Maurizio Falcone
Rome, Italy — Roberto Ferretti
Bayreuth, Germany — Lars Grüne
La Jolla, CA, USA — William M. McEneaney
September 2018

Contents

About the Editors

Maurizio Falcone has been a Professor of Numerical Analysis at the University of Rome "La Sapienza" since 2001 and has held visiting positions at several institutions including ENSTA (Paris), the IMA (Minneapolis), Paris 6 and 7, the Russian Academy of Sciences (Moscow and Ekaterinburg), and UCLA. He serves as associate editor for the journal *Dynamic Games and Applications* and has authored one monograph and more than 80 papers in international journals. His research interests include numerical analysis, control theory, and differential games.

Roberto Ferretti has been an Associate Professor of Numerical Analysis at Roma Tre University since 2001 and has been an invited professor at UCLA (USA), Universitet Goroda Pereslavlya (Russia), ENSTA ParisTech and IRMA (France), the TU Munich (Germany), and the UP Madrid (Spain). He has authored one monograph and more than 40 papers in international journals/volumes, on topics ranging from semi-Lagrangian schemes to optimal control, level set methods, image processing, and computational fluid dynamics.

Lars Grüne is a Professor of Applied Mathematics at the University of Bayreuth, Germany. He obtained his Ph.D. from the University of Augsburg in 1996 and completed his postdoctoral degree (Habilitation) at Goethe University in Frankfurt/Main in 2001. He has held visiting positions at Sapienza University in Rome (Italy) and at the University of Newcastle (Australia) and is currently Editor-in-Chief of the journal *Mathematics of Control, Signals, and Systems*. His research interests lie in the areas of mathematical systems theory and optimal control.

William M. McEneaney received his B.S. and M.S. degrees in Mathematics from Rensselaer Polytechnic Institute, followed by M.S. and Ph.D. degrees in Applied Mathematics from Brown University. He has held academic positions at Carnegie Mellon and North Carolina State University, prior to his current appointment at the

University of California, San Diego. His non-academic positions have included the Jet Propulsion Laboratory and Air Force Office of Scientific Research. His interests include stochastic control and games, max-plus algebraic numerical methods, and the principle of stationary action.

A Hamilton-Jacobi-Bellman Approach for the Numerical Computation of Probabilistic State Constrained Reachable Sets

Mohamed Assellaou and Athena Picarelli

Abstract Aim of this work is to characterise and compute the set of initial conditions for a system of controlled diffusion processes which allow to reach a terminal target satisfying pointwise state constraints with a given probability of success. Defining a suitable auxiliary optimal control problem, the characterization of this set is related to the solution of a particular Hamilton-Jacobi-Bellman equation. A semi-Lagrangian numerical scheme is defined and its convergence to the unique viscosity solution of the equation is proved. The validity of the proposed approach is then tested on some numerical examples.

Keywords Viscosity solutions · Reachable set · Discontinuous cost functions · Neumann boundary conditions

1 Introduction

We consider the control of stochastic differential equations in $\mathbb{R}^d$ of the following form

$$\begin{cases} \mathrm{d}X(s) = b(s, X(s), u(s))\mathrm{d}s + \sigma(s, X(s), u(s))\mathrm{d}B(s), \quad \forall s \in [t, T] \\ X(t) = x \end{cases}. \tag{1}$$

Given a fixed time horizon $T > 0$, we aim to characterize the set of initial states from which, with an assigned level of probability, it is possible to reach a target set at time T satisfying some state constraints along the whole interval $[t, T]$.

M. Assellaou
ENSTA ParisTech, Palaiseau Cedex, France

A. Picarelli (✉)
Department of Economical Sciences, University of Verona, Verona, Italy
e-mail: picarelli@univr.it

M. Falcone et al. (eds.), *Numerical Methods for Optimal Control Problems*,
Springer INdAM Series 29, https://doi.org/10.1007/978-3-030-01959-4_1

More precisely, let $\mathscr{C}$ and $\mathscr{K}$ be two non-empty subsets of $\mathbb{R}^d$ representing respectively the target set and the set of state constraints and let $\rho \in [0, 1)$. We define the state constrained backward reachable set under probability of success ρ as the set, hereafter denoted by Ω_t^ρ, of initial points $x \in \mathbb{R}^d$ for which the probability to steer the system (1) towards $\mathscr{C}$ maintaining the dynamics in the set $\mathscr{K}$ is higher than ρ, i.e.

$$\left\{ x \in \mathbb{R}^d \,:\, \exists u \in \mathscr{U},\ \mathbb{P}\big[X_{t,x}^u(\theta) \in \mathscr{K},\ \ \forall \theta \in [t, T] \ \text{ and } \ X_{t,x}^u(T) \in \mathscr{C}\big] > \rho \right\},$$

where $X_{t,x}^u(\cdot)$ represents the strong solution to (1) associated with the control $u \in \mathscr{U}$. Assumptions on the coefficients in (1) and on the set of controls $\mathscr{U}$ will be made clear in the next section. Such backward reachable sets play an important role in many applications, as the set Ω_t^ρ can be interpreted as a "safety region" for reaching $\mathscr{C}$ remaining in the set $\mathscr{K}$, with confidence ρ. It turns out that the set Ω_t^ρ can be characterized by means of the so-called *level set approach*. At the basis of this approach there is the idea to look at the set of interest, the set Ω_t^ρ in our case, as the level set of a certain function solution of a suitable partial differential equation (PDE). Such a characterization of the set is particularly useful in view of its numerical approximation, since it opens the way to the use of a wide choice of numerical methods designed for PDEs. Originally introduced in [25] to model front propagation problems, this approach immediately resulted in a very powerful method for studying backward reachable sets of continuous non-linear dynamical systems under very general conditions. In [16, 24] this idea is used to describe the reachable sets for deterministic problems. The link between stochastic target problems and level set approach is established in [26]. More recently, the level set approach has been extended to the case of state-constrained controlled systems [9, 10] and probabilistic reachability problems [6].

In our case, we will show that it is straightforward to see that

$$\Omega_t^\rho = \left\{ x \in \mathbb{R}^d \,:\, \vartheta(t, x) > \rho \right\}, \tag{2}$$

where ϑ is the value function associated to the following optimal control problem:

$$\vartheta(t, x) := \sup_{u \in \mathscr{U}} \mathbb{E}\left[\mathbb{1}_{\mathscr{C}}(X_{t,x}^u(T)) \bigwedge \min_{\theta \in [t,T]} \mathbb{1}_{\mathscr{K}}(X_{t,x}^u(\theta)) \right], \tag{3}$$

with the standard notation $a \wedge b := \min(a, b)$. In particular, equality (2) characterises the set Ω_t^ρ for $t \in [0, T]$ by means of the function ϑ.

We point out that, in the discrete time setting, a similar approach has been considered in [1, 2, 22]. In this case, the value function is obtained recursively by solving the dynamic programming principle. In the present paper, we are interested in the approximation of the probabilistic backward reachable sets for time-continuous stochastic processes by PDE techniques. In the non controlled framework, an alter-

native numerical algorithm consist in using Monte Carlo simulations to generate a set of trajectories starting from a given initial position. The percentage of trajectories reaching the target without violation of the state constraints gives an approximation of the probability of success when starting from this position. On the other hand, for linear stochastic systems, a bound for the probability of hitting a target can be obtained by using the enclosing hulls of the probability density function for time intervals, see [3, 4]. However, it is worth noticing that these approaches are used to calculate the probabilities of success but do not allow to define the entire set of points that have the same given probability. In addition, Monte-Carlo based methods often require a large number of simulations to obtain a good accuracy. We will use such simulations in Sect. 5 as a comparison to validate our approach. In the context of financial mathematics, the problem of characterizing the backward reachable set with a given probability was first introduced by Föllmer and Leukert [18]. This problem was also studied and converted into the class of stochastic target problems by Touzi, Bouchard and Elie in [12]. However in these references the possible presence of state constraints is not taken into account.

In order to apply a dynamic programming approach and characterize the value function ϑ as the unique viscosity solution of a Hamilton-Jacobi-Bellman (HJB) equation we face two main difficulties. First, the discontinuous cost functional given by the presence of the indicator functions would require to make use of the notion of discontinuous viscosity solutions. Establish uniqueness results in such a framework is usually a very hard task, so we propose here to work on a regularized version of problem (3). Second, the non commutativity between expectation and maximum operator makes problem (3) not satisfying the natural "Markovian structure" necessary to apply the dynamic programming arguments. We here follow the ideas in [8, 10, 19, 21] and define an auxiliary optimal control problem in an augmented state space and derive the HJB equation for this problem recovering the value function ϑ solution of the original problem at a later stage. The obtained HJB equation is defined in a domain and completed with mixed Dirichlet and oblique derivative boundary conditions. Derivative conditions (to be considered in the viscosity sense, see Definition 1) typically arise dealing with running maxima in the cost functional (see also [8, 10]), while the Dirichlet condition will be naturally satisfied pointwise by our value function. We discuss the numerical approximation of the obtained HJB equation. We introduce a semi-Lagrangian (SL) approximation scheme which incorporates the aforementioned boundary conditions and we prove its convergence to the viscosity solution following the framework in [7]. We recall that SL scheme for second order HJB equations have been introduced by Menaldi in [23] and then studied by Camilli and Falcone [13]. We refer to [15] and the references therein for an overview. Derivative boundary conditions have been added to the scheme in [10], while the case of mixed Dirichlet-derivative conditions has been recently studied in [21].

The paper is organised as follows. In Sect. 2 we present the problem and give some preliminary results. The regularized problem is introduced in Sect. 2.2. Section 3 is devoted to the development of the dynamic programming arguments and the HJB characterization. In Sect. 4 we discuss the numerical aspects and state the main convergence result. Numerical tests are presented in Sect. 5.

2 Formulation of the Problem and Preliminary Results

2.1 Problem Formulation

Let $\{\Omega, \mathscr{F}_t, \{\mathscr{F}_t\}_{t\geq 0}, \mathbb{P}\}$ be a filtered probability space and $B(\cdot)$ a given p-dimensional Brownian motion. Let $T > 0$. We denote by $\mathscr{U}$ the set of all progressively measurable processes valued in $U \subset \mathbb{R}^m$, U compact set. For any $u \in \mathscr{U}$, let us consider the following system of stochastic differential equations (SDEs) in $\mathbb{R}^d$:

$$\begin{cases} \mathrm{d}X(s) = b(s, X(s), u(s))\mathrm{d}s + \sigma(s, X(s), u(s))\mathrm{d}B(s), \quad \forall s \in [t, T] \\ X(t) = x. \end{cases} \tag{4}$$

The following classical assumptions will be considered on the coefficients b and σ:

(H1) $\sigma : [0, T] \times \mathbb{R}^d \times U \to \mathbb{R}^{d\times p}$ and $b : [0, T] \times \mathbb{R}^d \times U \to \mathbb{R}^d$ are continuous functions and there exists constant $L > 0$ such that

$$|b(t, x, u) - b(t, y, u)| + |\sigma(t, x, u) - \sigma(t, y, u)| <= L|x - y|.$$

for any $t \in [0, T]$, $x, y \in \mathbb{R}^d$ and $u \in U$.

It is well known that, under assumption (H1), for any $u \in \mathscr{U}$ there is a unique strong solution to (4) [27, p. 42, Thm. 6.3]. We denote by $X^u_{t,x}(\cdot)$ such a solution.

Let $\mathscr{C}$ and $\mathscr{K}$ be nonempty open sets in $\mathbb{R}^d$, representing respectively the target set and the set of state constraints. Let $\rho \in [0, 1)$ an assigned value of success probability. We define the backward reachable set under probability of success ρ, as the set Ω^ρ_t of initial points $x \in \mathbb{R}^d$ from which it starts a trajectory $X^u_{t,x}(\cdot)$ such that the probability to reach the target $\mathscr{C}$ at the final instant T satisfying the constraint $\mathscr{K}$ in the interval $[t, T]$ is grater than ρ, i.e.:

$$\Omega^\rho_t := \left\{ x \in \mathbb{R}^d \,:\, \exists u \in \mathscr{U},\ \mathbb{P}\big[X^u_{t,x}(\theta) \in \mathscr{K},\ \ \forall \theta \in [t, T] \ \text{ and } \ X^u_{t,x}(T) \in \mathscr{C}\big] > \rho \right\}.$$

For a given set $\mathscr{O} \subseteq \mathbb{R}^d$ we will denote by $\mathbb{1}_{\mathscr{O}}$ its indicator function, i.e.

$$\mathbb{1}_{\mathscr{O}}(x) := \begin{cases} 1 & \text{if } x \in \mathscr{O} \\ 0 & \text{otherwise.} \end{cases}$$

One can easily verify that

$$\mathbb{1}_{\mathscr{C}}(X^u_{t,x}(T)) \bigwedge \min_{\theta\in[t,T]} \mathbb{1}_{\mathscr{K}}(X^u_{t,x}(\theta)) = \begin{cases} 1 & \text{if } X^u_{t,x}(\theta) \in \mathscr{K} \forall \theta \in [t,T] \text{ and } X^u_{t,x}(T) \in \mathscr{C} \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

i.e. the expression on the left hand side of (5) is an indicator function for the event

$$X^u_{t,x}(\theta) \in \mathscr{K} \forall \theta \in [t,T] \text{ and } X^u_{t,x}(T) \in \mathscr{C}.$$

It follows that, for any $u \in \mathscr{U}$, $\mathbb{P}\left[X^u_{t,x}(\theta) \in \mathscr{K}, \ \forall \theta \in [t,T] \text{ and } X^u_{t,x}(T) \in \mathscr{C}\right]$ can be expressed by

$$\mathbb{E}\left[\mathbb{1}_{\mathscr{C}}(X^u_{t,x}(T)) \bigwedge \min_{\theta\in[t,T]} \mathbb{1}_{\mathscr{K}}(X^u_{t,x}(\theta))\right].$$

As a consequence, it is possible to describe the set Ω^ρ_t using optimal control tools just looking at the evolution of the level sets of the following value function:

$$\vartheta(t,x) := \sup_{u\in\mathscr{U}} \mathbb{E}\left[\mathbb{1}_{\mathscr{C}}(X^u_{t,x}(T)) \bigwedge \min_{\theta\in[t,T]} \mathbb{1}_{\mathscr{K}}(X^u_{t,x}(\theta))\right]. \tag{6}$$

Proposition 1 *Let assumption (H1) be satisfied. Then, for $t \in [0,T]$, we have:*

$$\Omega^\rho_t = \{x \in \mathbb{R}^d : \ \vartheta(t,x) > \rho\}.$$

Proof If $x \in \Omega^\rho_t$, thanks to equality (5)

$$\mathbb{E}\left[\mathbb{1}_{\mathscr{C}}(X^u_{t,x}(T)) \bigwedge \min_{\theta\in[t,T]} \mathbb{1}_{\mathscr{K}}(X^u_{t,x}(\theta))\right] > \rho$$

for some control $u \in \mathscr{U}$ and it follows $\vartheta(t,x) > \rho$.

Let us now suppose that $\vartheta(t,x) > \rho$. By the definition of the supremum and the fact that $\mathscr{U}$ is a non empty set, one has that, for some control $\bar{u} \in \mathscr{U}$,

$$\mathbb{E}\left[\mathbb{1}_{\mathscr{C}}(X^{\bar{u}}_{t,x}(T)) \bigwedge \min_{\theta\in[t,T]} \mathbb{1}_{\mathscr{K}}(X^{\bar{u}}_{t,x}(\theta))\right] > \rho$$

and then, using again (5), $x \in \Omega^\rho_t$.

Motivated by this result, we are going to focus on the characterization and numerical approximation of the function ϑ. Problem (6) is an optimal control problem with a discontinuous cost in a "minimum form". This is not a standard formulation in optimal control theory for two main reasons: first, the discontinuity of the cost functional prevents the characterization of (6) as the unique viscosity

solution of a HJB equation, second the loss of Markovian structure in the cost, due to the presence of the minimum operator inside the expectation, makes the dynamic programming arguments not directly applicable. We discuss the first issue in the next section.

2.2 Regularized Problem

The discontinuity introduced by the presence of the indicator functions and the consequent necessity of dealing with the notion of discontinuous viscosity solutions (see for instance [17, Section VII.4] for their definition) pose nontrivial issues when attempting to establish uniqueness results for the associated HJB equation. To overcome this difficulty, from now on we will work with a regularized version of the cost functional in (6). In particular, observing that the indicator functions $\mathbb{1}_{\mathscr{C}}$ and $\mathbb{1}_{\mathscr{K}}$ can be written as

$$\mathbb{1}_{\mathscr{C}}(z) = \begin{cases} 1 & \text{if } d_{\partial\mathscr{C}}(z) < 0 \\ 0 & \text{if } d_{\partial\mathscr{C}}(z) \geq 0 \end{cases}, \qquad \mathbb{1}_{\mathscr{K}}(z) = \begin{cases} 1 & \text{if } d_{\partial\mathscr{K}}(z) < 0 \\ 0 & \text{if } d_{\partial\mathscr{K}}(z) \geq 0 \end{cases}$$

where $d_{\partial\mathscr{C}}$ and $d_{\partial\mathscr{K}}$ are respectively the signed distance function to $\partial\mathscr{C}$ and $\partial\mathscr{K}$, we consider the following regularized functions $\phi^{\epsilon}_{\mathscr{C}}$ and $\phi^{\epsilon}_{\mathscr{K}}$ (see Fig. 1):

$$\phi^{\epsilon}_{\mathscr{C}}(x) := \min(1, \max(0, -\frac{1}{\epsilon} d_{\partial\mathscr{C}}(x))), \quad \phi^{\epsilon}_{\mathscr{K}}(x) := \min(1, \max(0, -\frac{1}{\epsilon} d_{\partial\mathscr{K}}(X)))$$

and the optimal control problem

$$\vartheta^{\epsilon}(t,x) := \sup_{u\in\mathscr{U}} \mathbb{E}\left[\phi^{\epsilon}_{\mathscr{C}}(X^{u}_{t,x}(T)) \bigwedge \min_{\theta\in[t,T]} \phi^{\epsilon}_{\mathscr{K}}(X^{u}_{t,x}(\theta))\right]. \tag{7}$$

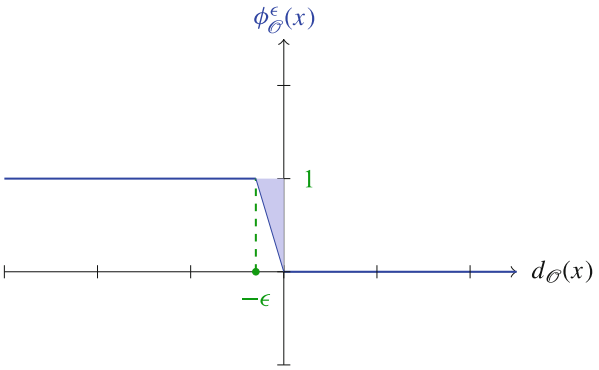


Fig. 1 Regularization of the indicator function in the case $\mathscr{O} = (-\infty, 0)$

Remark 1 Note that the choice of $\phi^\epsilon_\cdot$ is such that $\phi^\epsilon_\cdot \leq \mathbb{1}_\cdot$ which implies

$$\vartheta^\epsilon(t,x) > \rho \Rightarrow \vartheta(t,x) > \rho.$$

Hence if we are able to find a numerical approximation $\widetilde{\vartheta}^\epsilon$ of ϑ^ϵ such that $|\widetilde{\vartheta}^\epsilon - \vartheta^\epsilon| \leq \eta$ for some $\eta \geq 0$, we will have $\widetilde{\vartheta}^\epsilon(t,x) > \rho + \eta \Rightarrow \vartheta^\epsilon(t,x) > \rho \Rightarrow \vartheta(t,x) > \rho$.

This regularization allow us to deal with a continuous cost and also to obtain the following regularity result for the associated value function:

Proposition 2 *Let assumption (H1) be satisfied and let $\epsilon > 0$. The value function ϑ^ϵ is Lipschitz continuous with respect to x and $\frac{1}{2}$-Hölder continuous with respect to t, i.e. there exists a constant $L_\epsilon > 0$ such that*

$$|\vartheta^\epsilon(t,x) - \vartheta^\epsilon(s,y)| \leq L_\epsilon\big(|x-y| + |t-s|^{\frac{1}{2}}(1+|x|+|y|)\big)$$

for any $t, s \in [0,T]$, $x, y \in \mathbb{R}^d$.

Proof Let $0 \leq t \leq s \leq T$, $x, y \in \mathbb{R}^d$. Thanks to the property of minimum operator $|(a \wedge b) - (c \wedge d)| \leq |a-c| \vee |b-d|$, one has:

$$\begin{aligned}
&|\vartheta^\epsilon(t,x) - \vartheta^\epsilon(t,y)| \qquad (8)\\
&\leq \sup_{u\in\mathscr{U}} \mathbb{E}\Big[\Big|\big(\phi^\epsilon_{\mathscr{C}}(X^u_{t,x}(T)) \bigwedge \min_{\theta\in[t,T]} \phi^\epsilon_{\mathscr{K}}(X^u_{t,x}(\theta))\big) - \big(\phi^\epsilon_{\mathscr{C}}(X^u_{t,y}(T)) \bigwedge \min_{\theta\in[t,T]} \phi^\epsilon_{\mathscr{K}}(X^u_{t,y}(\theta))\big)\Big|\Big]\\
&\leq \sup_{u\in\mathscr{U}} \mathbb{E}\Big[\big|\phi^\epsilon_{\mathscr{C}}(X^u_{t,x}(T)) - \phi^\epsilon_{\mathscr{C}}(X^u_{t,y}(T))\big| \bigvee \max_{\theta\in[t,T]} \big|\phi^\epsilon_{\mathscr{K}}(X^u_{t,x}(\theta)) - \phi^\epsilon_{\mathscr{K}}(X^u_{t,y}(\theta))\big|\Big]
\end{aligned}$$

and

$$\begin{aligned}
&|\vartheta^\epsilon(t,x) - \vartheta^\epsilon(s,x)| \qquad (9)\\
&\leq \sup_{u\in\mathscr{U}} \mathbb{E}\Big[\big|\phi^\epsilon_{\mathscr{C}}(X^u_{s,X^u_{t,x}(s)}(T)) - \phi^\epsilon_{\mathscr{C}}(X^u_{s,x}(T))\big| \bigvee \max_{\theta\in[t,s]} \big|\phi^\epsilon_{\mathscr{K}}(X^u_{t,x}(\theta)) - \phi^\epsilon_{\mathscr{K}}(X)\big|\\
&\quad \bigvee \max_{\theta\in[s,T]} \big|\phi^\epsilon_{\mathscr{K}}(X^u_{s,X^u_{t,x}(s)}(\theta)) - \phi^\epsilon_{\mathscr{K}}(X^u_{s,x}(\theta))\big|\Big].
\end{aligned}$$

It can be easily verified that $\phi^\epsilon_{\mathscr{C}}$ and $\phi^\epsilon_{\mathscr{K}}$ are Lipschitz continuous functions with Lipschitz constant $1/\epsilon$. Then by (8) and (9) we get

$$|\vartheta^\epsilon(t,x) - \vartheta^\epsilon(t,y)| \leq \frac{1}{\epsilon} \sup_{u\in\mathscr{U}} \mathbb{E}\Big[\max_{\theta\in[t,T]} \big|X^u_{t,x}(\theta) - X^u_{t,y}(\theta)\big|\Big]$$

and

$$|\vartheta^\epsilon(t,x) - \vartheta^\epsilon(s,x)| \leq \frac{1}{\epsilon} \sup_{u\in\mathscr{U}} \mathbb{E}\Big[\max_{\theta\in[t,s]} \big|X^u_{t,x}(\theta) - x\big| \bigvee \max_{\theta\in[s,T]} \big|X^u_{s,X^u_{t,x}(s)}(\theta) - x\big|\Big].$$

Under assumption (H1), there exists some constant $C > 0$ such that for any $0 \leq t \leq s \leq T$, $x, y \in \mathbb{R}^d$ the unique strong solution to Eq. (4) satisfies

$$\mathbb{E}\Big[\sup_{\theta\in[t,s]}\big|X^u_{t,x}(\theta) - X^u_{t,y}(\theta)\big|^2\Big] \leq C|x-y|^2, \tag{10}$$

$$\mathbb{E}\Big[\sup_{\theta\in[t,s]}\big|X^u_{t,x}(\theta) - X^u_{s,x}(\theta)\big|^2\Big] \leq C(1+|x|^2)\,|t-s| \tag{11}$$

(see for instance [27, Theorem 6.3]).

Hence, the result follows just taking $L_\epsilon := C/\epsilon$.

Remark 2 It has been proved in [6, Theorem 3.1] that, if $\mathscr{K} = \mathbb{R}^d$ and $\mathscr{C}$ is a non empty, convex set with a C^1 regular boundary, under the uniform ellipticity condition, for some $\alpha > 0$, $\forall (t, x, u) \in (0, T) \times \mathbb{R}^d \times U$,

$$\sigma(t, x, u)\sigma(t, x, u)^T \geq \alpha 1_d, \tag{12}$$

where 1_d is the identity matrix, the following holds:

$$|\vartheta(t, x) - \vartheta^\epsilon(t, x)| \leq C\frac{1 + |x|^2 + |\log\epsilon|}{(T-t)^d}\epsilon \tag{13}$$

for some constant C depending only on α and the constants in assumption (H1). We conjecture that analogous estimates can be obtained in the general case $\mathscr{K} \neq \mathbb{R}^d$, but a rigorous proof of this fact is still material of ongoing research.

3 Dynamic Programming and Hamilton-Jacobi-Bellman Equation

Aim of this section is to characterize the function ϑ^ϵ as a (viscosity) solution to a suitable HJB equation. For doing this, we closely follow the dynamic programming arguments recently developed in [10, 19] for optimal control problems with a cost depending on a running maximum. Therefore, in order to directly use those results in our framework, we will rewrite the optimal control problem (7) by means of the cost functional

$$J(t, x, u) := \mathbb{E}\Big[-\phi^\epsilon_{\mathscr{C}}(X^u_{t,x}(T)) \bigvee \max_{\theta\in[t,T]} -\phi^\epsilon_{\mathscr{K}}(X^u_{t,x}(\theta))\Big] \tag{14}$$

such that the following holds

$$\vartheta^\epsilon(t, x) = -\inf_{u\in\mathscr{U}} J(t, x, u).$$

The presence of the maximum operator inside the expectation, makes the cost in (14) non-Markovian preventing the direct use of the Dynamic Programming Principle (DPP), which is the first fundamental result towards the HJB characterisation. A classical strategy to overcome this difficulty consists in adding an auxiliary variable y that, roughly speaking, gets rid of the non-Markovian component of the cost. This has been originally used in [8] where an approximation technique of the L^∞-norm is used, whereas in [10, 19] the HJB equation is derived without making use of any approximation.

Let us introduce the following value function:

$$w^\epsilon(t,x,y) := \inf_{u\in\mathscr{U}} \mathbb{E}\Big[-\phi^\epsilon_{\mathscr{C}}(X^u_{t,x}(T)) \bigvee \max_{\theta\in[t,T]} -\phi^\epsilon_{\mathscr{K}}(X^u_{t,x}(\theta)) \bigvee y \Big]. \quad (15)$$

Defining the process

$$Y^u_{t,x,y}(.) := \max_{s\in[t,.]} -\phi^\epsilon_{\mathscr{K}}(X^u_{t,x}(s)) \bigvee y,$$

the value function (15) can also be written as

$$w^\epsilon(t,x,y) = \inf_{u\in\mathscr{U}} \mathbb{E}\Big[-\phi^\epsilon_{\mathscr{C}}(X^u_{t,x}(T)) \bigvee Y^u_{t,x,y}(T) \Big].$$

Observe that the following property holds:

$$\vartheta^\epsilon(t,x) = -w^\epsilon(t,x,-1), \quad (16)$$

so from now on only the optimal control problem (15) will be taken into account, since the corresponding value of the function ϑ^ϵ can be derived by the previous equality. The following property is satisfied:

Proposition 3 *Let assumption (H1) be satisfied. Then, there exists a constant $C > 0$ such that for any $\epsilon > 0$, $t, s \in [0,T]$, $(x,y), (x',y') \in \mathbb{R}^{d+1}$ one has*

$$|w^\epsilon(t,x,y) - w^\epsilon(s,x',y')| \le \frac{C}{\epsilon}\Big(|x-x'| + |y-y'| + |t-s|^{\frac{1}{2}}(1+|x|\vee|x'|)\Big).$$

Moreover, for any family of stopping times $\{\tau^u, u\in\mathscr{U}\}$ with values in $[t,T]$ one has

$$w^\epsilon(t,x,y) = \inf_{u\in\mathscr{U}} \mathbb{E}\Big[w^\epsilon(\tau^u, X^u_{t,x}(\tau^u), Y^u_{t,x,y}(\tau^u)) \Big] \quad (17)$$

for any $(t,x,y) \in [0,T]\times\mathbb{R}^{d+1}$.

Proof The regularity of w^ϵ with respect to t and x can be proved as in Proposition 2, while the Lipschitzianity with respect to y is trivial.

Thanks to the regularity of w^ϵ, the DPP (17) follows by arguments similar to [11] observing that for the couple of variables $(X^u_{t,x}(\cdot), Y^u_{t,x,y}(\cdot))$ the following property holds:

$$\begin{pmatrix} X^u_{t,x}(s) \\ Y^u_{t,x,y}(s) \end{pmatrix} = \begin{pmatrix} X^u_{\theta, X^u_{t,x}(\theta)}(s) \\ Y^u_{\theta, X^u_{t,x}(\theta), Y^u_{t,x,y}(\theta)}(s) \end{pmatrix} \qquad \text{a.s.}$$

for any $t \le \theta \le s \le T$ (with θ possibly a stopping time). We remand to [10] for a sketch of the proof showing how the arguments in [11] adapt to our case.

3.1 HJB Equation

Proposition 3 is the main tool for proving the next result that characterizes w^ϵ as the unique solution, in the viscosity sense, of a suitable HJB equation. In the sequel we will restrict our domain to

$$D := \{(x, y) \in \mathbb{R}^{d+1} : -\phi^\epsilon_{\mathscr{K}}(x) < y < 0\}.$$

Indeed, the knowledge of w^ϵ in $\overline{D}$ is sufficient to characterize it everywhere thanks to the following relation:

$$\begin{aligned} &w^\epsilon(t, x, y) = y && \text{for any } y \ge 0 \\ &w^\epsilon(t, x, y) = w^\epsilon(x, -\phi^\epsilon_{\mathscr{K}}(x)) && \text{for any } y \le -\phi^\epsilon_{\mathscr{K}}(x). \end{aligned} \tag{18}$$

Based on this observation, it is sufficient to characterise w^ϵ in the domain $\overline{D}$. Letting

$$\Gamma_1 := \{(x, y) \in \overline{D} : y = 0\}; \quad \Gamma_2 := \{(x, y) \in \overline{D} : y = -\phi^\epsilon_{\mathscr{K}}(x)\}, \tag{19}$$

we are going to prove that w^ϵ is the unique solution (in the weak sense specified in Definition 1 below) of the following HJB equation with mixed derivative-Dirichlet boundary conditions:

$$\begin{cases} -\partial_t w + H(t, x, D_x w, D^2_x w) = 0 & [0, T) \times D \\ w = 0 & [0, T) \times \Gamma_1 \\ -\partial_y w = 0 & [0, T) \times \Gamma_2 \\ w(T, x, y) = w_0(x, y) & \overline{D} \end{cases} \tag{20}$$

with

$$H(t, x, p, Q) := \sup_{u \in U} \left\{ -b(t, x, u)p - \frac{1}{2} Tr[\sigma\sigma^T](t, x, u)Q \right\} \tag{21}$$

and

$$w_0(x, y) := -\phi^{\epsilon}_{\mathscr{C}}(x) \bigvee -\phi^{\epsilon}_{\mathscr{K}}(x) \bigvee y.$$

We point out that the derivative boundary condition $-\partial_y w = 0$ on Γ_2 is typically obtained in presence of a running maximum cost, see [8, 10], while the Dirichlet condition $w^{\varepsilon} = 0$ on Γ_1 is obtained by the very definition of w^{ϵ}. Observe also that the constant Dirichlet condition on Γ_1 is compatible with the homogeneous derivative condition on Γ_2. This prevents possible problems related with mixed boundary conditions at the junctions where different components of the boundary cross.

The fully nonlinearity and degeneracy of the equation requires to consider solutions in the viscosity sense (see [14] for an overview on the subject). This notion of solution requires also to specify in which sense boundary conditions are satisfied. In particular, we ask the Dirichlet conditions on Γ_1 to be satisfied in the strong sense, whereas the derivative conditions on Γ_2 are considered in the (weak) viscosity sense.

Definition 1 A USC function $\underline{w}$ (resp. LSC function $\overline{w}$) on $[0, T] \times \overline{D}$ is a viscosity sub-solution (resp. super-solution) of (20), if for every function $\varphi \in C^{1,2}([0, T] \times \overline{D})$, at each maximum (resp. minimum) point (t, x, y) of $\underline{w} - \varphi$ (resp. $\overline{w} - \varphi$) the following inequality holds

$$\begin{cases} -\partial_t \varphi + H(t, x, D_x\varphi, D_x^2\varphi) \leq 0 & [0, T) \times D \\ \underline{w} \leq 0 & [0, T) \times \Gamma_1 \\ \min\left(-\partial_y \varphi, -\partial_t \varphi + H(t, x, D_x\varphi, D_x^2\varphi)\right) \leq 0 & [0, T) \times \Gamma_2 \\ \underline{w}(T, x, y) \leq w_0(x, y) & \overline{D} \end{cases}$$

(resp.

$$\begin{cases} -\partial_t \varphi + H(t, x, D_x\varphi, D_x^2\varphi) \geq 0 & [0, T) \times D \\ \overline{w} \geq 0 & [0, T) \times \Gamma_1 \\ \max\left(-\partial_y \varphi, -\partial_t \varphi + H(t, x, D_x\varphi, D_x^2\varphi)\right) \geq 0 & [0, T) \times \Gamma_2 \\ \overline{w}(T, x, y) \geq w_0(x, y) & \overline{D}. \end{cases})$$

A continuous function w on $[0, T] \times \overline{D}$ is a viscosity solution of (20) if it is both a sub- and super-solution.

Theorem 1 *Let assumption (H1) be satisfied. Then, w^{ϵ} is the unique bounded and continuous viscosity solution of the HJB equation* (20)*.*

Proof The Dirichlet and terminal conditions are ensured by the very definition of w^{ϵ} and its continuity. In particular, the continuity allows the conditions to be considered in the strong sense. The proof of sub- and supersolution properties in $[0, T) \times (D \cup \Gamma_2)$ follows quite straightforward by the the arguments in [10, Theorem 3.2] and [21, Theorem 4.1].

Uniqueness of the solution relies on comparison results for sub and super solution. The proof can be found in [19, Appendix A]. We point out that the fact of considering Dirichlet conditions in a strong sense is an important requirement for the proof of the comparison principle.

4 Numerical Approximation

In this section we discuss an approximation scheme for the unique continuous viscosity solution w^ϵ to the equation

$$\partial_t w + H(t, x, D_x w, D_x^2 w) = 0 \qquad -\phi^\epsilon_{\mathscr{K}}(x) < y < 0,\ t \in (0, T] \tag{22a}$$

$$w = 0 \qquad y = 0,\ t \in (0, T] \tag{22b}$$

$$-\partial_y w = 0 \qquad y = -\phi^\epsilon_{\mathscr{K}}(x),\ t \in (0, T] \tag{22c}$$

with initial data

$$w(0, x, y) = w_0^\epsilon(x, y) \qquad -\phi^\epsilon_{\mathscr{K}}(x) \le y \le 0 \tag{22d}$$

(the convenient change of variable $t \to T - t$ has been here applied).

In [10, Section 4.1] a general convergence result for numerical schemes approximating HJB equations under oblique derivative boundary conditions such as (22c) is provided. Those arguments can be easily modified in order to prove convergence also in presence of the additional Dirichlet boundary condition (22b) (see also [21]). Following the ideas introduced in [10], we present here a semi-Lagrangian (SL) scheme for the approximation of (22). The same scheme will be used in the numerical experiments in Sect. 5.

Let $N \ge 1$ be an integer (number of time steps), and let

$$h := \frac{T}{N} \quad \text{and} \quad t_n := nh$$

for $n = 0, \dots, N$. Let $\Delta x = (\Delta x_1, \dots, \Delta x_d) \in (\mathbb{R}_+^*)^d$ and $\Delta y > 0$, and let $\mathscr{G}_\eta$ (where $\eta \equiv (\Delta x, \Delta y)$) be the space grid

$$\mathscr{G}_\eta := \left\{ (x_i, y_j) = (i\Delta x, j\Delta y), \text{for } (i, j) \in \mathbb{Z}^d \times \mathbb{Z} \right\}.$$

The grid is considered uniform for simplicity of presentation. We also assume that the discretization in the y coordinate is aligned with the boundary of the domain, this allows us to get the Dirichlet condition exactly.

We look for a fully discrete scheme for the viscosity solution of (22) on the time-space grid $\{t_0, \dots, t_N\} \times (\mathscr{G}_\eta \cap \overline{D})$. Following the ideas in [10, 21] the numerical

scheme is defined starting from a standard scheme for (22a), which is then mixed with a step of "projection" on Γ_2 and the use of the Dirichlet condition on Γ_1. The approximation of equation (22a) we consider is the SL scheme proposed by Camilli and Falcone [13] and also used in [15]. We recall that first schemes of this type have been introduced by Menaldi in [23].

Let $\sigma^u = \sigma(\cdot, \cdot, u)$ and $b^u = b(\cdot, \cdot, u)$, and let $(\sigma_k^u)_{k=1,\dots,p}$ denote the column vectors of σ^u. We consider the following operator $\mathscr{T}$:

$$\mathscr{T}(\varphi)(t, x, y) := \min_{u \in U} \frac{1}{2p} \Big(\sum_{k=1,\dots,2p} \big[\varphi(t, ., y)\big]\big(x + hb^u(t, x) + \sqrt{h}\bar{\sigma}_k^u(t, x)\big) \Big) \quad (23)$$

with the following vector definition in $\mathbb{R}^d$:

$$\bar{\sigma}_{2k-j}^u := \sqrt{p}\,(-1)^j \sigma_k^u \quad (24)$$

for $k = 1, \dots, p$ and $j \in \{0, 1\}$. Now $[\cdot] \equiv [\cdot]_x$ stands for a monotone, P_1 interpolation operator on the x-grid (x_i), satisfying in particular:

$$\begin{cases} (i)\ [\varphi](x_i) = \varphi(x_i), \text{ for any } i \in \mathbb{Z}^d, \\ (ii)\ |[\varphi](x) - \varphi(x)| \le C|\Delta x|^2 \|D_x^2 \varphi\|_\infty \text{ for any } \varphi \in C^2(\mathbb{R}^d, \mathbb{R}), \\ (iii) \text{ for any functions } \varphi, \psi : \mathbb{R}^d \to \mathbb{R}, \varphi \le \psi \Rightarrow [\varphi] \le [\psi]. \end{cases} \quad (25)$$

We point out that (23), if considered without interpolation, is a discretization in time of the Dynamic Programming Principle. In particular, such an approximation uses an Euler-Maruyama scheme (see [20] for instance) coupled with a finite state discretization of the Gaussian distribution to approximate the dynamics $X_{t,x}^u(\cdot)$.

The numerical scheme is defined as follows:

Algorithm Initialization step, for $n = 0$, for all i, j:

$$W_{i,j}^0 = w_0^\epsilon(x_i, y_j).$$

Then, for $n = 0, \dots, N-1$:

Step 1 Compute $W_{i,j}^{n+1} = \mathscr{T}(W)(t_n, x_i, y_j)$, for all $(x_i, y_j) \in \mathscr{G}_\eta \cap \overline{D}$;

Step 2 Assign $W_{i,j}^{n+1} = W_{i,j_{x_i}}^{n+1}$, for all $(x_i, y_j) : y_j \le -\phi_{\mathscr{K}}^\epsilon(x_i)$;

$W_{i,j}^{n+1} = y_j$, for all $(x_i, y_j) : y_j \ge 0$;.

where for every $x \in \mathbb{R}^d$, $j_x \in \mathbb{Z}$ is defined by

$$j_x := \min\left\{ j \in \mathbb{Z} : \ j\Delta y \ge -\phi_{\mathscr{K}}^\epsilon(x) \right\}$$

and we used the following short notation

$$W_{i,j}^n = W(t_n, x_i, y_j).$$

Hereafter we will denote by $W = (W^n_{ij})^{n=1\dots N}_{(i,j)\in\mathbb{Z}^{d+1}}$ the solution of the numerical scheme defined by the algorithm above on $\{t_0, \dots, t_N\} \times \mathscr{G}_\eta$. We point out that the necessity of defining W also at mesh points outside $\overline{D}$ comes from the fact that the SL scheme involves values outside the domain. However, this is not a issue in virtue of (18) (see Step 2 above).

We also denote by $W^{\eta,h}$ the continuous extension of W to $[0,T] \times \mathbb{R}^d \times \mathbb{R}$ obtained by linear interpolation.

Remark 3 The numerical solution $W^{\eta,h}$ is Lipschitz continuous in y with Lipschitz constant independent of η and h. This can be derived by the very definition of the operator $\mathscr{T}$ in (23). Indeed, given W^n $L-$Lipschitz continuous in y one can observe that

$$\begin{aligned}|W^{n+1}_{i,j} - W^{n+1}_{i,j'}| &= |\mathscr{T}(W)(t_n, x_i, y_j \vee y_{j_{x_i}}) - \mathscr{T}(W)(t_n, x_i, y_{j'} \vee y_{j_{x_i}})| \\ &\le L|(y_j \vee y_{j_{x_i}}) - (y_{j'} \vee y_{j_{x_i}})| \le L|y_j - y_{j'}|.\end{aligned}$$

Hence, being W^0 Lipschitz continuous, the same property holds for W^n for all $n = 1 \dots N$. Then, since the linear interpolation (used to pass from W to $W^{\eta,h}$) preserve Lipschitz constants, we can obtain the desired property.

Remark 4 The core of the scheme in Step 1 can be written as

$$S(t, x, y, W^{n+1}_{i,j}, W) := W^{n+1}_{i,j} - \mathscr{T}(W)(t_n, x_i, y_j) = 0.$$

It is immediate to verify that S is monotone in the sense of Barles and Souganidis [7], i.e. for every $h, \eta > 0$, $r \in \mathbb{R}$, for all function ϕ, ψ such that $\phi \ge \psi$, inequality

$$S(t, x, y, r, \phi) \le S(t, x, y, r, \psi)$$

holds.

The choice of $\bar{\sigma}^u_k$ in (24) leads to the following consistency estimate, for any $\varphi \in C^{2,4}((0,T) \times \mathbb{R}^d \times \mathbb{R})$:

$$\begin{aligned}&\left|\frac{1}{h} S(t, x, y, \varphi(t, x, y), \phi) - \left(\partial_t \varphi + H(t, x, D_x\varphi, D^2_x\varphi)\right)\right| \\ &\le C_1 \Big(|b^u(t,x)|^2 \|D^2_x\varphi\|_\infty + |b^u(t,x)||\sigma^u(t,x)|^2 \|D^3_x\varphi\|_\infty + |\sigma^u(t,x)|^4 \|D^4_x\varphi\|_\infty \\ &\quad \|\partial^2_{tt}\varphi\|_\infty \Big) h + C_2 \|D^2_x\varphi\|_\infty \frac{|\Delta x|^2}{h}.\end{aligned}$$

These are classical properties of SL schemes, see [15] for instance. In particular, the error term in $|\Delta x|^2/h$ comes the interpolation error estimate (ii) (observe that we do not need to interpolate with respect to y) and the term in h from classical Taylor

expansions. Then, in order to ensure consistency of the scheme, Δx and h have to be chosen so that $|\Delta x|^2/h \to 0$ as $\Delta x, h \to 0$. This usually leads to the choice $\Delta x \sim h$ in numerical simulations.

Moreover, it is easy to verify that the scheme admits a bounded solution in $\{t_0, \ldots, t_N\} \times (\mathscr{G}_\eta \cap \overline{D})$, so that the scheme is also stable.

We recall that monotonicity, consistency and stability are the fundamental properties necessary for proving convergence of numerical schemes in the framework of viscosity solutions, see [7].

Theorem 2 *Let assumption (H1) be satisfied. Let $W^{\eta,h}$ be the solution of the scheme defined by the Algorithm above, where $\mathscr{T}$ is the SL scheme (23)–(24). Then, if*

$$\frac{|\Delta x|^2}{h} \to 0 \qquad as \qquad \Delta x, h \to 0 \tag{26}$$

$W^{\eta,h}$ converges to w^ϵ in $[0, T] \times \overline{D}$ as $\eta, h \to 0$.

Proof The proof follows the strategy in [7] and [10]. Let us define for $(t, x, y) \in [0, T] \times \overline{D}$

$$\overline{W}(t, x, y) := \limsup_{\substack{[0,T]\times\overline{D}\ni(s,\xi,\gamma)\to(t,x,y)\\ \eta,h\to 0}} W^{\eta,h}(s, \xi, \gamma),$$

$$\underline{W}(t, x, y) := \liminf_{\substack{[0,T]\times\overline{D}\ni(s,\xi,\gamma)\to(t,x,y)\\ \eta,h\to 0}} W^{\eta,h}(s, \xi, \gamma).$$

One clearly has $\underline{W}(t, x, y) \le \overline{W}(t, x, y)$ for any $(t, x, y) \in [0, T] \times \overline{D}$. Convergence follows by the comparison principle once shown that $\overline{W}$ and $\underline{W}$ are respectively a sub- and supersolution of the HJB equation in the sense of Definition 1.

We sketch the proof of the subsolution property, the supersolution part can be proved in a similar way. Given a smooth test function φ, let $(\bar{t}, \bar{x}, \bar{y})$ be a maximum point for $(\underline{W} - \varphi)$, with $(\underline{W} - \varphi)(\bar{t}, \bar{x}, \bar{y}) = 0$, and let $(\eta_k, h_k, t_k, x_k, y_k)$ be such that $(t_k, x_k, y_k) \in [0, T] \times \overline{D}$, $\eta_k, h_k \to 0$, $(t_k, x_k, y_k) \to (\bar{t}, \bar{x}, \bar{y})$, $W^{\eta_k,h_k}(t_k, x_k, y_k) \to \underline{W}(\bar{t}, \bar{x}, \bar{y})$ and

$$(W^{\eta_k,h_k} - \varphi)(t_k, x_k, y_k) = \max(W^{\eta_k,h_k} - \varphi) = \delta_k \to 0$$

(the existence of such a sequence follows by classical arguments in viscosity theory).

If $(\bar{x}, \bar{y}) \in D$ the result follows as in [7] using the properties of monotonicity and consistency of the scheme in a sufficiently small neighborhood of $(\bar{x}, \bar{y})$ still contained in D.

If $(\bar{x}, \bar{y}) \in \Gamma_2$ one can work under the condition $-\partial_y \varphi(\bar{t}, \bar{x}, \bar{y}) > 0$, otherwise the subsolution property is automatically satisfied. In this case the result follows observing that, by the very definition of the scheme (see Step 2 of the algorithm) and its monotonicity, one can derive

$$\varphi(t_k, x_k, y_k) + \delta_k \leq \mathscr{T}(\varphi)(t_k, x_k, y_k \vee y_{j_{x_k}}) \leq \mathscr{T}(\varphi)(t_k, x_k, y_k)$$

so that the subsolution property follows again by the consistency of the scheme.

It remains to prove that $\underline{W}$ satisfies the Dirichlet condition pointwise on Γ_1. For this purpose it is worth to observe that $W^{\eta,h}$ is Lipschitz continuous in y (see Remark 3), i.e. there exists some constant $L > 0$ (independent of η, h, t, x) such that

$$|W^{\eta,h}(t, x, y) - W^{\eta,h}(t, x, y')| \leq L|y - y'|$$

for any $t \in [0, T]$, $x \in \mathbb{R}^d$, $y, y' \in \mathbb{R}$. Therefore one has

$$|W^{\eta,h}(s, \xi, \gamma) - (-1)| = |W^{\eta,h}(s, \xi, \gamma) - W^{\eta,h}(s, \xi, 0)| \leq L|\gamma|$$

so that on Γ_1 (i.e. for $\bar{y} = 0$)

$$\liminf_{\substack{[0,T]\times\overline{D}\ni(s,\xi,\gamma)\to(\bar{t},\bar{x},\bar{y}) \\ \eta,h\to 0}} W^{\eta,h}(s, \xi, \gamma) = -1.$$

5 Numerical Tests

In this section we present some numerical tests for probabilistic reachability problems in presence of state constraints. To solve the HJB equation (22), we use the fully discrete SL scheme introduced in Sect. 4 implemented on the ROC-HJ solver available at the link https://uma.ensta-paristech.fr/soft/ROC-HJ/. The minimum in (23) is performed on a subset of control values $\{u_1, \ldots, u_{N_u}\}$ that represents a discretization of U with a mesh size Δu. In all the simulations the regularization parameter will be chosen to be $\epsilon = 1.E - 08$.

5.1 Example 1

We consider the following controlled stochastic system:

$$\mathrm{d}X(s) = \left(\begin{pmatrix} -1 & -4 \\ 4 & -1 \end{pmatrix} X(s) + u(s) \right) \mathrm{d}s + \begin{pmatrix} 0.7 & 0 \\ 0 & 0.7 \end{pmatrix} \begin{pmatrix} \mathrm{d}B_1(s) \\ \mathrm{d}B_2(s) \end{pmatrix} \tag{27}$$

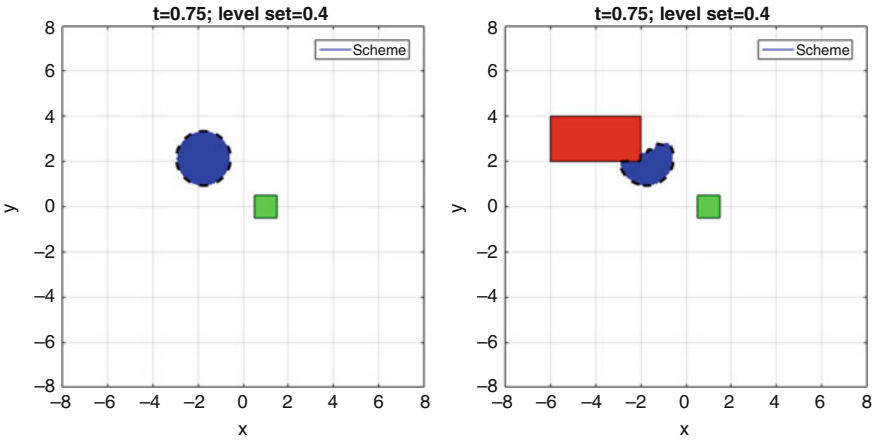


Fig. 2 (Example 1) Backward reachable sets at $t = 0.75$ for a time horizon $T = 1.75$ and $\rho = 0.4$ without (left) and with (right) obstacle. The target set, the obstacle and the backward reachable set $\Omega_{0.75}^{0.4}$ are represented respectively by the green square, the red rectangle and the blue region

where $u(s) = \begin{pmatrix} u_1(s) \\ u_2(s) \end{pmatrix}$, $u_i(s) \in [-0.1, 0.1]$, for $i = 1, 2$ and B_1, B_2 are two independent Brownian motions.

The linear system (27) has been used in [6] to validate the HJB approach in the characterization of an approximated probabilistic reachable set without state constraints and in [5] to illustrate an approximation of the probability of reaching a target by using enclosing hulls of probability density functions.

We set $T = 1.75$ and define the target $\mathscr{C} := (0.5, 1.5) \times (-0.5, 0.5)$ (green square in Fig. 2). The constraint is given by the presence of an obstacle, represented in Fig. 2 (right) by the red rectangle, i.e. $\mathscr{K} := \mathbb{R}^2 \setminus ([-6, -2] \times [2, 4])$. We compute the set Ω_t^ρ (blue region) for $t = 0.75$ and $\rho = 0.4$ in presence (Fig. 2, right) or not (Fig. 2, left) of the obstacle. To approximate the auxiliary function w^ϵ solution to (22), the numerical simulation is performed on a computational domain $[-8, 8]^2 \times [-1, 0]$. The corresponding values of ϑ^ϵ are then obtained using relation (16).

Figure 3 (top) shows the set Ω_t^ρ for $\rho = 0.4$ at different time $t \in \{0.25; 0\}$ in presence of the obstacle. Then, in Fig. 3 (bottom) we simulate different optimal paths starting from a given point of the backward reachable set using the algorithm described below.

Algorithm (Trajectory Reconstruction) Initialization: Set $X_0 = \bar{x}$.
For $k = 0$ to $N - 1$:

Step 1 Compute optimal control at $t = t_k$:

$$u^k = \arg\min_{u \in \{u_1, \ldots, u_{N_u}\}} \mathbb{E}[W(t_{k+1}, X_{k+1}^u, -1)]$$

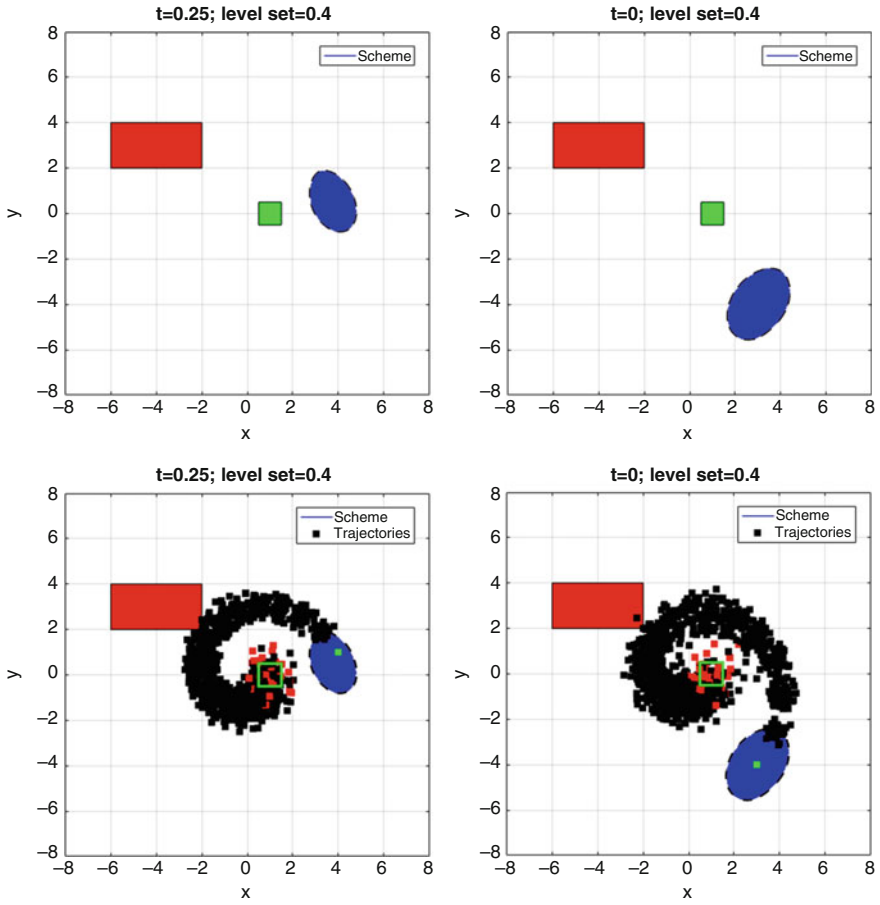


Fig. 3 (Example 1). Top: backward reachable set (blue region) at times $t \in \{0.25,\ 0\}$ for a final time horizon $T = 1.75$ in presence of the obstacle (red rectangle). Bottom: reconstruction of some optimal paths starting from point $\bar{x} = (4, 1)$ (bottom, left) and $\bar{x} = (3, -4)$ (bottom, right)

where for $u_i \in \{u_1, \ldots, u_{N_u}\}$

$$X_{k+1} := X_k + b(t_k, X_k, u_i)h + \sigma(t_k, X_k, u_i)\sqrt{h}\xi,$$

here $\xi := (\xi_1, \xi_2)$ with ξ_i $(i = 1, 2)$ random variables following a $N(0, 1)$ distribution.

Step 2 Compute the point of the optimal trajectory:

$$X_{k+1} := X_k + b(t_k, X_k, u^k)h + \sigma(t_k, X_k, u^k)\sqrt{h}\xi$$

where again $\xi_i \sim N(0, 1)$ for $i = 1, 2$.

Table 1 (Example 1) Percentage p of M simulated trajectories that reach the target set without hitting the obstacle, with a corresponding confidence interval (C.I.), and a Monte Carlo error estimate (MC-error)

	M	p	C.I.	MC-error
$\bar{x}_1$	6000	0.4630	(0.4504, 0.4756)	0.0126
	12,000	0.4624	(0.4535, 0.4713)	0.0089
	25,000	0.4603	(0.4541, 0.4664)	0.0062
	50,000	0.4618	(0.4574, 0.4661)	0.0045
	100,000	0.4628	(0.4597, 0.4659)	0.0031
$\bar{x}_2$	6000	0.3915	(0.3593, 0.4037)	0.0122
	12,000	0.3991	(0.3705, 0.4078)	0.0087
	25,000	0.4026	(0.3966, 0.4087)	0.0061
	50,000	0.4016	(0.3996, 0.4081)	0.0043
	100,000	0.4015	(0.3985, 0.4045)	0.0030

In order to validate our approach, we compare the value of the scheme in a given point with the percentage of trajectories reaching the target without hitting the obstacle. We consider the case $t = 0.25$ and two different starting points $\bar{x}_1 := (4.0, 1.0)^T$ and $\bar{x}_2 := (3.0, 0.0)^T$. The approximation of the level-set function obtained by numerically solving the corresponding HJB equation on the grid $\Delta x_1 = \Delta x_2 = 0.0125$, $\Delta y = 0.1$, $h = 0.025$ at points $\bar{x}_1$ and $\bar{x}_2$ is respectively $-W(t, \bar{x}_1, -1) \simeq 0.459 \pm 0.004$ and $-W(t, \bar{x}_2, -1) \simeq 0.404 \pm 0.009$.

The results of Monte Carlo simulations are reported in Table 1. One can conclude that the approximated value of the level set function belongs in each case to the confidence interval.

5.2 Example 2

We now test our method on the same example used in [10, Section 6]. Let us consider the following dynamics:

$$\mathrm{d}X(s) = u(s)\begin{pmatrix}1\\0\end{pmatrix}\mathrm{d}s + u(s)\sigma(X(s))\mathrm{d}B(s),\ s \geq t,$$

where B is a one-dimensional Brownian motion, $U = [0, 1] \subset \mathbb{R}$ and the volatility $\sigma(x)$ is given by

$$\sigma(x) := 5\, d_{\Theta}(x)\begin{pmatrix}0\\1\end{pmatrix}$$

where d_{Θ} denotes the distance function to the set

$$\Theta := \big\{(x_1, x_2),\ |x_2| \geq 0.3\big\}.$$

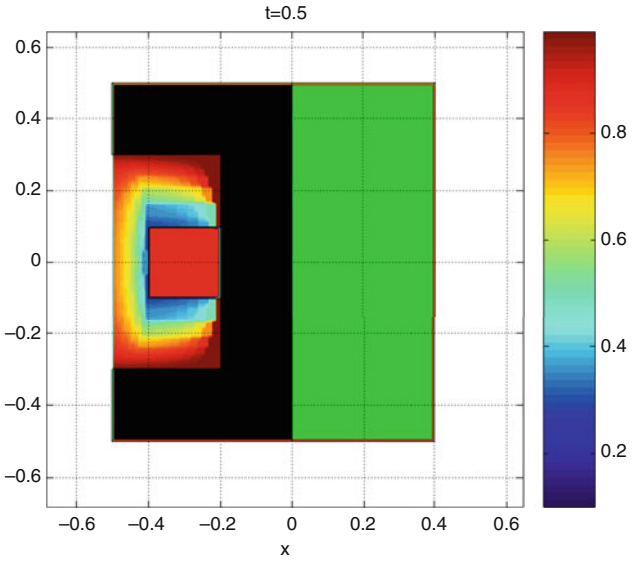


Fig. 4 (Example 2) Approximation of Ω_t^ρ at $t = 0$ for different levels $\rho \in [0, 1, 0.9]$ (indicated by the color bar) computed with $\Delta x_1 = \Delta x_2 = 0.005, \Delta y = 0.1, h = 0.01$ (same mesh parameters used in [10])

The target set is $\mathscr{C} = (0, 0.4) \times (-0.5, 0.5)$ (green rectangle in Fig. 4) and the state constraint is $\mathscr{K} = \mathbb{R}^2 \setminus ([-0.4, 0.2] \times [-0.1, 0.1])$ (i.e. the entire space except the red square obstacle in Fig. 4). We fix $T = 0.5$ and consider the computational domain $[-1, 1]^2 \times [-1, 0]$.

The strong degeneracy of the diffusion term in this example allowed in [10] to obtain the "almost sure" backward reachable set, corresponding here to the limit case $\rho = 1$. Figure 4 shows the approximation of Ω_t^ρ for $t = 0$ and different levels $\rho \in [0, 1, 0.9]$. The black region corresponds to the exact backward reachable set for $\rho = 1$. Indeed, due to the simple dynamics considered it is possible for this example to infer the exact set Ω_0^1, i.e. the set of points from where the target is reached and the constraint satisfied with probability one, see [10] for a further discussion. One can observe that as ρ approaches the value 1, we recover the results obtained in [10]. A loss of precision appears at corners. This was already noticed in [10] and it is due to the smoothing effects of the diffusion term (see [10, Figure 2, Section 6]) which can be reduced with the refinement of the mesh.

6 Conclusions

In this paper we have used the HJB theory for characterising the probabilistic backward reachable set for a system of controlled diffusions in presence of state constraints. We have shown that such a set is a level set of the value function

associated to a suitable optimal control problem. To deal with the discontinuity of the cost functional associated to this problem, arising from the use of indicator functions for representing probabilities, we have defined a regularised problem. Precise estimates of the error introduced by this regularisation are still object of ongoing research.

Following the approach in [10, 19, 21], for the regularised problem we have obtained a characterization by a HJB equation with mixed Dirichlet-derivative boundary conditions. We have defined a fully discrete SL approximation scheme and we have proved its convergence to the unique viscosity solution of the equation. Then, we have used such a scheme in order to validate our approach on some numerical tests. We focused on the examples studied in [6] and [10], adding state constraints to the first one and variable levels of probability to the second one. More complex tests on concrete models are a promising future direction of work.

Acknowledgements The authors are sincerely grateful to Olivier Bokanowski and Hasnaa Zidani for their guidance at the early stage of this paper.

References

1. Abate, A., Amin, S., Prandini, M., Lygeros, J., Sastry, S.: Computational approaches to reachability analysis of stochastic hybrid systems. In: Hybrid Systems. Lecture Notes in Computer Science, vol. 4416(1), pp. 4–17 (2007)
2. Abate, A., Prandini, M., Lygeros, J., Sastry, S.: Probabilistic reachability and safety for controlled discrete time stochastic hybrid systems. Automatica **44**, 2724–2734 (2008)
3. Althoff, M., Stursberg, O., Buss, M.: Safety assessment of autonomous cars using verification techniques. In: 2007 American Control Conference, pp. 4154–4159 (2007)
4. Althoff, M., Stursberg, O., Buss, M.: Safety assessment for stochastic linear systems using enclosing hulls of probability density functions. In: 2009 European Control Conference (ECC), pp. 625–630 (2009)
5. Althoff, M., Stursberg, O., Buss, M.: Safety assessment for stochastic linear systems using enclosing hulls of probability density functions. In: European Control Conference (ECC), pp. 625–630. IEEE (2009)
6. Assellaou, M., Bokanowski, O., Zidani, H.: Error estimates for second order hamilton-jacobi-bellman equations. approximation of probabilistic reachable sets. Discrete Contin. Dynam. Syst. Ser. A **35**(9), 3933–3964 (2015)
7. Barles, G., Souganidis, P.E.: Convergence of approximation schemes for fully nonlinear second order equations. Asymptot. Anal. **4**, 271–283 (1991)
8. Barron, E.N.: The Bellman equation for control of the running max of a diffusion and applications to lookback options. Appl. Anal. **48**, 205–222 (1993)
9. Bokanowski, O., Forcadel, N., Zidani, H.: Reachability and minimal times for state constrained nonlinear problems without any controllability assumption. SIAM J. Control Optim. **48**(7), 4292–4316 (2010)
10. Bokanowski, O., Picarelli, A., Zidani, H.: Dynamic programming and error estimates for stochastic control problems with maximum cost. Appl. Math. Optim. **71**(1), 125–163 (2015)
11. Bouchard, B., Touzi, N.: Weak dynamic programming principle for viscosity solutions. SIAM J. Control Optim. **49**(3), 948–962 (2011)
12. Bouchard, B., Elie, R., Touzi, N.: Stochastic target problems with controlled loss. SIAM J. Control Optim. **48**(5), 3123–3150 (2009)

13. Camilli, F., Falcone, M.: An approximation scheme for the optimal control of diffusion processes. RAIRO Modél. Math. Anal. Numér. **29**(1), 97–122 (1995)
14. Crandall, M.G., Ishii, H., Lions, P.L.: User's guide to viscosity solutions of second order partial differential equations. Bull. Am. Math. Soc. **27**(1), 1–67 (1992)
15. Debrabant, K., Jakobsen, E.R.: Semi-Lagrangian schemes for linear and fully non-linear diffusion equations. Math. Comp. **82**(283), 1433–1462 (2012)
16. Falcone, M., Giorgi, T., Loreti, P.: Level sets of viscosity solutions: some applications to fronts and rendez-vous problems. SIAM J. Appl. Math. **54**, 1335–1354 (1994)
17. Fleming, W.H., Soner, H.M.: Controlled Markov Processes and Viscosity Solutions, 2nd edn. Springer, New York (2006)
18. Föllmer, H., Leukert, P.: Quantile hedging. SIAM J. Comput. Phys. **3**(3), 251–273 (1999)
19. Grüne, L., Picarelli, A.: Zubov's method for controlled diffusions with state constraints. Nonlinear Differ. Equ. Appl. **22**(6), 1765–1799 (2015)
20. Kloeden, P.E., Platem, E.: Numerical Solution of Stochastic Differential Equations. Springer, Berlin/New York (1992)
21. Kröner, A., Picarelli, A., Zidani, Z.: Infinite horizon stochastic optimal control problems with running maximum cost. SIAM J. Control Optim. **56**(5), 3296–3319 (2018)
22. Margellos, K., Lygeros, J.: Hamilton-Jacobi formulation for Reach-avoid differential games. IEEE Trans. Autom. Control **56**, 1849–1861 (2011)
23. Menaldi, J.L.: Some estimates for finite difference approximations. SIAM J. Control Optim. **27**, 579–607 (1989)
24. Mitchell, I., Bayen, A., Tomlin, C.: A time-dependent Hamilton-Jacobi formulation of reachable sets for continuous dynamic games. IEEE Trans. Autom. Control **50**, 947–957 (2005)
25. Osher, S., Sethian, A.J.: Fronts propagating with curvature dependent speed: algorithms on Hamilton-Jacobi formulations. J. Comp. Phys. **79**, 12–49 (1988)
26. Soner, H., Touzi, N.: A stochastic representation for level set equations. Commun. Partial Differ. Equ. **27**(9–10), 2031–2053 (2002)
27. Yong, J., Zhou, X.Y.: Stochastic Controls. Applications of Mathematics (New York), vol. 43. Springer, New York (1999). Hamiltonian Systems and HJB Equations

An Iterative Solution Approach for a Bi-level Optimization Problem for Congestion Avoidance on Road Networks

Andreas Britzelmeier, Alberto De Marchi, and Matthias Gerdts

Abstract The paper introduces an iterative solution algorithm for a bi-level optimization problem arising in traffic control. The bi-level problem consists of a shortest path problem on the upper level, which aims at minimizing the total path cost of a set of cars in a road network. The cost coefficients in the shortest path problem represent the expected driving time on each edge, accounting for congestions, and depend on the solutions of a set of lower level optimal control problems, each one describing the behavior of a single minimum-time driven car. On the other hand, each lower level problem is built upon the path planned by the upper level. This leads to a strong coupling between upper level problem and lower level problem. This coupling is decomposed by an iterative procedure fixing either the costs or the paths in the upper level and the lower level, respectively. Numerical experiments illustrate the procedure and indicate that the iterative algorithm leads to suitable distribution of cars in the network.

Keywords Bi-level optimization · Traffic control · Time-optimal control · Network optimization · Iterative methods

1 Introduction

Increasing traffic loads due to a steadily growing population and rising commerce, poses a problem especially to urban areas. Nevertheless the economical aspect of CO_2 pollution is an imminent threat to the health of humans. Reducing traffic seems to be the main idea to solve these problems. However, banning cars from cities or cramming people into public transportations, seems not to be an attractive and productive solution. A different approach would be to reduce the total time a car needs to reach its destination in the sense that the flux of cars is optimized. Considering the introduction of automatic or autonomous cars, this could be

A. Britzelmeier · A. De Marchi · M. Gerdts (✉)
Department of Aerospace Engineering, Bundeswehr University Munich, Neubiberg, Germany
e-mail: andreas.britzelmeier@unibw.de; alberto.demarchi@unibw.de; matthias.gerdts@unibw.de

M. Falcone et al. (eds.), *Numerical Methods for Optimal Control Problems*,
Springer INdAM Series 29, https://doi.org/10.1007/978-3-030-01959-4_2

achieved by controlling the cars such that their paths and velocity is optimized with respect to avoid traffic jams. In this paper we propose a bi-level optimal control problem and an iterative scheme for optimizing the network-wide traffic flow. The upper level problem controls the overall vehicle distribution with an adaptive shortest path algorithm. The route planning for each car is based on shared costs, derived from coupling single cars behaviour. The lower level is concerned with providing optimal velocity profiles and density updates to the shortest path algorithm, such that speed limits are simulated. Palagachev and Gerdts [8] propose two approaches for solving a bi-level optimization problem. Either by treating the lower level problem as a parametric optimization problem, which is solved whenever it is required for the upper level, or by reducing the problem to a single level problem by replacing the lower level through its necessary conditions. A semi-analytic solution approach for minimum-time velocity profiles can be found in [1].

It should be noticed that, within the aforementioned problem, drivers can be seen as players in a differential game, insofar as they are aware of future traffic distribution starting from the current one. In this case, if a solution exists, it likely represents a global equilibrium among drivers on the network; it has been argued in [3] that this solution is strictly related to Wardrop's equilibrium [11]. More details and references can be found in [3].

The paper is organized as follows. Section 2 provides an overview on the bi-level optimization problem. Sections 3 and 4 formulate and propose numerical methods to solve the upper and lower level optimization problems. Section 5 discusses how the two levels interface with each other and finally in Sect. 6 we apply the iterative procedure and report numerical results.

2 Problem Formulation and Solution Approach

A road network can be represented by a graph $\mathbb{G} = (V, E)$ consisting of a vertex set V and an edge set E, see [2, 6]. An edge is the topological description of a road segment, and a vertex corresponds to an intersection. Edges have properties like, e.g., length, speed limit, maximum density (i.e. maximum number of vehicles per unit length). A set of cars C move on the graph $\mathbb{G}$, in the sense that cars are initially positioned on a vertex and aim at reaching another vertex by following a suitable path (i.e. a sequence of edges) on the graph $\mathbb{G}$. These agents interact at the microscopic scale, yielding macroscopic effects like congestions, traffic waves and self-organizational phenomena, compare [4, 6, 7, 10].

The problem here is how to plan a route for each and every car, from the initial to the desired point, taking into account traffic jam, driver's behavior, vehicle dynamics and speed limits. Drivers are supposed to aim at the minimum-time control of their own car, while obeying speed limits and constraints on the vehicle dynamics. Thus, the overall problem is here formulated as a bi-level optimization problem (BOP), where route planning is represented by the upper level optimization problem and the lower level optimization problem is adopted to predict how drivers will behave, given a certain path. The route planning, also referred to as upper level optimization problem (UL-OP), aims at finding the minimum-cost path for each car, given its

initial and final position. Instead, the lower level optimization problem (LL-OP) represents an optimal control problem with vehicle and road constraints. These two problems exchange information in the sense that they depend on each other. The UL-OP can be seen as constrained by solutions of the LL-OP, because the cost of each edge depends on the actual traffic jam, in terms of car density. On the other hand, the UL-OP affects the LL-OP, because the minimum-time control, and consequent optimal speed profile, depends on the planned path with corresponding length of edges and speed limits.

Some assumptions and simplifications are adopted throughout the present work: Road geometry is time-invariant; speed limits are considered constant in time and space on each single road segment.

In summary, the traffic control problem results in the following bi-level optimization problem whose details are described in Sects. 3 and 4:

The Upper Level Optimization Problem (UL-OP) reads as follows:

$$\begin{aligned} \textit{Minimize} \quad & \sum_{k \in C} c^k (x^k)^\top z^k && (1) \\ \textit{with respect to} \quad & (z^k, x^k, v^k, u^k), k \in C, \\ \textit{subject to} \quad & A z^k = b^k, z^k \geq 0, k \in C \\ & (x^k, v^k, u^k) \in \mathcal{M}(z^k), k \in C. \end{aligned}$$

Herein, $\mathcal{M}(z^k)$ denotes the set of minimizers of the following Lower Level Optimization Problem (LL-OP):

$$\begin{aligned} \textit{Minimize} \quad & T && && (2) \\ \textit{subject to} \quad & \dot{x}(t) = v(t), && \dot{v}(t) = f_k(v(t), u(t)), \\ & x(0) = 0, && x(T) = L_k, \\ & v(0) = v_k^0, && v(t) \in [0, \overline{v}_k(x(t))], \\ & u(t) \in \mathcal{U}_k. \end{aligned}$$

The index k indicates that the corresponding quantities depend on z^k. The function f_k represents the vehicle dynamics, the box $\mathcal{U}_k$ defines control constraints, $\overline{v}_k : \mathbb{R}_+ \to \mathbb{R}_+$ and L_k are speed limits and length of the driving path, respectively. Further c^k defines the edge costs, which are depending on the result x^k of the LL-OP. A denotes the (reduced) node-edge incidence matrix of the network. The set of cars is described by C. The vector b^k denotes a unit vector, which indicates the starting position in the network and z^k holds the path indicator variables.

There are basically a few main techniques for solving bi-level optimization problems. The first approach keeps the bi-level structure and treats the LL-OP as a parametric optimization problem, which is being solved whenever the solution algorithm for the UL-OP requires it [8]. The second technique, instead, is based on the formulation of first order necessary optimality conditions for the LL-OP. Then, the LL-OP is replaced by its necessary conditions, which are considered as constraints in the UL-OP. This reduces the bi-level problem into a single-level nonlinear optimization problem, but in general this is not equivalent to the original problem, since necessary conditions might be not sufficient [8]. A third approach is based on the substitution of the LL-OP with its value function. This generates an equivalent single-level optimization problem.

In this paper, we chose to follow an approach that resembles the first one discussed above, but we treat the two levels as coupled optimization problems, while iteratively solving one after the other. In general, during the iterative procedure, first the UL-OP is solved to compute the required input variables for the LL-OP. Further solving the LL-OP leads to an update of the weights of the UL-OP for the next iteration until a stopping criterion is satisfied. Considering such an iterative procedure, the LL-OP and UL-OP are solved the same number of times and the levels are treated as uncoupled problems, just coupled at the interface by the procedure itself. The procedure is explained in more detail in Algorithm 2.

Since we are not yet aware of any formal convergence result for such an iterative scheme, one purpose of this paper is to experimentally investigate if the procedure converges or if oscillations can be observed. Please note that the above bi-level problem is a hard problem and also the alternative second and third solution approaches mentioned before are very difficult to realize numerically owing to non-smoothness issues.

3 Upper Level: Route Planning

Let the road network be described through a directed graph $\mathbb{G} = (V, E, c, s, t)$, with vertices $V = \{1, 2, \ldots, n\}$ and edges E. For simplicity we assume that the vertices are numbered such that the initial vertex is given by $s := 1$ whereas the target vertex is $t := n$. The cost c_{ij} of each edge $(i, j) \in E$ is often associated to the length of the corresponding road segment, such that $c : E \to \mathbb{R}_+$ defines a cost function, see [2]. The shortest path problem for an individual car starting at s and moving to t can be formulated mathematically as follows, compare [9]:

$$\textit{Minimize} \quad \sum_{(i,j)\in E} c_{ij} z_{ij} \quad \textit{subject to} \quad Az = e^1, z_{ij} \geq 0, (i, j) \in E,$$

where z_{ij} is the load transported along the edge $(i, j) \in E$, e^1 is the canonical unit vector, and A denotes the reduced node-edge incidence matrix of $\mathbb{G}$. Note

that A is a totally unimodular matrix and hence the linear optimization problem possesses a binary solution with $z_{ij} \in \{0, 1\}$ for all $(i, j) \in E$. The shortest path then consists of all edges (i, j) with $z_{ij} = 1$. An efficient implementation for solving the above linear program is based on a primal-dual algorithm as described in, e.g. [9], and leads to the famous Dijkstra's algorithm [5] in Algorithm 1. Please note that extensions like the A* algorithm exist. After termination of Algorithm 1 $d(i)$ contains the length of a shortest path from s to i and $p(i)$ contains the predecessor of i on such a shortest path.

Algorithm 1: Dijkstra algorithm

Input: Set $W = \{s\}$, $d(s) = \{0\}$ and $d(i) = \infty$ for all $i \in V \setminus \{s\}$.
forall the $i \in V \setminus \{s\}$ *and* $(s, i) \in E$ **do**
  $\longrightarrow$ set $d(i) = c_{1i}$, $p(i) = s$
end
while $W \neq V$ **do**
  $\longrightarrow$ find $k \in V \setminus W$ where $d(k) = min\{d(i) : i \in V \setminus W\}$
  $\longrightarrow$ $W = W \cup k$
  forall the $i \in V \setminus W$ *with* $(k, i) \in E$ **do**
    if $d(i) > (d(k) + c_{ki})$ **then**
      $\longrightarrow$ $d(i) = d(k) + c_{ki}$
      $\longrightarrow$ $p(i) = k$
    end
  end
end

Now we are interested in minimizing the total path length, which is obtained by summing up the lengths of all individual shortest paths of the cars in the road network. To this end let $c^k = (c^k_{ij})_{(i,j)\in E} > 0$ denote the cost vector of car $k \in C$, $z^k = (z^k_{ij})_{(i,j)\in E}$ the corresponding path indicator variables, and $b^k = (b^k_i)_{i\in V}$ the unit vector that indicates the starting node of car $k \in C$. With this notation, the task to minimize the total path length for all cars in C yields the following upper level problem UL-OP:

$$\begin{aligned} \textit{Minimize} \quad & \sum_{k\in C} (c^k)^\top z^k = \sum_{k\in C} \sum_{(i,j)\in E} c^k_{ij} z^k_{ij} \\ \textit{subject to} \quad & Az^k = b^k, z^k \geq 0, k \in C. \end{aligned}$$

Please note that UL-OP is a separable optimization problem and its solution can be obtained by solving individual shortest path problems for all cars in C and summing up the lengths.

So far, we assumed that the cost vectors c^k, $k \in C$, are given vectors. This assumption will be dropped in the sequel by taking into account individual trajectories for each car on the shortest paths. To this end, the costs of each edge follow an evolution, depending on the congestion of the roads and therefore on the speed of the vehicles on the same edge e. Thus, the cost vectors will depend on the solution of lower level optimal control problems, which will be discussed in the following Sect. 4.

4 Lower Level: Minimum Time Driving

We aim at computing minimum-time trajectories on a given path in the road network. The vehicle dynamics are described by a second-order time-invariant linear system for simplicity. We take into account a linear drag force. The validity of this assumption significantly depends on the velocity regime, but it simplifies the derivation of a semi-analytical solution to the LL-OP. There exist results also accounting for both, linear and quadratic drag forces, see [1]. We point out that this simplification is not necessary for the proposed iterative scheme, but it reduces the computational time required for solving the lower level problem LL-OP.

Each individual car minimizes the time required to arrive at the destination subject to acceleration and speed limits. It is noticeable that at this level agents do not interact, in fact, no coupling between cars is present in LL-OP (2). This inaccuracy is more negligible as density gets lower and traffic congestions are avoided. In this section we focus on a single car. Each vehicle is characterized by its mass $m_D > 0$, its linear drag coefficient $c_D \geq 0$, its initial speed $v_0 \geq 0$ and its maximum braking and pushing forces $F_{\text{brake}} \in (-\infty, 0)$ and $F_{\text{push}} \in (0, +\infty)$. Let us introduce the drag parameter $c := c_D/m_D \geq 0$ and control bounds $\underline{u} := F_{\text{brake}}/m_D$ and $\overline{u} := F_{\text{push}}/m_D$. Let a path $p = (p_0, \ldots, p_N)$ with vertices $p_j \in V$, $j = 0, \ldots, N$, be given. With each edge $e_j = (p_j, p_{j+1})$ on the path we associate a (physical) distance ℓ_j, $j \in \{0, \ldots, N-1\}$. The total length of the path is then $L_p = \sum_{j=0}^{N-1} \ell_j$. We assume that a piecewise constant speed limit function $\overline{v} : [0, L_p] \to \mathbb{R}$ is given with $\overline{v}(x) := \overline{v}_j > 0$ for $x \in [a_j, a_j + \ell_j)$, $j \in \{0, \ldots, N-1\}$, and $a_j := \sum_{k=0}^{j-1} \ell_k$.

Each vehicle aims at solving the following path minimum-time optimization problem:

$$
\begin{aligned}
&\textit{Minimize} && T && && (3)\\
&\textit{subject to} && \dot{x}(t) = v(t), && \dot{v}(t) = u(t) - cv(t),\\
& && x(0) = 0, && x(T) = L_p,
\end{aligned}
$$

$$v(0) = v_0, \qquad v(t) \in [0, \overline{v}(x(t))],$$
$$u(t) \in [\underline{u}, \overline{u}].$$

Because of its particular structure, mostly the time cost and the edge-wise constant speed limit, it is possible to reduce Problem (3) to an ordered sequence of simpler edge minimum-time optimization problems. These have to be solved starting from the first edge and iterating until the end of path p. Let us consider edge $e = e_j$ with length $L := \ell_j > 0$, speed limit $\overline{v} := \overline{v}_j > 0$ and end-point speed limit $\overline{v}_T := \min\left(\overline{v}_j, \overline{v}_{j+1}\right) > 0$. On edge e we have to solve the following optimal control problem:

$$\begin{aligned} &\textit{Minimize} && T && (4)\\ &\textit{subject to} && \dot{x}(t) = v(t), && \dot{v}(t) = u(t) - cv(t),\\ &&& x(0) = 0, && x(T) = L,\\ &&& v(0) = v_0, && v(T) \in [0, \overline{v}_T],\\ &&& v(t) \in [0, \overline{v}], && u(t) \in [\underline{u}, \overline{u}]. \end{aligned}$$

Problem (4) resembles the minimum-time optimal control problem subject to velocity constraints and limited acceleration discussed in [1]. However, an additional constraint is present, that is the final speed constraint. In the following we focus on the solution of Problem (4) for the case $v_0 < \overline{v} > \overline{v}_T$, which is the most crucial case. An analogous derivation for the other cases is straightforward.

As suggested in [1], let us introduce the following auxiliary functions, both for numerical stability and notational clarity:

$$\mathcal{E}(t, w) := \frac{1 - e^{wt}}{w}, \qquad \mathcal{E}_2(t, w) := \frac{e^{wt} - 1 - wt}{w^2}. \tag{5}$$

Then, analogously to [1], we claim there exist two distinct time instants, denoted τ_1 and τ_2 and such that $0 < \tau_1 < \tau_2 < T$, that are switching times for the optimal control, whose expression reads

$$u(t) = \begin{cases} \overline{u}, & 0 < t < \tau_1,\\ c\overline{v}, & \tau_1 < t < \tau_2,\\ \underline{u}, & \tau_2 < t < T, \end{cases} \tag{6}$$

for a.e. $t \in [0, T]$. We like to emphasize that $u : [0, T] \to \mathbb{R}$ is uniquely identified by switching times and final time T. The optimal control (6) consists of an initial pushing phase, up to the maximum allowed speed, a second phase where speed is kept constant at the speed limit until the final braking phase. The structure of (6) resembles a bang-bang control, but it shows an intermediate phase due to the velocity constraint. Problem (4) is transformed into a boundary value

problem (BVP) collecting optimal control (6) and differential-algebraic constraints in (4). The unknowns of this BVP are switching times τ_1 and τ_2 and final time T. We remark that Problem (4) and the BVP are equivalent if and only if control (6) locally minimizes the Hamiltonian function of Problem (4), as claimed above (the proof is left to the reader). By considering the vehicle model and initial conditions in (4) along with the optimal control (6), it is possible to compute the time evolution of vehicle velocity and position, for $t \in [0, T]$, i.e.

$$v(t) = \begin{cases} v_0 e^{-ct} + \overline{u}\mathcal{E}(-t, c), & 0 \le t \le \tau_1, \\ v(\tau_1^-), & \tau_1 \le t \le \tau_2, \\ v(\tau_2^-)e^{-c(t-\tau_2)} + \underline{u}\mathcal{E}(\tau_2 - t, c), & \tau_2 \le t \le T, \end{cases} \tag{7}$$

$$x(t) = \begin{cases} x_0 + v_0\mathcal{E}(-t, c) + \overline{u}\mathcal{E}_2(-t, c), & 0 \le t \le \tau_1, \\ x(\tau_1^-) + v(\tau_1^-)(t - \tau_1), & \tau_1 \le t \le \tau_2, \\ x(\tau_1^-) + v(\tau_1^-)(\tau_2 - \tau_1) + v(\tau_1^-)\mathcal{E}(\tau_2 - t, c) + \underline{u}\mathcal{E}_2(\tau_2 - t, c), & \tau_2 \le t \le T. \end{cases} \tag{8}$$

The analytical solution of this Cauchy problem greatly simplifies the solution of the aforementioned BVP, transforming it into an equivalent non-linear system. This task can be achieved by enforcing boundary conditions and state constraints in (4) to speed profile and trajectory (7)–(8). In particular, the following conditions must be satisfied by the solution of Problem (4):

$$v(\tau_1) = \overline{v}, \qquad v(T) = \overline{v}_T, \qquad x(T) = L. \tag{9}$$

The first makes the pushing phase to stop when the speed limit is reached; similarly, the second constraint means that, at the final time T, the vehicle speed has to be as high as possible, otherwise it would not be a minimum-time speed profile. Finally, the third condition ensures that the final position is reached at the final time T. Conditions (9) can be rewritten by using (7)–(8), yielding the non-linear system $\varphi(z) = 0$, where $z = (\tau_1, \delta, T)^\top$, $\delta := T - \tau_2$, and $\varphi : \mathbb{R}^3 \to \mathbb{R}^3$ is defined by

$$\varphi(z) := \begin{pmatrix} v_0 e^{-c\tau_1} + \overline{u}\mathcal{E}(-\tau_1, c) - \overline{v} \\ \overline{v}e^{-c\delta} + \underline{u}\mathcal{E}(-\delta, c) - \overline{v}_T \\ x_0 + v_0\mathcal{E}(-\tau_1, c) + \overline{u}\mathcal{E}_2(-\tau_1, c) + \overline{v}(T - \delta - \tau_1) + \overline{v}\mathcal{E}(-\delta, c) + \underline{u}\mathcal{E}_2(-\delta, c) - L \end{pmatrix}. \tag{10}$$

It is possible to explicitly write the Jacobian φ' and then to take advantage of it by using Newton-type solvers to find $z^\star$ such that $\varphi(z^\star) = 0$, where

$$\varphi'(z) = \begin{bmatrix} (\overline{u} - cv_0)e^{-c\tau_1} & 0 & 0 \\ 0 & (\underline{u} - c\overline{v})e^{-c\delta} & 0 \\ v_0 e^{-c\tau_1} - \overline{v} + \overline{u}\mathcal{E}(-\tau_1, c) & \overline{v}(e^{-c\delta} - 1) + \underline{u}\mathcal{E}(-\delta, c) & \overline{v} \end{bmatrix}. \tag{11}$$

Non-linear solvers typically require an initial guess. A reasonable and easy-to-compute initial guess can be estimated by considering the limit $c \to 0^+$; in fact, typically the parameter c is small. Let us define $\varphi_0 : \mathbb{R}^3 \to \mathbb{R}^3$, such that $\varphi_0(z) := \lim_{c \to 0^+} \varphi(z)$ for any $z \in \mathbb{R}^3$.

Then, a reasonable initial guess is given by the solution of $\varphi_0(z^\star) = 0$, that is

$$z^\star = \left(\frac{\overline{v} - v_0}{\overline{u}}, \frac{\overline{v}_T - \overline{v}}{\underline{u}}, \frac{L - x_0}{\overline{v}} + \frac{(\overline{v} - v_0)^2}{2\,\overline{v}\,\overline{u}} - \frac{(\overline{v}_T - \overline{v})^2}{2\,\overline{v}\,\underline{u}} \right)^\top . \tag{12}$$

The following Sect. 5 describes how the lower level optimal control problems are coupled with the upper level shortest path problem in Sect. 3.

5 Levels Coupling

The interface between levels, namely UL-OP and LL-OP, plays a key role in the solution process of the bi-level optimization problem. In fact, this crucially affects the exchange of information among levels.

Considering the k-th car, the information flow from UL-OP to LL-OP consists of the ordered sequence of edge lengths and speed limits uniquely identified by the planned path p^k, that is the solution of UL-OP. These values constrain the LL-OP, both as boundary conditions and state constraints.

On the other hand, given optimal speed profiles $v^k(\cdot)$ and a trajectories $x^k(\cdot)$ for every car $k \in C$, an edge cost c^k, compare Sect. 3, has to be defined, based on an estimate of travel time, accounting for possible traffic jam and driver's behavior. Given the solutions to LL-OP for every car, one can reconstruct the number of cars $n_e(t)$ in any edge $e \in E$ as a function of time, $n_e : \mathbb{R}_+ \to \mathbb{R}_+$ with

$$n_e(t) := \text{card}\{k \mid x^k(t) \in e\} \tag{13}$$

(herein, we identified the edge e with its physical distance range for notational simplicity). For any edge $e \in E$, having length $L_e > 0$ and speed limit $\overline{v}_e > 0$, the edge density function $\rho_e : \mathbb{R}_+ \to \mathbb{R}_+$ is defined, such that $\rho_e(t) := n_e(t)/L_e$ for any t. Inspired by the LWR model in [7], that is a first-order PDE-based macroscopic model widely used for traffic flow, let us introduce also the edge speed function $v_e : \mathbb{R}_+ \to \mathbb{R}_+$, such that

$$v_e(t) := \overline{v}_e \left(1 - \frac{\rho_e(t)}{\overline{\rho}_e} \right) \tag{14}$$

for any t, where $\overline{\rho}_e > 0$ is the maximum edge density. Note that the edge speed v_e does not reflect vehicles speed along this edge, but it is just an estimate accounting for traffic jam (v_e is a non-increasing function of n_e and ρ_e). We notice also that for $n_e(t) = 1$, using Eq. (14), the edge speed $v_e(t)$ is lower than the speed limit $\overline{v}_e$,

which is not what we want to achieve. One possible way to fix this inaccuracy is to replace n_e with $\max(n_e - 1, 0)$, in order to make the driver not to interact with itself.

As an edge cost we consider an estimate of the time needed to run across the edge itself. To evaluate this time duration, a representative edge speed value is needed, here denoted by $\hat{v}_e$ and chosen to be

$$\hat{v}_e := (1-\theta)\frac{1}{T_\mathrm{h}}\int_0^{T_\mathrm{h}} v_e(t)\,\mathrm{d}t + \theta \min_{t\in[0,T_\mathrm{h}]} v_e(t) \tag{15}$$

given hyper-parameter $\theta \in [0, 1]$ and time horizon $T_\mathrm{h} > 0$. With this definition it always holds

$$0 \le \min_{t\in[0,T_\mathrm{h}]} v_e(t) \le \hat{v}_e \le \frac{1}{T_\mathrm{h}}\int_0^{T_\mathrm{h}} v_e(t)\,\mathrm{d}t \le \overline{v}_e$$

for any edge speed function v_e, in any edge $e \in E$. The hyper-parameter θ has been introduced to estimate an edge speed $\hat{v}_e$ representative of the predicted evolution of vehicle trajectories and their interactions. Note that this estimate may be really rough and in general it leads to sub-optimal solutions, especially when long edges are present.

The edge cost c_e, for $e \in E$, is expressed in terms of the time needed to travel along edge e, based on estimate $\hat{v}_e$. This cost is defined as the minimum-time run, plus an augmentation of the traffic-related time, to possibly give more importance to congestions, through a parameter $\lambda \ge 0$:

$$c_e := \frac{L_e}{\overline{v}_e} + \lambda\left(\frac{L_e}{\hat{v}_e} - \frac{L_e}{\overline{v}_e}\right). \tag{16}$$

Using (16) in the shortest path problem in Sect. 3 leads to a nonlinear coupling with the lower level problem LL-OP in Sect. 4. This coupling acts in both directions and the resulting bi-level optimization problem is very hard to solve in general. As a first approach towards its solution we propose the iterative procedure in Sect. 2, which results in the following Algorithm 2.

Numerical experiments are documented in the following Sect. 6.

6 Numerical Results

In the previous sections we presented the algorithms for solving the upper and lower level of the proposed bi-level optimization problem, the coupling of those levels, especially the cost function, was discussed in Sect. 5.

First we want to test the overall functionality of the proposed iterative bi-level algorithm. Thereafter, regarding the proposed parameters θ and λ in the cost function, which implies the connection from the lower to the upper level, we want

Algorithm 2: Iterative procedure as a method to solve BOP

Input: Road network $\mathbb{G} = (V, E, c, s, t)$, with cars position, speed and target, $\{s_j, v_j^0, t_j\}_{j \in C}$, parameters $\{c_j, \underline{u}_j, \overline{u}_j\}_{j \in C}$, hyper-parameters $\theta \in [0, 1]$, $\lambda \geq 0$.

$k \leftarrow 0$;
for $e \leftarrow E$ **do**
  $c_e^k \leftarrow L_e / \overline{v}_e$; // edge cost initialization
end
while *not converged* **do**
  for $j \leftarrow C$ **do**
    $p_j^k \leftarrow \texttt{shortestPath}\left(\{c_e^k\}_{e \in E}, s_j, t_j\right)$; // UL-OP
  end
  for $j \leftarrow C$ **do**
    $l_j^k \leftarrow \left\{L_e \mid e \in p_j^k\right\}$; // upper → lower
    $\overline{v}_j^k \leftarrow \left\{\overline{v}_e \mid e \in p_j^k\right\}$;
    $\left(x_j^k, v_j^k, u_j^k\right) \leftarrow \texttt{minTime}\left(l_j^k, \overline{v}_j^k, v_j^0, c_j, \underline{u}_j, \overline{u}_j\right)$; // LL-OP
  end
  for $e \leftarrow E$ **do**
    $c_e^{k+1} \leftarrow \texttt{edgeCost}\left(\{x_j^k\}_{j \in C}, \theta, \lambda\right)$; // lower → upper
  end
  $k \leftarrow k + 1$;
end

to analyze the impingement of these parameters on the numerical results as well as the convergence. Therefore we vary one parameter while fixing the other one and vice versa. Finally we take a closer look at the behaviour of a single car.

6.1 General Evaluation of the Bi-level Algorithm

The algorithms discussed above are implemented in a MATLAB program. For a first test we set the number of cars $n_c = 500$, $\theta = 0.5$ and $\lambda = 1000$, the drag is neglected ($c = 0$). The road network (Fig. 1) is randomly generated on a 2000 × 2000 [m] grid, the connections between the chosen gridpoints are derived through applying a Delaunay triangulation. The limits on the acceleration for the LL-OP is set to $u \in [-3, 2]$ [m/s^2], the maximum velocity therefore is chosen randomly for each car from a set $[10, 20]$ [m/s], as well as the initial speed $v_0 \in [6, 10]$ $[m/s]$, and the number of iterations $N_{iter} = 10$. Figure 2 shows the result of the bi-level algorithm as in the behaviour of the cost function and the evolution of the final time of every car. Considering the cost function, due to the weighing with λ, the meaning of the values is negligible. Nevertheless we notice a reduction in the cost for every car during the first 3 steps. The algorithm converges to different optimal solutions for sets of cars with the same costs. This can be explained such

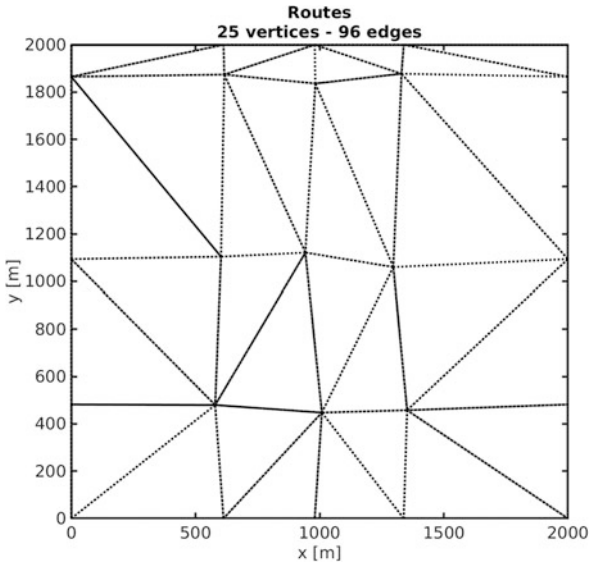


Fig. 1 Randomly generated road network, connections through Delaunay triangulation

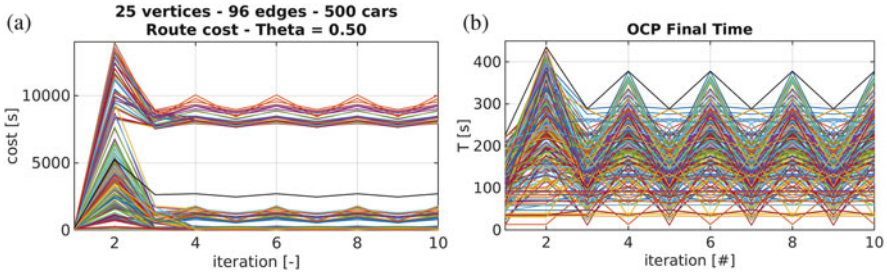


Fig. 2 Numerical tests for the evolution of the cost function and the final time over 10 iterations, $n_c = 500$ cars, $\theta = 0.5$ and $\lambda = 1000$. (**a**) Cost function. (**b**) Final time

that to avoid congestions the algorithm distributes the cars over the road network with respect to keeping the costs low. This leads to sets of vehicles with the same minimal cost to pass from start to their destination. However we also notice that there remains an oscillating behaviour, which seems to resemble two equally good solutions regarding the overall distribution of the vehicles. One solution however yields higher costs. This oscillating characteristic is also mirrored in the final time. In the first three steps the final time decreases. After that, the jumping between two solutions occurs. This oscillating effect was also observed in [3], herein the oscillations might occur between two competing equilibria, respectively Wardrop's equilibria.

Concluding, the algorithm finds optimal paths as well as velocity profiles for every car, while avoiding congestions, through consideration of the vehicle density on every edge which is taken into account as an update on the edge cost in every iteration.

6.2 *Influence of the Parameters θ and λ*

Considering the path planning in the UL-OP, which highly depends on the cost of the edges, the parameters θ and λ, which control the cost function, impact the result of the upper level path planning algorithm. Therefore we compare different parameter settings and analyze their effect on the cost function and the final time. Note that especially in the case of the cost function the values are not directly comparable, due to the different scaling factors. Hence we are more interested on the trend of the cost function itself.

Initially we examine λ, while fixing $\theta = 0.5$. The number of cars $n_c = 400$ is slightly reduced to speed up the computation. The other values remain as they were set in Sect. 6.1. Figure 3 shows the comparison of the progression of the final time and the cost function for $\lambda = 1$ and $\lambda = 1000$ over the iterations.

Comparing the cost profiles, the increase of λ and therefore emphasizing the congestion as an increase in the cost of certain edges, leads to a convergence in the cost function. Thus the algorithm generates bundles of cars with the same cost,

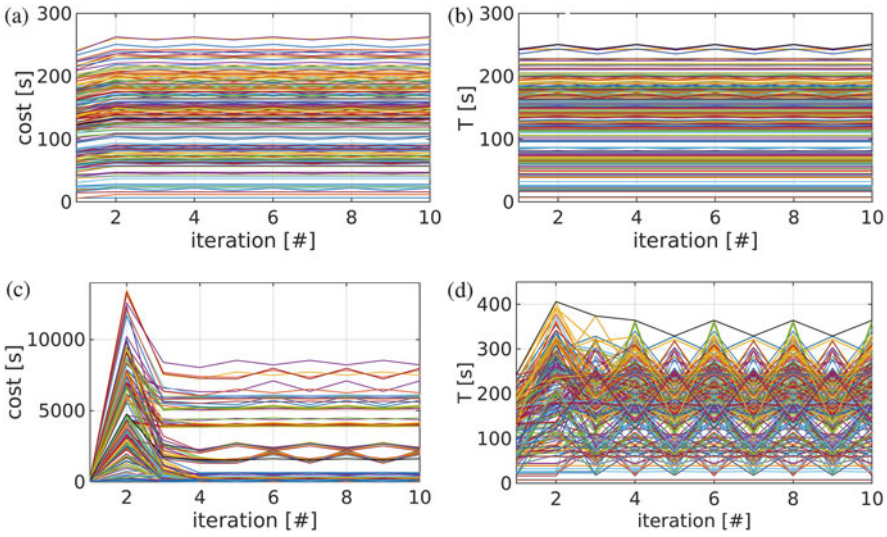


Fig. 3 Implication of the weight factor λ on the evolution of the cost function and the final time for every car, with fixed $\theta = 0.5$, $n_c = 400$, and 25 vertices—96 edges. (**a**) Cost function, $\lambda = 1$. (**b**) Final time, $\lambda = 1$. (**c**) Cost function, $\lambda = 1000$. (**d**) Final Time, $\lambda = 1000$

meaning multiple optima are achieved for such car bundles, and more important with a drastic decrease in the cost. Considering the final time, we notice an increase in the final time along side the increase in λ. For $\lambda = 1$ the cost functions as well as the final time remain almost constant, this is due to the underestimation of the traffic load. The traffic gets almost neglected, since the addition to the density on an edge is in the range of 0.1. Hence the increase in time for $\lambda = 1000$ is justified, since some cars get redirected on longer routes to their destination to avoid congestions. Through the stronger weight the traffic jam becomes emphasized. As a result we can draw the conclusion that a higher weight factor λ is recommended to achieve convergence and for a better distribution of the cars on the network.

Considering the hyper-parameter θ, which influences the estimated representative edge speed $\hat{v}_e$, a higher value of θ shifts the representative edge speed in the direction of the minimum edge velocity, whereas a lower θ emphasizes the mean velocity along the edge over time, see Eq. (15). The influence of θ on the cost function as well as the final time is shown in Fig. 4. Comparing the evolution of the cost function, we notice that the convergence and bundling effect grows with rising θ. However the magnitude of aberrations simultaneously rises, this effect can be countered by introducing additional constraints such that not only the average majority improves while others pay the price for it. Considering the evolution of the final time, the average final time decreases with increasing θ. With $\theta = 1$ the representative edge velocity is given through the minimum velocity value, which represents the worst case. The vehicles velocity on the same edge becomes devalued. This way the algorithm strives for a better distribution of the cars on the network, with the result that the vehicles on average reach their destination faster. We conclude that a higher value, respectively closer to $\theta_{max} = 1$ is recommended.

7 Conclusions

In this paper we presented an iterative algorithm for solving a bi-level optimal control problem. Furthermore we presented a model for a combined single car and network control through density updates and optimal time control. Considering the numerical results we could show that an increase in the hyper-parameters θ, λ affect the optimal solution and emphasize the convergence. The upper level control leads to an optimal distribution of cars among the edges of the network, such that in the lower level OCP an optimal speed profile for each car can be computed with the upper level solution as a constraint. Despite the increase in the final time, which results from longer paths due to a compromise for congestion avoidance, we showed that the density update on the edge cost affects the solution of each car and as a result to bundling of cars with the same cost.

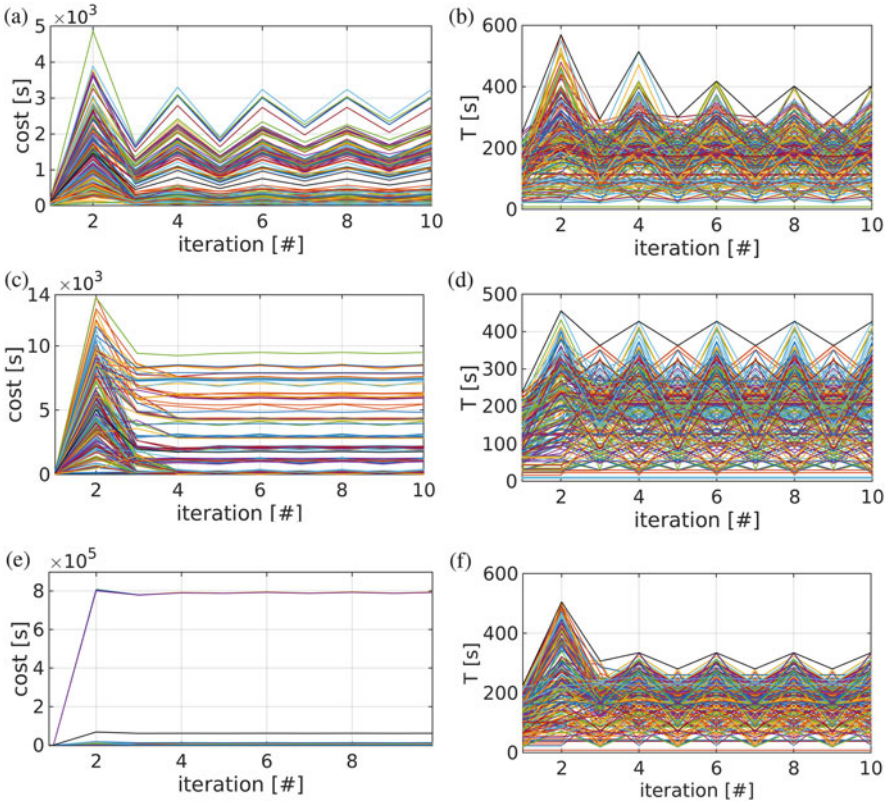


Fig. 4 Implication of the hyper-parameter θ on the evolution of the cost function and the final time for every car, with fixed $\lambda = 100$, $n_c = 400$, and 25 vertices—96 edges. (**a**) Cost function, $\theta = 0$. (**b**) Final time, $\theta = 0$. (**c**) Cost function, $\theta = 0.5$. (**d**) Final time, $\theta = 0.5$. (**e**) Cost function, $\theta = 1$. (**f**) Final time, $\theta = 1$

References

1. Bertolazzi, E., Frego, M.: Semi-analytical minimum time solution for the optimal control of a vehicle subject to limited acceleration. Optimal Control Appl. Methods **39**(2), 774–791 (2018)
2. Bressan, A., Čanić, S., Garavello, M., Herty, M., Piccoli, B.: Flows on networks: recent results and perspectives. EMS Surv. Math. Sci. **1**(1), 47–111 (2014)
3. Cristiani, E., Priuli, F.S.: A destination-preserving model for simulating Wardrop equilibria in traffic flow on networks. Netw. Heterogen Media **10**(4), 857–876 (2015)
4. Cristiani, E., Piccoli, B., Tosin, A.: How can macroscopic models reveal self-organization in traffic flow? In: 2012 IEEE 51st IEEE Conference on Decision and Control (CDC), pp. 6989–6994, 12 (2012)
5. Dijkstra, E.W.: A note on two problems in connexion with graphs. Numer. Math. **1**, 269–271 (1959)
6. Garavello, M., Piccoli, B.: Traffic Flow on Networks. AIMS Series on Applied Mathematics. American Institute of Mathematical Sciences, Springfield, MO (2006)
7. Lighthill, M., Whitham, J.: On kinematic waves. Proc. R. Soc. Lond. **229**(A), 281–345 (1955)

8. Palagachev, K.D., Gerdts, M.: Numerical approaches towards bi-level optimal control problems with scheduling tasks. In: Ghezzi, L., Hömberg, D., Landry, C. (eds.) Math for the Digital Factory. Mathematics in Industry, vol. 27, pp. 205–228. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-63957-4
9. Papadimitriou, C.H., Steiglitz, K.: Combinatorial Optimization: Algorithms and Complexity. Prentice-Hall, Inc., Upper Saddle River, NJ (1982)
10. Richards, P.I.: Shock waves on the highway. Oper. Res. **4**(1), 42–51 (1956)
11. Wardrop, J.G.: Some theoretical aspects of road traffic research. In: Proc. Inst. Civ. Eng. Part II **1**(3), 325–362 (1952)

Computation of Optimal Trajectories for Delay Systems: An Optimize-Then-Discretize Strategy for General-Purpose NLP Solvers

Simone Cacace, Roberto Ferretti, and Zahra Rafiei

Abstract We propose an "optimize-then-discretize" approach for the numerical solution of optimal control problems for systems with delays in both state and control. We first derive the optimality conditions and an explicit representation of the gradient of the cost functional. Then, we use explicit discretizations of the state/costate equations and employ general-purpose Non-Linear Programming (NLP) solvers, in particular Conjugate Gradient or Quasi-Newton schemes, to easily implement a descent method. Finally, we prove convergence of the algorithm to stationary points of the cost, and present some numerical simulations on model problems, including performance evaluation.

Keywords Delay systems · Optimality conditions · Numerical approximation · NLP solvers

1 Introduction

A large class of practical control systems in engineering, chemical processes and economics are modeled in presence of time delays, which introduce in the problem the additional difficulty of an intrinsically infinite-dimensional nature of the state space.

The first extension of the maximum principle to optimal control problems with a constant state delay has been given in Kharatishvili in [16], while a maximum principle for problems with multiple constant delays in state and control variables has been obtained by Halanay in [13]. More recent generalizations have considered the case of time-dependent delays in the state variables [2], and the presence of state

S. Cacace (✉) · R. Ferretti
Dipartimento di Matematica e Fisica, Università Roma Tre, Roma, Italy
e-mail: cacace@mat.uniroma3.it; ferretti@mat.uniroma3.it

Z. Rafiei
Department of Mathematics, Yazd University, Yazd, Iran
e-mail: z.rafiei@stu.yazd.ac.ir

M. Falcone et al. (eds.), *Numerical Methods for Optimal Control Problems*,
Springer INdAM Series 29, https://doi.org/10.1007/978-3-030-01959-4_3

constraints [1] and mixed control-state constraints [11]. Some of these results can be derived via a suitable technique (introduced in [12]) to recast a problem with discrete delays in the form of a non-delayed problem.

To our knowledge, the most complete numerical study of the problem is carried out in [11]. In this work, a "discretize-then-optimize" approach is used to first write the discrete approximation of the control problem, and then solve the Kuhn–Tucker optimality conditions for this latter problem. In this framework, the discretized state equation is treated as a set of equality constraints, which is complemented with the inequality constraints on the control and the state. The result is a set of optimality conditions involving (discretized) state and Kuhn–Tucker multipliers, which parallels the continuous conditions provided by the maximum principle. Among the recent "discretize-then-optimize" literature, we also quote [4], in which minimization is carried out on the discretized problem without making use of optimality conditions.

In this paper, we propose a somewhat different approach. Once simplified the problem by avoiding constraints involving the state, we treat it as a minimization with constraints on the control alone, in which the computation of the gradient (which is the base for Steepest Descent, Conjugate Directions or Quasi-Newton solvers) is carried out via an "optimize-then-discretize" strategy. The discrete gradient is obtained by discretizing the continuous form of the gradient, which is in some sense a byproduct of the maximum principle. A result of convergence for the approximate optimal solutions is also proved.

Note that the use of descent methods for optimal control problems dates back to the 60s [6, 7, 14, 15], whereas the finite difference discretization of delay differential equations has a relatively recent literature (see [3] for an up-to-date review). On the other hand, apart from [11], we are unaware of literature mixing the two techniques to treat optimal control problems for delay systems. Among the various techniques proposed, "optimize-then-discretize" strategies are relatively infrequent; however, within this line of research, it is worth to quote the dissertation [8], along with the discretization of the maximum principle proposed in [5].

The paper is structured as follows. In Sect. 2 the construction of the gradient for the cost functional is recalled. Sections 3 and 4 give a basic form for the discretization and a convergence result for the resulting sequence of approximate optimal controls. Last, Sect. 5 provides some numerical tests, and a performance assessment for the proposed algorithm.

2 Optimal Control Problems with Time Delays and Optimality Conditions

In this section, we present a formal derivation of the optimality conditions for control problems with time delays in both state and control, assuming enough regularity to perform the computations. In particular, we obtain an explicit expression for the gradient of the cost functional of the problem with respect to the control, which will

be useful for its numerical resolution. We refer the reader to [12] and [11] for a detailed and rigorous proof.

We consider the following delayed differential equation

$$\begin{cases} \dot{x}(t) = f(t, x(t), x(t-r), u(t), u(t-s)), & t \in [a, b] \\ x(t) = x_0(t), & t \in [a-r, a], \\ u(t) = u_0(t), & t \in [a-s, a], \end{cases} \tag{1}$$

where $f : [a, b] \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^n$ is the delayed dynamics, with delays r and s respectively in the state and control variables, $x(t) \in \mathbb{R}^n$ is the state and $u(t) \in \mathbb{R}^m$ is the control belonging to the class of admissible controls

$$\mathscr{U} = \{u : [a-s, b] \to \mathbb{U} \subseteq \mathbb{R}^m, u \text{ measurable}\}.$$

Moveover, $x_0(t) \in \mathbb{R}^n$ and $u_0(t) \in \mathbb{R}^m$ are given initial functions.

The optimal control problem consists in minimizing the functional

$$J(u) = \int_a^b (L(t, x(t), x(t-r), u(t), u(t-s))\, dt + g(x(b)), \tag{2}$$

among all the controls $u \in \mathscr{U}$ and subject to the delayed differential equation (1), where $L : [a, b] \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ is the running cost and $g : \mathbb{R}^n \to \mathbb{R}$ is the final cost.

For a generic scalar or vector function z depending on the delayed and undelayed states and controls, we will use in what follows the compact notation $z(t)$ in place of $z(t, x(t), x(t-r), u(t), u(t-s))$, and denote by z_x, z_y, z_u, z_v the partial derivatives of z with respect to the state, the delayed state, the control and the delayed control. To make explicit the dependency of the trajectories on the control, we will possibly denote by $x[u](t)$ the solution of (1).

We summarize below the set of basic assumptions which will be used throughout the paper.

Basic assumptions:

- The functions $f(x, y, u, v)$ and $L(x, y, u, v)$ are twice continuously differentiable with respect to all arguments, and f is Lipschitz continuous w.r.t. x and y;
- The delays r, s satisfy $r, s \geq 0$, $(r, s) \neq (0, 0)$ and $\frac{r}{s} \in \mathbb{Q}$ for $s > 0$ or $\frac{s}{r} \in \mathbb{Q}$ for $r > 0$;
- $g \in C^1(\mathbb{R}^n)$;
- Either $\mathbb{U}$ is a bounded convex set, or L is convex and coercive w.r.t. u and v.

We will now sketch the derivation of a formula for the gradient of J in the case of delayed control systems of the form (1).

First, by computing the variation of $x[u]$ with respect to u along the direction $\varphi : [a-s, b] \to \mathbb{R}^m$ such that $\varphi(t) = 0$ for $t \in [a-s, a]$, we find out that the function

$$\eta[u, \varphi](t) := \lim_{\theta \to 0} \frac{x[u + \theta\varphi](t) - x[u](t)}{\theta}$$

satisfies the following linearized delayed differential equation

$$\begin{cases} \dot{\eta}(t) = f_x(t)\eta(t) + f_y(t)\eta(t-r) + f_u(t)\varphi(t) + f_v(t)\varphi(t-s) & t \in [a, b] \\ \eta(t) = 0 & t \in [a-r, a] \end{cases} \tag{3}$$

We consider the variation of $J(u)$ with respect to u in direction φ: we get

$$< \delta J(u), \varphi > = \int_a^b \left[L_x(t) \cdot \eta(t) + L_y(t) \cdot \eta(t-r) + L_u(t) \cdot \varphi(t) + L_v(t) \cdot \varphi(t-s) \right] dt + g_x(x(b)) \cdot \eta(b) \, .$$

By the change of variable $t \leftarrow t - r$ and using the property $\eta = 0$ in $[a-r, a]$, we easily obtain that

$$\int_a^b L_y(t) \cdot \eta(t-r)\, dt = \int_a^b \chi_{[a,b-r]}(t) L_y(t+r) \cdot \eta(t)\, dt \, ,$$

where $\chi_{[a,b-r]}$ is the characteristic function of the interval $[a, b-r]$. Similarly, by the change of variable $t \leftarrow t - s$ and using the property $\varphi = 0$ in $[a-s, a]$, we obtain

$$\int_a^b L_v(t) \cdot \varphi(t-s)\, dt = \int_a^b \chi_{[a,b-s]}(t) L_v(t+s) \cdot \varphi(t)\, dt \, .$$

Then, we have

$$\begin{aligned} < \delta J(u), \varphi > = & \int_a^b \left[L_x(t) + \chi_{[a,b-r]}(t) L_y(t+r) \right] \cdot \eta(t) dt \\ & + \int_a^b \left[L_u(t) + \chi_{[a,b-s]}(t) L_v(t+s) \right] \cdot \varphi(t)\, dt + g_x(x(b)) \cdot \eta(b) \, . \end{aligned} \tag{4}$$

Now, we introduce the following adjoint equation with a final condition

$$\begin{cases} \dot{\lambda}^T(t) = -\left[L_x(t) + \lambda^T(t) f_x(t)\right] \\ \qquad\qquad -\chi_{[a,b-r]}(t)\left[L_y(t+r) + \lambda^T(t+r) f_y(t+r)\right], \\ \lambda^T(b) = g_x(x(b)). \end{cases} \quad t \in [a,b] \tag{5}$$

Note that, compared to the state equation, the adjoint equation is still delayed, but backward in time with negative delay $-r$.

We employ the adjoint equation to write the first integral in (4) in terms of λ: we get

$$\begin{aligned} < \delta J(u), \varphi > = & \int_a^b \left[-\dot{\lambda}^T(t) - \lambda(t)^T f_x(t) - \chi_{[a,b-r]}(t)\lambda^T(t+r) f_y(t+r)\right] \cdot \eta(t) dt \\ & + \int_a^b \left[L_u(t) + \chi_{[a,b-s]}(t) L_v(t+s)\right] \cdot \varphi(t) dt + g_x(x(b)) \cdot \eta(b). \end{aligned}$$

Integrating by parts and using the delayed equation (3) for η, it follows that

$$\begin{aligned} < \delta J(u), \varphi > = & -\lambda^T(b) \cdot \eta(b) + \int_a^b \lambda^T(t) \left[\dot{\eta}(t) - f_x(t)\eta(t)\right] dt \\ & - \int_a^b \chi_{[a,b-r]}(t)\lambda^T(t+r) f_y(t+r)\eta(t) dt \\ & + \int_a^b \left[L_u(t) + \chi_{[a,b-s]}(t) L_v(t+s)\right] \cdot \varphi(t) dt + g_x(x(b)) \cdot \eta(b) \\ = & -g_x(x(b)) \cdot \eta(b) \\ & + \int_a^b \lambda^T(t) \left[f_u(t)\varphi(t) + f_y(t)\eta(t-r) + f_v(t)\varphi(t-s)\right] dt \\ & - \int_a^b \chi_{[a,b-r]}(t)\lambda^T(t+r) f_y(t+r)\eta(t) dt \\ & + \int_a^b \left[L_u(t) + \chi_{[a,b-s]}(t) L_v(t+s)\right] \cdot \varphi(t) dt + g_x(x(b)) \cdot \eta(b). \end{aligned}$$

Again, by changing variables we obtain

$$\int_a^b \lambda^T(t) f_y(t)\eta(t-r) dt = \int_a^b \chi_{[a,b-r]}(t)\lambda^T(t+r) f_y(t+r)\eta(t) dt,$$

$$\int_a^b \lambda^T(t) f_v(t)\varphi(t-s) dt = \int_a^b \chi_{[a,b-s]}(t)\lambda^T(t+s) f_v(t+s)\varphi(t) dt,$$

which gives

$$< \delta J(u), \varphi > = \int_a^b \Big[\Big(L_u(t) + \lambda^T(t) f_u(t) \Big) + \chi_{[a,b-s]}(t) \Big(L_v(t+s) + \lambda^T(t+s) f_v(t+s) \Big) \Big] \cdot \varphi(t) dt.$$

Last, introducing the Hamiltonian

$$H(t, x, y, u, v, \lambda) = L(t, x, y, u, v) + \lambda^T f(t, x, y, u, v),$$

we end up with

$$< \delta J(u), \varphi >= \int_a^b \big[H_u(t) + \chi_{[a,b-s]}(t) H_v(t+s) \big] \cdot \varphi(t) dt. \tag{6}$$

Then, a control-constrained local minimum u^* satisfies, for any $u \in \mathscr{U}$, the well-known condition

$$< \delta J(u^*), u - u^* > \geq 0, \tag{7}$$

which reduces to $\delta J(u^*) = 0$ if the control is unconstrained ($\mathbb{U} = \mathbb{R}^m$).

A more general and rigorous version of the above computations leads to the following Pontryagin-type optimality conditions:

Theorem 1 ([11]) *Let the basic assumptions hold, and let $(\hat{x}, \hat{u})$ be locally optimal for the functional J in (2) subject to the delayed differential equation (1). Then, there exists an adjoint state function $\hat{\lambda} \in W^{1,\infty}([a, b], \mathbb{R}^n)$ such that the following conditions hold a.e. for $t \in [a, b]$:*

1. *Adjoint equation:*

$$\dot{\hat{\lambda}}^T(t) = -\hat{H}_x(t) - \chi_{[a,b-r]}(t) \hat{H}_y(t+r)$$

   where $\hat{H}_x$, $\hat{H}_y$ denote the partial derivatives of H computed on the optimal triple $(\hat{x}, \hat{u}, \hat{\lambda})$;
2. *Transversality:*

$$\hat{\lambda}^T(b) = g_x(\hat{x}(b));$$

3. *Minimum condition for Hamiltonian: for any $u \in \mathbb{U}$,*

$$\hat{H}(t) + \chi_{[a,b-s]}(t) \hat{H}(t+s) = \hat{H}(t, \hat{x}(t), \hat{x}(t-r), \hat{u}(t), \hat{u}(t-s), \hat{\lambda}(t)) + \chi_{[a,b-s]}(t) H(t+s, \hat{x}(t+s), \hat{x}(t+s-r),$$

$$\begin{aligned}
&\hat{u}(t+s-r), \hat{u}(t+s), \hat{u}(t), \hat{\lambda}(t+s)) \\
\leq\ &\hat{H}(t, \hat{x}(t), \hat{x}(t-r), u, \hat{u}(t-s), \hat{\lambda}(t)) \\
&+\chi_{[a,b-s]}(t) H(t+s, \hat{x}(t+s), \hat{x}(t+s-r), \\
&\hat{u}(t+s-r), \hat{u}(t+s), u, \hat{\lambda}(t+s)). \qquad (8)
\end{aligned}$$

Following [11], we remark that the rationality assumption on the delays r and s is crucial to handle the general lack of regularity of the solution to the delayed equation. Indeed, under this assumption, it is possible to make a partition of the time interval $[a, b]$ in an integer number of sub-intervals, and define suitable restrictions of the state and the control, such that the delayed equation can be recast as a system of ordinary (non-delayed) equations (see [12]). Accordingly, the optimal control problem with delays is transformed into a standard optimal control problem in higher dimension, in which the continuity of the unknowns at the end points of the sub-intervals is imposed as an additional constraint. The optimality conditions for the delayed case in Theorem 1 then follow, by applying standard optimality conditions and rebuilding the solution on the whole interval by merging the optimal solutions on the sub-intervals. We remark that an even more general version of Pontryagin's Maximum Principle for delayed control systems has been recently given in [18].

3 Discretization

Two basic strategies are available for solving numerically the optimal control problem under consideration. One is the so called "discretize-then-optimize" procedure, which first discretizes both the functional J in (2) and the state equation (1), then uses some NLP solver to obtain, working on the discrete problem, an approximation of the optimal triple $(\hat{x}, \hat{u}, \hat{\lambda})$.

Alternatively, one can an apply an "optimize-then-discretize" procedure, namely, first discretize the optimality conditions and then search for an approximation of the optimal triple. Again, a general solver can be employed.

Here, we still follow an "optimize-then-discretize" approach, but we do not solve the whole optimality system. We rather discretize the state and adjoint equations, then take advantage of the explicit expression of the gradient of J derived in the previous section. This allows one to build a descent method which is easy to implement. Indeed, it is enough to use a NLP solver only for the optimization in the control variables, passing iteratively the values of the functional J and its gradient depending on the updated state and costate. Note that, with respect to other techniques, we are not considering the state equation as a set of equality constraints, and the minimization is carried out with respect to the discrete control variables alone, thus reducing the dimension of the problem and avoiding equality constraints. Inequality constraints of the form $u(t) \in \mathbb{U}$ can easily be handled by a descent

method, at least in the case $\mathbb{U}$ is convex, as assumed here, and in particular for the typical case of box constraints.

As a start, we construct a uniform grid on the interval $[a, b]$, with $N + 1$ nodes $t_i = a + ih$, for $i = 0, \dots, N$ and $h = (b - a)/N$, and assume that the delays can be written as $r = kh$ and $s = lh$ for some integers k, l. For $i = 0, \dots, N$, we denote by $X_i \approx x(t_i) \in \mathbb{R}^n$, $U_i \approx u(t_i) \in \mathbb{R}^m$ and $\Lambda_i \approx \lambda(t_i) \in \mathbb{R}^n$ the approximations at grid points of respectively the state, control and adjoint variables. Via negative indices, the initial conditions are defined as $X_i = x_0(t_i) \in \mathbb{R}^n$ for $i = -k, \dots, 0$ and $U_i = u_0(t_i) \in \mathbb{R}^m$ for $i = -l, \dots, -1$.

In the simplest setting, the state equation can be discretized by means of the forward Euler scheme [3]:

$$\begin{cases} X_{i+1} = X_i + hf(t_i, X_i, X_{i-k}, U_i, U_{i-l}) & i = 0, \dots, N-1 \\ X_i = x_0(t_i) & i = -k, \dots, 0 \\ U_i = u_0(t_i) & i = -l, \dots, -1. \end{cases} \tag{9}$$

The adjoint equation can also be discretized using the Euler scheme with a negative step $-h$:

$$\begin{cases} \Lambda_i = \Lambda_{i+1} + h\,[H_x(t_i, X_i, X_{i-k}, U_i, U_{i-l}, \Lambda_i) \\ \qquad\quad + \chi_{[a,b-r]}(t_i) H_y(t_i + r, X_{i+k}, X_i, U_{i+k}, U_{i-l+k}, \Lambda_{i+k})\big] \\ \Lambda_N = g_x(X_N). \end{cases} \qquad i = N-1, \dots, 0 \tag{10}$$

Accordingly, a rectangle quadrature rule can be used to discretize both the functional J in (2) and its variation (6):

$$J(u) \approx J^h(U, X) := h \sum_{i=0}^{N-1} L(t_i, X_i, X_{i-k}, U_i, U_{i-l}) + g(X_N), \tag{11}$$

$$\begin{aligned} <\delta J(u), \varphi> \approx (J^h_u(U, X, \Lambda), \varphi)_h := h \sum_{i=0}^{N-1} [H_u(t_i, X_i, X_{i-k}, U_i, U_{i-l}, \Lambda_i) \\ + \chi_{[a,b-s]}(t_i) H_v(t_{i+s}, X_{i+l}, X_{i-k+l}, U_{i+l}, U_i, \Lambda_{i+l})\big]\varphi(t_i)\ . \end{aligned} \tag{12}$$

We have denoted here by $U \in \mathbb{R}^{mN}$ a vector collecting all discretized control variables, in the form

$$U = \begin{pmatrix} U_0 \\ \vdots \\ U_{N-1} \end{pmatrix}.$$

By the arbitrariness of φ we get, for $i = 0, \ldots, N-1$, the components of J_u^h as (row) vectors in $\mathbb{R}^m$, that we denote by J_u^h:

$$J_u^h(U, X, \Lambda)_i = h\,[H_u(t_i, X_i, X_{i-k}, U_i, U_{i-l}, \Lambda_i) + \chi_{[a,b-s]}(t_i) H_v(t_i + s, X_{i+l}, X_{i-k+l}, U_{i+l}, U_i, \Lambda_{i+l})]. \tag{13}$$

We will use in what follows the shorthand notations $J^h(U)$ and $J_u^h(U)$ whenever we do not need to stress the dependence of J^h and J_u^h on the state and the costate.

With this basic discretization, the same result could be obtained by a "discretize-then-optimize" strategy. However, (9)–(13) can also be discretized with more accurate as well as heterogeneous techniques, which might not lead to the same endpoint. In particular, we quote here the higher-order schemes used in [10], as well as the symplectic solvers described in [9, Section 5.3.2].

We finally choose our favourite NLP solver and implement the following iterative Algorithm 1.

Algorithm 1 Minimization algorithm

1: Assign an initial guess $U^{(0)}$, a tolerance ε, an integer $k_{\max}$ and set $k = 0$
2: **repeat**
3: Compute $X^{(k)}$ using the forward scheme (9)
4: Compute $\Lambda^{(k)}$ using the backward scheme (10)
5: Compute $J^h(U^{(k)}, X^{(k)})$ and $J_u^h(U^{(k)}, X^{(k)}, \Lambda^{(k)})$ using (11) and (13)
6: Update $U^{(k)}$ using the NLP solver with J^h, J_u^h and set $k = k + 1$
7: **until** $\left|J^h(U^{(k)}, X^{(k)}) - J^h(U^{(k-1)}, X^{(k-1)})\right| < \varepsilon$ or $k > k_{\max}$
8: Set $U^* = U^{(k)}$

Note that any constraint on the control (typically, a box constraint) can be enforced in Step 6. Here, convexity of $\mathbb{U}$ plays a crucial role, and inequality constraints could be treated, for example, by projection. On the other hand, handling nonconvex constraints would lead to more complex algorithms. Note also that the stopping condition

$$\left|J^h(U^{(k)}, X^{(k)}) - J^h(U^{(k-1)}, X^{(k-1)})\right| < \varepsilon$$

is applicable to both constrained and unconstrained problems.

4 Convergence

We prove now that the sequence of approximate optimal controls obtained via the previous procedure is a (locally) minimizing sequence for the exact cost functional. To this end, we define the piecewise constant counterpart of the discrete control U

as

$$u_*^h(t) = \sum_{i=0}^{N-1} U_i^* \chi_{[t_i, t_{i+1})}(t). \tag{14}$$

We can prove the following convergence result for the approximate optimal controls in the form (14). In addition to the basic assumptions, we assume that the set $\mathbb{U}$ is bounded and that the discrete optimal control U^* satisfies the constrained stationary point condition

$$(J_u^h(U^*), U - U^*)_h \geq 0 \tag{15}$$

for any $U \in \mathbb{U}^N$, where, according to (12),

$$\begin{aligned}(J_u^h(U^*), U - U^*)_h &= \sum_{i=0}^{N-1} J_u^h(U^*)_i (U_i - U_i^*) \\ &= h \sum_{i=0}^{N-1} \big[H_u(t_i, X_i, X_{i-k}, U_i^*, U_{i-l}^*, \Lambda_i) \\ &\quad + \chi_{[a,b-s]}(t_i) H_v(t_i + s, X_{i+l}, X_{i-k+l}, U_{i+l}^*, U_i^*, \Lambda_{i+l}) \big] (U_i - U_i^*).\end{aligned}$$

Theorem 2 *Let the basic assumptions hold, with a bounded set of control values $\mathbb{U}$. If U^* satisfies (15), then, for some constant C independent of h and any $u \in \mathscr{U}$:*

$$< \delta J(u_*^h), u - u_*^h > \geq -Ch, \tag{16}$$

and any convergent subsequence of u_*^h converges to a constrained stationary point of J as $h \to 0$.

Proof It is clear that, if a (sub)sequence $u_*^{h_k}$ of approximate optimal controls converges as $h \to 0$, then (16) implies the constrained minimum condition (7) in the limit. Therefore, we only need to prove (16). Throughout the various steps of the proof, we will use the symbol C to denote a constant, which might not be the same at any occurrence.

Step 1—uniform error estimate on x and λ First, note that the assumption that both delays r and s are multiples of h implies that, since u_*^h is constant on each time interval $[t_i, t_{i+1}]$, the related state $x_*^h(t) = x[u_*^h](t)$ is piecewise smooth. Moreover, the L^∞ bound on the control implies also a uniform bound on the state. In these conditions, the numerical error introduced in a single step is uniformly bounded by an $\mathscr{O}(h^2)$ and therefore, via the stability of the Euler scheme, for all admissible piecewise constant controls of the form (14) the global error satisfies

$$\left\| x_*^h(t_i) - X_i^* \right\| \leq Ch, \tag{17}$$

with X^* denoting the discrete evolutions associated to U^*, and $C = C(a, b, X_0, \mathbb{U})$. Consider now the costate equation (5) and its discretization (10). The costate λ_*^h associated to u_*^h and x_*^h is again smooth on each interval $[t_i, t_{i+1}]$, and by the same arguments as above, we get

$$\left\| \lambda_*^h(t_i) - \Lambda_i^* \right\| \leq Ch, \tag{18}$$

with Λ^* defined accordingly.

Step 2—error estimate on the directional derivative at u_*^h First, we introduce the shorthand notation

$$\begin{aligned} H_u^{*,h}(t_i) + \chi_{[a,b-s]}(t_i) H_v^{*,h}(t_i + s) := H_u(t_i, X_i^*, X_{i-k}^*, U_i^*, U_{i-l}^*, \Lambda_i^*) \\ + \chi_{[a,b-s]}(t_i) H_v(t_i + s, X_{i+l}^*, X_{i-k+l}^*, U_{i+l}^*, U_i^*, \Lambda_{i+l}^*), \end{aligned}$$

and denote by $H^*(t)$ the Hamiltonian computed on the triple $(x_*^h, \lambda_*^h, u_*^h)$ at the time t. By (17)–(18), and taking into account that the control arguments coincide, we can give for $t \in [t_i, t_{i+1}]$ the bound

$$\left| H_u^*(t) + \chi_{[a,b-s]}(t_i) H_v^*(t + s) - H_u^{*,h}(t_i) - \chi_{[a,b-s]}(t_i) H_v^{*,h}(t_i + s) \right| \leq Ch. \tag{19}$$

Note that, since U is arbitrary, (15) holds if and only if each of the terms in (15) satisfies

$$J_u^h(U^*)_i (U_i - U_i^*) = \left[H_u^{*,h}(t_i) + \chi_{[a,b-s]}(t_i) H_v^{*,h}(t_i + s) \right] (U_i - U_i^*) \geq 0 \tag{20}$$

for all $U_i \in \mathbb{U}$ and $i \in [0, N-1]$. Therefore, we have:

$$\begin{aligned} < \delta J(u_*^h), u - u_*^h > &= \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \left[H_u^*(t) + \chi_{[a,b-s]}(t_i) H_v^*(t + s) \right] (u(t) - U_i^*) dt \\ &= \mathcal{O}(h) + h \sum_{i=0}^{N-1} \left[H_u^{*,h}(t_i) + \chi_{[a,b-s]}(t_i) H_v^{*,h}(t_i + s) \right] \frac{1}{h} \int_{t_i}^{t_{i+1}} (u(t) - U_i^*) dt \\ &= \mathcal{O}(h) + h \sum_{i=0}^{N-1} \left[H_u^{*,h}(t_i) + \chi_{[a,b-s]}(t_i) H_v^{*,h}(t_i + s) \right] (U_i - U_i^*). \end{aligned} \tag{21}$$

In (21), the vector U_i denotes the integral mean

$$U_i = \frac{1}{h} \int_{t_i}^{t_{i+1}} u(t) dt$$

(clearly, the integral mean of the constant vector U_i^* is the vector itself), and the $\mathscr{O}(h)$ term is uniformly bounded as

$$|\mathscr{O}(h)| \leq C \operatorname{diam}(\mathbb{U})(b-a)h,$$

C being the same constant appearing in (19). Note that the convexity of $\mathbb{U}$ implies that $U_i \in \mathbb{U}$.

Finally, taking into account (20), (21) is equivalent to (16). □

We finally remark that, in case of higher order approximations of (9)–(13), the final rate of convergence turns out to be the lowest among the convergence rates of all discretizations. However, an increase of the accuracy would also require to define a higher order construction for the approximate optimal control (14) (e.g., a piecewise polynomial form). In this case, neither the uniform error estimate on the evolution nor the enforcement of the constraint $u_*^h \in \mathbb{U}$ are obvious, and a rigorous proof would need more technical arguments.

5 Numerical Simulations

In this section we solve numerically some model delayed optimal control problems taken from [11]. For performance comparison, but also to show that the proposed method can be easily implemented as a black box, we employ three different NLP solvers, namely Steepest-Descent with fixed step β (SD[β]), Limited-memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS) with 10 updates of the approximate Hessian, and Conjugate Gradient (CG), all with the same stopping tolerance $\varepsilon = 10^{-12}$. Looking at the convergence tables, it is apparent that the number of iteration of the various solvers has a weak dependence, if any at all, on the discretization steps.

For all the examples but the last one, the results have been compared with reference solutions obtained with an extremely fine discretization, and the rate of convergence of the optimal values as been computed. In all cases, we have obtained a first-order convergence, which confirms the theoretical analysis.

All tests were performed on a Lenovo Ultrabook X1 Carbon, using 1 CPU Intel Quad-Core i5-4300U 1.90 Ghz with 8 Gb Ram, running under the Linux Slackware 14.1 operating system. The algorithm is written in C++ and employs the LBFGS and CG solvers implemented in the *dlib C++ library* (www.dlib.net).

Test 1

Let $x : [0, 5] \to \mathbb{R}$ and $u : [0, 5] \to \mathbb{R}$. We want to minimize the quadratic cost

$$J = \int_0^5 \left(x^2(t) + u^2(t)\right) dt\,,$$

subject to

$$\begin{cases} \dot{x}(t) = x(t-2) + u(t-1) & t \in [0,5] \\ x(t) = 1 & t \in [-2,0] \\ u(t) = 0 & t \in [-1,0]. \end{cases}$$

The Hamiltonian is given by

$$H(x, y, u, v, \lambda) = x^2 + u^2 + \lambda(y + v)\,.$$

We get the following adjoint equation

$$\begin{cases} \dot{\lambda}(t) = -2x(t) - \chi_{[0,3]}(t)\lambda(t+2) & t \in [0,5] \\ \lambda(5) = 0 \end{cases}$$

and gradient

$$< \delta J(u), \varphi >= \int_0^5 \left(2u(t) + \chi_{[0,4]}(t)\lambda(t+1)\right)\varphi(t)dt\,.$$

Figure 1 shows the optimal solution for $N = 500$. In Table 1 we report the results obtained by the three solvers under grid refinement, including the number of iterations to reach convergence, the corresponding CPU times, the error and the rate of convergence with respect to a reference solution computed on a fine grid of 10^6 nodes, attaining the optimal value $J^* = 26.98748461$.

All the solvers compute the same solution up to the tolerance ε. We see that in this case LBFGS and CG are comparable in terms of iterations, but LBFGS performs better in terms of computational time. On the other hand, SD[0.01] requires a large number of iterations due to the restriction on the step size. This step is tuned by trial and error, and larger values would result in a lack of convergence.

Test 2
Let $x : [0,3] \to \mathbb{R}$ and $u : [0,3] \to \mathbb{R}$. We want to minimize

$$J = \int_0^3 \left(x^2(t) + u^2(t)\right) dt\,,$$

subject to

$$\begin{cases} \dot{x}(t) = x(t-1)u(t-2) & t \in [0,3] \\ x(t) = 1 & t \in [-1,0] \\ u(t) = 0 & t \in [-2,0]. \end{cases}$$

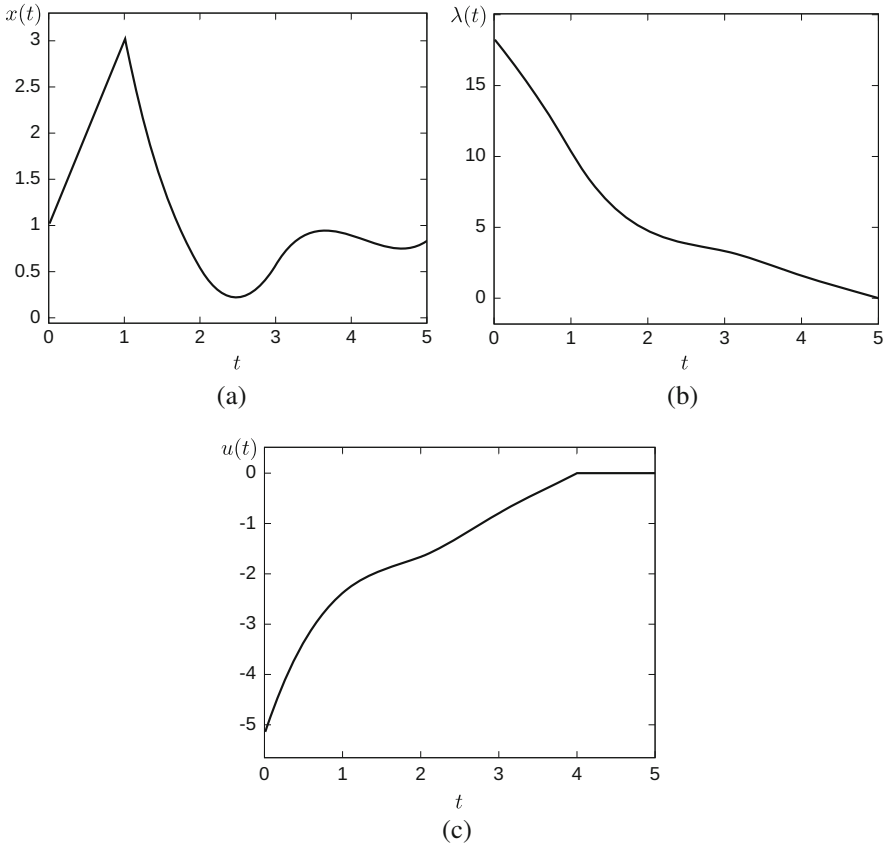


Fig. 1 Optimal solution for Test 1. (**a**) State x, (**b**) costate λ, (**c**) control u

The Hamiltonian is given by

$$H(x, y, u, v, \lambda) = x^2 + u^2 + \lambda y v\,.$$

We get the following adjoint equation

$$\begin{cases} \dot{\lambda}(t) = -2x(t) - \chi_{[0,2]}(t)\lambda(t+1)u(t-1) \quad t \in [0,3] \\ \lambda(3) = 0, \end{cases}$$

and gradient

$$< \delta J(u), \varphi >= \int_0^3 \left(2u(t) + \chi_{[0,1]}(t)\lambda(t+2)x(t+1)\right)\varphi(t)dt\,.$$

Table 1 Performance of the solvers for Test 1

N	h	CPU SD[0.01]	CPU LBFGS	CPU CG	Its SD[0.01]	Its LBFGS	Its CG	$\lvert J - J_0 \rvert$	Rate
500	0.01	0.008	0.0006	0.003	852	10	11	0.381	-
1000	0.005	0.019	0.001	0.006	850	10	9	0.190	1.004
2000	2.5e-03	0.036	0.002	0.011	849	11	12	0.094	1.003
4000	1.25e-03	0.07	0.0035	0.020	851	10	10	0.047	1.004
8000	6.25e-04	0.13	0.0064	0.09	845	10	16	0.023	1.006
16,000	3.125e-04	0.29	0.018	0.15	848	12	14	0.011	1.012

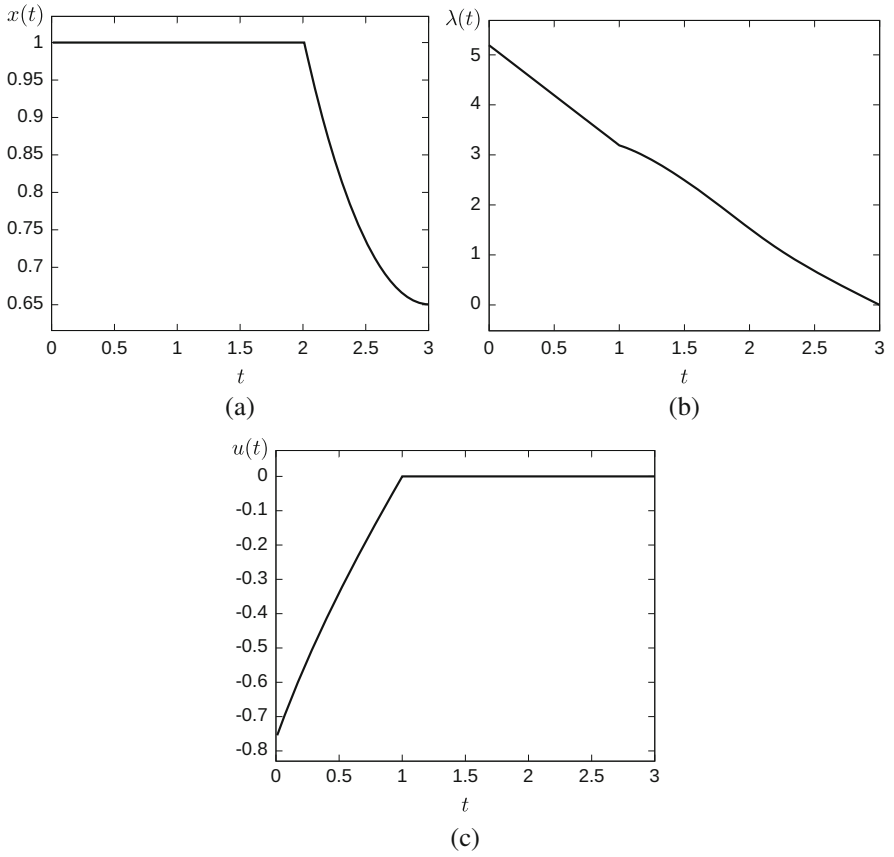


Fig. 2 Optimal solution for Test 2. (**a**) State x, (**b**) costate λ, (**c**) control u

Figure 2 shows the optimal solution for $N = 600$, whereas in Table 2 we report the results obtained by the three solvers. In this case the reference solution is computed on a fine grid of 1.5×10^6 nodes, attaining a value $J^* = 2.76159073$, which agrees up to the 6th significant digit with the analytical solution obtained in [11].

LBFGS is still the faster solver, both in terms of iterations and CPU times. On the other hand, in this case SD[0.1] performs better than CG in computational time, due to both a fortunate choice for the fixed step size, and the low complexity of a single iteration.

Test 3

Let $x : [0, 0.2] \to \mathbb{R}^3$ and $u : [0, 0.2] \to \mathbb{R}^2$. We want to minimize

$$J = \int_0^{0.2} \left(\|x(t)\|^2 + 0.01 \|u(t)\|^2 \right) dt \,,$$

Table 2 Performance of the solvers for Test 2

N	h	CPU SD[0.1]	CPU LBFGS	CPU CG	Its SD[0.1]	Its LBFGS	Its CG	$\lvert J - J_0 \rvert$	Rate
300	0.01	0.0007	0.0003	0.0018	87	6	7	0.0129	-
600	0.005	0.0013	0.0005	0.0035	87	6	6	0.0064	0.999
1200	2.5e-03	0.0027	0.0010	0.0058	88	6	6	0.0032	0.999
2400	1.25e-03	0.0053	0.0022	0.0105	88	6	6	0.0016	0.998
4800	6.25e-04	0.0108	0.0038	0.0298	88	6	7	0.0008	0.996
9600	3.125e-04	0.0228	0.0096	0.0780	87	7	6	0.0004	0.993
19,200	1.5625e-04	0.0449	0.0233	0.1283	88	7	5	0.0002	0.987

subject to

$$\begin{cases} \dot{x_1}(t) = -x_1(t) - R(x_1(t), x_2(t), x_3(t)) & t \in [0, 0.2] \\ \dot{x_2}(t) = -x_2(t) + 0.9u_2(t - 0.02) + 0.1u_2(t) & t \in [0, 0.2] \\ \dot{x_3}(t) = -2x_3(t) + 0.25R(x_1(t), x_2(t), x_3(t)) - 1.05u_1(t)x_3(t - 0.015) & t \in [0, 0.2] \\ x_1(0) = 0.49 \\ x_2(0) = -0.0002 \\ x_3(t) = -0.02 & t \in [-0.015, 0] \\ u_2(t) = 1 & t \in [-0.02, 0], \end{cases}$$

where

$$R(x_1, x_2, x_3) = (1 + x_1)(1 + x_2) \exp\left(\frac{25x_3}{1 + x_3}\right).$$

The Hamiltonian is given by

$$\begin{aligned} H(x_1, x_2, x_3, y_3, u_1, u_2, v_2, \lambda_1, \lambda_2, \lambda_3) = {} & x_1^2 + x_2^2 + x_3^2 + 0.01u_1^2 + 0.01u_2^2 \\ & + \lambda_1(-x_1 - R(x_1, x_2, x_3)) \\ & + \lambda_2(-x_2 + 0.9v_2 + 0.1u_2) \\ & + \lambda_3(-2x_3 + 0.25R(x_1, x_2, x_3) - 1.05u_1 y_3). \end{aligned}$$

We get the following adjoint system

$$\begin{cases} \dot{\lambda}_1(t) = -2x_1(t) + \lambda_1(t) - (-\lambda_1(t) + 0.25\lambda_3(t))R_{x_1}(x_1(t), x_2(t), x_3(t)) & t \in [0, 3] \\ \dot{\lambda}_2(t) = -2x_2(t) + \lambda_2(t) - (-\lambda_1(t) + 0.25\lambda_3(t))R_{x_2}(x_1(t), x_2(t), x_3(t)) & t \in [0, 3] \\ \dot{\lambda}_3(t) = -2x_3(t) + 2\lambda_3(t) - (-\lambda_1(t) + 0.25\lambda_3(t))R_{x_3}(x_1(t), x_2(t), x_3(t)) \\ \qquad +1.05\chi_{[0,0.185]}(t)\lambda_3(t + 0.015)u_1(t + 0.015) & t \in [0, 3] \\ \lambda_1(0.2) = 0 \\ \lambda_2(0.2) = 0 \\ \lambda_3(0.2) = 0, \end{cases}$$

and the gradient reads

$$\begin{aligned} < \delta J(u), \varphi > = {} & \int_0^{0.2} (0.02u_1(t) - 1.05\lambda_3(t)x_3(t - 0.015))\,\varphi_1(t)dt \\ & + \int_0^{0.2} (0.02u_2(t) + 0.1\lambda_2(t) \\ & + 0.9\chi_{[0,0.18]}(t)\lambda_2(t + 0.02))\varphi_2(t)dt. \end{aligned}$$

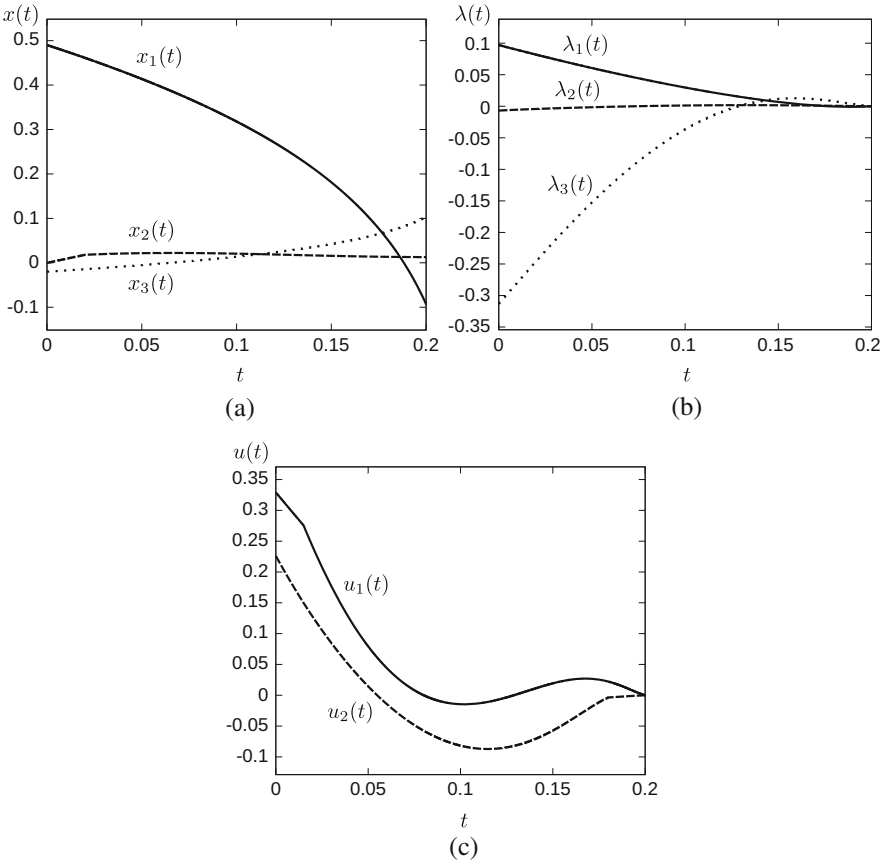


Fig. 3 Optimal solution for Test 3. (**a**) State (x_1, x_2, x_3), (**b**) costate $(\lambda_1, \lambda_2, \lambda_3)$, (**c**) control (u_1, u_2)

Figure 3 shows the optimal solution for $N = 400$. In Table 3 we report the results obtained by the three solvers. The reference solution for this test is computed on a fine grid of 10^6 nodes, attaining an optimal value $J^* = 0.02133289$.

Also this test confirms that LBFGS is the faster solver, despite the larger number of iterations with respect to CG. On the other hand, SD with fixed step size is relatively slow, but still remains the easiest method to implement without using third-party libraries.

Test 4

We consider a more challenging version of Test 3, namely to minimize

$$J = \int_0^{0.2} \left(\|x(t)\|^2 + 0.01 u_2(t)^2 \right) dt \,,$$

Table 3 Performance of the solvers for Test 3

N	h	CPU SD[1]	CPU LBFGS	CPU CG	Its SD[1]	Its LBFGS	Its CG	$\lvert J - J_0 \rvert$	Rate
400	5e − 04	0.05	0.007	0.03	490	15	6	8.93e − 05	-
800	2.5e − 04	0.08	0.01	0.06	490	17	6	4.46e − 05	1.001
1600	1.25e − 04	0.17	0.03	0.12	490	19	6	2.22e − 05	1.002
3200	6.25e − 05	0.35	0.07	0.19	490	18	4	1.11e − 05	1.004
6400	3.125e − 05	0.70	0.17	0.48	490	25	5	5.51e − 06	1.009
12,800	1.5625e − 05	1.41	0.36	1.01	490	26	5	2.71e − 06	1.019
25,600	7.8125e − 06	2.81	0.74	2.15	490	28	4	1.32e − 06	1.039
40,000	5e − 06	4.32	1.29	2.53	490	29	4	8.19e − 07	1.071

subject to

$$\begin{cases} \dot{x_1}(t) = -x_1(t) - R(x_1(t), x_2(t), x_3(t)) & t \in [0, 0.2] \\ \dot{x_2}(t) = -x_2(t) + 0.9u_2(t - 0.02) + 0.1u_2(t) & t \in [0, 0.2] \\ \dot{x_3}(t) = -2x_3(t) + 0.25R(x_1(t), x_2(t), x_3(t)) & \\ \qquad\qquad -u_1(t)x_3(t - 0.015)(x_3(t) + 0.125) & t \in [0, 0.2] \\ x_1(0) = 0.49 & \\ x_2(0) = -0.0002 & \\ x_3(t) = -0.02 & t \in [-0.015, 0] \\ x(0.2) = (0, 0, 0) & \\ |u_1(t)| \leq 500 & t \in [0, 0.2] \\ u_2(t) = 1 & t \in [-0.02, 0] \end{cases}$$

We observe that the functional J no longer depends on the first component u_1 of the control, which is replaced by the constraint $|u_1| \leq 500$. Moreover, the delayed equation for x_3 has an additional dependency on the state, on which a terminal condition is also imposed.

The convergence of numerical approximations for this example is extremely slow ([11] reports an optimal value $J^* = 0.011970541$ obtained in about 18 h for a grid with 16000 nodes). This is caused by the lack of coercivity of J w.r.t. u_1, along with the occurrence of singular arcs, which result in a very ill-conditioned discrete problem. In this setting, we expect that better result could be obtained by solving the entire optimality system, while carrying out a minimization with respect to control (as in the scheme under consideration) could result in a chattering solution.

Here, the terminal condition on the state is enforced via a standard penalization technique, i.e., by adding the term $C\|x(0.2)\|^2$ in the functional J. This yields the final condition $\lambda(0.2) = 2Cx(0.2)$ for the adjoint variables. In practice, we use an external loop, in which the penalty coefficient C is brought from 10 to 10^6, by iteratively increasing the penalization once the algorithm reaches the prescribed accuracy tolerance.

For a complete comparison with [11], we use a grid with exactly 16000 nodes and employ LBFGS, the NLP solver that performed better in all the previous tests. Figure 4 shows all the components of the computed optimal solution. The results are in good quantitative agreement with those presented in [11], except for the control u_1, which strongly chatters along the singular arcs. Including the whole loop for the penalization, we obtain the value $J^* = 0.011977486$ in about 14 min.

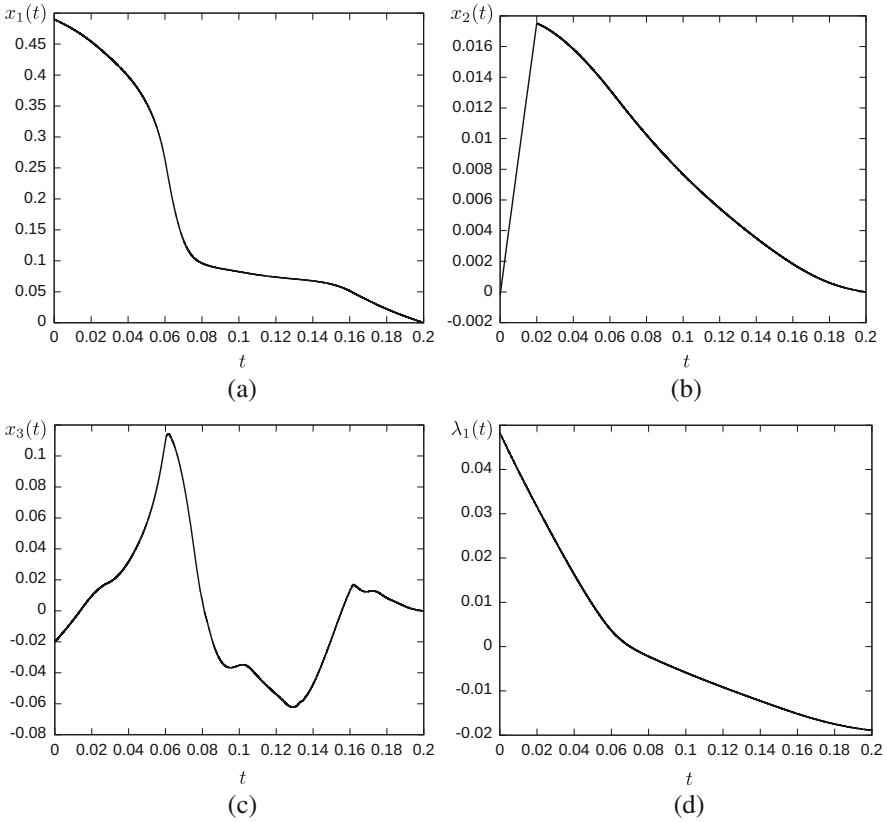


Fig. 4 Optimal solution for Test 4. (**a**), (**b**), (**c**) state (x_1, x_2, x_3), (**d**), (**e**), (**f**) costate $(\lambda_1, \lambda_2, \lambda_3)$, (**g**) and (**h**) control (u_1, u_2), and (**i**) regularized control u_1 via convolution

While chattering is clearly an undesired feature, convergence to the optimal cost is still guaranteed by Theorem 1. Moreover, the singular arcs are easily detectable, and suitable post-processing techniques might be applied to recover a smoother control (e.g., by low-pass filtering, see [17] and the references therein). Figure 4i shows a regularization of u_1 computed by convolution with a symmetric kernel (a moving average replacing the second component $U^*_{i,2}$ of the control by the average obtained over a symmetric window on the time steps $i - 10$ to $i + 10$). For this regularized control, we obtain a cost $J = 0.011983521$, which differs about 0.05% from the cost of the non-regularized control, and about 0.1% from the optimal value computed in [11].

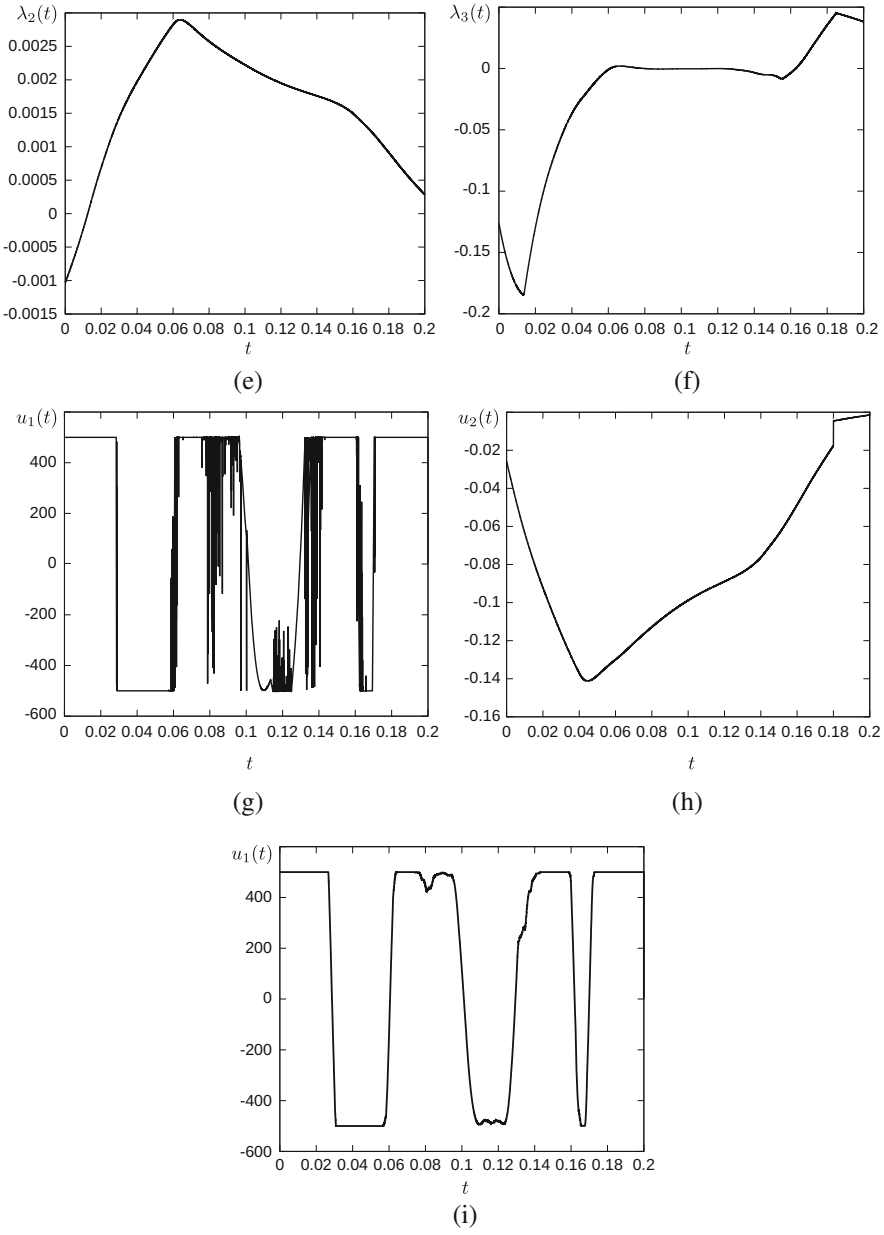


Fig. 4 (continued)

Acknowledgements The authors would like to thank anonymous reviewers for helpful comments which improved the presentation.

References

1. Angell, T.S., Kirsch, A.: On the necessary conditions for optimal control of retarded systems. Appl. Math. Optim. **22**, 117–145 (1990)
2. Banks, H.T.: Necessary conditions for control problems with variable time lags. SIAM J. Control **6**, 9–47 (1968)
3. Bellen, A., Zennaro, M.: Numerical Methods for Delay Differential Equations. Oxford University Press, Oxford (2013)
4. Betts, J.T., Campbell, S.L., Thompson, K.C.: Solving optimal control problems with control delays using direct transcription. Appl. Numer. Math. **108**, 185–203 (2016)
5. Bonalli, R., Hérissé, B., Trélat, E.: Solving nonlinear optimal control problems with state and control delays by shooting methods combined with numerical continuation on the delays (2017). arXiv:1709:04383
6. Bryson, A.E., Denham, W.F.: Optimum programming problems with inequality constraints. II. Solutions by Steepest Descent. Am. Inst. Aeronaut. Astronaut. J. **2**, 25–34 (1964)
7. Bryson, A.E., Denham, W.F., Dreyfus, S.E.: Optimal programming problems with inequality constraints. I. Necessary Conditions for Extremal Solutions. Am. Inst. Aeronaut. Astronaut. J. **1**, 2544–2550 (1963)
8. Burger, M.: Optimal Control of Dynamical Systems: Calculating Input Data for Multibody System Simulation. Verlag Dr. Hut, München (2011)
9. Gerdts, M.: Optimal Control of ODEs and DAEs. De Gruyter, Berlin (2012)
10. Göllmann, L., Maurer, H.: Theory and applications of optimal control problems with multiple time-delays. J. Ind. Manage. Optim. **10**, 413–441 (2014)
11. Göllmann, L., Kern, D., Maurer, H.: Optimal control problems with delays in state and control variables subject to mixed control-state constraints. Optimal Control Appl. Methods **30**, 341–365 (2009)
12. Guinn, T.: Reduction of delayed optimal control problems to nondelayed problems. J. Optim. Theory Appl. **18**, 371–377 (1976)
13. Halanay, A.: Optimal controls for systems with time lag. SIAM J. Control **6**, 215–234 (1968)
14. Kelley, H.J.: Gradient theory of optimal flight paths. Am. Rocket Soc. J. **30**, 947–954 (1960)
15. Kelley, H.J.: Guidance theory and extremal fields. IRE Trans. Autom. Control **7**, 75–82 (1962)
16. Kharatishvili, G.L.: A maximum principle in extremal problems with delays. Math. Theory Control 26–34 (1967)
17. Tseng, M.-L., Chen, M.-S.: Chattering reduction of sliding mode control by low-pass filtering the control signal. Asian J. Control **12**, 392–398 (2010)
18. Vinter, R.B.: State constrained optimal control problems with time delays. J. Math. Anal. Appl. **457**, 1696–1712 (2018)

POD-Based Economic Optimal Control of Heat-Convection Phenomena

Luca Mechelli and Stefan Volkwein

Abstract In the setting of energy efficient building operation, an optimal boundary control problem governed by the heat equation with a convection term is considered together with bilateral control and state constraints. The aim is to keep the temperature in a prescribed range with the least possible heating cost. In order to gain regular Lagrange multipliers a Lavrentiev regularization for the state constraints is utilized. The regularized optimal control problem is solved by a primal-dual active set strategy (PDASS) which can be interpreted as a semismooth Newton method and, therefore, has a superlinear rate of convergence. To speed up the PDASS a reduced-order approach based on proper orthogonal decomposition (POD) is applied. An a-posteriori error analysis ensures that the computed (suboptimal) POD solutions are sufficiently accurate. Numerical test illustrates the efficiency of the proposed strategy.

Keywords Convection-diffusion equation · Optimal control · State constraints · Primal-dual active set strategy · Model order reduction

1 Introduction

In this paper we consider a class of linear parabolic convection-diffusion equations which model, e.g., the evolution of the temperature inside a room, which we want to keep inside a constrained range. The boundary control implements heaters in the room, where, due to physical restrictions on the heaters, we have to impose bilateral control constraints. The goals are to minimize the heating cost while keeping the state (i.e., the temperature) inside the desired state constraints. In order to gain regular Lagrange multipliers, we utilize a Lavrentiev regularization for the state constraints; see [24]. Then, a primal-dual active set strategy (PDASS) can be applied, which has a superlinear rate of convergence [15] and a mesh-independent

L. Mechelli (✉) · S. Volkwein
University of Konstanz, Department of Mathematics and Statistics, Konstanz, Germany
e-mail: Luca.Mechelli@uni-konstanz.de; Stefan.Volkwein@uni-konstanz.de

M. Falcone et al. (eds.), *Numerical Methods for Optimal Control Problems*,
Springer INdAM Series 29, https://doi.org/10.1007/978-3-030-01959-4_4

property [16]. For the numerical solution of the equations we apply a Galerkin approximation combined with an implicit Euler scheme in time and, in order to speed-up the computation of optimal solutions, we build a reduced-order model based on proper orthogonal decomposition (POD); cf. [6, 13]. To have sufficiently accurate POD suboptimal solutions, we adapt the a-posteriori error analysis from [9]. Then, we are able to estimate the difference between the (unknown) optimal controls and their suboptimal POD approximations. For generating the POD basis, we need to solve the full system with arbitrary controls, this implies that the quality of the basis, which means how much the reduce order model solution captures the behavior of the full system one, depends on this initial choice for the controls. There are several techniques for improving the POD basis like, e.g., TR-POD [2], OS-POD [20] or adaptive strategies like in [1]. However, in this paper, we will only compare the quality of basis generated with arbitrary controls and with the idealized ones generated from the optimal finite element controls. Our motivation comes from the fact that we will utilize the proposed strategy within an economic model predictive control approach [10, Chapter 8], where the POD basis will be eventually updated during the closed-loop realization; cf. [22]. In contrast to [9] we consider economic costs, boundary controls, two-dimensional spatial domains and time- as well as space-dependent convection fields.

The paper is organized in the following way: in Sect. 2 we introduce our optimal control problem and how we deal with state and control constraints. The primal-dual active set strategy algorithm related to this problem is presented in Sect. 3. In Sect. 4 we briefly explain the POD method and the related a-posteriori error estimator is presented in Sect. 5. Numerical Tests are shown in Sect. 6. Finally, some conclusions are drawn in Sect. 7.

2 The Optimal Control Problem

2.1 The State Equation

Let $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$, be a bounded domain with Lipschitz-continuous boundary $\Gamma = \partial\Omega$. We suppose that Γ is split into two disjoint subsets Γ_{c} and Γ_{o}, where at least Γ_{c} has nonzero (Lebesgue) measure. Further, let $H = L^2(\Omega)$ and $V = H^1(\Omega)$ endowed with their usual inner products

$$\langle \varphi, \psi \rangle_H = \int_\Omega \varphi\psi \, \mathrm{d}\boldsymbol{x}, \quad \langle \varphi, \psi \rangle_V = \int_\Omega \varphi\psi + \nabla\varphi \cdot \nabla\psi \, \mathrm{d}\boldsymbol{x}$$

and their induced norms, respectively. For $T > 0$ we set $Q = (0, T) \times \Omega$, $\Sigma_{\mathsf{c}} = (0, T) \times \Gamma_{\mathsf{c}}$ and $\Sigma_{\mathsf{o}} = (0, T) \times \Gamma_{\mathsf{o}}$. By $L^2(0, T; V)$ we denote the space of measurable functions from $[0, T]$ to V, which are square integrable, i.e.,

$$\int_0^T \|\varphi(t)\|_V^2 \, \mathrm{d}t < \infty.$$

When t is fixed, the expression $\varphi(t)$ stands for the function $\varphi(t, \cdot)$ considered as a function in Ω only. The space $W(0, T)$ is defined as

$$W(0,T) = \left\{\varphi \in L^2(0,T;V) \,\middle|\, \varphi_t \in L^2(0,T;V')\right\},$$

where V' denotes the dual of V. The space $W(0, T)$ is a Hilbert space supplied with the common inner product; cf. [7, pp. 472–479]. For $m \in \mathbb{N}$ let $b_i : \Gamma_{\mathsf{c}} \to \mathbb{R}$, $1 \le i \le m$, denote given control shape functions. For $\mathcal{U} = L^2(0, T; \mathbb{R}^m)$ the set of admissible controls $u = (u_i)_{1 \le i \le m} \in \mathcal{U}$ is given as

$$\mathcal{U}_{\mathsf{ad}} = \left\{u \in \mathcal{U} \,\middle|\, u_{\mathsf{a}i} \le u_i(t) \le u_{\mathsf{b}i} \text{ for } i = 1, \ldots, m \text{ and a.e. in } [0,T]\right\},$$

where $u_{\mathsf{a}} = (u_{\mathsf{a}i})_{1 \le i \le m}$, $u_{\mathsf{b}} = (u_{\mathsf{b}i})_{1 \le i \le m} \in \mathbb{R}^m$ are lower and upper bounds, respectively, and 'a.e.' stands for 'almost everywhere'. Throughout the paper we identify the dual $\mathcal{U}'$ with $\mathcal{U}$. Then, for any control $u \in \mathcal{U}_{\mathsf{ad}}$ the state y is governed by the following *state equation*

$$\begin{aligned}
y_t(t,\boldsymbol{x}) - \Delta y(t,\boldsymbol{x}) + \boldsymbol{v}(t,\boldsymbol{x}) \cdot \nabla y(t,\boldsymbol{x}) &= 0 && \text{a.e. in } Q,\\
\frac{\partial y}{\partial \boldsymbol{n}}(t,\boldsymbol{s}) + y(t,\boldsymbol{s}) &= \sum_{i=1}^{m} u_i(t) b_i(\boldsymbol{s}) && \text{a.e. on } \Sigma_{\mathsf{c}},\\
\frac{\partial y}{\partial \boldsymbol{n}}(t,\boldsymbol{s}) + \gamma_{\mathsf{o}} y(t,\boldsymbol{s}) &= \gamma_{\mathsf{o}} y_{\mathsf{out}}(t) && \text{a.e. on } \Sigma_{\mathsf{o}},\\
y(0,\boldsymbol{x}) &= y_\circ(\boldsymbol{x}), && \text{a.e. in } \Omega.
\end{aligned} \tag{1}$$

We suppose the following hypotheses for the data in (1).

Assumption 2.1 *We assume that* $\gamma_o \ge 0$, $\boldsymbol{v} \in L^\infty(0,T;L^\infty(\Omega;\mathbb{R}^d))$ *with* $d \in \{1,2,3\}$, $y_{out} \in L^2(0,T)$, $y_\circ \in H$ *and* $b_1, \ldots, b_m \in L^\infty(\Gamma_c)$.

To write (1) in weak form we introduce the nonsymmetric, time-dependent bilinear form $a(t; \cdot, \cdot) : V \times V \to \mathbb{R}$

$$a(t;\varphi,\psi) = \int_\Omega \nabla\varphi \cdot \nabla\psi + \big(\boldsymbol{v}(t) \cdot \nabla\varphi\big)\psi \,\mathrm{d}\boldsymbol{x} + \int_{\Gamma_{\mathsf{c}}} \varphi\psi \,\mathrm{d}\boldsymbol{s} + \gamma_{\mathsf{o}} \int_{\Gamma_{\mathsf{o}}} \varphi\psi \,\mathrm{d}\boldsymbol{s}$$

for $\varphi, \psi \in V$ and the time-dependent linear functional $\mathscr{F}(t) : V \to V'$

$$\langle \mathscr{F}(t), \varphi \rangle_{V',V} = \gamma_{\mathsf{o}} y_{\mathsf{out}}(t) \int_{\Gamma_{\mathsf{o}}} \varphi \,\mathrm{d}\boldsymbol{s} \quad \text{for } \varphi \in V,$$

where $\langle \cdot, \cdot \rangle_{V',V}$ stands for the dual pairing between V and its dual space V'. Moreover, the linear operator $\mathscr{B} : \mathbb{R}^m \to V'$ is defined as

$$\langle \mathscr{B}u, \varphi \rangle_{V',V} = \sum_{i=1}^{m} u_i \int_{\Gamma_{\mathsf{c}}} b_i \varphi \,\mathrm{d}\boldsymbol{s} \quad \text{for all } \varphi \in V$$

for given $u = (u_i)_{1\le i\le m} \in \mathbb{R}^m$. Now, the state variable $y \in W(0,T)$ is called a *weak solution* to (1) if

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\langle y(t),\varphi\rangle_H + a(t;y(t),\varphi) &= \langle \mathscr{F}(t) + \mathscr{B}(u(t)),\varphi\rangle_{V',V} \ \forall \varphi \in V \text{ a.e. in } (0,T], \\ y(0) &= y_\circ \qquad\qquad \text{in } H \end{aligned} \tag{2}$$

is satisfied.

Lemma 2.1 *Let Assumption 2.1 hold. Then:*

1) *For almost all $t \in [0,T]$ the bilinear form satisfies*

$$\begin{aligned} \big|a(t;\varphi,\psi)\big| &\le \alpha\,\|\varphi\|_V\|\psi\|_V \qquad \forall \varphi,\psi \in V, \\ a(t;\varphi,\varphi) &\ge \alpha_1\,\|\varphi\|_V^2 - \alpha_2\,\|\varphi\|_H^2 \qquad \forall \varphi \in V \end{aligned}$$

   with constants α, $\alpha_1 > 0$ and $\alpha_2 \ge 0$.
2) *We have $\mathscr{F} \in L^2(0,T;V')$, and the linear operator $\mathscr{B}$ is bounded.*

Proof The claims follow by standard arguments; cf. [7] and [5], for instance. □

Theorem 2.1 *Suppose that Assumption 2.1 is satisfied. Then, (2) possesses a unique solution $y \in W(0,T)$ for every $u \in \mathcal{U}_{ad}$ satisfying the a-priori estimate*

$$\|y\|_{W(0,T)} \le c_y\big(\|y_\circ\|_H + \|y_{out}\|_{L^2(0,T)} + \|u\|_{\mathcal{U}}\big) \tag{3}$$

for a constant $c_y > 0$ which is independent of $y_\circ$, y_{out} and u.

Proof Existence of a unique solution to (2) follows directly from Lemma 2.1 and [7, pp. 512–520]. Moreover, the a-priori bound is shown in [25, Theorem 3.19]. □

Remark 2.1 We split the solution to (2) in one part, which depends on the fixed initial condition $y_\circ$ and the right-hand side $\mathscr{F}$, and another part depending linearly on the control variable. Let $\hat{y} \in W(0,T)$ be the unique solution to the problem

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\,\langle \hat{y}(t),\varphi\rangle_H + a(t;\hat{y}(t),\varphi) &= \langle \mathscr{F}(t),\varphi\rangle_{V',V} \qquad \forall \varphi \in V \text{ a.e. in } (0,T], \\ \hat{y}(0) &= y_\circ \qquad\qquad \text{in } H. \end{aligned}$$

We define the subspace

$$W_0(0,T) = \big\{\varphi \in W(0,T)\,\big|\,\varphi(0) = 0 \text{ in } H\big\}$$

endowed with the topology of $W(0,T)$. Let us now introduce the linear solution operator $\mathscr{S}: \mathcal{U} \to W_0(0,T)$: for $u \in \mathcal{U}$ the function $y = \mathscr{S}u \in W_0(0,T)$ is the unique solution to

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\langle y(t),\varphi\rangle_H + a(t;y(t),\varphi) = \langle \mathscr{B}(u(t)),\varphi\rangle_{V',V} \quad \forall \varphi \in V \text{ a.e. in } (0,T].$$

From $y \in W_0(0,T)$ it follows that $y(0) = 0$ in H. The boundedness of $\mathscr{S}$ follows from (3). Now, the solution to (2) can be expressed as $y = \hat{y} + \mathscr{S}u$. ◊

2.2 The State-Constrained Optimization Problem

We set $\mathcal{W} = L^2(0,T;H)$. Throughout the paper we identify the space $L^2(0,T;H)$ with $L^2(Q)$ and the dual $\mathcal{W}'$ with $\mathcal{W}$. Let $y \in W(0,T)$ be given and $\mathscr{E}: W(0,T) \to \mathcal{W}$ the canonical linear and bounded embedding operator. We deal with pointwise state constraints of the following type

$$y_{\mathsf{a}}(t,\boldsymbol{x}) \le \mathscr{E}y(t,\boldsymbol{x}) \le y_{\mathsf{b}}(t,\boldsymbol{x}) \quad \text{a.e. in } Q, \tag{4}$$

where $y_{\mathsf{a}}, y_{\mathsf{b}} \in \mathcal{W}$ are given lower and upper bounds, respectively. To gain regular Lagrange multipliers we utilize a Lavrentiev regularization. Let $\varepsilon > 0$ be a chosen regularization parameter and $w \in \mathcal{W}$ an additional (virtual) control. Then, (4) is replaced by the mixed control-state constraints

$$y_{\mathsf{a}}(t,\boldsymbol{x}) \le \mathscr{E}y(t,\boldsymbol{x}) + \varepsilon w(t,\boldsymbol{x}) \le y_{\mathsf{b}}(t,\boldsymbol{x}) \quad \text{a.e. in } Q.$$

We introduce the Hilbert space

$$\mathcal{X} = W(0,T) \times \mathcal{U} \times \mathcal{W}$$

endowed with the common product topology. The set of admissible solutions is given by

$$\mathcal{X}^{\varepsilon}_{\mathsf{ad}} = \big\{x = (y,u,w) \in \mathcal{X} \,\big|\, y = \hat{y} + \mathscr{S}u,\ y_{\mathsf{a}} \le \mathscr{E}y + \varepsilon w \le y_{\mathsf{b}} \text{ and } u \in \mathcal{U}_{\mathsf{ad}}\big\}.$$

The quadratic cost functional $J : \mathcal{X} \to \mathbb{R}$ is given by

$$\begin{aligned} J(x) = {} & \frac{\sigma_Q}{2}\int_0^T \|y(t) - y_Q(t)\|_H^2\,\mathrm{d}t + \frac{\sigma_T}{2}\,\|y(T) - y_T\|_H^2 \\ & + \frac{\sigma}{2}\sum_{i=1}^{m}\|u_i\|^2_{L^2(0,T)} + \frac{\sigma_w}{2}\,\|w\|^2_{\mathcal{W}} \qquad \text{for } x = (y,u,w) \in \mathcal{X}. \end{aligned}$$

Assumption 2.2 *Let the desired states satisfy* $y_Q \in L^2(0,T;H)$ *and* $y_T \in H$. *Furthermore,* $\varepsilon > 0$, σ_Q, $\sigma_T \geq 0$, *and* σ, $\sigma_w > 0$.

The optimal control problem is given by

$$\min J(x) \quad \text{subject to (s.t.)} \quad x \in \mathcal{X}_{\mathsf{ad}}^{\varepsilon}. \tag{$\mathbf{P}^{\varepsilon}$}$$

Remark 2.2 Following [19] one can consider the generalized problem

$$\begin{aligned} \min \frac{\sigma_Q}{2} \int_0^T \|y(t) - y_Q(t)\|_H^2 \,\mathrm{d}t + \frac{\sigma_T}{2} \|y(T) - y_T\|_H^2 \\ + \frac{\sigma}{2} \sum_{i=1}^m \|u_i\|_{L^2(0,T)}^2 + \frac{f(\varepsilon)}{2} \|w\|_{\mathcal{W}}^2 \end{aligned} \tag{5a}$$

subject to the modified state equations

$$\begin{aligned} y_t(t,\boldsymbol{x}) - \Delta y(t,\boldsymbol{x}) + \boldsymbol{v}(t,\boldsymbol{x}) \cdot \nabla y(t,\boldsymbol{x}) &= g(\varepsilon) w && \text{a.e. in } Q, \\ \frac{\partial y}{\partial \boldsymbol{n}}(t,\boldsymbol{s}) + y(t,\boldsymbol{s}) &= \sum_{i=1}^m u_i(t) b_i(\boldsymbol{s}) && \text{a.e. on } \Sigma_{\mathsf{c}}, \\ \frac{\partial y}{\partial \boldsymbol{n}}(t,\boldsymbol{s}) + \gamma_{\mathsf{o}} y(t,\boldsymbol{s}) &= \gamma_{\mathsf{o}} y_{\mathsf{out}}(t) && \text{a.e. on } \Sigma_{\mathsf{o}}, \\ y(0,\boldsymbol{x}) &= y_\circ(\boldsymbol{x}), && \text{a.e. in } \Omega \end{aligned} \tag{5b}$$

and to the inequality constraints

$$\begin{aligned} u_{\mathsf{a}i} \leq u_i(t) \leq u_{\mathsf{b}i} && \text{a.e. in } [0,T] \text{ for } i = 1, \ldots, m, \\ y_{\mathsf{a}}(t,\boldsymbol{x}) \leq \mathscr{E}y(t,\boldsymbol{x}) + h(\varepsilon) w(t,\boldsymbol{x}) \leq y_{\mathsf{b}}(t,\boldsymbol{x}) && \text{a.e. in } Q, \end{aligned} \tag{5c}$$

where f, g and h are chosen nonnegative functions defined for $\varepsilon \geq 0$. In [19] convergence of a solution $\bar{x}^{\varepsilon} = (\bar{y}^{\varepsilon}, \bar{u}^{\varepsilon}, \bar{w}^{\varepsilon}) \in \mathcal{X}$ is proved for $\varepsilon \to 0$ in the case of an elliptic state equation and unilateral state constraints. In our future work we will study the application of the arguments in [19] to our parabolic setting and to bilateral state constraints. ◊

Problem ($\mathbf{P}^{\varepsilon}$) can be formulated as pure control constrained problem. We set $\hat{y}_{\mathsf{a}} = y_{\mathsf{a}} - \mathscr{E}\hat{y} \in \mathcal{W}$ and $\hat{y}_{\mathsf{b}} = y_{\mathsf{b}} - \mathscr{E}\hat{y} \in \mathcal{W}$. Then, (4) can be formulated equivalently in the control variables u and w as follows:

$$\hat{y}_{\mathsf{a}}(t,\boldsymbol{x}) \leq (\mathscr{E}\mathscr{S}u)(t,\boldsymbol{x}) + \varepsilon w(t,\boldsymbol{x}) \leq \hat{y}_{\mathsf{b}}(t,\boldsymbol{x}) \quad \text{a.e. in } Q.$$

We define $\mathcal{Z} = \mathcal{U} \times \mathcal{W}$ and introduce the bounded and linear mapping

$$\mathscr{T}_\varepsilon : \mathcal{Z} \to \mathcal{Z}, \quad z = (u, w) \mapsto \mathscr{T}_\varepsilon(z) = \begin{pmatrix} u \\ \mathscr{E}\mathscr{S}u + \varepsilon w \end{pmatrix} = \begin{pmatrix} \mathscr{I}_\mathcal{U} & 0 \\ \mathscr{E}\mathscr{S} & \varepsilon \mathscr{I}_\mathcal{W} \end{pmatrix} \begin{pmatrix} u \\ w \end{pmatrix}, \tag{6}$$

where $\mathscr{I}_\mathcal{U} : \mathcal{U} \to \mathcal{U}$ and $\mathscr{I}_\mathcal{W} : \mathcal{W} \to \mathcal{W}$ stand for the identity operators in $\mathcal{U}$ and $\mathcal{W}$, respectively. Notice that $\mathscr{T}_\varepsilon$ is invertible and $\mathscr{T}_\varepsilon^{-1}$ is explicitly given as

$$\mathscr{T}_\varepsilon^{-1}(\mathfrak{u}, \mathfrak{w}) = \begin{pmatrix} \mathscr{I}_\mathcal{U} & 0 \\ -\varepsilon^{-1}\mathscr{E}\mathscr{S} & \varepsilon^{-1}\mathscr{I}_\mathcal{W} \end{pmatrix} \begin{pmatrix} \mathfrak{u} \\ \mathfrak{w} \end{pmatrix} = \left(\mathfrak{u}, \frac{1}{\varepsilon}(\mathfrak{w} - \mathscr{E}\mathscr{S}\mathfrak{u}) \right) \tag{7}$$

for all $\mathfrak{z} = (\mathfrak{u}, \mathfrak{w}) \in \mathcal{Z}$. With $z_\mathsf{a} = (u_\mathsf{a}, \hat{y}_\mathsf{a})$, $z_\mathsf{b} = (u_\mathsf{b}, \hat{y}_\mathsf{b}) \in \mathcal{Z}$ we define the closed, bounded, convex set of admissible controls as

$$\mathcal{Z}_{\mathsf{ad}}^\varepsilon = \left\{ z = (u, w) \in \mathcal{Z} \,\middle|\, z_\mathsf{a} \le \mathscr{T}_\varepsilon(z) \le z_\mathsf{b} \right\}$$

which depends—through $\mathscr{T}_\varepsilon$—from the regularization parameter ε. Let $\hat{y}_Q = y_Q - \hat{y} \in L^2(0, T; H)$ and $\hat{y}_T = y_T - \hat{y}(T) \in H$. Then, we introduce the reduced cost functional

$$\begin{aligned} \hat{J}(z) &= J(\hat{y} + \mathscr{S}u, u, w) \\ &= \frac{\sigma_Q}{2} \int_0^T \|(\mathscr{S}u)(t) - \hat{y}_Q(t)\|_H^2 \, \mathrm{d}t + \frac{\sigma_T}{2} \|(\mathscr{S}u)(T) - \hat{y}_T\|_H^2 \\ &\quad + \frac{\sigma}{2} \sum_{i=1}^m \|u_i\|_{L^2(0,T)}^2 + \frac{\sigma_w}{2} \|w\|_\mathcal{W}^2 \qquad \text{for } z = (u, w) \in \mathcal{Z}. \end{aligned}$$

Now ($\mathbf{P}^\varepsilon$) is equivalent to the following reduced problem

$$\min \hat{J}(z) \quad \text{s.t.} \quad z \in \mathcal{Z}_{\mathsf{ad}}^\varepsilon. \tag{$\hat{\mathbf{P}}^\varepsilon$}$$

Supposing Assumptions 2.1, 2.2 and applying standard arguments [21] one can prove that there exists a unique optimal solution $\bar{z} = (\bar{u}, \bar{w}) \in \mathcal{Z}_{\mathsf{ad}}^\varepsilon$ to ($\hat{\mathbf{P}}^\varepsilon$). The uniqueness follows from the strict convexity properties of the reduced cost functional on $\mathcal{Z}_{\mathsf{ad}}^\varepsilon$. Throughout this paper, a bar indicates optimality.

2.3 First-Order Optimality Conditions

First-order sufficient optimality conditions are formulated in the next theorem. The proof follows from Theorem 2.4 in [11].

Theorem 2.2 *Let Assumptions 2.1 and 2.2 hold. Suppose that the feasible set $\mathcal{Z}_{ad}^\varepsilon$ is nonempty and that $\bar{z} = (\bar{u}, \bar{w}) \in \mathcal{Z}_{ad}^\varepsilon$ is the solution to ($\hat{\mathbf{P}}^\varepsilon$) with associated optimal*

state $\bar{y} = \hat{y} + \mathscr{S}\bar{u}$. *Then, there exist unique Lagrange multipliers* $\bar{p} \in W(0,T)$ *and* $\bar{\beta} \in \mathcal{W}$, $\bar{\mu} = (\bar{\mu}_i)_{1\le i\le m} \in \mathcal{U}$ *satisfying the dual equations*

$$\begin{aligned} -\frac{\mathrm{d}}{\mathrm{d}t}\langle \bar{p}(t),\varphi\rangle_H + a(t;\varphi,\bar{p}(t)) + \langle \bar{\beta}(t),\varphi\rangle_H &= \sigma_Q \langle (y_Q - \bar{y})(t),\varphi\rangle_H \ \forall \varphi \in V, \\ \bar{p}(T) &= \sigma_T\big(y_T - \bar{y}(T)\big) \qquad \text{in } H \end{aligned} \tag{8}$$

a.e. in $[0,T]$ *and the optimality system*

$$\begin{aligned} \sigma\bar{u}_i - \int_{\Gamma_c} b_i\,\bar{p}\,\mathrm{d}s + \bar{\mu}_i = 0 \quad &\text{in } L^2(0,T) \text{ for } i = 1,\dots,m, \\ \sigma_w\bar{w} + \varepsilon\bar{\beta} = 0 \quad &\text{in } \mathcal{W}. \end{aligned} \tag{9}$$

Moreover,

$$\bar{\beta} = \max\big\{0, \bar{\beta} + \eta(\bar{y} + \varepsilon\bar{w} - y_b)\big\} + \min\big\{0, \bar{\beta} + \eta(\bar{y} + \varepsilon\bar{w} - y_a)\big\}, \tag{10a}$$

$$\bar{\mu}_i = \max\big\{0, \bar{\mu}_i + \eta_i(\bar{u}_i - u_{bi})\big\} + \min\big\{0, \bar{\mu}_i + \eta_i(\bar{u}_i - u_{ai})\big\} \tag{10b}$$

for $i = 1,\dots,m$ *and for arbitrarily chosen* $\eta, \eta_1, \dots, \eta_m > 0$, *where the max- and min-operations are interpreted componentwise in the pointwise everywhere sense.*

Remark 2.3 Analogous to Remark 2.1 we split the adjoint variable p into one part depending on the fixed desired states and into two other parts, which depend linearly on the control variable and on the multiplier β. Recall that $\hat{y}_Q$ as well as $\hat{y}_T$ are defined in Sect. 2.2. Let $\hat{p} \in W(0,T)$ denote the unique solution to the adjoint equation

$$\begin{aligned} -\frac{\mathrm{d}}{\mathrm{d}t}\langle \hat{p}(t),\varphi\rangle_H + a(t;\varphi,\hat{p}(t)) &= \sigma_Q\langle \hat{y}_Q(t),\varphi\rangle_H \qquad \forall \varphi \in V \text{ a.e. in } [0,T), \\ \hat{p}(T) &= \sigma_T\hat{y}_T \qquad\qquad \text{in } H. \end{aligned}$$

Further, we define the linear, bounded operators $\mathscr{A}_1 : \mathcal{U} \to W(0,T)$ and $\mathscr{A}_2 : \mathcal{W} \to W(0,T)$ as follows: for given $u \in \mathcal{U}$ the function $p = \mathscr{A}_1 u$ is the unique solution to

$$\begin{aligned} -\frac{\mathrm{d}}{\mathrm{d}t}\langle p(t),\varphi\rangle_H + a(t;\varphi,p(t)) &= -\sigma_Q\langle (\mathscr{S}u)(t),\varphi\rangle_H \quad \forall \varphi \in V \text{ a.e. in } [0,T), \\ p(T) &= -\sigma_T(\mathscr{S}u)(T) \qquad \text{in } H \end{aligned}$$

and for given $\beta \in \mathcal{W}$ the function $p = \mathscr{A}_2\beta$ uniquely solves

$$\begin{aligned} -\frac{\mathrm{d}}{\mathrm{d}t}\langle p(t),\varphi\rangle_H + a(\varphi,p(t)) &= -\langle \beta(t),\varphi\rangle_H \qquad \forall \varphi \in V \text{ a.e. in } [0,T), \\ p(T) &= 0 \qquad\qquad \text{in } H. \end{aligned}$$

In particular, the solution $\bar{p}$ to (8) is given by $\bar{p} = \hat{p} + \mathscr{A}_1\bar{u} + \mathscr{A}_2\bar{\beta}$. ◊

It follows from Theorem 2.2 that the first-order conditions for ($\hat{\mathbf{P}}^\varepsilon$) can be equivalently written as the nonsmooth nonlinear system

$$\sigma \bar{u}_i - \gamma_{\mathsf{c}} \int_{\Gamma_{\mathsf{c}}} b_i \bar{p} \, \mathrm{d}s + \bar{\mu}_i = 0, \quad i = 1, \ldots, m, \tag{11a}$$

$$\sigma_w \bar{w} + \varepsilon \bar{\beta} = 0, \tag{11b}$$

$$\bar{\mu}_i = \max\left\{0, \bar{\mu}_i + \eta_i(\bar{u}_i - u_{\mathsf{b}i})\right\} + \min\left\{0, \bar{\mu}_i + \eta_i(\bar{u}_i - u_{\mathsf{a}i})\right\}, \tag{11c}$$

$$\bar{\beta} = \max\left\{0, \bar{\beta} + \eta(\bar{y} + \varepsilon \bar{w} - y_{\mathsf{b}})\right\} + \min\left\{0, \bar{\beta} + \eta(\bar{y} + \varepsilon \bar{w} - y_{\mathsf{a}})\right\} \tag{11d}$$

with the unknowns $\bar{u}$, $\bar{w}$, $\bar{\beta}$ and $\bar{\mu}$.

Remark 2.4 Optimality system (11) can also be expressed as a variational inequality; cf. [17, 25]. Since the admissible set $\mathcal{Z}_{\mathsf{ad}}^\varepsilon$ is convex and the strictly convex reduced objective $\hat{J}$ is Fréchet-differentiable, first-order sufficient optimality conditions for ($\hat{\mathbf{P}}^\varepsilon$) are given as

$$\langle \nabla \hat{J}(\bar{z}), z - \bar{z} \rangle_{\mathcal{Z}} \geq 0 \quad \forall z \in \mathcal{Z}_{\mathsf{ad}}^\varepsilon, \tag{12}$$

where the gradient $\nabla \hat{J}$ of $\hat{J}$ at a given $z = (u, w) \in \mathcal{Z}_{\mathsf{ad}}^\varepsilon$ is

$$\nabla \hat{J}(z) = \begin{pmatrix} \left(\sigma u_i - \langle b_i, p(\cdot) \rangle_{L^2(\Gamma_{\mathsf{c}})}\right)_{1 \leq i \leq m} \\ \sigma_w w \end{pmatrix} \tag{13}$$

with $p = \hat{p} + \mathscr{A}_1 u$. ◊

3 The Primal-Dual Active Set Strategy (PDASS)

To solve ($\hat{\mathbf{P}}^\varepsilon$) we utilize a semismooth Newton method which can be interpreted as a primal-dual active set strategy; cf. [15, 18, 27]. For more details we refer to [9, 11]. Suppose that $z^k = (u^k, w^k) \in \mathcal{Z}$ is a current iterate for $k \in \mathbb{N}_0$. Then, we set $y^0 = \hat{y} + \mathscr{S} u^0$, $p^0 = \hat{p} + \mathscr{A}_1 u^0 - \sigma_w \mathscr{A}_2 w^0 / \varepsilon$,

$$y^k = \hat{y} + \mathscr{S} u^k, \qquad \beta^k = -\frac{\sigma_w}{\varepsilon} w^k,$$

$$p^k = \hat{p} + \mathscr{A}_1 u^k + \mathscr{A}_2 \beta^k, \qquad \mu_i^k = \int_{\Gamma_{\mathsf{c}}} b_i p^k \, \mathrm{d}s - \sigma u_i^k \text{ for } i = 1, \ldots, m.$$

Now we define the associated active sets

$$\begin{aligned}
\mathcal{A}^{\mathcal{U}}_{\mathsf{a}i}(z^k) &= \big\{t \in [0,T] \,\big|\, \mu_i^k + \sigma(u_i^k - u_{\mathsf{a}i}) < 0 \text{ a.e.}\big\}, \quad i = 1,\dots,m,\\
\mathcal{A}^{\mathcal{U}}_{\mathsf{b}i}(z^k) &= \big\{t \in [0,T] \,\big|\, \mu_i^k + \sigma(u_i^k - u_{\mathsf{b}i}) > 0 \text{ a.e.}\big\}, \quad i = 1,\dots,m,\\
\mathcal{A}^{\mathcal{W}}_{\mathsf{a}}(z^k) &= \left\{(t,\boldsymbol{x}) \in Q \,\middle|\, \beta^k + \frac{\sigma_w}{\varepsilon^2}\big(y^k + \varepsilon w^k - y_{\mathsf{a}}\big) < 0 \text{ a.e.}\right\},\\
\mathcal{A}^{\mathcal{W}}_{\mathsf{b}}(z^k) &= \left\{(t,\boldsymbol{x}) \in Q \,\middle|\, \beta^k + \frac{\sigma_w}{\varepsilon^2}\big(y^k + \varepsilon w^k - y_{\mathsf{b}}\big) > 0 \text{ a.e.}\right\}.
\end{aligned} \tag{14a}$$

The associated inactive sets are defined as

$$\begin{aligned}
\mathcal{I}^{\mathcal{U}}_i(z^k) &= [0,T] \setminus \big(\mathcal{A}^{\mathcal{U}}_{\mathsf{a}i}(z^k) \cup \mathcal{A}^{\mathcal{U}}_{\mathsf{b}i}(z^k)\big) \quad \text{for } i = 1,\dots,m,\\
\mathcal{I}^{\mathcal{W}}(z^k) &= Q \setminus \big(\mathcal{A}^{\mathcal{W}}_{\mathsf{a}}(z^k) \cup \mathcal{A}^{\mathcal{W}}_{\mathsf{b}}(z^k)\big).
\end{aligned} \tag{14b}$$

Now it turns out that the new state y^{k+1} and the new adjoint p^{k+1} are given by the two coupled problems

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\langle y^{k+1}(t),\varphi\rangle_H + a(y^{k+1}(t),\varphi) - \sum_{i=1}^m \chi_{\mathcal{I}^{\mathcal{U}}_i(z^k)}(t)\frac{1}{\sigma}\int_{\Gamma_{\mathsf{c}}} b_i\, p^{k+1}(t)\mathrm{d}\tilde{\boldsymbol{s}} \int_{\Gamma_{\mathsf{c}}} b_i\varphi \,\mathrm{d}\boldsymbol{s}\\
&\quad = \langle \mathscr{F}(t),\varphi\rangle_{V',V} + \sum_{i=1}^m \big(\chi_{\mathcal{A}^{\mathcal{U}}_{\mathsf{a}i}(z^k)}(t)u_{\mathsf{a}i}(t) + \chi_{\mathcal{A}^{\mathcal{U}}_{\mathsf{b}i}(z^k)}(t)u_{\mathsf{b}i}(t)\big)\int_{\Gamma_{\mathsf{c}}} b_i\varphi \,\mathrm{d}\boldsymbol{s}\\
&\qquad\qquad \forall \varphi \in V \text{ a.e. in } (0,T],\\
&y^{k+1}(0) = y_\circ.
\end{aligned}$$

and

$$\begin{aligned}
&-\frac{\mathrm{d}}{\mathrm{d}t}\,\langle p^{k+1}(t),\varphi\rangle_H + a(t;\varphi,p^{k+1}(t)) + \sigma_Q\,\langle y^{k+1}(t),\varphi\rangle_H\\
&\qquad + \frac{\sigma_w}{\varepsilon^2}\left\langle y^{k+1}(t)\big(\chi_{\mathcal{A}^{\mathcal{W}}_{\mathsf{a}}(z^k)}(t) + \chi_{\mathcal{A}^{\mathcal{W}}_{\mathsf{b}}(z^k)}(t)\big),\varphi\right\rangle_H\\
&\quad = \sigma_Q\,\langle y_Q(t),\varphi\rangle_H + \frac{\sigma_w}{\varepsilon^2}\left\langle y_{\mathsf{a}}(t)\chi_{\mathcal{A}^{\mathcal{W}}_{\mathsf{a}}(z^k)}(t) + y_{\mathsf{b}}(t)\chi_{\mathcal{A}^{\mathcal{W}}_{\mathsf{b}}(z^k)}(t),\varphi\right\rangle_H,\\
&\qquad\qquad \forall \varphi \in V \text{ a.e. in } [0,T),\\
&p^{k+1}(T) = \sigma_T\big(y_T - y^{k+1}(T)\big),
\end{aligned}$$

respectively, which can be expressed as

$$\begin{pmatrix} \mathscr{A}^k_{11} & \mathscr{A}^k_{12}\\ \mathscr{A}^k_{21} & \mathscr{A}^k_{22}\end{pmatrix}\begin{pmatrix} y^{k+1}\\ p^{k+1}\end{pmatrix} = \begin{pmatrix} \mathscr{Q}_1(z^k; y_\circ, u_{\mathsf{a}}, u_{\mathsf{b}}, b_i, \sigma, \gamma_{\mathsf{c}}, y_{\mathsf{out}})\\ \mathscr{Q}_2(z^k; y_{\mathsf{a}}, y_{\mathsf{b}}, y_Q, y_T, \varepsilon, \sigma_w)\end{pmatrix}. \tag{15}$$

We have $\mathscr{A}_{11}^k = \mathscr{A} + \tilde{\mathscr{A}}_{11}^k$ and $\mathscr{A}_{22}^k = \mathscr{A}^\star + \tilde{\mathscr{A}}_{22}^k$, where the k-independent operator $\mathscr{A}\colon W(0,T) \to L^2(0,T,V')$ is defined as

$$\langle \mathscr{A} y, \varphi \rangle_{L^2(0,T;V'),L^2(0,T;V)} = \int_0^T \langle y_t(t), \varphi(t) \rangle_{V',V} + a(t; y(t), \varphi(t))\,\mathrm{d}t$$

for $y \in W(0,T)$ and $\varphi \in L^2(0,T;V)$. The new control variable $z^{k+1} = (u^{k+1}, w^{k+1})$ is given by the linear system

$$\begin{aligned}
\int_{\Gamma_c} b_i\, p^{k+1}\,\mathrm{d}s - \sigma u_i^{k+1} &= 0 \qquad \text{in } \mathcal{I}_i^{\mathcal{U}}(z^k), \quad i = 1,\dots,m,\\
u_i^{k+1} &= u_{\mathsf{a}i} \quad \text{in } \mathcal{A}_{\mathsf{a}i}^{\mathcal{U}}(z^k), \quad i = 1,\dots,m,\\
u_i^{k+1} &= u_{\mathsf{b}i} \quad \text{in } \mathcal{A}_{\mathsf{b}i}^{\mathcal{U}}(z^k), \quad i = 1,\dots,m,\\
w^{k+1} &= 0 \qquad \text{in } \mathcal{I}^{\mathcal{W}}(z^k),\\
y^{k+1} + \varepsilon\, w^{k+1} &= y_{\mathsf{a}} \quad \text{in } \mathcal{A}_{\mathsf{a}}^{\mathcal{W}}(z^k),\\
y^{k+1} + \varepsilon\, w^{k+1} &= y_{\mathsf{b}} \quad \text{in } \mathcal{A}_{\mathsf{b}}^{\mathcal{W}}(z^k).
\end{aligned} \tag{16}$$

We resume the previous strategy in Algorithm 1.

Remark 3.1 Algorithm 1 has to be discretized for their numerical realizations. In our tests carried out in Sect. 6 we utilize the implicit Euler method for the time integration. For the spatial approximation we apply a finite element Galerkin scheme with piecewise linear finite elements on a triangular mesh. ◊

Algorithm 1 PDASS method for ($\hat{\mathbf{P}}^\varepsilon$)

1: Choose starting value $z^0 = (u^0, w^0) \in \mathcal{Z}$; set $k = 0$ and `flag = false`;
2: Determine $y^0 = \hat{y} + \mathscr{S}u^0$ and $p^0 = \hat{p} + \mathscr{A}_1 u^0 - \sigma_w \mathscr{A}_2 w^0/\varepsilon$;
3: **repeat**
4: Get $\mathcal{A}_{\mathsf{a}i}^{\mathcal{U}}(z^k)$, $\mathcal{A}_{\mathsf{b}i}^{\mathcal{U}}(z^k)$, $\mathcal{I}_i^{\mathcal{U}}(z^k)$, $i = 1,\dots,m$, and $\mathcal{A}_{\mathsf{a}}^{\mathcal{W}}(z^k)$, $\mathcal{A}_{\mathsf{b}}^{\mathcal{W}}(z^k)$, $\mathcal{I}^{\mathcal{W}}(z^k)$ from (14);
5: Compute the solution (y^{k+1}, p^{k+1}) by solving (15);
6: Compute $z^{k+1} = (u^{k+1}, w^{k+1}) \in \mathcal{Z}$ from (16);
7: Set $k = k + 1$;
8: **if** $\mathcal{A}_{\mathsf{a}1}^{\mathcal{U}}(z^k) = \mathcal{A}_{\mathsf{a}1}^{\mathcal{U}}(z^{k-1})$ **and** … **and** $\mathcal{A}_{\mathsf{a}m}^{\mathcal{U}}(z^k) = \mathcal{A}_{\mathsf{a}m}^{\mathcal{U}}(z^{k-1})$ **then**
9: **if** $\mathcal{A}_{\mathsf{b}1}^{\mathcal{U}}(z^k) = \mathcal{A}_{\mathsf{b}1}^{\mathcal{U}}(z^{k-1})$ **and** … **and** $\mathcal{A}_{\mathsf{b}m}^{\mathcal{U}}(z^k) = \mathcal{A}_{\mathsf{b}m}^{\mathcal{U}}(z^{k-1})$ **then**
10: **if** $\mathcal{A}_{\mathsf{a}}^{\mathcal{W}}(z^k) = \mathcal{A}_{\mathsf{a}}^{\mathcal{W}}(z^k)$ **and** $\mathcal{A}_{\mathsf{b}}^{\mathcal{W}}(z^k) = \mathcal{A}_{\mathsf{b}}^{\mathcal{W}}(z^{k-1})$ **then**
11: `flag = true`;
12: **end if**
13: **end if**
14: **end if**
15: **until** `flag = true`;

4 Proper Orthogonal Decomposition

For properly chosen admissible controls $z = (u, w) \in \mathcal{Z}_{\mathsf{ad}}^{\varepsilon}$ we set $y = \hat{y} + \mathscr{S}u$ and $p = \hat{p} + \mathscr{A}_1 u - \frac{\sigma_w}{\varepsilon}\mathscr{A}_2 w$. Then, we introduce the linear subspace

$$\mathcal{V} = \operatorname{span}\left\{ y(t),\, p(t) \,\middle|\, t \in [0,T] \right\} \subset V \tag{17}$$

with $\mathsf{d} = \dim \mathcal{V} \geq 1$. We call the set $\mathcal{V}$ the *snapshots subspace*. Let $\{\psi_i\}_{i=1}^{\mathsf{d}}$ denote an orthonormal basis for $\mathcal{V}$, then each snapshot can be expressed as

$$y(t) = \sum_{i=1}^{\mathsf{d}} \langle y(t), \psi_i \rangle_V \, \psi_i \quad \text{and} \quad p(t) = \sum_{i=1}^{\mathsf{d}} \langle p(t), \psi_i \rangle_V \, \psi_i \quad \text{a.e. in } [0,T] \tag{18}$$

The method of proper orthogonal decomposition (POD) consist in choosing an orthonormal basis $\{\psi_i\}_{i=1}^{\mathsf{d}}$ in $\mathcal{V}$ such that for every $\ell \in \mathbb{N}$ with $\ell \leq \mathsf{d}$ the mean square error between the snapshots y, p and their corresponding ℓ-th partial sum of (18) is minimized:

$$\begin{aligned} &\min \int_0^T \left\| y(t) - \sum_{i=1}^{\ell} \langle y(t), \psi_i \rangle_V \, \psi_i \right\|_V^2 + \left\| p(t) - \sum_{i=1}^{\ell} \langle p(t), \psi_i \rangle_V \, \psi_i \right\|_V^2 \mathrm{d}t \\ &\text{s.t. } \{\psi_i\}_{i=1}^{\ell} \subset V \text{ and } \langle \psi_i, \psi_j \rangle_V = \delta_{ij} \text{ for } 1 \leq i, j \leq \ell, \end{aligned} \tag{19}$$

where δ_{ij} is the Kronecker delta.

Definition 4.1 A solution $\{\psi_i\}_{i=1}^{\ell}$ to (19) is called a POD basis of rank ℓ. We define the subspace spanned by the first ℓ POD basis functions as $V^\ell = \operatorname{span}\{\psi_1, \ldots, \psi_\ell\}$.

Using a Lagrangian framework, the solution to (19) is characterized by the following optimality conditions (cf. [6, 13]):

$$\mathcal{R}\psi = \lambda\psi, \tag{20}$$

where the operator $\mathcal{R} : V \to V$ given by

$$\mathcal{R}\psi = \int_0^T \langle y(t), \psi \rangle_V \, y(t) + \langle p(t), \psi \rangle_V \, p(t) \, \mathrm{d}t \quad \text{for } \psi \in V$$

is compact, nonnegative and self-adjoint operator. Thus, there exist an orthonormal basis $\{\psi_i\}_{i\in\mathbb{N}}$ for V and an associated sequence $\{\lambda_i\}_{i\in\mathbb{N}}$ of nonnegative real numbers so that

$$\mathcal{R}\psi_i = \lambda_i \psi_i, \quad \lambda_1 \geq \cdots \geq \lambda_{\mathsf{d}} > 0 \quad \text{and} \quad \lambda_i = 0, \quad \text{for } i > d. \tag{21}$$

Moreover $\mathcal{V} = \operatorname{span}\{\psi_i\}_{i=1}^{\mathsf{d}}$. It can be also proved, see [6], that we have the following error formula for the POD basis $\{\psi_i\}_{i=1}^{\ell}$ of rank ℓ:

$$\int_0^T \Big\| y(t) - \sum_{i=1}^{\ell} \langle y(t), \psi_i \rangle_V \, \psi_i \Big\|_V^2 + \Big\| p(t) - \sum_{i=1}^{\ell} \langle p(t), \psi_i \rangle_V \, \psi_i \Big\|_V^2 \, \mathrm{d}t = \sum_{i=\ell+1}^{\mathsf{d}} \lambda_i .$$

Remark 4.1 For the numerical realization, the Hilbert space V has to be discretized by, e.g., piecewise finite elements and the integral over $[0, T]$ has to be replaced by a trapezoidal approximation; see [13]. ◊

If a POD basis $\{\psi_i\}_{i=1}^{\ell}$ of rank ℓ is computed, we can derive a reduced-order model for (2): for any $u \in \mathcal{U}$ the function $y^\ell = \mathscr{S}^\ell u \in W(0, T)$ is given by

$$\frac{\mathrm{d}}{\mathrm{d}t} \langle y^\ell(t), \psi \rangle_H + a(t; y^\ell(t), \psi) = \langle \mathscr{B}(u(t)), \psi \rangle_{V',V} \quad \forall \psi \in V^\ell \text{ a.e. in } (0, T]. \tag{22}$$

For any $u \in \mathcal{U}_{\mathsf{ad}}$ the POD approximation y^ℓ for the state solution is $y^\ell = \hat{y} + \mathscr{S}^\ell u$. Analogously a reduced-order model can be derived for the adjoint equation; see, e.g.,[13]. The POD Galerkin approximation of ($\hat{\mathbf{P}}^\varepsilon$) is given by

$$\min \hat{J}^\ell(z) = J(\hat{y} + \mathscr{S}^\ell u, z) \quad \text{s.t.} \quad z \in \mathcal{Z}_{\mathsf{ad}}^{\varepsilon,\ell}, \tag{$\hat{\mathbf{P}}^\ell$}$$

where the set of admissible controls is

$$\mathcal{Z}_{\mathsf{ad}}^{\varepsilon,\ell} = \big\{ z = (u, w) \in \mathcal{Z} \,\big|\, u \in \mathcal{U}_{\mathsf{ad}} \text{ and } \hat{y}_{\mathsf{a}} \le (\mathscr{E}\mathscr{S}^\ell u)(t, \boldsymbol{x}) + \varepsilon w(t, \boldsymbol{x}) \le \hat{y}_{\mathsf{b}} \big\}.$$

5 A-Posteriori Error Analysis

In this section we present an a-posteriori error estimate which is based on a perturbation argument [8] and has been already utilized in [26]. As done in [9], this estimate can be generalized for the mixed control-state constraints case. As first, suppose that Assumptions 2.1 and 2.2 hold. Recall that the linear, invertible operator $\mathscr{T}_\varepsilon$ has been introduced in (6). In particular, $z = (u, w)$ belongs to $\mathcal{Z}_{\mathsf{ad}}^{\varepsilon}$ if $\mathfrak{z} = (\mathfrak{u}, \mathfrak{w}) = \mathscr{T}(z) \in \mathfrak{Z}_{\mathsf{ad}}$ holds with the closed, bounded and convex subset

$$\mathfrak{Z}_{\mathsf{ad}} = \big\{ \mathfrak{z} = (\mathfrak{u}, \mathfrak{w}) \in \mathcal{Z} \,\big|\, u_{\mathsf{a}} \le \mathfrak{u} \le u_{\mathsf{b}} \text{ in } \mathcal{U} \text{ and } \hat{y}_{\mathsf{a}} \le \mathfrak{w} \le \hat{y}_{\mathsf{b}} \text{ in } \mathcal{W} \big\} \subset \mathcal{Z}.$$

Note that—compared to the definition of the admissible set $\mathcal{Z}_{\mathsf{ad}}^{\varepsilon}$—the set $\mathfrak{Z}_{\mathsf{ad}}$ does not depend on the solution operator $\mathscr{S}$ and on the regularization parameter ε. Now, we consider instead of ($\hat{\mathbf{P}}^\varepsilon$) the following optimal control problem

$$\min \hat{J}\big(\mathscr{T}_\varepsilon^{-1}\mathfrak{z}\big) \quad \text{s.t.} \quad \mathfrak{z} = (\mathfrak{u}, \mathfrak{w}) \in \mathfrak{Z}_{\mathsf{ad}}. \tag{$\hat{\mathbf{P}}^\varepsilon$}$$

If $\bar{z} = (\bar{u}, \bar{w})$ solves ($\hat{\mathbf{P}}^\varepsilon$), then $\bar{\mathfrak{z}} = \mathscr{T}_\varepsilon(\bar{z})$ is the solution to ($\hat{\mathbf{P}}^\varepsilon$). Conversely, if $\bar{\mathfrak{z}}$ solves ($\hat{\mathbf{P}}^\varepsilon$), then $\bar{z} = \mathscr{T}_\varepsilon^{-1}(\bar{\mathfrak{z}})$ is the solution to ($\hat{\mathbf{P}}^\varepsilon$). According to [9] we have the following result:

Theorem 5.1 *Suppose that Assumptions 2.1 and 2.2 hold. Let $\bar{z} = (\bar{u}, \bar{w})$ be the optimal solution to ($\hat{\mathbf{P}}^\varepsilon$).*

1) *$\bar{\mathfrak{z}} = \mathscr{T}_\varepsilon(\bar{z})$ is the solution to ($\hat{\mathbf{P}}^\varepsilon$).*
2) *Suppose that a point $\mathfrak{z}^{ap} = (\mathfrak{u}^{ap}, \mathfrak{w}^{ap}) \in \mathfrak{Z}_{ad}$ is computed. We set $z^{ap} = \mathscr{T}_\varepsilon^{-1}(\mathfrak{z}^{ap})$, i.e., $z^{ap} = (u^{ap}, w^{ap})$ fulfills $u^{ap} = \mathfrak{u}^{ap}$ and $w^{ap} = \varepsilon^{-1}(\mathfrak{w}^{ap} - \mathscr{E}\mathscr{S}\mathfrak{u}^{ap})$. Then, there exists a perturbation $\zeta = (\zeta^u, \zeta^w) \in \mathcal{Z}$, which is independent of $\bar{z}$, so that*

$$\|\bar{z} - z^{ap}\|_{\mathcal{Z}} \le \frac{1}{\sigma_z} \|\mathscr{T}_\varepsilon^\star \zeta\|_{\mathcal{Z}} \quad \text{with } \sigma_z = \min\{\sigma, \sigma_w\} > 0. \tag{23}$$

where $\mathscr{T}_\varepsilon^\star$ denotes the adjoint of the operator $\mathscr{T}_\varepsilon$; cf. (7).

Proof Since $\mathscr{T}_\varepsilon$ has a bounded inverse, part 1) follows. The second claim can be shown by adapting the proof of Proposition 1 in [9].

Remark 5.1

1) The perturbation ζ can be computed following [9, Section 1.5].
2) In our numerical realization the approximate solution z^{ap} is given by the POD suboptimal solution $\bar{z}^\ell = (\bar{u}^\ell, \bar{w}^\ell) \in \mathcal{Z}_{\mathsf{ad}}^{\varepsilon,\ell}$ to ($\hat{\mathbf{P}}^\ell$). Thus, we proceed as in [12, 26] and utilize (23) as an a-posteriori error estimate in the following manner: We set

$$\mathfrak{z}^{\mathsf{ap}} = (\mathfrak{u}^{\mathsf{ap}}, \mathfrak{w}^{\mathsf{ap}}) \in \mathcal{Z} \quad \text{with} \quad \mathfrak{u}^{\mathsf{ap}} = \bar{u}^\ell \text{ and } \mathfrak{w}^{\mathsf{ap}} = \varepsilon \bar{w}^\ell + \mathscr{E}\mathscr{S}^\ell \bar{u}^\ell. \tag{24}$$

From $\bar{z}^\ell \in \mathcal{Z}_{\mathsf{ad}}^{\varepsilon,\ell}$ we infer that $\mathfrak{z}^{\mathsf{ap}} \in \mathfrak{Z}_{\mathsf{ad}}$. It follows from (7) and (24) that

$$\begin{aligned} z^{\mathsf{ap}} = \mathscr{T}_\varepsilon^{-1}(\mathfrak{z}^{\mathsf{ap}}) &= \left(\mathfrak{u}^{\mathsf{ap}}, \varepsilon^{-1}\left(\mathfrak{w}^{\mathsf{ap}} - \mathscr{E}\mathscr{S}\mathfrak{u}^{\mathsf{ap}}\right)\right) \\ &= \left(\bar{u}^\ell, \bar{w}^\ell + \varepsilon^{-1}\mathscr{E}\big(\mathscr{S}^\ell - \mathscr{S}\big)\bar{u}^\ell\right) \end{aligned}$$

fulfills (23). Moreover, we found that

$$\bar{z} - z^{\mathsf{ap}} = \bar{z} - \bar{z}^\ell + \left(0, \varepsilon^{-1}\mathscr{E}\big(\mathscr{S} - \mathscr{S}^\ell\big)\bar{u}^\ell\right).$$

Consequently, (23) is not only an a-posteriori error estimate for $\bar{z} - \bar{z}^\ell$, but also for $\varepsilon^{-1}\mathscr{E}(\mathscr{S} - \mathscr{S}^\ell)\bar{u}^\ell$. ◊

6 Numerical Tests

All the tests in this section have been made on a Notebook Lenovo ThinkPad T450s with Intel Core i7-5600U CPU @ 2.60 GHz and 12 GB RAM. The codes are written in C language and we use the tools of PETSc, [3, 4], and SLEPc, [14, 23], for our numerical computations. In the tests we apply a discrete variant of Algorithm 1. For solving the linear system in step 5 of Algorithm 1, we use GMRES with an incomplete LU factorization as preconditioner. For all tests, $T = 1$ is chosen, and the domain Ω will be the unit square $(0, 1) \times (0, 1)$, where we supposed to have four 'heaters', which we call controls for simplicity, placed as shown in Fig. 1, with the following shape functions:

$$b_1(x) = \begin{cases} 1 & \text{if } x_1 = 0, \quad 0 \le x_2 \le 0.25, \\ 0 & \text{otherwise.} \end{cases} \qquad b_2(x) = \begin{cases} 1 & \text{if } 0.25 \le x_1 \le 0.5, \ x_2 = 1, \\ 0 & \text{otherwise.} \end{cases}$$

$$b_3(x) = \begin{cases} 1 & \text{if } x_1 = 1, \ 0.5 \le x_2 \le 0.75, \\ 0 & \text{otherwise.} \end{cases} \qquad b_4(x) = \begin{cases} 1 & \text{if } 0.5 \le x_1 \le 0.75, \ x_2 = 0, \\ 0 & \text{otherwise.} \end{cases}$$

We choose the physical parameter $\gamma_0 = 0.03$ and as initial condition $y_\circ(x) = |\sin(2\pi x_1)\cos(2\pi x_2)|$ for $x = (x_1, x_2) \in \Omega$, as shown in Fig. 1. The velocity field is chosen as $v(t, x) = (v_1(t, x), v_2(t, x))$ for all $t \in [0, T]$, with:

$$v_1(t, x) = \begin{cases} -1.6 & \text{if } t < 0.5, \ x \in \mathcal{V}_{\mathcal{F}_1}, \\ -0.6 & \text{if } t \ge 0.5, \ x \in \mathcal{V}_{\mathcal{F}_2}, \\ 0 & \text{otherwise} \end{cases} \qquad v_2(t, x) = \begin{cases} 0.5 & \text{if } t < 0.5, \ x \in \mathcal{V}_{\mathcal{F}_1}, \\ 1.5 & \text{if } t \ge 0.5, \ x \in \mathcal{V}_{\mathcal{F}_2}, \\ 0 & \text{otherwise} \end{cases}$$

and

$$\mathcal{V}_{\mathcal{F}_1} = \{x = (x_1, x_2) \,|\, 12x_2 + 4x_1 \ge 3, 12x_2 + 4x_1 \le 13\},$$
$$\mathcal{V}_{\mathcal{F}_2} = \{x = (x_1, x_2) \,|\, x_1 + x_2 \ge 0.5, x_1 + x_2 \le 1.5\}.$$

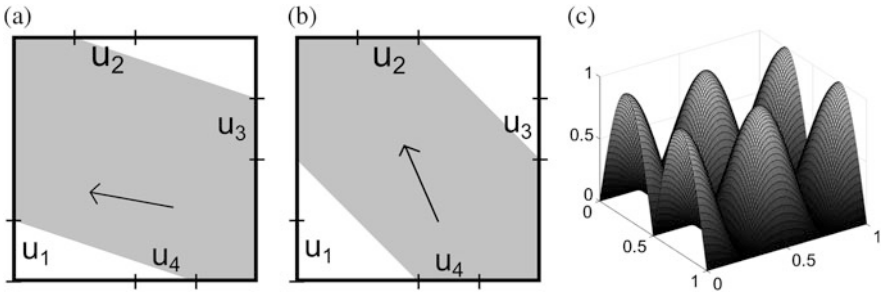


Fig. 1 Spatial domain Ω with the four boundary controls and the velocity fields (grey); initial condition $y_\circ(x)$. (**a**) $t < 0.5$. (**b**) $t \ge 0.5$. (**c**) $y_\circ(x)$

By these choices, this test represents the following scenario: the boundary controls are heaters and the velocity field, which is both space and time dependent, models the air flow in the room, which clearly changes in time. We also suppose that we have an outside temperature $y_{\mathsf{out}}(t) = -1$ for $t \in [0, 0.5)$ and $y_{\mathsf{out}}(t) = 1$ for $t \in [0.5, T]$. We fix as target $y_Q(t, x) = \min(2.0 + t, 3.0)$ and $y_T(\boldsymbol{x}) = y_Q(T, \boldsymbol{x})$, as state constraints $y_{\mathsf{a}}(t) = 0.5 + \min(2t, 2.0)$ and $y_{\mathsf{b}} = 3.0$. The time dependent lower constraints $y_{\mathsf{a}}(t)$ is chosen to gradually rise the temperature in time, in order to save heating. Moreover, we choose the control constraints $u_{\mathsf{a}i} = 0$ and $u_{\mathsf{b}i} = 7$ for $i = 1, \ldots, 4$. We build the POD basis in two different ways: the first POD basis (POD-M1) is built using the FE snapshots generated solving the state equation with the controls $u_i(t) = 3.5$ for $t \in [0, T]$ and $i = 1, \ldots, m$. The second POD basis (POD-M2) is constructed using the FE optimal control related to the considered test. We expect that the second basis will produce better results, since it contains information regarding the optimal solution. For the implicit Euler method we choose the equidistant time step $\Delta t = 0.01$. The spatial discretization is carried out by piecewise linear finite elements (FE) on a triangular mesh with $N_x = 625$ nodes.

6.1 Test 1: Economic Optimal Control

The cost functional weights are $\sigma_T = \sigma_Q = 0$ and $\sigma_w = \sigma = 1$. This choice is motivated by economic optimal control: we do not want to reach a target, but we focus our attention only on respecting the state constraints, keeping the controls as small as possible. For more information on economic optimal control we refer to [10, Chapter 8], for instance. In this test, as first, we study the behaviour of the PDASS for different values of ε. We will then analyse how this regularization parameter influences the POD approximation and the tightness of the error estimator. Finally, we will compare the POD-M1 and POD-M2 approximation for a fixed value of ε. As can be seen from Fig. 2 and as expected, when ε decreases the minimum temperature in the room gets progressively close to the lower constraints $y_a(t)$ at each time instance, while the average temperature and the maximum one remain for more time inside the constraints' range. The gradual decay of the temperature at the last time steps is due to the terminal condition for the dual variable p: from (8), since in this test $\sigma_T = 0$ holds, we have that $p(T) = 0$. Therefore, the computation of the optimal control, which is affected by this condition, lead to the previously noticed phenomena. As reported in Table 1, the number of PDASS iterations increases when ε decreases: when ε is small, the virtual control w is big in the active points, thus the algorithm employs more iterations to minimize the cost functional, where u and w have the same weights, respecting also the control constraints. It can be shown that $\varepsilon w = (y_{\mathsf{a}} - y)\, \chi_{\mathcal{A}^w_{\mathsf{a}}(z)} + (y_{\mathsf{b}} - y)\, \chi_{\mathcal{A}^w_{\mathsf{b}}(z)}$ holds, hence, the L^2-norm of εw can be used to measure how much the constraints are violated during all the evolution of the solution. As can be seen from Table 1, this value confirms what we already stated commenting Fig. 2. In Table 2, the relative errors between the solution computed with the POD-M2 approximation and the FE one are reported for the same number

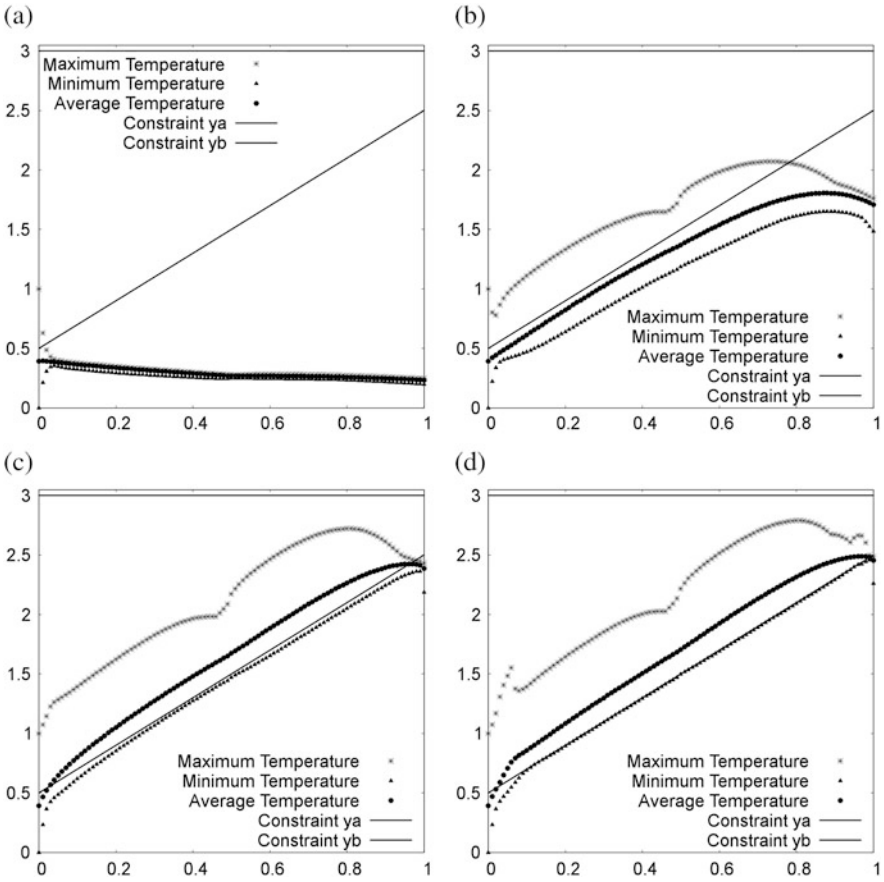


Fig. 2 Test 1: Temperature behaviour at each time-step for different ε. (**a**) $\varepsilon = 1$. (**b**) $\varepsilon = 0.1$. (**c**) $\varepsilon = 0.01$. (**d**) $\varepsilon = 0.001$

Table 1 Test 1: results for the FE discretization for different ε

Spatial discretization	ε	$\hat{J}(z)$	$\|\varepsilon w\|_{\mathcal{W}}$	Iterations
FE	1.0	0.931	1.3563	4
FE	0.1	7.584	0.2874	7
FE	0.01	9.066	0.0216	9
FE	0.001	120.329	0.0150	21

of basis and for different ε. In the last column, we have listed the values of the a-posteriori estimate for the difference $\|u^{\mathrm{FE}} - u^{\mathrm{POD}}\|$, which is defined as

$$\|u^{\mathrm{FE}} - u^{\mathrm{POD}}\|^2 = \sum_{i=1}^{m} \|u_i^{\mathrm{FE}} - u_i^{\mathrm{POD}}\|^2_{L^2(0,T)}.$$

Table 2 Test 1: results for the POD-M2 discretization for different ε and same number of basis

Spatial discretization	ε	rel-err(T)	rel-err	rel-err(Act.S.)	$\|u^{\mathrm{FE}} - u^{\mathrm{POD}}\|$	Err.Est.
POD-M2-10 Basis	1.000	0.002	0.003	0	0.0003	0.0004
POD-M2-10 Basis	0.100	0.006	0.004	0.001	0.0076	0.0167
POD-M2-10 Basis	0.010	0.004	0.007	0.024	0.3705	3.3604
POD-M2-10 Basis	0.001	0.700	0.648	0.465	7.359	$\simeq 2 \cdot 10^5$

We also need to clarify how we have computed the relative errors:

$$\begin{aligned}
\text{rel-err}(T) &= \|y^{\mathrm{FE}}(T) - y^{\mathrm{POD}}(T)\|_H / \|y^{\mathrm{FE}}(T)\|_H, \\
\text{rel-err} &= \|y^{\mathrm{FE}} - y^{\mathrm{POD}}\|_{L^2(0,T;H)} / \|y^{\mathrm{FE}}\|_{L^2(0,T;H)}, \\
\text{rel-err(Act.S.)} &= \left| \mathcal{A}^{\mathrm{FE}} \cup \mathcal{A}^{\mathrm{POD}} - \mathcal{A}^{\mathrm{FE}} \cap \mathcal{A}^{\mathrm{POD}} \right| / \left(N_x N_t \right),
\end{aligned}$$

where $\mathcal{A}^{FE} = \left(\mathcal{A}_{\mathsf{a}}^{\mathcal{W}} \cup \mathcal{A}_{\mathsf{b}}^{\mathcal{W}}\right)(z^{FE})$ and N_t is the number of time steps. The rel-err(Act.S.) in particular points out how much the active sets of state constraints related to the optimal solution computed with the reduced order model are far to the one computed in the FE discretization. As one can see, the POD approximation gets worse as ε decreases. For example, for $\varepsilon = 0.001$ the optimal control computed with the reduced order model is completely far from the one computed with the full order discretization. This is justified from the fact that there are more dynamics to approximate for smaller ε, since the number of iterations of the PDASS algorithm is greater. If we want to obtain, for example, an approximation error less than 0.01 in the case of POD-M2 we have to take at least 4 basis for $\varepsilon = 1$, 9 for $\varepsilon = 0.1$, 28 for $\varepsilon = 0.01$ and 58 for $\varepsilon = 0.001$. In addition, since in Theorem 5.1 w^{ap} depends on ε^{-1} and therefore also the error estimator, we have that its tightness depends on the regularization parameter. The previous statement is confirmed by the data reported in Table 2: the greater is ε the tighter is the error estimator. For example, for $\varepsilon = 1$ we have that it is only 1.3 times greater than the true error, instead it is 5.67 times the true one for $\varepsilon = 0.01$. From now to the end of the subsection, ε is fixed to 0.01. In Table 3 we present some results for Algorithm 1 for the FE and POD approximations using the two different strategies to build the POD bases. The norm of εw and also the cost functional gets closer to their values computed through the FE discretization as soon as the number of basis increases. Moreover, the PDASS algorithm applied to the reduced system converges almost in the same iterations' number of the full one. Even if we are able to solve the reduced linear system of Algorithm 1 around 80–100 times faster than the full one, the total algorithm speed-up is approximatively 4. This is due to the fact that we have to compute the active sets for the state constraints at each algorithm's iteration and this means that the reduced algorithm has to project into the FE discretization the approximated POD solution, compute the active sets, which costs $O(NN_t)$, and project back into the POD subspace those sets. To better compare POD-M1 and POD-M2 approaches, we

Table 3 Test 1: results for the FE and POD discretizations for $\varepsilon = 0.01$

Spatial discretization	POD basis elements	$\hat{J}(z)$	$\|\varepsilon w\|_{\mathcal{W}}$	rel-err(Act.S.)	Iterations	Speed-up
FE	–	9.066	0.0216	–	9	–
POD-M1	10	9.659	0.0339	0.127	10	3.91
POD-M1	15	9.123	0.0223	0.019	10	3.58
POD-M1	20	9.119	0.0221	0.010	9	3.48
POD-M2	10	9.181	0.0252	0.024	9	4.01
POD-M2	15	9.090	0.0229	0.014	9	3.90
POD-M2	20	9.076	0.0218	0.003	9	3.45

Table 4 Test 1: error values for the POD suboptimal solutions

Spatial discretization	POD basis elements	rel-err(T)	rel-err	$\|u^{\mathrm{FE}} - u^{\mathrm{POD}}\|$	Error estimator
POD-M1	10	0.068	0.115	1.344	7.620
POD-M1	15	0.003	0.004	0.174	2.361
POD-M1	20	0.003	0.003	0.136	1.549
POD-M2	10	0.004	0.007	0.371	3.360
POD-M2	15	0.003	0.002	0.128	1.321
POD-M2	20	0.001	0.001	0.065	0.179

also report the relative errors between the solution computed with the full and the reduced systems in Table 4. From this table, as expected, we can notice that the POD basis generated with the optimal solution performs better than the other basis: when the algorithm is getting closer to the optimal control, the information brought by the optimal snapshots is more helpful than the one brought by snapshots generated with an arbitrary control, which is usually far from the optimal one. This is also clear in Fig. 3, where we plot the differences between the optimal controls computed solving the full system and the reduced ones for 20 POD basis: the controls computed with POD-M2 are closer to the FE optimal controls than the ones obtained using the POD-M1 reduced system. This explains why we need an a-posteriori error estimator for the POD basis: we can estimate the quality of our basis and we can decide to consider a greater number of basis or to generate new basis from a different initial control. In Fig. 4, we show the comparison between the true error $\|u^{\mathrm{FE}} - u^{\mathrm{POD}}\|$ and the a-posteriori error estimator. Due to the previous discussion on the quality of the POD approximation, we can notice that as expected it is tighter for POD-M2 than for POD-M1 and it becomes for both approximations tighter and smaller as soon as the number of POD basis increases, although with some oscillation.

6.2 Test 2: Cost of Tracking Type

For the second test, we fix $\varepsilon = 0.1$ and we use the same data of Test 1, except for the cost functional weights which are chosen in the following way: $\sigma_T = \sigma_Q = 1$

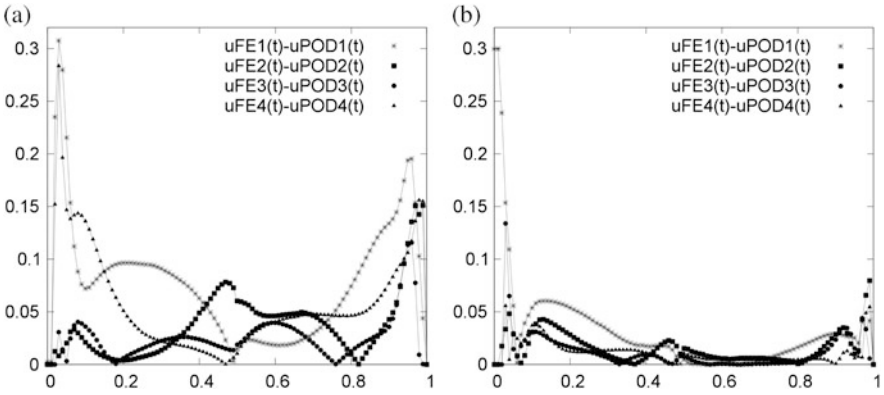


Fig. 3 Test 1: $|u^{FE}(t) - u^{POD}(t)|$ with $\ell = 20$ basis functions. (**a**) POD-M1. (**b**) POD-M2

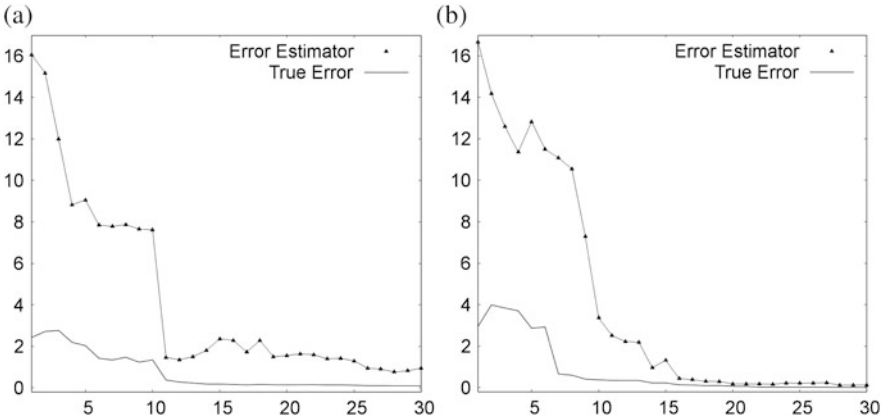


Fig. 4 Test 1: comparison between $\|u^{\mathrm{FE}} - u^{\mathrm{POD}}\|$ and its a-posteriori error estimate. (**a**) POD-M1. (**b**) POD-M2

and $\sigma = 0.01$. Regarding σ_w, we split the section in two parts: as first we study the model's behaviour when its value decreases, then we investigate the case $\sigma_w = 0$. Regarding this last condition, we want to point out that in the continuous model the terms connected to σ_w in the cost functional, adjoint equation and in the error estimator are zero: this means that w is not uniquely defined, since the only condition that w has to satisfy is $y_{\mathsf{a}}(t, \boldsymbol{x}) \leq \mathscr{E}y(t, \boldsymbol{x}) + \varepsilon w(t, \boldsymbol{x}) \leq y_{\mathsf{b}}(t, \boldsymbol{x})$ a.e. in Q, which clearly has no unique solution for fixed values of y_{a}, y_{b}, ε and y. By the way, due to the fact that $\sigma_w = 0$, we can observe that w is not more influencing the computation of the optimal control in the PDASS algorithm, so our optimal control will respect the control constraints and be the minimum of the reduced cost functional $\hat{J}$, but the solution may not be in the state constraints' range. From Table 5 we can noticed that the smaller σ_w is the more the algorithm focuses on reaching the target and the less on respecting the state constraints. In addition, when

Table 5 Test 2: results for the FE discretization for different σ_w

Spatial discretization	σ_w	$\hat{J}(z)$	$\|\varepsilon w\|_{\mathcal{W}}$	$\|y(T) - y_T\|$	$\|y - y_Q\|$	Iterations
FE	1.0000	0.318	0.015	0.159	0.618	9
FE	0.0100	0.311	0.036	0.156	0.624	5
FE	0.0001	0.309	0.161	0.155	0.623	4

Table 6 Test 2: results for the POD-M2 discretization for different σ_w and same number of basis

Spatial discretization	σ_w	rel-err(T)	rel-err	rel-err(Act.S.)	$\|u^{\text{FE}} - u^{\text{POD}}\|$	Err.Est.
POD-M2-10 Basis	1.0000	0.0019	0.0036	0.0014	0.1689	0.3051
POD-M2-10 Basis	0.0100	0.0013	0.0014	0.0007	0.0937	0.1456
POD-M2-10 Basis	0.0001	0.0013	0.0012	0.0005	0.0931	14.1898

σ_w decreases the conditions for the PDASS algorithm are less restrictive, therefore it uses less iteration to compute the solution. As showed in Table 6, also the POD-M2 approximation becomes better when σ_w gets smaller, but there is a worsening in the a-posteriori estimation: this is connected to the term σ_z^{-1} in (23), which makes the estimation increasing. For $\sigma_w = 0$ instead, we have a simplified error estimator, which produces better results compared to the case $\sigma_w > 0$ really small. As in Test 1, in Tables 7 and 8 we report the results of the finite elements solution (FE) and the reduced order ones (POD-M1,POD-M2) for $\sigma_w = 0$, with different choices of basis' number. As can be observed from Table 7, for this choice of parameters we have an improve of the speed-up gained in solving the reduced system, because in this context we do not have to compute the active sets for the state constraints. Therefore, we can have a speed-up for the algorithm similar to the one we get for solving the reduced linear system at each PDASS algorithm's step. In addition, the case $\sigma_w = 0.0001$ (or smaller) is equal to $\sigma_w = 0$, which is not surprising, since this means that already for this value of σ_w, we are almost ignoring the state constraints, due also to the choice on ε, but the advantage of taking $\sigma_w = 0$ is to have a tighter error estimator and a greater speed-up. As last, in this test it is confirmed that the number of POD basis functions needed to approximate the full order model really depends on the choice of the controls used for building the snapshots: we can see that for 4 basis, we can not capture in a good way the FE behaviour with POD-M1 basis, but with 10 basis we get results similar to POD-M2. The optimal trajectories at time $T = 1.0$ are reported in Fig. 5: we can notice that the FE and the POD-M2 ones are similar already for 7 basis, which is not the case for POD-M1.

7 Conclusions

With efficient building operation in mind, we have studied an optimal control problem of a parabolic convection-diffusion equations, with a time-dependent advection field, bilateral constraints for the boundary controls and pointwise state constraints, which have been treated with a Lavrentiev regularization. For solving

Table 7 Test 2: results for the FE and POD discretizations

Spatial discretization	POD basis elements	$\hat{J}(z)$	$\|y(T) - y_T\|$	$\|y - y_Q\|$	Iterations	Speed-up
FE	–	0.309	0.155	0.623	4	–
POD-M1	4	0.404	0.573	0.720	3	86.7
POD-M1	7	0.330	0.350	0.625	3	78.5
POD-M1	10	0.309	0.154	0.623	4	61.4
POD-M2	4	0.342	0.189	0.637	4	76.7
POD-M2	7	0.311	0.151	0.626	4	72.5
POD-M2	10	0.309	0.156	0.623	4	65.1

Table 8 Test 2: error values for the POD suboptimal solutions

Spatial discretization	POD basis elements	rel-err(T)	rel-err	$\|u^{\mathrm{FE}} - u^{\mathrm{POD}}\|$	Error estimator
POD-M1	4	0.091	0.087	1.711	6.170
POD-M1	7	0.056	0.026	0.421	0.781
POD-M1	10	0.002	0.002	0.103	0.166
POD-M2	4	0.044	0.062	1.416	5.770
POD-M2	7	0.004	0.004	0.200	0.379
POD-M2	10	0.001	0.001	0.093	0.142

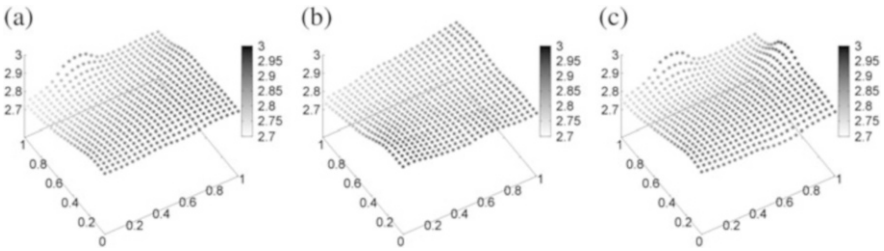


Fig. 5 Test 2: optimal trajectories at time $t = 1.0$. (**a**) FE. (**b**) POD-M1-7Basis. (**c**) POD-M2-7Basis

this optimal control problem we have applied the primal-dual active set strategy presented in [15], which has a super-linear rate of convergence. In order to speed-up the computational time of the algorithm, we have employed the POD method and utilized the a-posteriori error estimator in [9]. In the numerical test section, we have also shown how the variation of the regularization parameter ε and of the cost functional weight σ_w influences the behaviour of the solution and of the POD approximation. In addition, concerning the speed-up due to the POD method, we have noticed that this is reduced because of the computation of the state constraints' active sets, therefore it will be interesting in future work to treat the state constraints with other methods, e.g. the augmented Lagrangian algorithm. As shown in [22], the PDASS and its POD version can be combined with MPC, in order to face long-time horizon problems, which can be really costly to solve directly with the PDASS.

Acknowledgements The authors gratefully acknowledge support by the German Science Fund DFG grant VO 1658/4-1 *Reduced-Order Methods for Nonlinear Model Predictive Control*.

References

1. Afanasiev, K., Hinze, M.: Adaptive control of a wake flow using proper orthogonal decomposition. In: Shape Optimization and Optimal Design. Lecture Notes in Pure and Applied Mathematics, vol. 216, pp. 317–332. Marcel Dekker, New York (2001)
2. Arian, E., Fahl, M., Sachs, E.W.: Trust-region proper orthogonal decomposition for flow control. Technical Report 2000-25, ICASE (2000)

3. Balay, S., Gropp, W.D., Curfman McInnes, L., Smith, B.F.: Efficient management of parallelism in object oriented numerical software libraries. In: Arge, E., Bruaset, A.M., Langtangen, H.P. (eds.) Modern Software Tools in Scientific Computing, pp. 163–202. Birkhäuser Press, Basel (1997)
4. Balay, S., Abhyankar, S., Adams, M.F., Brown, J., Brune, P., Buschelman, K., Dalcin, L., Eijkhout, V., Gropp, W.D., Kaushik, D., Knepley, M.G., Curfman McInnes, L., Rupp, K., Smith, B.F., Zampini, S., Zhang, H.: PETSc Users Manual. ANL-95/11 - Revision 3.7. Argonne National Laboratory, Argonne (2016)
5. Banholzer, S., Beermann, D., Volkwein, S.: POD-based error control for reduced-order bicriterial PDE-constrained optimization. Annu. Rev. Control **44**, 226–237 (2017)
6. Berkooz, G., Holmes, P., Lumley, J.L.: Turbulence, Coherent Structures, Dynamical Systems and Symmetry. Cambridge Monographs on Mechanics. Cambridge University Press, Cambridge (1996)
7. Dautray, R., Lions, J.-L.: Mathematical Analysis and Numerical Methods for Science and Technology. Volume 5: Evolution Problems I. Springer, Berlin (2000)
8. Dontchev, A.L. , Hager, W.W., Poore, A.B., Yang, B.: Optimality, stability, and convergence in nonlinear control. Appl. Math. Optim. **31**, 297–326 (1995)
9. Grimm, E., Gubisch, M., Volkwein, S.: Numerical analysis of optimality-system POD for constrained optimal control. In: Recent Trends in Computational Engineering - CE2014: Optimization, Uncertainty, Parallel Algorithms, Coupled and Complex Problems. Lecture Notes in Computational Science and Engineering, vol. 105, pp. 297–317. Springer, Cham (2015)
10. Grüne, L., Pannek, J.: Nonlinear Model Predictive Control: Theory and Algorithms, 2nd edn. Springer, London (2017)
11. Gubisch, M.: Model order reduction techniques for the optimal control of parabolic partial differential equations with control and state constraints. Ph.D thesis, Department of Mathematics and Statistics, University of Konstanz. http://nbn-resolving.de/urn:nbn:de:bsz:352-0-355213 (2017)
12. Gubisch, M., Volkwein, S.: POD a-posteriori error analysis for optimal control problems with mixed control-state constraints. Comput. Optim. Appl. **58**, 619–644 (2014)
13. Gubisch, M., Volkwein, S.: Proper orthogonal decomposition for linear-quadratic optimal control. In: Ohlberger, M., Benner, P., Cohen, A., Willcox, K. (eds.) Model Reduction and Approximation: Theory and Algorithms, pp. 5–66. SIAM, Philadelphia (2017)
14. Hernandez, V., Roman, J.E., Vidal, V.: SLEPc: a scalable and flexible toolkit for the solution of eigenvalue problems. ACM Trans. Math. Softw. **31**(3), 351–362 (2005). http://dx.doi.org/10.1145/1089014.1089019
15. Hintermüller, M., Ito, K., Kunisch, K.: The primal-dual active set strategy as a semismooth Newton method. SIAM J. Optim. **13**, 865–888 (2002)
16. Hintermüller, M., Kopacka, I., Volkwein, S.: Mesh-independence and preconditioning for solving control problems with mixed control-state constraints. ESAIM: COCV **15**, 626–652 (2009)
17. Hinze, M. , Pinnau, R., Ulbrich, M., Ulbrich, S.: Optimization with PDE Constraints. Springer, Berlin (2009)
18. Ito, K., Kunisch, K.: Lagrange Multiplier Approach to Variational Problems and Applications. SIAM, Philadelphia (2008)
19. Krumbiegel, K., Rösch, A.: A virtual control concept for state constrained optimal control problems. Comput. Optim. Appl. **43**, 213–233 (2009)
20. Kunisch, K., Volkwein, S.: Proper orthogonal decomposition for optimality systems. ESAIM: M2AN **42**, 1–23 (2008)
21. Lions, J.L.: Optimal Control of Systems Governed by Partial Differential Equations. Springer, Berlin (1971)
22. Mechelli, L., Volkwein, S.: POD-based economic model predictive control for heat convection phenomena. In: Radu, F.A., Kumar, K., Berre, I., Nordbotten, J.M., Pop, I.S. (eds.) Numerical Mathematics and Advanced Applications ENUMATH 2017. Springer (2018)

23. Roman, J.E., Campos, C., Romero, E., Tomas, A.: SLEPc Users Manual. DSIC-II/24/02 – Revision 3.7. D. Sistemes Informàtics i Computació, Universitat Politècnica de València (2016)
24. Tröltzsch, F.: Regular Lagrange multipliers for control problems with mixed pointwise control-state constraints. SIAM J. Optim. **22**, 616–635 (2005)
25. Tröltzsch, F.: Optimal Control of Partial Differential Equations. Theory, Methods and Applications. American Mathematical Society, Providence (2010)
26. Tröltzsch, F., Volkwein, S.: POD a-posteriori error estimates for linear-quadratic optimal control problems. Comput. Optim. Appl. **44**, 83–115 (2009)
27. Ulbrich, M.: Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces. SIAM, Philadelphia (2011)

Order Reduction Approaches for the Algebraic Riccati Equation and the LQR Problem

Alessandro Alla and Valeria Simoncini

Abstract We explore order reduction techniques to solve the algebraic Riccati equation (ARE), and investigate the numerical solution of the linear-quadratic regulator problem (LQR). A classical approach is to build a low dimensional surrogate model of the dynamical system, for instance by means of balanced truncation, and then solve the corresponding ARE. Alternatively, iterative methods can be used to directly solve the ARE and use its approximate solution to estimate quantities associated with the LQR. We propose a class of Petrov-Galerkin strategies based on Krylov subspaces that simultaneously reduce the dynamical system while approximately solving the ARE by projection. This methodology significantly generalizes a recently developed Galerkin method, based on Krylov subspaces, by using a pair of projection spaces, as it is often done in model order reduction (MOR) of dynamical systems. Numerical experiments illustrate the advantages of the new class of methods over classical approaches when dealing with large matrices.

Keywords Model order reduction · Iterative schemes · Riccati equation · LQR

Part of this work was supported by the Indam-GNCS 2017 Project "Metodi numerici avanzati per equazioni e funzioni di matrici con struttura". The author "Valeria Simoncini" is a member of the Italian INdAM Research group GNCS.

A. Alla (✉)
Department of Mathematics, PUC-Rio, Rio De Janeiro, Brazil
e-mail: alla@mat.puc-rio.br

V. Simoncini
Alma Mater Studiorum - Universita' di Bologna, Bologna, Italy

IMATI-CNR, Pavia, Italy
e-mail: valeria.simoncini@unibo.it

M. Falcone et al. (eds.), *Numerical Methods for Optimal Control Problems*,
Springer INdAM Series 29, https://doi.org/10.1007/978-3-030-01959-4_5

1 Introduction

Optimal control problems for partial differential equations (PDEs) are an extremely important topic for many industrial applications in different fields, from aerospace engineering to economics. The problem has been investigated with different strategies as open-loop (see e.g. [23]) or closed-loop (see e.g. [18, 19]).

In this work we are interested in feedback control for linear dynamical systems and quadratic cost functionals which is known as the Linear Quadratic Regulator (LQR) problem. Although most models are nonlinear, LQR is still a very interesting and powerful tool, for instance in the stabilization of nonlinear models under perturbations, where a control in feedback form can be employed.

The computation of the optimal policy in LQR problems requires the solution of an algebraic Riccati equation (ARE), a quadratic matrix equation with the dimension of the dynamical system. This is a major bottleneck in the numerical treatment of the optimal control problem, especially for high dimensional systems such as those stemming from the discretization of a PDE.

Several powerful solution methods for the ARE have been developed throughout the years for small dynamical systems, based on spectral decompositions. The large scale case is far more challenging, as the whole spectral space of the relevant matrices cannot be determined because of memory and computational resource limitations. For these reasons, this algebraic problem is a very active research topic, and major contributions have been given in the past decade. Different approaches have been explored: variants of the Newton method have been largely employed in the past [11, 27], while only more recently reduction type methods based on Krylov subspaces have emerged as a feasible effective alternative; see, e.g., [12, 24, 38] and references therein. The recent work [36] shows that a Galerkin class of reduction methods onto Krylov subspaces for solving the ARE can be naturally set into the MOR framework for the original dynamical system. This fact is particularly striking because most literature has so far treated the solution of the Riccati equation as a *distinct* problem from the reduction process, whereas it is now clear that the MOR perspective provides a natural setting also for the solution of the Riccati equation. In the context of linear-quadratic optimal regulator problems, H_2 and H_∞ controller design and balancing-related model reduction often an approximation to the Riccati solution matrix is assumed to be available, from which other key quantities are determined; see the discussion in [36]. By exploiting the same space for the reduction of the model *and* the projection of the Riccati equation, it is possible to determine an approximate control function satisfying certain optimality property [36]. Here we deepen the MOR connection by using the Petrov-Galerkin Krylov subspaces commonly used in MOR to directly approximate the Riccati solution.

As already mentioned, the LQR problem is more complicated when dealing with PDEs because its discretization leads to a very large system of ODEs and, as a consequence, the numerical solution of the ARE is computationally more demanding. To significantly lower these computational costs and memory requirements,

model order reduction techniques can be employed. Here, we distinguish between two different concepts of reduction approaches.

A first methodology projects the dynamical system into a low dimensional system whose dimensions are much smaller than the original one; see, e.g., [14]. Therefore, the corresponding reduced ARE is practical and feasible to compute on a standard computer. The overall methodology thus performs a *first-reduce-then-solve* strategy. This approach has been investigated with different model reduction techniques like Balanced Truncation (BT) in e.g.[4], Proper Orthogonal Decomposition (POD, [39, 40]) in e.g. [5, 28] and via the interpolation of the rational functions, see e.g. [7, 15, 20]. A different approach has been proposed in [35] where the basis functions are computed from the solution of the high dimensional Riccati equation in a many query context. For the sake of completeness, we would like to mention that model order reduction has been applied to the Linear Quadratic Gaussian (LQG) problem which may involve the solution of two Riccati equations (see e.g. [9, 10, 16]).

We note that basis generation in the context of model order reduction for optimal control problems is an active research topic (see e.g. [2, 3, 28, 30]). Furthermore, the computation of the basis functions is made by a Singular Value Decomposition (SVD) of the high dimensional data which can be very expensive. One way to overcome this issue was proposed in [1] by means of randomized SVD which is a fast and accurate alternative to the SVD, and it is based on random samplings.

A second methodology follows a *reduce-while-solve* strategy. In this context, recent developments aim at reducing the original problem by subspace projection, and determining an approximate solution in a low dimensional approximation space. Proposed strategies either explicitly reduce the quadratic equation (see, e.g., [38] and references therein), or approximately solve the associated invariant subspace problem (see, e.g., [8] and its references). As already mentioned, these recently developed methods have shown to be effective alternatives to classical variants of the Newton method, which require the solution of a linear matrix equation at each nonlinear iteration; see, e.g., [13] for a general description.

The aim of this paper is to discuss and compare the aforementioned MOR methodologies for LQR problems. In particular, we compare the two approaches of reducing the dynamical system first versus building surrogate approximation of the ARE directly, using either Galerkin or Petrov-Galerkin projections. The idea of using a Petrov-Galerkin method for the ARE appears to be new, and naturally expands the use of two-bases type as typically employed for transfer function approximation.

To set the paper into perspective we start recalling LQR problem and its order reduction in Sect. 2. In Sect, 3 we describe reduction strategies of dynamical systems used in the small size case, such as proper orthogonal decomposition and balanced truncation. Section 4 discusses the new class of projection strategies that attack the Riccati equation, while delivering a reduced model for the dynamical system. Finally, numerical experiments are shown in Sect. 5 and conclusions are derived in Sect. 6.

2 The Linear-Quadratic Regulator Problem and Model Order Reduction

In this section we recall the mathematical formulation of the LQR problem. We refer the reader for instance to classical books such as e.g. [31] for a comprehensive description of what follows. We consider a linear time invariant system of ordinary differential equations of dimension n:

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \quad x(0) = x_0, \quad t > 0, \\ y(t) &= Cx(t) + Du(t), \end{aligned} \tag{1}$$

with $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$ and $D \in \mathbb{R}^{p\times m}$. Usually, $x(t) : [0, \infty] \to \mathbb{R}^n$ is called the state, $u(t) : [0, \infty] \to \mathbb{R}^m$ the input or control and $y(t) : [0, \infty] \to \mathbb{R}^p$ the output. Furthermore, we assume that A is passive. This may be viewed as a restrictive hypothesis, since the problems we consider only require that (A, B) are stabilizable and (A^T, C^T) controllable, however this is convenient to ensure that the methods we analyze are well defined. In what follows, without loss of generality, we will consider $D \equiv 0$. We also define the transfer function for later use:

$$G(s) = C(sI - A)^{-1}B. \tag{2}$$

Next, we define the quadratic cost functional for an infinite horizon problem:

$$J(u) := \int_0^\infty y(t)^T y(t) + u(t)^T Ru(t)\, dt, \tag{3}$$

where $R \in \mathbb{R}^{m\times m}$ is a symmetric positive definite matrix. The optimal control problem reads:

$$\min_{u\in\mathbb{R}^m} J(u) \quad \text{such that} \quad x(t) \quad \text{solves} \quad (1). \tag{4}$$

The goal is to find a control policy in feedback form as:

$$u(t) = -Kx(t) = -R^{-1}B^T Px(t), \tag{5}$$

with the feedback gain matrix $K \in \mathbb{R}^{m\times n}$ and $P \in \mathbb{R}^{n\times n}$ is the unique symmetric and positive (semi-)definite matrix that solves the following ARE:

$$A^T P + PA - PBR^{-1}B^T P + C^T C = 0, \tag{6}$$

which is a quadratic matrix equation for the unknown P.

We note that the numerical approximation of Eq. (6) can be very expensive for large n. Therefore, we aim at the reduction of the numerical complexity by projection methods.

Let us consider a general class of tall matrices $V, W \in \mathbb{R}^{n\times r}$, whose columns span some approximation spaces. We chose these two matrices such that they are biorthogonal, that is $W^T V = I_r$. Let us now assume that the matrix P, solution of (6), can be approximated as

$$P \approx W P_r W^T.$$

Then the residual matrix can be defined as

$$\mathscr{R}(P_r) = A^T W P_r W^T + W P_r W^T A - W P_r W^T B R^{-1} B^T W P_r W^T + C^T C.$$

The small dimensional matrix P_r can be determined by imposing the so-called Petrov-Galerkin condition, that is orthogonality of the residual with respect to range(V), which in matrix terms can be stated as $V^T \mathscr{R}(P_r) V = 0$. Substituting the residual matrix and exploiting the bi-orthogonality of V and W we obtain:

$$A_r^T P_r + P_r A_r - P_r B_r R^{-1} B_r^T P_r + C_r^T C_r = 0, \tag{7}$$

where

$$A_r = W^T A V, \quad B_r = W^T B, \quad C_r = C V.$$

It can be readily seen that Eq. (7) is again a matrix Riccati equation, in the unknown matrix $P_r \in \mathbb{R}^{r\times r}$, of much smaller dimension than P, provided that V and W generate small spaces. We refer to this equation as the *reduced Riccati equation*. The computation of P_r allows us to formally obtain the approximate solution $W P_r W^T$ to the original Riccati equation (6), although the actual product is never computed explicitly, as the approximation is kept in factorized form.

The optimal control for the reduced problem reads

$$u_r(t) = -K_r x_r(t) = -R^{-1} B_r^T P_r x_r(t)$$

with the reduced feedback gain matrix given by $K_r = K V \in \mathbb{R}^{m\times r}$. Note that this $u_r(t)$ is different from the one obtained by first approximately solving the Riccati equation and with the obtained matrix defining an approximation to $u(t)$; see [36] for a detailed discussion.

The Galerkin approach is obtained by choosing $V = W$ with orthonormal columns when imposing the condition on the residual.

To reduce the dimension of the dynamical system (1), we assume to approximate the full state vector as $x(t) \approx V x_r(t)$ with a basis matrix $V \in \mathbb{R}^{n\times r}$, where $x_r(t) : [0, \infty) \to \mathbb{R}^r$ are the reduced coordinates. Plugging this ansatz into the dynamical system (1), and requiring a so called Petrov-Galerkin condition yields

$$\begin{aligned} \dot{x}_r(t) &= A_r x_r(t) + B_r u(t), \quad x_r(0) = W^T x_0, \quad t > 0, \\ y_r(t) &= C_r x_r. \end{aligned} \tag{8}$$

The reduced transfer function is then given by:

$$G_r(s) = C_r(sI_r - A_r)^{-1}B_r. \tag{9}$$

The presented procedure is a generic framework for model reduction. It is clear that the quality of the approximation depends on the approximation properties of the reduced spaces. In the following sections, we will distinguish between the methods that directly compute V, W upon the dynamical systems (see Sect. 3) and those that readily reduce the ARE (see Sect. 4). In particular, for each method we will discuss both Galerkin and Petrov-Galerkin projections, to provide a complete overview of the methodology. The considered general Petrov-Galerkin approach for the ARE appears to be new.

3 Reduction of the Dynamical System

In this section we recall two well-known techniques as POD and BT to compute the projectors W, V starting from the dynamical systems.

3.1 Proper Orthogonal Decomposition

A common approach is based on the snapshot form of POD proposed in [39], which works as follows. We compute a set of snapshots $x(t_1), \ldots, x(t_k)$ of the dynamical system (1) corresponding to a prescribed input $\bar{u}(t)$ and different time instances $t_1, \ldots, t_k$ and define the POD ansatz of order r for the state $x(t)$ by

$$x(t) \approx \sum_{i=1}^{r} (x_r)_i(t)\psi_i, \tag{10}$$

where the basis vectors $\{\psi_i\}_{i=1}^r$ are obtained from the SVD of the snapshot matrix $X = [x(t_1), \ldots, x(t_k)]$, i.e. $X = \Psi\Sigma\Gamma^T$, and the first r columns of $\Psi = (\Psi_1, \ldots, \Psi_n)$ form the POD basis functions of rank r. Hence we choose the basis vectors $V = W = (\Psi_1, \ldots, \Psi_r)$ for the reduction in (8). Then, the evolution dynamics can be projected using a Galerkin method as in [29].

This technique strongly relies on the choice of a given input u, whose optimal selection is usually unknown. In this work, we decide to collect snapshots following the approach suggested in [28] as considers a linearization of the ARE (which corresponds to a Lyapunov equation). Therefore, the snapshots are computed by the following equation:

$$\dot{x}(t) = A^T x(t), \quad x(0) = c_i, \quad \text{for } i = 1, \ldots, p, \tag{11}$$

where c_i is the i-th column of the matrix C. The advantage of this approach is that Eq. (11) is able to capture the dynamics of the adjoint equation which is directly related to the optimality conditions, and we do not have to choose a reference input $\bar{u}(t)$. In order to obtain the POD basis, one has to simulate the high dimensional system and subsequently perform a SVD. As a consequence, the computational cost may become prohibitive for large scale problems. Algorithm 1 summarizes the method.

Algorithm 1 POD method to compute the reduced Riccati

Require: A, C, r
1: **for** $i = 1, \ldots, p$ **do**
2: Simulate system (11) with initial condition c_i.
3: Build the snapshots matrix $X = [X, x_i(t_1), \ldots, x_i(t_k)]$
4: **end for**
5: Compute the reduced SVD of $X = V \Sigma W^T$
6: Solve the reduced Riccati equation (7) for P_r.

3.2 Balanced Truncation

The BT method is a well-established model order reduction technique for linear time invariant systems (1). We refer to [4] for a complete description of the topic. It is based on the solution of the reachability Gramian R and the observability Gramian O which solve, respectively, the following Lyapunov equations

$$AR + RA^T + BB^T = 0, \quad A^{TO} + OA + C^T C = 0. \tag{12}$$

We determine the Cholesky factorization of the Gramians

$$R = \Phi \Phi^T \qquad O = \Upsilon \Upsilon^T. \tag{13}$$

Then, we compute the reduced SVD of the Hankel operator $\Upsilon^T \Phi$ and set

$$W = \Upsilon U \Sigma^{1/2}, \qquad V = \Upsilon V \Sigma^{1/2},$$

where $U, V \in \mathbb{R}^{n \times r}$ are the first r columns of the left and right singular vectors of the Hankel operator and $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$ matrix of the first r singular values.

The idea of BT is to neglect states that are both, hard to reach and hard to observe. This is done by eliminating states that correspond to low Hankel singular values σ_i. This method is very popular in the small case regime (e.g. $n \approx O(10^3)$) also because the whole procedure can be verified by a-priori error bounds in several system norms, and the Lyapunov equations can be solved very efficiently. In the large scale these equations need to be solved approximately; see, e.g., [12, 17].

In summary, the procedure first solves the two Lyapunov equations at a given accuracy for large matrices and then determines biorthogonal bases for the reduction

spaces by using a combined spectral decomposition of the obtained solution matrices. For consistency with the other projection strategies, in the large scale setting we solve Eq. (12) using adaptive rational Krylov subspaces as in [17]. We refer to [37] for alternatives. The algorithm is summarized in Algorithm 2.

Algorithm 2 BT method to compute the reduced Riccati

Require: A, B, C and the dimension of the reduced problem r
1: Compute R, O from (12) and their Cholesky factorization (13)
2: Compute the reduced SVD of the Hankel operator
3: Set $W = \Upsilon U \Sigma^{1/2}, \qquad V = \Upsilon V \Sigma^{1/2},$
4: Solve the reduced Riccati equation (7) for P_r.

4 Adaptive Reduction of the Algebraic Riccati Equation

In the previous section the reduced problem was obtained by a sequential procedure: first system reduction of a fixed order r and then solution of the reduced Riccati equation (7). A rather different strategy consists of determining the reduction bases *while* solving the Riccati equation. In this way, we combine both the reduction of the original system and of the ARE. While the reduction bases V and W are being generated by means of some iterative strategy, it is immediately possible to obtain a reduced Riccati equation by projecting the problem onto the current approximation spaces. The quality of the two spaces can be monitored by checking how well the Riccati equation is solved by means of its residual; if the approximation is not satisfactory, the spaces can be expanded and the approximation improved.

The actual space dimensions are not chosen a-priori, but tailored with the accuracy of the approximate Riccati solution. Mimicking what is currently available in the linear equation literature, the reduced problem can be obtained by imposing some constraints that uniquely identify an approximation. The idea is very natural and it was indeed presented in [24], where a standard Krylov basis was used as approximation space. However, only more recently, with the use of rational Krylov bases, a Galerkin approach has shown its real potential as a solver for the Riccati equation; see, e.g., [22, 38]. A more general Petrov-Galerkin approach was missing. We aim to fill this gap. In the following we give more details on these procedures.

Given an approximate solution $\tilde{P}$ of (6) written as $\tilde{P} = WYW^T$ for some Y to be determined, the former consists to require that the residual matrix is orthogonal to this same space, range(W), so that in practice $V = W$. The Petrov-Galerkin procedure imposes orthogonality with respect to the space range(V), where V is different from W, but with the same number of columns.

In [24] a first implementation of a Galerkin procedure was introduced, and the orthonormal columns of V spanning the (block) Krylov subspace $\mathscr{K}_r(A^T, C^T) = \text{range}([C^T, A^TC^T, \ldots, (A^T)^{r-1}C^T])$; see also [26] for a more detailed treatment and for numerical experiments. Clearly, this definition generates a sequence of

nested approximation spaces, that is $\mathscr{K}_r(A^T, C^T) \subseteq \mathscr{K}_{r+1}(A^T, C^T)$, whose dimension can be increased iteratively until a desired accuracy is achieved. More recently, in [22] and [38], rational Krylov subspaces have been used, again in the Galerkin framework. In particular, the special case of the *extended* Krylov subspace $\mathscr{K}_r(A^T, C^T) + \mathscr{K}_r((A^T)^{-1}, (A^T)^{-1}C^T)$ was discussed in [22], while the fully rational space

$$\mathscr{K}_r(A^T, C^T, \boldsymbol{\sigma}) := \text{range}([C^T, (A^T - \sigma_2 I)^{-1}C^T, \ldots, (A^T - \sigma_2 I)^{-1} \cdots (A^T - \sigma_r I)^{-1}C^T])$$

was used in [38]. The rational shift parameters $\boldsymbol{\sigma} = \{\sigma_2, \ldots, \sigma_r\}$ can be computed on the fly at low cost, by adapting the selection to the current approximation quality [17]. Note that $\dim(\mathscr{K}_r(A^T, C^T, \boldsymbol{\sigma})) \leq rp$, where p is the number of columns of C^T. In [38] it was also shown that a fully rational space can be more beneficial than the extended Krylov subspace for the Riccati equation. In the following section we are going to recall the general procedure associated with the Galerkin approach, and introduce the algorithm for the Petrov-Galerkin method, which to the best of the authors' knowledge is new. In both cases we use the fully rational Krylov subspace with adaptive choice of the shifts.

It is important to realize that $\mathscr{K}_r(A^T, C^T, \boldsymbol{\sigma})$ does not depend on the coefficient matrix BB^T of the second-order term in the Riccati equation. Nonetheless, experimental evidence shows good performance. This issue was analyzed in [33, 36] where however the use of the matrix B during the computation of the parameters was found to be particularly effective; a justification of this behavior was given in [36]. In the following we thus employ this last variant when using rational Krylov subspaces. More details will be given in the next section.

4.1 Galerkin and Petrov-Galerkin Riccati

In the Galerkin case, we will generate a matrix W whose columns span the rational Krylov subspace $\mathscr{K}_r(A^T, C^T, \boldsymbol{\sigma})$ in an iterative way, that is one block of columns at the time. This can be obtained by an Arnoldi-type procedure; see, e.g., [4]. The algorithm, hereafter GARK for Galerkin Adaptive Rational Krylov, works as follows:

Algorithm 3 GARK method to compute the reduced Riccati equation

Require: $A, C, \boldsymbol{\sigma}$
1: **for** $r = 1, 2, \ldots$ **do**
2: Expand the space $\mathscr{K}_r(A^T, C^T, \boldsymbol{\sigma})$;
3: Update the reduced matrices A_r, B_r and C_r with the newly generated vectors;
4: Solve the reduced Riccati equation for P_r;
5: Check the norm of the residual matrix $\mathscr{R}(P_r)$
6: If satisfied stop with P^r and the basis W of $\mathscr{K}_r(A^T, C^T, \boldsymbol{\sigma})$.
7: **end for**

The residual norm can be computed cheaply without the actual computation of the residual matrix; see, e.g., [38]. The parameters σ_j can be computed adaptively as the space grows; we refer the reader to [17] and [36] for more details.

In the general Petrov-Galerkin case, the matrix W is generated the same way, while we propose to compute the columns of V as the basis for the rational Krylov subspace $\mathscr{K}_r(A, B, \boldsymbol{\sigma})$; note that the starting block is now B, and the coefficient matrix is the transpose of the previous one. The two spaces are now constructed and expanded at the same time, so that the two bases can be enforced to be biorthogonal while they grow. For completeness, we report the algorithm in the Petrov-Galerkin setting in Algorithm 4 (hereafter PGARK for Petrov-Galerkin Adaptive Rational Krylov).

Algorithm 4 PGARK method to compute the reduced Riccati equation

Require: $A, B, C, \boldsymbol{\sigma}$
1: **for** $r = 1, 2, \ldots$ **do**
2: Expand the spaces $\mathscr{K}_r(A^T, C^T, \boldsymbol{\sigma})$, $\mathscr{K}_r(A, B, \boldsymbol{\sigma})$;
3: Update the reduced matrices A_r, B_r and C_r with the newly generated vectors;
4: Solve the reduced Riccati equation for P_r;
5: Check the norm of the residual matrix $\mathscr{R}(P_r)$
6: If satisfied stop with P^r and the basis W of $\mathscr{K}_r(A^T, C^T, \boldsymbol{\sigma})$.
7: **end for**

The parameters σ_j are computed for one space and used also for the other space. In this more general case, the formula for the residual matrix norm is not as cheap as for the Galerkin approach. We suggest the following procedure. We first recall that for rational Krylov subspace $\mathscr{K}_r(A, B, \boldsymbol{\sigma})$ the following relation holds (we assume here full dimension of the generated space after r iterations):

$$A^T W = W A_r + \hat{w} a_r^T, \quad a_r \in \mathbb{R}^{(r+1)m},$$

for certain vector $\hat{w}$ orthogonal to W, which changes as the iterations proceeds, that is as the number of columns W grows; we refer the reader to [32] for a derivation of this relation, which highlights that the distance of range(W) from an invariant subspace of A is measured in terms of a rank-one matrix. We write

$$\begin{aligned}
\mathscr{R}(P_r) &= A^T W P_r W^T + W P_r W^T A - W P_r W^T B R^{-1} B^T W P_r W^T + C^T C \\
&= W A_r P_r W^T + \hat{w} a_r^T P_r W^T + W P_r A_r^T W^T + W P_r a_r \hat{w}^T \\
&\quad - W P_r B_r R^{-1} B_r^T P_r W^T + W E E^T W^T \\
&= \hat{w} a_r^T P_r W^T + W P_r a_r \hat{w}^T \\
&= [W, \hat{w}] \begin{bmatrix} 0 & P_r a_r \\ a_r^T P_r & 0 \end{bmatrix} [W, \hat{w}]^T,
\end{aligned}$$

where we also used the fact that $C^T = WE$ for some matrix E. If $[W, \hat{w}]$ had orthonormal columns, as is the case for Galerkin, then $\|\mathscr{R}(P_r)\|^2 = 2\|P_r a_r\|^2$, which can be cheaply computed.

To overcome the nonorthogonality of $[W, \hat{w}]$, we suggest to perform a reduced QR factorization of $[W, \hat{w}]$ that maintains its columns orthogonal. This QR factorization does not have to be redone from scratch at each iteration, but it can be updated as the matrix W grows. If $[W, \hat{w}] = Q_W R_W$ with $R_W \in \mathbb{R}^{(r+1)m \times (r+1)m}$ upper triangular, then

$$\|\mathscr{R}(P_r)\| = \|R_W \begin{bmatrix} 0 & P_r a_r \\ a_r^T P_r & 0 \end{bmatrix} R_W^T\|.$$

The use of coupled (bi-orthogonal) bases has the recognized advantage of explicitly using both matrices C and B in the construction of the reduced spaces. This coupled basis approach has been largely exploited in the approximation of the dynamical system transfer function by solving a multipoint interpolation problem; see, e.g., [4] for a general treatment and [6] for a recent implementation. In addition, coupled bases can be used to simultaneously approximate both system Gramians leading to a large-scale BT strategy; see, e.g., [25] for early contributions using bi-orthogonal standard Krylov subspaces.[1] On the other hand, a Petrov-Galerkin procedure has several drawbacks associated with the construction of the two bases. More precisely, the two bi-orthogonal bases are generated by means of a Lanczos-type recurrence, which is known to have both stability and breakdown problems in other contexts such as linear system and eigenvalue solving. At any iteration it may happen that the new basis vectors w_j and v_j are actually orthogonal or quasi-orthogonal to each other, giving rise to a possibly incurable breakdown [21]. We have occasionally experienced this problem in our numerical tests, and it certainly occurs whenever $CB = 0$. In our specific context, an additional difficulty arises. The projected matrix $A_r = W^T A^T V$ is associated with a bilinear rather than a linear form, so that its field of values may be unrelated to that of A^T. As a consequence, it is not clear the type of hypotheses we need to impose on the data to ensure that the reduced Riccati equation (7) admits a unique stabilizable solution. Even in the case of A symmetric, the two bases will be different as long as $C \neq B^T$. All these questions are crucial for the robustness of the procedure and deserve a more throughout analysis which will be the topic of future research.

From an energy-saving standpoint, it is worth remarking that the Petrov-Galerkin approach uses twice as many memory allocations than the Galerkin approach, while performing about twice the number of floating point operations. In particular, constructing the two bases requires two system solves, with $A^T - s_j I$ and with $A - \bar{s}_j I$ respectively, at each iteration. Therefore, unless convergence is considerably

[1] We are unaware of any available implementation of rational Krylov subspace based approaches for large scale BT either with single or coupled bases, that simultaneously performs the balanced truncation while approximating the Gramians.

faster, the Petrov-Galerkin approach may not be superior to the Galerkin method in the solution of the ARE.

5 Numerical Experiments

In this section we present and discuss our numerical tests. We first compare the methods we have introduced in the previous sections on two test cases. Then we linger over the large scale implementation of BT we have adopted to make comparisons with projection based strategies, to highlight the difficulties that may arise when the approximation is performed in two separate steps.

We consider the discretization of the following linear PDE

$$\begin{aligned} w_t - \varepsilon \Delta w + \gamma w_x + \gamma w_y - cw = \mathbf{1}_{\Omega_B} u \quad & \text{in } \Omega \times (0, +\infty), \\ w(\cdot, 0) = w_0 \quad & \text{in } \Omega, \\ w(\cdot, t) = 0 \quad & \text{in } \partial\Omega \times (0, \infty), \end{aligned} \tag{14}$$

where $\Omega \subset \mathbb{R}^2$ is an open interval, $w : \Omega \times [0, \infty] \to \mathbb{R}^2$ denotes the state, and the parameters ε, γ and c are real positive constants. The initial value is w_0 and the function $\mathbf{1}_{\Omega_B}$ is the indicator function over the domain $\Omega_B \subset \mathbb{R}^2$. Note that we deal with zero Dirichlet boundary conditions. The problem in (14) includes the heat equation, for $\varepsilon \neq 0, \gamma = 0, c = 0$, reaction-diffusion equations for $\varepsilon \neq 0, \gamma = 0, c \neq 0$ and a class of convection-diffusion equations for $\varepsilon \neq 0, \gamma \neq 0$ and $c = 0$. Furthermore, we define an output of interest by:

$$s(t) := \frac{1}{|\Omega_C|} \int_{\Omega_C} w(x, t)\, dt, \tag{15}$$

where $\Omega_C \subset \mathbb{R}^2$. Space discretization of Eq. (14) by standard centered finite differences together with a rectangular quadrature rule for (15) lead to a system of the form (1). In general, the dimension n of the dynamical system (1) is rather large (i.e., $n \gg 1000$) and the numerical treatment of the corresponding ARE is computationally expensive or even unfeasible. Therefore, model order reduction is appropriate to lower the dimension of the optimal control problem (4). We will report experiments with small size problems, where all discussed methods can be employed, and with large size problems, where only the Krylov subspace based strategies are applied.

The numerical simulations reported in this paper were performed on a Mac-Book Pro with 1 CPU Intel Core i5 2.3 GHz and 8GB RAM and the codes are written in MatlabR2013a. In all our experiments, small dimensional Lyapunov and Riccati equations are solved by means of built-in functions of the Matlab Control Toolbox.

Whenever appropriate, the quality of the current approximation of the ARE is monitored by using the relative residual norm:

$$\mathtt{R}_P = \frac{\|\mathscr{R}(P_r)\|_F}{\|C\|_F^2}, \tag{16}$$

and the dimension of the surrogate model r is chosen such that $\mathtt{R}_P < 10^{-6}$. After the ARE solution is approximated the ultimate goal is to compute the feedback control (5). Therefore, we also report the error in the computation of feedback gain matrix K as the iterations proceed:

$$\mathscr{E}_K = \frac{\|K_r - K\|_F}{\|K\|_F}, \tag{17}$$

and we measure the quality of our surrogate model also by the $\mathscr{H}_2$-error

$$\mathscr{E}_G = \frac{\|G_r - G\|_{\mathscr{H}_2}}{\|G\|_{\mathscr{H}_2}}. \tag{18}$$

where

$$\|G(s)\|_{\mathscr{H}_2} := \frac{1}{2\pi}\left(\int_{-\infty}^{+\infty} \|G(i\omega)\|_F^2 d\omega\right)^{1/2}.$$

In particular, the approximation of the transfer function is one of the main targets of MOR, where the reduced system is used for analysis purposes, while the approximation of the feedback gain matrix is monitored to obtain a good control function.

5.1 Test 1: 2D Linear Heat Equation

In the first example we consider the linear reaction-diffusion equation. In (14) we chose $\gamma = 0, \varepsilon = 1, c = 400, \Omega = [0,1] \times [0,1]$, and $\Omega_B = [0.4, 0.6] \times [0.4, 0.6]$. In (1), the matrix A is obtained by centered five points finite difference discretization. We consider a small problem stemming from a spatial discretization step $\Delta x = 0.05$, leading to a system of dimension $n = 441$. The matrix C in (1) is given by the indicator function over the domain $\Omega_C = [0.3, 0.7] \times [0.3, 0.7]$ and $R \equiv I_m$ in (3).

The left panel of Fig. 1 shows the residual norm history of the reduced ARE (7) when the two projection matrices V, W are computed by each of the four algorithms explained in the previous sections. We can thus appreciate how the approximation proceeds as the reduced space is enlarged. We note that POD requires more basis

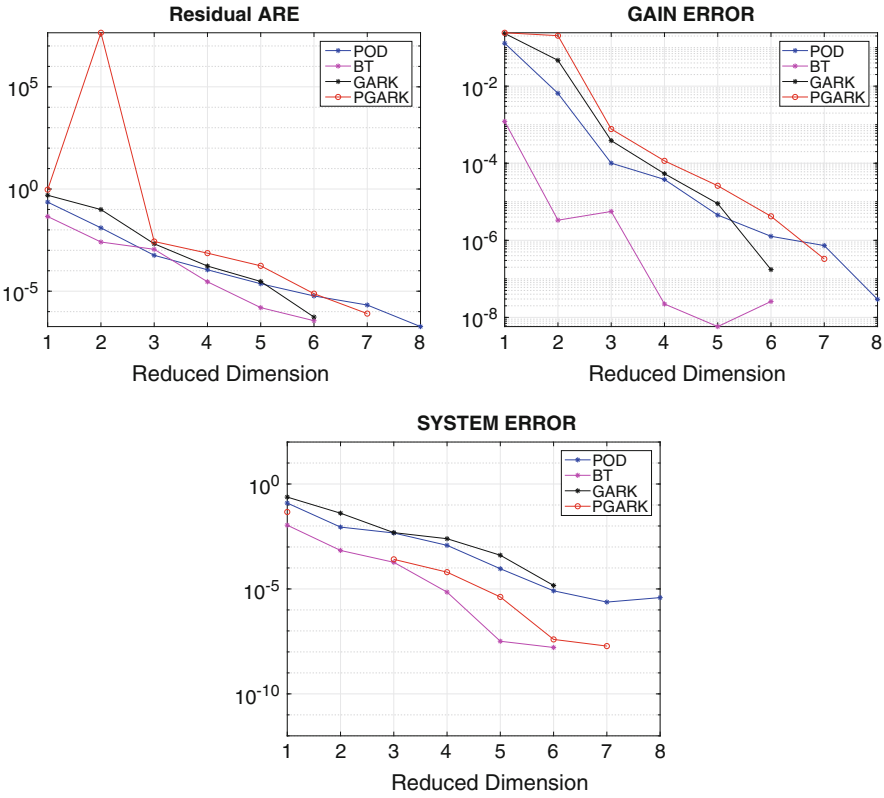


Fig. 1 Test 1: Convergence history of the relative residual norm $\mathtt{R}_P$ (left), Error $\mathscr{E}_K$ of the feedback gain matrix (middle), Error $\mathscr{E}_G$ for the approximation of reduced transfer function (right)

functions to achieve the desired tolerance for $\mathtt{R}_P$ than the other approaches, while the BT algorithm is the fastest. The POD method is a snapshot dependent method and thus it is crucially influenced by the choice of the initial input $u(t)$ and the results may be different for other choices of the snapshots set. All the other proposed techniques are, on the contrary, input/output independent. We note that under this setting the PGARK is not stable for the first reduced coordinates; see some additional comments on the issue in Test 2.

In the middle panel of Fig. 1, we show how well the feedback gain matrix K can be approximated with reduction methods. It is interesting to see that the basis functions computed by GARK and PGARK are able to approximate the matrix K very well. Furthermore, we note that BT does not decrease monotonically.

Finally, we would like to show the quality of the computed basis functions in the approximation of the dynamical system in the right panel of Fig. 1. In this example, BT approximates the transfer function very well with a lower number of basis functions.

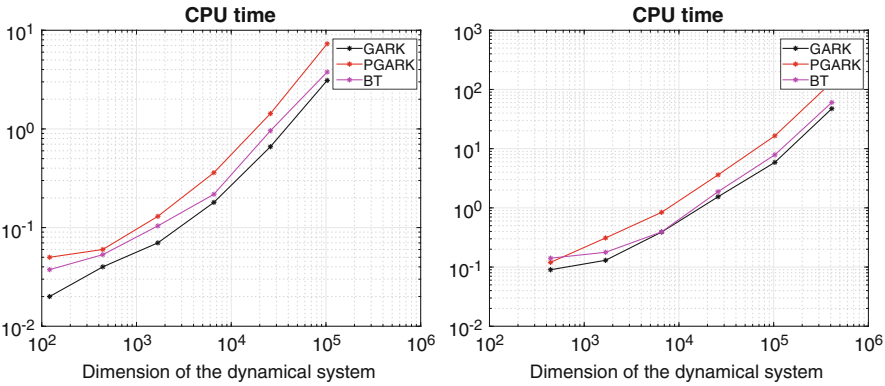


Fig. 2 CPU time as the problem dimension n increases for BT, GARK and PGARK. Left: Test 1. Right: Test 2

Last remark goes to the iterative methods GARK and PGARK . We showed that, although the basis functions are built upon information of the ARE, they are also able to approximate the dynamical systems and the feedback gain matrix. This is clearly not unexpected, as the generated approximation spaces are tightly related to classical model order reduction strategies with (rational) Krylov subspaces [4]. Nonetheless, as opposed to MOR methods, the quality of the generated spaces all leans on solving the Riccati equation, which also provides important quantities for the dynamical systems. This is a crucial point that motivates us to further investigate these methods in the context of the LQR problem. Projection methods are specifically designed to handle large dimension n. Together with the large scale version of BT, in Fig. 2 we report the CPU time of the iterative methods, for $\Delta x \in \{0.1, 0.05, 0.025, 0.0125, 0, 00625, 0.003125\}$. We note that the methods reach the desired accuracy in a few seconds even for $n = O(10^5)$. On the contrary, POD would be way more expensive since its cost heavily depends on the original dimension of the problem n, e.g., via the computation of the snapshots.

5.2 *Test 2: 2D Linear Convection-Diffusion Equation*

We consider the linear convection-diffusion equation in (14) with $\gamma = 50, \varepsilon = 1, c = 0, \Omega = [0, 2] \times [0, 2]$ and $\Omega_B = [0.2, 0.8] \times [0.2, 0.8]$. In (1), the matrix A is given by centered five points finite difference discretization plus an upwind approximation of the convection term (see e.g. [34]). The spatial discretization step is $\Delta x = 0.1$ and leads to a system of dimension $n = 441$. The matrix C in (1) is given by the indicator function over the domain $\Omega_C = [0.1, 0.9] \times [0.1, 0.9]$, and $R \equiv I_m$ in (3).

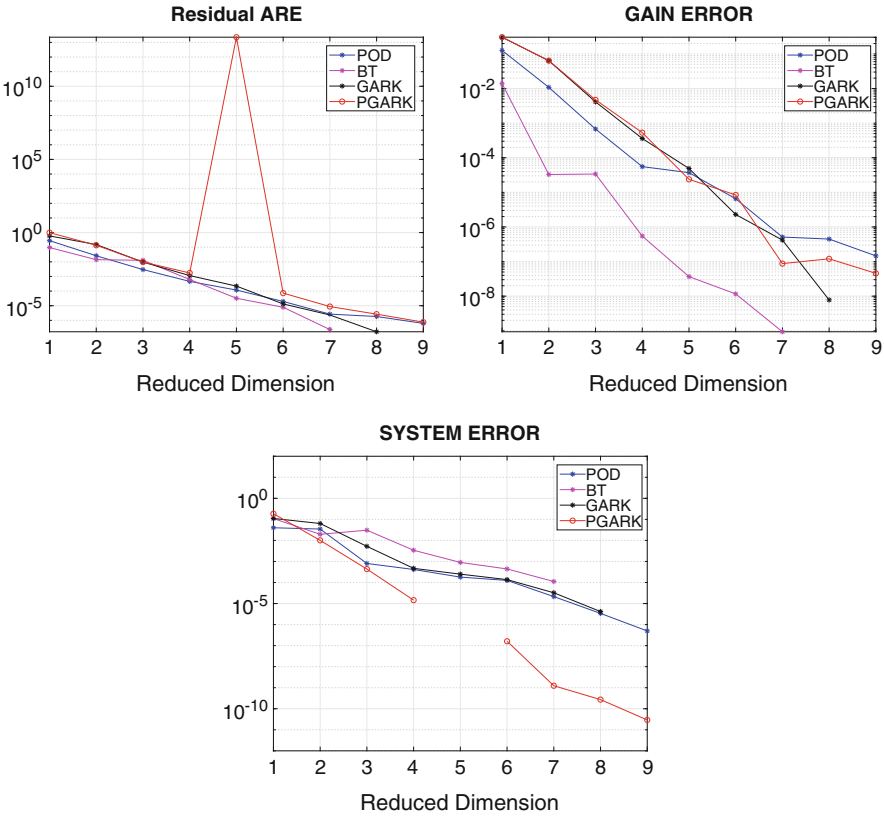


Fig. 3 Test 2: History of the relative residual norm (left), Error $\mathscr{E}_K$ of the feedback gain matrix (middle), Transfer function error $\mathscr{E}_G$ as the approximation space grows (right)

The left panel of Fig. 3 shows the residual norm history associated with the reduced ARE (7), with different projection techniques. We note that the methods converge with a different number of basis functions. We also note that in this example we can observe instability of the PGARK method as discussed in Sect. 4.1. As it is typically done in the algebraic linear system setting, in case of incidental quasi-orthogonality of the basis vectors in PGARK, one could "look a-ahead" and generate subsequent vectors, by accordingly modifying the implementation; see, e.g., [21] and references therein. We have not pursued this here, but it should be implemented if a robust piece of software is desired.

Middle panel of Fig. 3 reports the error in the approximation of the feedback gain matrix K. It is very interesting to observe that even on this convection dominated problem all the methods can reach an accuracy of order 10^{-6}.

Finally, we show the error in the approximation of the transfer function in the right panel of Fig. 3. In this example, we can see that the PGARK method performs better than the others but it is rather unstable; this well known instability problem

will be analyzed in future work. The discussion upon the quality of the basis function we had in Test 1 still hold true. The iterative methods are definitely a feasible alternative to well-known techniques as BT and POD.

In the right panel of Fig. 2, we show the CPU time of BT, GARK and PGARK for different dimensions n of the dynamical system. We note that, again, the Galerkin projection reaches the desired accuracy faster than the two Petrov-Galerkin methods (PGARK and BT). This is an interesting result, since it seems to show that generating two spaces is not strictly necessary to achieve good accuracy at high performance.

5.3 Test 3: A Discussion on Large Scale Balanced Truncation

In this experiment we report on some of the shortcomings we have experienced with the version of balanced truncation that we have implemented for handling the large scale setting; here the Lyapunov equations were solved using projection onto rational Krylov spaces. These problems mainly arise because of the two-step procedure: first the approximate solution of the two Lyapunov equations in (12), then the projection of the Riccati equation onto the spaces of the two obtained approximate Gramians. This is precisely what can be avoided in the projected Petrov-Galerkin approach. Indeed, while constructing the same spaces as those used by the Gramian solvers, GARK readily obtains an approximation to the sought after Riccati solution, without the intermediate approximation of the Gramians.

The first difficulty consists of choosing the stopping tolerance for iteratively solving the two Lyapunov equations, so that the two Gramians are good enough to produce an acceptable Riccati approximate solution. A tolerance simply smaller than that used in the stopping criterion for the Riccati equation is not sufficient. Rather, it should be a few orders of magnitude higher. In Table 1 we report what happens to the Riccati solution for the data in the two previous examples (here $n = 103{,}041$), for different values of the Lyapunov solvers tolerance.

For the sake of the analysis, we also considered a matrix A stemming from the five point stencil finite difference discretization of the operator

$$\mathscr{L}(w) = (e^{-xy} w_x)_x + (e^{xy} w_y)_y + 1/5(x + y) w_x \tag{19}$$

in the unit square, and the same values of B and C as in Test 1. Results are reported in the rightmost group of columns in Table 1.

Clearly, the Lyapunov equation tolerance granting convergence to the Riccati equation is data dependent. For the data in Test 2, requiring a tolerance one order of magnitude lower than the final one is enough, whereas this is not so for Test 1. For the operator $\mathscr{L}$ the situation is even more severe.

The second difficulty arises in case the iterative schemes for approximately computing the two Gramians converge to the requested accuracy in a quite different number of iterations. In this case, the Riccati equation can be projected onto a space

Table 1 Final achieved residual norm ($\mathtt{R}_P$) in the balanced truncation procedure, depending on the accuracy of the two Lyapunov solves

	Test 1			Test 2			$\mathscr{L}(w)$		
Lyap tol	# Lyap its	Riccati space dim	$\mathtt{R}_P$	# Lyap its	Riccati space dim	$\mathtt{R}_P$	# Lyap its	Riccati space dim	$\mathtt{R}_P$
10^{-5}	4–4	4	$8 \cdot 10^{-5}$	4–4	4	$1 \cdot 10^{-4}$	19–18	17	$4 \cdot 10^{-2}$
10^{-6}	5–5	5	$2 \cdot 10^{-5}$	6–6	5	$4 \cdot 10^{-7}$	21–20	19	$9 \cdot 10^{-4}$
10^{-7}	5–6	4	$9 \cdot 10^{-7}$	6–6	5	$4 \cdot 10^{-7}$	23–22	20	$1 \cdot 10^{-4}$
10^{-8}	6–6	4	$9 \cdot 10^{-7}$	6–7	5	$1 \cdot 10^{-7}$	25–25	22	$6 \cdot 10^{-5}$
10^{-9}	7–7	5	$7 \cdot 10^{-7}$	8–8	5	$1 \cdot 10^{-7}$	28–27	24	$8 \cdot 10^{-7}$

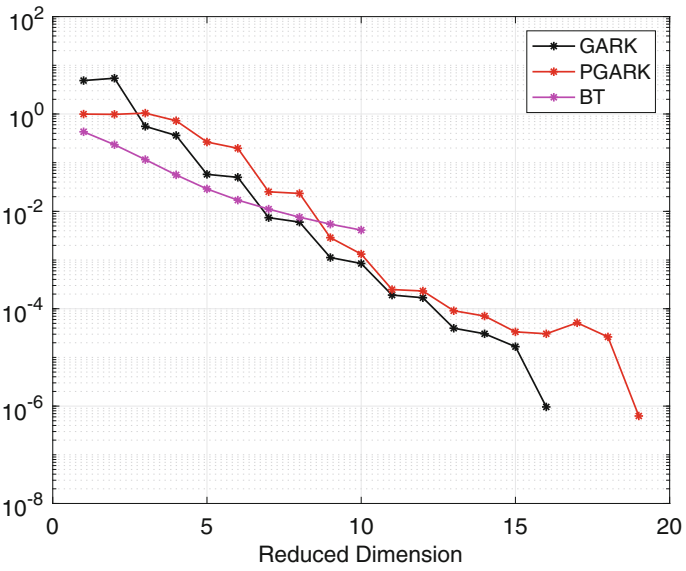


Fig. 4 Convergence history for GARK, PGARK and BT for $\mathscr{L}(w)$ in (19) and specific B and C

whose dimension is at most the one obtained by the lowest rank basis. The projection procedure then stops, irrespective of the final accuracy of the obtained approximate Riccati solution. This phenomenon is reported in Fig. 4, for the data associated with (19) and the selections $C = \mathbf{1}$ and $B = [-1, 1, -1, 1, \ldots, 1]$ (alternating ones). The employed Lyapunov solver for B converges in 14 iterations, with a solution of rank 10, while the solver for C takes 28 iterations, and yields a solution of rank 27. The balanced truncation procedure terminates after 10 iterations, with an unsatisfactory relative residual norm above 10^{-3}. The figure also reports the convergence of the projection methods, which are able to reach the desired accuracy with similar convergence history.

6 Conclusions

We have proposed a comparison of different model order reduction techniques for the ARE. We distinguished between two different strategies: (1) First reduction of the dynamical system complexity and then the solution of the corresponding reduced ARE whereas; (2) Simultaneous solution of the ARE and determination of the reduction spaces. The strength of the second strategy is its flexibility for very high dimensional problem, where problems in the class (1) may have memory and computational difficulties, or, in the case of large scale BT, parameter tuning issues. Experiments on both small and large dimensional problems confirm the promisingly good approximation properties of rational Krylov methods as compared to more standard approaches for the approximation of more challenging quantities in the LQR context.

References

1. Alla, A., Kutz, J.: Randomized model order reduction, submitted (2017). https://arxiv.org/pdf/1611.02316.pdf
2. Alla, A., Graessle, C., Hinze, M.: A posteriori snapshot location for POD in optimal control of linear parabolic equations. ESAIM Math Model. Numer Anal. (2018). https://doi.org/10.1051/m2an/2018009
3. Alla, A., Schmidt, A., Haasdonk, B.: Model order reduction approaches for infinite horizon optimal control problems via the HJB equation. In: Benner, P., Ohlberger, M., Patera, A., Rozza, G., Urban, K. (eds.) Model Reduction of Parametrized Systems, pp. 333–347. Springer International Publishing, Cham (2017)
4. Antoulas, A.C.: Approximation of Large-Scale Dynamical Systems, Advances in Design and Control. SIAM, Philadelphia (2005)
5. Atwell, J., King, B.: Proper orthogonal decomposition for reduced basis feedback controllers for parabolic equations. Math. Comput. Model. **33**, 1–19 (2001)
6. Barkouki, H., Bentbib, A.H., Jbilou, K.: An adaptive rational method Lanczos-type algorithm for model reduction of large scale dynamical systems. J. Sci. Comput. **67**, 221–236 (2015)
7. Beattie, C., Gugercin, S.: Model reduction by rational interpolation. In: Model Reduction and Approximation: Theory and Algorithms. SIAM, Philadelphia (2017)
8. Benner, P., Bujanović, Z.: On the solution of large-scale algebraic Riccati equations by using low-dimensional invariant subspaces. Linear Algebra Appl. **488**, 430–459 (2016)
9. Benner, P., Heiland, J.: LQG-balanced truncation low-order controller for stabilization of laminar flows. In: King, R. (ed.) Active Flow and Combustion Control 2014, vol. 127 of Notes on Numerical Fluid Mechanics and Multidisciplinary Design, pp. 365–379. Springer International Publishing, Cham (2015)
10. Benner, P., Hein, S.: MPC/LQG for infinite-dimensional systems using time-invariant linearizations. In: Tröltzsch, F., Hömberg, D. (eds.) System Modeling and Optimization. IFIP AICT, vol. 291, pp. 217–224. Springer, Berlin (2013)
11. Benner, P., Saak, J.: A Galerkin-Newton-ADI method for solving large-scale algebraic Riccati equations. Tech. Rep. 1253, DF Priority program (2010)
12. Benner, P., Saak, J.: Numerical solution of large and sparse continuous time algebraic matrix Riccati and Lyapunov equations: a state of the art survey. GAMM-Mitteilungen **36**, 32–52 (2013)

13. Benner, P., Li, J.-R., Penzl, T.: Numerical solution of large-scale Lyapunov equations, Riccati equations, and linear-quadratic optimal control problems. Numer. Linear Algebra Appl. **15**, 1–23 (2008)
14. Benner, P., Cohen, A., Ohlberger, M., Willcox, K. (eds.): Model Reduction and Approximation: Theory and Algorithms. Computational Science & Engineering. SIAM, Philadelphia (2017)
15. Borggaard, J., Gugercin, S.: Model reduction for DAEs with an application to flow control. In: King, R. (ed.) Active Flow and Combustion Control 2014. Notes on Numerical Fluid Mechanics and Multidisciplinary Design, vol. 127, pp. 381–396. Springer, Cham (2015)
16. Braun, P., Hernández, E., Kalise, D.: Reduced-order LQG control of a Timoshenko beam model. Bull. Braz. Math. Soc. N. Ser. **47**, 143–155 (2016)
17. Druskin, V., Simoncini, V.: Adaptive rational Krylov subspaces for large-scale dynamical systems. Syst. Control Lett. **60**, 546–560 (2011)
18. Falcone, M., Ferretti, R.: Semi-Lagrangian Approximation Schemes for Linear and Hamilton-Jacobi Equations. SIAM, Philadelphia (2014)
19. Grüne, L., Pannek, J.: Nonlinear Model Predictive Control. Theory and Algorithms. Springer, Berlin (2011)
20. Gugercin, S., Antoulas, A.C., Beattie, C.: $\mathscr{H}_2$ model reduction for large-scale linear dynamical systems. SIAM J. Matrix Anal. Appl. **30**, 609–638 (2008)
21. Gutknecht, M.H.: A completed theory of the unsymmetric Lanczos process and related algorithms. I. SIAM J. Matrix Anal. Appl. **13**, 594–639 (1992)
22. Heyouni, M., Jbilou, K.: An extended Block Krylov method for large-scale continuous-time algebraic Riccati equations. Electron. Trans. Numer. Anal. **33**, 53–62 (2008–2009)
23. Hinze, M., Pinnau, R., Ulbrich, M., Ulbrich, S.: Optimization with PDE Constraints. Mathematical Modelling: Theory and Applications. Springer, New York (2009)
24. Jaimoukha, I.M., Kasenally, E.M.: Krylov subspace methods for solving large Lyapunov equations. SIAM J. Numer. Anal. **31**, 227–251 (1994)
25. Jaimoukha, I.M., Kasenally, E.M.: Oblique projection methods for large scale model reduction. SIAM J. Matrix Anal. Appl. **16**, 602–627 (1995)
26. Jbilou, K.: Block Krylov subspace methods for large algebraic Riccati equations. Numer. Algorithms **34**, 339–353 (2003)
27. Kleinman, D.: On an iterative technique for Riccati equation computations. IEEE Trans. Autom. Control **13**, 114–115 (1968)
28. Kramer, B., Singler, J.: A POD projection method for large-scale algebraic Riccati equations. Numer. Algebra Control Optim. **4**, 413–435 (2016)
29. Kunisch, K., Volkwein, S.: Galerkin proper orthogonal decomposition methods for a general equation in fluid dynamics. SIAM J. Numer. Anal. **40**, 492–515 (2006)
30. Kunisch, K., Volkwein, S.: Proper orthogonal decomposition for optimality systems. Math. Model. Numer. Anal. **42**, 1–23 (2008)
31. Lancaster, P., Rodman, L.: Algebraic Riccati Equations. Oxford University Press, Oxford (1995)
32. Lin, Y., Simoncini, V.: Minimal residual methods for large scale Lyapunov equations. Appl. Numer. Math. **72**, 52–71 (2013)
33. Lin, Y., Simoncini, V.: A new subspace iteration method for the algebraic Riccati equation. Numer. Linear Algebra Appl. **22**, 26–47 (2015)
34. Quarteroni, A., Valli, A.: Numerical Approximation of Partial Differential Equations. Springer Series in Computational Mathematics. Springer, Berlin (1994)
35. Schmidt, A., Haasdonk, B.: Reduced basis approximation of large scale parametric algebraic Riccati equations. ESAIM Control Optim. Calc. Var. (2017). https://doi.org/10.1051/cocv/2017011
36. Simoncini, V.: Analysis of the rational Krylov subspace projection method for large-scale algebraic Riccati equations. SIAM J. Matrix Anal. Appl. **37**, 1655–1674 (2016)
37. Simoncini, V.: Computational methods for linear matrix equations. SIAM Rev. **58**, 377–441 (2016)

38. Simoncini, V., Szyld, D.B., Monslave, M.: On two numerical methods for the solution of large-scale algebraic Riccati equations. IMA J. Numer. Anal. **34**, 904–920 (2014)
39. Sirovich, L.: Turbulence and the dynamics of coherent structures. Parts I-II. Q. Appl. Math. **XVL**, 561–590 (1987)
40. Volkwein, S.: Model reduction using proper orthogonal decomposition. Lecture notes, University of Konstanz (2011)

Fractional PDE Constrained Optimization: Box and Sparse Constrained Problems

Fabio Durastante and Stefano Cipolla

Abstract In this paper we address the numerical solution of two Fractional Partial Differential Equation constrained optimization problems: the two-dimensional *semilinear Riesz Space Fractional Diffusion equation* with box or sparse constraints. Both a theoretical and experimental analysis of the problems is carried out. The algorithmic framework is based on the L-BFGS-B method coupled with a Krylov subspace solver for the box constrained problem within an *optimize-then-discretize* approach and on the semismooth Newton–Krylov method for the sparse one. Suitable preconditioning strategies by approximate inverses and Generalized Locally Toeplitz sequences are taken into account. The numerical experiments are performed with benchmarked software/libraries enforcing the reproducibility of the results.

Keywords Fractional differential equation · Constrained optimization · Preconditioner · Saddle matrix

1 Introduction

Partial fractional differential equations model different phenomena not appropriately modeled by partial differential equations with ordinary derivatives: from the models of viscoplasticity and viscoelasticity to the modeling of diffusion processes in porous media and, indeed, many other problems exhibiting *non-local properties*; see [30] for a gallery of possible applications. Thus, the study of their controllability and the research of efficient algorithms for this task are becoming always more relevant. Already the *discretize-then-optimize* framework [2, 16] and the *optimize-*

F. Durastante (✉)
Dipartimento di Informatica, Università di Pisa, Pisa (PI), Italy
e-mail: fabio.durastante@di.unipi.it

S. Cipolla
Dipartimento di Matematica, Università di Padova, Padova (PD), Italy
e-mail: cipolla@math.unipd.it

M. Falcone et al. (eds.), *Numerical Methods for Optimal Control Problems*,
Springer INdAM Series 29, https://doi.org/10.1007/978-3-030-01959-4_6

then-discretize one [1, 14] have been investigated in some directions. Particularly, in [14] we dealt with the FDE constrained optimization problem:

$$\begin{cases} \min J(y,u) = \dfrac{1}{2}\|y - z_d\|_2^2 + \dfrac{\lambda}{2}\|u\|_2^2, \\ \quad \text{subject to } -K_{x_1}\dfrac{\partial^{2\alpha} y}{\partial|x_1|^{2\alpha}} - K_{x_2}\dfrac{\partial^{2\beta} y}{\partial|x_2|^{2\beta}} + \mathbf{b}\cdot\nabla y + cy^{\zeta} = u, \\ \quad\quad y \equiv 0,\ (x_1,x_2) \in \partial\Omega, \end{cases} \tag{1}$$

where $\zeta \in \mathbb{N}$, $\mathbf{b} \in \mathscr{C}^1(\Omega, \mathbb{R}^2)$, $c \in \mathscr{C}(\Omega)$, $u \in \mathbb{L}^2(\Omega)$, $K_{x_1}, K_{x_2} \geq 0$ and $K_{x_1} + K_{x_2} > 0$, $\alpha, \beta \in (1/2, 1)$, $\Omega = [a,b] \times [c,d]$, in term of the *symmetric Riesz derivative* (Definition 1). This is an instance of the general problem:

$$\begin{cases} \min J(y,u) = \dfrac{1}{2}\|y - z_d\|_2^2 + \dfrac{\lambda}{2}\|u\|_2^2, \\ \quad \text{subject to } e(y,u) = 0, \end{cases} \tag{2}$$

where J and e are two continuously Fréchet derivable functionals such that,

$$J : Y \times U \to \mathbb{R}, \quad e : Y \times U \to W,$$

with Y, U and W reflexive Banach spaces. If we suppose that $e_y(\overline{y}, \overline{u}) \in \mathscr{B}(Y, W)$ is a bijection (where $\mathscr{B}(Y, W)$ is the set of bounded linear operators), using the Implicit Function Theorem, we can deduce the existence of a (locally) unique solution $y(u)$ to the state equation $e(y, u) = 0$. We can then reformulate the problem in the form

$$\min_{u \in U} f(u) = \min_{u \in U} J(y(u), u), \tag{3}$$

where $J(y(u), u)$ is the reduced cost functional.

In the applications not all the controls $u \in U$ are *admissible*, i.e., controls that satisfy further requirements are sought. The first example we consider here is to enforce $u \in \mathfrak{U}_{\text{ad}} \subseteq U \triangleq \mathbb{L}^2$ in Problem (1), where

$$\mathfrak{U}_{\text{ad}} \triangleq \{u \in U : u_a \leq u(x) \leq u_b \text{ a.e. in } \Omega, u_a, u_b \in \mathbb{R} \text{ and } u_a < u_b\} \subset U. \tag{4}$$

We have obtained in this way the *Box-constrained* problem:

$$\begin{cases} \min J(y,u) = \dfrac{1}{2}\|y - z_d\|_2^2 + \dfrac{\lambda}{2}\|u\|_2^2, \\ \quad \text{subject to } -K_{x_1}\dfrac{\partial^{2\alpha} y}{\partial|x_1|^{2\alpha}} - K_{x_2}\dfrac{\partial^{2\beta} y}{\partial|x_2|^{2\beta}} + \mathbf{b}\cdot\nabla y + cy^{\zeta} = u, \\ \quad\quad y \equiv 0,\ (x_1,x_2) \in \partial\Omega, \\ \quad\quad u \in \mathfrak{U}_{\text{ad}}. \end{cases} \tag{5}$$

Then we focus on the problem of finding *sparse controls* for the Problem (1) that amounts to substitute the regular quadratic-convex functional $J(y,u)$ in (1) with:

$$\begin{cases} \min \hat{J}(y,u) = \dfrac{1}{2}\|y - z_d\|_2^2 + \dfrac{\lambda}{2}\|u\|_2^2 + \eta\|u\|_1, \\ \quad \text{subject to } -K_{x_1}\dfrac{\partial^{2\alpha} y}{\partial|x_1|^{2\alpha}} - K_{x_2}\dfrac{\partial^{2\beta} y}{\partial|x_2|^{2\beta}} + \mathbf{b}\cdot\nabla y + cy^{\zeta} = u, \quad \text{for } \eta > 0, \\ \quad\quad y \equiv 0, (x_1,x_2) \in \partial\Omega, \end{cases} \tag{6}$$

being $\|\cdot\|_1$ the $\mathbb{L}^1$ norm. The non-differentiability of the $\|\cdot\|_1$ norm gives back a *semismooth* convex problem.

This paper is then divided as follows: in Sect. 2 we focus on the existence of solutions for Problem (5) and (6), then in Sect. 3 we provide an algorithmic framework in which these kinds of problems can be easily solved. In conclusion, we substantiate our proposals with some numerical examples in Sect. 4.

2 Existence of Solution with Box and Sparse Constraints

We need to face the numerical solution of the two-dimensional Riesz space fractional diffusion equation [30], that reads as

$$\begin{cases} -K_{x_1}\dfrac{\partial^{2\alpha} y}{\partial|x_1|^{2\alpha}} - K_{x_2}\dfrac{\partial^{2\beta} y}{\partial|x_2|^{2\beta}} + \mathbf{b}\cdot\nabla y + cy = u, & (x_1,x_2) \in \Omega, \\ y \equiv 0, & (x_1,x_2) \in \partial\Omega, \end{cases} \tag{7}$$

where $\mathbf{b} \in \mathscr{C}^1(\Omega,\mathbb{R}^2)$, $c \in \mathscr{C}(\Omega)$, $u \in \mathbb{L}^2(\Omega)$, $K_{x_1}, K_{x_2} \geq 0$ and $K_{x_1} + K_{x_2} > 0$, $\alpha, \beta \in (1/2, 1)$, $\Omega = [a,b] \times [c,d]$, where the symmetric Riesz fractional operator is defined as follows.

Definition 1 Given a function $u(x_1,x_2)$ and given $1/2 < \mu \leq 1$ and $n - 1 < 2\mu \leq n$, we define the *symmetric Riesz derivative* as,

$$\frac{\partial^{2\mu} u(x_1,x_2)}{\partial|x_1|^{2\mu}} = -c_{2\mu}\left({}_{\mathrm{RL}}D^{2\mu}_{a,x_1} + {}_{\mathrm{RL}}D^{2\mu}_{x_1,b}\right)u(x_1,x_2), \quad c_{2\mu} = \frac{1}{2\cos(\mu\pi)}$$

and

$$\begin{aligned} {}_{\mathrm{RL}}D^{2\mu}_{a,x}u(x_1,x_2) &= \frac{1}{\Gamma(n-2\mu)}\left(\frac{\partial}{\partial x_1}\right)^n \int_a^{x_1} \frac{u(\xi,x_2)d\xi}{(x_1-\xi)^{2\mu-n+1}}, \\ {}_{\mathrm{RL}}D^{2\mu}_{x,b}u(x_1,x_2) &= \frac{1}{\Gamma(n-2\mu)}\left(-\frac{\partial}{\partial x_1}\right)^n \int_{x_1}^b \frac{u(\xi,x_2)d\xi}{(\xi-x_1)^{2\mu-n+1}}, \end{aligned}$$

and analogously on the x_2 direction.

We have that the existence of both the solutions of Problems (5) and (6) is guaranteed by classic arguments; see, particularly, [15, Chapter 5, 6] and the discussion in [14] and references therein for adapting it to the case of fractional differential equations. Moreover, the construction of the optimality conditions and the selection of the computational framework is taken directly from [15].

We recall here only the two relevant results, extended to the case of fractional partial differential equation constraints, which are needed for establishing the computational procedures in Sect. 3.

Theorem 1 ([14]) *The optimality condition of the first order for problem* (1)*, given* $\lambda > 0$*, is expressed as:*

$$\nabla f(u) \equiv p + \lambda u = 0, \quad u \in U, \tag{8}$$

where $p \in W'$ *is obtained through the solution of the adjoint equation*

$$\begin{cases} -K_{x_1} \frac{\partial^{2\alpha} p}{\partial |x_1|^{2\alpha}} - K_{x_2} \frac{\partial^{2\beta} p}{\partial |x_2|^{2\beta}} - \nabla \cdot (p\mathbf{b}) + \zeta c y^{\zeta - 1} p = y - z_d, & (x_1, x_2) \in \Omega, \\ p \equiv 0, & (x_1, x_2) \in \partial\Omega. \end{cases} \tag{9}$$

and y *is the solution of equation* (7) *for a given control* $u \in U$*.*

By coupling together Theorem 1 and [15, Theorem 6.1] we can state the following result:

Theorem 2 *The optimality conditions for Problem* (6) *are given by:*

$$\begin{cases} -K_{x_1} \frac{\partial^{2\alpha} y}{\partial |x_1|^{2\alpha}} - K_{x_2} \frac{\partial^{2\beta} y}{\partial |x_2|^{2\beta}} + \mathbf{b} \cdot \nabla y + cy = u, & (x_1, x_2) \in \Omega, \\ y \equiv 0, & (x_1, x_2) \in \partial\Omega, \end{cases} \tag{10}$$

$$\begin{cases} -K_{x_1} \frac{\partial^{2\alpha} p}{\partial |x_1|^{2\alpha}} - K_{x_2} \frac{\partial^{2\beta} p}{\partial |x_2|^{2\beta}} - \nabla \cdot (p\mathbf{b}) + \zeta c y^{\zeta - 1} p = y - z_d, & (x_1, x_2) \in \Omega, \\ p \equiv 0, & (x_1, x_2) \in \partial\Omega. \end{cases} \tag{11}$$

$$p + \lambda u + v = 0, \tag{12}$$

$$u - \max(0, u + c(v - \eta)) - \min(0, u + c(v + \eta)) = 0, \qquad \forall c > 0, \tag{13}$$

where $p \in W'$ *is an adjoint status.*

Observe that the existence of the solution for Problem (6) the requirement $\lambda > 0$ is sufficient; as observed in [32], without this requirement, we need to add box-constraints to the formulation of (6) to ensure the existence of a solution.

3 Algorithms

Optimal control problem for differential equations, either with integer or fractional derivatives, in the *optimize-then-discretize* approach, is a *composite* computational problem, i.e., it requires the solution of at least two different kinds of problems:

1. a large scale optimization problem,
2. the numerical integration of several differential equations; one of them in this case is also nonlinear.

In the following Sects. 3.1–3.3 we discuss each of these computational phases separately showing where and how they connect. The focus will be producing an overall efficient strategy for the solution of the whole problem.

3.1 Finite Differences Discretization

It is possible to treat the discretization of the Riesz space-fractional differential equation (7) by means of the second order accurate *fractional centered derivative scheme* for Riesz derivative from [26] coupled with the usual finite difference scheme for the convective and reactive terms. Given $N \in \mathbb{N}$ we consider the uniform grid $\{x_{i,j} = (a + ih, c + jh)\}_{i,j}$ for $i, j = 0, \ldots, N$ with $h = {}^{b-a}/_{N} = {}^{d-c}/_{N}$ and the compact notation $y_{i,j} = y(x_{i,j})$, $\mathbf{b}_{i,j} = (b_1(x_{i,j}), b_2(x_{i,j}))$, $c_{i,j} = c(x_{i,j})$ and $u_{i,j} = u(x_{i,j})$, from which we obtain the following set of algebraic equations:

$$\begin{aligned} &\frac{1}{h^\alpha}\left(K_{x_1}\sum_{k=-\frac{b-x}{h}}^{\frac{x-a}{h}}\varsigma_i^{(\alpha)}y_{-k,j} + K_{x_2}\sum_{k=-\frac{b-x}{h}}^{\frac{x-a}{h}}\varsigma_j^{(\beta)}y_{i,-k}\right) + \\ &+b_{i,j}^{(1)}\frac{y_{i+1,j} - y_{i-1,j}}{2h} + b_{i,j}^{(2)}\frac{y_{i,j+1} - y_{i,j-1}}{2h} + c_{i,j}y_{i,j}^{\zeta} = u_{i,j} \end{aligned} \tag{14}$$

for each $i, j = 0, \ldots, N$ and the coefficients $\varsigma_k^{(\alpha)}$ are given by:

$$\varsigma_k^{(\alpha)} = \frac{(-1)^k \Gamma(\alpha+1)}{\Gamma(\alpha/2 - k + 1)\Gamma(\alpha/2 + k + 1)}, \tag{15}$$

and similarly for β. Collecting the terms in (14) we can represent it in matrix form[1] as:

$$H(\mathbf{y}; \mathbf{u}) \equiv \left[K_{x_1}(R_{x_1}^{(\alpha)} \otimes I) + K_{x_2}(I \otimes R_{x_2}^{(\beta)}) + B\right]\mathbf{y} + C\mathbf{y}^{\zeta} - \mathbf{u} = 0, \tag{16}$$

[1]The power of vectors is computed elementwise.

where I is the identity matrix relative to the grid size, $R_{x_1}^{(\alpha)}$ and $R_{x_2}^{(\beta)}$ are the dense Toeplitz matrix associated with the one-dimensional fractional order derivatives in the two directions, the B and C are respectively the evaluation on the relative nodes of the (i, j)-finite difference grid $\{x_{i,j}\}_{i,j}$ of the functions $\mathbf{b} = (b^{(1)}, b^{(2)})$, together with the convective terms obtained with centered differences, and c. The associated Jacobian, needed for the solution with the Newton method of (14), is then given simply by:

$$J_H(\mathbf{y}) = K_{x_1}(R_{x_1}^{(\alpha)} \otimes I) + K_{x_2}(I \otimes R_{x_2}^{(\beta)}) + B + \zeta C \operatorname{diag}(\mathbf{y}^{\zeta-1}). \tag{17}$$

The discretization of the adjoint equation (9) is obtained in the same way, we just need to observe that in this case, it becomes a linear equation in which the actual value of $\mathbf{y}$ is used as a coefficient. Thus, using again the same notation, we have the following parametric family of matrices discretizing the adjoint equation:

$$A'(\mathbf{y}) = K_{x_1}(R_{x_1}^{(\alpha)} \otimes I) + K_{x_2}(I \otimes R_{x_2}^{(\beta)}) + B' + \zeta C \operatorname{diag}(\mathbf{y}^{\zeta-1}) \tag{18}$$

where:

$$B' = \begin{bmatrix} B_1 & J_2 & & \\ -J_1 & B_2 & \ddots & \\ & \ddots & \ddots & J_N \\ & & -J_{N-1} & B_N \end{bmatrix},$$

with:

$$B_j = \begin{bmatrix} 0 & \frac{b_{j,2}^{(2)}}{2h} & & \\ -\frac{b_{j,1}^{(2)}}{2h} & \ddots & \ddots & \\ & \ddots & \ddots & \frac{b_{j,N}^{(2)}}{2h} \\ & & -\frac{b_{j,N-1}^{(2)}}{2h} & 0 \end{bmatrix}, \quad J_j = \begin{bmatrix} \frac{b_{j,1}^{(1)}}{2h} & 0 & 0 & 0 \\ 0 & \frac{b_{j,2}^{(1)}}{2h} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \frac{b_{j,N}^{(1)}}{2h} \end{bmatrix}.$$

Observe that this implies that the discretization matrix of the adjoint equation varies at each step of the minimization process.

3.2 Optimization Routines

3.2.1 Box-Constrained Optimization Problems

Concerning the optimization routines for Problem (5), we use two well known and benchmarked search direction methods: the *projected gradient* method [22] and

the *L-BFGS-B* method [13]. The projected gradient method represents a natural generalization of the gradient method, where the constraints are guaranteed through a projection operation on the set $\mathfrak{U}_{\text{ad}}$, i.e.,

$$\mathbf{u}^{(k+1)} = \Pi_{\mathfrak{U}_{\text{ad}}}(\mathbf{u}^{(k)} - \mu_k \nabla f(\mathbf{u}^{(k)})) \triangleq \max(u_a, \min(u_b, \mathbf{u}^{(k)} - \mu_k \nabla f(\mathbf{u}^{(k)}))),$$

where the steplength parameter μ_k is chosen by a projected line search rule (see [22]).

L-BFGS-B is a more sophisticated generalization of the Limited Memory BFGS method [24] since the (necessary) descent condition for a search direction is not preserved by projecting onto $\Pi_{\mathfrak{U}_{\text{ad}}}$; see [22]. This problem is solved in L-BFGS-B by a two phases approach: after a preliminary step where a set of *active constrained* is identified by a projected gradient strategy, a reduced Quasi-Newton quadratic model for the inactive constraints is used in order to generate the next search direction. We use the implementation from [24]. Thus, we need only the two procedures for computing both $f(\mathbf{u}^{(k)})$ and the gradient $\nabla f(\mathbf{u}^{(k)})$. As it is clear from Theorem 1, each gradient and function evaluation requires the solution of two FPDEs, and hence this turns out to be the dominating computational cost per step for the both the projected gradient and the L-BFGS-B methods. The computation of f requires the solution of the state equation (7) and is achieved by Newton method as

```
while res > tol && it <= itmax
        yold = y;
        d = JH(yold)\H(yold,u + F);
        y = yold - d;
        res = norm(H(y,u),2);
        it = it + 1;
end
f = 0.5*(y-z)'*(y-z) + lambda/2*(u'*u);
```

where the functions `JH(yold)` and `H(y,u)` computes (17) and (16); see Sect. 3.1 for the details. Then the gradient $\nabla f(\mathbf{u}^{(k)})$ is computed by using the solution `y` of the state equation computed at the previous step, together with the solution of the adjoint equation

```
v = y-z;
p = Adual(y)\(v+Fdual);
g = lambda*u+p;
```

where `Adual(y)` is the function generating the matrix discretization of the adjoint equation in (18).

The motivations for our choices are twofold: on one hand we believe that gradient type—i.e., first order methods—and Quasi-Newton methods—i.e., approximated second order methods—represent the state of the art of large scale optimization schemes for smooth problems, on the other, the choice of implementation-ready routines could enforce and simplify the reproducibility of our results.

3.2.2 Semismooth Newton Iteration

Using Theorem 2 it is clear that the solution of Problem (6) is indeed equivalent to the task of finding a zero of the function:

$$F(y,u,p,v)=\begin{bmatrix} -K_{x_1}\frac{\partial^{2\alpha}y}{\partial|x_1|^{2\alpha}} - K_{x_2}\frac{\partial^{2\beta}y}{\partial|x_2|^{2\beta}} + \mathbf{b}\cdot\nabla y + cy - u \\ -K_{x_1}\frac{\partial^{2\alpha}p}{\partial|x_1|^{2\alpha}} - K_{x_2}\frac{\partial^{2\beta}p}{\partial|x_2|^{2\beta}} - \nabla\cdot(p\mathbf{b}) + \zeta c y^{\zeta-1}p - y + z_d \\ p + \lambda u + v \\ u - \max(0, u + c(v-\eta)) - \min(0, u + c(v+\eta)) \end{bmatrix} \equiv 0. \tag{19}$$

The function F is semismooth due to the presence of the min/max functions, inheritance of the $\mathbb{L}^1$ norm we added in the functional. A canonical choice is the use of the semismooth Newton's iteration for finding a solution of Problem (19), see [15, 20, 31, 33]. We decided to follow the well consolidated practice and we refer the interested reader to the above references for more information on the semismooth Newton's method. For the sake of completeness we recall here just that the semismooth Newton's iteration can be expressed formally as the classic one, i.e., defining $\mathbf{w} = (y, u, p, v)^T$, we have

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - (J_F(\mathbf{w}^{(k)}))^{-1} F(\mathbf{w}^{(k)}),$$

being $J_F(\mathbf{w}^{(k)})$ the generalized Jacobian evaluated in the point $\mathbf{w}^{(k)}$. We recall, moreover, that under suitable standard assumptions, the semismooth Newton's method converges locally superlinearly.

3.3 *Fast and Reliable Solution of Sequences of FDEs*

As discussed in Sect. 3, the main cost of the optimization procedure is represented by the repeated solution of both the *state* and the *adjoint* equations from Theorem 1. The solution of such problems occur in both the computation of the reduced functional f, the gradient ∇f and thus also inside the linesearch procedure. This implies that to achieve a reasonable computational time we need to be able of solving in an *efficient* and *reliable* way both the nonlinear state equation (7) and the linear adjoint equation (9).

The fast solution of FDEs has attracted many efforts over time. For our aim, where repeated FDEs solutions are needed, the chosen framework is the one of *preconditioned iterative solvers*. In particular, preconditioned Krylov solvers can exploit both the *structure* of the problem and the information achieved from the solutions at the previous steps. Already in this setting we should mention the existence of several approaches: the ones based on *circulant* or *circulant-like* preconditioners [23, 27, 28], on *band-Toeplitz* preconditioners [17], or on *multigrid* preconditioner [25, 29] and on *incomplete factorizations* [10, 21].

In this case, we will use the preconditioners from [10] which exploit the *short-memory principle*, i.e., the decaying of the entries of the discretization matrices of the FDEs, in order to gain information on their inverse. The procedure is based on discarding elements of prescribed *small modulus* in the calculation of an approximate inverse of the matrices of the sequences $\{J_H(\mathbf{y}^{(k)})\}_k$ and $\{A'(\mathbf{y}^{(k)})\}_k$. This technique will produce explicit preconditioners for Krylov subspace methods called the approximate inverse preconditioners; see, e.g., [6, 11, 12]. We focus on *incomplete biconjugation* algorithms for non-Hermitian matrices, thus we consider the following factorization

$$A^{-1} = W D^{-1} Z^T,$$

where the matrices W and Z are lower triangular and are actually the inverse of the triangular factors in the usual LDU decomposition. We will use an algorithm based on a *biconjugate* Gram–Schmidt process for the bilinear form associated with A. In order to make this preconditioner cost effective, we need to enforce the sparsity in the inverse factors by carrying out such biconjugation process incompletely, i.e., applying a dropping rule based on the magnitude of the elements. In this way, we obtain an approximate inverse preconditioner in factorized form

$$A^{-1} \approx \tilde{W} \tilde{D}^{-1} \tilde{Z}^T. \tag{20}$$

We have set the focus on the incomplete biconjugation process from [12] since its left-looking/outer products formulation permits to obtain sparser factors under suitable conditions; see Fig. 1 for an example of the sparse factors $\tilde{W}, \tilde{Z}$ for one of the test problems. As we have done for the case without box-constraints in [14], we focus on this particular choice since it permits to use an effective strategy for *updating* preconditioners in (20) with a low computational effort. Observe that it is usually expensive to rebuild a new preconditioner for each new matrix, even in the low-cost framework we mentioned before, while on the other hand, reusing the same preconditioner cannot be appropriate if the matrices in the sequences change much from one step to another. In [10] the authors have specialized the update techniques from [4, 5, 8, 9] for updating sequences of matrices coming from discretization of FPDEs, i.e., if $J_H(\mathbf{y}^{(0)}) = LDU$ is the *reference* matrix of our sequence, we can consider a decomposition of the form:

$$\begin{aligned} J_H(\mathbf{y}^{(k)}) =& J_H(\mathbf{y}^{(0)}) + J_H(\mathbf{y}^{(k)}) - J_H(\mathbf{y}^{(0)}) \\ =& L(D + L^{-1}(J_H(\mathbf{y}^{(k)}) - J_H(\mathbf{y}^{(0)}))U^{-1})U, \end{aligned}$$

that, by using the approximation in (20), gives back an updated preconditioner of the form:

$$J_H(\mathbf{y}^{(k)})^{-1} \approx P^{-1} = \tilde{W}(D + g(\tilde{Z}^T(J_H(\mathbf{y}^{(k)}) - J_H(\mathbf{y}^{(0)}))\tilde{W}))^{-1}\tilde{Z}^T,$$

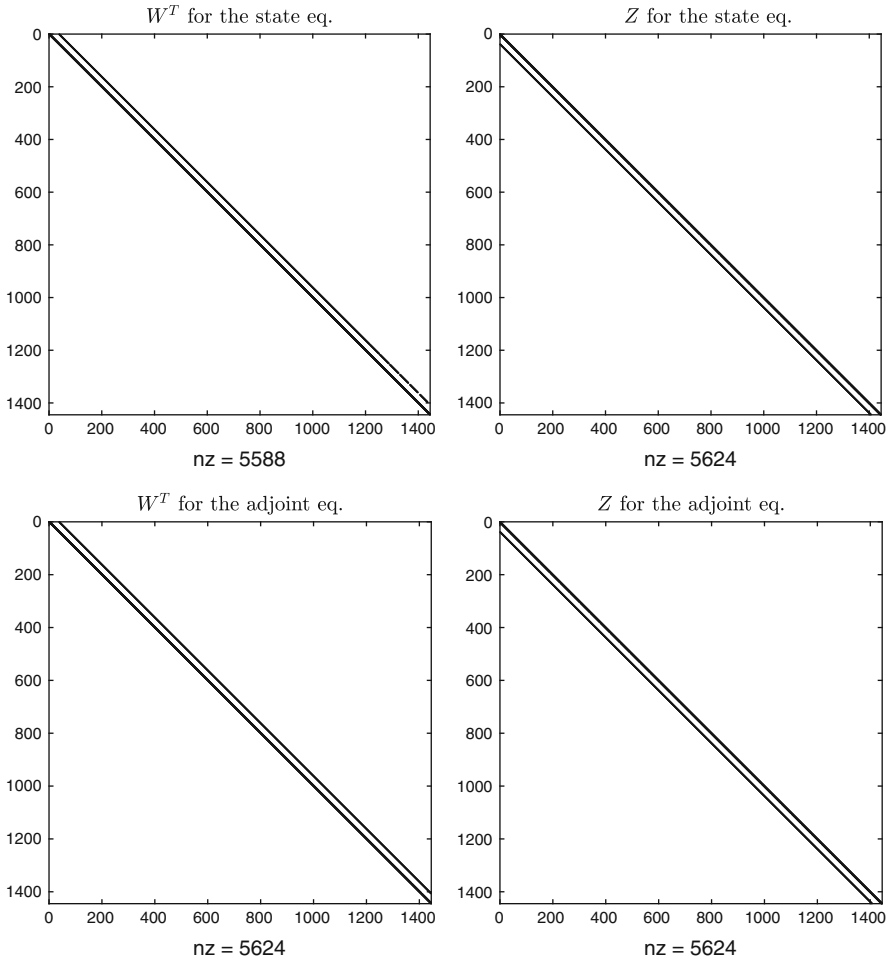


Fig. 1 Sparse factor (20) for the reference preconditioners for $J_H(\mathbf{y})$ and $A'(\mathbf{y})$ for Problem (5) with $\lambda = 1e-9$ and coefficients (25) and $\sigma_1 = \sigma_2 = 0.1$, $c_1 = 0.25$, $c_2 = 1$, $\alpha = 1.5$, $\beta = 1.3$, $u_a = -13$, and $u_b = 5$. The dropping tolerance for the construction is $\delta = 1e-1$

where g is a suitable sparsifying function, e.g., the extraction of the diagonal or of a banded approximation of its matrix argument, and we have exploited the link between the LDU factorization and the approximate inverse factorization; see again [5, 6, 9, 11] for the details.

3.3.1 Preconditioners for the Sequence of Jacobian in Semismooth Newton Method

The generalized Jacobian for $F(y, u, p, v)$ in (19), by means of the same notation and discretizations of Sect. 3.1 and assuming $c = \lambda^{-1}$, can be expressed as:

$$J_F(\mathbf{y}, \mathbf{u}, \mathbf{p}, \mathbf{v}) = \begin{bmatrix} M & -I & 0 & 0 \\ Y & 0 & A'(\mathbf{y}) & 0 \\ 0 & \lambda I & I & I \\ 0 & I - \chi_1 & 0 & -\chi_2 \end{bmatrix} \in \mathbb{R}^{4n^2 \times 4n^2} \tag{21}$$

with

$$\begin{aligned} M &= \left[K_{x_1}(R_{x_1}^{(\alpha)} \otimes I) + K_{x_2}(I \otimes R_{x_2}^{(\beta)}) + B \right] + \zeta C \operatorname{diag}(\mathbf{y}^{\zeta-1}), \\ Y &= \zeta(\zeta - 1) C \operatorname{diag}(\mathbf{y}^{\zeta-2}) - I, \\ \chi_1 &= \operatorname{diag}\left(\mathbb{1}_{\{\mathbf{u}+(\mathbf{v}-\eta)/\lambda > 0\}}(\mathbf{u}, \mathbf{v}) + \mathbb{1}_{\{\mathbf{u}+(\mathbf{v}+\eta)/\lambda < 0\}}(\mathbf{u}, \mathbf{v}) \right), \\ \chi_2 &= \frac{1}{\lambda} \operatorname{diag}\left(\mathbb{1}_{\{-\eta+\lambda u+v > 0\}}(\mathbf{u}, \mathbf{v}) + \mathbb{1}_{\{\eta+\lambda u+v < 0\}}(\mathbf{u}, \mathbf{v}) \right), \end{aligned}$$

where $\mathbb{1}_C$ denotes the characteristic function of the set C computed in an elementwise way; in Fig. 2 we report the pattern of J_F for several iteration of the semismooth Newton algorithm, i.e., for several values of $(\mathbf{y}, \mathbf{u}, \mathbf{p}, \mathbf{v})$.

This is indeed a sequence of large and sparse linear systems with *nonsymmetric saddle matrices* (see the review [7] for general information on problems with this structure), for which preconditioning is necessary. Ignoring the structure of J_F and applying one of the factorization algorithms from Sect. 3.3, combined with some form of reordering to promote its existence and stability, will not work: usually both heavy fill-in phenomena and unstable pivot sequences tend to occur; see again [7, 11, 18]. For this case we will proceed in a different way. We start from the concept of *spectral approximation* for the matrix sequences $\{M \in \mathbb{R}^{n^2 \times n^2}\}_n$ and $\{A'(\mathbf{y}) \in \mathbb{R}^{n^2 \times n^2}\}_n$ to obtain a matrix sequence approximating $\{J_F \in \mathbb{R}^{4n^2 \times 4n^2}\}_n$ more favorable for preconditioning. This approach is inspired by the work in [17, 25] for the discretization of the fractional differential equations. It uses the tools of *Generalized Locally Toeplitz* (GLT) sequences from [19]. In here we recall just the properties we need in the following steps; see again [19] and references therein for a full account of the theory and the precise definition of GLT sequences.

Definition 2 (Asymptotic Eigenvalue Distribution) Given a sequence of matrices $\{X_n\}_n \in \mathbb{C}^{d_n \times d_n}$ with $d_n = \dim X_n \xrightarrow{n \to +\infty} \infty$ monotonically and a μ-measurable function $f : D \to \mathbb{R}$, with $\mu(D) \in (0, \infty)$, we say that the sequence

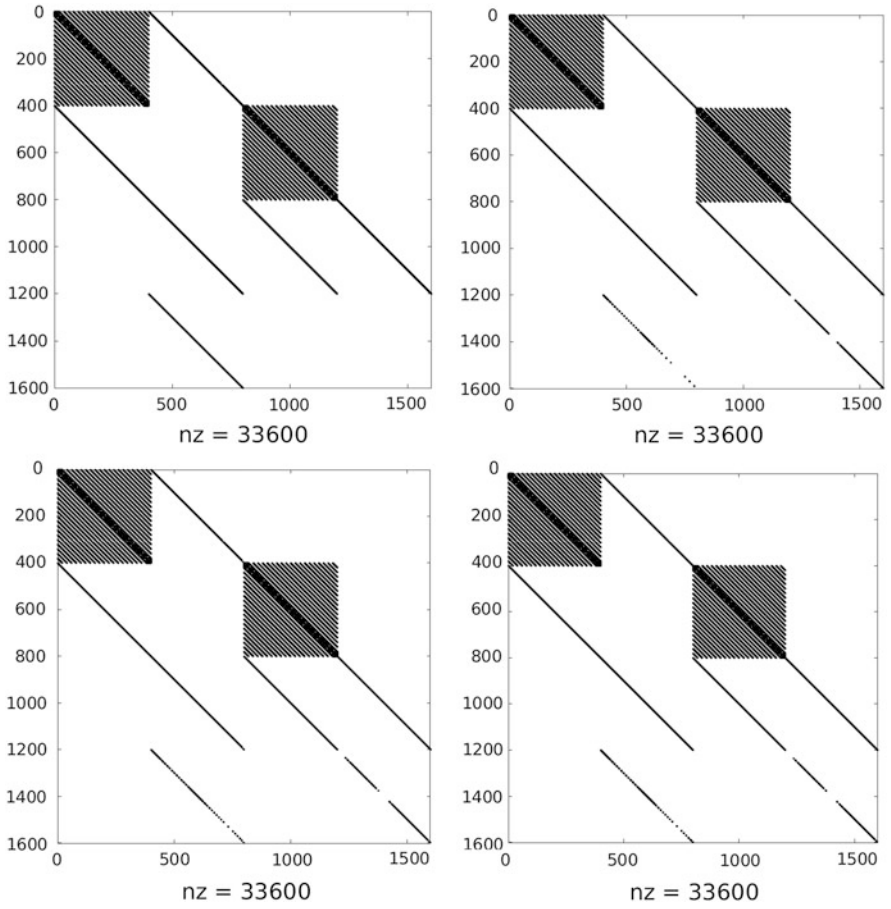


Fig. 2 Example of the structure of Jacobian matrices J_F for several iterations of the semismooth Newton algorithm

$\{X\}_n$ is distributed in the sense of the eigenvalues as the function f and write $\{X_n\}_n \sim_\lambda f$ if and only if, for any F continuous with bounded support, we have

$$\lim_{n\to\infty} \frac{1}{d_n} \sum_{j=0}^{d_n} F(\lambda_j(X_n)) = \frac{1}{\mu(D)} \int_D F(f(t))dt,$$

where $\lambda_j(\cdot)$ indicates the j-th eigenvalue.

Definition 3 (Asymptotic Singular Values Distribution) Given a sequence of matrices $\{X_n\}_n \in \mathbb{C}^{d_n \times d_n}$ with $d_n = \dim X_n \overset{n\to+\infty}{\longrightarrow} \infty$ monotonically and a μ-measurable function $f : D \to \mathbb{R}$, with $\mu(D) \in (0, \infty)$, we say that the sequence $\{X\}_n$ is distributed in the sense of the singular values as the function f and write

$\{X_n\}_n \sim_\sigma f$ if and only if, for any F continuous with bounded support, we have

$$\lim_{n\to\infty} \frac{1}{d_n} \sum_{j=0}^{d_n} F(\sigma_j(X_n)) = \frac{1}{\mu(D)} \int_D F(|f(t)|)dt,$$

where $\sigma_j(\cdot)$ indicates the j-th singular value.

Proposition 1 (GLT Sequences Properties)

GLT1 $\{A_n\}_n \sim_{GLT} \chi \Rightarrow \{A_n\}_n \sim_\sigma \chi$. *Moreover, if* $\{A_n\}_n$ *is a sequence of Hermitian matrices* $\Rightarrow \{A_n\}_n \sim_\lambda \chi$;

GLT2 $\{A_n\}_n \sim_{GLT} \chi$ *and* $A_n = X_n + Y_n$ *with each* X_n *Hermitian, norm bounded* $\|X_n\|_2 = O(1)$, *and* $\|Y_n\|_2 \to 0 \Rightarrow \{A_n\}_n \sim_\lambda \chi$;

GLT3 Sequences of Toeplitz matrices are such that $\{T_n(f)\}_n \sim_{GLT} f$ *if* $f \in \mathbb{L}^1[-\pi, \pi]$; *similarly diagonal sampling matrix sequence of size n generated by a continuous (also Riemann-integrable)* $a : [0, 1] \to \mathbb{C}$ *is:*

$$\left\{ D_n(a) = \text{diag}\left(a\left(\frac{j}{n}\right)\right)_{j=1}^n \right\}_n \sim_{GLT} a;$$

$\{Z_n\}_n \sim_{GLT} 0$ *if and only if* $\{Z_n\} \sim_\sigma 0$;

GLT4 The set of GLT matrices is a $*$*-algebra:*

If $\{A_n\}_n \sim_{\text{GLT}} \kappa$ *and* $\{B_n\}_n \sim_{\text{GLT}} \xi$ *then*

- $\{A_n^*\}_n \sim_{\text{GLT}} \overline{\kappa}$,
- $\{\alpha A_n + \beta B_n\}_n \sim_{\text{GLT}} \alpha\kappa + \beta\xi$ *for all* $\alpha, \beta \in \mathbb{C}$,
- $\{A_n B_n\}_n \sim_{\text{GLT}} \kappa\xi$.

Now we are going to use these instruments to find a preconditioner P for (21) whose blocks are formed by "simpler spectral approximations" of the matrix sequences $\{M \in \mathbb{R}^{n^2\times n^2}\}_n$ and $\{A'(\mathbf{y}) \in \mathbb{R}^{n^2\times n^2}\}_n$, i.e., by matrices whose spectral distributions (in the sense of Definition 2) approximate the spectral distributions of $\{M \in \mathbb{R}^{n^2\times n^2}\}_n$ and $\{A'(\mathbf{y}) \in \mathbb{R}^{n^2\times n^2}\}_n$ giving an overall saddle matrix P that is easier to invert.

Proposition 2 *The sequence of matrices* $\{h^\alpha R_{x_1}^{(\alpha)}\}_n$ *obtained from* (14) *is a GLT sequence with* $\{h^\alpha R_{x_1}^{(\alpha)}\}_n \sim_{\text{GLT}} g_\alpha(\theta)$ *with*

$$g_\alpha(\theta) = \Re\left(-(e^{i\theta} - e^{-i\theta})^\alpha\right). \tag{22}$$

Proof The proof is straightforward by **GLT3**. Matrices $R_{x_1}^{(\alpha)}$ are symmetric Toeplitz matrices whose coefficients are given by (15), moreover the sequence $\{\varsigma_k^{(\alpha)}\}_k$ is in ℓ^1 since

$$\alpha \in (1, 2),\ |\varsigma_k^{(\alpha)}| \sim \frac{1}{\pi}\left|\frac{\Gamma(\alpha+1)}{k^{1+\alpha}}\right|, \quad k \to +\infty.$$

Thus, the ℓ^1-norm converges by asymptotic confrontation with an absolutely summable series. At last we have that $R_{x_1}^{(\alpha)} = T_n(g_\alpha(\theta))$ by direct inspection since the $\varsigma_k^{(\alpha)}$ are the Fourier coefficients of $g_\alpha(\theta)$; see also [26].

Observe that by **GLT3** and **GLT1** we have also $R_{x_1}^{(\alpha)} \sim_\lambda g_\alpha(\theta)$. Therefore, the GLT symbol is a spectral symbol that gives a good approximation of the eigenvalues of the matrix. In Fig. 3a, b we have given a representation of: (1) the symbols $g_\alpha(\theta)$ for several values of $\alpha \in (1, 2]$; (2) the comparison of the eigenvalues (computed numerically) of $R_{x_1}^{(\alpha)}$; (3) the monotonic representative of the symbol $g_\alpha(\theta)$. Observe now that, by **GLT4**, the matrix $\{h^\alpha R_{x_1}^{(\alpha)} + h^\beta R_{x_2}^{(\beta)}\}_n$ is such that

$$\{h^\alpha R_{x_1}^{(\alpha)} + h^\beta R_{x_2}^{(\beta)}\}_n \sim_{\text{GLT}} g_\alpha(\theta_1) + g_\beta(\theta_2) \equiv g_{\alpha,\beta}(\theta_1, \theta_2),$$

and again, by the fact that it is symmetric, we get also that $g_{\alpha,\beta}(\theta_1, \theta_2)$ is also a spectral distribution; see Fig. 3c. At last we conclude our spectral analysis observing that $\{M\}, \{A'(\mathbf{y})\} \sim_{\text{GLT}} g_{\alpha,\beta}(\theta_1, \theta_2)$ and that this relation holds also spectrally by using property **GLT2**, i.e., $\{M\}, \{A'(\mathbf{y})\} \sim_\lambda g_{\alpha,\beta}(\theta_1, \theta_2)$.

Finally, to propose our *spectral* preconditioner, we observe that

$$\lim_{\substack{\alpha\to 2\\ \beta\to 2}} g_{\alpha,\beta}(\theta_1, \theta_2) = (2 - 2\cos(\theta_1)) + (2 - 2\cos(\theta_2)),$$

that is the symbol of the classic five-point discretization of the Laplace operator; see, e.g., [19]. The first preconditioner we want to consider for the sequence of matrices J_F is the one given by:

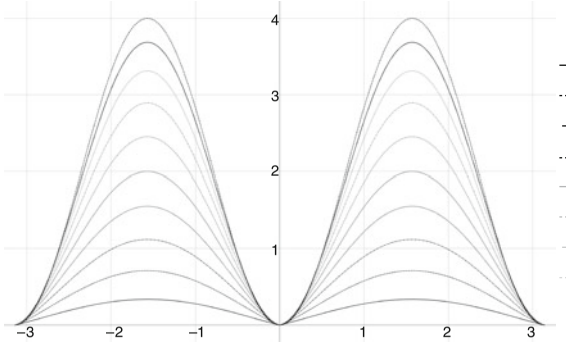
$$P^{(1)} = \begin{bmatrix} L & -I & 0 & 0 \\ Y & 0 & L & 0 \\ 0 & \lambda I & I & I \\ 0 & I - \chi_1 & 0 & -\chi_2 \end{bmatrix}. \tag{23}$$

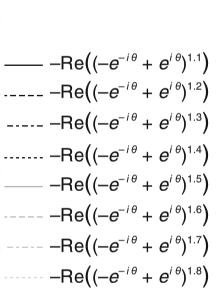
It is obtained from J_F by simply substituting the matrices M and $A'(\mathbf{y})$ with: $L = K_{x_1}(T^{(\alpha)} \otimes I) + K_{x_2}(I \otimes T^{(\beta)})$ where $T^{(\alpha)} = h^{-\alpha} T_n(2 - 2\cos(\theta))$ and similarly for $T^{(\beta)}$.

Another preconditioner obtained with a further degree of approximation it the one given by

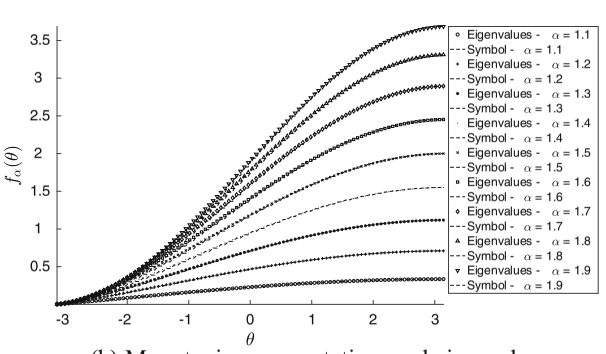
$$P^{(2)} = \begin{bmatrix} L & -I & 0 & 0 \\ -I & 0 & L & 0 \\ 0 & \lambda I & I & I \\ 0 & I - \chi_1 & 0 & -\chi_2 \end{bmatrix}, \tag{24}$$

in which we have also approximated the Y matrix by $-I$. This is interesting since, in this way, we can consider a permutation of $P^{(2)}$ into a saddle matrix with a symmetric $(1, 1)$-block. This is achieved by firstly substituting the adjoint state p with $\tilde{p} = -p$ in all the derivations of Sects. 2, 3.2.2 and 3.1. The permuted version

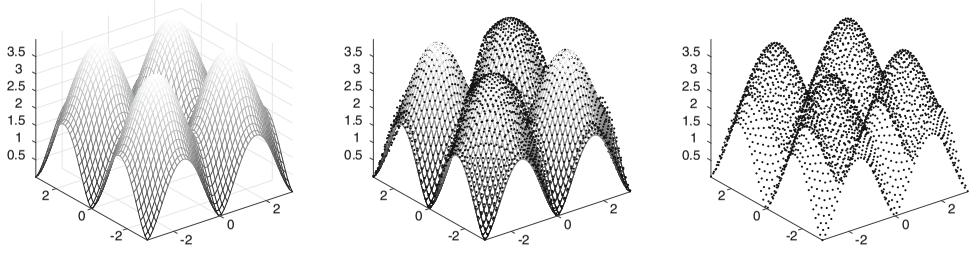




(a) Symbols $g_\alpha(\theta)$



(b) Monotonic representatives and eigenvalues



(c) Symbol and computed eigenvalues for the 2D case. The surface represents the evaluation of the symbol, while the dots are the numerical eigenvalues.

Fig. 3 Symbols $g_\alpha(\theta)$ for several values of α (**a**); monotonic representatives and eigenvalues of the matrix $h^\alpha R_{x_1}^{(\alpha)}$ (**b**); symbols and eigenvalues for $h^\alpha R_{x_1}^{(\alpha)} + h^\beta R_{x_2}^{(\beta)}$ (**c**). Observe that, considering the symmetry properties of the symbols, we can restrict the analysis to the sets $[0, \pi]$ and $[0, \pi]^2$ respectively. We have plotted the symbols on the whole domain just to showcase the full definition, while for all practical purposes the restricted ones are sufficient

of the preconditioner is then:

$$\hat{P}^{(2)} = \left[\begin{array}{ccc|c} I & 0 & L & 0 \\ 0 & \lambda I & -I & I \\ L & -I & 0 & 0 \\ \hline 0 & I-\chi_1 & 0 & -\chi_2 \end{array}\right] = \begin{bmatrix} A & B_1^T \\ B_2 & -C \end{bmatrix} \in \mathbb{R}^{4n^2 \times 4n^2}$$

This is indeed the discretization of the Jacobian for the problem

$$\begin{cases} \min \hat{J}(y,u) = \dfrac{1}{2}\|y - z_d\|_2^2 + \dfrac{\lambda}{2}\|u\|_2^2 + \eta\|u\|_1, \\ \text{subject to } -\nabla^2 y = u, \\ \quad y \equiv 0, (x_1, x_2) \in \partial\Omega, \end{cases} \quad \text{for } \lambda, \eta > 0,$$

that can be further approximated by means of the preconditioner proposed in [31]. The application of the preconditioner $\hat{P}^{(2)}$ can be done either in a direct way or in the form of a multi-iterative scheme in which the application of $P^{(2)}$ is approximated iteratively with a Krylov preconditioned method as in [31].

Remark 1 Observe that we could have used the spectral approximations from Proposition 2 also for building preconditioners for the Box-constrained case, i.e., to solve the sequences of linear systems generated by the *state* and *adjoint* equations. This would have implied going toward a multi-iterative scheme in which we need to use another iterative method to apply these *spectrally approximated* preconditioners which are not anymore explicit preconditioners; see [17, 25] for some solutions in this direction. In here we preferred a more straightforward approach. We computed *directly*, from the matrices of the problems, sparse preconditioners in explicit form such that their application is performed via low-cost matrix-vector product; see again [10, 14] for these issues.

4 Numerical Examples

This section is divided in two parts:

- in Sect. 4.1 we cover an example of a box-constrained problem (5) using algorithms from Sect. 3.2.1 in conjunction with the preconditioners from Sect. 3.3;
- in Sect. 4.2 we discuss the solution of a sparsity constrained problem (6) with the *semismooth Newton* algorithm and preconditioners for the Jacobians from Sect. 3.3.1.

The results presented here have been obtained on a laptop running Linux with 8 Gb memory and CPU Intel® Core™ i7-4710HQ CPU with clock 2.50 GHz, while the GPU is a NVIDIA GeForce GTX 860M. The scalar code is written and executed in MATLAB R2016b, while for the GPU we use C++ with Cuda compilation tools, release 7.5, V7.5.17 and the CUSP library v0.5.1 [3].

4.1 Box-Constrained Problem

We consider the following choice of coefficients for Eq. (7) within the optimal control Problem (5):

$$\begin{aligned} K_{x_1} = 0.5,\ K_{x_2} = 1.5,\ c(x_1, x_2) = 1 + \exp(-x_1^2 - x_2^2), \\ \mathbf{b} = (\Gamma(\beta+1)x^{\beta+\alpha}y^{\alpha+1}, \Gamma(\alpha+1)y^{\beta+\alpha}x^{\beta+1}), \end{aligned} \tag{25}$$

on the domain $\Omega = [0, 1]^2$. The results are obtained for the desired state

$$\begin{aligned} z_d(x_1, x_2) =& c_1 \exp\left(-\frac{(x - 1/4)^2}{2\sigma_1^2} - \frac{(y - 3/4)^2}{2\sigma_1^2}\right) \\ &- c_2 \exp\left(-\frac{(x - 3/4)^2}{2\sigma_2^2} - \frac{(y - 1/4)^2}{2\sigma_2^2}\right). \end{aligned} \tag{26}$$

In this case we will focus *only* on the effects of the projected optimization algorithms from Sect. 3.2.1 and give the results in Tables 1 and 2. The solution of the linear

Table 1 Box constrained problem (5) with coefficient (25) and desired state (26); $\sigma_1 = \sigma_2 = 0.1$, $c_1 = 0.25$, $c_2 = 1$, $u_a = -13$, and $u_b = 5$

			Projected gradient			L-BFGS-B $M = 1$			L-BFGS-B $M = 10$			L-BFGS-B $M = 100$		
λ	α	β	nfg	it	$\|\nabla f\|_\infty$	nfg	it	$\|\nabla f\|_\infty$	nfg	it	$\|\nabla f\|_\infty$	nfg	it	$\|\nabla f\|_\infty$
1e−3	1.1	1.8	3002	2001	3.62e−03	17	7	1.00e−03	20	9	1.14e−03	20	9	1.14e−03
	1.2	1.5	3002	2001	2.18e−03	33	14	1.47e−03	27	12	1.44e−03	27	12	1.44e−03
	1.5	1.2	3002	2001	1.51e−03	48	22	3.78e−04	28	13	3.94e−04	28	13	4.51e−04
	1.7	1.3	3002	2001	2.60e−03	36	17	2.64e−03	22	10	2.60e−03	22	10	2.60e−03
	1.5	1.5	3002	2001	2.95e−03	27	12	2.93e−03	22	10	2.91e−03	22	10	2.91e−03
1e−6	1.1	1.8	3002	2001	8.69e−03	72	35	1.10e−02	64	31	1.10e−02	64	31	1.10e−02
	1.2	1.5	3002	2001	5.47e−03	68	52	8.57e−03	65	31	8.74e−03	53	25	8.78e−03
	1.5	1.2	3002	2001	4.19e−03	99	46	4.46e−03	52	25	4.54e−03	48	23	4.44e−03
	1.7	1.3	3002	2001	6.02e−03	70	33	8.49e−03	42	20	8.66e−03	38	18	8.58e−03
	1.5	1.5	3002	2001	6.26e−03	56	27	9.45e−03	38	18	9.56e−03	36	17	9.50e−03
1e−9	1.1	1.8	3002	2001	8.70e−03	74	36	1.10e−02	60	29	1.10e−02	62	30	1.10e−02
	1.2	1.5	3002	2001	5.47e−03	94	44	8.52e−03	59	28	8.81e−03	57	27	8.90e−03
	1.5	1.2	3002	2001	4.19e−03	109	49	4.51e−03	54	26	4.51e−03	50	24	4.44e−03
	1.7	1.3	3002	2001	5.77e−03	88	43	8.50e−03	38	18	8.66e−03	38	18	8.62e−03
	1.5	1.5	3002	2001	6.27e−03	51	24	9.52e−03	36	17	9.54e−03	36	17	9.48e−03

In the table the number of function and gradient evaluations is reported (nfg) together with the number of iterations of the optimization procedure (it). Observe that the *Projected gradient* algorithms reaches the maximum number of iterations every time

Table 2 Timings in second achieved with the preconditioner from Sect. 3.3 for the examples given in Table 1

			L-BFGS-B $M = 1$				L-BFGS-B $M = 10$				L-BFGS-B $M = 100$			
			AINV				AINV				AINV			
λ	α	β	Direct	1e−1	1e−2	I	Direct	1e−1	1e−2	I	Direct	1e−1	1e−2	I
1e−3	1.1	1.8	3.7	1.4	1.8	2.0	4.1	1.5	1.9	2.2	4.2	1.6	2.1	2.3
	1.2	1.5	7.4	2.2	3.3	3.5	5.8	1.7	2.5	2.8	6.0	1.9	2.7	2.9
	1.5	1.2	11.5	3.7	6.3	6.4	8.8	2.0	3.6	3.5	7.9	2.2	3.6	3.7
	1.7	1.3	7.6	2.3	3.9	4.1	4.5	1.4	2.4	2.5	4.6	1.4	2.7	2.6
	1.5	1.5	6.2	1.8	2.8	2.9	4.6	1.4	2.6	2.5	5.0	1.6	2.2	2.3
1e−6	1.1	1.8	16.4	5.7	7.2	9.1	14.2	5.1	6.9	11.0	13.7	5.1	8.7	7.9
	1.2	1.5	20.7	7.3	11.0	9.0	15.0	4.1	6.1	6.5	13.5	3.5	5.0	5.3
	1.5	1.2	29.5	7.7	18.3	14.9	14.3	3.9	6.6	9.0	12.9	3.6	6.7	6.4
	1.7	1.3	17.8	4.9	7.3	9.3	9.8	2.6	4.6	4.7	10.0	2.4	4.5	4.5
	1.5	1.5	13.1	3.5	7.8	6.0	8.3	2.4	3.8	3.9	9.0	2.3	3.8	4.5
1e−9	1.1	1.8	16.2	5.7	7.3	8.8	17.3	4.9	6.4	7.3	17.6	5.4	7.8	7.5
	1.2	1.5	23.1	6.1	11.0	10.2	14.5	3.8	6.2	6.8	12.4	4.0	5.5	5.5
	1.5	1.2	26.5	9.4	14.2	13.1	15.0	4.0	9.4	7.1	14.4	3.7	6.6	6.8
	1.7	1.3	21.9	6.0	11.2	7.9	9.7	2.4	4.1	4.2	8.4	2.4	4.2	4.6
	1.5	1.5	12.3	3.5	7.5	5.5	8.0	2.4	3.7	4.2	7.3	2.3	3.6	4.2

The results are obtained with the left-preconditioned BiCGstab with AINV preconditioners with drop-tolerances 1e−1 and 1e−2; timings needed by the direct method (Direct) and by the unpreconditioned method (I) are given as term of comparison

systems, in both the Newton method for the *state equation* and the *adjoint equation* is achieved with the BiCGstab algorithm [34]. The type of linear problem solved is exactly the same encountered in [14], thus we register the same performances, see again Table 2. Particularly, we observe that AINV preconditioner with coarser dropping outperforms both the direct method and the method preconditioned with a preconditioner with a finer drop-tolerance. Clearly, this is due to the fact that the increase in density of the AINV(1e−2) factors is not worth the gain in number of iterations since denser factors implies higher timings for the matrix-vector operations; see again Fig. 1.

The results collected in Table 1 confirm that, as the theoretical framework predicts, approximated second order methods outperforms gradient type methods. This validates the use of quasi-Newton method also in the FPDE-constrained framework. Moreover, we observe that allowing for larger memory consumption, i.e., $M = 100$, does neither result in a drastic improvement of the optimization's routine performances nor in the achieved timings; see an example of the convergence history in Fig. 4 and again the timings in Table 2. A depiction of the desired state (26), the control and the resulting solution with that control is given in Fig. 5.

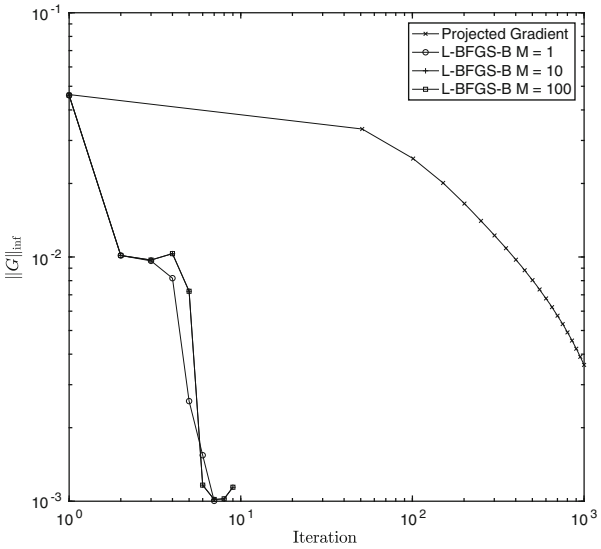


Fig. 4 Convergence behavior for problem (5) with $\lambda = 1e-3$ and coefficients (25). Where $\sigma_1 = \sigma_2 = 0.1$, $c_1 = 0.25$, $c_2 = 1$, $\alpha = 1.5$, $\beta = 1.3$, $u_a = -13$, and $u_b = 5$

4.2 Sparse Constrained Problem

We consider Problem (6) with the same coefficients in (25), except for $c(x_1, x_2) = 1$ and the desired state

$$z_d(x, y) = \sin(2\pi x_1)\sin(2\pi x_2)\frac{\exp(2x_1)}{6}. \tag{27}$$

We face the solution of the linear systems with both a direct solver (Matlab's "\"), to have a reference time, and the BiCGstab preconditioned on the left with preconditioner $P^{(1)}$ and $P^{(2)}$ from (23) and (24); the BiCGstab is set to achieve a tolerance of `1e-6`. In Table 3 we report also the number of the semismooth Newton iterations ($\text{IT}^{\text{Newton}}$), the mean number of iterations needed for the solutions of the linear systems with the Jacobians and the overall time of solution of the problem measured in seconds. A "‡" is reported when the reference direct solver goes out of memory. Attempts of solution with the BiCGstab method without preconditioner are not reported since in that case convergence is never reached. We observe that $P^{(1)}$ and $P^{(2)}$ behave practically in the same way, with $P^{(2)}$ showing timings that are slightly greater than $P^{(1)}$. Moreover, we observe also that the approximated spectral preconditioners $P^{(1)}$ and $P^{(2)}$ are quite independent from the choice of

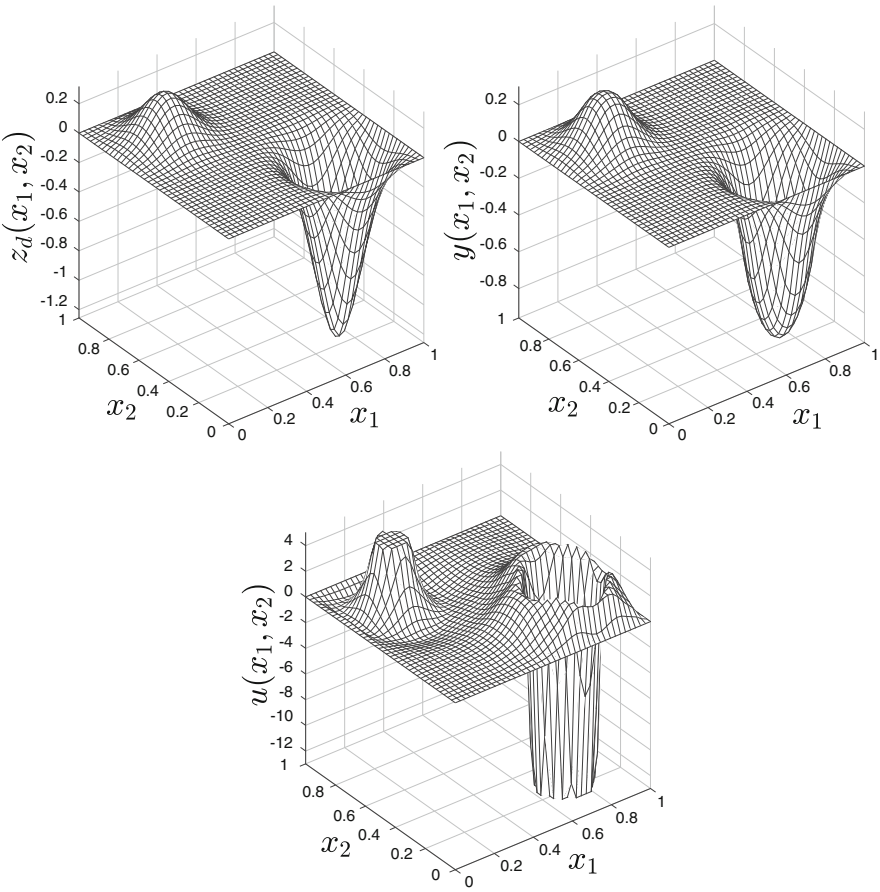


Fig. 5 Desired state z_d from (26), computed solution $\mathbf{y}$ with the control $\mathbf{u}$ for problem (5) with $\lambda = 1e-9$ and coefficients (25). Where $\sigma_1 = \sigma_2 = 0.1$, $c_1 = 0.25$, $c_2 = 1$, $\alpha = 1.5$, $\beta = 1.3$, $u_a = -13$, and $u_b = 5$

the λ parameter, while the number of iterations needed by the semismooth Newton method increase for smaller values of λ, indeed this has to be expected since the function we are minimizing tends to become *less convex*.

In Fig. 6 we plot the desired state (26), the sparse control, in which it is easy to see the zone where it takes value zero, and the obtained solution.

Table 3 Sparse constrained problem (6): mean number of iterations of the relative solver for the various preconditioners for the semismooth Newton method

	Direct		$P^{(1)}$			$P^{(2)}$		
n	$\text{IT}^{\text{Newton}}$	T(s)	$\text{IT}^{\text{Newton}}$	$\text{IT}^{\text{Avg.}}$	T(s)	$\text{IT}^{\text{Newton}}$	$\text{IT}^{\text{Avg.}}$	T(s)
(a) $\alpha = \beta = 1.5$, $\lambda = 1e-3$, $\eta = 5e-3$								
20	6	1.39e−01	6	3.0	3.67e−01	6	3.0	3.47e−01
40	6	2.32e+00	7	3.5	2.34e+00	7	3.5	2.33e+00
60	6	1.89e+01	6	4.0	5.45e+00	6	4.0	5.53e+00
80	6	9.55e+01	7	4.5	1.43e+01	6	4.5	1.48e+01
100	‡	‡	7	5.0	2.85e+01	7	5.0	2.81e+01
120	‡	‡	7	5.0	4.55e+01	7	5.0	4.54e+01
140	‡	‡	7	5.0	7.02e+01	7	5.0	6.99e+01
160	‡	‡	7	5.5	1.08e+02	7	5.5	1.03e+02
(b) $\alpha = \beta = 1.5$, $\lambda = 1e-6$, $\eta = 5e-3$								
20	10	6.45e−01	10	3.0	7.78e−01	10	3.0	5.93e−01
40	13	5.82e+00	14	4.0	5.04e+00	14	4.0	5.00e+00
60	16	5.74e+01	16	4.0	1.73e+01	16	4.0	1.67e+01
80	16	2.82e+02	16	4.5	3.44e+01	16	4.5	3.42e+01
100	‡	‡	17	4.5	7.92e+01	17	4.5	7.76e+01
120	‡	‡	20	4.5	1.43e+02	20	4.5	1.44e+02
140	‡	‡	20	5.0	2.18e+02	20	5.0	2.14e+02
160	‡	‡	20	5.0	3.03e+02	20	5.0	3.12e+02
(c) $\alpha = \beta = 1.5$, $\lambda = 1e-9$, $\eta = 5e-3$								
20	10	2.06e−01	10	2.5	5.08e−01	10	2.0	4.53e−01
40	15	6.59e+00	15	3.0	4.13e+00	15	3.0	4.12e+00
60	19	6.85e+01	19	3.5	1.62e+01	19	3.5	1.57e+01
80	24	4.07e+02	24	4.0	4.66e+01	24	4.0	4.67e+01
100	‡	‡	29	4.0	1.17e+02	29	4.0	1.17e+02
120	‡	‡	33	4.0	2.10e+02	34	4.0	2.19e+02
140	‡	‡	37	4.5	3.48e+02	37	4.5	3.60e+02
160	‡	‡	41	4.5	5.94e+02	41	4.5	6.04e+02
(d) $\alpha = 1.2$, $\beta = 1.8$, $\lambda = 1e$-3, $\eta = 5e-3$								
20	5	1.14e−01	5	4.5	4.69e−01	5	4.5	4.35e−01
40	5	1.81e+00	5	6.0	2.59e+00	5	6.0	2.70e+00
60	6	1.84e+01	6	6.5	9.13e+00	6	6.5	8.94e+00
80	6	9.28e+01	6	7.0	1.99e+01	6	7.0	1.90e+01
100	‡	‡	6	8.5	3.68e+01	6	8.0	3.80e+01
120	‡	‡	6	8.5	6.76e+01	6	8.5	6.77e+01
140	‡	‡	6	9.0	1.07e+02	6	9.0	1.08e+02
160	‡	‡	6	9.5	1.50e+02	6	9.5	1.70e+02

(continued)

Table 3 (continued)

	Direct		$P^{(1)}$			$P^{(2)}$		
n	$\text{IT}^{\text{Newton}}$	T(s)	$\text{IT}^{\text{Newton}}$	$\text{IT}^{\text{Avg.}}$	T(s)	$\text{IT}^{\text{Newton}}$	$\text{IT}^{\text{Avg.}}$	T(s)
(e) $\alpha = 1.2$, $\beta = 1.8$, $\lambda = 1e-6$, $\eta = 5e-3$								
20	8	1.75e−01	9	3.0	5.72e−01	9	3.0	5.29e−01
40	11	4.28e+00	11	4.0	3.72e+00	11	4.0	3.94e+00
60	11	3.41e+01	11	5.0	1.26e+01	12	5.0	1.40e+01
80	11	1.53e+02	11	5.0	2.56e+01	11	5.5	2.57e+01
100	‡	‡	12	6.5	5.81e+01	12	6.0	6.30e+01
120	‡	‡	11	6.5	9.63e+01	11	6.5	9.90e+01
140	‡	‡	12	7.0	1.73e+02	12	7.0	1.72e+02
160	‡	‡	11	7.5	2.31e+02	11	7.5	2.44e+02
(f) $\alpha = 1.2$, $\beta = 1.8$, $\lambda = 1e-8$, $\eta = 5e-3$								
20	8	1.84e−01	8	2.5	4.42e−01	8	2.5	4.26e−01
40	13	5.06e+00	13	3.5	3.87e+00	13	3.5	3.98e+00
60	17	5.77e+01	17	4.0	1.57e+01	17	4.0	1.55e+01
80	‡	‡	21	4.0	4.17e+01	21	4.0	4.22e+01
100	‡	‡	25	4.0	9.52e+01	25	4.0	9.15e+01
120	‡	‡	29	4.5	1.75e+02	29	4.5	1.78e+02
140	‡	‡	33	5.0	3.32e+02	33	4.5	3.39e+02
160	‡	‡	28	5.5	4.50e+02	28	5.5	4.44e+02
(g) $\alpha = 1.7$, $\beta = 1.3$, $\lambda = 1e-3$, $\eta = 5e-3$								
20	6	1.48e−01	6	4.0	4.78e−01	6	4.0	4.68e−01
40	7	2.59e+00	7	4.5	2.98e+00	7	4.5	2.99e+00
60	7	2.20e+01	7	5.0	7.89e+00	7	5.0	7.98e+00
80	7	1.11e+02	7	5.0	1.89e+01	7	5.5	1.85e+01
100	‡	‡	7	5.5	3.25e+01	7	5.5	3.56e+01
120	‡	‡	7	6.0	5.51e+01	7	6.0	5.51e+01
140	‡	‡	7	6.0	7.96e+01	7	6.0	8.29e+01
160	‡	‡	7	6.0	1.16e+02	7	6.0	1.21e+02
(h) $\alpha = 1.7$, $\beta = 1.3$, $\lambda = 1e-6$, $\eta = 5e-3$								
20	11	2.64e−01	12	4.0	8.84e−01	12	4.0	8.29e−01
40	16	6.82e+00	16	4.5	6.81e+00	17	4.5	7.28e+00
60	20	6.86e+01	20	5.0	2.65e+01	20	5.0	2.50e+01
80	‡	‡	24	5.5	6.84e+01	24	5.5	6.60e+01
100	‡	‡	21	6.0	1.18e+02	21	6.0	1.10e+02
120	‡	‡	22	6.5	2.05e+02	22	6.5	2.02e+02
140	‡	‡	25	6.5	3.71e+02	25	7.0	3.59e+02
160	‡	‡	24	7.0	4.98e+02	24	7.0	5.10e+02

(continued)

Table 3 (continued)

	Direct		$P^{(1)}$			$P^{(2)}$		
n	$\text{IT}^{\text{Newton}}$	T(s)	$\text{IT}^{\text{Newton}}$	$\text{IT}^{\text{Avg.}}$	T(s)	$\text{IT}^{\text{Newton}}$	$\text{IT}^{\text{Avg.}}$	T(s)
(i) $\alpha = 1.7$, $\beta = 1.3$, $\lambda = 1e-8$, $\eta = 5e-3$								
20	11	2.56e−01	11	5.5	7.49e−01	11	5.5	7.31e−01
40	17	7.92e+00	17	5.0	1.00e+01	17	5.0	7.20e+00
60	24	8.84e+01	24	4.0	2.43e+01	24	4.0	2.69e+01
80	‡	‡	30	4.5	6.93e+01	30	4.5	7.13e+01
100	‡	‡	37	5.0	1.59e+02	37	5.0	1.62e+02
120	‡	‡	43	5.0	3.29e+02	43	5.0	3.57e+02
160	‡	‡	56	5.5	9.19e+02	56	5.5	9.53e+02

For BiCGstab (with left preconditioning) iteration are given in the Matlab convention, in which there is half iteration for each matrix-vector product inside the algorithm, and rounded consequently. A "‡" is reported when the reference direct solver goes out of memory

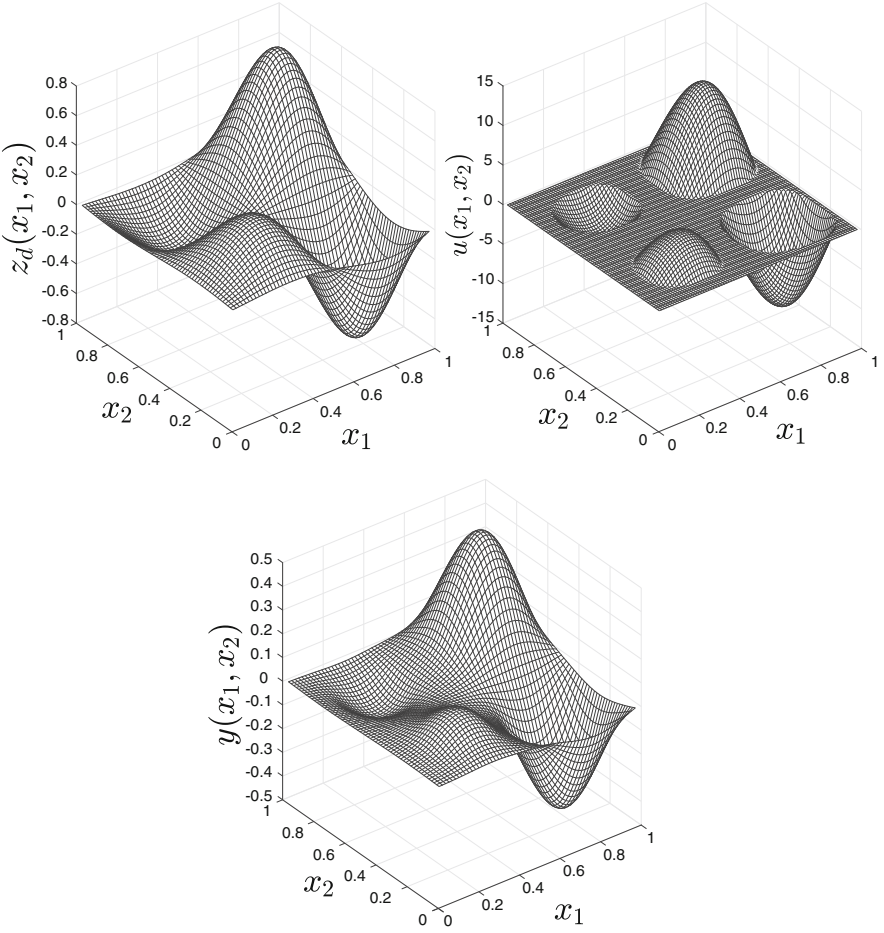


Fig. 6 Solution of problem (6) with sparsity constraint and desired state (26)

References

1. Annunziato, M., Borzì, A., Magdziarz, M., Weron, A.: A fractional Fokker–Planck control framework for subdiffusion processes. Optim. Control. Appl. Methods **37**(2), 290–304 (2016). https://doi.org/10.1002/oca.2168
2. Antil, H., Otarola, E.: A FEM for an optimal control problem of fractional powers of elliptic operators. SIAM J. Control. Optim. **53**(6), 3432–3456 (2015)
3. Bell, N., Garland, M.: Cusp: Generic Parallel Algorithms for Sparse Matrix and Graph Computations (2015). http://cusplibrary.github.io/. Version 0.5.1
4. Bellavia, S., Bertaccini, D., Morini, B.: Nonsymmetric preconditioner updates in Newton–Krylov methods for nonlinear systems. SIAM J. Sci. Comput. **33**(5), 2595–2619 (2011)
5. Benzi, M., Bertaccini, D.: Approximate inverse preconditioning for shifted linear systems. BIT **43**(2), 231–244 (2003)
6. Benzi, M., Tuma, M.: A sparse approximate inverse preconditioner for nonsymmetric linear systems. SIAM J. Sci. Comput. **19**(3), 968–994 (1998)
7. Benzi, M., Golub, G.H., Liesen, J.: Numerical solution of saddle point problems. Acta Numer. **14**, 1–137 (2005)
8. Bertaccini, D.: Efficient preconditioning for sequences of parametric complex symmetric linear systems. Electron. Trans. Numer. Anal. **18**, 49–64 (2004)
9. Bertaccini, D., Durastante, F.: Interpolating preconditioners for the solution of sequence of linear systems. Comput. Math. Appl. **72**(4), 1118–1130 (2016)
10. Bertaccini, D., Durastante, F.: Solving mixed classical and fractional partial differential equations using short-memory principle and approximate inverses. Numer. Algorithms **74**(4), 1061–1082 (2017)
11. Bertaccini, D., Durastante, F.: Iterative Methods and Preconditioning for Large and Sparse Linear Systems with Applications. Chapman & Hall/CRC Monographs and Research Notes in Mathematics. CRC Press, Boca Raton (2018)
12. Bertaccini, D., Filippone, S.: Sparse approximate inverse preconditioners on high performance GPU platforms. Comput. Math. Appl. **71**(3), 693–711 (2016)
13. Byrd, R.H., Lu, P., Nocedal, J., Zhu, C.: A limited memory algorithm for bound constrained optimization. SIAM J. Sci. Comput. **16**(5), 1190–1208 (1995)
14. Cipolla, S., Durastante, F.: Fractional PDE constrained optimization: an optimize-then-discretize approach with L-BFGS and approximate inverse preconditioning. Appl. Numer. Math. **123**, 43–57 (2018). https://doi.org/10.1016/j.apnum.2017.09.001
15. De los Reyes, J.C.: Numerical PDE-Constrained Optimization. Springer, Cham (2015)
16. Dolgov, S., Pearson, J.W., Savostyanov, D.V., Stoll, M.: Fast tensor product solvers for optimization problems with fractional differential equations as constraints. Appl. Math. Comput. **273**, 604–623 (2016)
17. Donatelli, M., Mazza, M., Serra-Capizzano, S.: Spectral analysis and structure preserving preconditioners for fractional diffusion equations. J. Comput. Phys. **307**, 262–279 (2016)
18. Duff, I.S., Erisman, A.M., Reid, J.K.: Direct Methods for Sparse Matrices. Oxford University Press, Oxford (2017)
19. Garoni, C., Serra-Capizzano, S.: Generalized Locally Toeplitz Sequences: Theory and Applications, vol. 1. Springer, Berlin (2017). https://doi.org/10.1007/978-3-319-53679-8
20. Hinze, M., Pinnau, R., Ulbrich, M., Ulbrich, S.: Optimization with PDE Constraints, vol. 23. Springer Science & Business Media, Berlin (2008)
21. Jin, X.Q., Lin, F.R., Zhao, Z.: Preconditioned iterative methods for two-dimensional space-fractional diffusion equations. Commun. Comput. Phys. **18**(2), 469–488 (2015)
22. Kelley, C.T.: Iterative Methods for Optimization. SIAM, Philadelphia (1999)
23. Lei, S.L., Sun, H.W.: A circulant preconditioner for fractional diffusion equations. J. Comput. Phys. **242**, 715–725 (2013)
24. Liu, D.C., Nocedal, J.: On the limited memory BFGS method for large scale optimization. Math. Program. **45**(1), 503–528 (1989)

25. Moghaderi, H., Dehghan, M., Donatelli, M., Mazza, M.: Spectral analysis and multigrid preconditioners for two-dimensional space-fractional diffusion equations. J. Comput. Phys. **350**(Suppl. C), 992–1011 (2017). https://doi.org/10.1016/j.jcp.2017.08.064
26. Ortigueira, M.D.: Riesz potential operators and inverses via fractional centred derivatives. Int. J. Math. Math. Sci. **2006**, 12 pp. (2006)
27. Pan, J., Ke, R., Ng, M.K., Sun, H.W.: Preconditioning techniques for diagonal-times-Toeplitz matrices in fractional diffusion equations. SIAM J. Sci. Comput. **36**(6), A2698–A2719 (2014)
28. Pan, J., Ng, M., Wang, H.: Fast preconditioned iterative methods for finite volume discretization of steady-state space-fractional diffusion equations. Numer. Algorithms **74**(1), 153–173 (2017)
29. Pang, H.K., Sun, H.W.: Multigrid method for fractional diffusion equations. J. Comput. Phys. **231**(2), 693–703 (2012)
30. Podlubny, I.: Fractional Differential Equations: An Introduction to Fractional Derivatives, Fractional Differential Equations, to Methods of Their Solution and Some of Their Applications, vol. 198. Academic, London (1998)
31. Porcelli, M., Simoncini, V., Stoll, M.: Preconditioning PDE-constrained optimization with L^1-sparsity and control constraints. Comput. Math. Appl. **74**(5), 1059–1075 (2017)
32. Stadler, G.: Elliptic optimal control problems with $\mathbb{L}^1$-control cost and applications for the placement of control devices. Comput. Optim. Appl. **44**(2), 159 (2009)
33. Ulbrich, M.: Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces. SIAM, Philadelphia (2011)
34. Van der Vorst, H.A.: Bi-CGSTAB: a fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems. SIAM J. Sci. Comput. **13**(2), 631–644 (1992)

Control, Shape, and Topological Derivatives via Minimax Differentiability of Lagrangians

Michel C. Delfour

Abstract In Control Theory, the *semidifferential* of a *state* constrained *objective function* can be obtained by introducing a *Lagrangian* and an *adjoint state*. Then the initial problem is equivalent to the one-sided derivative of the minimax of the Lagrangian with respect to a positive parameter t as it goes to 0. In this paper, we revisit the results of Sturm (On shape optimization with non-linear partial differential equations. Doctoral thesis, Technische Universität of Berlin, 2014; SIAM J Control Optim 53(4):2017–2039, 2015) recently extended by Delfour and Sturm (J Convex Anal 24(4):1117–1142, 2017; Delfour and Sturm, Minimax differentiability via the averaged adjoint for control/shape sensitivity. In: Proc. of the 2nd IFAC Workshop on Control of Systems Governed by Partial Differential Equations, IFAC-PaperOnLine, vol 49-8, pp 142–149, 2016) from the single valued case to the case where the solutions of the *state/averaged adjoint state* equations are not unique. New simpler conditions are given in term of the *standard adjoint* and extended to the multivalued case. They are applied to the computation of *semidifferentials* with respect to the control and the shape and the topology of the domain. The *shape derivative* is a *differential* while the *topological derivative* usually obtained by expansion methods is not. It is a *semidifferential*, that is, a *one-sided directional derivative* in the directions contained in the *adjacent tangent cone* obtained from dilatations of points, curves, surfaces and, potentially, microstructures (Delfour, Differentials and semidifferentials for metric spaces of shapes and geometries. In: Bociu L, Desideri JA, Habbal A (eds) System Modeling and Optimization. Proc. 27th IFIP TC7 Conference, CSMO 2015, Sophia-Antipolis, France. AICT Series, pp 230–239. Springer, Berlin, 2017; Delfour, J Convex Anal 25(3):957–982, 2018) by using the notion of d-dimensional *Minkowski content*. Examples of such sets are the *rectifiable sets* (Federer, Geometric measure theory. Springer, Berlin, 1969) and the *sets of positive reach* (Federer, Trans Am Math Soc 93:418–419, 1959).

M. C. Delfour (✉)
Centre de recherches mathématiques and Département de mathématiques et de statistique, Université de Montréal, Montréal, QC, Canada
e-mail: delfour@crm.umontreal.ca

M. Falcone et al. (eds.), *Numerical Methods for Optimal Control Problems*, Springer INdAM Series 29, https://doi.org/10.1007/978-3-030-01959-4_7

Keywords Minimax · One-sided differentiability · Minimax · Lagrangian · Standard adjoint · Control · Shape · Topological derivatives · Rectifiability

1 Introduction

This paper is motivated by the generic notions of *shape* and *topological derivatives* that have proven to be both pertinent and useful from the theoretical and numerical points of view (for instance, see the recent book of Novotny and Sokołowski [17] and its bibliography). The *shape derivative* is a differential (see [4, 12, 16]) while the *topological derivative* rigorously introduced by Sokolowski and Zochowski [18] is only a *semidifferential* (one sided directional derivative) which is usually obtained by the *method* of *matched* and *compound expansions*. The lack of linearity of the one-sided directional derivative with respect to the direction arises from the fact that the tangent space to the underlying metric spaces of "geometries" is only a cone. For instance, the set of measurable subsets of an hold-all can be identified with the metric Abelian group of its *characteristic functions*. By using the notion of *d-dimensional Minkowski content* for *d-rectifiable sets* [14] or *sets of positive reach* of Federer [13], we can characterize as *bounded measures* elements of the tangent cone that are only half tangents but not full tangents. In that spirit, the definition of a topological derivative as a semidifferential was extended in [5, 6] from the dilatation of a point to dilatations of curves, surfaces, and, potentially, microstructures.

An important advantage of the use of *semidifferentials* over *expansion methods* is that theorems on the one-sided differentiation of the minimax of a Lagrangian can be used to get the semidifferential of *state constrained objective functions* (see [3, 9, 10, 12]). By using the notion of *averaged adjoint* introduced by Sturm [19, 20], the minimax problem need not be related to a saddle point: non-convex objective functions and non-linear state equations can be directly handled. Recently, a simpler and more general version of the original condition of Sturm [19, 20] was given by Delfour and Sturm [7, 8] and extended from the single valued case to the case where the solutions of the state/averaged adjoint state equations are not unique.

In this paper we revisit the one-sided derivative of the minimax of a Lagrangian with respect to a positive parameter t and provide new conditions that only require the existence of the *standard adjoint state* at $t = 0$ instead of the *averaged adjoint state* at $t \geq 0$ near 0. A simple example is given to illustrate the application of the new conditions to the computation of semidifferentials with respect to perturbations in the control and in the shape or topology of the underlying geometric domain. Finally, the paper is completed by extending the results from the single valued case to the case where the solutions of the state/adjoint state equations are not unique.

2 Examples of Derivatives with Respect to a Control, Shape, or Topological Variable

In order to better motivate and appreciate the main results of this paper, we consider a *very simple* example to illustrate how the computation of the one-sided directional derivative (semidifferential) of a state constrained objective function with respect to a control, shape, or topological variable is amenable to the one-sided derivative of the minimax of a Lagrangian with respect to a parameter $t \geq 0$.

It was not possible to make detailed comparisons with the many results and techniques available in the literature or to explore new or more complex applications within the limits of a conference paper. For instance, one of the referees pointed out the potential for problems involving *cavities in elastic solids* (see Lewiński-Sokołowski [15]) where the *capacity* rather than the *volume* would be more relevant. This is clearly beyond the scope of this paper,

Given a bounded open *domain* Ω in $\mathbb{R}^3$ with Lipschitz boundary Γ, a *control function* $a \in L^2(\mathbb{R}^3)$, the *state* $u = u(a, \Omega) \in H_0^1(\Omega)$ is the solution of the variational *state equation*

$$\exists u \in H_0^1(\Omega), \quad \int_\Omega \nabla u \cdot \nabla \psi - a\, \psi \, dx = 0, \quad \forall \psi \in H_0^1(\Omega), \tag{1}$$

where $x \cdot y$ denotes the inner product of x and y in $\mathbb{R}^3$. Given a *target function* $z \in L^2(\mathbb{R}^3)$, associate with $u(a, \Omega)$ the *objective function*

$$f(a, z, \Omega) \stackrel{\text{def}}{=} \int_\Omega \frac{1}{2} |u(a, \Omega) - z|^2 \, dx, \tag{2}$$

which depends on a, z, and Ω. Our purpose is to find the expressions of the one-sided directional derivatives with respect to a, z, and Ω. In the first two cases, the functions a and z live in Banach or Fréchet vector spaces. In contrast, the variable domains Ω live in spaces that are, at best, groups with a metric [4, 12]. In such spaces, the tangent space is a cone in 0 which is not necessarily linear as in the case of infinite dimensional Riemann or Finsler manifolds. When only half-tangents are available, only one-sided directional derivatives of a function can be expected. Two types of perturbations of a domain Ω will be considered:

1. perturbation of Ω by a group of diffeomorphisms of the Euclidean space $\mathbb{R}^N$ [11, 12, 16] that lead to the notion of *shape derivative* [9, 10, 21–23];
2. topological perturbations of Ω by removing dilations of d-dimensional closed subsets E of Ω that lead to the notion of *topological derivative* [5, 6, 17, 18].

2.1 Directional Derivative with Respect to a Control Variable

First consider the directional derivative of f with respect to the *control function* a. We get an unconstrained minimax formulation by introducing the Lagrangian

$$G(a,\varphi,\psi) \overset{\text{def}}{=} \int_\Omega \frac{1}{2}|\varphi - z|^2\,dx + \int_\Omega \nabla\varphi\cdot\nabla\psi - a\,\psi\,dx$$

$$f(a) = \inf_{\varphi\in H^1(\Omega)} \sup_{\psi\in H^1(\Omega)} G(a,\varphi,\psi).$$

Associate with the perturbed control function $a+tb$, $t\geq 0$, the *state* $u^t \in H_0^1(\Omega)$ which is the solution of the state equation

$$\int_\Omega \nabla u^t\cdot\nabla\psi - (a+tb)\,\psi\,dx = 0, \quad \forall\psi\in H_0^1(\Omega). \tag{3}$$

The semidifferential of $f(a)$ in the direction $b\in L^2(\Omega)$,

$$df(a;b) \overset{\text{def}}{=} \lim_{t\searrow 0}\frac{f(a+tb)-f(a)}{t}, \qquad f(a+tb) = \int_\Omega \frac{1}{2}|u^t - z|^2\,dx,$$

is obtained by a similar minimax formulation. Given the *t-dependent Lagrangian*

$$L(t,\varphi,\psi) \overset{\text{def}}{=} \int_\Omega \frac{1}{2}|\varphi - z|^2\,dx + \int_\Omega \nabla\varphi\cdot\nabla\psi - (a+tb)\,\psi\,dx,$$

it is readily seen that

$$g(t) \overset{\text{def}}{=} \inf_{\varphi\in H_0^1(\Omega)} \sup_{\psi\in H_0^1(\Omega)} L(t,\varphi,\psi) = f(a+tb),$$

$$dg(0) \overset{\text{def}}{=} \lim_{t\searrow 0}\frac{g(t)-g(0)}{t} = df(a;b).$$

2.2 Shape Derivative via the Velocity Method as a Differential

It is now well-established that the *Velocity Method*[1] developed by Zolésio [23] in 1979 is naturally associated with the construction of the *Courant metrics* on special groups of C^k-diffeomorphisms by Micheletti [16] in 1972. It turns out that the tangent space to those groups is linear and is made up of the velocities that generate

[1] See also Zolésio [21, 22] for the introduction of velocities in 1976.

the diffeomorphisms [4, 12]. In that case, a differential with respect to the velocity can be expected when the various functions involved are themselves differentiable.

In the *Velocity Method*, the domain $\Omega \subset \mathbb{R}^N$ is perturbed by a family of diffeomorphisms T_t of $\mathbb{R}^N$, $t \geq 0$, generated by sufficiently smooth velocity fields $V(t)$:

$$\frac{dx}{dt}(t; X) = V(t, x(t; X)),\ x(0; X) = X, \quad T_t(X) \stackrel{\text{def}}{=} x(t; X),\ t \geq 0,\ X \in \mathbb{R}^N .$$

Denote by $\Omega_t \stackrel{\text{def}}{=} T_t(\Omega)$ the perturbed domain. The state equation and objective function at $t \geq 0$ become: to find $u_t \in H_0^1(\Omega_t)$ such that

$$\int_{\Omega_t} \nabla u_t \cdot \nabla \psi - a\, \psi \, dx = 0, \quad \forall \psi \in H_0^1(\Omega_t), \quad f(\Omega_t) \stackrel{\text{def}}{=} \int_{\Omega_t} \frac{1}{2}|u_t - z|^2 \, dx. \tag{4}$$

Introducing the composition $u^t = u_t \circ T_t$ to work in the fixed space $H_0^1(\Omega)$:

$$\int_{\Omega} \left[A(t)\nabla u^t \cdot \nabla \psi - a \circ T_t\, \psi\, J_t\right] dx = 0, \quad \forall \psi \in H_0^1(\Omega), \tag{5}$$

$$A(t) = J_t\, DT_t{}^{-1}\, (DT_t{}^{-1})^*,\ J_t = \det DT_t, \quad DT_t \text{ is the Jacobian matrix},$$

$$g(t) \stackrel{\text{def}}{=} f(\Omega_t) = \int_{\Omega_t} \frac{1}{2}\,|u_t - z|^2\, dx = \int_{\Omega} \frac{1}{2}\,|u^t - z \circ T_t|^2\, J_t\, dx. \tag{6}$$

This technique is known as *Function Space Parametrization* [12, Chap. 10, sec. 6.3, p. 565] as opposed to *Function Space Embedding* that necessitates working with multivalued solutions (u^0, p^0) of the state/adjoint state equations and Theorem 4.1 or its Corollary in Sect. 4. Here, the t-dependent Lagrangian is

$$L(t, \varphi, \psi) \stackrel{\text{def}}{=} \int_{\Omega}\left[\frac{1}{2}|\varphi - z \circ T_t|^2\, J_t + A(t)\nabla\varphi \cdot \nabla\psi - a \circ T_t\, J_t\, \psi\right] dx \tag{7}$$

$$g(t) = \inf_{\varphi \in H_0^1(\Omega)} \sup_{\psi \in H_0^1(\Omega)} L(t, \varphi, \psi),\ dg(0) \stackrel{\text{def}}{=} \lim_{t \searrow 0}(g(t) - g(0))/t = df(\Omega; V(0)).$$

2.3 Topological Derivative via Dilatations as a Semidifferential

The rigorous introduction of the *topological derivative* in 1999 by Sokołowski and Zochowski [18])[2] provided a broader spectrum of notions of *derivative with*

[2]See also the book by Novotny-Sokołowski [17] and its bibliography for a review of past contributions to the field.

respect to a set. Initially, topological perturbations were induced by creating a hole corresponding to removing a small closed ball of radius r and center $e \in \Omega$ from a domain Ω. The ball can be seen as an r-dilatation $E_r = \{x \in \mathbb{R}^N : d_E(x) \leq r\}$ of the set $E = \{e\}$. It turns out that to make sense of a one-sided directional derivative, we have to use the *auxiliary variable* t equal to the volume $\mathrm{m}_N(E_r)$ in $\mathbb{R}^N$ of the ball of radius r and not the *dilatation parameter* r itself (m_N, the Lebesgue measure). This point of view differs from the widespread techniques of considering the topological derivative as a term in an expansion of the objective function with respect to r.

In that perspective, it is natural to seek to extend that construction to removing an r-dilatation of a curve, a surface, a submanifold, or a microstructure $E \subset \Omega$. As shown in [6], this idea readily extends to families of *d-rectifiable* closed subsets E of Ω of dimension d, $1 \leq d \leq N-1$, whose *Hausdorff measure* $H^d(E)$ is finite and to *sets of positive reach*. For the convenience of the reader, we recall definitions and theorems about of the *d-dimensional Minkowski content* and the *d-rectifiability* below. Sets of positive reach will not be used in this paper.

Notation $|\cdot|$denotes the Euclidean norm in $\mathbb{R}^N$. Given a closed subset E of $\mathbb{R}^N$ and $r \geq 0$, the distance function d_E and the *r-dilatation* E_r of E are defined as follows

$$d_E(x) \stackrel{\text{def}}{=} \inf_{e \in E} |x - e|, \quad E_r \stackrel{\text{def}}{=} \left\{ x \in \mathbb{R}^N : d_E(x) \leq r \right\}. \tag{8}$$

Given an open subset Ω of $\mathbb{R}^N$, $\mathscr{D}(\Omega)$ denotes the set of infinitely differentiable functions with compact support in Ω.

2.3.1 Tangent Space to the Group of Characteristic Functions

To be specific let m_N be the Lebesgue measure in $\mathbb{R}^N$. Identify the set of Lebesgue measurable subsets Ω of $\mathbb{R}^N$ with the set of their *characteristic functions* χ_Ω:

$$\mathrm{X}(\mathbb{R}^{\mathrm{N}}) \stackrel{\text{def}}{=} \left\{ \chi_\Omega : \Omega \subset \mathbb{R}^{\mathrm{N}} \text{ Lebesgue measurable} \right\} \subset L^\infty(\mathbb{R}^{\mathrm{N}}).$$

The *symmetric difference* operation $\Omega_2 \Delta \Omega_1$ induces an *Abelian group* structure:

$$\Omega_2 \Delta \Omega_1 \stackrel{\text{def}}{=} (\Omega_2 \backslash \Omega_1) \cup (\Omega_1 \backslash \Omega_2) \quad \Rightarrow \ \chi_{\Omega_2 \Delta \Omega_1}(x) = |\chi_{\Omega_2}(x) - \chi_{\Omega_1}(x)|,$$

where $\chi_\varnothing = 0$ is the neutral element and χ_Ω is its own *inverse*. The group $\mathrm{X}(\mathbb{R}^N)$ is a closed subset without interior of the Banach space $L^\infty(\mathbb{R}^{\mathrm{N}})$ and of the Fréchet spaces $L^p_{\mathrm{loc}}(\mathbb{R}^{\mathrm{N}})$, $1 \leq p < \infty$. For $L^\infty(\mathbb{R}^N)$ the metric is

$$\rho([\Omega_2], [\Omega_1]) \stackrel{\text{def}}{=} \|\chi_{\Omega_2} - \chi_{\Omega_1}\|_{L^\infty(\mathbb{R}^N)} = \|\chi_{\Omega_2 \Delta \Omega_1}\|_{L^\infty(\mathbb{R}^N)};$$

for the Fréchet space $L^p_{\text{loc}}(\mathbb{R}^N)$, use the seminorms on bounded open subsets $D \subset R^N$

$$\|\chi_{\Omega_2 \Delta \Omega_1}\|_{L^p(D)} = \|\chi_{\Omega_2} - \chi_{\Omega_1}\|_{L^p(D)}.$$

The *adjacent tangent cone*[3] to $\mathrm{X}(\mathbb{R}^N)$ always exists but is not a linear space.

Definition 2.1 (Aubin and Frankowska [2, pp. 126–128]) The *adjacent tangent cone* to A at $a \in \overline{A}$ is the set

$$T_a^\flat A \stackrel{\text{def}}{=} \left\{ v \in \mathbb{R}^N : \forall \{t_n \searrow 0\},\ \exists \{x_n\} \subset A \text{ such that } \lim_{n\to\infty} \frac{x_n - a}{t_n} = v \right\}$$

($T_a^\flat A$ is a closed cone and $T_a^\flat \overline{A} = T_a^\flat A$). □

As a result, for a function defined on $\mathrm{X}(\mathbb{R}^N)$, we can only expect a *semidifferential*, that is, a one-sided directional derivative which is *not necessarily linear* with respect to the directions that are *half* or *semi-tangents* in the cone. In fact, we shall see below that the adjacent tangent cone to $\mathrm{X}(\mathbb{R}^N)$ contains semi-tangents which are bounded measures associated with some d-dimensional closed subsets E of $\mathbb{R}^N$.

Example 2.1 Let $N \geq 1$, $e \in \mathbb{R}^N$, $E = \{e\}$, $\dim E = 0$, $E_r = \overline{B_r(e)}$. The function $\phi \mapsto \phi(e) : \mathscr{D}(\mathbb{R}^N) \to \mathbb{R}$ is a measure. The *auxiliary variable* t is chosen as the volume of $\overline{B_r(e)}$. Given an open subset Ω in $\mathbb{R}^N$ and $e \in \Omega$, the perturbed sets are

$$t \mapsto \Omega_t \stackrel{\text{def}}{=} \Omega \backslash \overline{B_r(e)} = \Omega \backslash B_{\sqrt[N]{t/\alpha_N}}(e),\ \alpha_N = \text{volume of the ball of radius 1 in } \mathbb{R}^N.$$

The trajectory $t \mapsto \chi_{\Omega_t}$ is continuous in $\mathrm{X}(\mathbb{R}^N)$. Given $\phi \in \mathscr{D}(\mathbb{R}^N)$, the weak limit of the differential quotient $(\chi_{\Omega_t} - \chi_\Omega)/t$ is

$$\begin{aligned}\frac{1}{t}\left[\int_{\Omega_t} \phi\, dx - \int_\Omega \phi\, dx\right] &= -\frac{1}{\mathrm{m}_N(B_{\sqrt[N]{t/\alpha_N}}(e))} \int_{B_{\sqrt[N]{t/\alpha_N}}(e)} \chi_\Omega\, \phi\, dx \\ &= -\frac{1}{\mathrm{m}_N(B_r(e))} \int_{B_r(e)} \chi_\Omega\, \phi\, dx \to -\phi(e).\end{aligned}$$

The limit is a distribution (measure), that is, $-\delta(e)$, the negative of the Dirac delta function at e. It generates a *half tangent* since for all $\rho > 0$

$$\frac{1}{t}\left[\int_{\Omega_{\rho t}} \phi\, dx - \int_\Omega \phi\, dx\right] \to -\rho\phi(e),$$

[3] We use the terminology *adjacent tangent cone* of Aubin and Frankowska [2, pp. 126–128] for the Dubovitski-Milyutin tangent cone.

but not a full tangent. Note that we can also create points by introducing the perturbed sets $\Omega_t = \Omega \cup B_{(t/\alpha_N)^{1/N}}(b)$ to get $+\phi(b)$, $b \in \mathbb{R}^N \setminus \overline{\Omega}$. □

Example 2.2 Let E be a compact C^2-submanifold of $\mathbb{R}^N$ of dimension d, $1 \leq d \leq N-1$, such that the $H^d(E) < \infty$. In that case $\partial E = E$ and $\mathrm{m}_N(E) = 0$. For instance, E could be a closed non-intersecting C^2-curve in $\mathbb{R}^3$ without boundary. Assuming that there exists $\varepsilon > 0$ such that $d_E^2 \in C^2(U_\varepsilon(E))$, $U_\varepsilon(E) = \{x \in \mathbb{R}^N : d_E(x) < \varepsilon\}$, the *projection* onto E is $p_E(x) = x - \frac{1}{2}\nabla d_E^2(x)$, $Dp_E(x) = I - \frac{1}{2}D^2 d_E^2(x)$, $\mathrm{Im}Dp_E(x)$ is the *tangent space* at $x \in E$ to E, and $d = \dim E(x) = \dim(\mathrm{Im}Dp_E(x))$.

Let $r > 0$, E_r be the *r-dilatation*, and $t = \alpha_{N-d}\, r^{N-d}$ be the *auxiliary variable*, where α_{N-d} is the volume of the $(N-d)$-dimensional ball of radius 1. Given $\varepsilon > 0$ such that $U_\varepsilon(E) = \{x \in \mathbb{R}^N : d_E(x) < \varepsilon\} \subset \Omega$, the perturbed sets will be

$$t \mapsto \Omega_t \overset{\text{def}}{=} \Omega \setminus E_r = \Omega \setminus E_{(t/\alpha_{N-d})^{N-d}}.$$

Given $\phi \in \mathscr{D}(\mathbb{R}^N)$, the weak limit of the differential quotient $(\chi_{\Omega_t} - \chi_\Omega)/t$ is

$$\frac{1}{t}\left[\int_{\Omega_t} \phi\, d\mathrm{m}_N - \int_\Omega \phi\, d\mathrm{m}_N\right] = -\frac{1}{t}\int_{E_{(t/\alpha_{N-d})^{N-d}}} \chi_\Omega\, \phi\, d\mathrm{m}_N$$

$$= -\frac{1}{\alpha_{N-d}\, r^{N-d}}\int_{E_r} \chi_\Omega\, \phi\, d\mathrm{m}_N \to -\int_E \phi\, dH^d.$$

This distribution (measure) is again a *half or semi-tangent* since for $\rho > 0$

$$\frac{1}{t}\left[\int_{\Omega_{\rho t}} \phi\, d\mathrm{m}_N - \int_\Omega \phi\, d\mathrm{m}_N\right] \to -\rho\int_E \phi\, dH^d.$$

□

2.3.2 The d-Dimensional Minkowski Content and d-Rectifiable Sets

The semi-tangents that we have constructed are directly related to the notion of *d-dimensional Minkowski content* [14]

$$M^d(E) \overset{\text{def}}{=} \lim_{r \searrow 0} \frac{\mathrm{m}_N(E_r)}{\alpha_{N-d}\, r^{N-d}}, \quad \alpha_{N-d} = \text{ volume of the unit ball in } \mathbb{R}^{N-d}, \tag{9}$$

for general topological perturbations obtained by dilation of smooth submanifolds E of dimension d in $\mathbb{R}^N$. The case of $E = \{e\}$ corresponds to $d = 0$, while $d = 1$ corresponds to a curve and $d = 2$ to a surface. We shall see below that $M^d(E)$ is equal to $H^d(E)$, the d-dimensional Hausdorff measure in $\mathbb{R}^N$, for compact *d-rectifiable subsets* E of $\mathbb{R}^N$. We recall some basic notions. For details and other motivating examples the reader is referred to the recent papers of Delfour [5, 6].

Definition 2.2 Given d, $0 \leq d \leq N$, the *d-dimensional upper and lower Minkowski contents* of a set E are defined through an r-dilatation of the set E as follows

$$M^{*d}(E) \stackrel{\text{def}}{=} \limsup_{r \searrow 0} \frac{\mathrm{m}_N(E_r)}{\alpha_{N-d}\, r^{N-d}}, \quad M^d_*(E) \stackrel{\text{def}}{=} \liminf_{r \searrow 0} \frac{\mathrm{m}_N(E_r)}{\alpha_{N-d}\, r^{N-d}}, \tag{10}$$

where m_N is the Lebesgue measure in $\mathbb{R}^N$ and α_{N-d} is the volume of the ball of radius one in $\mathbb{R}^{N-d}$. Since the dilatation E_r does not distinguish between E and its closure, it can be assumed that E is closed in $\mathbb{R}^N$. When the two limits exist and are equal, we say that E admits a d-dimensional *Minkowski content* and their common value will be denoted $M^d(E)$. □

Intuitively, $M^d(E)$ is some *measure* of the d-dimensional "area" or "volume" of an object E in $\mathbb{R}^N$. It plays a role similar to the *d-dimensional Hausdorff or Radon measure* in $\mathbb{R}^N$, but it is generally not a measure.

Thanks to the pioneering and seminal work of Federer [14], the previous constructions can be readily extended to the dilation of *d-rectifiable sets*. Further extensions of the notion of d-rectifiability can be found in Ambrosio et al. [1].

Definition 2.3 (Federer [14, pp. 251–252]) Let E be a subset of a metric space X. $E \subset X$ is *d-rectifiable* if it is the image of a compact subset K of $\mathbb{R}^d$ by a Lipschitz continuous function $f : \mathbb{R}^d \to X$.

Theorem 2.1 ([14, p. 275]) *If $E \subset \mathbb{R}^N$ is compact and d-rectifiable, then $M^d(E) = H^d(E)$.*

In this paper, we are interested in closed subsets E of $\mathbb{R}^N$ such that the d-dimensional *Minkowski content* (9) exists and for which M^d would be a *measure* such that the following limit

$$\phi \mapsto \int_E \phi \, dM^d = \lim_{r \searrow 0} \frac{1}{\alpha_{N-d}\, r^{N-d}} \int_{\mathbb{R}^N} \chi_{E_r} \phi \, d\mathrm{m}_N : \mathscr{D}(\mathbb{R}^N) \to \mathbb{R} \tag{11}$$

makes sense. Therefore, in applications to d-dimensional objects, $0 \leq d \leq N$, the appropriate choice of *auxiliary variable* is the volume $t = \alpha_{N-d}\, r^{N-d}$ of the ball of radius r in $\mathbb{R}^{N-d}$, that is, $r = (t/\alpha_{N-d})^{1/(N-d)}$, and

$$\phi \mapsto \int_E \phi \, dM^d = \lim_{t \searrow 0} \frac{1}{t} \int_{E_{(t/\alpha_{N-d})^{1/(N-d)}}} \phi \, d\mathrm{m}_N : \mathscr{D}(\mathbb{R}^N) \to \mathbb{R}. \tag{12}$$

Given a Lebesgue measurable set $\Omega \subset \mathbb{R}^N$ and a d-rectifiable closed subset E such that $\mathrm{m}_N(E) = 0$, several perturbations can be introduced:

$$\Omega_t = \Omega \backslash E_r, \quad \Omega_t = \Omega \cup E_r, \text{ and } \Omega_t = \Omega \triangle E_r, \quad t = \alpha_{N-d}\, r^{N-d}, \tag{13}$$

depending on whether E_r is removed, added, or both removed and added. In each case a continuous trajectory $t \mapsto \chi_{\Omega_t}$ is obtained in $\mathrm{X}(\mathbb{R}^N)$ such that

$$\chi_{\Omega_t} \to \chi_\Omega \ \text{ in } L^p_{\mathrm{loc}}(\mathbb{R}^N), \quad 1 \le p < \infty.$$

2.3.3 Back to the Example

Going back to the example for the state equation (1) and the objective function (2), we consider the topological derivative with respect to the dilatation of a point e from Novotny and Sokolowski [17] and extend it to dilatations of curves and surfaces.

Given a bounded open subset Ω of $\mathbb{R}^3$ with Lipschitz boundary Γ and a compact d-rectifiable (see Definition 2.3 and Theorem 2.1) subset $E \subset \Omega$, $0 \le d \le 2$ ($d = 0$ for $E = \{e\}$, $d = 1$ for a curve and $d = 2$ for a surface), consider the family of *perturbed state equations*: to find $u^t = u^t(E) \in H^1_0(\Omega)$ such that[4]

$$\forall \psi \in H^1_0(\Omega), \quad \int_\Omega \nabla u^t \cdot \nabla \psi \, dx = \int_\Omega \left[a - (1-\gamma)\,\chi_{E_r}\right] \psi \, dx \tag{14}$$

parametrized by the *auxiliary variable* $t = \alpha_{3-d} r^{3-d}$, where α_{3-d} is the volume of the unit ball in $\mathbb{R}^{3-d}$ and χ_A denotes the characteristic function of a set A. The *objective function* for $t \ge 0$ becomes

$$J(t) \overset{\text{def}}{=} \int_\Omega \frac{1}{2} |u^t(E) - z|^2 \, dx, \quad z \in L^2(\Omega). \tag{15}$$

The t-dependent *Lagrangian* is now

$$L(t, \varphi, \psi) \overset{\text{def}}{=} \int_\Omega \frac{1}{2} |\varphi - z|^2 + \nabla \varphi \cdot \nabla \psi \, dx - \left[a - (1-\gamma)\,\chi_{E_r}\right] \psi \, dx$$

for $(\varphi, \psi) \in H^1_0(\Omega) \times H^1_0(\Omega)$. Finally, the computation of the topological derivative takes the same form as in the previous examples:

$$g(t) \overset{\text{def}}{=} \inf_{\varphi \in H^1_0(\Omega)} \sup_{\psi \in H^1_0(\Omega)} L(t, \varphi, \psi), \, dg(0) \overset{\text{def}}{=} \lim_{t \searrow 0} \frac{g(t) - g(0)}{t} = \lim_{t \searrow 0} \frac{J(t) - J(0)}{t}.$$

It is readily seen that for $E = \{e\}$, $e \in \Omega$, and $\psi \in H^2(\Omega) \cap H^1_0(\Omega)$,

$$\frac{L(t, \varphi, \psi) - L(0, \varphi, \psi)}{t} = (1-\gamma) \frac{1}{|B_r(e)|} \int_{B_r(e)} \chi_\Omega \, \psi \, dx \to (1-\gamma)\, \psi(e)$$

[4] In [17], $a = \chi_\Omega$, the characteristic function of Ω.

and $d_t L(0, \varphi, \psi) = (1 - \gamma)\, \psi(e)$ by the Lebesgue differentiation theorem. For a compact d-rectifiable set $E \subset \Omega$ ($d = 1$ for a curve and $d = 2$ for a surface) and $\psi \in H^2(\Omega) \cap H_0^1(\Omega)$

$$d_t L(0, \varphi, \psi) = (1 - \gamma) \int_E \psi \, dH^d. \tag{16}$$

3 Minimax for State Constrained Objective Functions

3.1 Abstract Framework

In this section we recall the framework used in [19, 20] and extended in [7, 8] for the mutivalued case. A *Lagrangian* is a function of the form

$$(t, x, y) \mapsto G(t, x, y) : [0, \tau] \times X \times Y \to \mathbb{R}, \quad \tau > 0,$$

where Y is a *vector space*, X is a non empty subset of a vector space, and $y \mapsto G(t, x, y)$ is *affine*. Associate with the *parameter* t the *parametrized minimax*

$$t \mapsto g(t) \stackrel{\text{def}}{=} \inf_{x \in X} \sup_{y \in Y} G(t, x, y) : [0, \tau] \to \mathbb{R} \quad \text{and} \quad dg(0) \stackrel{\text{def}}{=} \lim_{t \searrow 0} \frac{g(t) - g(0)}{t}. \tag{17}$$

When the limits exist we shall use the following compact notation:

$$d_t G(0, x, y) \stackrel{\text{def}}{=} \lim_{t \searrow 0} \frac{G(t, x, y) - G(0, x, y)}{t}$$

$$\varphi \in X, \quad d_x G(t, x, y; \varphi) \stackrel{\text{def}}{=} \lim_{\theta \searrow 0} \frac{G(t, x + \theta\varphi, y) - G(t, x, y)}{\theta}$$

$$\psi \in Y, \quad d_y G(t, x, y; \psi) \stackrel{\text{def}}{=} \lim_{\theta \searrow 0} \frac{G(t, x, y + \theta\psi) - G(t, x, y)}{\theta}.$$

The notation $t \searrow 0$ and $\theta \searrow 0$ means that t and θ go to 0 by strictly positive values.

Since $G(t, x, y)$ is affine in y, for all $(t, x) \in [0, \tau] \times X$,

$$\forall y, \psi \in Y, \quad d_y G(t, x, y; \psi) = G(t, x, \psi) - G(t, x, 0) = d_y G(t, x, 0; \psi).$$

The *state equation* at $t \geq 0$:

$$\text{to find } x^t \in X \text{ such that for all } \psi \in Y,\ d_y G(t, x^t, 0; \psi) = 0. \tag{18}$$

The set of solutions (*states*) x^t at $t \geq 0$ is denoted

$$E(t) \overset{\text{def}}{=} \{x^t \in X : \forall \varphi \in Y,\ d_y G(t, x^t, 0; \varphi) = 0\}.$$

The *(standard) adjoint state equation* at $t \geq 0$ is

$$\text{to find } p^t \in Y \text{ such that } \forall \varphi \in X, \quad d_x G(t, x^t, p^t; \varphi) = 0 \tag{19}$$

and the set of solutions will be denoted $Y(t, x^t)$. Finally, the set of minimizers for the minimax is given by

$$X(t) \overset{\text{def}}{=} \left\{ x^t \in X : g(t) = \inf_{x \in X} \sup_{y \in Y} G(t, x, y) = \sup_{y \in Y} G(t, x^t, y) \right\}. \tag{20}$$

Lemma 3.1 (Constrained Infimum and Minimax)

(i) $\inf_{x \in X} \sup_{y \in Y} G(t, x, y) = \inf_{x \in E(t)} G(t, x, 0)$.
(ii) *The minimax* $g(t) = +\infty$ *if and only if* $E(t) = \varnothing$. *Hence* $X(t) = X$.
(iii) *If* $E(t) \neq \varnothing$, *then* $g(t) < +\infty$ *and*

$$X(t) = \{x^t \in E(t) : G(t, x^t, 0) = \inf_{x \in E(t)} G(t, x, 0)\} \subset E(t). \tag{21}$$

Proof Cf., for instance, [8]. □

Hypothesis (H0) *Let X be a vector space.*

(i) *For all* $t \in [0, \tau]$, $x^0 \in X(0)$, $x^t \in X(t)$, *and* $y \in Y$, *the function*

$$s \mapsto G(t, x^0 + s(x^t - x^0), y) : [0, 1] \to \mathbb{R} \tag{22}$$

is absolutely continuous. This implies that, for almost all s, the derivative exists and is equal to $d_x G(t, x^0 + s(x^t - x^0), y; x^t - x^0)$ and that it is the integral of its derivative. In particular,

$$G(t, x^t, y) = G(t, x^0, y) + \int_0^1 d_x G(t, x^0 + s(x^t - x^0), y; x^t - x^0)\, ds. \tag{23}$$

(ii) *For all* $t \in [0, \tau]$, $x^0 \in X(0)$, $x^t \in X(t)$, $y \in Y$, $\varphi \in X$, *and almost all* $s \in (0, 1)$, $d_x G(t, x^0 + s(x^t - x^0), y; \varphi)$ *exists*
and the function $s \mapsto d_x G(t, x^0 + s(x^t - x^0), y; \varphi)$ *belongs to* $L^1(0, 1)$.

Definition 3.1 (Sturm [19, 20]) Given $x^0 \in X(0)$ and $x^t \in X(t)$, the *averaged adjoint state equation* is

$$\text{to find } y^t \in Y,\ \forall \varphi \in X, \quad \int_0^1 d_x G(t, x^0 + s(x^t - x^0), y^t; \varphi)\, ds = 0 \tag{24}$$

and the set of solutions is denoted $Y(t, x^0, x^t)$. □

$Y(0, x^0, x^0)$ clearly reduces to the set of *standard adjoint states* $Y(0, x^0)$ at $t = 0$.

3.2 Original Condition of Sturm and Its First Extension

Theorem 3.1 ([19], [20, Thm. 3.1]) *Consider the Lagrangian functional*

$$(t, x, y) \mapsto G(t, x, y) : [0, \tau] \times X \times Y \to \mathbb{R}, \quad \tau > 0,$$

where X and Y are vector spaces and the function $y \mapsto G(t, x, y)$ is affine. Let (H0) *and the following hypotheses be satisfied:*

(H1) *for all $t \in [0, \tau]$, $g(t)$ is finite, $X(t) = \{x^t\}$ and $Y(t, x^0, x^t) = \{y^t\}$ are singletons;*
(H2) *$d_t G(t, x^0, y)$ exists for all $t \in [0, \tau]$ and all $y \in Y$;*
(H3) *the following limit exists*[5]

$$\lim_{s \searrow 0,\, t \searrow 0} d_t G(s, x^0, y^t) = d_t G(0, x^0, y^0). \tag{25}$$

Then, $dg(0)$ exists and $dg(0) = d_t G(0, x^0, y^0)$.

This theorem was extended in [7, 8] with a weakening of *(H2)* and *(H3)* that resulted in the appearance of an extra term in the expression of $dg(0)$. An example of a *topological derivative* where that extra term is not zero was given in [7].

Theorem 3.2 (Singleton Case [7, 8]) *Consider the Lagrangian functional*

$$(t, x, y) \mapsto G(t, x, y) : [0, \tau] \times X \times Y \to \mathbb{R}, \quad \tau > 0,$$

where X and Y are vector spaces and the function $y \mapsto G(t, x, y)$ is affine. Let (H0) *and the following hypotheses be satisfied:*

(H1) *for all $t \in [0, \tau]$, $g(t)$ is finite, $X(t) = \{x^t\}$ and $Y(t, x^0, x^t) = \{y^t\}$ are singletons;*
(H2') *$d_t G(0, x^0, y^0)$ exists;*

[5] Condition (H3) is typical of what can be found in the literature (see, for instance, [3]).

(H3') *the following limit exists*

$$R(x^0, y^0) \overset{def}{=} \lim_{t \searrow 0} d_y G\left(t, x^0, 0; \frac{y^t - y^0)}{t}\right). \tag{26}$$

Then, $dg(0)$ exists and $dg(0) = d_t G(0, x^0, y^0) + R(x^0, y^0)$.

Notice that, under condition (H2'), condition (H3') is optimal since

$$dg(0) \text{ exists} \iff \lim_{t \searrow 0} d_y G\left(t, x^0, 0; \frac{y^t - y^0}{t}\right) \text{ exists.}$$

Hypotheses (H2') and (H3') are weaker and more general than (H2) and (H3). Indeed, it is readily seen that if (H2)–(H3) are verified, then (H2')–(H3') are verified with $R(x^0, y^0) = 0$:

(H2') it is only assumed that $d_t G(0, x^0, y^0)$ exists. Hypothesis (H2) assumes that $d_t G(t, x^0, y)$ exists for all $t \in [0, \tau]$ and $y \in Y$.

(H3') Hypothesis (H3) assumes that

$$\lim_{s \searrow 0,\, t \searrow 0} d_t G(s, x^0, y^t) = d_t G(0, x^0, y^0)$$

which implies that $R(x^0, y^0) = 0$ (see the proof of Theorem 3.1 and Remark 3.2 in [8] for details). Thus, condition (H3') with $R(x^0, y^0) = 0$ is definitely weaker and more general than (H3) since it extends to cases where $R(x^0, y^0)$ is not zero.

Since $d_x G$ and $d_x d_y G$ both exist, Hypothesis (H3') can be rewritten as follows

$$\begin{aligned}
&d_y G\left(t, x^0, 0; \frac{y^t - y^0}{t}\right) \\
&= d_y G\left(t, x^0, 0; \frac{y^t - y^0}{t}\right) - d_y G\left(t, x^t, 0; \frac{y^t - y^0}{t}\right) \\
&= \int_0^1 d_x d_y G\left(t, \theta x^0 + (1-\theta)x^t, 0; \frac{y^t - y^0}{t^\alpha}; \frac{x^0 - x^t}{t^{1-\alpha}}\right) d\theta,
\end{aligned}$$

for some $\alpha \in [0, 1]$. For instance, with $\alpha = 1/2$, it would be sufficient to find bounds on the differential quotients

$$(y^t - y^0)/t^{1/2} \text{ and } (x^t - x^0)/t^{1/2}$$

which is less demanding than finding a bound on $(x^t - x^0)/t$ or $(y^t - y^0)/t$. When the integral with respect to θ can be taken inside, the expressions simplify

$$d_y G\left(t, x^0, 0; \frac{y^t - y^0}{t}\right) = d_x d_y G\left(t, \frac{x^0 + x^t}{2}, 0; \frac{y^t - y^0}{t^\alpha}; \frac{x^0 - x^t}{t^{1-\alpha}}\right)$$

$$\lim_{t \searrow 0} d_y G\left(t, x^0, 0; \frac{y^t - y^0}{t}\right) = \lim_{t \searrow 0} d_x d_y G\left(t, \frac{x^0 + x^t}{2}, 0; \frac{y^t - y^0}{t^\alpha}; \frac{x^0 - x^t}{t^{1-\alpha}}\right).$$

It is a second order condition without assuming second order derivatives in x.

3.3 A New Condition with the Standard Adjoint at $t = 0$

The use of the *averaged adjoint* revealed the possible occurrence of an *extra term* and provided a simpler expression of the former hypothesis (H3). It turns out that the extra term can also be obtained by using the *standard adjoint* at $t = 0$ significantly simplifying the checking of that condition.

Theorem 3.3 (Singleton Case) *Consider the Lagrangian functional*

$$(t, x, y) \mapsto G(t, x, y) : [0, \tau] \times X \times Y \to \mathbb{R}, \quad \tau > 0,$$

where X and Y are vector spaces and the function $y \mapsto G(t, x, y)$ is affine. Let (H0) *and the following hypotheses be satisfied:*

(H1) *for all $t \in [0, \tau]$, $g(t)$ is finite, $X(t) = \{x^t\}$ and $Y(0, x^0) = \{p^0\}$ are singletons;*
(H2") *$d_t G(0, x^0, y^0)$ exists;*
(H3") *the following limit exists*

$$R(x^0, p^0) \stackrel{def}{=} \lim_{t \searrow 0} \int_0^1 d_x G\left(t, x^0 + \theta(x^t - x^0), p^0; \frac{x^t - x^0}{t}\right) d\theta. \qquad (27)$$

Then, $dg(0)$ exists and $dg(0) = d_t G(0, x^0, p^0) + R(x^0, p^0)$.

Proof Recalling that $g(t) = G(t, x^t, y)$ and $g(0) = G(0, x^0, y)$ for any $y \in Y$, then for the standard adjoint state p^0 at $t = 0$

$$g(t) - g(0) = G(t, x^t, p^0) - G(t, x^0, p^0) + \Big(G(t, x^0, p^0) - G(0, x^0, p^0)\Big).$$

Dividing by $t > 0$

$$\frac{g(t) - g(0)}{t} = \frac{G(t, x^t, p^0) - G(t, x^0, p^0)}{t} + \frac{G(t, x^0, p^0) - G(0, x^0, p^0)}{t}$$
$$= \int_0^1 d_x G\left(t, x^0 + \theta(x^t - x^0), p^0; \frac{x^t - x^0}{t}\right) d\theta + \frac{G(t, x^0, p^0) - G(0, x^0, p^0)}{t}.$$

Therefore, in view of Hypothesis (H2"), the limit $dg(0)$ exists if and only if the limit of the first term exists. Therefore

$$dg(0) = \lim_{t \searrow 0} \int_0^1 d_x G\left(t, x^0 + \theta(x^t - x^0), p^0; \frac{x^t - x^0}{t}\right) d\theta + d_t G(0, x^0, p^0)$$

and the existence of the limit of the first term replaces hypothesis (H3'). □

Remark 3.1

(i) Hypothesis (H3") seems *less demanding* than (H3') since it only requires the existence of the *standard adjoint* p^0 at $t = 0$ while Hypotheses (H3) and (H3') necessitate the existence of the averaged adjoint y^t and the study of the differential quotient $(y^t - y^0)/t$ for small $t \geq 0$ going to 0.

(ii) (Separation Principle) Theorem 3.3 also *separates* the study of the state/adjoint state system (x^0, p^0) at $t = 0$ from the study of the differential quotient $(x^t - x^0)/t$ of the state. In the examples, the pair (x^0, p^0) is independent of the fact that we compute a semidifferential with respect to the control, the shape or the topology. □

3.4 Application of Theorem 3.3 to the Examples

3.4.1 Directional Derivative with Respect to the Control

Recall that, given the direction $b \in L^2(\Omega)$ and the perturbation $a + tb$ of the control a, we want to compute $df(a; b) = \lim_{t \searrow 0}(f(a + tb) - f(a))/t$. The *state* $u^t \in H_0^1(\Omega)$ at $t \geq 0$ is solution of the state equation

$$\int_\Omega \nabla u^t \cdot \nabla \psi - (a + tb)\, \psi \, dx = 0, \quad \forall \psi \in H_0^1(\Omega), \tag{28}$$

and the *t-Lagrangian* is

$$L(t, \varphi, \psi) \stackrel{\text{def}}{=} \int_\Omega \frac{1}{2}|\varphi - z|^2 \, dx + \int_\Omega \nabla \varphi \cdot \nabla \psi - (a + tb)\, \psi \, dx.$$

It is readily seen that

$$d_y L(t,\varphi,\psi;\psi') = \int_\Omega \nabla\varphi\cdot\nabla\psi' - (a+tb)\,\psi'\,dx$$

$$d_x L(t,\varphi,\psi;\varphi') = \int_\Omega (\varphi - z)\,\varphi' + \nabla\varphi'\cdot\nabla\psi\,dx, \quad d_t L(t,\varphi,\psi) = -\int_\Omega b\,\psi\,dx.$$

Observe that the *derivative of the state* $\dot{u} \in H_0^1(\Omega)$ exists since

$$\int_\Omega \nabla\left(\frac{u^t - u^0}{t}\right)\cdot\nabla\psi - b\,\psi\,dx = 0, \quad \forall\psi \in H_0^1(\Omega), \tag{29}$$

implies that $(u^t - u^0)/t = \dot{u} \in H_0^1(\Omega)\cap H^2(\Omega)$ is solution of

$$\int_\Omega \nabla\dot{u}\cdot\nabla\psi - b\,\psi\,dx = 0, \quad \psi \in H_0^1(\Omega). \tag{30}$$

The *adjoint* $p^0 \in H_0^1(\Omega)\cap H^2(\Omega)$ is solution of

$$\int_\Omega (u^0 - z)\,\varphi + \nabla p^0\cdot\nabla\varphi\,dx = 0, \quad \forall\varphi \in H_0^1(\Omega). \tag{31}$$

It remains to check that the limit as t goes to 0 exists in (27):

$$\begin{aligned} R(t) &\overset{\text{def}}{=} \int_0^1 d_x G\left(t, u^0 + \theta(u^t - u^0), p^0; \frac{u^t - u^0}{t}\right) d\theta \\ &= \int_\Omega \left(\frac{u^t + u^0}{2} - z\right)\left(\frac{u^t - u^0}{t}\right) + \nabla p^0\cdot\nabla\left(\frac{u^t - u^0}{t}\right) dx \\ &= \int_\Omega \left(\frac{u^t + u^0}{2} - z\right)\left(\frac{u^t - u^0}{t}\right) - (u^0 - g)\left(\frac{u^t - u^0}{t}\right) dx \\ &= \int_\Omega \left(\frac{u^t - u^0}{2}\right)\left(\frac{u^t - u^0}{t}\right) dx = \frac{t}{2}\int_\Omega |\dot{u}|^2\,dx \to 0 \end{aligned}$$

by using Eq. (31) for p^0. Therefore, by Theorem 3.3,

$$df(a;b) = -\int_\Omega b\,p^0\,dx, \quad p^0 \in H_0^1(\Omega)\cap H^2(\Omega) \tag{32}$$

$$-\Delta u^0 = a \text{ in } \Omega,\ u^0 = 0 \text{ on } \Gamma, \quad -\Delta p^0 = u^0 - z \text{ in } \Omega,\ p^0 = 0 \text{ on } \Gamma. \tag{33}$$

3.4.2 Shape Derivative

Recall that the t-dependent Lagrangian is given by expression (7)

$$L(t,\varphi,\psi)=\int_\Omega\left[\frac{1}{2}|\varphi-z\circ T_t|^2\,J_t+A(t)\nabla\varphi\cdot\nabla\psi-a\circ T_t\,J_t\,\psi\right]dx \tag{34}$$

$$A(t)=J_t\,DT_t^{-1}\,(DT_t^{-1})^*,\ J_t=\det\,DT_t,\qquad DT_t \text{ is the Jacobian matrix.}$$

The state equation at $t\geq 0$ and the adjoint state equation at $t=0$ are

$$u^t\in H_0^1(\Omega),\ \forall\psi\in H_0^1(\Omega),\quad \int_\Omega\left\{A(t)\nabla u^t\cdot\nabla\psi-J_t\,(a\circ T_t)\,\psi\right\}dx=0, \tag{35}$$

$$p^0\in H_0^1(\Omega),\ \forall\varphi\in H_0^1(\Omega),\quad \int_\Omega\left\{\nabla p^0\cdot\nabla\varphi+(u^0-z)\,\varphi\right\}dx=0. \tag{36}$$

The pair $(u^0,p^0)\in H_0^1(\Omega)\cap H^2(\Omega)\times H_0^1(\Omega)\cap H^2(\Omega)$ is the solution of the same system (33) as in the previous example.

For $V\in C^0([0,\tau];C_0^1(\mathbb{R}^N,\mathbb{R}^N))$ and the diffeomorphism $T_t(V)=T_t$

$$\frac{d}{dt}T_t(X)=V(t,T_t(X)),\quad T_0(X)=X,\quad \frac{dT_t}{dt}=V(t)\circ T_t,\quad T_0=I, \tag{37}$$

where $V(t)(X)=V(t,X)$ and I is the identity matrix on $\mathbb{R}^{\mathrm{N}}$. Moreover,

$$\frac{d}{dt}DT_t=DV(t)\circ T_t\,DT_t,\quad DT_0=I,\quad \frac{d}{dt}J_t=\operatorname{div}V(t)\circ T_t\,J_t,\quad J_0=1, \tag{38}$$

where $DV(t)$ and DT_t are the Jacobian matrices of $V(t)$ and T_t. For $k\geq 1$, $C_0^k(\mathbb{R}^N,\mathbb{R}^N)$ is the space of k times continuously differentiable functions from $\mathbb{R}^N$ to $\mathbb{R}^N$ going to zero at infinity; for $k=0$, $C_0^0(\mathbb{R}^N,\mathbb{R}^N)$ is the space of continuous functions from $\mathbb{R}^N$ to $\mathbb{R}^N$ going to zero at infinity. We shall also use the notation

$$V_t\overset{\text{def}}{=}V(t)\circ T_t,\ \ V_t(X)=V(t,T_t(X)),\ \ f(t)\overset{\text{def}}{=}T_t-I,\ \ f(t)(X)=T_t(X)-X.$$

We regroup the main properties from [12, Thm. 4.4, Chap. 4, p. 189].

Lemma 3.2 *Assume that* $V\in C^0([0,\tau];C_0^1(\mathbb{R}^N,\mathbb{R}^N))$, *then*

$$f\in C^1\left([0,\tau];C_0^1(\mathbb{R}^N,\mathbb{R}^N)\right). \tag{39}$$

For $\tau>0$ *sufficiently small* $J_t=\det DT_t=|\det DT_t|=|J_t|$, $0\leq t\leq\tau$, *and there*

exist constants $0 < \alpha < \beta$ *such that*

$$\forall \xi \in \mathbb{R}^N, \quad \alpha|\xi|^2 \le A(t)\,\xi \cdot \xi \le \beta|\xi|^2 \text{ and } \alpha \le J_t \le \beta. \tag{40}$$

(i) *As* t *goes to zero,* $V_t \to V(0)$ *in* $C_0^1(\mathbb{R}^N, \mathbb{R}^N)$, $DT_t \to I$ *in* $C_0^0(\mathbb{R}^N, \mathbb{R}^N)$, $J_t \to 1$ *in* $C_0^0(\mathbb{R}^N, \mathbb{R})$,

$$\frac{DT_t - I}{t} \text{ is bounded in } C_0^0(\mathbb{R}^N, \mathbb{R}^N), \quad \frac{J_t - 1}{t} \text{ is bounded in } C_0^0(\mathbb{R}^N).$$

(ii) *As* t *goes to zero,*

$$A(t) \to I \text{ in } C_0^0(\mathbb{R}^N, \mathbb{R}^N), \quad \frac{A(t) - I}{t} \text{ is bounded in } C_0^0(\mathbb{R}^N, \mathbb{R}^N),$$

$$A'(t) = \operatorname{div} V_t\, I - DV_t - DV_t^* \to A'(0) = \operatorname{div} V(0) - DV(0) - DV(0)^* \text{ in } C_0^0(\mathbb{R}^N, \mathbb{R}^N).$$

where DV_t *is the Jacobian matrix of* V_t, *and* DV_t^* *is the transpose of* DV_t.

(iii) *Given* $h \in H^1(\mathbb{R}^N)$, *as* t *goes to zero,*

$$h \circ T_t \to h \text{ in } L^2(\Omega), \quad \frac{h \circ T_t - h}{t} \text{ is bounded in } L^2(\Omega)$$

$$\nabla h \cdot V_t \to \nabla h \cdot V(0) \text{ in } L^2(\Omega).$$

Since the bilinear forms associated with (35) and (36) are coercive, there exists a unique u^t and a unique pair (u^0, p^0) solution of the system (35)–(36). Hence

$$\forall t \in [0, \tau], \quad X(t) = \{u^t\},\ Y(0, u^0) = \{p^0\} \tag{41}$$

are singletons. So assumption (H1) is verified.

To check (H2") we use expression (34) with $\varphi = u^0$ and $\psi = p^0$:

$$d_t G(t, u^0, p^0) = \int_\Omega \left\{ \frac{1}{2}(u^0 - z \circ T_t)^2 \operatorname{div} V_t - (u^0 - z \circ T_t)\, \nabla z \cdot V_t\, J_t \right\} dx$$
$$+ \int_\Omega \left\{ A'(t)\, \nabla u^0 \cdot \nabla p^0 - (a \circ T_t \operatorname{div} V_t - \nabla a \cdot V_t J_t)\, \psi \right\} dx.$$

Using Lemma 3.2, we can let t go to zero in the above expression and

$$d_t G(0, u^0, p^0) = \int_\Omega \left\{ \frac{1}{2}(u^0 - z)^2 \operatorname{div} V(0) - (u^0 - z)\, \nabla z \cdot V(0) \right\} dx$$
$$+ \int_\Omega \left\{ A'(0)\, \nabla u^0 \cdot \nabla p^0 - (\operatorname{div} V(0)\, a + \nabla a \cdot V(0))\ p^0 \right\} dx. \tag{42}$$

So, condition (H2") is satisfied.

To check condition (H3"), we need the x-derivative of $L(t, \varphi, \psi)$

$$L(t, \varphi, \psi) = \int_\Omega \left[\frac{1}{2} |\varphi - z \circ T_t|^2 \, J_t + A(t) \, \nabla\varphi \cdot \nabla\psi - a \circ T_t \, \psi \, J_t \right] dx \tag{43}$$

$$d_x L(t, \varphi, \psi; \varphi') = \int_\Omega \left[(\varphi - z \circ T_t) \; J_t \, \varphi' + A(t) \, \nabla\psi \cdot \nabla\varphi' \right] dx \tag{44}$$

$$\begin{aligned} R(t) &\stackrel{\text{def}}{=} \int_0^1 d_x L \left(t, u^0 + \theta \, (u^t - u^0), p^0; \frac{u^t - u^0}{t} \right) d\theta \\ &= \int_\Omega \left[\left(\frac{u^0 + u^t}{2} - z \circ T_t \right) J_t \, \frac{u^t - u^0}{t} + A(t) \, \nabla p^0 \cdot \nabla \left(\frac{u^t - u^0}{t} \right) \right] dx. \end{aligned} \tag{45}$$

By substituting $\varphi = (u^t - u^0)/t$ in the adjoint equation for p^0,

$$\int_\Omega (u^0 - z) \, \frac{u^t - u^0}{t} + \nabla p^0 \cdot \nabla \left(\frac{u^t - u^0}{t} \right) dx = 0, \tag{46}$$

we can rewrite the expression of $R(t)$ as follows

$$\begin{aligned} R(t) = \int_\Omega & \frac{A(t) - I}{t} \, \nabla p^0 \cdot \nabla \left(u^t - u^0 \right) \\ & + \frac{J_t - 1}{t} \left(\frac{u^0 + u^t}{2} - z \circ T_t \right) \left(u^t - u^0 \right) dx + \frac{t}{2} \int_\Omega \left| \frac{u^t - u^0}{t} \right|^2 dx. \end{aligned}$$

From this, we get the following estimate

$$\begin{aligned} |R(t)| \le & \left\| \frac{A(t) - I}{t} \right\|_{C[0,\tau]} \|\nabla p^0\| \, \|\nabla(u^t - u^0)\| \\ & + \left\| \frac{J_t - 1}{t} \right\|_{C[0,\tau]} \left\| \frac{u^0 + u^t}{2} - z \right\| \, \|u^t - u^0\| + \frac{t}{2} \left\| \frac{u^t - u^0}{t} \right\|^2. \end{aligned}$$

By Lemma 3.2 the terms $(A(t) - I)/t$ and $(J_t - 1)/t$ are uniformly bounded. To conclude that the limit of $R(t)$ exists and is zero, it remains to show that $u^t \to u^0$ in $H_0^1(\Omega)$-strong and that the $L^2(\Omega)$ norm of $(u^t - u^0)/t$ is bounded.

From the state equations (35) of u^t and u^0, for all $\psi \in H_0^1(\Omega)$

$$\begin{aligned} & \int_\Omega \nabla(u^t - u^0) \cdot \nabla\psi \, dx \\ & = \int_\Omega \left[J_t \, a \circ T_t - a \right] \psi \, dx - \int_\Omega \left[A(t) - I \right] \nabla u^t \cdot \nabla\psi \, dx \\ & = \int_\Omega (J_t - 1) \, a \circ T_t \psi dx + \int_\Omega \left[a \circ T_t - a \right] \psi dx - \int_\Omega \left[A(t) - I \right] \nabla u^t \cdot \nabla\psi \, dx. \end{aligned}$$

Substitute $\psi = u^t - u^0$ to obtain the following estimate

$$\left\|\nabla(u^t - u^0)\right\|^2 \leq \|J_t - 1\|_C \, \|a \circ T_t\| \, \|u^t - u^0\| + \|a \circ T_t - a\| \, \|u^t - u^0\| + \|A(t) - I\|_C \, \|\nabla u^t\| \, \|\nabla(u^t - u^0)\|.$$

Since Ω is a bounded open Lipschitzian domain, there exists a constant such that $\|u^t - u^0\| \leq c(\Omega) \, \|\nabla(u^t - u^0)\|$ and

$$\left\|\nabla(u^t - u^0)\right\| \leq \|J_t - 1\|_C \|a \circ T_t\| c(\Omega) + \|a \circ T_t - a\| c(\Omega) + \|A(t) - I\|_C \|\nabla u^t\|.$$

But the right-hand side of this inequality goes to zero as t goes to zero. Therefore, $u^t \to u^0$ in $H_0^1(\Omega)$. Finally, going back to the last inequality and dividing by $t > 0$

$$\left\|\nabla\left(\frac{u^t - u^0}{t}\right)\right\| \leq \left\|\frac{J_t - 1}{t}\right\|_C \|a \circ T_t\| \, c(\Omega) + \left\|\frac{a \circ T_t - a}{t}\right\| \, c(\Omega) + \left\|\frac{A(t) - I}{t}\right\|_C \|\nabla u^t\|.$$

Since the right-hand side of the above inequality is bounded, $\nabla(u^t - u^0)/t$ is bounded. This means that $(u^t - u^0)/t$ is bounded in $H_0^1(\Omega)$, and, hence, $(u^t - u^0)/t$ is bounded in $L^2(\Omega)$.

As a result the $R(x^0, y^0)$ term is zero and the expression of the shape derivative is given by (42).

3.4.3 Topological Derivative

The pair $(u^0, p^0) \in H_0^1(\Omega) \cap H^2(\Omega) \times H_0^1(\Omega) \cap H^2(\Omega)$ is the solution of the same system (33) as in the previous examples. For $t \geq 0$, the state equation (14) has a unique solution in $H_0^1(\Omega) \cap H^2(\Omega)$. The expression of $d_t L(0, u^0, p^0)$ was given in (16). Therefore, (H1) and (H2") are verified. As for (H3") we need the expression

$$d_x L(t, \hat{\varphi}, \hat{\psi}; \varphi) = \int_\Omega \left(\hat{\varphi} - z\right) \varphi + \nabla\hat{\psi} \cdot \nabla\varphi \, dx \tag{47}$$

to check the limit of the term in Hypothesis (H3") as t goes to zero

$$\begin{aligned} R(t) &\stackrel{\text{def}}{=} \int_0^1 d_x L\left(t, u^0 + \theta(u^t - u^0), p^0; \frac{u^t - u^0}{t}\right) d\theta \\ &= \int_0^1 \int_\Omega \left(u^0 + \theta(u^t - u^0) - z\right) \frac{u^t - u^0}{t} + \nabla p^0 \cdot \nabla \left(\frac{u^t - u^0}{t}\right) dx\, d\theta \\ &= \int_\Omega \left(\frac{u^0 + u^t}{2} - z\right) \frac{u^t - u^0}{t} + \nabla p^0 \cdot \nabla \left(\frac{u^t - u^0}{t}\right) dx \\ &= \int_\Omega \left(\frac{u^0 - z + u^t - z}{2}\right) \frac{u^t - u^0}{t} + \nabla p^0 \cdot \nabla \left(\frac{u^t - u^0}{t}\right) dx. \end{aligned} \tag{48}$$

By substituting $\varphi = (u^t - u^0)/t$ in the adjoint equation for p^0

$$\int_\Omega (u^0 - z) \frac{u^t - u^0}{t} + \nabla p^0 \cdot \nabla \left(\frac{u^t - u^0}{t}\right) dx = 0, \tag{49}$$

we can simplify the expression for $R(t)$ as follows

$$\begin{aligned} R(t) &= \int_\Omega \left(\frac{u^0 - z + u^t - z}{2}\right) \frac{u^t - u^0}{t} - (u^0 - z) \frac{u^t - u^0}{t} dx \\ &= \int_\Omega \left(\frac{u^t - u^0}{2}\right) \frac{u^t - u^0}{t} dx = \frac{1}{2} \left\| \frac{u^t - u^0}{t^{1/2}} \right\|^2. \end{aligned} \tag{50}$$

To complete the proof, we have to show that $(u^t - u^0)/t^{1/2} \to 0$ in $L^2(\Omega)$-strong. By subtracting the state equation (14) at t from the one at 0

$$\int_\Omega \nabla \left(\frac{u^t - u^0}{t}\right) \cdot \nabla \psi \, dx = -(1 - \gamma) \frac{1}{t} \int_\Omega \chi_{E_r} \psi \, dx. \tag{51}$$

Substitute $\psi = u^t - u^0$ in the last equation and use the fact that, for a Lipschitzian domain Ω, there exists a constant $c(\Omega)$ such that $\|v\| \le c(\Omega) \|\nabla v\|$

$$\begin{aligned} \left\| \nabla \left(\frac{u^t - u^0}{t^{1/2}}\right) \right\|^2 &= -(1 - \gamma) \frac{1}{t} \int_{E_r} \chi_\Omega \, (u^t - u^0) \, dx \\ &\le |1 - \gamma| \left[\frac{1}{t} \int_{E_r} dx\right]^{1/2} \left\| \frac{u^t - u^0}{t^{1/2}} \right\| \\ &\le |1 - \gamma| \, H_d(E)^{1/2} \, c(\Omega) \left\| \nabla \left(\frac{u^t - u^0}{t^{1/2}}\right) \right\|. \end{aligned}$$

Finally, we get a bound on the norm of the gradient

$$\|\nabla((u^t - u^0)/t^{1/2})\| \leq |1-\gamma|\, H_d(E)^{1/2}\, c(\Omega). \tag{52}$$

Hence, there exists $w \in H_0^1(\Omega)$ and a sequence $\{t_n\}$ going to 0 such that

$$\frac{u^{t_n} - u^0}{{t_n}^{1/2}} \rightharpoonup w \text{ in } H_0^1(\Omega)\text{-weak} \quad \Rightarrow \quad \frac{u^{t_n} - u^0}{{t_n}^{1/2}} \to w \text{ in } L^2(\Omega)\text{-strong}.$$

Furthermore, we can show that $w = 0$. For $E = \{e\} \subset \Omega$, and $\psi \in \mathscr{D}(\Omega)$, from (51)

$$\underbrace{\int_\Omega \nabla\left(\frac{u^{t_n} - u^0}{{t_n}^{1/2}}\right)\cdot\nabla\psi\, dx}_{\to \int_\Omega \nabla w\cdot\nabla\psi\, dx} = -\underbrace{{t_n}^{1/2}}_{\to 0}(1-\gamma)\underbrace{\frac{1}{|B_{r_n}(e)|}\int_{B_{r_n}(e)}\chi_\Omega\,\psi\, dx}_{\to \chi_\Omega(e)\,\psi(e)=\psi(e)}$$

$$\Rightarrow\ \forall\psi \in H_0^1(\Omega), \quad \int_\Omega \nabla w\cdot\nabla\psi\, dx = 0 \quad \Rightarrow\ w = 0. \tag{53}$$

Since the limit is independent of the choice of the sequence, the limit exists as $t \to 0$

$$R(t) = \frac{1}{2}\frac{\|u^t - u^0\|^2}{t} \to \frac{1}{2}\|w\|^2 = 0.$$

All the hypotheses of Theorem 3.3 are now verified. Coming back to the t-derivative, since $p^0 \in H^2(\Omega)\cap H_0^1(\Omega)$, for $e \in \Omega$

$$dJ(\Omega;\delta_{\{e\}}) = d_t L(0,u^0,p^0) = (1-\gamma)\,\chi_\Omega(a)\,p^0(e) = (1-\gamma)\,p^0(e). \tag{54}$$

When E is a curve or a surface, we also have $w = 0$: from (51) with $\psi \in \mathscr{D}(\Omega)$

$$\underbrace{\int_\Omega \nabla\left(\frac{u^t - u^0}{t^{1/2}}\right)\cdot\nabla\psi\, dx}_{\to \int_\Omega \nabla w\cdot\nabla\psi\, dx} = -\underbrace{t^{1/2}}_{\to 0}(1-\gamma)\underbrace{\frac{1}{t}\int_{E_r}\chi_\Omega\,\psi\, dx}_{\to \int_E \chi_\Omega\,\psi\, dH_d}$$

$$\Rightarrow\ \forall\psi \in H_0^1(\Omega), \quad \int_\Omega \nabla w\cdot\nabla\psi\, dx = 0 \quad \Rightarrow\ w = 0. \tag{55}$$

Therefore, the limit exists as $t \to 0$

$$R(t) = \frac{1}{2}\frac{\|u^t - u^0\|^2}{t} \to \frac{1}{2}\|w\|^2 = 0$$

and all the hypotheses of Theorem 3.3 are verified. Coming back to the t-derivative, since $p^0 \in H^2(\Omega) \cap H_0^1(\Omega)$,

$$dJ(\Omega; \delta_E) = d_t L(0, u^0, p^0) = (1-\gamma) \int_E \chi_\Omega \, p^0 \, dH_d = (1-\gamma) \int_E p^0 \, dH_d, \tag{56}$$

where (u^0, p^0) is solution of the coupled system (33).

4 Minimax Theorems in the Mutivalued Case

We give a general theorem for the existence and expressions of $dg(0)$ in the multivalued case where only a right-hand side derivative of g is expected. The Corollary can be applied to PDE problems with non-homogeneous Dirichlet boundary conditions. The trick introduced in 1991 [10] to get around the requirement that the spaces X and Y be fixed, was to use extensions of (x^t, p^t) from Ω_t to $\mathbb{R}^N$ to work in $H^2(\mathbb{R}^N) \times H^2(\mathbb{R}^N)$. Such extensions to $\mathbb{R}^N$ are not unique, but their restrictions to Ω_t are (see also [12, sec. 6.3, Chap. 10. pp. 564–570]).

Theorem 4.1 *Consider the Lagrangian*

$$(t, x, y) \mapsto G(t, x, y) : [0, \tau] \times X \times Y \to \mathbb{R}, \quad \tau > 0,$$

where X and Y are vector spaces, the function $y \mapsto G(t, x, y)$ is affine. Let (H0) and the following hypotheses be satisfied:

(H1) *for all $t \in [0, \tau]$, $X(t) \neq \varnothing$, $g(t)$ is finite, and for all $x \in X(0)$, $Y(0, x) \neq \varnothing$;*
(H2) *for each x in $X(0)$ and $y \in Y(0, x)$, $d_t G(0, x, y)$ exists;*
(H3) *for each $x \in X(0)$, there exists a function $p \mapsto R(x, p) : Y(0, x) \to \mathbb{R}$ such that for each sequence $t_n \to 0$, $0 < t_n \leq \tau$,*

(i) *there exists $x^0 \in X(0)$ such that for each $p \in Y(0, x^0)$, there exist a subsequence $\{t_{n_k}\}$ of $\{t_n\}$, $x^{t_{n_k}} \in X(t_{n_k})$ such that*

$$\liminf_{k \to \infty} \int_0^1 d_x G \left(t_{n_k}, x^0 + \theta(x^{t_{n_k}} - x^0), p^0; \frac{x^{t_{n_k}} - x^0}{t_{n_k}} \right) d\theta \geq R(x^0, p^0), \tag{57}$$

(ii) *for each $x^0 \in X(0)$, there exist $p^0 \in Y(0, x^0)$, a subsequence $\{t_{n_k}\}$ of $\{t_n\}$, $x^{t_{n_k}} \in X(t_{n_k})$ such that*

$$\limsup_{k \to \infty} \int_0^1 d_x G \left(t_{n_k}, x^0 + \theta(x^{t_{n_k}} - x^0), p^0; \frac{x^{t_{n_k}} - x^0}{t_{n_k}} \right) d\theta \leq R(x^0, p^0). \tag{58}$$

Then, $dg(0)$ exists and there exist $x^0 \in X(0)$ and $y^0 \in Y(0, x^0)$ such that

$$dg(0) = d_t G(0, x^0, y^0) + R(x^0, y^0) = \sup_{y \in Y(0,x^0)} d_t G(0, x^0, y) + R(x^0, y) = \inf_{x \in X(0)} \sup_{y \in Y(0,x)} d_t G(0, x, y) + R(x, y). \tag{59}$$

Proof

(i) Since $g(t)$ is finite, $E(t) \neq \varnothing$. Moreover, since $X(t) \neq \varnothing$, $X(t)$ is the set of minimizers of $G(t, x, 0)$ over $E(t)$. The differential quotient can be written as follows: for all $x \in X(0)$, $p \in Y(0, x)$,

$$\begin{aligned}\frac{g(t) - g(0)}{t} &= \frac{G(t, x^t, p) - G(0, x, p)}{t} \\ &= \frac{G(t, x^t, p) - G(t, x, p)}{t} + \frac{G(t, x, p) - G(0, x, p)}{t} \\ &= \int_0^1 d_x G\left(t, x + \theta(x^t - x), p; \frac{x^t - x}{t}\right) d\theta \\ &\quad + \frac{G(t, x, p) - G(0, x, p)}{t}.\end{aligned}$$

At this juncture, introduce the liminf and limsup of the differential quotient

$$\underline{d}g(0) \stackrel{\text{def}}{=} \liminf_{t \searrow 0} \frac{g(t) - g(0)}{t}, \quad \overline{d}g(0) \stackrel{\text{def}}{=} \limsup_{t \searrow 0} \frac{g(t) - g(0)}{t}.$$

We show that they exist and are equal.

(ii) There exists a sequence $t_n \to 0$, $0 < t_n \leq \tau$, such that $(g(t_n) - g(0))/t_n \to \underline{d}g(0)$. By Hypothesis (H3) (i), there exists $x^0 \in X(0)$ such that for all $p^0 \in Y(0, x^0)$, there exists a subsequence $\{t_{n_k}\}$ of $\{t_n\}$, there exists $x^{t_{n_k}} \in X(t_{n_k})$ such that

$$\begin{aligned}\underline{d}g(0) &\geq \liminf_{k \to \infty} \int_0^1 d_x G\left(t_{n_k}, x^0 + \theta(x^{t_{n_k}} - x^0), p^0; \frac{x^{t_{n_k}} - x^0}{t_{n_k}}\right) d\theta \\ &\quad + \lim_{k \to \infty} \frac{G(t_{n_k}, x^0, p^0) - G(0, x^0, p^0)}{t_{n_k}} \\ &\geq \liminf_{k \to \infty} \int_0^1 d_x G\left(t_{n_k}, x^0 + \theta(x^{t_{n_k}} - x^0), p^0; \frac{x^{t_{n_k}} - x^0}{t_{n_k}}\right) d\theta \\ &\quad + d_t G(0, x^0, p^0) \\ &\geq R(x^0, p^0) + d_t G(0, x^0, p^0).\end{aligned}$$

Therefore, there exists $x^0 \in X(0)$ such that for all $p^0 \in Y(0, x^0)$

$$\underline{d}g(0) \geq d_t G(0, x^0, p^0) + R(x^0, p^0).$$

So, we can take the sup with respect to $p^0 \in Y(0, x^0)$

$$\exists x^0 \in X(0) \text{ such that } \quad \underline{d}g(0) \geq \sup_{y \in Y(0,x^0)} d_t G(0, x^0, y) + R(x^0, y)$$

and we obtain our first estimate

$$\underline{d}g(0) \geq \sup_{y \in Y(0,x^0)} d_t G(0, x^0, y) + R(x^0, y) \geq \inf_{x \in X(0)} \sup_{y \in Y(0,x)} d_t G(0, x, y) + R(x, y). \tag{60}$$

(iii) There exists a sequence $t_n \to 0$, $0 < t_n \leq \tau$, such that $\lim_{n\to\infty}(g(t_n) - g(0))/t_n = \overline{d}g(0)$. By Hypothesis (H3) (ii), for all $x^0 \in X(0)$ there exist $p^0 \in Y(0, x^0)$, a subsequence $\{t_{n_k}\}$ of $\{t_n\}$, $x^{t_{n_k}} \in X(t_{n_k})$ such that

$$\begin{aligned}\overline{d}g(0) \leq & \limsup_{k\to\infty} \int_0^1 d_x G\left(t_{n_k}, x^0 + \theta(x^{t_{n_k}} - x^0), p^0; \frac{x^{t_{n_k}} - x^0}{t_{n_k}}\right) d\theta \\ & + \lim_{k\to\infty} \frac{G(t_{n_k}, x^0, p^0) - G(0, x^0, p^0)}{t_{n_k}} \leq R(x^0, p^0) + d_t G(0, x^0, y^0).\end{aligned}$$

Therefore, for all $x^0 \in X(0)$ there exists $p^0 \in Y(0, x^0)$ such that $\overline{d}g(0) \leq R(x^0, p^0) + d_t G(0, x^0, p^0)$. Since for all $x^0 \in X(0)$ there exists $p^0 \in Y(0, x^0)$ such that

$$\begin{aligned}&\overline{d}g(0) \leq d_t G(0, x^0, p^0) + R(x^0, p^0) \\ \Rightarrow\ &\overline{d}g(0) \leq d_t G(0, x^0, p^0) + R(x^0, p^0) \leq \sup_{y \in Y(0,x^0)} d_t G(0, x^0, y) + R(x^0, y) \\ \Rightarrow\ &\forall x^0 \in X(0), \quad \overline{d}g(0) \leq \sup_{y \in Y(0,x^0)} d_t G(0, x^0, y) + R(x^0, y)\end{aligned}$$

and we can take the infimum of the right-hand side over all $x^0 \in X(0)$,

$$\overline{d}g(0) \leq \inf_{x^0 \in X(0)} \sup_{y \in Y(0,x^0)} d_t G(0, x^0, y) + R(x^0, y). \tag{61}$$

(iv) Combining (61) and (60), there exists $\hat{x}^0 \in X(0)$ such that

$$\begin{aligned}\underline{d}g(0) \geq & \sup_{y \in Y(0,\hat{x}^0)} d_t G(0, \hat{x}^0, y) + R(0, \hat{x}^0, y) \\ \geq & \inf_{x \in X(0)} \sup_{y \in Y(0,x)} d_t G(0, x, y) + R(x, y) \geq \overline{d}g(0) \geq \underline{d}g(0).\end{aligned}$$

Therefore, $dg(0)$ exists and there exists $\hat{x}^0 \in X(0)$ such that

$$dg(0) = \sup_{y \in Y(0,\hat{x}^0)} d_t G(0, \hat{x}^0, y) + R(\hat{x}^0, y) = \inf_{x \in X(0)} \sup_{y \in Y(0,x)} d_t G(0, x, y) + R(x, y). \tag{62}$$

But we can get more. From part (iii), we have shown that

$$\exists \hat{x}^0 \in X(0),\ \forall y \in Y(0, \hat{x}^0), \quad \underline{d}g(0) \geq d_t G(0, \hat{x}^0, y) + R(\hat{x}^0, y). \tag{63}$$

From part (iii), we have shown that

$$\forall x^0 \in X(0),\ \exists y^0 \in Y(0, x^0), \quad \overline{d}g(0) \leq d_t G(0, x^0, y^0) + R(x^0, y^0).$$

In particular, for $\hat{x}^0$, there exists $\hat{y}^0 \in Y(0, \hat{x}^0)$ such that

$$\overline{d}g(0) \leq d_t G(0, \hat{x}^0, \hat{y}^0) + R(\hat{x}^0, \hat{y}^0).$$

Taking $y = \hat{y}^0$ in (63), $\underline{d}g(0) \geq d_t G(0, \hat{x}^0, \hat{y}^0) + R(\hat{x}^0, \hat{y}^0)$ and

$$\underline{d}g(0) \geq d_t G(0, \hat{x}^0, \hat{y}^0) + R(\hat{x}^0, \hat{y}^0) \geq \overline{d}g(0) \quad \Rightarrow dg(0) = d_t G(0, \hat{x}^0, \hat{y}^0) + R(\hat{x}^0, \hat{y}^0)$$

and the conclusion of the theorem.

□

This specialization of the theorem was used to compute the shape derivative of an objective function constrained by a partial differential equation with non-homogeneous Dirichlet boundary conditions in [12, sec. 6, pp. 562–570].

Corollary 4.1 *Consider the Lagrangian*

$$(t, x, y) \mapsto G(t, x, y) : [0, \tau] \times X \times Y \to \mathbb{R}, \quad \tau > 0,$$

where X and Y are vector spaces, the function $y \mapsto G(t, x, y)$ is affine. Let (H0) and the following hypotheses be satisfied:

(H1) *for all t in $[0, \tau]$, $X(t) \neq \varnothing$ and $g(t)$ is finite, and for each $x \in X(0)$, $Y(0, x) \neq \varnothing$;*
(H2) *for all x in $X(0)$ and $p \in Y(0, x)$, $d_t G(0, x, p)$ exists;*
(H3") *there exist $x^0 \in X(0)$ and $p^0 \in Y(0, x^0)$ such that the following limit exists*

$$R(x^0, p^0) \overset{def}{=} \lim_{t \searrow 0} \int_0^1 d_x G\left(t, x^0 + \theta(x^t - x^0), p^0; \frac{x^t - x^0}{t}\right) d\theta. \tag{64}$$

Then, $dg(0)$ exists and there exist $x^0 \in X(0)$ and $p^0 \in Y(0, x^0)$ such that $dg(0) = d_t G(0, x^0, p^0) + R(x^0, p^0)$.

References

1. Ambrosio, L., Fusco, N., Pallara, D.: Functions of Bounded Variation and Free Discontinuity Problems. The Clarendon Press, Oxford University Press, New York (2000)
2. Aubin, J.-P., Frankowska, H.: Set-Valued Analysis. Birkhäuser, Boston (1990)
3. Correa, R., Seeger, A.: Directional derivatives of a minimax function. Nonlinear Anal. Theory Methods Appl. **9**, 13–22 (1985)
4. Delfour M.C.: Metrics spaces of shapes and geometries from set parametrized functions. In: Pratelli, A., Leugering, G. (eds.) New Trends in Shape Optimization. International Series of Numerical Mathematics, vol. 166, pp. 57–101. Birkhäuser, Basel (2015)
5. Delfour, M.C.: Differentials and semidifferentials for metric spaces of shapes and geometries. In: Bociu, L., Desideri, J.A., Habbal, A. (eds.) System Modeling and Optimization. Proc. 27th IFIP TC7 Conference, CSMO 2015, Sophia-Antipolis, France. AICT Series, pp. 230–239. Springer, Berlin (2017)
6. Delfour, M.C.: Topological derivative: a semidifferential via the Minkowski content. J. Convex Anal. **25**(3), 957–982 (2018)
7. Delfour, M.C., Sturm, K.: Minimax differentiability via the averaged adjoint for control/shape sensitivity. In: Proc. of the 2nd IFAC Workshop on Control of Systems Governed by Partial Differential Equations, IFAC-PaperOnLine, vol. 49-8, pp. 142–149 (2016)
8. Delfour, M.C., Sturm, K.: Parametric semidifferentiability of minimax of Lagrangians: averaged adjoint approach. J. Convex Anal. **24**(4), 1117–1142 (2017)
9. Delfour, M.C., Zolésio, J.P.: Shape sensitivity analysis via min max differentiability. SIAM J. Control Optim. **26**, 834–862 (1988)
10. Delfour, M.C., Zolésio, J.P.: Velocity method and Lagrangian formulation for the computation of the shape Hessian. SIAM J. Control Optim. **29**(6), 1414–1442 (1991)
11. Delfour, M.C., Zolésio, J.P.: Shapes and Geometries: Analysis, Differential Calculus and Optimization. SIAM Series on Advances in Design and Control. Society for Industrial and Applied Mathematics, Philadelphia (2001)
12. Delfour, M.C., Zolésio, J.P.: Shapes and Geometries: Metrics, Analysis, Differential Calculus, and Optimization. SIAM Series on Advances in Design and Control, 2nd edn. Society for Industrial and Applied Mathematics, Philadelphia (2011)
13. Federer, H.: Curvature measures. Trans. Am. Math. Soc. **93**, 418–419 (1959)
14. Federer, H.: Geometric Measure Theory. Springer, Berlin (1969)
15. Lewiński, T., Sokołowski, J.: Energy change due to the appearance of cavities in elastic solids. Int. J. Solids Struct. **40**(7), 1765–1803
16. Micheletti, A.M.: Metrica per famiglie di domini limitati e proprietà generiche degli autovalori. Ann. Scuola Norm. Sup. Pisa (3) **26**, 683–694 (1972)
17. Novotny, A.A., Sokołowski, J.: Topological Derivatives in Shape Optimization. Interaction of Mechanics and Mathematics. Springer, Heidelberg (2013)
18. Sokolowski, J., Zochowski, A.: On the topological derivative in shape optimization. SIAM J. Control Optim. **37**(4), 1251–1272 (1999)
19. Sturm, K.: On shape optimization with non-linear partial differential equations. Doctoral thesis, Technische Universität of Berlin (2014)
20. Sturm, K.: Minimax Lagrangian approach to the differentiability of non-linear PDE constrained shape functions without saddle point assumption. SIAM J. Control Optim. **53**(4), 2017–2039 (2015)
21. Zolésio, J.P.: Un résultat d'existence de vitesse convergente dans des problèmes d'identification de domaine. C. R. Acad. Sci. Paris Sér. A-B (11) **283**, A855–A858 (1976)
22. Zolésio, J.P.: An optimal design procedure for optimal control support. In Auslender, A. (ed.) Convex Analysis and Its Applications (Proc. Conf., Muret-le-Quaire, 1976). Lecture Notes in Economics and Mathematical Systems, vol. 144, pp. 207–219. Springer, Berlin (1977)
23. Zolésio, J.P.: Identification de domaines par déformation. Thèse de doctorat d'état, Université de Nice (1979)

Minimum Energy Estimation Applied to the Lorenz Attractor

Arthur J. Krener

Abstract Minimum Energy Estimation is a way of filtering the state of a nonlinear system from partial and inexact measurements. It is a generalization of Gauss' method of least squares. Its application to filtering of control systems goes back at least to Mortensen who called it Maximum Likelyhood Estimation. For linear, Gaussian systems it reduces to maximum likelihood estimation (aka Kalman Filtering) but this is not true for nonlinear systems. We prefer the name Minimum Energy Estimation (MEE) that was introduced by Hijab. Both Mortensen and Hijab dealt with systems in continuous time, we extend their methods to discrete time systems and show how Taylor polynomial techniques can lessen the computational burden. The degree one version is equivalent to the Extended Kalman Filter in Information form. We apply this and the degree three version to problem of estimating the state of the three dimensional Lorenz Attractor from a one dimensional measurement.

Keywords Infinite horizon stochastic optimal control · Finite horizon stochastic optimal control · Stochastic Hamilton–Jacobi–Bellman equations · Stochastic algebraic Riccati equations · Stochastic differential Riccati equations · Stochastic linear quadratic regulator

1 Introduction

We consider the problem of estimating the state of a discrete time nonlinear control system from noisy knowledge of the dynamics, noisy and partial past measurements of the state, noisy knowledge of the initial condition, and exact knowledge of the control input.

An algorithm that does the state estimation is called a filter if one assumes that the noises are stochastic and the stochastic reasoning is used to derive the algorithm.

A. J. Krener (✉)
Naval Postgraduate School, Monterey, CA, USA
e-mail: ajkrener@nps.edu

M. Falcone et al. (eds.), *Numerical Methods for Optimal Control Problems*,
Springer INdAM Series 29, https://doi.org/10.1007/978-3-030-01959-4_8

The foremost example of this is the discrete time Kalman filter for systems where the dynamics and observations are linear and the noises are Gaussian, see [3]. The Extended Kalman Filter [3] and its variant the Unscented Kalman Filter [5] are widely used for nonlinear systems. They assume that the nonlinear system can be locally approximated by a linear system and the noises and errors can be locally approximated by Gaussian random variables.

If deterministic reasoning is used to derive the estimation algorithm then the result is usually called an observer. An observer for an autonomous linear discrete time system can be constructed by using output injection to place the poles of the linear error dynamics inside the unit disk [2]. If the system is nonlinear then under certain conditions a nonlinear output injection and a nonlinear change of state coordinates can linearize the error dynamics and place its poles inside the unit disk [6, 10].

Most filters and observers consist of a copy of the plant dynamics, including the known control terms, modified by a gain times the so-called innovation. The innovation is the difference between the current measurement and what the filter or observer thinks the measurement should be based on its current state estimate.

Another approach is to use high gain to construct an observer that is approximately a multiple differentiator [7]. The drawback of this approach is that the size of the gains grows exponentially with the number of differentiations necessary to determine the full state and so a high gain observer can be very sensitive to measurement noise. High gain can be very useful when the current state estimate is far from the true state because then the innovation has a high signal to noise ratio. But high gain can be very detrimental when the current state estimate is close to the true state for then the innovation has a low signal to noise ratio and the high gain accentuates the noise.

Interestingly one of the first examples of an observer was developed by Gauss to predict the orbit of the dwarf planet Ceres in 1801. Gauss used the method of least squares to make his prediction. In 1968 Mortensen used a similar method of least squares to develop his Maximum Likelyhood estimator [11] and this method was extended by Hijab [4] who called it Minimum Energy Estimation.

Both Mortensen and Hijab worked with continuous time systems. In [9] we extended their method to discrete time systems and used a Taylor polynomial approach to simplify the necessary calculations. We shall show if the Taylor polynomials are stopped at degree one then our method reduces to an Extended Kalman Filter but in the form of an Information Filter [1]. The purpose of this paper is to show by example that going to higher degree terms we can improve the accuracy of the estimates. The example that we choose is the Lorenz Attractor. Because this a chaotic system the estimation problem is challenging. The degree one minimum energy estimator (EKF) does a reasonable job but the degree three estimator does substantially better.

2 Minimum Energy Estimation in Discrete Time

Consider a discrete time nonlinear system

$$\begin{aligned} x^+ &= f(t,x,u) \\ y &= h(t,x,u) \\ x(0) &= x^0 \end{aligned} \tag{1}$$

where the state $x(t) \in \mathbb{R}^{n\times 1}$, the control $u(t) \in \mathbb{R}^{m\times 1}$, the measurement $y(t) \in \mathbb{R}^{p\times 1}$ and $x^+(t) = x(t+1)$. We assume that f and h are sufficiently smooth to have Taylor polynomials of the desired degree. The problem is to estimate $x(t)$ from $u(s), 0 \le s \le t-1$, from $y(s), 1 \le s \le t-1$ and from some inexact knowledge about x^0. We denote this estimate by $\hat{x}(t|t-1)$. We also consider estimating $x(t)$ using the additional measurement $y(t)$, we denote this estimate $\hat{x}(t|t)$.

The standard approach is add noises to the model, a driving noise $v(t) \in \mathbb{R}^{k\times 1}$, an observation noise $w(t) \in \mathbb{R}^{p\times 1}$ and an initial condition noise $\bar{x}^0 \in \mathbb{R}^{n\times 1}$ to obtain

$$\begin{aligned} x^+ &= f(t,x,u) + g(t,x,u)v \\ y &= h(t,x,u) + w \\ x(0) &= \hat{x}^0 + \bar{x}^0 \end{aligned} \tag{2}$$

where $\hat{x}^0$ is our estimate of $x(0)$ based on prior information.

The stochastic version of this approach is to assume that $v(t)$ and $w(t)$ are independent, white Gaussian noise processes and $\bar{x}^0$ is an independent Gaussian random vector of mean zero and known covariance. Then the conditional density of the state given the past controls $u(s), 0 \le s \le t-1$ and past measurements $y(s), 1 \le s \le t-1$ satisfies an integral-difference equation that is very difficult to solve. About the only time it can be solved is when the system is linear for then the conditional density is Gaussian and its mean and covariance can be computed by a Kalman filter.

The minimum energy approach is to assume the noises are deterministic but unknown. One finds the noise triple $\bar{z}^0, v(\cdot), w(\cdot)$ that minimizes

$$\min_{\bar{z}^0,v,w} \frac{1}{2}\left\{ \alpha^t \|\bar{z}^0\|^2_{P0} + \sum_{s=0}^{t-1} \alpha^{t-s} \|v(s)\|^2_{Q(s)} + \sum_{s=1}^{\tau} \alpha^{t-s} \|w(s)\|^2_{R(s)} \right\} \tag{3}$$

subject to

$$\begin{aligned} z^+ &= f(s,z(s),u(s)) + g(s,z(s),u(s))v(s) \\ w(s) &= y(s) - h(s,z(s),u(s)) \\ z(0) &= \hat{x}^0 + \bar{z}^0 \end{aligned}$$

where $\tau = t-1$ or $\tau = t$, $P^0 \geq 0,\ Q(s) > 0,\ R(s) > 0$ and

$$\|\bar{z}^0\|^2_{P^0} = (\bar{z}^0)' P^0 \bar{z}^0$$
$$\|v(s)\|^2_{Q(s)} = v'(s) Q(s) v(s)$$
$$\|w(s)\|^2_{R(s)} = w'(s) R(s) w(s)$$

The control $u(s), 0 \leq s \leq t-1$ and the observation $y(s), 1 \leq s \leq \tau$ are the actual control and measurement sequences and $0 < \alpha \leq 1$ is a forgetting factor. The smaller the forgetting factor α the more weight is placed on the most recent observations. Assuming that $z(s)$ is the minimizing state trajectory then the Minimum Energy Estimation (MEE) is $\hat{x}(t|\tau) = z(t)$.

The heuristic behind MEE is that (3) (or some variant) is the energy in the noise triple. Nature in her attempt to confuse us chooses the three noises in a parsimonious fashion consistent with the past controls and past measurements. If the system is linear and $\alpha = 1$ then MEE filtering is identical to Kalman filtering provided that the weights of the norms in (3) are the inverses of covariances of the noises in the stochastic version. Then the MEE is a Kalman Filter in Innovation form [1].

It has been shown that under suitable conditions the continuous time MEE is globally convergent [8]. that is, the MEE estimate converges to the state of the model (1) regardless of the initial conditions of the model and the filter assuming the driving and observation noises are zero. We conjecture that a similar result holds in discrete time MEE.

Since we know $u(t)$ we can simplify notation by redefining

$$f(t,x) = f(t,x,u(t))$$
$$g(t,x) = g(t,x,u(t))$$
$$h(t,x) = h(t,x,u(t))$$

The discrete time version $\pi(x,t|\tau)$ of the Mortensen function is defined by a family of optimization problems

$$\pi(x,t|\tau) = \min_{\bar{z}^0,v,w} \frac{1}{2} \left\{ \alpha^t \|\bar{z}^0\|^2_{P^0} + \sum_{s=0}^{t-1} \alpha^{t-1-s} \|v(s)\|^2_{Q(s)} + \sum_{s=1}^{\tau} \alpha^{t-s} \|w(s)\|^2_{R(s)} \right\}$$

subject to

$$z^+ = f(s,z) + g(s,z)v$$
$$w = y - h(s,z)$$
$$z(0) = \hat{x}^0 + \bar{z}^0$$
$$z(t) = x$$

If $x \mapsto f(t,x)$ is invertible for every $t \geq 0$ then there is at least one noise triple $\bar{z}^0, v(\cdot), w(\cdot)$ that satisfies the constraints. That triple is found by setting $v(t) = 0$ and mapping the dynamics backward from $z(t) = x$ to find $w(\cdot)$ and $\bar{z}^0$. If the terminal constraint $z(t) = x$ is not achievable by any noise triple then we set $\pi(x,t|\tau) = \infty$.

One can take a dynamic programming approach to computing the Mortensen function. The prediction step from $(t|t)$ to $(t+1|t)$ is accomplished by solving a family of optimization problems indexed by x,

$$\pi(x,t+1|\tau) = \min_{z,v} \left(\alpha \pi(z,t|\tau) + \frac{1}{2} \|v\|^2_{Q(t)} \right) \tag{4}$$

where the minimum is over all z, v that satisfy the constraint

$$x = f(s,z) + g(s,z)v \tag{5}$$

The $(t+1|t)$ minimum energy estimate is then

$$\hat{x}(t+1|t) = \text{argmin}_x \pi(x,t+1|t) \tag{6}$$

The assimilation step from $(t+1|t)$ to $(t+1|t+1)$ is

$$\pi(x,t+1|t+1) = \pi(x,t+1|t) + \frac{1}{2} \|y(t+1) - h(t+1,x)\|^2_{R(t+1)} \tag{7}$$

The $(t+1|t+1)$ minimum energy estimate is then

$$\hat{x}(t+1|t+1) = \text{argmin}_x \pi(x,t+1|t+1) \tag{8}$$

The estimation algorithm proceeds as follows

1. Assume $\pi(x,0|0)$ and $\hat{x}(0|0) = \text{argmin}_x \pi(x,0|0)$ are known.
2. Given $\pi(x,t|t)$ compute $\pi(x,t+1|t)$ by solving (4) subject to (5).
3. Solve (6) to find $\hat{x}(t+1|t)$.
4. Solve (7) to find $\pi(x,t+1|t+1)$.
5. Solve (8) to get $\hat{x}(t+1|t+1)$.
6. Increment t by one and go to Step 2.

Steps 3–5 are relatively straightforward but Step 2 is difficult so in the next section we take a Taylor polynomial approach to it.

3 Taylor Polynomial Approach

The Taylor polynomial approach to approximating $\pi(x, t+1|t)$ starts by assuming that the Taylor polynomial of $\pi(x, t|t)$ around $\hat{x}(t|t)$ of degree $d+1$ is approximately known,

$$\pi^{[0:d+1]}(x, t|t) = \pi^{[0]}(t|t) + \frac{1}{2}(\bar{x}(t|t))' P(t|t)\bar{x}(t|t) + \pi^{[3]}(\bar{x}(t|t), t|t) + \pi^{[4]}(\bar{x}(t|t), t|t) + \cdots + \pi^{[d+1]}(\bar{x}(t|t), t|t)$$

where $\bar{x}(t|t) = x - \hat{x}(t|t)$ and $^{[k]}$ denotes a homogeneous polynomial term of degree k. We assume that $\pi^{[0:d+1]}(x, t|t) \geq 0$ and define

$$\hat{x}(t|t) = \text{argmin}_x \pi^{[0:d+1]}(x, t|t) \tag{9}$$

We are assuming that a unique minimum exists, but an interesting question for future research is what happens otherwise.

From (4) and (9) it is not hard to see that minimum of $\pi^{[0:d+1]}(x, t+1|t)$ occurs at $z = \hat{x}(t|t)$, $v = 0$ and

$$\hat{x}(t+1|t) = f(t, \hat{x}(t|t)) \tag{10}$$

so we have completed Step 3 before Step 2.

To complete Step 2 we compute an approximation of the Taylor polynomial of $\pi(x, t+1|t)$ around the point $\hat{x}(t+1|t)$,

$$\pi^{[0:d+1]}(x, t+1|t) = \pi^{[0]}(t+1|t)\frac{1}{2}(\bar{x}(t+1|t))' P(t|t)\bar{x}(t+1|t) + \pi^{[3]}(\bar{x}(t+1|t), t|t) + \pi^{[4]}(\bar{x}(t+1|t), t|t) + \cdots + \pi^{[d+1]}(\bar{x}(t+1|t), t|t)$$

where $\bar{x}(t+1|t) = x - \hat{x}(t+1|t)$.

Step 2 is a family of constrained minimization problems indexed by x, so we add the constraint (5) to the criterion (4) using a Lagrange multiplier λ to get a family of unconstrained problems,

$$\min_{z,v,\lambda} \left\{ \alpha\pi(z, t|t) + \frac{1}{2}\|v\|^2_{Q(t)} + \lambda'(x)\left(x - f(t, z) - g(t, z)v\right) \right\} \tag{11}$$

The three minimization variables are functions of x and t, to simplify notation we only show their dependence on x, $z = z(x)$, $v = v(x)$, $\lambda = \lambda(x)$.

To get the first order necessary conditions we set to zero the partials of (11) with respect to z, v, λ. Setting to zero the partial of (11) with respect to λ yields the constraint (5).

Setting to zero the partial of (11) with respect to v yields

$$0 = v'(x)Q(t) - \lambda'(x)g(t,z)$$

Because $Q(t)$ is assumed to be invertible this can be solved for $v(x)$ as a function of $z(x),\ \lambda(x)$

$$v(x) = Q^{-1}(t)g'(t,z(x))\lambda(x) \tag{12}$$

Then the constraint (5) becomes

$$x = f(t,z(x)) + g(t,z(x))Q^{-1}(t)g'(t,z(x))\lambda(x) \tag{13}$$

Setting to zero the partial of (11) with respect to z yields

$$0 = \alpha\frac{\partial\pi}{\partial z}(z(x),t|t) - \lambda'(x)\left(\frac{\partial f}{\partial z}(t,z(x)) + \frac{\partial}{\partial z}\left(g(t,z(x))v(x)\right)\right)$$

Because of (12) this becomes

$$\begin{aligned} 0 = \alpha\frac{\partial\pi}{\partial z}(z(x),t|t) - \lambda'(x)\frac{\partial f}{\partial z}(t,z(x)) \\ -\frac{\partial}{\partial z}\left(\lambda'(x)g(t,z(x))Q^{-1}(t)g'(t,z(x))\lambda(x)\right) \end{aligned} \tag{14}$$

At time t for each state x we must solve the two nonlinear equations (13), (14) in the two unknowns $z(x),\ \lambda(x)$. Then (12) yields $v(x)$ and we can compute $\pi(x,t+1|t)$ and $\hat{x}(t+1|t)$ from

$$\pi(x,t+1|t) = \alpha\pi(z(x),t|t) + \frac{1}{2}\|v(x)\|^2_{Q(t)} \tag{15}$$

$$\hat{x}(t+1|t) = \text{argmin}_x\,\pi(x,t+1|t) \tag{16}$$

Solving these equations is a daunting task so we turn to a Taylor polynomial approach. For simplicity of exposition we assume $g(t,z) = G(t)$. The general case is notationally more complicated but it follows in a similar fashion. To simplify notation let

$$\bar{Q}(t) = G(t)Q^{-1}(t)G'(t)$$

We know that

$$z(\hat{x}(t+1|t)) = \hat{x}(t|t), \qquad v(\hat{x}(t+1|t)) = 0, \qquad \lambda(\hat{x}(t+1|t)) = 0$$

Define $\bar{z} = z - \hat{x}(t|t)$. We expand $\bar{z}(x),\ v(x),\ \lambda(x)$ in Taylor polynomials of degree d in $\bar{x}(t+1|t) = x - \hat{x}(t+1|t)$.

$$\bar{z} = Z\bar{x}(t+1|t) + z^{[2]}(\bar{x}(t+1|t)) + z^{[3]}(\bar{x}(t+1|t)) + \cdots + z^{[d]}(\bar{x}(t+1|t))$$
$$v = V\bar{x}(t+1|t) + v^{[2]}(\bar{x}(t+1|t)) + v^{[3]}(\bar{x}(t+1|t)) + \cdots + v^{[d]}(\bar{x}(t+1|t))$$
$$\lambda = \Lambda\bar{x}(t+1|t) + \lambda^{[2]}(\bar{x}(t+1|t)) + \lambda^{[3]}(\bar{x}(t+1|t)) + \cdots + \lambda^{[d]}(\bar{x}(t+1|t))$$

We also expand $f(t,z)$ and $\pi(z,t|t)$ in Taylor polynomials in $\bar{z} = z - \hat{x}(t|t)$,

$$f^{[0:d]}(t,z) = f^{[0]}(t) + F(t)\bar{z} + f^{[2]}(t,\bar{z}) + \cdots + f^{[d]}(t,\bar{z})$$
$$\pi^{[0:d+1]}(z,t|t) = \pi^{[0]}(t|t) + \frac{1}{2}\bar{z}'P(t|t)\bar{z} + \pi^{[3]}(\bar{z},t|t) + \cdots + \pi^{[d+1]}(\bar{z},t|t)$$

Then the constraint (13) becomes

$$\bar{x} = f^{[0:d]}(t,z) - f^{[0]}(t) + \bar{Q}(t)\lambda(x) \tag{17}$$

and (14) becomes

$$0 = \alpha\frac{\partial\pi^{[0:d+1]}}{\partial z}(z,t|t) - \lambda'(x)\frac{\partial f^{[0:d]}}{\partial z}(t,z) \tag{18}$$

We collect from these equations the terms linear in $\bar{x}$ to obtain

$$\bar{x} = F(t)Z(t)\bar{x} + \bar{Q}(t)\Lambda(t)\bar{x}$$
$$0 = \alpha P(t|t)Z(t)\bar{x} - F'(t)\Lambda(t)\bar{x}$$

These equations must hold for any $\bar{x}$ so $Z(t),\ \Lambda(t)$ satisfy

$$\begin{bmatrix} I \\ 0 \end{bmatrix} = \mathscr{H}(t)\begin{bmatrix} Z(t) \\ \Lambda(t) \end{bmatrix} \tag{19}$$

where

$$\mathscr{H}(t) = \begin{bmatrix} F(t) & \bar{Q}(t) \\ \alpha P(t|t) & -F'(t) \end{bmatrix} \tag{20}$$

The solvability of these equations was discussed in [9] where this theorem is proven.

Theorem *Suppose that $P(0|0)$ is positive definite and at each $t \geq 0$, the pair $F(t), G(t)$ is stabilizable in the continuous time sense then $\mathscr{H}(t)$ given by (20) is invertible for each $t \geq 0$.*

Assuming $\mathscr{H}(t)$ is invertible, we collect the second order terms in $\bar{x}$ from (17) and (18),

$$\mathscr{H}(t)\begin{bmatrix} z^{[2]}(t,\bar{x}) \\ \lambda^{[2]}(t,\bar{x}) \end{bmatrix} = -\begin{bmatrix} k^{[2]}(t,\bar{x})) \\ l^{[2]}(t,\bar{x}) \end{bmatrix}$$

where

$$\begin{aligned} k^{[2]}(t,\bar{x}) &= f^{[2]}(t, Z(t)\bar{x}) \\ l^{[2]}(t,\bar{x}) &= \alpha\left(\frac{\partial \pi^{[3]}}{\partial z}(t, Z(t)\bar{x})\right)' - \left(\frac{\partial f^{[2]}}{\partial z}(t, Z(t)\bar{x})\right)' \Lambda(t)\bar{x} \end{aligned}$$

Again the solvability of these equations depends on the invertibility of $\mathscr{H}(t)$.

The degree three terms in (17) and (18) are

$$\mathscr{H}(t)\begin{bmatrix} z^{[3]}(t,\bar{x}) \\ \lambda^{[3]}(t,\bar{x}) \end{bmatrix} = -\begin{bmatrix} k^{[3]}(t,\bar{x}) \\ l^{[3]}(t,\bar{x}) \end{bmatrix}$$

where

$$\begin{aligned} k^{[3]}(t,\bar{x}) &= f^{[3]}(t, Z(t)\bar{x}) + \left(f^{[2]}(t, Z(t)\bar{x} + z^{[2]}(t,\bar{x}))\right)^{[3]} \\ l^{[3]}(t,\bar{x}) &= \alpha\left(\left(\frac{\partial \pi^{[3]}}{\partial z}(t, Z(t)\bar{x} + z^{[2]}(t,\bar{x})\right)^{[3]}\right)' + \alpha\left(\frac{\partial \pi^{[4]}}{\partial z}(t, Z(t)\bar{x})\right)' \\ &\quad - \left(\frac{\partial f^{[2]}}{\partial z}(t, Z(t)\bar{x})\right)' \lambda^{[2]}(t,\bar{x}) - \left(\frac{\partial f^{[2]}}{\partial z}(t, z^{[2]}(t,\bar{x})\right)' \Lambda(t)\bar{x} \\ &\quad - \left(\frac{\partial f^{[3]}}{\partial z}(t, Z(t)\bar{x})\right)' \Lambda(t)\bar{x} \end{aligned}$$

and $(\cdot)^{[d]}$ indicates the degree d terms of the enclosed expression. The higher degree terms are found in a similar fashion.

Suppose we have found $\bar{z}$ and λ as polynomials in $\bar{x}$ to degree d then (12) yields an expansion of $v(x)$ to degree d in $\bar{x}$.

We plug this into (15) and assuming $\pi(x, t|t)$ is of degree $d+1$ we obtain an approximation of the Taylor polynomial of $\pi(x, t+1|t)$ to degree $d+1$ in $\bar{x} = x - \hat{x}(t+1|t)$

$$\pi^{[0:d+1]}(x, t+1|t) = \alpha\pi^{[0:d+1]}(z^{[0:d]}(x), t|t) + \frac{1}{2}\|v^{[1:d]}(x)\|^2_{Q(t)}$$

This completes the Prediction Steps 2 and 3.

Step 4, the first Assimilation Step, requires the Taylor polynomial of the measurement function $h(t,x)$ in $\bar{x}(t+1|t) = x - \hat{x}(t+1|t)$

$$h^{[0:d]}(t+1,x) = h^{[0]}(t+1) + H(t+1)\bar{x}(t+1|t) + h^{[2]}(t+1,\bar{x}(t+1|t)) + \cdots + h^{[d]}(t+1,\bar{x}(t+1|t))$$

The expected value of $y(t+1)$ is

$$\hat{y}(t+1|t) = h(\hat{x}(t+1|t)) \;=\; h^{[0]}(t+1)$$

If $y(t+1)$ is the actual measurement at time $t+1$ then the innovation is

$$\bar{y}(t+1|t) = y(t+1) - h^{[0]}(t+1)$$

The formula (7) suggests that the Taylor polynomial of $\pi(x,t+1|t+1)$ at $\hat{x}(t+1|t)$ to degree $d+1$ can be approximated by

$$\pi^{[0:d+1]}(x,t+1|t+1) = \pi^{[0:d+1]}(x,t+1|t) + \frac{1}{2}\left(\|y(t+1) - h^{[0:d]}(t+1,x)\|^2_{R(t+1)}\right)^{[0:d+1]} \tag{21}$$

and then

$$\hat{x}(t+1|t+1) = \text{argmin}_x \pi^{[0:d+1]}(x,t+1|t+1) \tag{22}$$

Assuming $\hat{x}(t+1|t+1)$ is close to $\hat{x}(t+1|t)$ it can readily be found by several iterations of Newton's method starting at $\hat{x}(t+1|t)$.

The final assimilation step is to convert $\pi^{[0:d+1]}(x,t+1|t+1)$ given by (21) from a polynomial in $\bar{x}(t+1|t) = x - \hat{x}(t+1|t)$ to a polynomial in $x - \hat{x}(t+1|t+1)$.

4 The Relation Between the MEE and the EKF

In this section we show that the polynomial Minimum Energy estimator described above reduces to an Extended Kalman Filter when $d = 1$ and $\alpha = 1$. In fact this MEE is the extended version of the so called Information Filter, see [1, Section 6.3]. Recall the EKF for (2) where we assume that v and w are zero mean independent white Gaussian sequences of covariances $\mathscr{Q}(t)$ and $\mathscr{R}(t)$ and $\bar{x}^0$ is an independent zero mean Gaussian vector of covariance $\mathscr{P}^0$. As before for simplicity of notation we assume that $g(t,x) = G(t)$.

Let $\hat{\xi}(t|\tau)$ be the EKF estimate of $x(t)$ and $\mathscr{P}(t|\tau)$ be its approximate error covariance where $\tau = t-1$ or $\tau = t$. Given $\hat{\xi}(t|t)$ and $\mathscr{P}(t|t)$ the Prediction Step of the EKF is

$$\hat{\xi}(t+1|t) = f(t, \hat{\xi}(t|t)) \tag{23}$$

$$\mathscr{P}(t+1|t) = F(t)\mathscr{P}(t|t)F'(t) + G(t)\mathscr{Q}(t)G'(t) \tag{24}$$

where $F(t) = \frac{\partial f}{\partial x}(t, \hat{\xi}(t|t))$. We immediately recognize that (10) and (23) are the same state prediction formula.

Recall that Z and Λ are the solution of (19) or equivalently

$$I = F(t)Z + \bar{Q}(t)\Lambda$$
$$0 = P(t|t)Z - F'(t)\Lambda$$

Applying these to (15) yields

$$\begin{aligned} P(t+1|t) &= Z'P(t|t)Z + \Lambda'\bar{Q}(t)\Lambda \\ &= Z'F'(t)\Lambda + \Lambda'\bar{Q}(t)\Lambda \\ &= (I - \Lambda'\bar{Q}(t))\Lambda + \Lambda'\bar{Q}\Lambda \\ &= \Lambda \end{aligned}$$

By direct calculation (19) also implies

$$Z = P^{-1}(t|t)F'(t)\Lambda$$
$$I = \left(F(t)P^{-1}(t|t)F'(t) + \bar{Q}(t)\right)\Lambda$$

so

$$P^{-1}(t+1|t) = \Lambda^{-1} = F(t)P^{-1}(t|t)F'(t) + \bar{Q}(t)$$

By comparing this with (24) we conclude that $\mathscr{P}(t+1|t) = P^{-1}(t+1|t)$ so the Prediction Steps of the EKF and the MEE are identical.

The Assimilation Step of the EKF is

$$\hat{x}(t|t) = \hat{x}(t|t-1) + \mathscr{K}(t)(y(t) - h^{[0]}(t))$$
$$\mathscr{P}(t|t) = (I - \mathscr{K}(t)H(t))\mathscr{P}(t|t-t)$$

where

$$h^{[0]}(t) = h(t, \hat{x}(t|t-1))$$
$$H(t) = \frac{\partial h}{\partial x}(t, \hat{x}(t|t-1))$$

and the Kalman gain is

$$\mathscr{K}(t) = \mathscr{P}(t|t-1)H'(t)\left(\mathscr{R}(t) + H(t)\mathscr{P}(t|t-1)H'(t)\right)^{-1}$$

The Assimilation Step of the MEE is given by (21) and (22). When $d = 1$ (21) reduces to

$$\begin{aligned}\pi^{[0:2]}(x,t|t) &= \pi^{[0]}(t|t-1) + \frac{1}{2}(x-\hat{x}(t|t-1))'P(t|t-1)(x-\hat{x}(t|t-1)) \\ &\quad + \frac{1}{2}\|y(t) - h^{[0]}(t) - H(t)\left(x-\hat{x}(t|t-1)\right)\|^2_{R(t)} \\ &= \pi^{[0]}(t|t-1) + \frac{1}{2}\|y(t) - h^{[0]}(t)\|^2_{R(t)} \\ &\quad -(y(t) - h^{[0]}(t))'R(t)H(t)\left(x-\hat{x}(t|t-1)\right) \\ &\quad \frac{1}{2}(x-\hat{x}(t|t-1))'\left(P(t|t-1) + H'(t)R(t)H(t)\right)(x-\hat{x}(t|t-1))\end{aligned}$$

Since $\pi^{[0:2]}(x,t|t)$ is quadratic its minimum is found by setting to zero its first derivative with respect to x,

$$\begin{aligned}\hat{x}(t|t) &= \hat{x}(t|t-1) \\ &\quad + \left(P(t|t-1) + H'(t)R(t)H(t)\right)^{-1} H'(t)R(t)(y(t) - h^{[0]}(t))\end{aligned}$$

so the MEE gain is

$$K(t) = \left(P(t|t-1) + H'(t)R(t)H(t)\right)^{-1} H'(t)R(t)$$

Recall the matrix inversion lemma

$$(A+BDC)^{-1} = A^{-1} - A^{-1}B\left(D^{-1} + CA^{-1}B\right)^{-1}CA^{-1}$$

This implies that

$$\begin{aligned}\left(P(t|t-1) + H(t)R(t)H'(t)\right)^{-1} &= P^{-1}(t|t-1) - P^{-1}(t|t-1)H(t) \\ \times \left(H'(t)P^{-1}(t|t-1)H(t) + R^{-1}(t)\right)^{-1} &H'(t)P^{-1}(t|t-1)\end{aligned}$$

We multiply by $H(t)R(t)$ on the right to get

$$\begin{aligned}\left(P(t|t-1) + H(t)R(t)H'(t)\right)^{-1} H(t)R(t) &= P^{-1}(t|t-1) - P^{-1}(t|t-1)H(t) \\ \times \left(H'(t)P^{-1}(t|t-1)H(t) + R^{-1}(t)\right)^{-1} &H'(t)P^{-1}(t|t-1)H(t)R(t)\end{aligned}$$

Now

$$H'(t)P^{-1}(t|t-1)H(t)R(t) = \left(H'(t)P^{-1}(t|t-1)H(t)\right)R^{-1}(t) - I$$

so

$$\begin{aligned} K(t) &= \left(P(t|t-1) + H(t)R(t)H'(t)\right)^{-1} H'(t)R(t) \\ &= P^{-1}(t|t-1)H(t)\left(H'(t)P^{-1}(t|t-1)H(t) + R^{-1}(t)\right)^{-1} \\ &= \mathscr{P}(t|t-1)H'(t)\left(\mathscr{R}(t) + H(t)\mathscr{P}(t|t-1)H'(t)\right)^{-1} \\ &= \mathscr{K}(t) \end{aligned}$$

5 Example: Lorenz Attractor

We apply the degree one (EKF) and degree three MEE filters to a difficult problem, estimating the state of the chaotic three dimensional Lorenz Attractor from a one dimensional measurement.

The dynamics of the Lorenz Attractor is given by the differential equation

$$\begin{aligned} \dot{x}_1 &= \sigma(x_2 - x_1) \\ \dot{x}_2 &= x_1(\rho - x_3 - x_2) \\ \dot{x}_3 &= x_1 x_2 - \beta x_3 \end{aligned}$$

with the standard parameter values

$$\sigma = 10, \qquad \rho = 28, \qquad \beta = \frac{8}{3}$$

It is known that system exhibits chaotic behaviour as seen Fig. 1 which was generated by Matlab's ode45.m using its default settings.

We approximate this continuous time dynamics by an Euler discretization with a time step of $dt = 0.01$. The discrete time dynamics also exhibits chaotic behaviour as seen Fig. 2.

But the Euler time step of $dt = 0.01$ is too large for the discrete time dynamics to accurately approximate the continuous time dynamics. Figure 3 shows the difference between the continuous time and discrete time trajectories started at the same initial condition $x^0 = (0.1, 0.1, 0.1)'$. Notice the scale of the differences, they are of the same order of magnitude as size of the attractors themselves.

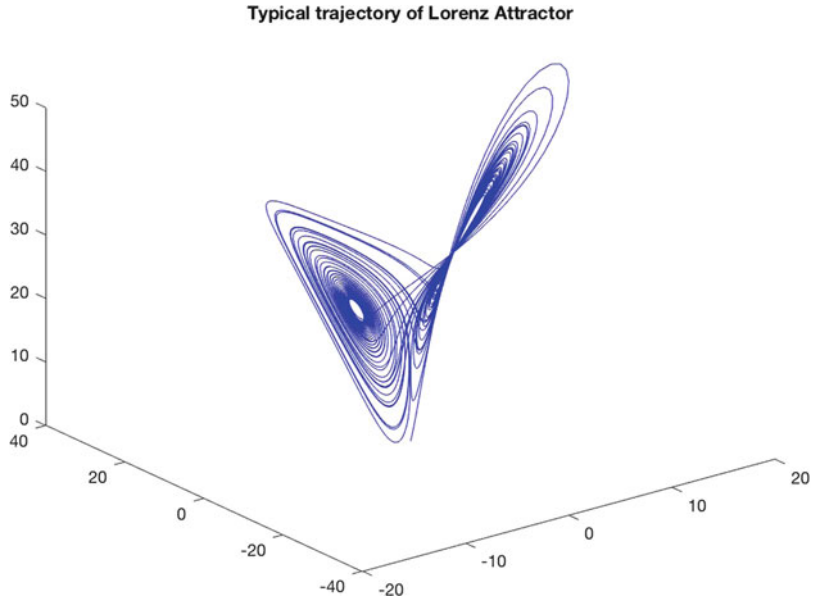


Fig. 1 Typical trajectory of continuous time Lorenz Attractor

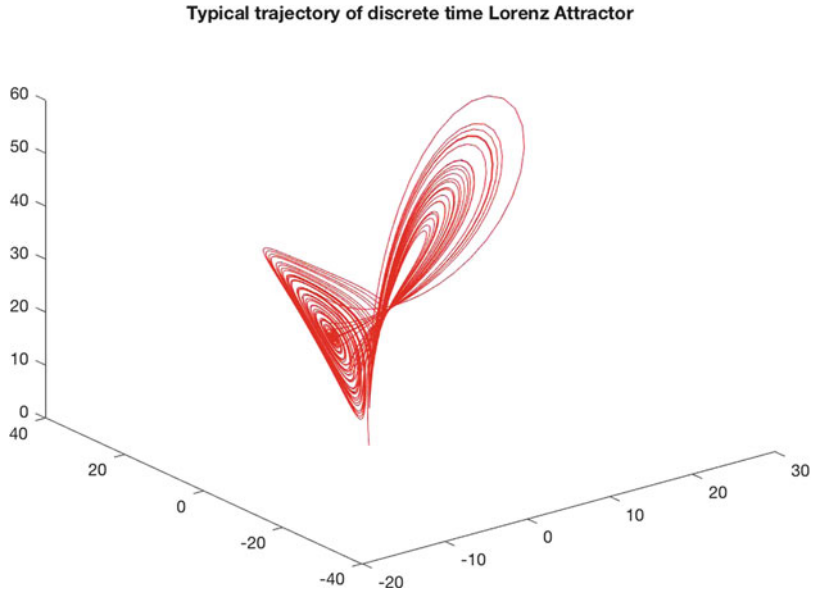


Fig. 2 Typical trajectory of discrete time Lorenz Attractor

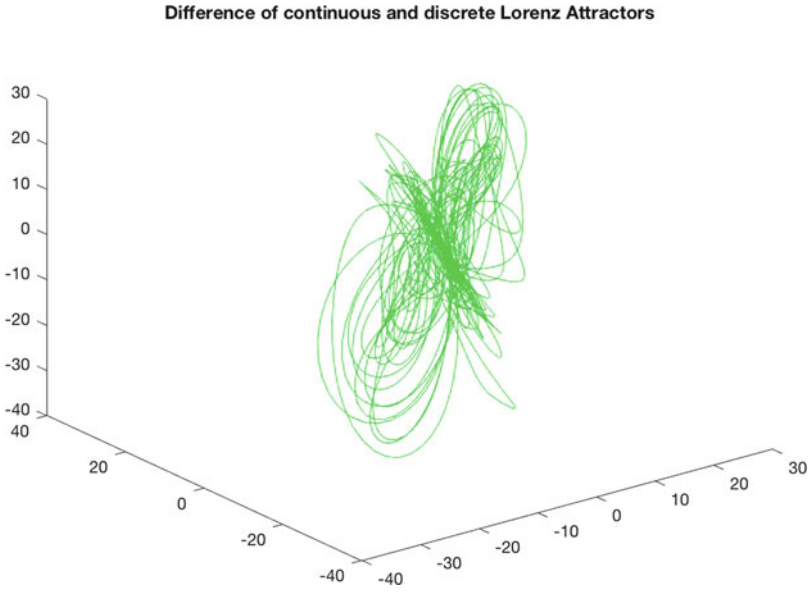


Fig. 3 Difference between continuous and discrete time Lorenz Attractors

We define the observation by

$$y(t) = x_1(t) + x_2(t) + 0.1w(t)$$

where $t = 0 : dt : T$ and $T = 50$. The coefficient of the measurement noise was chosen to be $0.1 = dt^{1/2}$ so that total variance of the noise in one unit of time is one.

We shall use the discrete time dynamics to design the degree one and three MEEs but we shall simulate these estimators using the actual measurements from the continuous time system. This mismatch between continuous and discrete in effect introduces a considerable amount of driving noise to the estimation problems.

The parameters of the filters are

$$\begin{aligned} G &= I^{3\times 3} \\ Q &= dtI^{3\times 3} = 0.01I^{3\times 3} \\ R &= dt = 0.01 \\ \alpha &= 0.999 \end{aligned}$$

The discount factor α was chosen because $0.999^{100} \approx 0.9$. This means that the filter gently forgets the past observations.

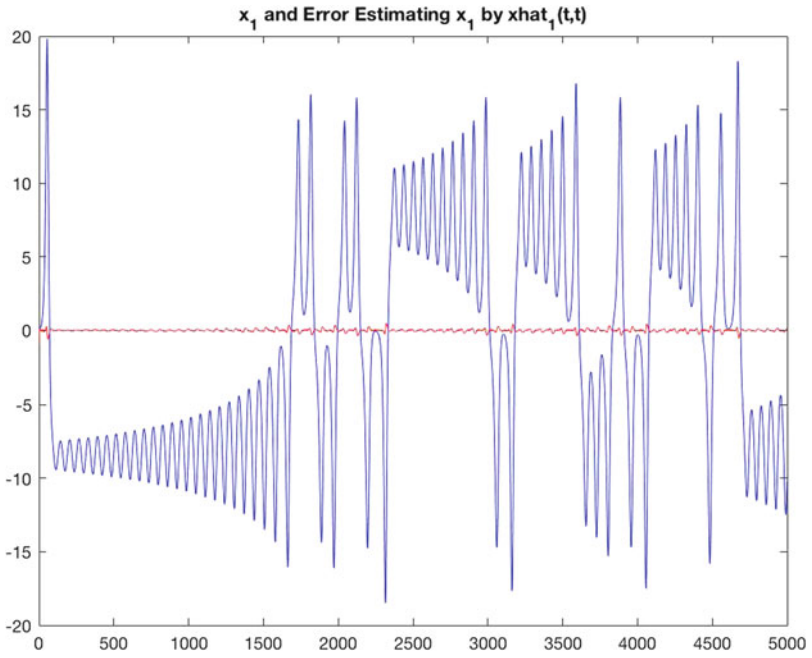


Fig. 4 $x_1(t)$ (blue) and $\tilde{x}_1(t|t)$ (red)

Figures 4, 5, and 6 shows the trajectories $x_i(t)$ of the continuous time dynamics and the corresponding errors $\tilde{x}_i(t) = x_i(t) - \hat{x}_i(t|t)$ of the degree three MEE which was designed based on the discrete time dynamics.

As one can see the degree three MEE does quite well. The immediate question is whether the degree one MEE (EKF) performs similarly. Figure 7 shows the norm of the error of the degree three MEE minus the norm of the error of the degree one MEE. Occasionally the degree three error norm is larger but most of the time the degree three error norm is smaller. The average error norm of the degree one MEE is 0.4828 while the average error norm of the degree three MEE is 0.2946, a difference of 0.1882. This is about a 40% error reduction. So the degree three MEE substantially outperforms the degree one MEE.

6 Conclusion

We have presented the Minimum Energy Estimator for discrete time systems and its Taylor polynomial approximation. The degree one and degree three MEEs were applied to estimating the state of the three dimensional Lorenz Attractor from a one dimensional measurement. Both performed reasonably well but the degree three MEE substantially outperformed the degree one MEE.

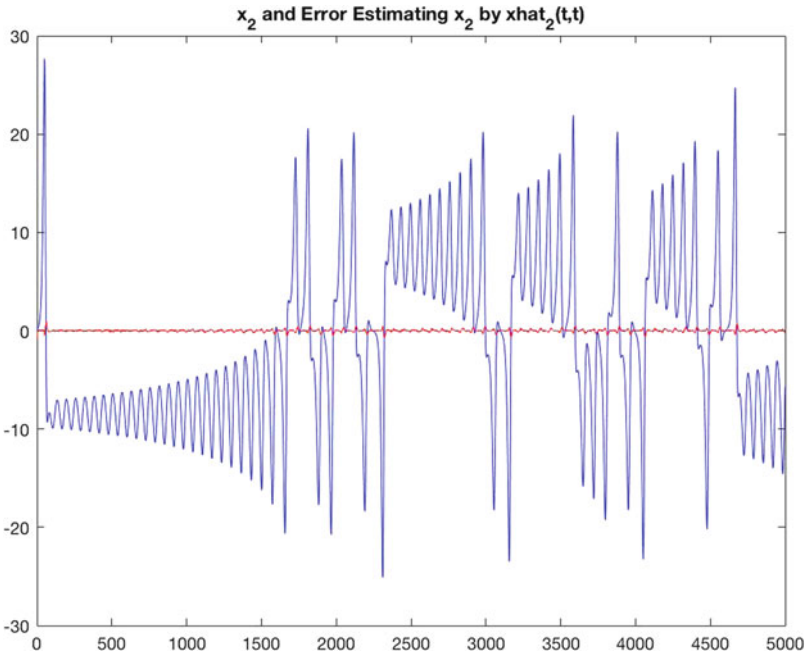


Fig. 5 $x_2(t)$ (blue) and $\tilde{x}_3(t|t)$ (red)

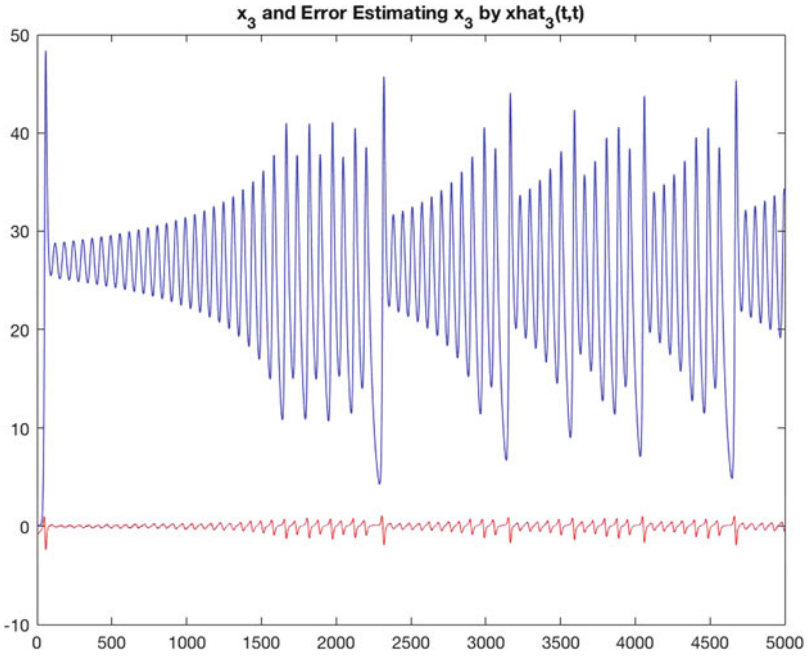


Fig. 6 $x_3(t)$ (blue) and $\tilde{x}_3(t|t)$ (red)

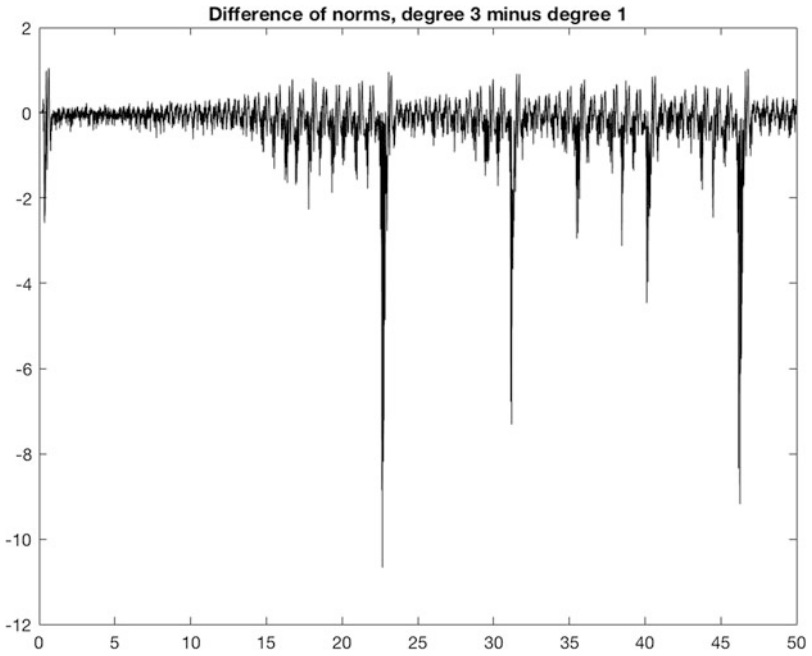


Fig. 7 Degree three error minus degree one error

References

1. Anderson, B.D.O., Moore, J.B.: Optimal Filtering. Prentice Hall, Englewood Cliffs (1979)
2. Chen, C.-T.: Introduction to Linear System Theory. Holt, Rinehart and Winston, New York (1970)
3. Gelb, A.: Applied Optimal Estimation, MIT Press, Cambridge (1974)
4. Hijab, O.B.: Minimum energy estimation. PhD Thesis, University of California, Berkeley (1980)
5. Julier, S.J., Uhlmann, J.K., Durrant-Whyte, H.F.: A new approach to filtering nonlinear systems. In: Proceedings of American Control Conference (1995)
6. Kazantis, N., Kravaris, C.: Nonlinear observer design using Lyapunov's auxiliary theorem. Syst. Control Lett. **34**, 241–247 (1998)
7. Khalil, H.K.: High-Gain Observers in Nonlinear Feedback Control. SIAM, Philadelphia (2017)
8. Krener, A.J.: The convergence of the minimum energy estimator. In: Kang, W., Xiao, M., Borges, C. (eds.) New Trends in Nonlinear Dynamics and Control, and Their Applications, pp. 187–208. Springer, Heidelberg (2003)
9. Krener, A.J.: Minimum energy estimation and moving horizon estimation. In: Proceedings of the 2015 CDC (2015)
10. Krener, A.J., Respondek, W.: Nonlinear observers with linearizable error dynamics. SIAM J. Control Optim. **23**, 197–216 (1985)
11. Mortensen, R.E.: Maximum likelihood recursive nonlinear filtering. J. Optim. Theory Appl. **2**, 386–394 (1968)

Probabilistic Max-Plus Schemes for Solving Hamilton-Jacobi-Bellman Equations

Marianne Akian and Eric Fodjo

Abstract We consider fully nonlinear Hamilton-Jacobi-Bellman equations associated to diffusion control problems involving a finite set-valued (or switching) control and possibly a continuum-valued control. In previous works (Akian and Fodjo, A probabilistic max-plus numerical method for solving stochastic control problems. In: 55th Conference on Decision and Control (CDC 2016), Las Vegas, 2016 and From a monotone probabilistic scheme to a probabilistic max-plus algorithm for solving Hamilton-Jacobi-Bellman equations. In: Kalise, D., Kunisch, K., Rao, Z. (eds.) Hamilton-Jacobi-Bellman Equations. Radon Series on Computational and Applied Mathematics, vol. 21. De Gruyter, Berlin 2018), we introduced a lower complexity probabilistic numerical algorithm for such equations by combining max-plus and numerical probabilistic approaches. The max-plus approach is in the spirit of the one of McEneaney et al. (Idempotent method for continuous-time stochastic control and complexity attenuation. In: Proceedings of the 18th IFAC World Congress, 2011, pp 3216–3221. Milano, Italie, 2011), and is based on the distributivity of monotone operators with respect to suprema. The numerical probabilistic approach is in the spirit of the one proposed by Fahim et al. (Ann Appl Probab 21(4):1322–1364, 2011). A difficulty of the latter algorithm was in the critical constraints imposed on the Hamiltonian to ensure the monotonicity of the scheme, hence the convergence of the algorithm. Here, we present new probabilistic schemes which are monotone under rather weak assumptions, and show error estimates for these schemes. These estimates will be used in further works to study the probabilistic max-plus method.

M. Akian (✉)
INRIA and CMAP, École polytechnique CNRS, Palaiseau Cedex, France
e-mail: Marianne.Akian@inria.fr

E. Fodjo
I-Fihn Consulting, Paris, France

INRIA and CMAP, École polytechnique CNRS, Palaiseau Cedex, France
e-mail: eric.fodjo@polytechnique.edu

M. Falcone et al. (eds.), *Numerical Methods for Optimal Control Problems*,
Springer INdAM Series 29, https://doi.org/10.1007/978-3-030-01959-4_9

Keywords Stochastic control · Hamilton-Jacobi-Bellman equations · Max-plus numerical methods · Tropical methods · Probabilistic schemes

1 Introduction

We consider a finite horizon diffusion control problem on $\mathbb{R}^d$ involving at the same time a "discrete" control taking its values in a finite set $\mathscr{M}$, and a "continuum" control taking its values in some subset $\mathscr{U}$ of a finite dimensional space $\mathbb{R}^p$ (for instance a convex set with nonempty interior), which we next describe.

Let T be the horizon. The state $\xi_s \in \mathbb{R}^d$ at time $s \in [0, T]$ satisfies the stochastic differential equation

$$d\xi_s = f^{\mu_s}(\xi_s, u_s)ds + \sigma^{\mu_s}(\xi_s, u_s)dW_s \ , \tag{1}$$

where $(W_s)_{s\geq 0}$ is a d-dimensional Brownian motion on a filtered probability space $(\Omega, \mathscr{F}, (\mathscr{F}_s)_{0\leq s\leq T}, P)$. The control processes $\mu := (\mu_s)_{0\leq s\leq T}$ and $u := (u_s)_{0\leq s\leq T}$ take their values in the sets $\mathscr{M}$ and $\mathscr{U}$ respectively and they are admissible if they are progressively measurable with respect to the filtration $(\mathscr{F}_s)_{0\leq s\leq T}$. We assume that, for all $m \in \mathscr{M}$, the maps $f^m : \mathbb{R}^d \times \mathscr{U} \to \mathbb{R}^d$ and $\sigma^m : \mathbb{R}^d \times \mathscr{U} \to \mathbb{R}^{d\times d}$ are continuous and satisfy properties implying the existence of the process $(\xi_s)_{0\leq s\leq T}$ for any admissible control processes μ and u.

Given an initial time $t \in [0, T]$, the control problem consists in maximizing the following payoff:

$$J(t, x, \mu, u) := \mathbb{E}\left[\int_t^T e^{-\int_t^s \delta^{\mu_\tau}(\xi_\tau, u_\tau)d\tau} \ell^{\mu_s}(\xi_s, u_s)ds \right.$$
$$\left. + e^{-\int_t^T \delta^{\mu_\tau}(\xi_\tau, u_\tau)d\tau}\psi(\xi_T) \mid \xi_t = x\right] \ ,$$

where, for all $m \in \mathscr{M}$, $\ell^m : \mathbb{R}^d \times \mathscr{U} \to \mathbb{R}$, $\delta^m : \mathbb{R}^d \times \mathscr{U} \to \mathbb{R}$, and $\psi : \mathbb{R}^d \to \mathbb{R}$ are given continuous maps. We then define the value function of the problem as the optimal payoff:

$$v(t, x) = \sup_{\mu,u} J(t, x, \mu, u) \ ,$$

where the maximization holds over all admissible control processes μ and u.

Let $\mathbb{S}_d$ denotes the set of symmetric $d \times d$ matrices and let us denote by $\leq$ the Loewner order on $\mathbb{S}_d$ ($A \leq B$ if $B - A$ is nonnegative). The Hamiltonian $\mathscr{H} : \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{S}_d \to \mathbb{R}$ of the above control problem is defined as:

$$\mathscr{H}(x, r, p, \Gamma) := \max_{m\in\mathscr{M}} \mathscr{H}^m(x, r, p, \Gamma) \ , \tag{2a}$$

with

$$\mathscr{H}^m(x, r, p, \Gamma) := \max_{u \in \mathscr{U}} \mathscr{H}^{m,u}(x, r, p, \Gamma) \ , \tag{2b}$$

$$\mathscr{H}^{m,u}(x, r, p, \Gamma) := \frac{1}{2} \operatorname{tr}\left(\sigma^m(x, u)\sigma^m(x, u)^{\mathsf{T}} \Gamma\right) + f^m(x, u) \cdot p - \delta^m(x, u)r + \ell^m(x, u) \ . \tag{2c}$$

Under suitable assumptions the value function $v : [0, T] \times \mathbb{R}^d \to \mathbb{R}$ is the unique (continuous) viscosity solution of the following Hamilton-Jacobi-Bellman equation

$$-\frac{\partial v}{\partial t} - \mathscr{H}(x, v(t, x), Dv(t, x), D^2 v(t, x)) = 0, \quad x \in \mathbb{R}^d, \ t \in [0, T), \tag{3a}$$

$$v(T, x) = \psi(x), \quad x \in \mathbb{R}^d, \tag{3b}$$

satisfying also some growth condition at infinity (in space). On may for instance assume the boundedness and Lipschitz continuity of the maps σ^m, f^m, δ^m, ℓ^m, ψ, and the compactness of $\mathscr{U}$, and look for a bounded solution, see for instance [8]. Other possible assumptions, which are satisfied in the case of Hamiltonians $\mathscr{H}^m$ associated to linear quadratic problems, are the linear or quadratic growth of the above maps, with their local Lipschitz continuity with a linearly growing Lipschitz constant, together with some additional condition like small horizon T or strong stability, see for instance [6, 13].

In [7], Fahim, Touzi and Warin proposed a probabilistic numerical method to solve such fully nonlinear partial differential equations (3), inspired by their backward stochastic differential equation interpretation given by Cheridito, Soner, Touzi and Victoir in [5]. This method consists in two steps, the first one being a time discretization of the partial differential equation using the Euler discretization of the stochastic differential equation of an uncontrolled diffusion (thus different from the controlled one). The second step of the method is based on the simulation of the discretized diffusion and linear regression estimations which can be seen as an alternative to a space discretization.

In [10, 11, 14], McEneaney, Kaise and Han proposed an idempotent numerical method which works at least when the Hamiltonians with fixed discrete control, $\mathscr{H}^m$, correspond to linear quadratic control problems. This method is based on the distributivity of the (usual) addition operation over the supremum (or infimum) operation, and on a property of invariance of the set of quadratic forms. It computes in a backward manner the value function $v(t, \cdot)$ at time t as a supremum of quadratic forms. However, as t decreases, that is as the algorithm advances, the number of quadratic forms generated by the method increases exponentially (and even become infinite if the Brownian is not discretized in space) and some pruning is necessary to reduce the complexity of the algorithm.

In [1], we introduced an algorithm combining the two above methods at least in their spirit. The algorithm applies the first step (the time discretization) of the method of [7] to the HJB equations obtained when the discrete control is fixed. Then using the simulation of as many uncontrolled stochastic processes as discrete controls, it applies a max-plus type space discretization in the spirit of the method of [10, 11, 14]. Then, without any pruning, the number of quadratic forms representing the value function is bounded by the sampling size [1]. Hence, the complexity of the algorithm is bounded polynomially in the number of discretization time steps and the sampling size, see [1].

The convergence of the probabilistic max-plus algorithm proposed in [1] is based, as for the one of [7], on the monotonicity of the time discretization scheme. In particular, see [7], this monotonicity allows one to apply the theorem of Barles and Souganidis [4]. However, for this monotonicity to hold, critical constraints are imposed on the Hamiltonians: the diffusion matrices $\sigma^m(x,u)\sigma^m(x,u)^\mathsf{T}$ need at the same time to be bounded from below (with respect to the Loewner order) by a symmetric positive definite matrix a and bounded from above by $(1+2/d)a$. Such a constraint is restrictive, in particular it may not hold even when the matrices $\sigma^m(x,u)$ do not depend on x and u but take different values for $m \in \mathscr{M}$. In [9], Guo, Zhang and Zhuo proposed a monotone scheme exploiting the diagonal part of the diffusion matrices and combining a usual finite difference scheme to the scheme of [7]. This scheme can be applied in more general situations than the one of [7], but still does not work for general control problems. In [2], we proposed a new probabilistic discretization scheme of the second order derivatives which allowed us to obtain the monotonicity of the time discretization of HJB equations (3) with bounded coefficients and an ellipticity condition. Indeed, the monotonicity holds when the first order terms of the HJB equation are dominated by the second order ones.

Here, we propose a new probabilistic scheme for the first order derivatives which is in the spirit of the upwind discretization used by Kushner for optimal control problems, see for instance [12]. This allows one to solve also degenerate equations or to use time discretizations based on the simulation of a diffusion with same variance as the controlled process.

As soon as the convergence of the algorithm holds, one may expect to obtain estimates on the error leading to bounds on the complexity as a function of the error. Both depend on the error of the time discretization on the one hand, and the error of the "space discretization" on the other hand. We shall only study here the error of the time discretization, for which we obtain error estimates similar to the ones in [7], using the results of Barles and Jakobsen [3]. We shall also show how to adapt the method of [1, 2] with the new time discretization scheme.

The paper is organized as follows. In Sect. 2, we recall the scheme of [7]. Then, monotone probabilistic discretizations of second order and first order derivatives are presented in Sect. 3, with error estimates for regular functions. These discretizations and error estimates are applied to Hamilton-Jacobi-equations in Sect. 4, for which the error on a bounded Lipschitz solution is obtained by using the results of Barles and Jakobsen [3]. In Sect. 5, we recall the algorithm of [1, 2] and show how it can be combined with the scheme of Sect. 4.

2 The Probabilistic Time Discretization of Fahim, Touzi and Warin

Let us first recall the first step of the probabilistic numerical scheme proposed by Fahim, Touzi and Warin in [7], which can be viewed as a time discretization.

Let h be a time discretization step such that T/h is an integer. We denote by $\mathscr{T}_h = \{0, h, 2h, \ldots, T-h\}$ and $\overline{\mathscr{T}}_h = \{0, h, 2h, \ldots, T\}$ the set of discretization times of $[0, T)$ and $[0, T]$ respectively. Let $\mathscr{H}$ be any Hamiltonian of the form (2). Let us decompose $\mathscr{H}$ as the sum of the (linear) generator $\mathscr{L}$ of a diffusion (with no control) and of a nonlinear elliptic Hamiltonian $\mathscr{G}$, that is $\mathscr{H} = \mathscr{L} + \mathscr{G}$ with

$$\mathscr{L}(x, r, p, \Gamma) = \mathscr{L}(x, p, \Gamma) := \frac{1}{2}\operatorname{tr}(a(x)\Gamma) + \underline{f}(x) \cdot p \ ,$$

$a(x) = \underline{\sigma}(x)\underline{\sigma}(x)^\mathsf{T}$ and $\mathscr{G}$ such that $a(x)$ is positive definite and $\partial_\Gamma \mathscr{G}$ is positive semidefinite, for all $x \in \mathbb{R}^d$, $r \in \mathbb{R}$, $p \in \mathbb{R}^d$, $\Gamma \in \mathbb{S}_d$. Denote by $\hat{X}$ the Euler discretization of the diffusion with generator $\mathscr{L}$:

$$\hat{X}(t+h) = \hat{X}(t) + \underline{f}(\hat{X}(t))h + \underline{\sigma}(\hat{X}(t))(W_{t+h} - W_t) \ . \tag{4}$$

The time discretization of (3) proposed in [7] has the following form:

$$v^h(t, x) = T_{t,h}(v^h(t+h, \cdot))(x), \quad t \in \mathscr{T}_h \ , \tag{5}$$

with

$$T_{t,h}(\phi)(x) = \mathscr{D}^0_{t,h}(\phi)(x) + h\mathscr{G}(x, \mathscr{D}^0_{t,h}(\phi)(x), \mathscr{D}^1_{t,h}(\phi)(x), \mathscr{D}^2_{t,h}(\phi)(x)) \ , \tag{6}$$

where, for $i = 0, 1, 2$, $\mathscr{D}^i_{t,h}(\phi)$ is the approximation of the ith differential of $e^{h\mathscr{L}}\phi$ obtained using the following scheme:

$$\mathscr{D}^i_{t,h}(\phi)(x) = \mathbb{E}(D^i\phi(\hat{X}(t+h)) \mid \hat{X}(t) = x) \tag{7a}$$

$$= \mathbb{E}(\phi(\hat{X}(t+h))\mathscr{P}^i_{t,x,h}(W_{t+h} - W_t) \mid \hat{X}(t) = x) \ , \tag{7b}$$

where, D^i denotes the ith differential operator, and for all t, x, h, i, $\mathscr{P}^i_{t,x,h}$ is the polynomial of degree i in the variable $w \in \mathbb{R}^d$ given by:

$$\mathscr{P}^0_{t,x,h}(w) = 1 \ , \tag{8a}$$

$$\mathscr{P}^1_{t,x,h}(w) = (\underline{\sigma}(x)^\mathsf{T})^{-1}h^{-1}w \ , \tag{8b}$$

$$\mathscr{P}^2_{t,x,h}(w) = (\underline{\sigma}(x)^\mathsf{T})^{-1}h^{-2}(ww^\mathsf{T} - hI)(\underline{\sigma}(x))^{-1} \ , \tag{8c}$$

where I is the $d \times d$ identity matrix. Note that the second equality in (7) holds for all ϕ with exponential growth [7, Lemma 2.1].

In [7], the convergence of the time discretization scheme (5) is proved by using the theorem of Barles and Souganidis of [4], under the above assumptions together with the critical assumption that $\partial_\Gamma \mathscr{G}$ is lower bounded by some positive definite matrix (for all $x \in \mathbb{R}^d,\ r \in \mathbb{R},\ p \in \mathbb{R}^d,\ \Gamma \in \mathbb{S}_d$) and that $\mathrm{tr}(a(x)^{-1}\partial_\Gamma \mathscr{G}) \leq 1$.

Indeed, let us say that an operator T between any partially ordered sets $\mathscr{F}$ and $\mathscr{F}'$ of real valued functions (for instance the set of bounded functions from some set Ω to $\mathbb{R}$, or $\mathbb{R}^n$) is *L-almost monotone*, for some constant $L \geq 0$, if

$$\phi, \psi \in \mathscr{F},\ \phi \leq \psi \implies T(\phi) \leq T(\psi) + L \sup(\psi - \phi)\ , \tag{9}$$

and that it is *monotone*, when this holds for $L = 0$.

The above conditions together with the boundedness of $\partial_p \mathscr{G}$ are used to show (in Lemma 3.12 and 3.14 of [7]) that the operator $T_{t,h}$ is a Ch-almost monotone operator over the set of Lipschitz continuous functions from $\mathbb{R}^d$ to $\mathbb{R}$. Then, this property, together with other technical assumptions, are used to obtain the assumptions of the theorem of Barles and Souganidis of [4], and also estimates in the same spirit as in [3].

In [1], we proposed to bypass the critical constraint, by assuming that the Hamiltonians $\mathscr{H}^m$ (but not necessarily $\mathscr{H}$) satisfy the critical constraint, and applying the above scheme to the Hamiltonians $\mathscr{H}^m$.

In [2], we proposed an approximation of $\mathbb{E}(D^2\phi(\hat{X}(t+h)) \mid \hat{X}(t) = x)$ or $D^2\phi(x)$ that we recall in the next section. It is expressed as a conditional expectation as in (7b) but depend on the derivatives of $\mathscr{G}$ with respect to Γ at the given point, via the matrices $\sigma^m(x,u)$ of the control problem. Below, we also propose an approximation of $\mathbb{E}(D\phi(\hat{X}(t+h)) \mid \hat{X}(t) = x)$ or $D\phi(x)$ which is monotone in itself and thus allows one to consider the case where the derivatives of $\mathscr{G}$ with respect to Γ are zero or degenerate nonnegative matrices.

3 Monotone Probabilistic Approximation of First and Second Order Derivatives and Their Estimates

We first describe the approximation of the second order derivatives proposed in [2]. Consider any matrix $\Sigma \in \mathbb{R}^{d\times\ell}$ with $\ell \in \mathbb{N}$ and let us denote by $\Sigma_{\cdot j},\ j = 1, \ldots \ell$, its columns. We denote by $\mathscr{C}^k([0,T] \times \mathbb{R}^d)$ or simply $\mathscr{C}^k$ the set of functions from $[0,T] \times \mathbb{R}^d$ to $\mathbb{R}$ with continuous partial derivatives up to order k in t and x, and by $\mathscr{C}^k_{\mathrm{b}}([0,T] \times \mathbb{R}^d)$ or $\mathscr{C}^k_{\mathrm{b}}$ the subset of functions with bounded such derivatives. Then, for any $v \in \mathscr{C}^2$, we have

$$\frac{1}{2}\,\mathrm{tr}(\underline{\sigma}(x)\Sigma\Sigma^{\mathsf{T}}\underline{\sigma}^{\mathsf{T}}(x)\,D^2v(t,x)) = \frac{1}{2}\sum_{j=1}^{\ell} \Sigma_{\cdot j}^{\mathsf{T}}\underline{\sigma}^{\mathsf{T}}(x)\,D^2v(t,x)\underline{\sigma}(x)\Sigma_{\cdot j}\ . \tag{10}$$

For any integer k, consider the polynomial:

$$\mathscr{P}^2_{\Sigma,k}(w) = \sum_{j=1}^{\ell} \|\Sigma_{.j}\|_2^2 \left(c_k \left(\frac{[\Sigma^\mathsf{T} w]_j}{\|\Sigma_{.j}\|_2} \right)^{4k+2} - d_k \right) \quad , \tag{11a}$$

with

$$c_k := \frac{1}{(4k+2)\mathbb{E}\left[N^{4k+2}\right]} \quad , \quad d_k := \frac{1}{4k+2} \quad , \tag{11b}$$

where N is a standard normal random variable, and where we use the convention that the jth term of the sum is zero when $\|\Sigma_{.j}\|_2 = 0$. This is the sum of the same expression defined for each column $\Sigma_{.j}$ instead of Σ.

Let $v \in \mathscr{C}^4_\mathrm{b}$, and $\hat{X}$ as in (4), then, the following expression is an approximation of (10) with an error in $O(h)$ uniform in t and x [2, Th. 1.3.1]:

$$h^{-1}\mathbb{E}\left[v(t+h, \hat{X}(t+h)) \mathscr{P}^2_{\Sigma,k}(h^{-1/2}(W_{t+h} - W_t)) \mid \hat{X}(t) = x \right] \quad . \tag{12}$$

In order to obtain error estimates, we need the more precise following result. For p and q two integers and ϕ a function from $[0, T] \times \mathbb{R}^d$ to $\mathbb{R}$ with partial derivatives up to order p in t and q in x, we introduce the following notation:

$$|\partial_t^p D^q \phi| = \sup_{\substack{(t,x)\in[0,T]\times\mathbb{R}^d \\ (\beta_i)_i \in \mathbb{N}^d,\ \sum_i \beta_i = q}} \left| \frac{\partial^{i+q}\phi}{\partial t^p \partial x_1^{\beta_1} \dots \partial x_d^{\beta_d}}(t,x) \right|$$

In the sequel, $\|\cdot\|$ will denote any norm on $\mathbb{R}^d$ or on $\mathbb{R}^{d\times d}$. Also $[x]_i$ will denote the ith coordinate of any vector $x \in \mathbb{R}^d$, and $[A]_{ij}$ will denote the (i, j) entry of any matrix $A \in \mathbb{R}^{d\times\ell}$.

Theorem 1 *Let $\hat{X}$ be as in (4), and denote $W_h^t = W_{t+h} - W_t$. Consider any matrix $\Sigma \in \mathbb{R}^{d\times\ell}$ with $\ell \le d$. Assume that $\underline{f}$ and $\underline{\sigma}$ are bounded by some constant C uniformly in (t and) x, and let M be an upper bound of $\|\Sigma\Sigma^\mathsf{T}\|$. Then, there exists $K = K(C, M) > 0$ such that, for all $v \in \mathscr{C}^4_b([0, T] \times \mathbb{R}^d)$, we have, for all $(t, x) \in \mathscr{T}_h \times \mathbb{R}^d$,*

$$\begin{aligned}
&\left| h^{-1}\mathbb{E}\left[v(t+h, \hat{X}(t+h)) \mathscr{P}^2_{\Sigma,k}(h^{-1/2} W_h^t) \mid \hat{X}(t) = x \right] \right. \\
&\quad \left. - \frac{1}{2} \operatorname{tr}(\underline{\sigma}(x)\Sigma\Sigma^\mathsf{T}\underline{\sigma}^\mathsf{T}(x)\, D^2 v(t,x)) \right| \\
&\le K(1+\sqrt{h})^4 \Big[h(|\partial_t^1 D^2 v| + |\partial_t^0 D^3 v| + |\partial_t^0 D^4 v|) + \\
&\quad h\sqrt{h}|\partial_t^1 D^3 v| + h^2|\partial_t^2 D^2 v| + h^2\sqrt{h}|\partial_t^3 D^1 v| + h^3|\partial_t^4 D^0 v| \Big] \quad .
\end{aligned}$$

Sketch of Proof The proof follows from the following lemma and the property that $[h^{-1/2}W_h^t]_i$ are independent standard normal variables and that normal random variables have all their moments finite. □

Lemma 1 *Let v, W_h^t and Σ be as in Theorem 1. For all $(t,x) \in \mathscr{T}_h \times \mathbb{R}^d$, we have*

$$\begin{aligned}
&h^{-1}\mathbb{E}\left[v(t+h,\hat{X}(t+h))\mathscr{P}^2_{\Sigma,k}(h^{-1/2}W_h^t) \mid \hat{X}(t)=x\right]\\
&=\frac{1}{2}\operatorname{tr}(\underline{\sigma}(x)\Sigma\Sigma^T\underline{\sigma}^T(x)\,D^2v(t,x))+\frac{h}{2}\operatorname{tr}(\underline{\sigma}(x)\Sigma\Sigma^T\underline{\sigma}^T(x)\,\frac{\partial D^2v(t,x)}{\partial t})\\
&\quad+\frac{h}{2}\sum_{i,j,p}\left(\frac{\partial^3 v}{\partial x_i\partial x_j\partial x_p}(t,x)[\underline{\sigma}(x)\Sigma\Sigma^T\underline{\sigma}^T(x)]_{ij}[\underline{f}(x)]_p\right)\\
&\quad+\mathbb{E}\left[M^4(v,h,t,x,W_h^t)\mathscr{P}^2_{\Sigma,k}(h^{-1/2}W_h^t)\right]\ , \qquad (13)
\end{aligned}$$

where, for each h,t,x, M^4 is a continuous function of W_h^t such that:

$$\begin{aligned}
M^4(v,h,t,x,W_h^t)\le&\frac{h^3}{24}|\frac{\partial^4 v}{\partial t^4}|\\
&+\frac{h^2}{6}\sum_{i=1}^d|\frac{\partial^4 v}{\partial t^3\partial x_i}|\,|[\underline{f}(x)h+\underline{\sigma}(x)W_h^t]_i|\\
&+\frac{h}{4}\sum_{i,j}|\frac{\partial^4 v}{\partial t^2\partial x_i\partial x_j}|\,|[\underline{f}(x)h+\underline{\sigma}(x)W_h^t]_i[\underline{f}(x)h+\underline{\sigma}(x)W_h^t]_j|\\
&+\frac{1}{6}\sum_{i,j,p}|\frac{\partial^4 v}{\partial t\partial x_i\partial x_j\partial x_p}|\,|[\underline{f}(x)h+\underline{\sigma}(x)W_h^t]_i[\underline{f}(x)h+\underline{\sigma}(x)W_h^t]_j\\
&\qquad[\underline{f}(x)h+\underline{\sigma}(x)W_h^t]_p|\\
&+\frac{1}{24h}\sum_{i,j,p,q}|\frac{\partial^4 v}{\partial x_i\partial x_j\partial x_p\partial x_q}|\,|[\underline{f}(x)h+\underline{\sigma}(x)W_h^t]_i\\
&\qquad[\underline{f}(x)h+\underline{\sigma}(x)W_h^t]_j[\underline{f}(x)h+\underline{\sigma}(x)W_h^t]_p[\underline{f}(x)h+\underline{\sigma}(x)W_h^t]_q|\ .
\end{aligned}$$

where $|\frac{\partial^{p+q}v}{\partial t^p\partial x_{i_1}\cdots\partial x_{i_q}}|$ is any bound of the corresponding derivative in the segment between (t,x) and $(t+h,x+\underline{f}(x)h+\underline{\sigma}(x)W_h^t)$ and is thus bounded by $|\partial_t^p D^q v|$.

Sketch of Proof Applying at the point $(t+h,\hat{X}(t+h))$ a Taylor expansion of v around (t,x) to order 3, and denoting by M^4 the rest of this expansion, we get that M^4 is continuous with respect to $(t+h,\hat{X}(t+h))$, as the difference of two continuous functions, and thus can be seen as a function of (h,t,x,W_h^t) that is continuous with respect to W_h^t. Moreover, it satisfies the bound stated in the lemma. For any fixed (t,x), the above Taylor expansion is a polynomial of degree 3 in the

variables h and $[\underline{f}(x)h+\underline{\sigma}(x)W_h^t]_i$. One then need to compute the expectation of the product of any monomial in these variables with $h^{-1}\mathscr{P}^2_{\Sigma,k}(h^{-1/2}W_h^t)$. Following the arguments used in the sketch of proof of [2, Theorem 1.3.1], it is sufficient to consider the case when $\ell = 1$ (by using the sum expression of $\mathscr{P}^2_{\Sigma,k}$) and Σ is the unit vector $(1,0,\ldots,0)^\mathsf{T}$ (by considering a scaling and a change of variable by a unitary matrix the first column of which is proportional to Σ, since the property that a vector has independent standard normal random coordinates is invariant by any unitary change of variable). Since $h^{-1/2}W_h^t$ is a d-dimensional vector with independent standard normal random coordinates, $\mathscr{P}^2_{\Sigma,k}(h^{-1/2}W_h^t) = c_k(([h^{-1/2}W_h^t]_1)^{4k+2} - \mathbb{E}([h^{-1/2}W_h^t]_1^{4k+2}))$ has zero expectation and after multiplication with any monomial in W_h^t with odd total degree, or with odd degree in one of the variables $[W_h^t]_i$, its expectation is again zero. This eliminates all the terms of the Taylor expansion except the ones which involve monomials with degree 2 in the $[\underline{\sigma}(x)W_h^t]_i$. These terms come from monomials with degree 2 or 3 in the $[\underline{f}(x)h + \underline{\sigma}(x)W_h^t]_i$ and lead to the three first terms of the sum in (13). □

Let us also introduce the following approximation of the first order derivatives. For any vector $g \in \mathbb{R}^d$, consider the piecewise linear function $\mathscr{P}^1_g$ on $\mathbb{R}^d$:

$$\mathscr{P}^1_g(w) = 2(g_+ \cdot w_+ + g_- \cdot w_-) \ , \tag{14}$$

where for any vector $\mu \in \mathbb{R}^d$, μ_+, $\mu_- \in \mathbb{R}^d$ are defined such that $[\mu_+]_i = \max([\mu]_i, 0)$, $[\mu_-]_i = -\min([\mu]_i, 0)$. Note that $\mathscr{P}^1_g$ is nonnegative. We shall show that

$$\mathbb{E}\Big[(v(t+h, \hat{X}(t+h)) - v(t,x))\mathscr{P}^1_g(h^{-1}W_h^t)\Big] \tag{15}$$

is a monotone approximation of

$$(\underline{\sigma}(x)g) \cdot Dv(x) \ .$$

Before this, let us note that if $\underline{\sigma}(x)$ is the identity matrix, $\underline{f}(x) = 0$ and $[h^{-1/2}W_h^t]_i$ are discretized by independent random variables taking the values 1 and -1 with probability $1/2$, then the discretization $\mathscr{D}^1_{t,h}(v(t+h,\cdot))(x)$ defined in (7b) is equivalent to a centered discretization of $Dv(x)$ with space step $\Delta x = h^{1/2}$, whereas (15) corresponds to the Kushner (upwind) discretization [12]

$$\sum_{i=1}^d \left[[g_i]_+ \frac{v(t+h, x+h^{1/2}e_i) - v(t,x)}{h^{1/2}} + [g_i]_- \frac{v(t+h, x-h^{1/2}e_i) - v(t,x)}{h^{1/2}} \right] \ .$$

As for Theorem 1, the following result is deduced from Lemma 2 using the boundedness of the moments of normal random variables.

Theorem 2 *Let $\hat{X}$ as in (4), and denote $W_h^t = W_{t+h} - W_t$. Consider any vector $g \in \mathbb{R}^d$. Assume that $\underline{f}$ and $\underline{\sigma}$ are bounded by some constant C uniformly in (t and) x, and let M be an upper bound of $\|g\|$. Then, there exists $K = K(C, M) > 0$ such that, for all $v \in \mathscr{C}_b^2([0,T] \times \mathbb{R}^d)$, we have, for all $(t,x) \in \mathscr{T}_h \times \mathbb{R}^d$,*

$$\begin{aligned}\left|(\underline{\sigma}(x)g)\cdot Dv - \mathbb{E}\Big[(v(t+h,\hat{X}(t+h)) - v(t,x))\mathscr{P}_g^1(h^{-1}W_h^t)\Big]\right| \\ \leq K(1+\sqrt{h})^2\Big[\sqrt{h}(|\partial_t^1 D^0 v| + |\partial_t^0 D^1 v| + |\partial_t^0 D^2 v|) \\ + h(|\partial_t^1 D^1 v|) + h\sqrt{h}|\partial_t^2 D^0 v|\Big]\ .\end{aligned}$$

Lemma 2 *Let v, W_h^t and g be as in Theorem 2. For all $(t,x) \in \mathscr{T}_h \times \mathbb{R}^d$, we have*

$$\begin{aligned}(\underline{\sigma}(x)g)\cdot Dv =& \mathbb{E}\Big[(v(t+h,\hat{X}(t+h)) - v(t,x))\mathscr{P}_g^1(h^{-1}W_h^t)\Big] \\ & - h(\frac{\partial v}{\partial t}(t,x) + \underline{f}(x)\cdot Dv(t,x))\mathbb{E}[\mathscr{P}_g^1(h^{-1}W_h^t)] \\ & - \mathbb{E}[M^2(v,h,t,x,W_h^t)\mathscr{P}_g^1(h^{-1}W_h^t)]\ ,\end{aligned}$$

where, for each h,t,x, M^2 is a continuous function of W_h^t such that:

$$\begin{aligned}|M^2(v,h,t,x,W_h^t)| \leq & \frac{h^2}{2}|\frac{\partial^2 v}{\partial t^2}| \\ & + h\sum_{i=1}^d |\frac{\partial^2 v}{\partial t \partial x_i}|\ |[\underline{f}(x)h + \underline{\sigma}(x)W_h^t]_i| \\ & + \frac{1}{2}\sum_{i,j=1}^d |\frac{\partial^2 v}{\partial x_i \partial x_j}|\ |[\underline{f}(x)h + \underline{\sigma}(x)W_h^t]_i[\underline{f}(x)h + \underline{\sigma}(x)W_h^t]_j|\ ,\end{aligned}$$

where $|\frac{\partial^{p+q} v}{\partial t^p \partial x_{i_1}\cdots\partial x_{i_q}}|$ is as in Lemma 1.

Sketch of Proof Similarly to the proof of Lemma 1, applying at the point $(t+h, \hat{X}(t+h))$ a Taylor expansion of v around (t,x) to order 1, and denoting by M^2 the rest of this expansion, we get that M^2 satisfies the conditions of the lemma. Then the result follows from $\mathbb{E}\Big[((\underline{\sigma}(x)W_h^t)\cdot Dv)\mathscr{P}_g^1(h^{-1}W_h^t)\Big] = \mathbb{E}\Big[(W_h^t \cdot (\underline{\sigma}(x)^\mathsf{T} Dv))\mathscr{P}_g^1(h^{-1}W_h^t)\Big] = g\cdot(\underline{\sigma}(x)^\mathsf{T} Dv) = (\underline{\sigma}(x)g)\cdot Dv$. □

We shall also need the following bound, that can be proved along the same lines as the previous theorems. We do not give the proof since it can be bypassed by using alternatively the proof of Lemma 3.22 in [7].

Lemma 3 *Let $\mathscr{L}$, $\hat{X}$ and $\mathscr{D}^0_{t,h}$ be as in Sect. 2. Denote $W^t_h = W_{t+h} - W_t$. Assume that $\underline{f}$ and $\underline{\sigma}$ are bounded by some constant C uniformly in (t and) x. Then, there exists $K = K(C) > 0$ such that, for all $v \in \mathscr{C}^4_b([0,T] \times \mathbb{R}^d)$, we have, for all $(t,x) \in \mathscr{T}_h \times \mathbb{R}^d$,*

$$\begin{aligned}
&\left|h^{-1}(\mathscr{D}^0_{t,h}(v(t+h,\cdot)) - v(t,x)) - (\partial^1_t v + \mathscr{L}(x, Dv(t,x), D^2v(t,x)))\right| = \\
&\left|h^{-1}(\mathbb{E}(v(t+h,\hat{X}(t+h)) \mid \hat{X}(t) = x) - v(t,x)) - (\partial^1_t v + \mathscr{L}(x, Dv(t,x), D^2v(t,x)))\right| \\
&\le K(1+\sqrt{h})^4\Big[h(|\partial^0_t D^2 v| + |\partial^1_t D^1 v| + |\partial^2_t D^0 v| + |\partial^0_t D^3 v| + |\partial^1_t D^2 v| + |\partial^0_t D^4 v|) \\
&\quad + h\sqrt{h}|\partial^1_t D^3 v| + h^2(|\partial^2_t D^2 v| + |\partial^2_t D^1 v| + |\partial^3_t D^0 v|) \\
&\quad + h^2\sqrt{h}|\partial^3_t D^1 v| + h^3|\partial^4_t D^0 v|\Big].
\end{aligned}$$

4 Monotone Probabilistic Schemes for HJB Equations

We shall apply the above approximations of the first and second order derivatives in (3) in the same way as in [2]. Let us decompose the Hamiltonian $\mathscr{H}^{m,u}$ of (2c) as $\mathscr{H}^{m,u} = \mathscr{L}^m + \mathscr{G}^{m,u}$ with

$$\mathscr{L}^m(x,p,\Gamma) := \frac{1}{2}\operatorname{tr}\left(a^m(x)\Gamma\right) + \underline{f}^m(x)\cdot p \ ,$$

and $a^m(x) = \underline{\sigma}^m(x)\underline{\sigma}^m(x)^\mathsf{T}$, and denote by $\hat{X}^m$ the Euler discretization of the diffusion with generator $\mathscr{L}^m$. We may choose the same linear operator $\mathscr{L}^m$ for different values of m, which is the case in Algorithm 1 below. Assume that $a^m(x)$ is positive definite and that $a^m(x) \le \sigma^m(x,u)\sigma^m(x,u)^\mathsf{T}$ for all $x \in \mathbb{R}^d$, $u \in \mathscr{U}$, and denote by $\Sigma^m(x,u)$ any $d \times \ell$ matrix such that

$$\sigma^m(x,u)\sigma^m(x,u)^\mathsf{T} - a^m(x) = \underline{\sigma}^m(x)\Sigma^m(x,u)\Sigma^m(x,u)^\mathsf{T}\underline{\sigma}^m(x)^\mathsf{T} \ . \tag{16}$$

One may use for instance a Cholesky factorization of the matrix $\underline{\sigma}^m(x)^{-1}(\sigma^m(x,u)\sigma^m(x,u)^\mathsf{T} - a^m(x))(\underline{\sigma}^m(x))^\mathsf{T})^{-1}$ in which zero columns are eliminated to obtain a rectangular matrix $\Sigma^m(x,u)$ of size $d \times \ell$ when the rank of the initial matrix is equal to $\ell < d$.

Denote also by $g^m(x,u)$ the d-dimensional vector such that

$$f^m(x,u) - \underline{f}^m(x) = \underline{\sigma}^m(x)g^m(x,u) \ . \tag{17}$$

Define

$$\mathscr{G}^m_1(x,p,g) := (\underline{\sigma}^m(x)g)\cdot p \tag{18a}$$

$$\mathscr{G}^m_2(x,\Gamma,\Sigma) := \frac{1}{2}\operatorname{tr}\left(\underline{\sigma}^m(x)\Sigma\Sigma^\mathsf{T}\underline{\sigma}^m(x)^\mathsf{T}\Gamma\right) \tag{18b}$$

so that

$$\mathscr{G}^{m,u}(x,r,p,\Gamma) = \ell^m(x,u) - \delta^m(x,u)r + \mathscr{G}_1^m(x,p,g^m(x,u)) + \mathscr{G}_2^m(x,\Gamma,\Sigma^m(x,u)) \quad .$$

Applying Theorems 1 and 2 and Lemma 3, we deduce the following result which shows the consistency of the scheme (19), together with estimates that are necessary to apply the results of Barles and Jakobsen in [3].

Theorem 3 *Let $\underline{\sigma}^m$, $\underline{f}^m$, $\hat{X}^m$ and $\mathscr{L}^m$ be as above. Let us consider the following "discretization" operators from the set of functions from $\overline{\mathscr{T}_h} \times \mathbb{R}^d$ to $\mathbb{R}$ to the set of functions from $\mathscr{T}_h \times \mathbb{R}^d$ to $\mathbb{R}$:*

$$\mathscr{D}^0_{t,h,m}(\phi)(t,x) := \mathbb{E}\left[\phi(t+h, \hat{X}^m(t+h)) \mid \hat{X}(t) = x\right]$$

$$\mathscr{D}^1_{t,h,m,g}(r,\phi)(t,x) := \mathbb{E}\left[(\phi(t+h, \hat{X}^m(t+h)) - r)\mathscr{P}^1_g(h^{-1}(W_{t+h} - W_t)) \mid \hat{X}(t) = x\right]$$

$$\mathscr{D}^2_{t,h,m,\Sigma,k}(\phi)(t,x) := h^{-1}\mathbb{E}\left[\phi(t+h, \hat{X}^m(t+h))\mathscr{P}^2_{\Sigma,k}(h^{-\frac{1}{2}}(W_{t+h} - W_t)) \mid \hat{X}^m(t) = x\right]$$

with $\mathscr{P}^1_g$ and $\mathscr{P}^2_{\Sigma,k}$ as in (14) and (11) respectively.

Then, consider the following discretization of (3):

$$\mathscr{K}(h,t,x,v^h(t,x),v^h) = 0, \quad (t,x) \in \mathscr{T}_h \times \mathbb{R}^d \quad , \tag{19}$$

where v^h is a map from $\overline{\mathscr{T}_h} \times \mathbb{R}^d$ to $\mathbb{R}$, and $\mathscr{K}$ is defined by:

$$\mathscr{K}(h,t,x,r,\phi) = -\max_{m\in M,\, u\in\mathscr{U}} \Big\{ h^{-1}(\mathscr{D}^0_{t,h,m}(\phi)(t,x) - r) + \ell^m(x,u) - \delta^m(x,u)r + \mathscr{D}^1_{t,h,m,g^m(x,u)}(r,\phi)(t,x) + \mathscr{D}^2_{t,h,m,\Sigma^m(x,u),k}(\phi)(t,x) \Big\} \quad .$$

Assume that $\underline{\sigma}^m$, $\underline{f}^m$, g^m and Σ^m are bounded maps (in x and u). Then, there exists K depending on these bounds, such that, for any $0 < \epsilon \leq 1$, $\tilde{K}$ and $v \in \mathscr{C}_b^\infty$ satisfying

$$|\partial_t^p D^q v| \leq \tilde{K}\epsilon^{1-2p-q} \quad \text{for all } p,q \in \mathbb{N} \quad , \tag{20}$$

we have,

$$|\mathscr{K}(h,t,x,v(t,x),v) + \frac{\partial v}{\partial t}(t,x) + \mathscr{H}(x,v(t,x),Dv(t,x),D^2v(t,x))| \leq E(\tilde{K},h,\epsilon) \quad ,$$

for all $t \in \mathscr{T}_h$ and $x \in \mathbb{R}^d$, with

$$E(\tilde{K},h,\epsilon) = K\tilde{K}\left(h\epsilon^{-3}(1+\sqrt{h})^4(1+\sqrt{h}\epsilon^{-1})^4 + \sqrt{h}\epsilon^{-1}(1+\sqrt{h})^2(1+\sqrt{h}\epsilon^{-1})^2\right) \quad .$$

Sketch of Proof When m and u are frozen, the expression of $\mathscr{K}(h,t,x,v(t,x),v)+\frac{\partial v}{\partial t}(t,x)+\mathscr{H}(x,v(t,x),Dv(t,x),D^2v(t,x))$ is the sum of three terms, which are precisely the ones that are bounded in Theorems 1 and 2 and Lemma 3, respectively. Using that the difference of maxima of two functions is less or equal to the maximum of the difference of these functions, and that the bounds involved in Theorems 1 and 2 and Lemma 3 can be taken uniform in m and u, since $\underline{\sigma}^m$, $\underline{f}^m$, g^m and Σ^m are bounded maps (in x and u), we get a similar bound for the true expression of $\mathscr{K}(h,t,x,v(t,x),v)+\frac{\partial v}{\partial t}(t,x)+\mathscr{H}(x,v(t,x),Dv(t,x),D^2v(t,x))$. Then using the assumption (20) on v, we deduce the result of the theorem. □

Lemma 4 *Denote*

$$\begin{aligned}T^N_{t,h,m,u}(\phi)(x) =& \mathscr{D}^0_{t,h,m}(\phi)(t,x)+h\big\{\ell^m(x,u)\\ &+\mathscr{D}^1_{t,h,m,g^m(x,u)}(0,\phi)(t,x)+\mathscr{D}^2_{t,h,m,\Sigma^m(x,u),k}(\phi)(t,x)\big\}\\ T^D_{t,h,m,u}(x) =& 1+h\delta^m(x,u)+h\mathbb{E}\left[\mathscr{P}^1_{g^m(x,u)}(h^{-1}(W_{t+h}-W_t))\right]\ .\end{aligned}$$

If $\delta^m\geq 0$, or if δ^m is lower bounded and h is small enough, then $T^D_{t,h,m,u}(x)\geq 1/2$ for all $x\in\mathbb{R}^d$ and we can define $T_{t,h}$ as:

$$T_{t,h}(\phi)(x)=\sup_{m\in M,\,u\in\mathscr{U}}\frac{T^N_{t,h,m,u}(\phi)(x)}{T^D_{t,h,m,u}(x)}\ . \tag{21}$$

Moreover, any solution v^h of the discretized equation (19), if it exists, also satisfies the iterative equation (5) and is thus unique.

Conversely, if for all t,x, there exists a constant $C_{t,x}>0$, such that the supremum in (21) is the same as the supremum restricted to the actions m,u such that $T^D_{t,h,m,u}(x)\leq C_{t,x}$, then the solution of the iterative equation (5) is the unique solution of the discretized equation (19).

The latter assumption holds in particular if δ^m is upper bounded and g^m is bounded with respect to u, in which case $C_{t,x}=1+O(\sqrt{h})$. It also holds when ℓ^m is quadratic and strictly concave in u and all other functions g^m, σ^m and δ^m are linearly growing in u.

Proof Since $\mathscr{P}^1_{g^m(x,u)}$ is nonnegative, we have $T^D_{t,h,m,u}(x)\geq 1-Ch$ where $-C$ is a lower bound of δ^m. Then, for h small enough this is $\geq 1/2$, and $T_{t,h}$ is well defined. Using the definitions of the lemma, the operator $\mathscr{K}$ of Theorem 3 can be rewritten as

$$\mathscr{K}(h,t,x,r,\phi)=-h^{-1}\max_{m\in M,\,u\in\mathscr{U}}\left(T^N_{t,h,m,u}(\phi(t+h,\cdot))(x)-T^D_{t,h,m,u}(x)r\right)\ . \tag{22}$$

Assume first that $\mathscr{K}(h,t,x,r,\phi)=0$ for some h,t,x,r,ϕ, and denote $\phi_{+h}=\phi(t+h,\cdot)$. This implies that $T^N_{t,h,m,u}(\phi_{+h})(x)-rT^D_{t,h,m,u}(x)\leq 0$ for all $m\in M$, $u\in\mathscr{U}$, hence $r\geq T^N_{t,h,m,u}(\phi_{+h})(x)/T^D_{t,h,m,u}(x)$ for all $m\in M$, $u\in\mathscr{U}$. Taking the maximum, we deduce that $r\geq T_{t,h}(\phi_{+h})(x)$. Moreover, for all $\epsilon>0$, there exists m and u such that $T^N_{t,h,m,u}(\phi_{+h})(x)-rT^D_{t,h,m,u}(x)\geq -\epsilon$ and since $T^D_{t,h,m,u}(x)\geq 1/2$, we deduce that $T^N_{t,h,m,u}(\phi_{+h})(x)/T^D_{t,h,m,u}(x)\geq r-2\epsilon$, hence $T_{t,h}(\phi_{+h})(x)\geq r-2\epsilon$. Since the latter inequality holds for all $\epsilon>0$, this shows that $r=T_{t,h}(\phi_{+h})(x)$ and so we proved that $\mathscr{K}(h,t,x,r,\phi)=0$ implies $r=T_{t,h}(\phi_{+h})(x)$. Applying this implication to a solution v^h of (19), we obtain the first assertion of the lemma.

Using the same arguments as above, and using that one can restrict the set of actions m,u so that $T^D_{t,h,m,u}(x)$ is bounded above, we obtain the reverse implication: $r=T_{t,h}(\phi_{+h})(x)$ implies $\mathscr{K}(h,t,x,r,\phi)=0$. Applying this implication to the solution v^h to the iterative equation (5), we obtain the second assertion of the lemma. □

Remark 1 Recall that in [2], we proposed the following similar but different operator

$$\widetilde{T}_{t,h}(\phi)(x)=\max_{m\in M,\ u\in\mathscr{U}}\widetilde{T}_{t,h,m,u}(\phi)(x) \tag{23a}$$

$$\widetilde{T}_{t,h,m,u}(\phi)(x)=\mathscr{D}^0_{t,h,m}(\phi)(t,x)(1-\delta^m(x,u)h)+h\big\{\ell^m(x,u) \tag{23b}$$

$$+\tilde{\mathscr{D}}^1_{t,h,m,g^m(x,u)}(\phi)(t,x)+\mathscr{D}^2_{t,h,m,\Sigma^m(x,u),k}(\phi)(t,x)\big\}\big\}, \tag{23c}$$

with

$$\tilde{\mathscr{D}}^1_{t,h,m,g}(\phi)(t,x):=\mathbb{E}\left[\phi(t+h,\hat{X}^m(t+h))g\cdot(h^{-1}(W_{t+h}-W_t))\mid\hat{X}(t)=x\right].$$

This operator coincides with the one of Lemma 4 when $\delta^m(x,u)$ and $g^m(x,u)$ are zero. It coincides with the operator (6) proposed in [7], when $k=0$, and $\mathscr{L}^m=\mathscr{L}$ does not depend on m, see [2]. Note that when $\delta^m(x,u)\neq 0$, and $g^m(x,u)$ is zero, one need to replace $-\delta^m(x,u)r$ by $-\delta^m(x,u)\mathscr{D}^0_{t,h,m}(\phi)(t,x)$ in the expression of $\mathscr{K}$ in order to recover the operators of [7] and [2]. When the sign of δ^m is not fixed or δ^m is not lower bounded, one can replace $-\delta^m(x,u)r$ by

$$-\delta^m(x,u)_+r+\delta^m(x,u)_-\mathscr{D}^0_{t,h,m}(\phi)(t,x)$$

in the expression of $\mathscr{K}$ so that in all cases, any solution of the discretized equation (19) satisfies the iterative equation (5) with $T_{t,h}$ defined by (21) and

$$\begin{aligned}T^N_{t,h,m,u}(\phi)(x)=&\mathscr{D}^0_{t,h,m}(\phi)(t,x)+h\big\{\ell^m(x,u)+\delta^m(x,u)_-\mathscr{D}^0_{t,h,m}(\phi)(t,x)\\&+\mathscr{D}^1_{t,h,m,g^m(x,u)}(0,\phi)(t,x)+\mathscr{D}^2_{t,h,m,\Sigma^m(x,u),k}(\phi)(t,x)\big\}\big\}\\T^D_{t,h,m,u}(x)=&1+h\delta^m(x,u)_++h\mathbb{E}\left[\mathscr{P}^1_{g^m(x,u)}(h^{-1}(W_{t+h}-W_t))\right].\end{aligned}$$

In [2, Theorem 1.3.3], we proved that the operator $T_{t,h}$ is monotone for h small enough over the set of bounded continuous functions $\mathbb{R}^d \to \mathbb{R}$, under the assumption that $\bar{a} < 4k+2$ with $\bar{a}$ an upper bound of $\mathrm{tr}(\Sigma^m(x,u)\Sigma^m(x,u)^\mathsf{T})$ (for all x and u) and that δ^m is upper bounded, and that there exists a bounded map $\tilde{g}^m$ such that $g^m(x,u) = \Sigma^m(x,u)\tilde{g}^m(x,u)$. This was already a generalization of [7, Lemma 3.12], since the latter corresponds to the case where $k=0$. Here, the boundedness of $\tilde{g}^m$ will not be needed, so that one can apply the result to degenerate matrices $\Sigma^m(x,u)\Sigma^m(x,u)^\mathsf{T}$. Also δ^m need not be upper bounded at this point because the expression of $\mathscr{K}$ uses $-\delta^m(x,u)r$ instead of $-\delta^m(x,u)\mathscr{D}^0_{t,h,m}(\phi)(t,x)$.

Theorem 4 *Let $\mathscr{K}$ be as in Theorem 3. Assume that the map $\mathrm{tr}(\Sigma^m(x,u)\Sigma^m(x,u)^\mathsf{T})$ is upper bounded in x and u and let $\bar{a}$ be an upper bound. Assume also that δ^m is lower bounded. Then, for k such that $\bar{a} \le 4k+2$, $\mathscr{K}$ is monotone in the sense of [3]. Also, there exists h_0 such that the operator $T_{t,h}$ of Lemma 4 is monotone for $h \le h_0$ over the set of bounded continuous functions $\mathbb{R}^d \to \mathbb{R}$.*

Proof Adapting the definition of monotonicity of [3, **(S1)**] to our setting (backward equations and a time discretization only), we need to prove that there exists $\lambda, \mu \ge 0$, $h_0 > 0$ such that if $h \le h_0$, v, v' are bounded continuous functions from $\overline{\mathscr{T}_h} \times \mathbb{R}^d$ to $\mathbb{R}$ such that $v \le v'$ and $\psi(t) = e^{\mu(T-t)}(a + b(T-t)) + c$ with $a, b, c \ge 0$, then:

$$\mathscr{K}(h,t,x,r+\psi(t),v+\psi) \ge \mathscr{K}(h,t,x,r,v') + b/2 - \lambda c \text{ in } \mathscr{T}_h \times \mathbb{R}^d \quad . \tag{24}$$

Let us first show the inequality for $\psi = 0$. Using the notations of Lemma 4, we have

$$\begin{aligned} T^N_{t,h,m,u}(\phi)(x) =& h\ell^m(x,u) \\ &+ \mathbb{E}\left[\phi(\hat{X}^m(t+h))\mathscr{P}^{h,m,u,x}(h^{-1/2}(W_{t+h}-W_t)) \mid \hat{X}^m(t) = x\right] \quad , \end{aligned}$$

where

$$\mathscr{P}^{h,m,u,x}(w) = 1 + h\mathscr{P}^1_{g^m(x,u)}(h^{-1/2}w) + \mathscr{P}^2_{\Sigma^m(x,u),k}(w) \quad .$$

Since $\mathscr{P}^1_g \ge 0$ for all g and $\mathscr{P}^2_\Sigma \ge -\frac{\mathrm{tr}(\Sigma\Sigma^\mathsf{T})}{4k+2}$ for all Σ, we get that $\mathscr{P}^{h,m,u,x}(w) \ge 1 - \frac{\bar{a}}{4k+2}$. Assume now that $\bar{a} \le 4k+2$. Then, $\mathscr{P}^{h,m,u,x}(w) \ge 0$, so if $v \le v'$, then $T^N_{t,h,m,u}(v(t+h,\cdot)) \le T^N_{t,h,m,u}(v'(t+h,\cdot))$ and by (22), $\mathscr{K}(h,t,x,r,v) \ge \mathscr{K}(h,t,x,r,v')$.

To show (24), it is now sufficient to show the same inequality for $v = v'$. We have

$$\begin{aligned} \mathscr{K}(h,t,x,r+\psi(t),v+\psi) - \mathscr{K}(h,t,x,r,v) \ge -\max_{m\in M,\, u\in\mathscr{U}} \Big\{ & h^{-1}(\psi(t+h)-\psi(t)) \\ & - \delta^m(x,u)\psi(t) + (\psi(t+h)-\psi(t))\mathbb{E}[\mathscr{P}^1_{g^m(x,u)}(h^{-1}(W_{t+h}-W_t))] \Big\} \quad . \end{aligned}$$

Let us take for λ an upper bound of $-\delta^m$. From $\psi(t+h)-\psi(t)\leq 0$, and $\mathscr{P}_g^1 \geq 0$ for all g, we deduce

$$\begin{aligned}
&\mathscr{K}(h,t,x,r+\psi(t),v+\psi)-\mathscr{K}(h,t,x,r,v)\\
&\quad\geq -h^{-1}(\psi(t+h)-\psi(t))-\lambda\psi(t)\\
&\quad= be^{\mu(T-t-h)}+e^{\mu(T-t)}(\frac{1-e^{-\mu h}}{h}-\lambda)(a+b(T-t))-\lambda c\\
&\quad\geq b-\lambda c\ ,
\end{aligned}$$

if $1-e^{-\mu h}\geq \lambda h$. Taking $\mu>\lambda$, there exists h_0 such that $1-e^{-\mu h}\geq \lambda h$ for all $h\leq h_0$, leading to the previous inequality and so to (24) for $v=v'$. This shows that $\mathscr{K}$ is monotone in the sense of [3].

Since $\mathscr{P}_g^1 \geq 0$ for all g, and $\lambda \geq -\delta^m$, we get also that $T^D_{t,h,m,u}(x)\geq 1-\lambda h$ and so $T^D_{t,h,m,u}(x)>0$ for $h\leq h_0$ if $h_0<1/\lambda$. Since we already proved that $T^N_{t,h,m,u}$ is monotone, for all m,u, we obtain that the operator $T_{t,h}$ of Lemma 4 is well defined and monotone for $h\leq h_0$ over the set of bounded continuous functions $\mathbb{R}^d\to\mathbb{R}$. □

We shall say that an operator T between any sets $\mathscr{F}$ and $\mathscr{F}'$ of partially ordered sets of real valued functions, which are stable by the addition of a constant function (identified to a real number), is *additively α-subhomogeneous* if

$$\lambda\in\mathbb{R},\lambda\geq 0,\ \phi\in\mathscr{F}\implies T(\phi+\lambda)\leq T(\phi)+\alpha\lambda\ . \tag{25}$$

Lemma 5 *Assume that δ^m is lower bounded in x and u and let $T_{t,h}$ be as in Lemma 4. Then, there exists $h_0>0$ such that for $h\leq h_0$, $T_{t,h}$ is additively α_h-subhomogeneous over the set of bounded continuous functions $\mathbb{R}^d\to\mathbb{R}$, for some constant $\alpha_h=1+Ch$ with $C\geq 0$.*

Proof If λ is an upper bound of $-\delta^m$, take $C=2\lambda$ and h_0 such that $1-\lambda h_0\geq 1/2$. □

With the monotonicity, the α_h-subhomogeneity implies the α_h-Lipschitz continuity of the operator, which allows one to show easily the stability as follows, see [2, Corollary 1.3.5] for the proof.

Corollary 1 *Let the assumptions and conclusions of Theorems 3 and 4 hold and assume also that ψ and ℓ^m are bounded. Then, there exists a unique function v^h on $\mathscr{T}_h\times\mathbb{R}^d$ satisfying (19) or equivalently (5) with $T_{t,h}$ as in Lemma 4 and $v^h(T,x)=\psi(x)$ for all $x\in\mathbb{R}^d$. Moreover v^h is bounded (independently of h).*

Note that the assumptions can be summarized in "all the maps ψ, δ^m, ℓ^m, $\underline{\sigma}^m$, $\underline{f}^m$, g^m and Σ^m are bounded". This implies that f^m and $\sigma^m(\sigma^m)^\mathsf{T}$ are bounded, and, if σ^m is symmetric then σ^m is also bounded, but we do not need this directly.

Corollary 2 *Let the assumptions and conclusions of Corollary 1 hold. Assume also that all the maps ψ, δ^m, ℓ^m, $\underline{\sigma}^m$, $\underline{f}^m$, g^m and Σ^m are continuous with respect to $x \in \mathbb{R}^d$, uniformly in x and $u \in \mathscr{U}$. Then the unique solution v^h of (19), with the initial condition $v^h(T,x) = \psi(x)$ for all $x \in \mathbb{R}^d$, is uniformly continuous on $\overline{\mathscr{T}_h} \times \mathbb{R}^d$.*

Proof Since $\overline{\mathscr{T}_h}$ is finite, we just need to show that $v^h(t,\cdot)$ is uniformly continuous on $\mathbb{R}^d$ for all $t \in \overline{\mathscr{T}_h}$. Since $v^h(T,\cdot) = \psi$ which is already bounded and uniformly continuous on $\mathbb{R}^d$, we only need to show that the operator $T_{t,h}$ of Lemma 4 sends the set of bounded and uniformly continuous functions on $\mathbb{R}^d$ to itself. From the proof of Corollary 1, it sends bounded functions to bounded functions. So, it is sufficient to show that $T^D_{t,h,m,u}$ is uniformly continuous, uniformly in $u \in \mathscr{U}$ and that $T^N_{t,h,m,u}$ sends bounded uniformly continuous functions on $\mathbb{R}^d$ to functions that are uniformly continuous in x uniformly in $u \in \mathscr{U}$. The first property is due to the uniform continuity of δ^m and g^m uniformly in $u \in \mathscr{U}$. For the second one, one uses that if $\hat{X}^m(t) = x$, then $\hat{X}^m(t+h) = x + \underline{f}^m(x)h + \underline{\sigma}^m(x)(W_{t+h} - W_t)$ which is uniformly continuous in x, for all given values of $W_{t+h} - W_t$, since $\underline{\sigma}^m$ and $\underline{f}^m$ are uniformly continuous in x. Hence, when ϕ is bounded and uniformly continuous with respect to x, then $\phi(\hat{X}^m(t+h))$ is bounded and uniformly continuous with respect to x, for all given values of $W_{t+h} - W_t$. Since all moments of $W_{t+h} - W_t$ are finite and the maps ℓ^m, g^m and Σ^m are uniformly continuous with respect to $x \in \mathbb{R}^d$, uniformly in $u \in \mathscr{U}$, we deduce that $T^N_{t,h,m,u}(\phi)$ is uniformly continuous in x, uniformly in $u \in \mathscr{U}$. □

The previous result shows that the map v^h can be extended in a continuous function over $[0,T] \times \mathbb{R}^d$. Then, the convergence of the scheme can be obtained as in [2] by applying the theorem of Barles and Souganidis [4]:

Corollary 3 *Let the assumptions of Corollary 2 hold. Assume also that (3) has a strong uniqueness property for viscosity solutions and let v be its unique viscosity solution. Let v^h be the unique solution of (19), with the initial condition $v^h(T,x) = \psi(x)$ for all $x \in \mathbb{R}^d$. Let us extend v^h on $[0,T] \times \mathbb{R}^d$ as a continuous and piecewise linear function with respect to t. Then, when $h \to 0^+$, v^h converges to v locally uniformly in $t \in [0,T]$ and $x \in \mathbb{R}^d$.*

To apply the theorem of Barles and Jakobsen [3], we also need the following regularity result (corresponding to **(S2)** in [3]) which is comparable to the previous one.

Lemma 6 *Let the assumptions of Corollary 2 hold. Assume also that δ^m is bounded. Then, for all continuous and bounded function v on $\overline{\mathscr{T}_h} \times \mathbb{R}^d$, the function $(t,x) \mapsto \mathscr{K}(h,t,x,v(t,x),v)$ is bounded and continuous in $\mathscr{T}_h \times \mathbb{R}^d$. Moreover, the function $r \mapsto \mathscr{K}(h,t,x,r,v)$ is uniformly continuous for bounded r, uniformly in $(t,x) \in \mathscr{T}_h \times \mathbb{R}^d$.*

Proof Using the arguments of the proof of Corollary 2 and the rewriting of $\mathscr{K}$ in (22), one gets that $x \mapsto \mathscr{K}(h,t,x,r,v)$ is continuous in x, uniformly in r

bounded. Also since δ^m and g^m are bounded, then $T^D_{t,h,m,u}$ is bounded, so $r \mapsto \mathscr{K}(h,t,x,r,v)$ is uniformly continuous in r bounded, uniformly in $x \in \mathbb{R}^d$. Also, since v is bounded and continuous, this implies that $x \mapsto \mathscr{K}(h,t,x,v(t,x),v)$ is bounded and continuous in $\mathbb{R}^d$. Since $\mathscr{T}_h$ is a finite set, the assertions of the lemma follow. □

We also need the following assumptions which correspond to the assumptions with same names in [3].

For a function v defined on $\mathbb{R}^d$, $|v|_0$ and $|v|_1$ will denote respectively the norm on the space of bounded functions (that is the sup-norm) and the norm on the space of bounded Lipschitz continuous functions on $\mathbb{R}^d$ (that is the sup-norm plus the minimal Lipschitz constant). More generally, for a function defined on $Q = [0,T]\times\mathbb{R}^d$, $|v|_0$ will denote the sup-norm, while $|v|_1$ will denote a norm on the space of bounded functions that are Lipschitz continuous with respect to x and $1/2$-Hölder continuous with respect to t:

$$|v|_0 = \sup_{(t,x)\in Q} |v(t,x)| \ , \quad |v|_1 = |v|_0 + \sup_{\substack{(t,x)\in Q\\(t',x')\in Q'\\(t,x)\neq(t',x')}} \frac{|v(t',x')-v(t,x)|}{(t'-t)^{1/2}+|x'-x|} \ .$$

(A1) There exists a constant $K > 0$, such that

$$|\phi|_1 \le K$$

for $\phi = \psi$ and for all the maps $\phi = h(\cdot,u)$ with h being any coordinate of the maps $f^m, \sigma^m, \delta^m, \ell^m$, and any $m \in \mathscr{M}$ and $u \in \mathscr{U}$.

(A2) For every $\delta > 0$, there is a finite subset $\mathscr{U}_F$ of $\mathscr{U}$ such that for any $u \in \mathscr{U}$, there exists $u_F \in \mathscr{U}_F$ such that

$$|h(\cdot,u) - h(\cdot,u_F)|_0 \le \delta$$

for all the maps h being any coordinate of the maps $f^m, \sigma^m, \delta^m, \ell^m$, and any $m \in \mathscr{M}$.

Applying [3, Theorem 3.1], we obtain the following estimations which are of the same order as the ones obtained for usual explicit finite difference schemes with Δx in the order of $\sqrt{h}$ [3] or for the scheme of [7].

Corollary 4 *Let the assumptions of Lemma 6 hold. Assume also* **(A1)** *and* **(A2)**. *Let v be the unique viscosity solution of (3) and v^h be the unique solution of (19), with the initial condition $v^h(T,x) = \psi(x)$ for all $x \in \mathbb{R}^d$. Then, there exists C_1, C_2 depending on $|v|_1$ such that, for all $(t,x) \in \overline{\mathscr{T}_h} \times \mathbb{R}^d$, we have*

$$-C_1 h^{1/10} \le (v^h - v)(t,x) \le C_2 h^{1/4} \ .$$

5 The Probabilistic Max-Plus Method

In [7], the solution v^h of the time discretization (5) of the partial differential equation (3) is obtained by using the following method which can be compared to a space discretization. The conditional expectations in (7) are approximated by any probabilistic method such as a regression estimator: after a simulation of the processes W_t and $\hat{X}(t)$, one applies at each time $t \in \mathscr{T}_h$ a regression estimation to find the value of $\mathscr{D}^i_{t,h}(v^h(t+h,\cdot))$ at the points $\hat{X}(t)$ by using the values of $v^h(t+h, \hat{X}(t+h))$ and $W_{t+h} - W_t$. The regression can be done over a finite dimensional linear space approximating the space of bounded Lipschitz continuous functions, for instance the linear space of functions that are polynomial with a certain degree on some "finite elements". Hence, the value function $v^h(t,\cdot)$ is obtained by an estimation of it at the simulated points $\hat{X}(t)$. This method can also be used for the scheme (5) obtained in the previous section, since the new one also involves conditional expectations.

In the probabilistic max-plus method proposed in [1] and used in [2], the aim was to replace the (large) finite dimensional linear space of functions used in the regression estimations by the max-plus linear space of max-plus linear combinations of functions that belong to a small dimensional linear space (such as the space of quadratic forms). The idea is that stochastic control problems involve at the same time an expectation which is a linear operation and a maximization which is a max-plus linear operation. Note that a direct regression estimation on such a non linear space is difficult. We rather used the distributivity property of monotone operators over suprema operations, recalled in Theorem 5 below, a property which generalizes the one shown in Theorem 3.1 of McEneaney et al. [14], see also [11]. This allowed us to reduce the regression estimations to the small dimensional linear space of quadratic forms.

The algorithm of [1] was based on the scheme of [7], that is (5) with $T_{t,h}$ as in (6). The one of [2] was based on (5) with $\widetilde{T}_{t,h}$ of (23) instead of $T_{t,h}$ and k large enough in such a way that the scheme is monotone. Here, we shall explain how the algorithm can be adapted to the case of the discretization of Theorem 3, that is to $T_{t,h}$ as in (21).

In the sequel, we denote $\mathscr{W} = \mathbb{R}^d$ and $\mathscr{D}$ the set of measurable functions from $\mathscr{W}$ to $\mathbb{R}$ with at most some given growth or growth rate (for instance with at most exponential growth rate), assuming that it contains the constant functions.

Theorem 5 ([1, Theorem 4]) *Let G be a monotone additively α-subhomogeneous operator from $\mathscr{D}$ to $\mathbb{R}$, for some constant $\alpha > 0$. Let $(Z, \mathfrak{A})$ be a measurable space, and let $\mathscr{W}$ be endowed with its Borel σ-algebra. Let $\phi : \mathscr{W} \times Z \to \mathbb{R}$ be a measurable map such that for all $z \in Z$, $\phi(\cdot, z)$ is continuous and belongs to $\mathscr{D}$. Let $v \in \mathscr{D}$ be such that $v(W) = \sup_{z\in Z} \phi(W, z)$. Assume that v is continuous and bounded. Then,*

$$G(v) = \sup_{\bar{z}\in\overline{Z}} G(\bar{\phi}^{\bar{z}})$$

where $\bar{\phi}^{\bar{z}} : \mathscr{W} \to \mathbb{R},\ W \mapsto \phi(W, \bar{z}(W))$, *and*

$$\overline{Z} = \{\bar{z} : \mathscr{W} \to Z,\ \textit{measurable and such that } \bar{\phi}^{\bar{z}} \in \mathscr{D}\}.$$

To explain the algorithm, assume that the final reward ψ of the control problem can be written as the supremum of a finite number of quadratic forms. Denote $\mathscr{Q}_d = \mathbb{S}_d \times \mathbb{R}^d \times \mathbb{R}$ (recall that $\mathbb{S}_d$ is the set of symmetric $d \times d$ matrices) and let

$$q(x, z) := \frac{1}{2} x^{\mathsf{T}} Q x + b \cdot x + c, \quad \text{with } z = (Q, b, c) \in \mathscr{Q}_d \ , \tag{26}$$

be the quadratic form with parameter z applied to the vector $x \in \mathbb{R}^d$. Then for $g_T = q$, we have

$$v^h(T, x) = \psi(x) = \sup_{z \in Z_T} g_T(x, z)$$

where Z_T is a finite subset of $\mathscr{Q}_d$.

The application of the operator $T_{t,h}$ of Lemma 4 to a (continuous) function $\phi : \mathbb{R}^d \to \mathbb{R}, x \mapsto \phi(x)$ can be written, for each $x \in \mathbb{R}^d$, as

$$T_{t,h}(\phi)(x) = \max_{m \in \mathscr{M}} G^m_{t,h,x}(\tilde{\phi}^m_{t,h,x}) \ , \tag{27a}$$

where

$$S^m_{t,h} : \mathbb{R}^d \times \mathscr{W} \to \mathbb{R}^d,\ (x, W) \mapsto S^m_{t,h}(x, W) = x + \underline{f}^m(x)h + \underline{\sigma}^m(x)W \ , \tag{27b}$$

$$\tilde{\phi}^m_{t,h,x} = \phi(S^m_{t,h}(x, \cdot)) \in \mathscr{D} \quad \text{if } \phi \in \mathscr{D} \ , \tag{27c}$$

and $G^m_{t,h,x}$ is the operator from $\mathscr{D}$ to $\mathbb{R}$ given by

$$G^m_{t,h,x}(\tilde{\phi}) = \max_{u \in \mathscr{U}} \frac{G^N_{t,h,x,m,u}(\tilde{\phi})}{T^D_{t,h,m,u}(x)} \ , \tag{28}$$

with

$$\begin{aligned}
G^N_{t,h,x,m,u}(\tilde{\phi}) &= D^0_{t,h}(\tilde{\phi}) + h\big\{\ell^m(x, u) + D^1_{t,h,g^m(x,u)}(\tilde{\phi}) + D^2_{t,h,\Sigma^m(x,u),k}(\tilde{\phi})\big\}, \quad (29)\\
D^0_{t,h}(\tilde{\phi}) &= \mathbb{E}(\tilde{\phi}(W_{t+h} - W_t)) \ ,\\
D^1_{t,h,g}(\tilde{\phi}) &= \mathbb{E}(\tilde{\phi}(W_{t+h} - W_t)\mathscr{P}^1_g(h^{-1}(W_{t+h} - W_t)) \ ,\\
D^2_{t,h,\Sigma,k}(\tilde{\phi})(x) &= h^{-1}\mathbb{E}\left[\tilde{\phi}(W_{t+h} - W_t)\mathscr{P}^2_{\Sigma,k}(h^{-1/2}(W_{t+h} - W_t))\right] \ ,
\end{aligned}$$

$g^m(x,u)$ and $\Sigma^m(x,u)$, as in Sect. 4, and $\mathscr{P}^1_g$ and $\mathscr{P}^2_{\Sigma,k}$ as in (14) and (11) respectively. Indeed, the Euler discretization $\hat{X}^m$ of the diffusion with generator $\mathscr{L}^m$ satisfies

$$\hat{X}^m(t+h) = S^m_{t,h}(\hat{X}^m(t), W_{t+h} - W_t) \ . \tag{30}$$

Recall from [2] that we can also write the operator $\widetilde{T}_{t,h}$ of (23) in a similar way:

$$\widetilde{T}_{t,h}(\phi)(x) = \max_{m\in\mathscr{M}} \widetilde{G}^m_{t,h,x}(\tilde{\phi}^m_{t,h,x}) \ , \tag{31}$$

with

$$\widetilde{G}^m_{t,h,x}(\tilde{\phi}) = \max_{u\in\mathscr{U}} \widetilde{G}_{t,h,x,m,u}(\tilde{\phi}) \ , \tag{32}$$

$$\begin{aligned}\widetilde{G}_{t,h,x,m,u}(\tilde{\phi}) &= D^0_{t,h}(\tilde{\phi})(1-\delta^m(x,u)h) \\ &\quad + h\big\{\ell^m(x,u) + \widetilde{D}^1_{t,h,g^m(x,u)}(\tilde{\phi}) + D^2_{t,h,\Sigma^m(x,u),k}(\tilde{\phi})\big\} \ , \\ \widetilde{D}^1_{t,h,g}(\tilde{\phi}) &= \mathbb{E}(\tilde{\phi}(W_{t+h}-W_t)g\cdot(h^{-1}(W_{t+h}-W_t)) \ .\end{aligned} \tag{33}$$

Using the same arguments as for Theorem 4 and Lemma 5, one can obtain the stronger property that for $h \le h_0$, all the operators $G^m_{t,h,x}$ belong to the class of monotone additively α_h-subhomogeneous operators from $\mathscr{D}$ to $\mathbb{R}$. This allows us to apply Theorem 5. In [2], we shown the following result, see also [1, Theorem 2] concerning $\widetilde{T}_{t,h}$.

Theorem 6 ([2, Theorem 1.4.2], Compare with [11, 14, Theorem 5.1]) *Consider the control problem of Sect. 1. Assume that $\mathscr{U} = \mathbb{R}^p$ and that for each $m \in \mathscr{M}$, δ^m and σ^m are constant, σ^m is nonsingular, f^m is affine with respect to (x,u), ℓ^m is quadratic with respect to (x,u) and strictly concave with respect to u, and that ψ is the supremum of a finite number of quadratic forms. Consider the scheme (5), with $\widetilde{T}_{t,h}$ of (23) instead of $T_{t,h}$, $\underline{\sigma}^m$ constant and nonsingular, Σ^m constant and nonsingular and $\underline{f}^m$ affine. Assume that the operators $\widetilde{G}^m_{t,h,x}$ of (32) belong to the class of monotone additively α_h-subhomogeneous operators from $\mathscr{D}$ to $\mathbb{R}$, for some constant $\alpha_h = 1 + Ch$ with $C \ge 0$. Assume also that the value function v^h of (5) belongs to $\mathscr{D}$ and is locally Lipschitz continuous with respect to x. Then, for all $t \in \mathscr{T}_h$, there exists a set Z_t and a map $g_t : \mathbb{R}^d \times Z_t \to \mathbb{R}$ such that for all $z \in Z_t$, $g_t(\cdot, z)$ is a quadratic form and*

$$v^h(t,x) = \sup_{z\in Z_t} g_t(x,z) \ . \tag{34}$$

Moreover, the sets Z_t satisfy $Z_t = \mathscr{M} \times \{\bar{z}_{t+h} : \mathscr{W} \to Z_{t+h} \mid \text{Borel measurable}\}$.

Theorem 6 uses Theorem 5 together with the property that, for each m, the operator $\widetilde{T}^m_{t,h}$, such that $\widetilde{T}^m_{t,h}(\phi)(x) = \widetilde{G}^m_{t,h,x}(\tilde{\phi}^m_{t,h,x})$, sends a random quadratic form that is upper bounded by a deterministic quadratic form into a quadratic form. This means that if $\bar{z}$ is a measurable function from $\mathscr{W}$ to $\mathscr{Q}_d$ and $Z \in \mathscr{Q}_d$ is such that $q(x, \bar{z}(W)) \leq q(x, Z)$ for all $x \in \mathbb{R}^d$ and $W \in \mathscr{W}$, and $\tilde{q}_x$ denotes the measurable map $\mathscr{W} \to \mathbb{R}$, $W \mapsto q(S^m_{t,h}(x, W), \bar{z}(W))$, where q is as in (26), then the function $x \mapsto \widetilde{G}^m_{t,h,x}(\tilde{q}_x)$ is a quadratic form, that is it can be written as $q(x, z)$ for some $z \in \mathscr{Q}_d$, see [2, Lemma 1.4.3].

If we replace the operator $\widetilde{T}_{t,h}$ by $T_{t,h}$ of Lemma 4, the previous property does not hold because of the expressions g^+ and g^- and so one cannot deduce directly a result like Theorem 6. However, one can still obtain the following result:

Lemma 7 *Let us consider the notations and assumptions of Theorem 6, except that $\widetilde{T}_{t,h}$ is replaced by the operator $T_{t,h}$ of Lemma 4. For each m, consider the operator $T^m_{t,h}$ such that $T^m_{t,h}(\phi)(x) = G^m_{t,h,x}(\tilde{\phi}^m_{t,h,x})$ with $G^m_{t,h,x}$ as in (28). Let $\bar{z}$ be a measurable function from $\mathscr{W}$ to $\mathscr{Q}_d$ and $Z \in \mathscr{Q}_d$ be such that $q(x, \bar{z}(W)) \leq q(x, Z)$ for all $x \in \mathbb{R}^d$ and $W \in \mathscr{W}$, where q is as in (26). Let $\tilde{q}_x$ be the map $\mathscr{W} \to \mathbb{R}$, $W \mapsto q(S^m_{t,h}(x, W), \bar{z}(W))$. Then, the function $\bar{q} : x \mapsto G^m_{t,x,h}(\tilde{q}_x)$ is upper bounded by a quadratic map. The same property holds for lower bounds.*

Moreover, there exists $C > 0$, independent of h such that if the map $\bar{z}$ is constant, that is deterministic, and $\|\bar{z}\| \leq K$ for some norm on $\mathscr{Q}_d$, then there exists $z \in \mathscr{Q}_d$ such that $\|z - \bar{z}\| \leq C(K+1)^2 h$ and

$$|\bar{q}(x) - q(x, z)| \leq C(K+1)^3 h^{3/2}(\|x\|^2 + 1)^{3/2}, \quad \textit{for all } x \in \mathbb{R}^d \ .$$

Proof Let $\bar{z}$, $Z \in \mathscr{Q}_d$, $\tilde{q}_x$ and $\bar{q}$ be as in the first part of the lemma. Consider the map $\phi(x) = q(x, Z)$. It satisfies $-CK(1 + \|x\|^2) \leq \phi(x) \leq CK(1 + \|x\|^2)$ as soon as $\|Z\| \leq K$. Here and below $\|\cdot\|$ denotes a norm on $\mathscr{Q}_d$ and C is any positive constant independent of $h \leq 1$. Since $G^m_{t,h,x}$ is monotone, we get that $\bar{q}(x) = G^m_{t,x,h}(\tilde{q}_x) \leq G^m_{t,h,x}(\tilde{\phi}^m_{t,h,x}) = T^m_{t,h}(\phi)(x)$ and a similar result holds for a lower bound. Due to the assumptions on the parameters of the problem, it is easy to show that for any (deterministic) quadratic form ϕ, $\widetilde{T}^m_{t,h}(\phi)$ is a quadratic form, for $\widetilde{T}^m_{t,h}$ defined as in (23), see below. Hence, to obtain the two assertions of the lemma, it is sufficient to show that, for any quadratic form ϕ with norm K, $T^m_{t,h}(\phi)$ is bounded above and below by quadratic forms, the norm of which depend on K, $\widetilde{T}^m_{t,h}(\phi)$ is a quadratic form such that the norm of its difference with ϕ is bounded by $C(K+1)^2 h$, and that we have

$$|T^m_{t,h}(\phi)(x) - \widetilde{T}^m_{t,h}(\phi)(x)| \leq C(K+1)^3 h^{3/2}(\|x\|^2 + 1)^{3/2}, \quad \text{for all } x \in \mathbb{R}^d \ .$$

Using that $\mathscr{P}_g^1(W) = g \cdot W + |g| \cdot (|W| - \mathbb{E}(|W|) + |g| \cdot \mathbb{E}(|W|)$, we deduce

$$\frac{G^N_{t,h,x,m,u}(\tilde{\phi})}{T^D_{t,h,m,u}(x)} - D^0_{t,h}(\tilde{\phi}) = \frac{\widetilde{G}_{t,h,x,m,u}(\tilde{\phi}) - D^0_{t,h}(\tilde{\phi}) + R_{t,h,g^m(x,u)}(\tilde{\phi})}{T^D_{t,h,m,u}(x)} \quad , \tag{35}$$

where

$$R_{t,h,g}(\tilde{\phi}) = \mathbb{E}\left[\tilde{\phi}(W^t_h)|g| \cdot (|W^t_h| - \mathbb{E}(|W^t_h|))\right] \quad \text{with } W^t_h = W_{t+h} - W_t \ .$$

Due to the assumptions on the coefficients and on the scheme, $S^m_{t,h}(x, W)$ is affine with respect to (x, W), Σ^m is constant and nonsingular, g^m is affine in (x, u), and δ^m is constant. Hence the map $\tilde{\phi}^m_{t,h,x}(W)$ is a quadratic function of (x, W). Applying expectations with appropriate factors, we obtain that $D^0_{t,h}(\tilde{\phi}^m_{t,h,x})$ is a quadratic form, such that the norm of $D^0_{t,h}(\tilde{\phi}^m_{t,h,x}) - \phi(x)$ is bounded by CKh, and that $D^2_{t,h,\Sigma^m(x,u),k}(\tilde{\phi}^m_{t,h,x})$ is a constant (in x and u) which can be bounded by CK.

Since the coordinates of W^t_h are independent and with zero expectation, we also get that the first order term $\widetilde{D}^1_{t,h,g^m(x,u)}(\tilde{\phi}^m_{t,h,x})$ in (33) is equal to the scalar product of $g^m(x, u)$, which is affine in (x, u), with an affine function of x, the norm of which is bounded by CK. We deduce that

$$\widetilde{G}_{t,h,x,m,u}(\tilde{\phi}^m_{t,h,x}) - D^0_{t,h}(\tilde{\phi}^m_{t,h,x}) = h(\ell^m(x, u) + \Psi(x, u)) \ , \tag{36}$$

where Ψ is quadratic in x and u with second order derivatives in u equal to 0, that $D^0_{t,h}(\tilde{\phi}^m_{t,h,x})$ is quadratic in x, and that their norms are bounded by CK. Taking the supremum with respect to u in the previous expression, we deduce that $\widetilde{T}^m_{t,h}(\phi) = \widetilde{G}^m_{t,h,x}(\tilde{\phi}^m_{t,h,x})$ is quadratic in x, and that the norm of its difference with ϕ is bounded by $C(K + 1)^2 h$.

Since g^m has linear growth, $|R_{t,h,g^n(x,u)}(\tilde{\phi}^m_{t,h,x})| \leq C(1 + \|u\| + \|x\|)\|\mathbb{E}\left[\tilde{\phi}^m_{t,h,x}(W^t_h)(|W^t_h| - \mathbb{E}(|W^t_h|))\right]\|$. Again due to the properties of $\tilde{\phi}^m_{t,h,x}$ and W^t_h, we get that the second factor in the former inequality is constant and is bounded by $CKh^{3/2}$.

Altogether, we obtain

$$\frac{G^N_{t,h,x,m,u}(\tilde{\phi}^m_{t,h,x})}{T^D_{t,h,m,u}(x)} - D^0_{t,h}(\tilde{\phi}^m_{t,h,x}) \leq \frac{h(\ell^m(x, u) + \Psi(x, u)) + CKh^{3/2}(1 + \|u\| + \|x\|)}{T^D_{t,h,m,u}(x)} \ .$$

Then, using $CKh^{1/2}\|u\| \leq \|u\|^2\epsilon/2 + C^2K^2h/(2\epsilon)$, for $\epsilon > 0$ small enough, and similarly for $\|x\|$, and using that $T^D_{t,h,m,u}(x) \geq 1 + h\delta^m(x, u) \geq 1/2$ for h small enough, we deduce that the right hand side of the above inequality is bounded above by a quadratic form in x, so does the supremum with respect to

u of the left hand side. Since $D^0_{t,h}(\tilde\phi^m_{t,h,x})$ is a quadratic form, we deduce that $G^m_{t,h,x}(\tilde\phi^m_{t,h,x}) = T^m_{t,h}(\phi)(x)$ is bounded above by a quadratic form. Moreover the norm of this quadratic form is bounded by $K + C(K+1)^2 h$. A similar lower bound is obtained with the same arguments.

To obtain the second assertion of the lemma, we shall use the following equation

$$\frac{G^N_{t,h,x,m,u}(\tilde\phi)}{T^D_{t,h,m,u}(x)} - \widetilde G^m_{t,h,x}(\tilde\phi) = \frac{\widetilde G_{t,h,x,m,u}(\tilde\phi) - \widetilde G^m_{t,h,x}(\tilde\phi) + \widetilde R_{t,h,g^m(x,u)}(\tilde\phi)}{T^D_{t,h,m,u}(x)} \ , \tag{37}$$

where

$$\widetilde R_{t,h,g^m(x,u)}(\tilde\phi) = (D^0_{t,h}(\tilde\phi) - \widetilde G^m_{t,h,x}(\tilde\phi))(T^D_{t,h,m,u}(x) - 1) + R_{t,h,g^m(x,u)}(\tilde\phi) \ .$$

Using (36), we get that for $\tilde\phi = \tilde\phi^m_{t,h,x}$, $\widetilde G_{t,h,x,m,u}(\tilde\phi) - \widetilde G^m_{t,h,x}(\tilde\phi)$ can be written in the form $-h(u-L(x))^T Q(u-L(x))$ where L is affine with a norm bounded by CK and Q is a positive definite matrix, independent of K. Hence, there exists $\beta > 0$ such that $\widetilde G_{t,h,x,m,u}(\tilde\phi) - \widetilde G^m_{t,h,x}(\tilde\phi)$ is bounded above by $-\beta h\|u - L(x)\|^2$. Using (36) again, we obtain that $D^0_{t,h}(\tilde\phi) - \widetilde G^m_{t,h,x}(\tilde\phi)$ is a quadratic form, the norm of which is bounded by $C(K+1)^2 h$. Moreover, $T^D_{t,h,m,u}(x) - 1 = \delta^m h + C\sqrt{h}\|g^m(x,u\|$ for some norm (the 1-norm) on $\mathbb{R}^d$ and $|R_{t,h,g^m(x,u)}(\tilde\phi)| \le CKh^{3/2}\|g^m(x,u)\|$. We deduce that $|\widetilde R_{t,h,g^m(x,u)}(\tilde\phi)| \le Ch^2(K+1)^2(1+\|x\|^2) + C(K+1)^2h^{3/2}(1+\|x\|^2)\|g^m(x,u)\|$. Then, using that $T^D_{t,h,m,u}(x) \ge 1 + h\delta^m(x,u) \ge 1/2$ for h small enough, and that $y \mapsto y/(a+y)$ is increasing with respect to $y > 0$, for any $a > 0$, we obtain

$$\begin{aligned}\frac{|\widetilde R_{t,h,g^m(x,u)}(\tilde\phi)|}{T^D_{t,h,m,u}(x)} &\le Ch^2(K+1)^2(1+\|x\|^2) + \frac{C(K+1)^2h^{3/2}(1+\|x\|^2)\|g^m(x,u)\|}{1+C\sqrt{h}\|g^m(x,u\|} \\ &\le Ch^2(K+1)^2(1+\|x\|^2) + \frac{C(K+1)^2h(1+\|x\|^2)A(x,u)}{1+A(x,u)} \ ,\end{aligned}$$

for any bound $A(x,u)$ of $h^{1/2}\|g^m(x,u)\|$. Since $\|g^m(x,u\| \le C(K+1)(1+\|x\|) + \|u - L(x)\|$, we can take $A(x,u) = C(K+1)h^{1/2}(1+\|x\|^2)^{1/2} + \frac{\epsilon}{2C(K+1)^2h(1+\|x\|^2)}\|u-L(x)\|^2 + \frac{C(K+1)^2h^2(1+\|x\|^2)}{2\epsilon}$ for any $\epsilon > 0$. Then, bounding above separately the three terms of the sum in $A(x,u)/(1+A(x,u))$ by lower bounding $1 + A(x,u)$, and using the same upper bound $A(x,u)$ of $h^{1/2}\|g^m(x,u)\|$ in the expression of the first summand in (37), and that $\widetilde G_{t,h,x,m,u}(\tilde\phi) - \widetilde G^m_{t,h,x}(\tilde\phi) \le -\beta h\|u - L(x)\|^2 \le 0$, we deduce for $\epsilon = 2\beta h/C$:

$$\frac{G^N_{t,h,x,m,u}(\tilde\phi)}{T^D_{t,h,m,u}(x)} - \widetilde G^m_{t,h,x}(\tilde\phi) \le C(K+1)^3[h(1+\|x\|^2)]^{3/2} \ .$$

Then, taking the supremum over u, we obtain

$$G^m_{t,h,x}(\tilde{\phi}) - \widetilde{G}^m_{t,h,x}(\tilde{\phi}) \leq C(K+1)^3[h(1+\|x\|^2)]^{3/2} \ .$$

For the reverse inequality, using that $\widetilde{G}_{t,h,x,m,u}(\tilde{\phi}) - \widetilde{G}^m_{t,h,x}(\tilde{\phi}) = 0$ for $u = L(x)$, and applying the above bound of $\widetilde{R}_{t,h,g^m(x,u)}(\tilde{\phi})$ to $u = L(x)$, we get directly that

$$G^m_{t,h,x}(\tilde{\phi}) - \widetilde{G}^m_{t,h,x}(\tilde{\phi}) \geq -C(K+1)^3[h(1+\|x\|^2)]^{3/2} \ . \qquad \square$$

If the last inequality of Lemma 7 were true for random maps $\bar{z}$, then one may expect to obtain Eq. (34) of Theorem 6 up to an error in $O(\sqrt{h}(1+\|x\|^2)^{3/2})$. Note that, for this bound to be true, one would also need to show the following Lipschitz property for $T_{t,h}$: if $\phi(x) - \phi'(x) \leq K(1+\|x\|^2)^{3/2}$ for all $x \in \mathbb{R}^d$, and ϕ and ϕ' have a given quadratic growth, then $T_{t,h}(\phi)(x) - T_{t,h}(\phi')(x) \leq (1+Ch)K(1+\|x\|^2)^{3/2}$ for all $x \in \mathbb{R}^d$. Such an estimation would justify rigorously the application of the same algorithm as in [2] for the operator of Lemma 4, that we recall below for completeness. Recall that in the same spirit as in [7], we proposed in [1] and [2] to compute the expression of the maps $v^h(t,\cdot)$ by using simulations of the processes $\hat{X}^m$. These simulations are not only used for regression estimations of conditional expectations, which are computed there only in the case of random quadratic forms, by optimizing over the set of quadratic forms, but they are also used to fix the "discretization points" x at which the optimal quadratic forms in the expression (34) are computed. Since the last inequality of Lemma 7 is not true in general for random maps $\bar{z}$, the regression estimations may not give a good approximation of the expectation, which may lead to a large final error of the algorithm. However, we still expect the algorithm to give a good estimate of the value function, with a relatively small complexity.

Algorithm 1 ([2, Algorithm 1.4.4])
Input: A constant ϵ giving the precision, a time step h and a horizon time T such that T/h is an integer, a 3-tuple $N = (N_{\text{in}}, N_x, N_w)$ of integers giving the numbers of samples, such that $N_x \leq N_{\text{in}}$, a subset $\overline{\mathscr{M}} \subset \mathscr{M}$ and a projection map $\pi : \mathscr{M} \to \overline{\mathscr{M}}$. A finite subset Z_T of $\mathscr{Q}_d$ such that $|\psi(x) - \max_{z \in Z_T} q(x,z)| \leq \epsilon$, for all $x \in \mathbb{R}^d$, and $\#Z_T \leq \#\overline{\mathscr{M}} \times N_{\text{in}}$. The operators $T_{t,h}$, $S^m_{t,h}$ and $G^m_{t,x,h}$ as in (27)–(28) for $t \in \mathscr{T}_h$ and $m \in \mathscr{M}$, with $\mathscr{L}^m$ (and thus $S^m_{t,h}$) depending only on $\pi(m)$.
Output: The subsets Z_t of $\mathscr{Q}_d$, for $t \in \mathscr{T}_h \cup \{T\}$, and the approximate value function $v^{h,N} : (\mathscr{T}_h \cup \{T\}) \times \mathbb{R}^d \to \mathbb{R}$.

- *Initialization:* Let $\hat{X}^m(0) = \hat{X}(0)$, for all $m \in \overline{\mathscr{M}}$, where $\hat{X}(0)$ is random and independent of the Brownian process. Consider a sample of $(\hat{X}(0), (W_{t+h} - W_t)_{t \in \mathscr{T}_h})$ of size N_{in} indexed by $\omega \in \Omega_{N_{\text{in}}} := \{1, \ldots, N_{\text{in}}\}$, and denote, for each $t \in \mathscr{T}_h \cup \{T\}$, $\omega \in \Omega_{N_{\text{in}}}$, and $m \in \overline{\mathscr{M}}$, $\hat{X}^m(t,\omega)$ the value of $\hat{X}^m(t)$ induced by this sample satisfying (30). Define the function $v^{h,N}(T,\cdot)$ by $v^{h,N}(T,x) = \max_{z \in Z_T} q(x,z)$, for $x \in \mathbb{R}^d$, with q as in (26).

- For $t = T - h, T - 2h, \dots, 0$ apply the following three steps:

  (1) Choose a random sampling $\omega_{i,1}$, $i = 1, \dots, N_x$ among the elements of $\Omega_{N_{\text{in}}}$ and independently a random sampling $\omega'_{1,j}$ $j = 1, \dots, N_w$ among the elements of $\Omega_{N_{\text{in}}}$, then take the product of samplings, that is consider $\omega_{(i,j)} = \omega_{i,1}$ and $\omega'_{(i,j)} = \omega'_{1,j}$ for all i and j, leading to $(\omega_\ell, \omega'_\ell)$ for $\ell \in \Omega_{N_{\text{rg}}} := \{1, \dots, N_x\} \times \{1, \dots, N_w\}$.

  Induce the sample $\hat{X}^m(t, \omega_\ell)$ (resp. $(W_{t+h} - W_t)(\omega'_\ell)$) for $\ell \in \Omega_{N_{\text{rg}}}$ of $\hat{X}^m(t)$ with $m \in \overline{\mathscr{M}}$ (resp. $W_{t+h} - W_t$). Denote by $\mathscr{W}_t^N \subset \mathscr{W}$ the set of $(W_{t+h} - W_t)(\omega'_\ell)$ for $\ell \in \Omega_{N_{\text{rg}}}$.

  (2) For each $\omega \in \Omega_{N_{\text{in}}}$ and $m \in \overline{\mathscr{M}}$, denote $x_t = \hat{X}^m(t, \omega)$ and construct $z_t \in \mathscr{Q}_d$ depending on ω and m as follows:

  (a) Choose $\bar{z}_{t+h} : \mathscr{W}_t^N \to Z_{t+h} \subset \mathscr{Q}_d$ such that, for all $\ell \in \Omega_{N_{\text{rg}}}$, we have

$$v^{h,N}(t + h, S^m_{t,h}(x_t, (W_{t+h} - W_t)(\omega'_\ell))) = q\left(S^m_{t,h}(x_t, (W_{t+h} - W_t)(\omega'_\ell)), \bar{z}_{t+h}((W_{t+h} - W_t)(\omega'_\ell))\right) .$$

  Extend $\bar{z}_{t+h}$ as a measurable map from $\mathscr{W}$ to $\mathscr{Q}_d$. Let $\tilde{q}_{t,h,x}$ be the element of $\mathscr{D}$ given by $W \in \mathscr{W} \mapsto q(S^m_{t,h}(x, W), \bar{z}_{t+h}(W))$.

  (b) For each $\bar{m} \in \mathscr{M}$ such that $\pi(\bar{m}) = m$, compute an approximation of $x \mapsto G^{\bar{m}}_{t,h,x}(\tilde{q}_{t,h,x})$ by a linear regression estimation on the set of quadratic forms using the sample $(\hat{X}^m(t, \omega_\ell), (W_{t+h} - W_t)(\omega'_\ell))$, with $\ell \in \Omega_{N_{\text{rg}}}$, and denote by $z_t^{\bar{m}} \in \mathscr{Q}_d$ the parameter of the resulting quadratic form.

  (c) Choose $z_t \in \mathscr{Q}_d$ optimal among the $z_t^{\bar{m}} \in \mathscr{Q}_d$ at the point x_t, that is such that $q(x_t, z_t) = \max_{\pi(\bar{m})=m} q(x_t, z_t^{\bar{m}})$.

  (3) Denote by Z_t the set of all the $z_t \in \mathscr{Q}_d$ obtained in this way, and define the function $v^{h,N}(t, \cdot)$ by

$$v^{h,N}(t, x) = \max_{z \in Z_t} q(x, z) \quad \forall x \in \mathbb{R}^d .$$

Recall that no computation is done at Step (3), which gives only a formula to be able to compute the value function at each time step and state x by using the sets Z_t.

Contrarily to what happened in [2], the map $x \mapsto G^{\bar{m}}_{t,h,x}(\tilde{q}_{t,h,x})$ is not necessarily a quadratic form. If the last inequality of Lemma 7 were true for that map, then for x in a bounded set and h small enough, it can be approximated by a quadratic form, and the regression estimation over the set of quadratic forms gives an approximation of order $O(h\sqrt{h})$. If all these estimations hold, this would only add an error in $O(\sqrt{h})$ to the value function at time 0. In [1, Proposition 5], under suitable assumptions, we shown the convergence $\lim_{N_{\text{in}}, N_{\text{rg}} \to \infty} v^{h,N}(t, x) = v^h(t, x)$. Here, we may expect that $\limsup_{N_{\text{in}}, N_{\text{rg}} \to \infty} |v^{h,N}(t, x) - v^h(t, x)| \leq C\sqrt{h}$. However a

further study is needed to obtain a precise estimation of the error depending on N_{in}, N_{rg} and h.

Acknowledgements Marianne Akian was partially supported by the ANR project MALTHY, ANR-13-INSE-0003, by ICODE, and by PGMO, a joint program of EDF and FMJH (Fondation Mathématique Jacques Hadamard).

References

1. Akian, M., Fodjo, E.: A probabilistic max-plus numerical method for solving stochastic control problems. In: 55th Conference on Decision and Control (CDC 2016), Las Vegas (2016). arXiv:1605.02816
2. Akian, M., Fodjo, E.: From a monotone probabilistic scheme to a probabilistic max-plus algorithm for solving Hamilton-Jacobi-Bellman equations. In: Kalise, D., Kunisch, K., Rao, Z. (eds.) Hamilton-Jacobi-Bellman Equations. Radon Series on Computational and Applied Mathematics, vol. 21. De Gruyter, Berlin (2018). ArXiv:1709.09049
3. Barles, G., Jakobsen, E.R.: Error bounds for monotone approximation schemes for parabolic Hamilton-Jacobi-Bellman equations. Math. Comput. **76**(260), 1861–1893 (2007). http://dx.doi.org/10.1090/S0025-5718-07-02000-5
4. Barles, G., Souganidis, P.E.: Convergence of approximation schemes for fully nonlinear second order equations. Asymptotic Anal. **4**(3), 271–283 (1991)
5. Cheridito, P., Soner, H.M., Touzi, N., Victoir, N.: Second-order backward stochastic differential equations and fully nonlinear parabolic PDEs. Commun. Pure Appl. Math. **60**(7), 1081–1110 (2007). http://dx.doi.org/10.1002/cpa.20168
6. Da Lio, F., Ley, O.: Uniqueness results for second-order Bellman-Isaacs equations under quadratic growth assumptions and applications. SIAM J. Control Optim. **45**(1), 74–106 (2006). https://doi.org/10.1137/S0363012904440897
7. Fahim, A., Touzi, N., Warin, X.: A probabilistic numerical method for fully nonlinear parabolic PDEs. Ann. Appl. Probab. **21**(4), 1322–1364 (2011). http://dx.doi.org/10.1214/10-AAP723
8. Fleming, W.H., Soner, H.M.: Controlled Markov Processes and Viscosity Solutions. Springer, Berlin (1993)
9. Guo, W., Zhang, J., Zhuo, J.: A monotone scheme for high-dimensional fully nonlinear PDEs. Ann. Appl. Probab. **25**(3), 1540–1580 (2015). http://dx.doi.org/10.1214/14-AAP1030
10. Kaise, H., McEneaney, W.M.: Idempotent expansions for continuous-time stochastic control: compact control space. In: Proceedings of the 49th IEEE Conference on Decision and Control. Atlanta (2010)
11. Kaise, H., McEneaney, W.M.: Idempotent expansions for continuous-time stochastic control. SIAM J. Control Optim. **54**(1), 73–98 (2016). https://doi.org/10.1137/140971038
12. Kushner, H.J., Dupuis, P.G.: Numerical Methods for Stochastic Control Problems in Continuous Time. Applications of Mathematics (New York), vol. 24. Springer, New York (1992). https://doi.org/10.1007/978-1-4684-0441-8
13. McEneaney, W.M.: Max-Plus Methods for Nonlinear Control and Estimation. Systems & Control: Foundations & Applications. Birkhäuser, Boston (2006)
14. McEneaney, W.M., Kaise, H., Han, S.H.: Idempotent method for continuous-time stochastic control and complexity attenuation. In: Proceedings of the 18th IFAC World Congress, 2011, pp. 3216–3221. Milano, Italie (2011)

An Adaptive Max-Plus Eigenvector Method for Continuous Time Optimal Control Problems

Peter M. Dower

Abstract An adaptive max-plus eigenvector method is proposed for approximating the solution of continuous time nonlinear optimal control problems. At each step of the method, given a set of quadratic basis functions, a standard max-plus eigenvector method is applied to yield an approximation of the value function of interest. Using this approximation, an approximate level set of the back substitution error defined by the Hamiltonian is tessellated according to where each basis function is active in approximating the value function. The polytopes obtained, and their vertices, are sorted according to this back substitution error, allowing "worst-case" basis functions to be identified. The locations of these basis functions are subsequently evolved to yield new basis functions that reduce this worst-case. Basis functions that are inactive in the value function approximation are pruned, and the aforementioned steps repeated. Underlying algebraic properties associated with max-plus linearity, dynamic programming, and semiconvex duality are provided as a foundation for the development, and the utility of the proposed method is illustrated by example.

Keywords Optimal control · Dynamic programming · Semiconvexity · Max-plus algebra · Max-plus eigenvector method · Basis adaptation

1 Introduction

Continuous time optimal feedback control of nonlinear dynamical systems remains largely impractical due to serious computational obstacles associated with numerically solving the attendant Hamilton-Jacobi-Bellman (HJB) partial differential equation (PDE). Rather than addressing this HJB PDE, *max-plus methods* [1–3] exploit algebraic properties of the Lax-Oleinik semigroup of evolution operators,

P. M. Dower (✉)
Department of Electrical and Electronic Engineering, University of Melbourne, Melbourne, VIC, Australia
e-mail: pdower@unimelb.edu.au

M. Falcone et al. (eds.), *Numerical Methods for Optimal Control Problems*,
Springer INdAM Series 29, https://doi.org/10.1007/978-3-030-01959-4_10

defined via dynamic programming [4], in order to yield sparse approximations for the value function of interest. The investigation of a specific sparse approximation provides the foundation for the adaptive method presented here.

The Lax-Oleinik semigroup is a one-parameter semigroup of evolution operators that describes all possible finite horizon value functions for optimal control problems associated with a specific running payoff and space of terminal payoffs. Attendant max-plus linearity and semiconvexity properties admit a pair of state convolution representations for these evolution operators, and yield a respective pair of *max-plus fundamental solution semigroups* that equivalently describe value function propagation for all time horizons and admissible terminal payoffs. Elements of these semigroups are *max-plus linear max-plus integral operators*, defined with respect to corresponding bivariate kernels [1, 5, 6] that also define semigroups. Reduced complexity approximate evolution of elements of these kernel semigroups, defined with respect to a truncated problem specific sparse *basis*, yields a subclass of *max-plus eigenvector methods* for approximating value functions and optimal controls.

Basis selection is an essential step in applying max-plus eigenvector methods. Typically, basis selection is implemented either manually (via problem specific insights), using brute-force (via computationally expensive grids of basis functions), or using randomization, see [1, 3, 7]. In the development reported here, an *adaptive max-plus eigenvector method* is proposed, in which a standard max-plus eigenvector method [1, 8] is encapsulated within a back substitution error calculation and basis adaptation iteration. In this iteration, basis functions are added and pruned as a consequence of the value function approximation obtained by the aforementioned standard max-plus eigenvector method at each step. In identifying where new basis functions are required, a *Voronoi tessellation* of a subset of the state space is identified, with the subset of interest being an approximation of a level set of the back substitution error defined via the Hamiltonian [9]. Each element of this tessellation corresponds to a convex polytope of states in which a unique existing basis function is active in the value function approximation. Using this tessellation, the "worst-case" basis functions are identified by approximately evaluating the back substitution error on each polytope. By examining the dependence of this back substitution error on the location of these worst-case basis functions, new basis functions are added so as to reduce the expected error. Basis functions that are inactive in the value function approximation are pruned. The resulting value function approximation, tessellation and sorting, and basis adaptation iterations are illustrated by example.

In terms of organization, the specific class of optimal control problems of interest is provided in Sect. 2. This is followed in Sect. 3 by the introduction and application of the max-plus algebra and semiconvex duality to the development of a pair of max-plus fundamental solution semigroups. Section 4 subsequently describes a standard max-plus eigenvector method, the aforementioned approximate back substitution error calculation and basis adaptation, and a summary of the resulting algorithm implemented. Section 5 provides illustrations of its application by example, followed by some brief concluding remarks. An appendix recording some useful additional technical details is also included.

Throughout, $\mathbb{R}$ ($\mathbb{R}_{\geq 0}$) denotes the real (non-negative) numbers, $\overline{\mathbb{R}} \doteq \mathbb{R} \cup \{-\infty\} \cup \{+\infty\}$ denotes the extended reals, and $\mathbb{N}$ denotes the natural numbers. $\mathbb{R}^n$ denotes n-dimensional Euclidean space, given $n \in \mathbb{N}$, while $\mathscr{B}_y(r) \subset \mathbb{R}^n$ denotes the open ball of radius $r \in \mathbb{R}_{>0}$, centred at $y \in \mathbb{R}^n$. The inner product on $\mathbb{R}^n$ is written as $\langle \, , \rangle$. The space of $n \times m$ real matrices is denoted by $\mathbb{R}^{n\times m}$, $n, m \in \mathbb{N}$. Two subsets of $\mathbb{R}^{n\times n}$ of interest are defined, with the dependence on $n \in \mathbb{N}$ suppressed, by

$$\Sigma \doteq \left\{ P \in \mathbb{R}^{n\times n} \,\middle|\, P = P' \right\}, \qquad \Sigma_{\mathsf{M}} \doteq \left\{ P \in \Sigma \,\middle|\, P - \mathsf{M} > 0 \right\}, \qquad \mathsf{M} \in \Sigma,$$

in which P' denotes the transpose of P, and $P - \mathsf{M} > 0$ denotes positive definiteness of P with respect to M. A function $\psi : \mathbb{R}^n \to \overline{\mathbb{R}}$ is convex if its epigraph $\{(x, \alpha) \in \mathbb{R}^n \times \mathbb{R} \mid \psi(x) \leq \alpha\}$ is convex [10]. It is *lower closed* if $\psi = \mathrm{cl}^-\, \psi$, in which cl^- is the *lower closure* defined with respect to the lower semicontinuous envelope lsc by

$$\mathrm{cl}^-\, \psi(x) \doteq \begin{cases} \mathrm{lsc}\, \psi(x), & \mathrm{lsc}\, \psi(x) > -\infty \text{ for all } x \in \mathbb{R}^n, \\ -\infty, & \text{otherwise,} \end{cases}$$

for all $x \in \mathbb{R}^n$. Similarly, ψ is concave if $-\psi$ is convex, and *upper closed* if $-\psi$ is lower closed, see [10, pp. 15–17]. The *convex hull* of a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is defined to be the largest convex function that lower bounds f, see [10]. Given $\mathsf{K} \in \Sigma$, spaces of (uniformly) semiconvex and semiconcave functions are defined (respectively) by

$$\mathscr{S}_+^{\mathsf{K}} \doteq \left\{ \psi : \mathbb{R}^n \to \overline{\mathbb{R}} \;\middle|\; \psi + \tfrac{1}{2} \langle \cdot, \mathsf{K} \cdot \rangle \text{ convex, lower closed} \right\},$$

$$\mathscr{S}_-^{\mathsf{K}} \doteq \left\{ a : \mathbb{R}^n \to \overline{\mathbb{R}} \;\middle|\; a - \tfrac{1}{2} \langle \cdot, \mathsf{K} \cdot \rangle \text{ concave, upper closed} \right\}.$$

2 Optimal Control Problem

Attention is restricted to infinite horizon continuous time nonlinear optimal control problems of a form considered in [1]. An element in this class is identified by an infinite horizon value function $W : \mathbb{R}^n \to \overline{\mathbb{R}}$, specified using its finite horizon counterpart $W_t : \mathbb{R}^n \to \overline{\mathbb{R}}$, $t \in \mathbb{R}_{\geq 0}$. In particular,

$$W(x) \doteq \sup_{t\geq 0} W_t(x) = \lim_{t\to\infty} W_t(x), \qquad W_t(x) \doteq [\mathsf{S}_t\, \psi_0](x), \tag{1}$$

for all initial states $x \in \mathbb{R}^n$ and time horizons $t \in \mathbb{R}_{\geq 0}$, in which S_t denotes the corresponding dynamic programming evolution operator, and $\psi_0 : \mathbb{R}^n \to \mathbb{R}$ is the *zero* terminal payoff defined by $\psi_0(y) \doteq 0$ for all $y \in \mathbb{R}^n$. S_t is defined explicitly by

$$[\mathsf{S}_t\, \psi](x) \doteq \sup_{w \in \mathscr{W}[0,t]} \{ I_t(x, w) + \psi([\chi(x, w)]_t) \}, \qquad \psi : \mathbb{R}^n \to \overline{\mathbb{R}}, \tag{2}$$

for all $t \in \mathbb{R}_{\geq 0}$, $x \in \mathbb{R}^n$, in which $\mathscr{W}[0,t] \doteq \mathscr{L}_2([0,t];\mathbb{R}^m)$ is the input space, and I_t and χ denote the integrated running payoff and state trajectory maps defined by

$$I_t(x,w) \doteq \int_0^t l([\chi(x,w)]_s) - \frac{\gamma^2}{2}|w_s|^2\, ds\,, \tag{3}$$

$$[\chi(x,w)]_s \doteq \xi_s\,, \quad \begin{cases} \dot{\xi}_r = f(\xi_r) + \sigma(\xi_r)\, w_r\,, & r \in [0,s]\,,\ s \in [0,t]\,, \\ \xi_0 = x \in \mathbb{R}^n\,, \end{cases} \tag{4}$$

for any $x \in \mathbb{R}^n$ and $w \in \mathscr{W}[0,t]$. In (3), (4), $\gamma \in \mathbb{R}_{\geq 0}$ is a fixed gain parameter, and $\xi_s \in \mathbb{R}^n$ and $w_s \in \mathbb{R}^m$ denote the state and input of the associated nonlinear dynamics at time $s \in [0,t]$. Note in particular that the map $t \mapsto \mathsf{S}_t\, \psi_0$ is monotone, guaranteeing existence of the limit in (1).

Standard assumptions [1, p. 59] are adopted regarding the problem data $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$, $\sigma : \mathbb{R}^n \to \mathbb{R}^{n\times m}$, $l : \mathbb{R}^n \to \mathbb{R}$, and gain $\gamma \in \mathbb{R}_{\geq 0}$. They amount to a sector bound and an incremental exponential stability property for f, boundedness of σ along with invertibility of elements of its range, a quadratic growth bound on ℓ, and selection of a sufficiently large γ. Under these assumptions, the value functions W_t, $t \in \mathbb{R}_{\geq 0}$, and W of (1) are semiconvex [1, Theorem 4.9, p. 67, and Corollary 4.11, p. 69], and may be characterized [1, Lemma 3.14, p. 50, and Theorem 4.2, p. 60] as unique continuous viscosity solutions of the respective non-stationary and stationary HJB PDEs

$$0 = \frac{\partial W_t}{\partial t} + H(x, \nabla_x W_t(x))\,, \qquad W_0 = \Psi_0\,, \tag{5}$$

$$0 = H(x, \nabla_x W(x))\,, \qquad W(0) = 0\,, \tag{6}$$

for all $t \in \mathbb{R}_{\geq 0}$, $x \in \mathbb{R}^n$, in which the Hamiltonian $H : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is defined by

$$H(x,p) \doteq -l(x) - \langle p, f(x)\rangle - \frac{1}{2\gamma^2}\langle p, \sigma(x)\,\sigma(x)'\, p\rangle \tag{7}$$

for all $x, p \in \mathbb{R}^n$.

3 Max-Plus Fundamental Solution Semigroups

Max-Plus Algebra and Dynamic Programming The *(complete) max-plus algebra* [1, 11–14], denoted by the triple $(\overline{\mathbb{R}}, \oplus, \otimes)$, is a commutative semifield over $\overline{\mathbb{R}}$ equipped with addition $\oplus$ and multiplication $\otimes$ operations defined by

$$a \oplus b \doteq \max(a,b)\,, \quad a \otimes b \doteq a + b\,,$$

for all $a, b \in \overline{\mathbb{R}}$. The additive and multiplicative identities are $\mathbf{0} \doteq -\infty$ and $\mathbf{1} \doteq 0$ respectively, with

$$a \oplus \mathbf{0} = \max(a, -\infty) = a\,, \quad a \otimes \mathbf{0} = a - \infty = \mathbf{0}\,, \quad a \otimes \mathbf{1} = a + 0 = a\,, \qquad a \in \overline{\mathbb{R}}.$$

It is an idempotent algebra as the $\oplus$ operation is idempotent (i.e. $a \oplus a = a$), and a semifield as no additive inverse exists (i.e. there is no "$\ominus$" operation such that $a \oplus b \ominus b = a$). The max-plus integral of a function $f : \mathscr{Y} \to \overline{\mathbb{R}}$ over a set $\Omega \subset \mathscr{Y}$ is defined by

$$\int_{\Omega}^{\oplus} f(y)\, dy \doteq \sup_{y \in \Omega} f(y)\,.$$

Note further that

$$f(x) = \int_{\Omega}^{\oplus} \delta^{-}(x, y) \otimes f(y)\, dy, \quad \delta^{-}(x, y) \doteq \delta_{y}^{-}(x) \doteq \begin{cases} 0, & x = y, \\ -\infty, & x \neq y, \end{cases} \tag{8}$$

for all $x, y \in \Omega$, in which $\delta^{-} : \mathbb{R}^n \times \mathbb{R}^n \to \overline{\mathbb{R}}$ is the *max-plus delta function*. With $y \in \mathbb{R}^n$ and $\mathsf{K} \in \Sigma$ fixed arbitrarily, $\delta^{-}(\cdot, y) - \frac{1}{2}\langle \cdot\,, \mathsf{K}\,\cdot \rangle$ is concave and upper closed, so that $\delta^{-}(\cdot, y) \in \mathscr{S}_{-}^{\mathsf{K}}$, and by symmetry, $\delta^{-}(y, \cdot) \in \mathscr{S}_{-}^{\mathsf{K}}$.

The *min-plus algebra* is defined analogously over $\overline{\mathbb{R}}$, with min replacing max.

A max-plus vector space is a vector space over the max-plus algebra, and is often referred to as a moduloid [11], or an idempotent semimodule [13, 14]. A min-plus vector space is analogously defined with respect to the min-plus algebra. Both types of vector space contain the $\pm\infty$ functions, as their respective algebras are semifields over $\overline{\mathbb{R}}$. Spaces of semiconvex (semiconcave) functions are max-plus (min-plus) vector spaces, see for example [1, Theorem 2.7, p. 14].

The max-plus algebra is particularly useful for optimal control problems of the form (1). The dynamic programming evolution operator (2) involved naturally defines a one-parameter semigroup of integral operators that are linear with respect to the max-plus algebra [1, Theorem 4.5, p. 66]. The crucial properties concerning the dynamic programming evolution operator (2) for the optimal control problem (1) may be summarized as follows.

Assumption 1 *There exists an invertible* $\mathsf{M} \in \Sigma$ *and* $\tau^* \in \mathbb{R}_{>0}$ *such that the dynamic programming evolution operator (2) satisfies* $\mathsf{S}_\tau : \mathscr{S}_{+}^{-\mathsf{M}} \to \mathscr{S}_{+}^{-\mathsf{M}}$ *for all* $\tau \in [0, \tau^*]$.

Theorem 2 *Given* $\mathsf{M} \in \Sigma$ *as per Assumption 1, the dynamic programming evolution operator* S_t *of (2) satisfies*

$$\mathsf{S}_t : \mathscr{S}_+^{-\mathsf{M}} \to \mathscr{S}_+^{-\mathsf{M}}, \quad [\mathsf{S}_t\,\psi](x) = \int_{\mathscr{W}[0,t]}^{\oplus} I_t(x,w) \otimes \psi([\xi(x,w)]_t)\,dw\,,$$
$$\mathsf{S}_\tau\,\mathsf{S}_t = \mathsf{S}_{\tau+t}\,,\ \mathsf{S}_0 = \mathsf{I}\,, \quad \mathsf{S}_t(\psi \oplus c \otimes \phi) = \mathsf{S}_t\psi \oplus c \otimes \mathsf{S}_t\,\phi\,, \tag{9}$$

for all $\tau, t \in \mathbb{R}_{\geq 0}$, $\psi, \phi \in \mathscr{S}_+^{-\mathsf{M}}$, $c \in \overline{\mathbb{R}}$.

Proof Fix $\mathsf{M} \in \Sigma$ and $\tau^* \in \mathbb{R}_{>0}$ as per Assumption 1, and any $\psi \in \mathscr{S}_+^{-\mathsf{M}}$, $t \in \mathbb{R}_{\geq 0}$. Define $\kappa_t \doteq \lfloor \frac{t}{\tau^*} \rfloor \in \mathbb{N} \cup \{0\}$ and $\tilde{t} \doteq t - \kappa_t\,\tau^* \in [0, \tau^*]$. Again by Assumption 1, note that $\mathsf{S}_{\tilde{t}} : \mathscr{S}_+^{-\mathsf{M}} \to \mathscr{S}_+^{-\mathsf{M}}$. Hence, by induction, $\mathsf{S}_t = \mathsf{S}_{\tilde{t}} \circ [\mathsf{S}_{\tau^*}]^{\circ\kappa_t} : \mathscr{S}_+^{-\mathsf{M}} \to \mathscr{S}_+^{-\mathsf{M}}$, in which $[\cdot]^{\circ\kappa_t}$ denotes the composition of $[\cdot]$ with itself $\kappa_t - 1$ times. That is, the first assertion in (9) holds. For the remaining assertions in (9), the integral operator form of the second assertion is immediate by inspection of (2), while the semigroup property of the third assertion follows by dynamic programming. Finally, (max-plus) linearity as per the fourth assertion follows by inspection of (2), see [1, Theorem 4.5, p. 66]. □

Semiconvex Duality *Semiconvex duality* (sometimes referred to as *max-plus duality*) is a duality between the semiconvex and semiconcave function spaces, defined via the semiconvex transform [15]. The semiconvex transform is a generalization of the Legendre-Fenchel transform, where convexity is weakened to semiconvexity via a relaxation of affine to quadratic support. The quadratic support functions involved are defined via the bivariate basis function $\varphi : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ given by

$$\varphi(x,z) \doteq \frac{1}{2}\,\langle x - z,\, \mathsf{M}\,(x - z)\rangle \tag{10}$$

for all $x, z \in \mathbb{R}^n$, in which $\mathsf{M} \in \Sigma$ is as per Assumption 1. With $-\mathsf{K} \in \Sigma_{\mathsf{M}} \cup \{\mathsf{M}\}$, the duality is defined via the semiconvex transform D_φ and its inverse D_φ^{-1}, with

$$a = \mathsf{D}_\varphi\,\psi\,, \quad \psi = \mathsf{D}_\varphi^{-1}\,a\,, \tag{11}$$

for all $\psi \in \mathscr{S}_+^{\mathsf{K}}$, $a \in \mathscr{R}_-^{\mathsf{K}}$, in which

$$\mathsf{D}_\varphi\,\psi \doteq -\int_{\mathbb{R}^n}^{\oplus} \varphi(x,\cdot) \otimes [-\psi(x)]\,dx\,, \qquad \psi \in \mathrm{dom}\,(\mathsf{D}_\varphi) \doteq \mathscr{S}_+^{\mathsf{K}}\,, \tag{12}$$

$$\mathsf{D}_\varphi^{-1}\,a \doteq \int_{\mathbb{R}^n}^{\oplus} \varphi(\cdot,z) \otimes a(z)\,dz\,, \qquad a \in \mathrm{dom}\,(\mathsf{D}_\varphi^{-1}) \doteq \mathscr{R}_-^{\mathsf{K}}\,, \tag{13}$$

and

$$\mathscr{R}_-^{\mathsf{K}} \doteq \operatorname{ran}(\mathsf{D}_\varphi) = \left\{ a : \mathbb{R}^n \to \overline{\mathbb{R}} \;\middle|\; \exists\, \psi \in \mathscr{S}_+^{\mathsf{K}} \text{ such that } \mathsf{D}_\varphi\, \psi = a \right\}. \tag{14}$$

The required properties of D_φ and D_φ^{-1} are recorded in the following theorem, c.f. [1, Theorem 2.9, p. 16].

Theorem 3 *With $-\mathsf{K} \in \Sigma_{\mathsf{M}} \cup \{\mathsf{M}\}$ fixed, D_φ and D_φ^{-1} of (12) and (13) satisfy the following properties:*

1) $\mathit{dom}\,(\mathsf{D}_\varphi^{-1}) = \mathit{ran}\,(\mathsf{D}_\varphi) = \mathscr{R}_-^{\mathsf{K}} \subset \mathscr{R}_-^{-\mathsf{M}} \equiv \mathscr{S}_-^{-\mathsf{M}}$;
2) $\mathit{ran}\,(\mathsf{D}_\varphi^{-1}) = \mathit{dom}\,(\mathsf{D}_\varphi) = \mathscr{S}_+^{\mathsf{K}} \subset \mathscr{S}_+^{-\mathsf{M}}$;
3) $\mathsf{D}_\varphi\, \mathsf{D}_\varphi^{-1} = \mathsf{I}$ on $\mathit{dom}\,(\mathsf{D}_\varphi^{-1}) = \mathscr{R}_-^{\mathsf{K}}$; and
4) $\mathsf{D}_\varphi^{-1}\, \mathsf{D}_\varphi = \mathsf{I}$ on $\mathit{dom}\,(\mathsf{D}_\varphi) = \mathscr{S}_+^{\mathsf{K}}$.

Proof See the Appendix. □

Remark 4 Theorems 2 and 3 motivate adoption of $\mathsf{K} \doteq -\mathsf{M}$ throughout for simplicity, in which $\mathsf{M} \in \Sigma$ is as per Assumption 1. □

The max-plus vector space $\mathscr{S}_+^{-\mathsf{M}}$ has a countable basis, see [1, Theorem 2.13, p. 20], given by $\{\psi_i\}_{i\in\mathbb{N}}$, with elements $\psi_i : \mathbb{R}^n \to \mathbb{R}$ defined via (10) by

$$\psi_i \doteq \varphi(\cdot, z_i)\,, \tag{15}$$

in which $\{z_i\}_{i\in\mathbb{N}}$ is a countable dense subset of $\mathbb{R}^n$. The semiconvex transform and its inverse (12) and (13) subsequently satisfy (see Lemma 18 in the Appendix)

$$\mathsf{D}_\varphi\, \psi = -\mathrm{cl}^- \operatorname{co}\left(-\bigoplus_{i\in\mathbb{N}} \psi_i(0) \otimes \delta_{z_i}^-(\cdot) \otimes [\mathsf{D}_\varphi \psi](z_i) \right) - \varphi(0, \cdot)\,, \tag{16}$$

$$\mathsf{D}_\varphi^{-1}\, a = \bigoplus_{i\in\mathbb{N}} \psi_i(\cdot) \otimes a(z_i)\,, \tag{17}$$

in which $\delta_{z_i}^-(\cdot)$ is a max-plus delta function (8), and cl^- and co denote respectively the lower closure and convex hull (as introduced in Sect. 1). Identity (16) specifies the semiconvex transform $\mathsf{D}_\varphi \psi$ everywhere on $\mathbb{R}^n$ in terms of its evaluation on the dense subset $\{z_i\}_{i\in\mathbb{N}} \subset \mathbb{R}^n$, i.e. in terms of $\{[\mathsf{D}_\varphi \psi](z_i)\}_{i\in\mathbb{N}}$. The lower closure and convex hull operations indicated ensure that the map defined by

$$z \mapsto [\mathsf{D}_\varphi \psi](z) + \varphi(0, z) = [\mathsf{D}_\varphi \psi](z) - \frac{1}{2}\, \langle z,\, -\mathsf{M}\, z \rangle$$

is concave and upper closed, as required for $\mathsf{D}_\varphi \psi \in \mathscr{S}_-^{-\mathsf{M}}$. By inspection of (16) and (17), the sequence $\{a(z_i)\}_{i\in\mathbb{N}} = \{[\mathsf{D}_\varphi\, \psi](z_i)\}_{i\in\mathbb{N}}$ completely describes $a \in \mathscr{S}_-^{-\mathsf{M}}$. Hence, it is sufficient to write (11) using the coordinate representation

$$e_i \doteq a(z_i) = [\mathsf{D}_\varphi\, \psi](z_i) = -\int_{\mathbb{R}^n}^{\oplus} \psi_i(x) \otimes [-\psi(x)]\, dx\,, \quad \psi \in \mathscr{S}_+^{-\mathsf{M}},\ i \in \mathbb{N}\,,$$
$$\psi = \bigoplus_{i\in\mathbb{N}} \psi_i(\cdot) \otimes e_i\,, \qquad e \in \overline{\mathbb{R}}^{\infty}\,, \tag{18}$$

in which $\overline{\mathbb{R}}^{\infty}$ denotes the product space of $\overline{\mathbb{R}}$ with itself countably infinite times.

A semiconvex function $\psi \in \mathscr{S}_+^{-\mathsf{M}}$ may be approximated by truncating the sums in (17) and (18) to a finite number $\nu \in \mathbb{N}$ of terms. This yields approximations $\widehat{a} \in \mathscr{S}_-^{-\mathsf{M}}$ and $\widehat{\psi} \in \mathscr{S}_+^{-\mathsf{M}}$ of $a = \mathsf{D}_\varphi \psi \in \mathscr{S}_-^{-\mathsf{M}}$ and $\psi \in \mathscr{S}_+^{-\mathsf{M}}$, satisfying

$$\widehat{a} = \mathsf{D}_\varphi\, \widehat{\psi}\,, \quad \widehat{\psi} = \mathsf{D}_\varphi^{-1}\, \widehat{a}\,, \tag{19}$$

$$\widehat{a} \doteq \widehat{\mathsf{D}}_\varphi^{\nu}\, \psi \doteq -\mathrm{cl}^{-}\,\mathrm{co}\left(-\bigoplus_{i=1}^{\nu} \psi_i(0) \otimes \delta_{z_i}^{-}(\cdot) \otimes [\mathsf{D}_\varphi\, \psi](z_i)\right) - \varphi(0,\cdot)\,. \tag{20}$$

As per (18), as $\widehat{a}(z_i) = [\mathsf{D}_\varphi\, \psi](z_i)$, $i \in \mathbb{N}_{\le\nu}$, completely describes $\widehat{a}$ via (20),

$$\widehat{e}_i \doteq \widehat{a}(z_i) = [\mathsf{D}_\varphi\, \widehat{\psi}](z_i) = -\int_{\mathbb{R}^n}^{\oplus} \psi_i(x) \otimes [-\widehat{\psi}(x)]\, dx\,, \quad \widehat{\psi} \in \mathscr{S}_+^{-\mathsf{M}},\ i \in \mathbb{N}_{\le\nu}\,,$$
$$\widehat{\psi} = \bigoplus_{i=1}^{\nu} \psi_i(\cdot) \otimes \widehat{e}_i\,, \qquad \widehat{e} \in \overline{\mathbb{R}}^{\nu}\,. \tag{21}$$

Max-Plus Fundamental Solution Semigroups A *max-plus fundamental solution semigroup* [5, 6, 16–19] is a one-parameter semigroup of *max-plus linear max-plus integral operators* that can be used to propagate a finite horizon value function $\mathsf{S}_t\, \psi$, or its semiconvex dual, to longer time horizons $t \in \mathbb{R}_{\ge 0}$. Given the max-plus or min-plus vector space $\mathscr{V} \in \{\mathscr{S}_+^{-\mathsf{M}}, \mathscr{S}_-^{-\mathsf{M}}\}$, an element $\mathsf{F}_t^{\oplus} : \mathscr{V} \to \mathscr{V}$ of a max-plus fundamental solution semigroup is an operator of the form

$$\mathsf{F}_t^{\oplus}\, \phi \doteq \int_{\mathbb{R}^n}^{\oplus} F_t(\cdot, y) \otimes \phi(y)\, dy\,, \quad t \in \mathbb{R}_{\ge 0}\,,\ \phi \in \mathscr{V}, \tag{22}$$

in which $F_t : \mathbb{R}^n \times \mathbb{R}^n \to \overline{\mathbb{R}}$ denotes its bivariate kernel. The dynamic programming evolution operator (2) and the semiconvex transform (12) can be combined to yield primal and dual space max-plus fundamental solution semigroups $\{\mathsf{G}_t^{\oplus}\}_{t\in\mathbb{R}_{\ge 0}}$ and $\{\mathsf{B}_t^{\oplus}\}_{t\in\mathbb{R}_{\ge 0}}$, corresponding respectively to $\mathscr{V} \doteq \mathscr{S}_+^{-\mathsf{M}}$ and $\mathscr{V} \doteq \mathscr{S}_-^{-\mathsf{M}}$. The associated

kernels are specified in terms of an auxiliary value function $S_t : \mathbb{R}^n \times \mathbb{R}^n \to \overline{\mathbb{R}}$, which is defined with respect to S_t and φ of (2) and (10) by

$$S_t(x, z) \doteq [\mathsf{S}_t\, \varphi(\cdot, z)](x)\,, \tag{23}$$

for all $t \in \mathbb{R}_{\geq 0}$, $x, z \in \mathbb{R}^n$. The following is a consequence of Assumption 1.

Assumption 5 *$S_\tau(x, \cdot)$, $S_\tau(\cdot, z) \in \mathscr{S}_+^{-\mathsf{M}}$ for all $\tau \in [0, \tau^*]$, $x, z \in \mathbb{R}^n$.*

Given any $t \in \mathbb{R}_{\geq 0}$, kernels G_t, $B_t : \mathbb{R}^n \times \mathbb{R}^n \to \overline{\mathbb{R}}$ of the aforementioned max-plus linear max-plus integral operators $\mathsf{G}_t^{\oplus}$ and $\mathsf{B}_t^{\oplus}$ are defined via Assumption 5 by

$$G_t(x, y) \doteq \begin{cases} [\mathsf{D}_\varphi\, S_t(x, \cdot)](y)\,, & t \in [0, \tau^*]\,, \\ G_{\tau_t}^{\circledast k_t}(x, y)\,, & t > \tau^*\,, \end{cases} \qquad B_t(y, z) \doteq \begin{cases} [\mathsf{D}_\varphi\, S_t(\cdot, z)](y)\,, & t \in [0, \tau^*]\,, \\ B_{\tau_t}^{\circledast k_t}(y, z)\,, & t > \tau^*\,, \end{cases}$$

$$k_t \doteq \left\lceil \frac{t}{\tau^*} \right\rceil\,, \quad \tau_t \doteq \frac{t}{k_t}\,, \tag{24}$$

for all $x, y, z \in \mathbb{R}^n$, in which $F_{\tau_t}^{\circledast k_t}$ denotes the convolution defined for $F \in \{G, B\}$ by

$$[F_{\tau_t}^{\circledast k_t}](x, y) \doteq \int_{\mathbb{R}^n}^{\oplus} \int_{\mathbb{R}^n}^{\oplus} \cdots \int_{\mathbb{R}^n}^{\oplus} F_{\tau_t}(x, \eta_1) \otimes F_{\tau_t}(\eta_1, \eta_2) \otimes \cdots \otimes F_{\tau_t}(\eta_{k_t - 1}, y)\, d\eta_{k_t - 1} \cdots d\eta_2\, d\eta_1\,. \tag{25}$$

Theorem 6 *Given Assumptions 1 and 5, and max-plus linear max-plus integral operators $\mathsf{G}_t^{\oplus}$, $\mathsf{B}_t^{\oplus}$ defined by (22) with respect to the kernels G_t, B_t of (24),*

$$\mathsf{G}_t^{\oplus}\, \psi = \mathsf{S}_t\, \psi = \mathsf{D}_\varphi^{-1}\, \mathsf{B}_t^{\oplus}\, \mathsf{D}_\varphi\, \psi \tag{26}$$

for all $t \in \mathbb{R}_{\geq 0}$, $\psi \in \mathscr{S}_+^{-\mathsf{M}}$, in which S_t is as per (2). Furthermore,

$$\mathsf{G}_{\tau+t}^{\oplus} = \mathsf{G}_\tau^{\oplus}\, \mathsf{G}_t^{\oplus}\,, \quad \mathsf{B}_{\tau+t}^{\oplus} = \mathsf{B}_\tau^{\oplus}\, \mathsf{B}_t^{\oplus}\,, \tag{27}$$

for all $\tau, t \in \mathbb{R}_{\geq 0}$.

Proof Fix any $t \in [0, \tau^*]$, with $\tau^* \in \mathbb{R}_{>0}$ as per Assumptions 1 and 5, $\psi \in \mathscr{S}_+^{-\mathsf{M}}$, and $x \in \mathbb{R}^n$. Note by Assumption 5 and definition (24) that

$$S_t(x, z) = [\mathsf{D}_\varphi^{-1} G_t(x, \cdot)](z) \tag{28}$$

for all $z \in \mathbb{R}^n$. Recalling (2) and applying (12), (13), (28), and Theorem 2, and noting that φ of (10) is symmetric (i.e. $\varphi(y,z) = \varphi(z,y)$ for all $y, z \in \mathbb{R}^n$),

$$[\mathsf{S}_t\,\psi](x) = \left[\mathsf{S}_t\,\mathsf{D}_\varphi^{-1}\,\mathsf{D}_\varphi\,\psi\right](x) = \left[\mathsf{S}_t\int_{\mathbb{R}^n}^{\oplus}\varphi(\cdot,z)\otimes[\mathsf{D}_\varphi\,\psi](z)\,dz\right](x)$$

$$= \int_{\mathbb{R}^n}^{\oplus}[\mathsf{S}_t\,\varphi(\cdot,z)](x)\otimes[\mathsf{D}_\varphi\,\psi](z)\,dz = \int_{\mathbb{R}^n}^{\oplus}S_t(x,z)\otimes[\mathsf{D}_\varphi\,\psi](z)\,dz \tag{29}$$

$$= \int_{\mathbb{R}^n}^{\oplus}[\mathsf{D}_\varphi^{-1}G_t(x,\cdot)](z)\otimes[\mathsf{D}_\varphi\,\psi](z)\,dz = \int_{\mathbb{R}^n}^{\oplus}\int_{\mathbb{R}^n}^{\oplus}\varphi(z,y)\otimes G_t(x,y)\,dy\otimes[\mathsf{D}_\varphi\,\psi](z)\,dz$$

$$= \int_{\mathbb{R}^n}^{\oplus}G_t(x,y)\otimes\int_{\mathbb{R}^n}^{\oplus}\varphi(y,z)\otimes[D_\varphi\,\psi](z)\,dz\,dy = \int_{\mathbb{R}^n}^{\oplus}G_t(x,y)\otimes[\mathsf{D}_\varphi^{-1}\,\mathsf{D}_\varphi\,\psi](y)\,dy$$

$$= \int_{\mathbb{R}^n}^{\oplus}G_t(x,y)\otimes\psi(y)\,dy = [\mathsf{G}_t^{\oplus}\,\psi](x)\,, \tag{30}$$

which is the left-hand equality in (26). Similarly, applying the definition (24) of kernel B_t in (29),

$$[\mathsf{S}_t\,\psi](x) = \int_{\mathbb{R}^n}^{\oplus}S_t(x,z)\otimes[\mathsf{D}_\varphi\,\psi](z)\,dz = \int_{\mathbb{R}^n}^{\oplus}[\mathsf{D}_\varphi^{-1}B_t(\cdot,z)](x)\otimes[\mathsf{D}_\varphi\,\psi](z)\,dz$$

$$= \int_{\mathbb{R}^n}^{\oplus}\int_{\mathbb{R}^n}^{\oplus}\varphi(x,y)\otimes B_t(y,z)\,dy\otimes[\mathsf{D}_\varphi\,\psi](z)\,dz$$

$$= \int_{\mathbb{R}^n}^{\oplus}\varphi(x,y)\otimes\int_{\mathbb{R}^n}^{\oplus}B_t(y,z)\otimes[\mathsf{D}_\varphi\,\psi](z)\,dz\,dy = \int_{\mathbb{R}^n}^{\oplus}\varphi(x,y)\otimes[\mathsf{B}_t^{\oplus}\,\mathsf{D}_\varphi\,\psi](y)\,dy$$

$$= [\mathsf{D}_\varphi^{-1}\,\mathsf{B}_t^{\oplus}\,\mathsf{D}_\varphi\,\psi](x)\,, \tag{31}$$

which is the right-hand equality in (26).

For $t > \tau^*$, define $k_t \in \mathbb{N}$ and $\tau_t \in [0, \tau^*]$ as per (24). Applying Theorem 2, definition (24) of G_t, and (30),

$$\mathsf{S}_t\,\psi = [\mathsf{S}_{\tau_t}]^{\circ k_t}\,\psi = [\mathsf{G}_{\tau_t}^{\oplus}]^{\circ k_t}\,\psi = \int_{\mathbb{R}^n}^{\oplus}G_{\tau_t}^{\circledast k_t}(\cdot,y)\otimes\psi(y)\,dy$$

$$= \int_{\mathbb{R}^n}^{\oplus}G_t(\cdot,y)\otimes\psi(y)\,dy = \mathsf{G}_t\,\psi\,. \tag{32}$$

in which $[\cdot]^{\circ k_t}$ denotes self-composition $k_t - 1$ times, as per the proof of Theorem 2. Similarly applying Theorem 2, definition (24) of B_t, and (31),

$$\mathsf{S}_t\,\psi = [\mathsf{S}_{\tau_t}]^{\circ k_t}\,\psi = [\mathsf{D}_\varphi^{-1}\,\mathsf{B}_{\tau_t}\,\mathsf{D}_\varphi]^{\circ k_t}\,\psi = \mathsf{D}_\varphi^{-1}\,[\mathsf{B}_{\tau_t}^{\oplus}]^{\circ k_t}\,\mathsf{D}_\varphi\,\psi = \mathsf{D}_\varphi^{-1}\,\mathsf{B}_t\,\mathsf{D}_\varphi\,\psi\,.$$

The semigroup properties (27) follow analogously by Theorem 2 and (26). □

Remark 7 The identity operator $\mathsf{I} : \mathscr{S}_+^{-\mathsf{M}} \to \mathscr{S}_+^{-\mathsf{M}}$, defined by $\mathsf{I}\,\psi \doteq \psi$ for all $\psi \in \mathscr{S}_+^{-\mathsf{M}}$, is also a max-plus linear max-plus integral operator of the form (22). Recalling (8), its bivariate kernel is the max-plus delta function δ^-, with

$$\mathsf{I}\,\psi = \mathsf{I}^{\oplus}\,\psi \doteq \int_{\mathbb{R}^n}^{\oplus} \delta^-(\cdot, y) \otimes \psi(y)\, dy\,, \tag{33}$$

for all $\psi \in \mathscr{S}_+^{-\mathsf{M}}$. An alternative representation for the kernel G_t of (24) may be derived for any $t \in \mathbb{R}_{\geq 0}$ using (2) and (33). With $t \in [0, \tau^*]$ and $\psi \in \mathscr{S}_+^{-\mathsf{M}}$, applying definitions (24) and (33) of G_t and $\mathsf{I}^{\oplus}$, along with max-plus linearity of S_t from Theorem 2, yields

$$\begin{aligned}
&[\mathsf{D}_\varphi^{-1}\, G_t(x, \cdot)](y) = S_t(x, y) = [\mathsf{S}_t\, \varphi(\cdot, y)](x) = [\mathsf{S}_t\, \mathsf{I}^{\oplus}\, \varphi(\cdot, y)](x) \\
&= \left[\mathsf{S}_t \int_{\mathbb{R}^n}^{\oplus} \delta^-(\cdot, z) \otimes \varphi(z, y)\, dz\right](x) = \int_{\mathbb{R}^n}^{\oplus} [\mathsf{S}_t\, \delta^-(\cdot, z)](x) \otimes \varphi(z, y)\, dz \\
&= \int_{\mathbb{R}^n}^{\oplus} \varphi(y, z) \otimes [\mathsf{S}_t\, \delta^-(\cdot, z)](x)\, dz = [\mathsf{D}_\varphi^{-1}\, T_t(x, \cdot)](y)\,,
\end{aligned}$$

in which $T_t : \mathbb{R}^n \times \mathbb{R}^n \to \overline{\mathbb{R}}$ is defined by $T_t(x, y) \doteq [\mathsf{S}_t\, \delta^-(\cdot, y)](x)$, for all $x, y \in \mathbb{R}^n$. Hence, applying Theorem 3, i.e. $\mathsf{D}_\varphi\, \mathsf{D}_\varphi^{-1} = \mathsf{I} = \mathsf{I}^{\oplus}$,

$$G_t(x, y) = [\mathsf{D}_\varphi\, \mathsf{D}_\varphi^{-1} T_t(x, \cdot)](y) = T_t(x, y) = [\mathsf{S}_t\, \delta^-(\cdot, y)](x)$$

for all $x, y \in \mathbb{R}^n$, in which $\mathsf{S}_t\, \delta^-(\cdot, y)$ defines the value function for an optimal two-point boundary value problem with terminal state $y \in \mathbb{R}^n$. Recalling (24) and Assumption 5, this representation may be extended from $t \in [0, \tau^*]$ to any $t \in \mathbb{R}_{\geq 0}$. □

Theorem 6 stipulates that $\{\mathsf{G}_t^{\oplus}\}_{t \in \mathbb{R}_{\geq 0}}$ and $\{\mathsf{B}_t^{\oplus}\}_{t \in \mathbb{R}_{\geq 0}}$ define one-parameter semigroups of max-plus linear max-plus integral operators of the form (22), with elements of each defined via their respective kernels G_t and B_t as per (24) for any $t \in \mathbb{R}_{\geq 0}$. These define *max-plus primal* and *max-plus dual space fundamental solution semigroups* respectively [5, 16–19]. For a fixed $\tau \in \mathbb{R}_{>0}$, Theorem 6 further states that either semigroup may be used to iteratively represent the finite and infinite horizon value functions $W_{k\tau} \doteq \mathsf{S}_{k\tau}\, \psi_0$ and W of (1) corresponding to the terminal payoff $\psi_0 \in \mathscr{S}_+^{-\mathsf{M}}$ for all $k \in \mathbb{Z}_{\geq 0}$. Using the dual space semigroup in this way underpins the max-plus eigenvector method of [1].

Remark 8 Max-plus primal space fundamental solution semigroups have been exploited in nonlinear filtering problems [15, p. 692], and for constructing solution representations for the gravitational N-body problem [16, Section 1.2, p. 2901], differential Riccati equations [19, Section 3, p. 15], and two-point boundary value problems involving wave equations [20, Section 3, p. 2159]. Similarly, max-plus

dual space fundamental solution semigroups have been exploited in developing standard max-plus eigenvector methods [1, e.g. Sections 4.4 and 4.5, p. 70], and for constructing solution representations for differential Riccati equations [5, Theorems 4.6 and 4.7, p. 926], [18, Section 3, p. 976], and non-quadratic regulator problems [17, Section 3, p. 701]. As indicated, it is the max-plus dual space fundamental solution semigroup underlying the standard max-plus eigenvector method of [1] that is particular interest here. □

4 Adaptive Max-Plus Eigenvector Method

Max-Plus Eigenvector Method In applying the max-plus dual space fundamental solution semigroup $\{\mathsf{B}_t^{\oplus}\}_{t\geq 0}$, a sequence of finite horizon value functions $\{W_{k\tau}\}_{k\in\mathbb{N}}$ defined by (1) for a priori fixed $\tau \in \mathbb{R}_{>0}$ can be equivalently represented via Theorem 6 by a sequence of semiconcave functions $\{a_k\}_{k\in\mathbb{N}} \subset \mathscr{S}_-^{\mathsf{M}}$. In particular [1],

$$W_{k\tau} = \mathsf{D}_{\varphi}^{-1}\, a_k\,, \quad a_k = \mathsf{B}_{k\tau}^{\oplus}\, a_0 = \mathsf{B}_{\tau}^{\oplus}\, a_{k-1} = [\mathsf{B}_{\tau}^{\oplus}]^{\circ k}\, a_0\,, \quad a_0 \doteq \mathsf{D}_{\varphi}\, \psi_0\,, \tag{34}$$

for all $k \in \mathbb{N}\cup\{0\}$, in which it is recalled that $[\cdot]^{\circ k}$ denotes composition $k-1$ times. Defining an infinitely dimensioned vector e_k component-wise by $[e_k]_i \doteq a_k(z_i)$ for all $i, k \in \mathbb{N}$, with respect to the dense set $\{z_i\}_{i\in\mathbb{N}} \subset \mathbb{R}^n$ in (18), it follows that

$$\begin{aligned}[e_k]_i = a_k(z_i) = [\mathsf{B}_{\tau}^{\oplus}\, a_{k-1}](z_i) &= \int_{\mathbb{R}^n}^{\oplus} B_\tau(z_i, z) \otimes a_{k-1}(z)\, dz \\ &= \bigoplus_{j\in\mathbb{N}} B_\tau(z_i, z_j) \otimes a_{k-1}(z_j) = \bigoplus_{j\in\mathbb{N}} [B_\tau]_{ij} \otimes [e_{k-1}]_j = [B_\tau \otimes e_{k-1}]_i\,,\end{aligned} \tag{35}$$

for all $k, i \in \mathbb{N}$, in which (abusing notation) B_τ is both the kernel of $\mathsf{B}_{\tau}^{\oplus}$ and its representation as a compatibly (infinite) dimensioned square matrix, defined element-wise by $[B_\tau]_{ij} \doteq B_\tau(z_i, z_j)$ for $i, j \in \mathbb{N}$. Note also that the final $\otimes$ in (35) denotes (with a further abuse of notation) a suitably generalized matrix-vector max-plus multiplication operation. Combining (18), (34) and (35), the infinite horizon value function W of (1) is equivalently given by

$$\begin{aligned} W = \mathsf{D}_{\varphi}^{-1}\, a_\infty = \bigoplus_{i\in\mathbb{N}} \psi_i(\cdot) \otimes [e_\infty]_i\,, \quad e_\infty \doteq \lim_{k\to\infty} e_k\,, \\ e_k = B_\tau \otimes e_{k-1}\,, \quad [e_0]_i = [\mathsf{D}_{\varphi}\, \psi_0](z_i)\,, \quad k, i \in \mathbb{N}\,. \end{aligned} \tag{36}$$

An approximation $\widehat{W}$ of W follows by truncating the basis $\{\psi_i\}_{i\in\mathbb{N}}$ of (15) to a finite cardinality subset $\{\psi_i\}_{i\in\mathbb{N}_{\leq\nu}}$, $\nu\in\mathbb{N}$, with (21), (34), and (35) yielding

$$\widehat{W} \doteq \bigoplus_{i=1}^{\nu} \psi_i(\cdot)\otimes[\widehat{e}_\infty]_i\,, \quad \widehat{e}_\infty \doteq \lim_{k\to\infty}\widehat{e}_k\,, \tag{37}$$
$$\widehat{e}_k = \widehat{B}_\tau\otimes\widehat{e}_{k-1}\,,\ [\widehat{e}_0]_i = [\widehat{\mathsf{D}}_\varphi\,\psi_0](z_i)\,,\ k\in\mathbb{N}\,,\ i\in\mathbb{N}_{\leq\nu}\,,$$

in which $\widehat{B}_\tau \in \overline{\mathbb{R}}^{\nu\times\nu}$ and $\widehat{e}_k \in \overline{\mathbb{R}}^{\nu}$. Iteration (37) is referred to as a *max-plus eigenvector method* [1, 8]. It is applicable to the numerical computation of finite and infinite horizon value functions of the form (1), and is applied here for the latter.

Remark 9 The idempotent property of the $\oplus$ operation is crucial in establishing that the limit $\widehat{e}_\infty$ in (37), where it exists and is finite, is achieved in a finite number of steps. In particular, recalling the proof of [1, Theorem 4.22, p. 81], elements of the vector $[\widehat{e}_k]_i = [(\widehat{B}_\tau)^{\otimes k}\otimes 0]_i \doteq [\widehat{B}_\tau\otimes\cdots\otimes\widehat{B}_\tau\otimes 0]_i$ generated by (37) correspond to the optimal cost over all length k paths traversed from the ith vertex of a precedence graph, defined by $\widehat{B}_\tau$. As $\nu\in\mathbb{N}$, this graph is finite. Furthermore, paths longer than ν necessarily contains loops, and it may be shown that these loops contribute negative cost. Consequently, longer paths cannot yield maximum cost, which implies convergence of (37) in a finite number of steps. (Note also that one vertex must correspond to a basis function located at $0\in\mathbb{R}^n$, see [1, Lemma 4.20, p. 77].)

With $\widehat{e}_\infty$ denoting the limit as per (37), note by inspection that $0\otimes\widehat{e}_\infty = \mathbf{1}\otimes\widehat{e}_\infty = \widehat{B}_\tau\otimes\widehat{e}_\infty$, so that $\widehat{e}_\infty$ is an eigenvector of $\widehat{B}_\tau$ corresponding to eigenvalue $\mathbf{1}(\doteq 0)$. This eigenvector is unique up to a max-plus multiplicative constant, see [1, Theorem 4.22, p. 81 and Corollary 4.23, p. 82]. □

Remark 10 Elements $[\widehat{B}_\tau]_{ij} = \widehat{B}_\tau(z_i, z_j)$, $i, j\in\mathbb{N}_{\leq\nu}$, of the matrix $\widehat{B}_\tau$ appearing in (37) are approximated via numerical integration of a related nonlinear ODE, see [1, Section 5.3.1, p. 122], [8]. The maximum allowable errors in this approximation, in the presence of which the convergence and eigenvector properties in Remark 9 are retained, are specified in [1, Theorem 5.9, p. 107]. □

Tessellation and Sorting Given a convex and bounded subset $\mathscr{Y}\subset\mathbb{R}^n$ of the state space, $x\in\mathscr{Y}$, and truncated basis $\mathscr{B}\doteq\{\psi_i\}_{i\in\mathbb{N}_{\leq\nu}}$, $\nu\in\mathbb{N}$, evaluation of (37) via the $\oplus$ operation indicates that a *unique* basis function ψ_i is *active* at x if $\psi_i(x)+[\widehat{e}_\infty]_i > \psi_j(x)+[\widehat{e}_\infty]_j$ for all $j\in\mathbb{N}_{\leq\nu}$, $j\neq i$. Indeed, the (possibly empty) subset $\mathscr{Y}_i\subset\mathscr{Y}\subset\mathbb{R}^n$ of states for which the i^{th} basis function is active in (37) is given by

$$\mathscr{Y}_i \doteq \left\{ x\in\mathscr{Y} \,\middle|\, \Gamma_{ij}(x,\widehat{e}_\infty) > 0\ \forall\ j\in\mathbb{N}_{\leq\nu},\ j\neq i \right\},\quad i\in\mathbb{N}_{\leq\nu}, \tag{38}$$

in which Γ_{ij} is defined for $i, j \in \mathbb{N}_{\leq \nu}$ by

$$\Gamma_{ij}(x, \widehat{e}_\infty) \doteq \langle x,\, \mathsf{M}(z_j - z_i)\rangle - \left[[\widehat{e}_\infty]_j + \frac{1}{2}\langle z_j,\, \mathsf{M}\, z_j\rangle\right] + [\widehat{e}_\infty]_i + \frac{1}{2}\langle z_i,\, \mathsf{M}\, z_i\rangle\,. \tag{39}$$

By inspection of (38) and (39), $\mathscr{Y}_i$ defines the interior of an intersection of a family of half spaces for each $i \in \mathbb{N}_{\leq \nu}$. Consequently, $\overline{\mathscr{Y}_i}$ is a closed convex polytope, $\cup_{i\in\mathbb{N}_{\leq\nu}} \overline{\mathscr{Y}_i} = \mathscr{Y}$, and the family $\{\overline{\mathscr{Y}_i}\}_{i\in\mathbb{N}_{\leq\nu}}$ defines a *Voronoi tessellation* of $\mathscr{Y}$.

The utility of specific basis functions in approximating the value function (1) on a subset $\mathscr{X}$ of $\mathscr{Y}$ can be assessed using the Hamiltonian (7). In particular, this Hamiltonian can be used to define a back substitution error [9] associated with the value function approximation $\widehat{W}$ of (37). By definition (38), with $x \in \mathscr{Y}_i$,

$$\begin{aligned} \widehat{W}(x) &= \psi_i(x) + [\widehat{e}_\infty]_i = \frac{1}{2}\langle x - z_i, \mathsf{M}(x - z_i)\rangle + [\widehat{e}_\infty]_i\,, \\ \nabla_x \widehat{W}(x) &= \nabla_x \psi_i(x) = p(x, z_i) \doteq \mathsf{M}(x - z_i)\,, \end{aligned} \tag{40}$$

so that the aforementioned back substitution error $h(\cdot, z_i) : \mathscr{Y}_i \to \mathbb{R}$ is defined by

$$h(\cdot, z_i) \doteq H(\cdot, p(\cdot, z_i)) \tag{41}$$

for any $i \in \mathbb{N}_{\leq \nu}$. Given a target bound $\delta_H \in \mathbb{R}_{>0}$ on the magnitude of this back substitution error, it is convenient to define the subset

$$\mathscr{X}_i \doteq \mathscr{Y}_i \cap \operatorname{co} \widehat{\mathscr{B}}_0(r_H^*)\,, \quad i \in \mathbb{N}_{\leq \nu}\,, \tag{42}$$

where $\widehat{\mathscr{B}}_0(r_H^*) \subset \mathbb{R}^n$ is a star-shaped neighbourhood of radius $r_H^* \in \mathbb{R}_{\geq 0}$, centred at $0 \in \mathbb{R}^n$, that is defined by

$$\begin{aligned} &\widehat{\mathscr{B}}_0(r) \doteq \left\{\rho\, \widehat{v} \in \mathbb{R}^n \,\middle|\, \widehat{v} \in \widehat{\mathscr{V}}_q\,,\ \rho \in [0, r]\right\}, \quad \widehat{\mathscr{V}}_q \doteq \{\widehat{v}_i \in \mathbb{R}^n \mid \|\widehat{v}_i\| = 1\,,\ i \in \mathbb{N}_{\leq q}\}\,, \\ &r_H^* \doteq \min_{i\in\mathbb{N}_{\leq\nu}} r_i^*\,, \quad r_i^* \doteq \sup\left\{r \in \mathbb{R}_{\geq 0} \,\middle|\, |h(y, z_i)| \leq \delta_H\ \forall\ y \in \mathscr{Y}_i \cap \widehat{\mathscr{B}}_0(r)\right\}, \end{aligned} \tag{43}$$

in which the normalized vertex set $\widehat{\mathscr{V}}_q$ is fixed a priori for some $q \in \mathbb{N}$. Note in particular that r_H^* is the maximal radius for which this star-shaped neighbourhood is within the magnitude back substitution error level set defined by δ_H. In this respect, $\operatorname{co} \widehat{\mathscr{B}}_0(r_H^*)$ is an approximation for the largest ball that fits within this level set.

As $\overline{\mathscr{Y}_i}$, $i \in \mathbb{N}_{\leq \nu}$ and $\operatorname{co} \widehat{\mathscr{B}}_0(r_H^*)$ are convex polytopes, it follows by (42) that $\overline{\mathscr{X}_i}$ is a convex polytope for each $i \in \mathbb{N}_{\leq \nu}$. The vertices $\mathsf{V}(\mathscr{X}_i)$ of $\overline{\mathscr{X}_i}$ can subsequently be sorted with respect to $h(\cdot, z_i)$, with the sorted set of vertices defined iteratively by

$$\begin{aligned} &[\mathsf{V}_i^*]_j \doteq [\mathsf{V}_i^*]_{j-1} \cup \{[x_i^*]_j\}\,, \quad [x_i^*]_j \doteq \underset{x \in \mathsf{V}(\mathscr{X}_i) \setminus [\mathsf{V}_i^*]_{j-1}}{\arg\max} \frac{1}{2} |h(x, z_i)|^2\,, \\ &[\mathsf{V}_i^*]_0 \doteq \emptyset\,, \end{aligned} \tag{44}$$

for all $j \in \mathbb{N}_{\leq K_i}$, where $K_i \in \mathbb{N}$ is at most the total number of vertices of $\overline{\mathscr{X}_i}$. The set of polytopes $\{\overline{\mathscr{X}_i}\}_{i \in \mathbb{N}_{\leq \nu}}$ defining the tessellation can subsequently be sorted with respect to their "worst-case" vertex, i.e. $[V_i^*]_1$ for all $i \in \mathbb{N}_{\leq \nu}$.

Remark 11 A refinement of the approximation $\widehat{\mathscr{B}}_0(r_H^*)$ of (43) for the Hamiltonian back substitution error level set is provided in [21]. There, an annulus in the state space is characterized and computed that approximates the boundary of the largest ball contained within the Hamiltonian back substitution error level set. □

An example of a Voronoi tessellation (42) corresponding to the max-plus dual space coordinate representation $\widehat{e}_\infty$ of a value function approximation generated by the standard max-plus eigenvector method (37) is illustrated in Fig. 1. There, the basis functions (10) employed are located at the x marked locations corresponding to $\{z_i\}_{i \in \mathbb{N}_{\leq \nu}}$, $\nu \doteq 89$. The set $\widehat{\mathscr{V}}$ selected in constructing the approximately circular boundary of the tessellation shown consists of $q \doteq 36$ unit vectors rotated through

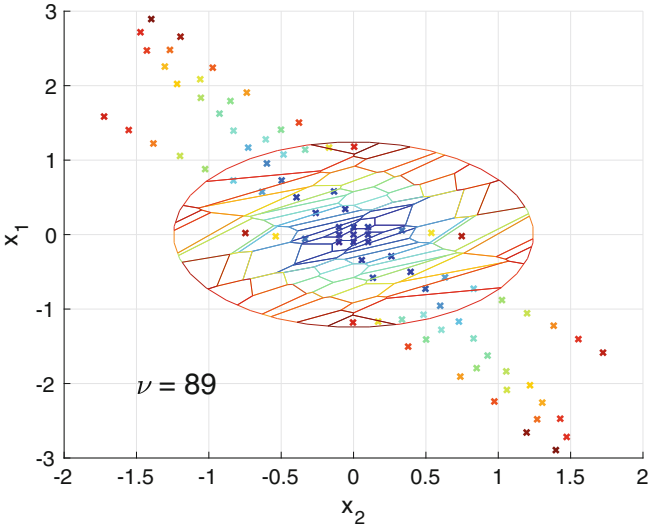


Fig. 1 Voronoi tessellation corresponding to a specific dual approximation $\widehat{e}_\infty$ and basis

multiples of $\pi/18$ radians in $\mathbb{R}^2$. Each component polytope shown corresponds to a specific basis function x, with pairings illustrated via matching colours.

Remark 12 Where $\sigma(x) \doteq \sigma \in \mathbb{R}^{n\times m}$ and $l(x) \doteq \frac{1}{2}\langle x,\ C'\,C\,x\rangle$ for all $x \in \mathscr{Y}$, (7) implies that the back substitution error (41) is $h(x, z_i) = -q_{z_i}(x) - r_{z_i}(x)$ for all $x \in \mathscr{X}_i$, $i \in \mathbb{N}_{\leq \nu}$, with $q_{z_i}, r_{z_i} : \mathscr{X}_i \to \mathbb{R}$ defined with respect to $A \doteq Df(0) \in \mathbb{R}^{n\times n}$, $F(x) \doteq f(x) - A\,x$, and $\Gamma(\mathsf{M}) \doteq A'\,\mathsf{M} + \mathsf{M}\,A + \frac{1}{\gamma^2}\,\mathsf{M}\,\sigma\,\sigma'\,\mathsf{M} + C'\,C$, by

$$q_{z_i}(x) \doteq \frac{1}{2}\langle x - z_i,\, \Gamma(\mathsf{M})\,(x - z_i)\rangle + \frac{1}{2}\langle z_i,\, C'\,C\,z_i\rangle + \langle x - z_i,\, (\mathsf{M}\,A + C'\,C)\,z_i + \mathsf{M}\,F(z_i)\rangle$$

$$r_{z_i}(x) \doteq \langle x - z_i,\, \mathsf{M}\,(F(x) - F(z_i))\rangle\,, \tag{45}$$

for all $x \in \mathscr{X}_i$. If the value function (1) corresponding to the linearized problem is finite, there exists an invertible $\mathsf{M} \in \Sigma$ in (10), (45) such that $\Gamma(\mathsf{M}) \in \Sigma_{>0}$. Consequently, q_{z_i} is convex. Hence, $h(\cdot, z_i)$ is concave on $\mathscr{X}_i$ if r_{z_i} is convex on $\mathscr{X}_i$, whereupon its infimum is attained at a vertex of $\overline{\mathscr{X}_i}$, i.e. at an element of $\mathsf{V}(\mathscr{X}_i)$. □

Basis Adaptation Using the sorted vertex and polytope sets generated via (44), the "worst-case" polytopes and corresponding basis functions can be identified. New basis functions can subsequently be evolved from existing basis function locations $z_i \in [\mathsf{V}_i^*]_{K_i}$, $i \in \mathbb{N}_{\leq \nu}$, via an ODE of the form

$$z_i^+ \doteq Z(\Pi_{\eta_i^+})\,, \quad \begin{cases} \dot{\Pi}_\eta = F(\Pi_\eta)\,, & \eta \in \mathbb{R}_{\geq 0}\,, \\ \Pi_0 = G(z_i)\,, \end{cases} \tag{46}$$

in which functions F, G, Z, and parameter $\eta_i^+ \in (0, \bar{\eta}]$ remain to be determined, and $\bar{\eta} \in \mathbb{R}_{>0}$ is fixed a priori.

Stopping Condition for (46) In integrating (46), a stopping condition based on the desired *ripple* in the Hamiltonian back substitution error may be used to determine η_i^+. Without loss of generality, consider the first vertex $y \doteq [x_i^*]_1 \in \mathsf{V}(\mathscr{X}_i)$ in the sequence (44). Should a new basis function be located at $z_i(\eta_i)$ for some $\eta_i \in (0, \bar{\eta}]$, and be rendered active at y a subsequent application of the standard max-plus eigenvector method (37), the magnitude of the Hamiltonian at y is revised to $|h(y, z_i(\eta))|$, with the magnitude of the change induced there being $|h(y, z_i(\eta)) - h(y, z_i)|$. In view of (42), a candidate for η_i^+ in (46) is thus

$$\eta_i^+ \doteq \sup\left\{\eta \in (0, \bar{\eta}] \,\middle|\, \begin{array}{c} |h(y, z_\eta)| < (1 - \mu)\,\delta_H\,, \\ |h(y, z_\eta) - h(y, z_i)| < \mu\,\delta_H \end{array}\right\}, \tag{47}$$

where $\mu \in (0, 1)$, $\bar{\eta} \in \mathbb{R}_{>0}$ are fixed. By definition, this choice of η_i^+ establishes a reduction in the Hamiltonian magnitude, while limiting the ripple induced by the switch to the new basis function, see Fig. 2. The upper bound $\bar{\eta}$ ensures that a new basis function located at $z_i^+ = z_i(\eta_i^+)$ is sufficiently close to z_i so as to allow

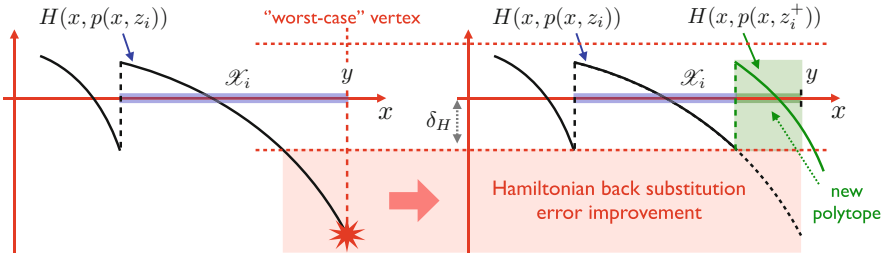


Fig. 2 Hamiltonian back substitution error improvement via addition of a new basis function, located at z_i^+, and its corresponding polytope

information flow between its location and that of the other basis functions in the subsequent application of the max-plus eigenvector method, see [1, Remark 4.24, p. 83]. It may be noted that only one new row and one new column of $\widehat{B}_\tau$ need be computed per basis function added, see also Remark 10.

Gradient Descent Yielding (46) By inspection of (7), (41), $h(y, \cdot)$ is differentiable, so that

$$D_z \, \frac{1}{2}|h(y,z)|^2 \, \zeta = \langle \zeta, \, h(y,z) \, \nabla_z h(y,z) \rangle$$

for all $\zeta \in \mathbb{R}^n$, in which the Riesz representation $\nabla_z h(y,z)$ of the Fréchet derivative of $h(y, \cdot)$ at $z \in \mathbb{R}^n$ is given by

$$\nabla_z h(y,z) = -\mathsf{M} \, \nabla_p H(y, \, p(y,z)) = \mathsf{M}\,[f(y) + \frac{1}{\gamma^2} \, g(y) \, g(y)' \, p(y,z)] \qquad (48)$$

for all $y, z \in \mathbb{R}^n$. Given some fixed $\epsilon \in \mathbb{R}_{>0}$, a normalized gradient descent direction motivates a candidate for system (46), defined by

$$F(\zeta) \doteq -\operatorname{sgn}(h(y,\zeta)) \, \frac{\nabla_z h(y,\zeta)}{\|\nabla_z h(y,\zeta)\| + \epsilon} \, , \quad G(\zeta) \doteq \zeta \, , \quad Z(\zeta) \doteq \zeta \, , \qquad (49)$$

for all $\zeta \in \mathbb{R}^n$. By inspection of (48), F of (49) is related to the flow of the optimal dynamics, either backward or forward in time depending on the sign of the Hamiltonian and M. The Hamiltonian squared may be reduced, or at least not increased, by adding (locally) a new basis function in direction $F(z_i)$ from the existing basis function located at $z_i \in \mathbb{R}^n$ as per (10). Its precise location can be determined by applying (47) and (49) in (46).

Characteristics Yielding (46) Given fixed $z_i \in \mathbb{R}^n$, and H as per (7), let $y \in \mathscr{X}_i$ be such that $0 = H(y, \mathsf{M}(y - z_i))$. Define ξ, π, ζ via the Cauchy problem

$$\begin{aligned} \dot{\xi}_\eta &= \nabla_p H(\xi_\eta, \pi_\eta), & \xi_0 &= y, \\ \dot{\pi}_\eta &= -\nabla_x H(\xi_\eta, \pi_\eta) & \pi_0 &= \mathsf{M}(y - z_i), \\ \zeta_\eta &= \xi_\eta - \mathsf{M}^{-1}\pi_\eta, \end{aligned} \tag{50}$$

for all $\eta \in [0, \bar{\eta}]$, $\bar{\eta} \in \mathbb{R}_{>0}$, in which $\nabla_p H(x, p)$, $x, p \in \mathbb{R}^n$, denotes the Riesz representation of the Fréchet derivative of $H(x, \cdot)$ at p, and likewise for $\nabla_x H(x, p)$ at x. By inspection, (50) defines a characteristic curve of (6), so that

$$0 = H(y, \mathsf{M}(y - z_i)) = H(\xi_0, \pi_0) = H(\xi_\eta, \pi_\eta) = H(\xi_\eta, \mathsf{M}(\xi_\eta - \zeta_\eta))$$

for all $\eta \in [0, \bar{\eta}]$. Note in particular that $y \in \mathscr{X}_i$ corresponds (by definition) to a state for which the Hamiltonian back substitution error is zero, while the characteristics generated by (50) show how this state and the corresponding basis location z_i may be evolved to other locations without increasing this error. This motivates the introduction of new basis functions defined along these characteristics, via an alternative candidate for system (46), defined in terms of $\Pi \doteq (\xi, \pi) \in \mathscr{Z}^2$ and $y \in \mathscr{X}_i$ by

$$F(\Pi) \doteq \begin{pmatrix} \nabla_p H(\xi, \pi) \\ -\nabla_x H(\xi, \pi) \end{pmatrix}, \quad G(z_i) \doteq \begin{pmatrix} y \\ \mathsf{M}(y - z_i) \end{pmatrix}, \quad Z(\Pi) \doteq \xi - \mathsf{M}^{-1}\pi. \tag{51}$$

New basis functions are thus added along trajectories corresponding to the approximate optimal dynamics, as a consequence of the characteristic curves employed.

Remark 13 An intrinsic difficulty with the standard max-plus eigenvector method (37) is that there is no guarantee that a specific basis function located away from the origin will be rendered *active* in the subsequent value function approximation. Consequently, in the basis adaptation iteration proposed, there is likewise no guarantee that an added basis function will be rendered *active* in the subsequent value function approximation. This of course applies irrespective of whether gradient descent (46), (47), (49) or characteristics (46), (47), (51) is applied. Notwithstanding this difficulty, it is known that the value function approximation error obtained is non-increasing with increasing density of the set of basis function locations $\{z_i\}_{i \in \mathbb{N}_{\leq \nu}}$, and that this error converges to zero in limit of a densely defined basis, see [1, Theorem 5.14, p. 118]. Consequently, without pruning, the basis adaptation iteration must at worst preserve the prior value function approximation error.

Algorithm The aforementioned steps of max-plus eigenvector based value function approximation, back substitution error level set tessellation and basis sorting, and basis adaptation can be combined to yield an adaptive max-plus eigenvector method for approximately solving optimal control problems of the form (1), see

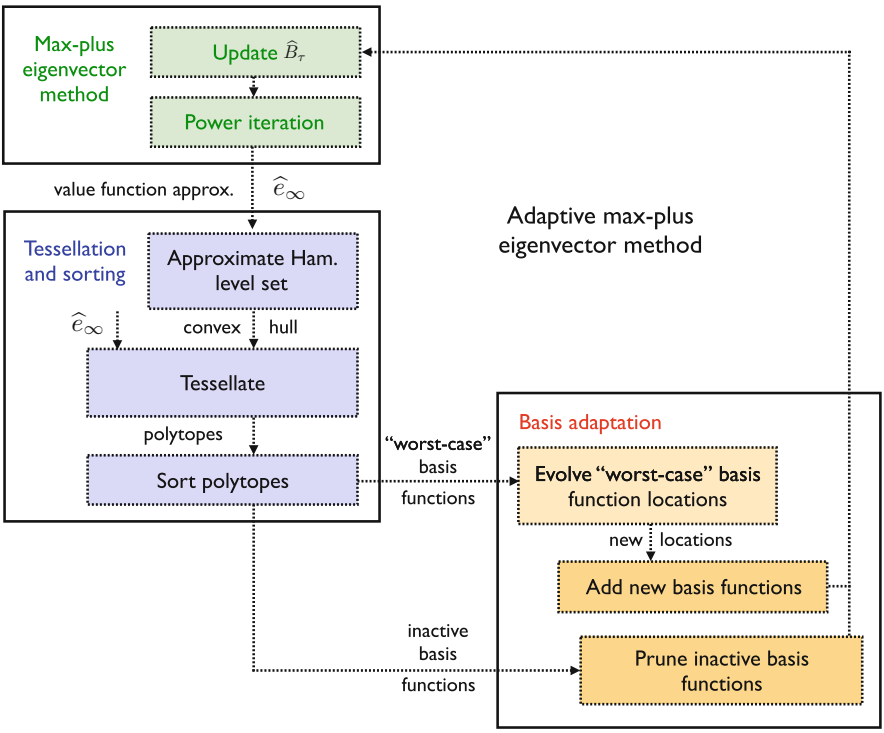


Fig. 3 Algorithmic overview of the adaptive max-plus eigenvector method

Fig. 3. Within the main adaptation loop of the associated algorithm, the basis (10) available at the current step is used to compute an approximation $\widehat{W}$ of the value function W of (1), using the standard max-plus eigenvector method (37). Based on this approximation, the back substitution error is computed via the attendant Hamiltonian (7), and a level set corresponding to a target error is approximated via a convex polytope in the state space. A Voronoi tessellation of this level set approximation is subsequently computed, in which each component convex polytope corresponds to a subset of states on which a single basis function is *active* in the value function approximation. The polytope vertices, and subsequently the polytopes themselves, are sorted as per (44) according to their back substitution error. The "worst-case" polytopes and their associated basis functions are subsequently identified, and their respective locations evolved in directions that reduce the expected back substitution error, yielding locations of new basis functions to be added. Basis functions that are inactive in the value function approximation on the level set are pruned. The basis is updated accordingly, and the steps above iterated.

Remark 14 The time complexity of the aforementioned algorithm is dominated by computation of *(i)* the matrix $\widehat{B}_\tau$ in the standard max-plus eigenvector method (37), and *(ii)* the Voronoi tessellation of the approximate Hamiltonian level set.

In *(i)*, and as indicated in Remark 10, computation of $\widehat{B}_\tau \in \overline{\mathbb{R}}^{\nu \times \nu}$ for a basis of cardinality $\nu \in \mathbb{N}$ requires integration of a separate nth order ODE for every entry of $\widehat{B}_\tau$, where $n \in \mathbb{N}$ is the state dimension. Using a standard Runge-Kutta scheme [1, p. 122], the time complexity for computation of $\widehat{B}_\tau$ thus exhibits quadratic growth in ν. This is problematic as the basis adaptation iteration proceeds and ν increases. However, by storing $\widehat{B}_\tau$ from the previous iteration, and restricting computation to new rows and columns corresponding to those basis functions added in the most recent basis adaptation iteration, this quadratic growth can be controlled.

In *(ii)*, computation of the Voronoi tessellation requires vertex enumeration for a set of at most ν polytopes, defined via (38). A benchmark pivoting algorithm [22, 23] that implements this vertex enumeration suffers from a curse-of-dimensionality, with worst-case time complexity of

$$O\left(n\,\nu\,(\nu-1)\,(\nu+n-1)\,\binom{\nu-1}{n}\right).$$

State-of-the-art vertex enumeration algorithms provide some improvement, albeit with a similarly afflicted worst-case time complexity of $O(\nu^{\lceil n/2 \rceil})$, see [24, p. 168] and the references cited therein. In practice however, the worst case is rarely encountered, as the number of hyperplanes defining facets for individual polytopes is typically far fewer than the basis cardinality ν, while degenerate vertices/hyperplanes can in-principle be controlled by basis pruning. Together, these aspects can significantly reduce the combinatorial term dominating the worst-case described. □

Remark 15 A secondary issue affecting application of the standard max-plus eigenvector method (37) within the main loop of adaptive basis algorithm concerns the choice of the short time horizon τ. In particular, applicability of the error analysis of [1, Ch.5] requires that τ satisfy a lower bound, see for example [1, (5.91), p. 121], that scales with the inverse of basis cardinality ν. As a consequence, it may be required that τ be reduced as the adaptive basis iteration proceeds. Anecdotal improvements in convergence of the basis adaptation algorithm have been observed in examples by reducing τ in this way. Further details are omitted. □

5 Examples

Two examples are considered, corresponding to linear and nonlinear dynamics.

Linear Dynamics Suppose the dynamics and running cost are given by

$$f(x) \doteq \begin{pmatrix} -1 & 0 \\ -1 & -1 \end{pmatrix} x\,,\ \sigma(x) \doteq \begin{pmatrix} 0.5 \\ 0 \end{pmatrix},$$

$$l(x) \doteq \frac{1}{2}\,|x|^2\,,\ \gamma \doteq 2\,,\ \mathsf{M} \doteq 0.2\,\mathsf{I}_2\,, \tag{52}$$

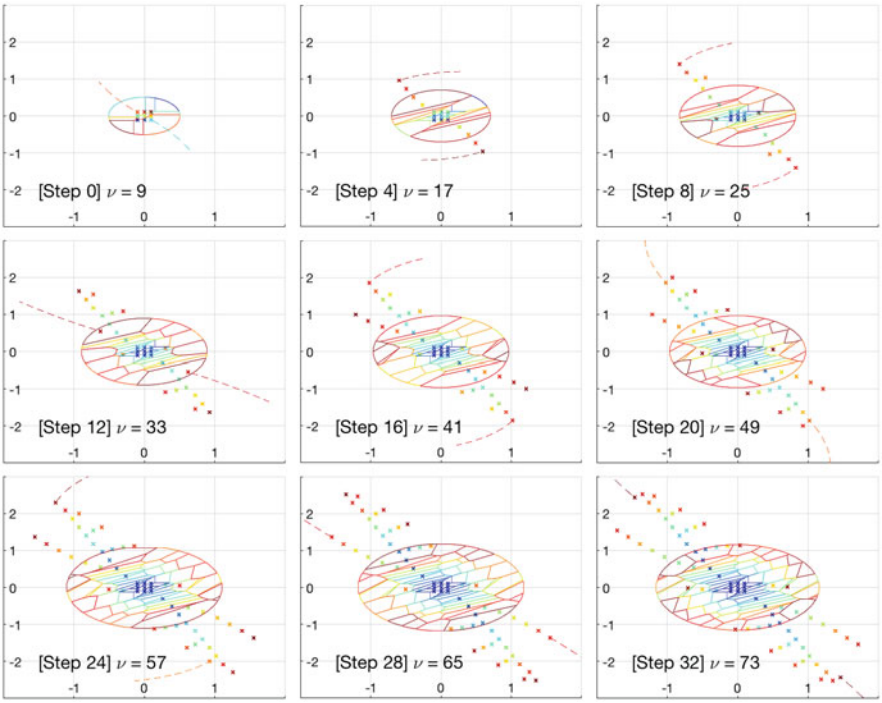


Fig. 4 Linear dynamics—basis evolution using characteristics (46), (47), (51)

for all $x \in \mathbb{R}^2$, in which $\mathsf{I}_2 \in \mathbb{R}^{2\times 2}$ denotes the identity. The algorithm of Fig. 3 is employed to approximate the value function. Within each iteration of this algorithm, the max-plus eigenvector method (37) is applied with $\tau \doteq 0.5$. The characteristics approach (46), (47), (51) is used to evolve the basis involved, with a target back substitution error of $\delta_H \doteq 0.1$, and using $\mu \doteq 0.25$ and $\bar{\eta} \doteq 0.5$. Figures 4, 5, and 6 illustrate evolution of the basis (and Voronoi tessellation), Hamiltonian, and value function approximation respectively, via (42), (43), (44) with $q \doteq 36$. The basis functions used are of the form (10), with M fixed as per (52), and centres $\{z_i\}_{i\in\mathbb{N}_{\leq\nu}}$ located at the x marks shown. The initial basis is as depicted in the top left panel of Fig. 4. The dashed lines indicate two basis function evolution trajectories, each originating from the location z_i, $i \in \mathbb{N}_{\leq\nu}$, of a corresponding worst-case basis function. The trajectories are obtained by integrating (46) via (51) up to the stopping condition η_i^+ specified by (47). Two new basis functions are added per iteration, at the worst-case locations z_i^+ thus obtained.

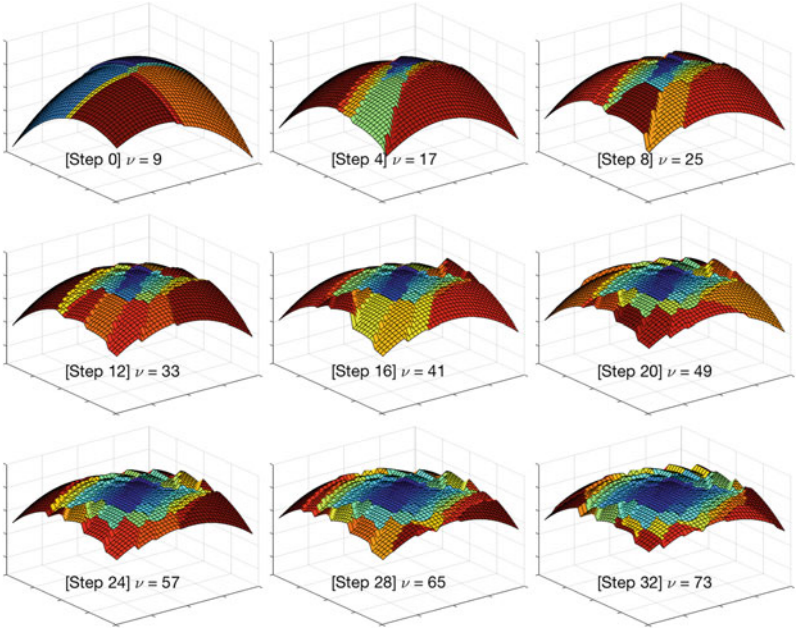


Fig. 5 Linear dynamics—Hamiltonian evolution using characteristics (46), (47), (51)

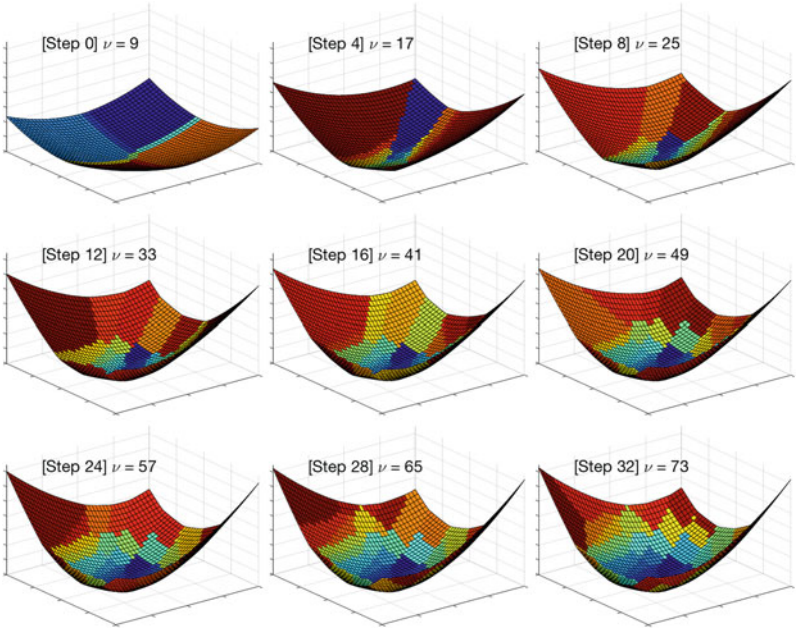


Fig. 6 Linear dynamics—value function evolution using characteristics (46), (47), (51)

Nonlinear Dynamics Suppose the dynamics and running cost are given by

$$f(x) \doteq \begin{pmatrix} -2\,x_1\,[1+\frac{1}{2}\,\tan^{-1}(3x_2^2/2)] \\ \frac{1}{2}\,x_1 - 3\,x_2\,\exp(-x_1/3) \end{pmatrix},\ \sigma(x) \doteq \sigma \doteq \mathsf{I}_2,$$

$$l(x) \doteq \frac{1}{2}\,|x|^2,\ \gamma \doteq 1,\ \mathsf{M} \doteq -0.1\,\mathsf{I}_2, \tag{53}$$

for all $x = (x_1, x_2) \in \mathbb{R}^2$, see [1, p. 127]. The algorithm of Fig. 3 is again employed to approximate the value function, with the encapsulated max-plus eigenvector method (37) applied with $\tau \doteq 0.1$. The gradient descent approach (46), (47), (49) is used to evolve the basis, with target back substitution error $\delta_H \doteq 0.1$. Figures 7, 8, and 9 and illustrate evolution of the basis, Hamiltonian, and value function approximation respectively. The settings are otherwise analogous to the preceding linear example. It is observed that the basis is comparatively uniformly distributed, a feature that was also observed in applying characteristics (46), (47), (51) to the same problem data. The details are omitted.

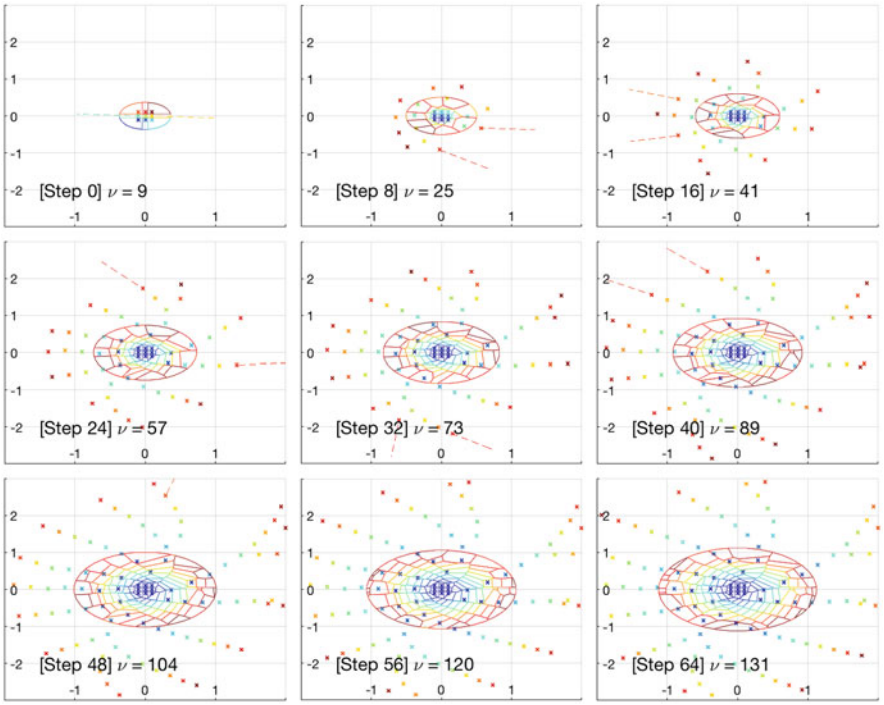


Fig. 7 Nonlinear dynamics—basis evolution using gradient descent (46), (47), (49)

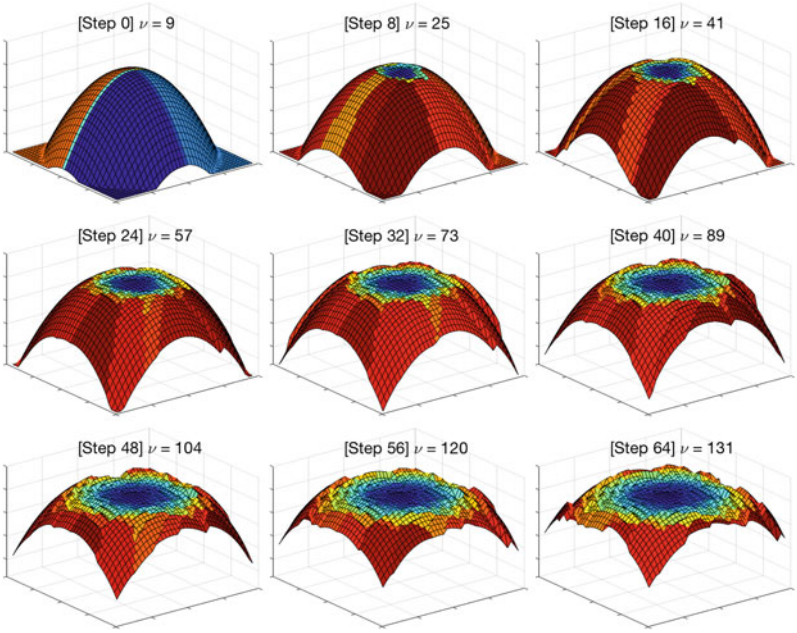


Fig. 8 Nonlinear dynamics—Hamiltonian evolution using gradient descent (46), (47), (49)

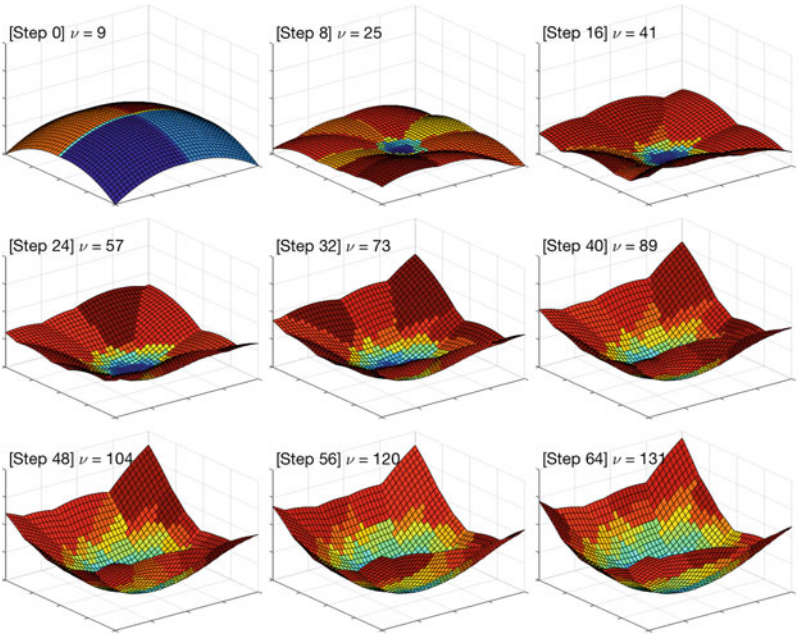


Fig. 9 Nonlinear dynamics—value evolution using gradient descent (46), (47), (49)

6 Conclusion

An adaptive max-plus eigenvector method is proposed, based on an iteration involving a standard max-plus eigenvector method, a tessellation and sorting step (for "worst-case" basis function identification), and a basis adaptation step. The proposed method is illustrated via two simple examples.

Acknowledgements This research was partially supported by AFOSR/AOARD grant FA2386-16-1-4066.

Appendix

In the statement of Theorem 3, the semiconvex transform D_φ is as specified by (12), and its candidate inverse is as per (13). For convenience, formally define

$$\mathsf{D}_\varphi^\sharp\, a \doteq \int_{\mathbb{R}^n}^{\oplus} \varphi(\cdot, z) \otimes a(z)\, dz\,, \qquad a \in \mathrm{dom}\,(\mathsf{D}_\varphi^\sharp) \doteq \mathscr{R}_-^{\mathsf{K}}\,, \tag{54}$$

in which $\mathscr{R}_-^{\mathsf{K}}$ is as per (14).

Lemma 16 *With $-\mathsf{K} \in \Sigma_{\mathsf{M}} \cup \{\mathsf{M}\}$ fixed, D_φ and $\mathsf{D}_\varphi^\sharp$ of (12) and (54) satisfy the following properties:*

1) $\mathrm{dom}\,(\mathsf{D}_\varphi^\sharp) = \mathrm{ran}\,(\mathsf{D}_\varphi) = \mathscr{R}_-^{\mathsf{K}} \subset \mathscr{S}_-^{-\mathsf{M}}$;
2) $\mathrm{ran}\,(\mathsf{D}_\varphi^\sharp) = \mathrm{dom}\,(\mathsf{D}_\varphi) = \mathscr{S}_+^{\mathsf{K}} \subset \mathscr{S}_+^{-\mathsf{M}}$;
3) $\mathsf{D}_\varphi\, \mathsf{D}_\varphi^\sharp = \mathsf{I}$ on $\mathrm{dom}\,(\mathsf{D}_\varphi^\sharp) = \mathscr{R}_-^{\mathsf{K}}$;
4) $\mathsf{D}_\varphi^\sharp\, \mathsf{D}_\varphi = \mathsf{I}$ on $\mathrm{dom}\,(\mathsf{D}_\varphi) = \mathscr{S}_+^{\mathsf{K}}$;
5) $\mathscr{R}_-^{-\mathsf{M}} = \mathscr{S}_-^{-\mathsf{M}}$;
6) If $-\mathsf{K} > \mathsf{M}$ then $\mathscr{S}_+^{-\mathsf{M}} \setminus \mathscr{S}_+^{\mathsf{K}} \neq \emptyset$ and $\mathscr{R}_-^{-\mathsf{M}} \setminus \mathscr{R}_-^{\mathsf{K}} \neq \emptyset$.

The following observations are useful in establishing Lemma 16.

1. Given invertible $\mathsf{M} \in \Sigma$,

$$\langle x,\, z\rangle + \frac{1}{2}\langle x,\, \mathsf{M}\, x\rangle = \frac{1}{2}\langle x - \zeta(z),\, \mathsf{M}\,(x - \zeta(z))\rangle - \frac{1}{2}\langle \zeta(z),\, \mathsf{M}\,\zeta(z)\rangle \tag{55}$$

   for all $x, z \in \mathbb{R}^n$, where $\zeta(z) \doteq -\mathsf{M}^{-1}\, z \in \mathbb{R}^n$.
2. Given $\mathsf{O}_1, \mathsf{O}_2 \in \Sigma$ satisfying $\mathsf{O}_1 < \mathsf{O}_2$,

$$\mathscr{S}_+^{\mathsf{O}_1} \subset \mathscr{S}_+^{\mathsf{O}_2}\,, \qquad \mathscr{S}_-^{\mathsf{O}_1} \subset \mathscr{S}_-^{\mathsf{O}_2}\,. \tag{56}$$

Proof of Lemma 16 Assertion 1) By definitions (14) and (54),

$$\operatorname{ran}(\mathsf{D}_\varphi) = \mathscr{R}_-^{\mathsf{K}} = \operatorname{dom}(\mathsf{D}_\varphi^\sharp).$$

Fix an arbitrary $a \in \mathscr{R}_-^{\mathsf{K}}$. Applying definition (14), there exists a $\psi \in \mathscr{S}_+^{\mathsf{K}}$ such that $\mathsf{D}_\varphi\, \psi = a$. Define $\psi_+, \psi_\pm : \mathbb{R}^n \to \overline{\mathbb{R}}$ by

$$\psi_+(x) \doteq \psi(x) + \frac{1}{2}\langle x,\, \mathsf{K}\, x\rangle\,, \quad \psi_\pm(x) \doteq \psi_+(x) + \frac{1}{2}\langle x,\, -(\mathsf{K}+\mathsf{M})\, x\rangle\,, \tag{57}$$

for all $x \in \mathbb{R}^n$, and note that

$$\psi_\pm(x) = \psi(x) - \frac{1}{2}\langle x,\, \mathsf{M}\, x\rangle \tag{58}$$

for all $x \in \mathbb{R}^n$. By inspection of (57), ψ_+ is convex and lower closed on $\mathbb{R}^n$, as $\psi \in \mathscr{S}_+^{\mathsf{K}}$. Hence, $\psi_\pm$ is also convex and lower closed on $\mathbb{R}^n$, as $-(\mathsf{K}+\mathsf{M}) \geq 0$ by definition of $-\mathsf{K} \in \Sigma_{\mathsf{M}} \cup \{\mathsf{M}\}$. Hence, the convex conjugate $\psi_\pm^* : \mathbb{R}^n \to \overline{\mathbb{R}}$ of $\psi_\pm$ is also convex and lower closed [25, Theorem 5, p. 16], with

$$\psi_\pm^*(z) \doteq \int_{\mathbb{R}^n}^{\oplus} \langle z,\, x\rangle \otimes (-\psi_\pm(x))\, dx \tag{59}$$

for all $z \in \mathbb{R}^n$. Define $\widetilde{a} : \mathbb{R}^n \to \overline{\mathbb{R}}$ by

$$\widetilde{a}(\zeta) \doteq -\psi_\pm^*(-\mathsf{M}\,\zeta) - \frac{1}{2}\langle \zeta,\, \mathsf{M}\,\zeta\rangle \tag{60}$$

for all $\zeta \in \mathbb{R}^n$. Note that $\widetilde{a}$ is upper closed, as $\psi_\pm^*$ and $\frac{1}{2}\langle \cdot,\, \mathsf{M}\,\cdot\rangle$ are lower closed. Furthermore, $\zeta \mapsto \tilde{a}(\zeta) - \frac{1}{2}\langle \zeta,\, -\mathsf{M}\,\zeta\rangle$ is concave, as $\psi_\pm^*$ is convex. Hence, $\widetilde{a} \in \mathscr{S}_-^{-\mathsf{M}}$. Recalling that $\psi_\pm^*$ is the convex conjugate of $\psi_\pm$, (58), (60) imply that

$$\begin{aligned}
\widetilde{a}(\zeta) &= -\psi_\pm^*(-\mathsf{M}\,\zeta) - \frac{1}{2}\langle \zeta,\, \mathsf{M}\,\zeta\rangle = -\int_{\mathbb{R}^n}^{\oplus} \langle -\mathsf{M}\,\zeta,\, x\rangle \otimes (-\psi_\pm(x))\, dx - \frac{1}{2}\langle \zeta,\, \mathsf{M}\,\zeta\rangle \\
&= -\int_{\mathbb{R}^n}^{\oplus} \frac{1}{2}\langle x,\, \mathsf{M}\, x\rangle - \langle \mathsf{M}\,\zeta,\, x\rangle + \frac{1}{2}\langle \zeta,\, \mathsf{M}\,\zeta\rangle - \psi(x)\, dx \\
&= (\mathsf{D}_\varphi\, \psi)(\zeta) = a(\zeta)
\end{aligned} \tag{61}$$

for all $\zeta \in \mathbb{R}^n$. That is, $a = \widetilde{a} \in \mathscr{S}_-^{-\mathsf{M}}$, and as $a \in \mathscr{R}_-^{\mathsf{K}}$ is arbitrary, $\mathscr{R}_-^{\mathsf{K}} \subset \mathscr{S}_-^{-\mathsf{M}}$.

2) By definitions (12) and (54), $\operatorname{dom}(\mathsf{D}_\varphi) = \mathscr{S}_+^{\mathsf{K}}$ and $\operatorname{dom}(\mathsf{D}_\varphi^\sharp) = \mathscr{R}_-^{\mathsf{K}}$. Fix an arbitrary $a \in \operatorname{dom}(\mathsf{D}_\varphi^\sharp) = \mathscr{R}_-^{\mathsf{K}}$. Following the proof of Assertion 1), there exists a $\psi \in \mathscr{S}_+^{\mathsf{K}}$ such that $\mathsf{D}_\varphi \psi = a$, defining a convex and lower closed $\psi_\pm : \mathbb{R}^n \to \overline{\mathbb{R}}$ as

per (57), (58). Note further that $a = \widetilde{a} \in \mathscr{S}_{-}^{-\mathsf{M}}$, where $\widetilde{a}$ is defined as per (60). As $\psi_{\pm} = \psi_{\pm}^{**}$, see [25, Theorem 5, p. 16], (54), (58) and (60) imply that

$$\begin{aligned}\psi(x) &= (\psi_{\pm}^{*})^{*}(x) + \frac{1}{2}\langle x,\, \mathsf{M}\, x\rangle = \int_{\mathbb{R}^n}^{\oplus} \langle x,\, z\rangle + \frac{1}{2}\langle x,\, \mathsf{M}\, x\rangle - \psi_{\pm}^{*}(z)\, dz \\ &= \int_{\mathbb{R}^n}^{\oplus} \frac{1}{2}\langle x-\zeta,\, \mathsf{M}\,(x-\zeta)\rangle + \left[-\psi_{\pm}^{*}(-\mathsf{M}\,\zeta) - \frac{1}{2}\langle \zeta,\, \mathsf{M}\,\zeta\rangle\right] d\zeta \\ &= \int_{\mathbb{R}^n}^{\oplus} \frac{1}{2}\langle x-\zeta,\, \mathsf{M}\,(x-\zeta)\rangle \otimes \widetilde{a}(\zeta)\, d\zeta = (\mathsf{D}_{\varphi}^{\sharp}\,\widetilde{a})(x) = (\mathsf{D}_{\varphi}^{\sharp}\, a)(x) \end{aligned} \tag{62}$$

where the third equality uses (55) and the change of variable $\zeta \doteq -\mathsf{M}^{-1} z \in \mathbb{R}^n$. Hence, $\mathrm{ran}\,(\mathsf{D}_{\varphi}^{\sharp}) \subset \mathscr{S}_{+}^{\mathsf{K}}$, as $\psi = \mathsf{D}_{\varphi}^{\sharp}\, a \in \mathscr{S}_{+}^{\mathsf{K}}$, and $a \in \mathrm{dom}\,(\mathsf{D}_{\varphi}^{\sharp})$ is arbitrary.

Alternatively, fix an arbitrary $\psi \in \mathscr{S}_{+}^{\mathsf{K}}$, and construct $\psi_{\pm}$ directly using (57), (58). Subsequently define $a = \widetilde{a} \in \mathscr{S}_{-}^{-\mathsf{M}}$ via (60), and note that $\psi = \mathsf{D}_{\varphi}^{\sharp}\, a$ as per (62). That is, $\mathscr{S}_{+}^{\mathsf{K}} \subset \mathrm{ran}\,(\mathsf{D}_{\varphi}^{\sharp})$, as $\psi \in \mathscr{S}_{+}^{\mathsf{K}}$ is arbitrary. Recalling the earlier conclusion $\mathrm{ran}\,(\mathsf{D}_{\varphi}^{\sharp}) \subset \mathscr{S}_{+}^{\mathsf{K}}$ yields $\mathscr{S}_{+}^{\mathsf{K}} = \mathrm{ran}\,(\mathsf{D}_{\varphi}^{\sharp})$. Finally, as $-\mathsf{K} \in \Sigma_{\mathsf{M}} \cup \{\mathsf{M}\}$, i.e. $\mathsf{K} \leq -\mathsf{M}$, (56) implies that $\mathscr{S}_{+}^{\mathsf{K}} \subset \mathscr{S}_{+}^{-\mathsf{M}}$ as required.

3), 4) Applying Assertions 1) and 2), the compositions $\mathsf{D}_{\varphi}\,\mathsf{D}_{\varphi}^{\sharp} : \mathscr{R}_{-}^{\mathsf{K}} \to \mathscr{R}_{-}^{\mathsf{K}}$ and $\mathsf{D}_{\varphi}^{\sharp}\,\mathsf{D}_{\varphi} : \mathscr{S}_{+}^{\mathsf{K}} \to \mathscr{S}_{+}^{\mathsf{K}}$ are well-defined. Applying (61) and (62) yields $a = \mathsf{D}_{\varphi}\,\mathsf{D}_{\varphi}^{\sharp}\, a$ and $\psi = \mathsf{D}_{\varphi}^{\sharp}\,\mathsf{D}_{\varphi}\,\psi$ for any $a \in \mathscr{R}_{-}^{\mathsf{K}} = \mathrm{dom}\,(\mathsf{D}_{\varphi}^{\sharp})$, $\psi \in \mathscr{S}_{+}^{\mathsf{K}} = \mathrm{dom}\,(\mathsf{D}_{\varphi})$, i.e.,

$$\begin{aligned} \mathsf{D}_{\varphi}\,\mathsf{D}_{\varphi}^{\sharp} &= \mathsf{I}\,, & \mathrm{dom}\,(\mathsf{D}_{\varphi}\,\mathsf{D}_{\varphi}^{\sharp}) &= \mathscr{R}_{-}^{\mathsf{K}}\,, \\ \mathsf{D}_{\varphi}^{\sharp}\,\mathsf{D}_{\varphi} &= \mathsf{I}\,, & \mathrm{dom}\,(\mathsf{D}_{\varphi}^{\sharp}\,\mathsf{D}_{\varphi}) &= \mathscr{S}_{+}^{\mathsf{K}}\,. \end{aligned}$$

5) Fix any $a \in \mathscr{S}_{-}^{-\mathsf{M}}$. Define $a_{\mp} : \mathbb{R}^n \to \overline{\mathbb{R}}$ by

$$a_{\mp}(\zeta) = -\left[a(\zeta) - \frac{1}{2}\langle \zeta,\, -\mathsf{M}\,\zeta\rangle\right],$$

for all $\zeta \in \mathbb{R}^n$. Observe that $a_{\mp}$ is convex and lower closed by definition of a, so that $a_{\mp} = a_{\mp}^{**}$, see [10, Theorem 5, p. 16]. Hence,

$$\begin{aligned} a(y) &= -(a_{\mp}^{*})^{*}(y) - \frac{1}{2}\langle y,\, \mathsf{M}\, y\rangle = -\int_{\mathbb{R}^n}^{\oplus} \langle y,\, z\rangle + \frac{1}{2}\langle y,\, \mathsf{M}\, y\rangle - a_{\mp}^{*}(z)\, dz \\ &= -\int_{\mathbb{R}^n}^{\oplus} \frac{1}{2}\langle y-\zeta,\, \mathsf{M}\,(y-\zeta)\rangle - \left[\frac{1}{2}\langle \zeta,\, \mathsf{M}\,\zeta\rangle + a_{\mp}^{*}(\mathsf{M}\,\zeta)\right] d\zeta \end{aligned} \tag{63}$$

where the final equality follows by application of (55). Define $\psi_{\mp} : \mathbb{R}^n \to \overline{\mathbb{R}}$ by

$$\psi_{\mp}(\zeta) \doteq \frac{1}{2}\langle \zeta,\, \mathsf{M}\,\zeta\rangle + a^*_{\mp}(\mathsf{M}\,\zeta) \tag{64}$$

for all $\zeta \in \mathbb{R}^n$. As $a^*_{\mp}$ is also convex and lower closed [10], $\psi_{\mp} \in \mathscr{S}_+^{-\mathsf{M}}$. Substituting (64) in (63), and recalling (12) with $-\mathsf{K} \doteq \mathsf{M} \in \Sigma_{\mathsf{M}} \cup \{\mathsf{M}\}$ selected in the definition of dom (D_φ),

$$a(y) = -\int_{\mathbb{R}^n}^{\oplus} \frac{1}{2}\langle y-\zeta,\, \mathsf{M}\,(y-\zeta)\rangle \otimes (-\psi_{\mp}(\zeta))\, d\zeta = (\mathsf{D}_\varphi\, \psi_{\mp})(y)$$

for all $y \in \mathbb{R}^n$. Hence, $a \in \mathscr{R}_-^{-\mathsf{M}}$ by inspection of (14). As $a \in \mathscr{S}_-^{-\mathsf{M}}$ is arbitrary, it follows immediately that $\mathscr{S}_-^{-\mathsf{M}} \subset \mathscr{R}_-^{-\mathsf{M}}$. However, applying Assertions 1) and 2) for $\mathsf{K} = -\mathsf{M}$ yields that $\mathscr{R}_-^{-\mathsf{M}} \subset \mathscr{S}_-^{-\mathsf{M}}$, so that the claimed property holds.

6) Let $\mathsf{D}_\varphi^{\mathsf{M}}$ and $\mathsf{D}_\varphi^{\mathsf{M}\sharp}$ denote the operators D_φ and $\mathsf{D}_\varphi^\sharp$ of (12) and (54) with respective domains given by $\mathrm{dom}\,(\mathsf{D}_\varphi^{\mathsf{M}}) = \mathscr{S}_+^{-\mathsf{M}}$ and $\mathrm{dom}\,(\mathsf{D}_\varphi^{\mathsf{M}\sharp}) = \mathscr{R}_-^{-\mathsf{M}}$. Fix any $-\mathsf{K} \in \Sigma_{\mathsf{M}}$, i.e. so that $-\mathsf{K} > \mathsf{M}$. Observe that $\mathscr{S}_+^{-\mathsf{M}} \setminus \mathscr{S}_+^{\mathsf{K}} \neq \emptyset$, as it contains $\frac{1}{2}\langle \cdot,\, \mathsf{N}\,\cdot\rangle$ for any $\mathsf{N} \in \Sigma$ satisfying $\mathsf{M} < \mathsf{N} < -\mathsf{K}$. Fix any $\psi \in \mathscr{S}_+^{-\mathsf{M}} \setminus \mathscr{S}_+^{\mathsf{K}}$. By Assertions 1) through 5) applied to $\mathsf{D}_\varphi^{\mathsf{M}}$ and $\mathsf{D}_\varphi^{\mathsf{M}\sharp}$, there exists an $a \in \mathscr{S}_-^{-\mathsf{M}} = \mathscr{R}_-^{-\mathsf{M}} \supset \mathscr{R}_-^{\mathsf{K}}$ such that $\mathsf{D}_\varphi^{\mathsf{M}}\,\psi = a$ and $\mathsf{D}_\varphi^{\mathsf{M}\sharp}\,a = \psi$. Suppose that $a \in \mathscr{R}_-^{\mathsf{K}} = \mathrm{dom}\,(\mathsf{D}_\varphi^\sharp) \subset \mathrm{dom}\,(\mathsf{D}_\varphi^{\mathsf{M}\sharp})$. By (12) and Assertion 4), applied to D_φ and $\mathsf{D}_\varphi^\sharp$,

$$\mathsf{D}_\varphi^\sharp\, a = \mathsf{D}_\varphi^\sharp\, \mathsf{D}_\varphi^{\mathsf{M}}\, \psi = \mathsf{D}_\varphi^\sharp\, \mathsf{D}_\varphi\, \psi = \psi\,.$$

Hence, $\psi \in \mathrm{ran}\,(\mathsf{D}_\varphi^\sharp) = \mathscr{S}_+^{\mathsf{K}}$, by assertion 2), which contradicts $\psi \in \mathscr{S}_+^{-\mathsf{M}} \setminus \mathscr{S}_+^{\mathsf{K}}$. That is, $a \in \mathscr{R}_-^{-\mathsf{M}}$ and $a \notin \mathscr{R}_-^{\mathsf{K}}$, so that $a \in \mathscr{R}_-^{-\mathsf{M}} \setminus \mathscr{R}_-^{\mathsf{K}} \neq \emptyset$. □

Corollary 17 *Given $-\mathsf{K} \in \Sigma_{\mathsf{M}} \cup \{\mathsf{M}\}$, and φ as per (10), the semiconvex transform $\mathsf{D}_\varphi : \mathscr{S}_+^{\mathsf{K}} \to \mathscr{R}_-^{\mathsf{K}}$ of (12) has a well-defined inverse $\mathsf{D}_\varphi^{-1} : \mathscr{R}_-^{\mathsf{K}} \to \mathscr{S}_+^{\mathsf{K}}$ given by*

$$\mathsf{D}_\varphi^{-1}\, a = \mathsf{D}_\varphi^\sharp\, a\,, \qquad\qquad a \in dom\,(\mathsf{D}_\varphi^{-1}) \doteq dom\,(\mathsf{D}_\varphi^\sharp) = \mathscr{R}_-^{\mathsf{K}}\,,$$

where $\mathsf{D}_\varphi^\sharp$ is as per (54).

Proof of Corollary 17 and Theorem 3 Immediate by inspection of Lemma 16. □

Lemma 18 *The asserted forms (16), (17) of the semiconvex transform and its inverse (12), (13) hold.*

Proof Given any $\psi \in \mathscr{S}_+^{-\mathsf{M}}$, Theorem 3 implies that $\mathsf{D}_\varphi\, \psi \in \mathscr{S}_-^{-\mathsf{M}}$. Define $a_\mp : \mathbb{R}^n \to \overline{\mathbb{R}}$ by

$$a_\mp(z) \doteq -[\mathsf{D}_\varphi\, \psi](z) - \varphi(0, z), \tag{65}$$

for all $z \in \mathbb{R}^n$, and observe that $a_\mp$ is convex and lower closed, i.e. $a_\mp = \mathrm{cl}^-\,\mathrm{co}\, a_\mp$, see [10]. Motivated by the form of the argument of the convex hull operation in the right-hand side of (16), define $\widetilde{a}_\mp : \mathbb{R}^n \to \overline{\mathbb{R}}$ by $\widetilde{a}_\mp(z) \doteq -\bigoplus_{i\in\mathbb{N}} \psi_i(0) \otimes \delta_{z_i}^-(z) \otimes (\mathsf{D}_\varphi\, \psi)(z_i)$ for all $z \in \mathbb{R}^n$. Note by definition (8) of $\delta_{z_i}^-$ that

$$\widetilde{a}_\mp(z) = \begin{cases} -[\mathsf{D}_\varphi\, \psi](z_i) - \varphi(0, z_i), & z = z_i\,,\ i \in \mathbb{N}\,, \\ +\infty\,, & \text{otherwise,} \end{cases} \tag{66}$$

for all $z \in \mathbb{R}^n$. As $\{z_i\}_{i\in\mathbb{N}}$ is dense in $\mathbb{R}^n$, inspection of (65) and (66) yields that $a_\mp = \mathrm{cl}^-\,\mathrm{co}\, \widetilde{a}_\mp$. Recalling the definition of $\widetilde{a}_\mp$ subsequently yields (16).

The remaining assertion (17) is an immediate consequence of (13), as $\{z_i\}_{i\in\mathbb{N}}$ is dense in $\mathbb{R}^n$. □

References

1. McEneaney, W.: Max-Plus Methods for Nonlinear Control and Estimation. Systems & Control: Foundations & Application. Birkhauser, Basel (2006)
2. McEneaney, W.: A curse-of-dimensionality-free numerical method for solution of certain HJB PDEs. SIAM J. Control Optim. **46**(4), 1239–1276 (2007)
3. Akian, M., Gaubert, S., Lakhoua, A.: The max-plus finite element method for solving deterministic optimal control problems: basic properties and convergence analysis. SIAM J. Control Optim. **47**(2), 817–848 (2008)
4. Bellman, R.: Dynamic Programming. Princeton University Press, Princeton (1957)
5. McEneaney, W.: A new fundamental solution for differential Riccati equations arising in control. Automatica **44**, 920–936 (2008)
6. Dower, P., McEneaney, W., Zhang, H.: Max-plus fundamental solution semigroups for optimal control problems. In: Proceedings of SIAM Conference on Control Theory and Its Applications (Paris), 2015, pp. 368–375 (2015)
7. Qu, Z.: A max-plus based randomized algorithm for solving a class of HJB PDEs. In: Proceedings of 53rd IEEE Conference on Decision and Control (Los Angeles, CA) (2014)
8. Dower, P.: An approximation arising in max-plus based optimal stopping. In: Proceedings of Australian Control Conference (Sydney), pp. 271–276 (2012)
9. Grune, L., Dower, P.: Hamiltonian based a posteriori error estimation for Hamilton-Jacobi-Bellman equations, Technical report. Universitat Bayreuth. https://epub.uni-bayreuth.de/id/eprint/3578 (2018)
10. Rockafellar, R.: Conjugate Duality and Optimization. SIAM Regional Conference Series in Applied Mathematics, vol. 16. SIAM, Philadelphia (1974)
11. Baccelli, F., Cohen, G., Olsder, G., Quadrat, J.-P.: Synchronization and Linearity. Wiley, New York (1992)
12. Kolokoltsov, V., Maslov, V.: Idempotent Analysis and Applications. Kluwer Publishing House, Dordrecht (1997)

13. Litvinov, G., Maslov, V., Shpiz, G.: Idempotent functional analysis: an algebraic approach. Math. Notes **69**(5), 696–729 (2001)
14. Cohen, G., Gaubert, S., Quadrat, J.-P.: Duality and separation theorems in idempotent semimodules. Linear Algebra Appl. **379**, 395–422 (2004)
15. Fleming, W., McEneaney, W.: A max-plus-based algorithm for a Hamilton-Jacobi-Bellman equation of nonlinear filtering. SIAM J. Control Optim. **38**(3), 683–710 (2000)
16. McEneaney, W., Dower, P.: The principle of least action and fundamental solutions of mass-spring and n-body two-point boundary value problems. SIAM J. Control Optim. **53**(5), 2898–2933 (2015)
17. Zhang, H., Dower, P.: Max-plus fundamental solution semigroups for a class of difference Riccati equations. Automatica **52**, 103–110 (2015)
18. Dower, P., McEneaney, W.: A max-plus dual space fundamental solution for a class of operator differential Riccati equations. SIAM J. Control Optim. **53**(2), 969–1002 (2015)
19. Dower, P., Zhang, H.: A max-plus primal space fundamental solution for a class of differential Riccati equations. Math. Control Signals Syst. **29**(3), 1–33 (2017) [Online]. http://dx.doi.org/10.1007/s00498-017-0200-2
20. Dower, P., McEneaney, W.: Solving two-point boundary value problems for a wave equation via the principle of stationary action and optimal control. SIAM J. Control Optim. **55**(4), 2151–2205 (2017)
21. Dower, P.: Basis adaptation for a max-plus eigenvector method arising in optimal control. In: Proceedings of 23rd International Symposium on Mathematical Theory of Networks and Systems (Hong Kong), pp. 350–355 (2018)
22. Avis, D., Fukuda, K.: A pivoting algorithm for convex hulls and vertex enumeration of arrangements and polyhedra. Discret. Comput. Geom. **8**, 295–313 (1992)
23. Bremner, D., Fukuda, K., Marzetta, A.: Primal-dual methods for vertex and facet enumeration. Discret. Comput. Geom. **20**, 333–357 (1998)
24. de Berg, M., Cheong, O., van Kreveld, M., Overmars, M.: Computational Geometry: Algorithms and Applications, 3rd ed. Springer, Berlin (2008)
25. Rockafellar, R., Wets, R.: Variational Analysis. Springer, Berlin (1997)

Diffusion Process Representations for a Scalar-Field Schrödinger Equation Solution in Rotating Coordinates

William M. McEneaney and Ruobing Zhao

Abstract A particular class of Schrödinger initial value problems is considered, wherein a particle moves in a scalar field centered at the origin, and more specifically, the distribution associated to the solution of the Schrödinger equation has negligible mass in the neighborhood of the origin. The Schrödinger equation is converted to the dequantized form, and a non-inertial frame centered along the trajectory of a classical particle is employed. A solution approximation as a series expansion in a small parameter is obtained through the use of complex-valued diffusion-process representations, where under a smoothness assumption, the expansion converges to the true solution. In the case of an expansion up through only the cubic terms in the space variable, there exist approximate solutions that are periodic with the period of a classical particle, but with an additional secular perturbation. The computations required for solution up to a finite order are purely analytical.

Keywords Stochastic control · Schrödinger equation · Hamilton–Jacobi · Stationary action · Staticization · Complex-valued diffusion

1 Introduction

Diffusion representations have long been a useful tool in solution of second-order Hamilton-Jacobi partial differential equations (HJ PDEs), cf. [7, 10] among many others. The bulk of such results apply to real-valued HJ PDEs, that is, to HJ PDEs where the coefficients and solutions are real-valued. The Schrödinger equation is complex-valued, although generally defined over a real-valued space domain, which presents difficulties for the development of stochastic control representations. In [17, 18], a representation for the solution of a Schrödinger-equation initial value

W. M. McEneaney (✉) · R. Zhao
University of California, San Diego, La Jolla, CA, USA
e-mail: wmceneaney@ucsd.edu; ruz015@eng.ucsd.edu

M. Falcone et al. (eds.), *Numerical Methods for Optimal Control Problems*,
Springer INdAM Series 29, https://doi.org/10.1007/978-3-030-01959-4_11

problem over a scalar field was obtained as a stationary value for a complex-valued diffusion process control problem. Although there is substantial existing work on the relation of stochastic processes to the Schrödinger equation (cf. [9, 14, 20, 21, 26]), the approach considered in [17, 18] is along a slightly different path, closer to [2–5, 13, 16]. However, the representation in [17, 18] employs stationarity of the payoff [19] rather than optimization of the payoff, where stationarity can be used to overcome the limited-duration constraints of methods that use optimization of the payoff.

Here we discuss a particular problem class, and use *diffusion representations as a tool for approximate solution of the Schrödinger equation.* We will consider a specific type of weak field problem. Suppose we have a particle in a scalar field centered at the origin, but in the special case where the particle is sufficiently far from the origin that the distribution associated to the corresponding Schrödinger equation has negligible density near the origin. More specifically, let the particle mass be denoted by m, and let $\hbar$ denote Planck's constant. The simplest scalar-field example, which can be instructive if only purely academic, is the quadratic-field case, generating the quantum harmonic oscillator. Of somewhat more interest is the case where one has the potential energy generated by the field interacting with the particle taking the form $\bar{V}(x) = -\bar{c}/|x|$. Let the solution of the Schrödinger equation at time, t, and position, x, be denoted by $\psi(t,x)$, and consider the associated distribution given by $\tilde{P}(t,x) \doteq [\psi^*\psi](t,x)$. Formally speaking, as $\hbar/m \downarrow 0$, one expects that in some sense $\tilde{P}(t,\cdot)$ approaches a Dirac-delta function centered at $\xi(t)$, where $m\ddot{\xi}(t) = -\nabla_x \bar{V}(\xi(t))$. Consequently, we will consider a non-inertial frame where the origin will be centered at $\xi(t)$ for all t. In particular, we consider a case where $\xi(t)$ follows a circular orbit with constant angular velocity, i.e., $\xi(t) = \hat{\delta}\big(\cos(\omega t), \sin(\omega t)\big)$ where $\hat{\delta} \in (0,\infty)$. (In the interests of space and reduction of clutter, where it will not lead to confusion, we will often write (x_1, x_2) in place of $(x_1, x_2)^T$, etc.) Although such motion can be generated by a two-dimensional harmonic oscillator, we will focus mainly on the $\bar{V}(x) = -\bar{c}/|x|$ class, in which case $\omega \doteq [\bar{c}/(m\hat{\delta}^3)]^{1/2}$. We suppose that $\hat{\delta}$ is sufficiently large such that $\tilde{P}(t,x) \ll 1$ for $|x| < \hat{\delta}/2$, and thus that one may approximate $\bar{V}$ in the vicinity of $\xi(t)$ by a finite number of terms in a power series expansion centered at $\xi(t)$. We will use a set of complex-valued diffusion representations to obtain an approximation to the resulting Schrödinger equation solution. If the solution is holomorphic in x and a small parameter, then the approximate solution converges as the number of terms in the set of diffusion representations approaches infinity.

The analysis will be carried out only in the case of a holomorphic field approximation. As our motivation is the case where $\hat{\delta}$ is large relative to the associated position distribution, one expects that the case of a $-\bar{c}/|x|$ potential may be sufficiently well-modeled by a finite number of terms in a power series expansion. However, an analysis of the errors induced by such an approximation to a $-\bar{c}/|x|$ potential is beyond the scope of this already long paper, and may be addressed in a later effort; the focus here is restricted to the diffusion-representation based method of solution approximation method *given* such an approximation to the potential. We remark that in the case of a quadratic potential, we recover the quantum harmonic

oscillator solution. Also, in the case of $\bar{V}(x) = -\bar{c}/|x|$, as $\hat{\delta} \to \infty$, the solution approaches that of the free particle case. The computations required for solution up to any finite polynomial-in-space order may be performed analytically.

In Sect. 2, we review the Schrödinger initial value problem, and the dequantized form of the problem. The solution to the dequantized form of the problem will be approximated through the use of diffusion representations; the solution to the originating Schrödinger initial value problem is recovered by a simple transformation. As it is used in Sect. 2, we briefly recall the stat operator in Sect. 3.1. In Sect. 3.2, the dequantized form will be converted into a form over a rotating and translating reference frame centered at the position of a classical particle following a circular trajectory generated by the central field. Then, in Sect. 3.3, we discuss equivalent forms over a complex space domain, and over a double-dimension real-valued domain. Classical existence, uniqueness and smoothness results will be applied to the problem in this last form. These will then be transferred to the original form as a complex-valued solution over a real space domain. In Sect. 4, we indicate the expansion of the solution in a small parameter related to the inverse of the distance to the origin of the field. A power series representation will be used, where this will be over both space and the small parameter. In particular, we will assume that at each time, the solution will be holomorphic over space and the small parameter. The functions in the expansion are solutions to corresponding HJ PDEs, where these are also indicated here. The HJ PDE for the first term, say $k = 0$, has a closed-form solution, and this is given in Sect. 5. Then, in Sect. 6, it is shown that for $k \geq 1$, given the solutions to the preceding terms, the HJ PDE for the kth term takes a linear parabolic form, with a corresponding diffusion representation. It is shown that diffusion representation may be used to obtain the solution of the $k+1$ HJ PDE given the solutions to the k-and-lower HJ PDE solutions. The required computations may be performed analytically. In Sect. 7, this method is applied to obtain the next term in the expansion in the case of a cubic approximation of the classic $1/r$ type of potential, and additional terms may be obtained similarly.

2 Dequantization

We recall the Schrödinger initial value problem, given as

$$0 = i\hbar\psi_t(s,x) + \tfrac{\hbar^2}{2m}\Delta_x\psi(s,x) - \psi(s,x)\bar{V}(x), \quad (s,x) \in \mathscr{D}, \tag{1}$$

$$\psi(0,x) = \psi_0(x), \quad x \in \mathbb{R}^n, \tag{2}$$

where initial condition ψ_0 takes values in $\mathbb{C}$, Δ_x denotes the Laplacian with respect to the space (second) variable, $\mathscr{D} \doteq (0,t) \times \mathbb{R}^n$, and subscript t will denote the derivative with respect to the time variable (the first argument of ψ here) regardless of the symbol being used for time in the argument list. We also let $\overline{\mathscr{D}} \doteq [0,t) \times \mathbb{R}^n$. We consider the Maslov dequantization of the solution of the Schrödinger equation (cf. [15]), *which similar to a standard log transform,* is

$S : \overline{\mathscr{D}} \to \mathbb{C}$ given by $\psi(s,x) = \exp\{\frac{i}{\hbar}S(s,x)\}$. Note that $\psi_t = \frac{i}{\hbar}\psi S_t$, $\psi_x = \frac{i}{\hbar}\psi S_x$ and $\Delta_x \psi = \frac{i}{\hbar}\psi \Delta_x S - \frac{1}{\hbar^2}\psi |S_x|_c^2$ where for $y \in \mathbb{C}^n$, $|y|_c^2 \doteq \sum_{j=1}^n y_j^2$. (We remark that notation $|\cdot|_c^2$ is not intended to indicate a squared norm; the range is complex.) We find that (1)–(2) become

$$0 = -S_t(s,x) + \tfrac{i\hbar}{2m}\Delta_x S(s,x) + H^0(x, S_x(s,x)), \quad (s,x) \in \mathscr{D}, \tag{3}$$

$$S(0,x) = \bar{\phi}(x), \quad x \in \mathbb{R}^n, \tag{4}$$

where $H : \mathbb{R}^n \times \mathbb{C}^n \to \mathbb{C}$ is the Hamiltonian given by

$$H^0(x,p) = -\big[\tfrac{1}{2m}|p|_c^2 + \bar{V}(x)\big] = \operatorname*{stat}_{v \in \mathbb{C}^n} \big\{ v \cdot p + \tfrac{m}{2}|v|_c^2 - \bar{V}(x) \big\}, \tag{5}$$

and stat is defined in Sect. 3.1. We look for solutions in the space

$$\mathscr{S} \doteq \{S : \overline{\mathscr{D}} \to \mathbb{C} \mid S \in C_p^{1,2}(\mathscr{D}) \cap C(\overline{\mathscr{D}})\}, \tag{6}$$

where $C_p^{1,2}$ denotes the space of functions which are continuously differentiable once in time and twice in space, and which satisfy a polynomial-growth bound.

3 Preliminaries

In this section, we collect condensed discussions of relevant classical material as well as some recently obtained results and definitions.

3.1 Stationarity Definitions

Recall that classical conservative systems obey the stationary action principle, where the path taken by the system is that which is a stationary point of the action functional. For this and other reasons, as in the definition of the Hamiltonian given in (5), we find it useful to develop additional notation and nomenclature. Specifically, we will refer to the search for stationary points more succinctly as *staticization*, and we make the following definitions. Suppose $(\mathscr{Y}, |\cdot|)$ is a generic normed vector space over $\mathbb{C}$ with $\mathscr{G} \subseteq \mathscr{Y}$, and suppose $F : \mathscr{G} \to \mathbb{C}$. We say $\bar{y} \in \operatorname{argstat}\{F(y) \mid y \in \mathscr{G}\}$ if $\bar{y} \in \mathscr{G}$ and either $\limsup_{y \to \bar{y}, y \in \mathscr{G}\setminus\{\bar{y}\}} |F(y) - F(\bar{y})|/|y - \bar{y}| = 0$, or there exists $\delta > 0$ such that $\mathscr{G} \cap B_\delta(\bar{y}) = \{\bar{y}\}$ (where $B_\delta(\bar{y})$ denotes the ball of radius δ around $\bar{y}$). If $\operatorname{argstat}\{F(y) \mid y \in \mathscr{G}\} \neq \emptyset$, we define the possibly set-valued stat^s operator by

$$\operatorname*{stat}^s_{y \in \mathscr{G}} F(y) \doteq \operatorname{stat}^s\{F(y) \mid y \in \mathscr{G}\} \doteq \big\{ F(\bar{y}) \,\big|\, \bar{y} \in \operatorname{argstat}\{F(y) \mid y \in \mathscr{G}\} \big\}.$$

If $\text{argstat}\{F(y) \mid y \in \mathscr{G}\} = \emptyset$, $\text{stat}^s_{y\in\mathscr{G}} F(y)$ is undefined. We will also be interested in a single-valued stat operation. In particular, if there exists $a \in \mathbb{C}$ such that $\text{stat}^s_{y\in\mathscr{G}} F(y) = \{a\}$, then $\text{stat}_{y\in\mathscr{G}} F(y) \doteq a$; otherwise, $\text{stat}_{y\in\mathscr{G}} F(y)$ is undefined. At times, we may abuse notation by writing $\bar{y} = \text{argstat}\{F(y) \mid y \in \mathscr{G}\}$ in the event that the argstat is the set $\{\bar{y}\}$. For further discussion, we refer the reader to [19]. The following is immediate from the above definitions.

Lemma 1 *Suppose $\mathscr{Y}$ is a Hilbert space, with open set $\mathscr{G} \subseteq \mathscr{Y}$, and that $F : \mathscr{G} \to \mathbb{C}$ is Fréchet differentiable at $\bar{y} \in \mathscr{G}$ with Riesz representation $F_y(\bar{y}) \in \mathscr{Y}$ Then, $\bar{y} \in \text{argstat}\{F(y) \mid y \in \mathscr{G}\}$ if and only if $F_y(\bar{y}) = 0$.*

3.2 The Non-inertial Frame

As noted in the introduction, we suppose a central scalar field such that a particular solution for the motion of a classical particle in the field takes the form $\xi(t) = \hat{\delta}\big(\cos(\omega t), \sin(\omega t)\big)$ where $\hat{\delta}, \omega \in (0, \infty)$. In particular, we concentrate on the potential $\bar{V}(x) = -\bar{c}/|x|$, in which case $\omega \doteq [\bar{c}/(m\hat{\delta}^3)]^{1/2}$. We consider a two-dimensional space model and a non-inertial frame centered at $\xi(t)$ for all $t \in (0, \infty)$, with the first basis axis in the positive radial direction and the second basis vector in the direction of the velocity of the particle. Let positions in the non-inertial frame be denoted by $z \in \mathbb{R}^2$, where the transformation between frames at time $t \in \mathbb{R}$ is given by

$$z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = G_{\omega t} x - \begin{pmatrix} \hat{\delta} \\ 0 \end{pmatrix} \doteq \begin{pmatrix} \cos(\omega t) & \sin(\omega t) \\ -\sin(\omega t) & \cos(\omega t) \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \begin{pmatrix} \hat{\delta} \\ 0 \end{pmatrix}. \tag{7}$$

We will denote this transformation as $z = z^*(x)$, with its inverse denoted similarly as $x = x^*(z)$, where $x^*(z) = (G_{\omega t})^T (z + (\hat{\delta}, 0)^T)$.

For $z \in \mathbb{R}^2$, define $V(z) \doteq \bar{V}(x^*(z))$ and $\phi(z) \doteq \bar{\phi}(x^*(z))$. Then, $\tilde{S}^f : \overline{\mathscr{D}} \to \mathbb{C}$ defined by $\tilde{S}^f(s, z) \doteq \hat{\tilde{S}}^f(s, x^*(z))$ is a solution of the forward-time dequantized HJ PDE problem given by

$$0 = -S_t(s, z) + \frac{i\hbar}{2m} \Delta_z S(s, z) - (A_0 z + b_0)^T S_z(s, z) - \frac{1}{2m} |S_z(s, z)|_c^2 - V(z), \quad (s, z) \in \mathscr{D}, \tag{8}$$

$$S(0, z) = \phi(z), \quad z \in \mathbb{R}^2, \quad \text{where} \quad A_0 \doteq \omega \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \text{ and } b_0 \doteq -\omega\hat{\delta} \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \tag{9}$$

if and only if $\hat{\tilde{S}}^f$ is a solution of (3)–(4). (We remark that one may see [24] for further discussion of non-inertial frames in the context of the Schrödinger equation.) In order to apply the diffusion representations as an aid in solution, we will find it

helpful to reverse the time variable, and hence we look instead, and equivalently, at the Hamilton-Jacobi partial differential equation (HJ PDE) problem given by

$$0 = S_t(s,z) + \tfrac{i\hbar}{2m}\Delta_z S(s,x) - (A_0 z + b_0)^T S_z(s,z) - \tfrac{1}{2m}|S_z(s,z)|_c^2 - V(z), \quad (s,z) \in \mathscr{D}, \tag{10}$$

$$S(t,z) = \phi(z), \quad z \in \mathbb{R}^n. \tag{11}$$

In this last form, we will fix $t \in (0,\infty)$, and allow s to vary in $(0,t]$.

3.3 Extensions to the Complex Domain

Various details of extensions to the complex domain must be considered prior to the development of the representation. This material is rather standard, but it is required for the main development. Models (1)–(2), (3)–(4) and (10)–(11) are typically given as HJ PDE problems over real space domains. However, as in Doss et al. [1–3], we will find it convenient to change the domain to one where the space components lie over the complex field. We also extend the domain of the potential to $\mathbb{C}^2$, i.e., $V : \mathbb{C}^2 \to \mathbb{C}$, and we will abuse notation by employing the same symbol for the extended-domain functions. Throughout, for $k \in \mathbb{N}$, and $z \in \mathbb{C}^k$ or $z \in \mathbb{R}^k$, we let $|z|$ denote the Euclidean norm. Let $\mathscr{D}_{\mathbb{C}} \doteq (0,t) \times \mathbb{C}^2$ and $\overline{\mathscr{D}}_{\mathbb{C}} = (0,t] \times \mathbb{C}^2$, and define

$$\mathscr{S}_{\mathbb{C}} \doteq \{S : \overline{\mathscr{D}}_{\mathbb{C}} \to \mathbb{C} |\ S \text{ is continuous on } \overline{\mathscr{D}}_{\mathbb{C}}, \text{continuously differentiable in time on } \mathscr{D}_{\mathbb{C}}, \text{and holomorphic on } \mathbb{C}^2 \text{ for all } r \in (0,t]\}, \tag{12}$$

$$\mathscr{S}_{\mathbb{C}}^{p} \doteq \{S \in \mathscr{S}_{\mathbb{C}} |\ S \text{ satisfies a polynomial growth condition in space, uniformly on } (0,t]\}. \tag{13}$$

The extended-domain form of problem (10)–(11) is

$$0 = S_t(s,z) + \tfrac{i\hbar}{2m}\Delta_z S(s,z) - (A_0 z + b_0)^T S_z(s,z) - \tfrac{1}{2m}|S_z(s,z)|_c^2 - V(z), \quad (s,z) \in \mathscr{D}_{\mathbb{C}}, \tag{14}$$

$$S(t,z) = \phi(z), \quad z \in \mathbb{C}^2. \tag{15}$$

Remark 1 We remark that a holomorphic function on $\mathbb{C}^2$ is uniquely defined by its values on the real part of its domain. In particular, $\tilde{S} : \overline{\mathscr{D}} \to \mathbb{C}$ uniquely defines its extension to a time-indexed holomorphic function over complex space, say $\bar{S} : \overline{\mathscr{D}}_{\mathbb{C}} \to \mathbb{C}$, if the latter exists. Consequently, although (10)–(11) form an HJ PDE problem for a complex-valued solution over real time and real space, (14)–(15) is an equivalent formulation, under the assumptions that a holomorphic solution exists and one has uniqueness for both.

Throughout the remainder, we will assume the following.

$$V, \phi : \mathbb{C}^2 \to \mathbb{C} \text{ are holomorphic on } \mathbb{C}^2. \tag{A.1}$$

Remark 2 The assumption on V requires a remark. Recall that we are interested here in a class of problems where $\hat{\delta}$ is large in the sense that the distribution associated to the solution of the Schrödinger initial value problem has only very small probability mass outside a ball of radius less than $\hat{\delta}$. If $\bar{V}$ is of the $\bar{c}/|x|$ form, one would use only a finite number of terms in the power series expansion around $z = 0$. The focus here is on a diffusion-representation based method for approximate solution of the Schrödinger initial value problem given a holomorphic potential. The errors introduced by the use of a truncated power series for a $\bar{c}/|x|$-type potential for large $\hat{\delta}$ are outside the scope of the discussion.

We will refer to a linear space over the complex [real] field as a complex [real] space. Although (14)–(15) form an HJ PDE problem for a complex-valued solution over real time and complex space, there is an equivalent formulation as a real-valued solution over real time and a double-dimension real space. We will find such formulations to be helpful in the analysis to follow. Further, although it is natural to work with complex-valued state processes in this problem domain, in order to easily apply many of the existing results regarding existence, uniqueness and moments, we will also find it handy to use a "vectorized" real-valued representation for the complex-valued state processes. We begin from the standard mapping of the complex plane into $\mathbb{R}^2$, denoted here by $\mathscr{V}_{00} : \mathbb{C} \to \mathbb{R}^2$, with $\mathscr{V}_{00}(z) \doteq (x, y)^T$, where $x = \mathbf{Re}(z)$ and $y = \mathbf{Im}(z)$. This immediately yields the mapping $\mathscr{V}_0 : \mathbb{C}^n \to \mathbb{R}^{2n}$ given by $\mathscr{V}_0(x + iy) \doteq (x^T, y^T)^T$, where component-wise, $(x_j, y_j)^T = \mathscr{V}_{00}(x_j + iy_j)$ for all $j \in]1, n[$, where throughout, for integer $a \leq b$, we let $]a, b[\doteq \{a, a+1, \ldots b\}$. Also, as remarked above, in the interests of a reduction of cumbersome notation, we will frequently abuse notation by writing (x, y) in place of $(x^T, y^T)^T$ when the meaning is clear. Lastly, we may decompose any function in $\mathscr{S}_{\mathbb{C}}$, say $F \in \mathscr{S}_{\mathbb{C}}$, as

$$(\bar{R}(r, \mathscr{V}_0(z)), \bar{T}(r, \mathscr{V}_0(z)))^T \doteq \mathscr{V}_{00}(F(r, z)), \tag{16}$$

where $\bar{R}, \bar{T} : \overline{\mathscr{D}}_2 \doteq (0, t] \times \mathbb{R}^{2n} \to \mathbb{R}$, and we also let $\mathscr{D}_2 \doteq (0, t) \times \mathbb{R}^{2n}$. For later reference, it will be helpful to recall some standard relations between derivative components, which are induced by the Cauchy-Riemann equations. For all $(r, z) = (r, x + iy) \in (0, t) \times \mathbb{C}^2$ and all $j, k, \ell \in]1, n[$, and suppressing the arguments for reasons of space we have

$$\mathbf{Re}[F_{z_j, z_k}] = \bar{R}_{x_j, x_k} = -\bar{R}_{y_j, y_k} = \bar{T}_{y_j, x_k} = \bar{T}_{x_j, y_k}, \tag{17}$$

$$\mathbf{Im}[F_{z_j, z_k}] = -\bar{R}_{x_j, y_k} = -\bar{R}_{y_j, x_k} = -\bar{T}_{y_j, y_k} = \bar{T}_{x_j, x_k}. \tag{18}$$

4 An Expansion

We now reduce our problem class to the two-dimensional space case (i.e., $n = 2$). We will expand the desired solutions of our problems, and use these expansions as a means for approximation of the solution. First, we consider holomorphic V in the form of a finite or infinite power series. In the simple example case where $\bar{V}$ generates the quantum harmonic oscillator, one may take $\bar{V}(x) = \frac{\bar{c}^q}{2}[x_1^2 + x_2^2]$, in which case

$$V(z) = \frac{\bar{c}^q}{2}\hat{\delta}^2 + \bar{c}^q\hat{\delta}z_1 + \frac{\bar{c}^q}{2}[z_1^2 + z_2^2].$$

The scalar field of most interest takes the form $-\bar{V}(x) = \bar{c}/|x|$, yielding $-V(z) = \bar{c}/|z + (\hat{\delta}, 0)|$. In this case, recalling that this effort focuses on the case where $\hat{\delta}$ is large relative to the radius of the "non-negligible" portion of the probability distribution associated to the solution, we consider only a truncated power series, and let $\breve{V}^K(z)$ denote the partial sum containing only terms up to order $K+2 < \infty$ in z. We will be interested in the dependence of the potential and the resulting solutions in the parameter $\hat{\epsilon} \doteq 1/\hat{\delta}$. We also recall from Sect. 3.2 that $\omega \doteq [\bar{c}/(m\hat{\delta}^3)]^{1/2}$, or $\bar{c} = m\omega^2\hat{\delta}^3$. We explicitly indicate the expansion up to the fourth-order term in z and the form of higher-order terms. One finds,

$$-\breve{V}^2(z) = -\sum_{k=0}^{2}\hat{\epsilon}^k\hat{V}^k(z), \tag{19}$$

$$-\hat{V}^0(z) = m\omega^2\big[\hat{\delta}^2 - \hat{\delta}z_1 + (z_1^2 - z_2^2/2)\big],$$

$$-\hat{V}^1(z) = m\omega^2\big[-z_1^3 + 3z_1z_2^2/2\big], \quad -\hat{V}^2(z) = m\omega^2\big[z_1^4 - 3z_1^2z_2^2 + 3z_2^4/8\big],$$

and more generally, for $k > 1$, $-\hat{V}^k(z) = m\omega^2\Big[\sum_{j=0}^{k+2} c^V_{k+2,j}z_1^j z_2^{k-j}\Big]$, for proper choice of coefficients $c^V_{k,j}$.

Here, we find it helpful to explicitly consider the dependence of $\tilde{S}$ and $\bar{S}$ (solutions of (10)–(11) and (14)–(15), respectively) on $\hat{\epsilon}$, where for convenience of exposition, we also allow $\hat{\epsilon}$ to take complex values. Abusing notation, we let $\tilde{S} : \overline{\mathscr{D}} \times \mathbb{C} \to \mathbb{C}$ and $\bar{S} : \overline{\mathscr{D}}_{\mathbb{C}} \times \mathbb{C} \to \mathbb{C}$, and denote the dependence on their arguments as $\tilde{S}(s, z, \hat{\epsilon})$ and $\bar{S}(s, z, \hat{\epsilon})$. We let $\breve{\mathscr{D}} \doteq \mathscr{D} \times \mathbb{C}$, $\breve{\overline{\mathscr{D}}} \doteq \overline{\mathscr{D}} \times \mathbb{C}$, $\breve{\mathscr{D}}_{\mathbb{C}} \doteq \mathscr{D}_{\mathbb{C}} \times \mathbb{C}$ and $\breve{\overline{\mathscr{D}}}_{\mathbb{C}} \doteq \overline{\mathscr{D}}_{\mathbb{C}} \times \mathbb{C}$, where we recall that the physical-space components are now restricted to the two-dimensional case. We also let

$$\breve{\mathscr{S}}_{\mathbb{C}} \doteq \{S : \breve{\overline{\mathscr{D}}}_{\mathbb{C}} \to \mathbb{C} |\ S \text{ is continuous on } \breve{\overline{\mathscr{D}}}_{\mathbb{C}}, \text{ continuously differentiable in time on } \breve{\mathscr{D}}_{\mathbb{C}}, \text{ and } S(r, \cdot, \cdot) \text{ is holomorphic on } \mathbb{C}^2 \times \mathbb{C} \text{ for all } r \in (0, t]\}, \tag{20}$$

$$\breve{\mathscr{S}}^{p}_{\mathbb{C}} \doteq \{S \in \breve{\mathscr{S}}_{\mathbb{C}} \mid S \text{ satisfies a polynomial growth condition in space, uniformly on } (0,t]\}. \tag{21}$$

We will make the following assumption throughout the sequel.

$$\text{There exists a unique solution, } \bar{S} \in \breve{\mathscr{S}}_{\mathbb{C}}, \text{ to (14)–(15).} \tag{A.2}$$

We also let the power series expansion for ϕ be arranged as

$$\phi(z) = \sum_{k=0}^{\infty} \hat{\epsilon}^k \phi^k(z) \doteq \phi^0(z) + \sum_{k=1}^{\infty} \hat{\epsilon}^k \sum_{l=0}^{\infty} \sum_{j=0}^{\infty} b^{\phi}_{k+2,l,j} z_1^j z_2^{l-j}, \tag{22}$$

where $\phi^0(z)$ is quadratic in z. We consider the following terminal value problems. The zeroth-order problem is

$$0 = S^0_t + \tfrac{i\hbar}{2m} \Delta_z S^0 - \left(A_0 z + b_0\right)^T S^0_z - \tfrac{1}{2m} |S^0_z|^2_c - \hat{V}^0, \qquad (s,z) \in \mathscr{D}_{\mathbb{C}}, \tag{23}$$

$$S^0(t,z) = \phi^0(z), \qquad z \in \mathbb{C}^2. \tag{24}$$

For $k \geq 1$, the kth terminal value problem is

$$0 = S^k_t + \tfrac{i\hbar}{2m} \Delta_z S^k - \left(A_0 z + b_0 + \tfrac{1}{m} S^0_z\right)^T S^k_z - \tfrac{1}{2m} \sum_{\kappa=1}^{k-1} \left(S^{\kappa}_z\right)^T S^{k-\kappa}_z - \hat{V}^k, \quad (s,z) \in \mathscr{D}_{\mathbb{C}}, \tag{25}$$

$$S^k(t,z) = \phi^k(z), \qquad z \in \mathbb{C}^2. \tag{26}$$

Note that for $k \geq 1$, given the $\hat{S}^{\kappa}$ for $\kappa < k$, (25) is a linear, parabolic, second-order PDE, while zeroth-order case (23) is a nonlinear, parabolic, second-order PDE. Also note that (23) is (25) in the case of $k = 0$, but as its form is different, it is worth breaking it out separately. It is also worth noting here that if the S^k are all polynomial in z of order up to k, then the right-hand side of (25) is polynomial in z of order up to k, as is the right-hand side of (26).

Theorem 1 *Assume there exists a unique solution, $\hat{S}^0$, in $\breve{\mathscr{S}}_{\mathbb{C}}$ to (23)–(24), and that for each $k \geq 1$, there exists a unique solution, $\hat{S}^k$, in $\breve{\mathscr{S}}_{\mathbb{C}}$ to (25)–(26). Then, $\bar{S} = \sum_{k=0}^{\infty} \hat{\epsilon}^k \hat{S}^k$.*

Remark 3 It is worth noting here that if the S^k are all polynomial in z of order up to $k+2$, then for each k, the right-hand side of (25) is polynomial in z of order up to $k+2$, as is the right-hand side of (26). That is, with the expansion in powers of $\hat{\epsilon} = \hat{\delta}^{-1}$, the resulting constituent HJ PDE problems indexed by k are such that one might hope for polynomial-in-z solutions of order $k+2$, and this hope will be realized further below.

Proof Let $\bar{\mathbb{N}} \doteq \mathbb{N} \cup \{0\}$. By Assumption (A.2), $\bar{S}$ has a unique power series expansion on $\bar{\bar{\mathscr{D}}}_{\mathbb{C}}$, which we denote by

$$\bar{S}(s,z,\hat{\epsilon}) = \sum_{k=0}^{\infty} \hat{\epsilon}^k \tilde{c}^k(s,z) \doteq \sum_{k=0}^{\infty} \hat{\epsilon}^k \sum_{l=0}^{\infty} \sum_{j=0}^{\infty} \tilde{\tilde{c}}_{k,l,j}(s) z_1^j z_2^{l-j},$$

where the $\tilde{\tilde{c}}_{k,l,j}(\cdot) : (0,t] \to \mathbb{C}$ form a time-indexed set of coefficients, and obviously, the $\tilde{c}^k(\cdot,\cdot) : \overline{\mathscr{D}}_{\mathbb{C}} \to \mathbb{C}$ are given by $\tilde{c}^k(s,z) = \sum_{l=0}^{\infty} \sum_{j=0}^{\infty} \tilde{\tilde{c}}_{k,l,j}(s) z_1^j z_2^{l-j}$ for all $k \in \bar{\mathbb{N}}$. For all $k \in \bar{\mathbb{N}}$, define the notation $\tilde{c}^{-k}(\cdot,\cdot) \doteq \sum_{j=k+1}^{\infty} \hat{\epsilon}^{j-(k+1)} \tilde{c}^j(\cdot,\cdot)$. Also define $V^{-k} \doteq \hat{\epsilon}^{-(k+1)}\big[V - \sum_{j=0}^{k} \hat{\epsilon}^j \hat{V}^j\big]$ and $\phi^{-k} \doteq \sum_{j=k+1}^{\infty} \hat{\epsilon}^{j-(k+1)} \phi^j = \hat{\epsilon}^{-(k+1)}\big[\phi - \sum_{j=0}^{k} \hat{\epsilon}^j \phi^j\big]$ for all $k \in \bar{\mathbb{N}}$. Recall that $\bar{S}$ is the unique solution in $\breve{\mathscr{S}}_{\mathbb{C}}$ of (14)–(15). By (15),

$$\tilde{c}^k(t,z) = \phi^k(z) \quad \text{and} \quad \tilde{c}^{-k}(t,z) = \phi^{-k}(z) \quad \forall\, z \in \mathbb{C}^2. \tag{27}$$

Separating the $\tilde{c}^0$ and $\tilde{c}^{-0}$ components of $\bar{S}$ in (14) yields

$$\begin{aligned} 0 =& \tilde{c}_t^0 + \tfrac{i\hbar}{2m} \Delta_z \tilde{c}^0 - (A_0 z + b_0)^T \tilde{c}_z^0 - \tfrac{1}{2m} |\tilde{c}_z^0|_c^2 - \hat{V}^0 \\ &+ \hat{\epsilon} \Big\{ \tilde{c}_t^{-0} + \tfrac{i\hbar}{2m} \Delta_z \tilde{c}^{-0} - (A_0 z + b_0 + \tfrac{1}{m} \tilde{c}_z^0)^T \tilde{c}_z^{-0} - \tfrac{\hat{\epsilon}}{2m} |\tilde{c}_z^{-0}|_c^2 - V^{-0} \Big\}. \end{aligned} \tag{28}$$

Now, note that as $\bar{S}(s,\cdot,\cdot)$ is holomorphic for all $s \in (0,t]$, we have $\bar{S}_z(s,\cdot,\cdot)$ and $\Delta_z \bar{S}(s,\cdot,\cdot)$ holomorphic for all $s \in (0,t]$. Further, by standard results on the composition of holomorphic mappings, noting that $g : \mathbb{C}^2 \to \mathbb{C}$ given by $g(z) \doteq |z|_c^2 = z^T z$ is holomorphic, we see that $|\tilde{S}_z(s,\cdot,\cdot)|_c^2 = g\big(\tilde{S}_z(s,\cdot,\cdot)\big)$ is holomorphic for all $s \in (0,t]$. Combining these insights, we see that with $S = \bar{S}$ all terms on the right-hand side of (14), with the exception of S_t are holomorphic in $(z,\hat{\epsilon})$, which implies that $\bar{S}_t(s,\cdot,\cdot)$ is holomorphic for all $s \in (0,t]$. Consequently, for any $s \in (0,t]$, the right-hand side of (14) with $S = \bar{S}$ has a unique power series expansion. This implies that, as (28) is satisfied for all $\hat{\epsilon} \in \mathbb{C}$, we must have

$$0 = \tilde{c}_t^0 + \tfrac{i\hbar}{2m} \Delta_z \tilde{c}^0 - (A_0 z + b_0)^T \tilde{c}_z^0 - \tfrac{1}{2m} |\tilde{c}_z^0|_c^2 - \hat{V}^0, \tag{29}$$

$$0 = \tilde{c}_t^{-0} + \tfrac{i\hbar}{2m} \Delta_z \tilde{c}^{-0} - (A_0 z + b_0 + \tfrac{1}{m} \tilde{c}_z^0)^T \tilde{c}_z^{-0} - \tfrac{\hat{\epsilon}}{2m} |\tilde{c}_z^{-0}|_c^2 - V^{-0}. \tag{30}$$

By (27), (29) and the assumptions, $\tilde{c}^0 = \hat{S}^0$.

Next, separating the $\tilde{c}^1$ and $\tilde{c}^{-1}$ components, (30) implies

$$\begin{aligned} 0 =& \tilde{c}_t^1 + \tfrac{i\hbar}{2m} \Delta_z \tilde{c}^1 - (A_0 z + b_0 + \tfrac{1}{m} \tilde{c}_z^0)^T \tilde{c}_z^1 - \hat{V}^1 \\ &+ \hat{\epsilon} \Big\{ \tilde{c}_t^{-1} + \tfrac{i\hbar}{2m} \Delta_z \tilde{c}^{-1} - (A_0 z + b_0 + \tfrac{1}{m} \tilde{c}_z^0)^T \tilde{c}_z^{-1} - V^{-1} - \tfrac{\hat{\epsilon}}{2m} |\tilde{c}_z^{-1}|_c^2 \Big\}. \end{aligned} \tag{31}$$

Similar to the $k = 0$ case, as (31) is satisfied for all $\hat{\epsilon} \in \mathbb{C}$, we have

$$0 = \tilde{c}_t^1 + \tfrac{i\hbar}{2m}\Delta_z \tilde{c}^1 - (A_0 z + b_0 + \tfrac{1}{m}\tilde{c}_z^0)^T \tilde{c}_z^1 - \hat{V}^1, \tag{32}$$

$$0 = \tilde{c}_t^{-1} + \tfrac{i\hbar}{2m}\Delta_z \tilde{c}^{-1} - (A_0 z + b_0 + \tfrac{1}{m}\tilde{c}_z^0)^T \tilde{c}_z^{-1} - V^{-1} - \tfrac{1}{2m}\sum_{\kappa=1}^{0} (\tilde{c}_z^\kappa)^T \tilde{c}_z^{k-\kappa} - \tfrac{\hat{\epsilon}}{2m}|\tilde{c}_z^{-1}|_c^2, \tag{33}$$

where the zero-valued penultimate term on the right-hand side of (33) is included because analogous terms will appear with non-zero value in higher-order expansion equations. By (27), (32) and the assumptions, $\tilde{c}^1 = \hat{S}^1$. Proceeding inductively, one finds $\tilde{c}^k = \hat{S}^k$ for all $k \in \bar{\mathbb{N}}$, which yields the assertion. □

4.1 An Alternate Assumption

It may be worth noting the following reformulation and assumption. Let $\tilde{g}_{\hat{\delta}} : \mathbb{C}^2 \to \mathbb{C}^2$ and $\hat{g}_{\hat{\delta}} : \mathbb{R} \to \mathbb{R}$ be given by $\tilde{g}_{\hat{\delta}}(z) \doteq (1/\hat{\delta})z$ and $\hat{g}_{\hat{\delta}}(s) \doteq s/\hat{\delta}^2$. Let $\tilde{s} \doteq \hat{g}_{\hat{\delta}}(s) = s/\hat{\delta}^2$ and $\tilde{z} \doteq \tilde{g}_{\hat{\delta}}(z) = (1/\hat{\delta})z$. Note that under this change of variables, the angular rate becomes $\hat{\omega} = \frac{d\theta}{d\tilde{s}} = \frac{d\theta}{ds}\frac{ds}{d\tilde{s}} = \hat{\delta}^2\omega$, and where the units of $\hbar$ are such that the resulting scaling is the identity. Let $\tilde{\bar{S}}(\tilde{s}, \tilde{z}) \doteq \bar{S}(\hat{g}_{\hat{\delta}}^{-1}(\tilde{s}), \tilde{g}_{\hat{\delta}}^{-1}(\tilde{z})) = \bar{S}(\hat{g}_{\hat{\delta}}^{-1}(\tilde{s}), \tilde{g}_{\hat{\delta}}^{-1}(\tilde{z}), \hat{\epsilon})$ for all $(s, z) \in \overline{\mathscr{D}}$, where we recall the abuse of notation regarding explicit inclusion of the third argument in $\bar{S}$. Note that $\tilde{\bar{S}}_{\tilde{s}}(\tilde{s}, \tilde{z}) = \bar{S}_s(\hat{g}_{\hat{\delta}}^{-1}(\tilde{s}), \tilde{g}_{\hat{\delta}}^{-1}(\tilde{z}))\frac{\hat{g}_{\hat{\delta}}^{-1}(\tilde{s})}{d\tilde{s}} = \hat{\delta}^2 \bar{S}_s(s, z)$, with similar expressions for the space derivatives. The HJ PDE problem for $\tilde{\bar{S}}$, corresponding to (14)–(15) for $\bar{S}$, is

$$0 = S_{\tilde{s}}(\tilde{s}, \tilde{z}) + \tfrac{i\hbar}{2m}\Delta_{\tilde{z}} S(\tilde{s}, \tilde{z}) - \hat{\omega}(\bar{A}_0 \tilde{z} + \bar{b}_0)^T S_{\tilde{z}}(\tilde{s}, \tilde{z}) - \tfrac{1}{2m}|S_{\tilde{z}}(\tilde{s}, \tilde{z})|_c^2 - \tilde{V}(\tilde{z}), \qquad (\tilde{s}, \tilde{z}) \in (0, \tilde{t}) \times \mathbb{C}^2, \tag{34}$$

$$S(\tilde{t}, \tilde{z}) = \tilde{\phi}(\tilde{z}), \quad \tilde{z} \in \mathbb{C}^2, \qquad \bar{A}_0 \doteq \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \qquad \bar{b}_0 \doteq -\begin{pmatrix} 0 \\ 1 \end{pmatrix}, \tag{35}$$

$\tilde{t} = t/\hat{\delta}^2$, $\tilde{\phi}(\tilde{z}) \doteq \phi(\tilde{g}_{\hat{\delta}}^{-1}(\tilde{z}))$ and $\frac{1}{\hat{\delta}^2}\tilde{V}(\tilde{z}) \doteq V(\tilde{g}_{\hat{\delta}}^{-1}(\tilde{z})) = V(z)$.

Note that in the case of a truncated expansion of a potential of form $-\bar{V}(x) = \bar{c}/|x|$, one obtains $-\tilde{V}(\tilde{z}) = -\sum_{k=0}^{K} \hat{\tilde{V}}^k(\tilde{z}) - \hat{\tilde{V}}^0(\tilde{z}) = m\hat{\omega}^2\big[1 - \tilde{z}_1 + (\tilde{z}_1^2 - \tilde{z}_2^2/2)\big]$ and

$$-\hat{\tilde{V}}^k(\tilde{z}) = m\hat{\omega}^2 \sum_{j=0}^{k+2} c_{k+2,j}^V \tilde{z}_1^j \tilde{z}_2^{k+2-j} \quad \text{for } k \geq 1.$$

In particular, one should note that the change of variables leads to a lack of $\hat{\epsilon}^k$ in the expansion of the potential. With this reformulation in hand, consider the following assumption, where we note that $\breve{\mathscr{S}}_{\mathbb{C}}$ in (A.2) is replaced by $\mathscr{S}_{\mathbb{C}}$ in (A.2′).

$$\text{There exists a unique solution, } \tilde{\tilde{S}} \in \mathscr{S}_{\mathbb{C}} \text{ to (34)–(35).} \tag{A.2′}$$

Corollary 1 *Assume (A.2′) in place of (A.2). Assume there exists a unique solution, $\hat{S}^0$, in $\breve{\mathscr{S}}_{\mathbb{C}}$ to (23)–(24), and that for each $k \geq 1$, there exists a unique solution, $\hat{S}^k$, in $\breve{\mathscr{S}}_{\mathbb{C}}$ to (25)–(26). Then, $\bar{S} = \sum_{k=0}^{\infty} \hat{\epsilon}^k \hat{S}^k$.*

Proof Let $\tilde{\tilde{S}}$ satisfy (A.2′). Fix an arbitrary $\tilde{s} \in (0, \tilde{t})$, and let $D > 0$. Let $\mathscr{P}(D)$ denote the polydisc in $\mathbb{C}^2$ of multiradius $\bar{D} \doteq (D, D)$. By standard results (cf.[23]), for all $\tilde{z} \in \mathscr{P}(D)$,

$$\tilde{\tilde{S}}(\tilde{s}, \tilde{z}) = \sum_{l=0}^{\infty} \sum_{j=0}^{l} \frac{\tilde{\tilde{S}}_{z_1^j z_2^{l-j}}(\tilde{s}, 0)}{j!(l-j)!} \tilde{z}_1^j \tilde{z}_2^{l-j},$$

which through application of the Cauchy integral formula,

$$= \sum_{l=0}^{\infty} \sum_{j=0}^{l} \frac{1}{(2\pi i)^2} \int_{\partial \mathscr{P}(D)} \frac{\tilde{\tilde{S}}(\tilde{s}, \zeta_1, \zeta_2)}{\zeta_1^{j+1} \zeta_2^{l-j+1}} \, d\zeta_1 \, d\zeta_2 \tilde{z}_1^j \tilde{z}_2^{l-j} \quad \forall \, (\tilde{s}, \tilde{z}) \in (0, \infty) \times \mathbb{C}^2, \tag{36}$$

where $\partial \mathscr{P}(D) \doteq \{\zeta \in \mathbb{C}^2 \,|\, |\zeta_1| = D, \ |\zeta_2| = D\}$. For each $\tilde{s} \in (0, \tilde{t})$, we may express the Taylor series representation for $\tilde{\tilde{S}}$ as $\tilde{\tilde{S}}(\tilde{s}, \tilde{z}) = \sum_{l=0}^{\infty} \sum_{j=0}^{l} \tilde{\tilde{c}}_{l,j}(\tilde{s}) \tilde{z}_1^j \tilde{z}_2^{l-j}$ for all $\tilde{z} \in \mathbb{C}^2$. Let $0 \leq j \leq l < \infty$. Then, by (36) and the uniqueness of the Taylor expansion, we see that

$$\tilde{\tilde{c}}_{l,j}(\tilde{s}) = \frac{1}{(2\pi i)^2} \int_{\partial \mathscr{P}(D)} \frac{\tilde{\tilde{S}}(\tilde{s}, \zeta_1, \zeta_2)}{\zeta_1^{j+1} \zeta_2^{l-j+1}} \, d\zeta_1 \, d\zeta_2,$$

and the right-hand side is independent of $D \in (0, \infty)$. Further, letting $\zeta_\kappa = De^{i\theta_k}$ for $\kappa \in \{1, 2\}$ and $\zeta \in \partial \mathscr{P}(D)$, this becomes

$$\tilde{\tilde{c}}_{l,j}(\tilde{s}) = \frac{1}{(2\pi i)^2} \int_0^{2\pi} \int_0^{2\pi} \frac{-\tilde{\tilde{S}}(\tilde{s}, De^{i\theta_1}, De^{i\theta_2})}{D^l \exp\{i[(j+1)\theta_1 + (l-j+1)\theta_2]\}} \, d\theta_1 \, d\theta_2.$$

Let $\{\tilde{s}_n\} \subset (0, \tilde{t})$ be a sequence such that $\tilde{s}_n \to \tilde{s} \in (0, \tilde{t})$. By the Bounded Converegence Theorem, for any $0 \le j \le l < \infty$,

$$\lim_{n\to\infty} \tilde{\tilde{c}}_{l,j}(\tilde{s}_n) = \frac{1}{(2\pi i)^2} \lim_{n\to\infty} \int_0^{2\pi} \int_0^{2\pi} \frac{-\tilde{\tilde{S}}(\tilde{s}_n, De^{i\theta_1}, De^{i\theta_2})}{D^l \exp\{i[(j+1)\theta_1 + (l-j+1)\theta_2]\}} \, d\theta_1 \, d\theta_2$$

$$= \frac{1}{(2\pi i)^2} \int_0^{2\pi} \int_0^{2\pi} \frac{-\tilde{\tilde{S}}(\tilde{s}, De^{i\theta_1}, De^{i\theta_2})}{D^l \exp\{i[(j+1)\theta_1 + (l-j+1)\theta_2]\}} \, d\theta_1 \, d\theta_2 = \tilde{\tilde{c}}_{l,j}(\tilde{s}),$$

and we see that each $\tilde{\tilde{c}}_{l,j}(\cdot)$ is continuous.

Similarly, for $0 \le j \le l < \infty$,

$$\lim_{h\to 0} \frac{\tilde{\tilde{c}}_{l,j}(\tilde{s}+h) - \tilde{\tilde{c}}_{l,j}(\tilde{s})}{h} = \lim_{h\to 0} \frac{1}{(2\pi i)^2} \int_0^{2\pi} \int_0^{2\pi} \frac{-1}{D^l \exp\{i[(j+1)\theta_1 + (l-j+1)\theta_2]\}}$$

$$\cdot \frac{\tilde{\tilde{S}}(\tilde{s}+h, De^{i\theta_1}, De^{i\theta_2}) - \tilde{\tilde{S}}(\tilde{s}, De^{i\theta_1}, De^{i\theta_2})}{h} \, d\theta_1 \, d\theta_2.$$

Recalling that $\tilde{\tilde{S}}$ is continuously differentiable on $(0, \tilde{t})$, we find that the integrand is bounded, and another application of the Bounded Convergence Theorem yields

$$\lim_{h\to 0} \frac{\tilde{\tilde{c}}_{l,j}(\tilde{s}+h) - \tilde{\tilde{c}}_{l,j}(\tilde{s})}{h} = \frac{1}{(2\pi i)^2} \int_0^{2\pi} \int_0^{2\pi} \frac{-\tilde{\tilde{S}}_t(\tilde{s}, De^{i\theta_1}, De^{i\theta_2})}{D^l e^{i[(j+1)\theta_1 + (l-j+1)\theta_2]}} \, d\theta_1 \, d\theta_2,$$

and we see that $\tilde{\tilde{c}}_{l,j} \in C^1(0, \tilde{t})$.

Now, let $\bar{S}(s, z) \doteq \tilde{\tilde{S}}(\hat{g}_{\hat{\delta}}(s), \tilde{g}_{\hat{\delta}}(z))$ for all $(s, z) \in (0, t] \times \mathbb{C}^2$, and let $\hat{\epsilon} = 1/\hat{\delta}$. By Theorem 1, it is sufficient to show that $\bar{S}$ satisfies Assumption (A.2). We have $\bar{S}(s, z) = \sum_{l=0}^{\infty} \sum_{j=0}^{l} \tilde{\tilde{c}}_{l,j}(s) \hat{\epsilon}^l z_1^j z_2^{l-j}$ for all $(s, z) \in (0, t] \times \mathbb{C}^2$. Letting $\hat{S}^l(s, z) \doteq \sum_{j=0}^{l} \tilde{\tilde{c}}_{l,j}(s) z_1^j z_2^{l-j}$ for $l \in \mathbb{N}$, the smoothness assertions of the corollary then follow directly from the above and the composition of analytic functions. The existence and uniqueness are also easily demonstrated, and the steps are omitted. □

5 Periodic $\hat{S}^0$ Solutions

In order to begin computation of the terms in the expansion of Theorem 1, we must obtain a solution of the complex-valued, second-order, nonlinear HJ PDE problem given by (23)–(24). We note that we continue to work with the case of dimension $n = 2$ here. We will choose the initial condition, ϕ^0, such that the resulting solution will be periodic with frequency that is an integer multiple of ω, where we include the case where the multiple is zero (i.e., the steady-state case). We also note that we are seeking periodic solutions, $\hat{S}^0$ that are themselves clearly physically meaningful.

Recall that the original, forward-time solution, $\tilde{S}^f$, of (8)–(9) is a solution of the dequantized version of the original Schrödinger equation. Let $\tilde{\psi}^f(s,z) \doteq \exp\{\frac{i}{\hbar}\tilde{S}^f\}$ for all $(s,z) \in \overline{\mathscr{D}}^f \doteq [0,t) \times \mathbb{R}^2$. Recall also that for physically meaningful solutions, at each $s \in [0,t)$, $\tilde{P}^f(s,\cdot) : \mathbb{R}^2 \to \mathbb{R}$ given by $\tilde{P}(s,\cdot) \doteq [\psi^*\psi](s,\cdot)$ represents an unnormalized density associated to the particle at time s. Let $\tilde{R}^f, \tilde{T}^f : \overline{\mathscr{D}}^f \to \mathbb{R}$ be given by $\tilde{R}^f(s,z) \doteq \mathbf{Re}[\tilde{S}^f(s,z)]$ and $\tilde{T}^f(s,z) \doteq \mathbf{Im}[\tilde{S}^f(s,z)]$ for all $(s,z) \in \overline{\mathscr{D}}^f$. Then, $\tilde{P}^f(s,z) = \exp\{\frac{-2}{\hbar}\tilde{T}^f(s,z)\}$ for all $(s,z) \in \overline{\mathscr{D}}^f$. This suggests that we should seek $\tilde{S}^f$ such that $\exp\{\frac{-2}{\hbar}\tilde{T}^f(s,\cdot)\}$ represents an unnormalized probability density for all $s \in [0,t)$.

Although the goal in this section is to generate a set of physically meaningful periodic solutions to the zeroth-order term, we do not attempt a full catalog of all possible such solutions. Let $\hat{S}^{0,f}(s,z) \doteq \hat{S}^0(t-s,z)$ for all $(s,z) \in \overline{\mathscr{D}}^f$. As we seek $\hat{S}^0(t-s,\cdot)$ that are quadratic, we let the resulting time-dependent coefficients be defined by

$$\hat{S}^{0,f}(s,z) = \tfrac{1}{2} z^T Q(s) z + \Lambda^T(s) z + \rho(s). \tag{37}$$

It should be noted here that the condition that $\exp\{\frac{-\hbar}{2}\tilde{T}^f(s,\cdot)\}$ represent an unnormalized density implies that the imaginary part of $Q(s)$ should be nonnegative definite for all $s \in [0,t)$, which is a significant restriction on the set of allowable solutions.

As $\hat{S}^{0,f}(s,\cdot)$ is quadratic, its values over $\mathbb{C}^2$ are fully defined by its values over $\mathbb{R}^2$, and hence it is sufficient to solve the problem on the real domain. The forward-time version of (23)–(24), with domain restricted to $\overline{\mathscr{D}}^f$ is

$$0 = -S_t^{0,f} + \tfrac{i\hbar}{2m}\Delta_z S^{0,f} - \big(A_0 z + b_0\big)^T S_z^{0,f} - \tfrac{1}{2m}|S_z^{0,f}|_c^2 - \hat{V}^0, \quad (s,z) \in (0,t) \times \mathbb{R}^2, \tag{38}$$

$$S^{0,f}(0,z) = \phi^0(z) \quad \forall\, z \in \mathbb{R}^2. \tag{39}$$

Remark 4 It is worth noting that any solution of form (37) to (38)–(39) is the unique solution in $\mathscr{S}^p_{\mathbb{C}}$, and in particular, where this uniqueness is obtained through a controlled-diffusion representation [17, 18].

Substituting form (37) into (38), and collecting terms, yields the system of ordinary differential equations (ODEs) given as

$$\frac{d}{ds}Q(s) = -\big(A_0^T Q(s) + Q(s)A_0\big) - \tfrac{1}{m}Q^2(s) + m\omega^2 T^V, \tag{40}$$

$$\frac{d}{ds}\Lambda(s) = -\big(A_0^T + \tfrac{1}{m}Q(s)\big)\Lambda + \omega\hat{\delta}Q(s)u^2 - m\omega^2\hat{\delta}u^1, \tag{41}$$

$$\frac{d}{ds}\rho(s) = \tfrac{i\hbar}{2m}\,\mathrm{tr}[Q(s)] + \omega\hat{\delta}(u^2)^T\Lambda(s) - \tfrac{1}{2m}\Lambda^T(s)\Lambda(s) + m\omega^2\hat{\delta}^2, \tag{42}$$

$$T^V = \begin{bmatrix} 2 & 0 \\ 0 & -1 \end{bmatrix}, \qquad u^1 \doteq \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad u^2 \doteq \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \tag{43}$$

where $Q : [0,t) \to \mathbb{C}^{2\times 2}$, $\Lambda : [0,t) \to \mathbb{C}^2$ and $\rho : [0,t) \to \mathbb{C}$. Throughout, we assume that $Q(s)$ is symmetric for all $s \in [0,t)$. Note that if $Q(s)$ is nonsingular for all $s \in [0,t)$, then (37) may also be written as

$$\hat{S}^{0,f}(s,z) = \tfrac{1}{2}\big(z + Q^{-1}(s)\Lambda(s)\big)^T Q(s)\big(z + Q^{-1}(s)\Lambda(s)\big) + \rho(s) - \Lambda^T(s)Q^{-1}(s)\Lambda(s),$$

where we see that $-Q^{-1}(s)\Lambda(s)$ may be interpreted as a mean of the associated distribution at each time s. Consequently, we look for solutions with $-Q^{-1}(s)\Lambda(s) \in \mathbb{R}^2$ for all s.

One may use a Bernoulli-type substitution as a means for seeking solutions of (40). That is, suppose $Q(s) = W(s)U^{-1}(s)$, where $U(s)$ is nonsingular for all $s \in [0,t)$. Then, without loss of generality, we may take $W(0) = Q(0)$, $U(0) = \mathscr{I}_{2\times 2}$. The resulting system of ODEs is

$$\frac{d}{ds}\begin{pmatrix} U \\ W \end{pmatrix} = \mathscr{B}\begin{pmatrix} U \\ W \end{pmatrix}, \qquad \mathscr{B} \doteq \begin{bmatrix} 0 & \omega & 1/m & 0 \\ -\omega & 0 & 0 & 1/m \\ 2m\omega^2 & 0 & 0 & \omega \\ 0 & -m\omega^2 & -\omega & 0 \end{bmatrix}.$$

Employing the Jordan canonical form, one obtains the solution as

$$\Big(U(s)^T, W(s)^T\Big)^T = RPe^{J\omega s}P^{-1}R^{-1}\left(\mathscr{I}_{2\times 2}, Q(0)\right)^T, \tag{44}$$

where

$$P = \begin{bmatrix} 0 & 2 & -i & i \\ -3 & 0 & 2 & 2 \\ 3 & 0 & -1 & -1 \\ 0 & -1 & i & -i \end{bmatrix}, \qquad P^{-1} = \begin{bmatrix} 0 & 1/3 & 2/3 & 0 \\ 1 & 0 & 0 & 1 \\ -i/2 & 1/2 & 1/2 & -i \\ i/2 & 1/2 & 1/2 & i \end{bmatrix},$$

$$e^{J\omega s} = \begin{bmatrix} 1 & \omega s & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \exp\{i\omega s\} & 0 \\ 0 & 0 & 0 & \exp\{-i\omega s\} \end{bmatrix}, \qquad R = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & m\omega & 0 \\ 0 & 0 & 0 & m\omega \end{bmatrix}.$$

We remark that, as our goal here is the generation of periodic solutions that may be used as a basis for the expansion to follow, and as this work is already of substantial length, we will not discuss the question of stability of the above solutions, to perturbations within the class of physically meaningful $Q = U^{-1}W$.

Note that we seek solutions that generate periodic densities $\tilde{P}^f(s,\cdot)$, and that the $(1,2)$ entry of $e^{J\omega s}$ has secular behavior. Examining (44), we see that a sufficient condition for avoidance of secular growth/decay of Q, is that entries in the second row of $P^{-1}R^{-1}(\mathscr{I}_{2\times 2}, Q(0))^T$ be zero. One easily sees that this corresponds to $Q_{2,1}(0) = -m\omega$ and $Q_{2,2}(0) = 0$, and considering here only symmetric Q, we take $Q_{1,2}(0) = -m\omega$. That is, we have

$$Q(0) = \begin{bmatrix} \bar{k}_0 im\omega & -m\omega \\ -m\omega & 0 \end{bmatrix}, \tag{45}$$

for some $\bar{k}_0 \in \mathbb{C}$. Propagating the resulting solutions, we find that the imaginary part of $\bar{k}_0$ being nonnegative is necessary and sufficient for satisfaction of the condition that $\mathbf{Im}[Q(s)]$ be nonnegative-definite for all s. We also note that with such initial condition, $Q_{1,2}$, $Q_{2,1}$, $Q_{2,2}$ remain constant for all s, while the real and imaginary parts of $Q_{1,1}$ are periodic. That is, $Q(s)$ takes the form

$$Q(s) = \begin{bmatrix} im\omega p(s) & -m\omega \\ -m\omega & 0 \end{bmatrix} \quad \forall\, s \in [0,t), \tag{46}$$

where $p(s) = [\bar{k}_1^+ e^{2i\omega s} + \bar{k}_1^-]/[\bar{k}_1^+ e^{2i\omega s} - \bar{k}_1^-]$ with $\bar{k}_1^+ \doteq \bar{k}_0 + 1$ and $\bar{k}_1^- \doteq \bar{k}_0 - 1$.

One may seek steady-state solutions by substitution of form (45) into the right-hand side of (40), and setting this to be zero. One easily finds that the unique steady state solution (among those with $\mathbf{Im}[\bar{k}_0] \geq 0$) is

$$Q(s) = \bar{Q}^0 \doteq \begin{bmatrix} im\omega & -m\omega \\ -m\omega & 0 \end{bmatrix} \quad \forall\, s \in [0,t). \tag{47}$$

Next we consider the linear term in $\hat{S}^{0,f}$, where this satisfies (41). We focus on the steady-state Q case of (47). Substituting (47) into (41) yields

$$\dot{\Lambda} = \begin{bmatrix} -i\omega & 2\omega \\ 0 & 0 \end{bmatrix} \Lambda - 2m\omega^2 \hat{\delta} u^1.$$

This has a steady-state solution in the case that $-i\Lambda_1(0) + 2\Lambda(0) = 2m\omega\hat{\delta}$, or equivalently, the one-parameter set of steady-state solutions given by $\Lambda(s) = \bar{\Lambda}^0 \doteq \left(id, m\omega\hat{\delta} - d/2\right)^T$ for $d \in \mathbb{C}$. This includes, in particular, the cases $\bar{\Lambda}^0 = (0, m\omega\hat{\delta})^T$ (i.e., $d = 0$) and $\bar{\Lambda}^0 = m\omega\hat{\delta}(-2i, 2)^T$ (i.e., $d = -2m\omega\hat{\delta}$), where this latter case is obtained if one requires $(\bar{Q}^0)^{-1}\bar{\Lambda}^0$ to be real valued. We also remark that more generally, the solution is given for all $s \in [0,t)$ by

$$\Lambda(s) = \begin{bmatrix} -i\exp\{-i\omega s\} & 2i[\exp\{-i\omega s\} - 1] \\ 0 & 0 \end{bmatrix} \Lambda(0) + 2i[1 - \exp\{-i\omega s\}] m\omega\hat{\delta} u^1.$$

Lastly, we turn to the zeroth-order term. Note that the one may allow secular growth in the real part of $\rho(\cdot)$ with no effect on the associated probability distribution, as is standard in solutions of the quantum harmonic oscillator. Continuing to focus on the steady-state solution, but allowing a real-valued secular zeroth-order term, we substitute the above steady-state quadratic and linear coefficients into (42). This yields

$$\dot{\rho} = \frac{-\hbar\omega}{2} + m\omega^2\left[\frac{3\hat{\delta}^2}{2} + \frac{3d^2}{4(m\omega)^2}\right] \doteq \bar{c}_1(d),$$

and we see that this is purely real if and only if $d \in \mathbb{R}$, and we have

$$\Lambda(s) = \bar{\Lambda}^0 \doteq \left(id, m\omega\hat{\delta} - d/2\right)^T, \qquad \rho^0(s) = \rho^0(0) + \bar{c}_1(d)s \quad \forall\, s \in [0,t). \tag{48}$$

We will restrict ourselves to the simple, steady-state case (modulo the real part of ρ^0) given by (47), (48) with $\bar{k}_0 = 1$, $d = 0$, for our actual computations of succeeding terms in the expansion. However, the theory will be sufficiently general to encompass the periodic case as well.

6 Diffusion Representations for Succeeding Terms

As noted above, we will use diffusion representations to obtain the solutions to the HJ PDEs (25)–(26) that define the succeeding terms in the expansion, i.e., to obtain the $\hat{S}^k$ for $k \in \mathbb{N}$. In order to achieve this goal, we need to define the complex-valued diffusion dynamics and the expected payoffs that will yield the $\hat{S}^k$. The representation result naturally employs the Itô integral rule. As the dynamics are complex-valued, we need an extension of the Itô rule to that process domain. In a similar fashion to that of Sect. 3.3, we use the Itô rule for the double-dimension real case to obtain the rule for the complex case. Once the Itô rule is established, the proof of the representation is straightforward. However, additional effort is require to generate the machinery by which the the actual solutions are computed, where the machinery relies mainly on computation of moments for the diffusion process.

6.1 The Underlying Stochastic Dynamics

We let $(\Omega, \mathscr{F}, P)$ be a probability triple, where Ω denotes a sample space, $\mathscr{F}$ denotes a σ-algebra on Ω, and P denotes a probability measure on $(\Omega, \mathscr{F})$. Let $\{\mathscr{F}_s \,|\, s \in [0,t]\}$ denote a filtration on $(\Omega, \mathscr{F}, P)$, and let $B.$ denote an $\mathscr{F}.$-adapted Brownian motion taking values in $\mathbb{R}^n$. We will be interested in diffusion processes

given by the linear stochastic differential equation (SDE) in integral form

$$\zeta_r = \zeta_r^{(s,z)} = z + \int_s^r -\big(A_0\zeta_\rho + b_0 + \tfrac{1}{m}\hat{S}_z^0(\rho,\zeta_\rho)\big)\,d\rho + \sqrt{\tfrac{\hbar}{m}}\tfrac{1+i}{\sqrt{2}}\int_s^r dB_\rho$$
$$\doteq z + \int_s^r \lambda(\rho,\zeta_\rho)\,d\rho + \sqrt{\tfrac{\hbar}{m}}\tfrac{1+i}{\sqrt{2}}\,B_r^\Delta \quad \forall\, r\in[s,t], \tag{49}$$

where $z\in\mathbb{C}^2$, $s\in[0,t)$, $B_r^\Delta \doteq B_r - B_s$ for $r\in[s,t)$, and

$$\lambda(\rho,z) \doteq -[A_0 z + b_0 + \tfrac{1}{m}S_z^0(\rho,z)] = -[A_0 z + b_0 + \tfrac{1}{m}Q(\rho)z + \tfrac{1}{m}\Lambda(\rho)]$$
$$\doteq -A_{>0}(\rho)z - b_{>0}(\rho). \tag{50}$$

Let $\bar{f}:[0,t]\times\mathbb{C}^2\to\mathbb{C}^2$, and suppose there exists $K_{\bar{f}}<\infty$ such that $|\bar{f}(s,z^1)-\bar{f}(s,z^2)|\le K_{\bar{f}}|z^1-z^2|$ for all $(s,z^1),(s,z^2)\in\overline{\mathscr{D}}_{\mathbb{C}}$. For $(s,z)\in\overline{\mathscr{D}}_{\mathbb{C}}$, consider the complex-valued diffusion, $\zeta_\cdot\in\mathscr{X}_s$, given by

$$\zeta_r = \zeta_r^{(s,z)} = z + \int_s^r \bar{f}(\rho,\zeta_\rho)\,d\rho + \int_s^r \tfrac{1+i}{\sqrt{2}}\sigma\,dB_\rho, \tag{51}$$

where $\sigma\in\mathbb{R}^{n\times n}$, and note that this is a slight generalization of (49). For $s\in(0,t]$, let

$$\mathscr{X}_s \doteq \{\zeta : [s,t]\times\Omega\to\mathbb{C}^2 \mid \zeta \text{ is } \mathscr{F}_\cdot\text{-adapted, right-continuous and such that } \mathbb{E}\sup_{r\in[s,t]}|\zeta_r|^m<\infty\ \forall m\in\mathbb{N}\}. \tag{52}$$

We supply $\mathscr{X}_s$ with the norm $\|\zeta\|_{\mathscr{X}_s} \doteq \max_{m\in]1,\bar{M}[}\big[\mathbb{E}\sup_{r\in[s,t]}|\zeta_r|^m\big]^{1/m}$. It is important to note here that complex-valued diffusions have been discussed elsewhere in the literature; see for example, [25] and the references therein.

We also define the isometric isomorphism, $\mathscr{V}:\mathscr{X}_s\to\mathscr{X}_s^v$ by $[\mathscr{V}(\zeta)]_r \doteq [\mathscr{V}(\xi+i\nu)]_r \doteq (\xi_r^T,\nu_r^T)^T$ for all $r\in[s,t]$ and $\omega\in\Omega$, where

$$\mathscr{X}_s^v \doteq \{(\xi,\nu) : [s,t]\times\Omega\to\mathbb{R}^{2n} \mid (\xi,\nu) \text{ is } \mathscr{F}_\cdot\text{-adapted, right-continuous and such that } \mathbb{E}\sup_{r\in[s,t]}[|\xi_r|^m+|\nu_r|^m]<\infty\ \forall m\in\mathbb{N}\}, \tag{53}$$

$$\|(\xi,\nu)\|_{\mathscr{X}_s^v} \doteq \max_{m\in]1,\bar{M}[}\Big[\mathbb{E}\sup_{r\in[s,t]}(|\xi_r|^m+|\nu_r|^m)\Big]^{1/m}. \tag{54}$$

Under transformation by $\mathscr{V}$, (51) becomes

$$\begin{pmatrix}\xi_r\\ \nu_r\end{pmatrix} = \begin{pmatrix}x\\ y\end{pmatrix} + \int_s^r \hat{f}(\rho,\xi_\rho,\nu_\rho)\,d\rho + \int_s^r \tfrac{1}{\sqrt{2}}\hat{\sigma}\,dB_\rho \quad \forall\, r\in[s,t], \tag{55}$$

where $\hat{f}(\rho, \xi_\rho, \nu_\rho) \doteq \left((\mathbf{Re}[\bar{f}(\rho, \xi_\rho + i\nu_\rho)])^T, (\mathbf{Im}[\bar{f}(\rho, \xi_\rho + i\nu_\rho)])^T\right)^T$ and $\hat{\sigma} \doteq (1, 1)^T$. Throughout, concerning both real and complex stochastic differential equations, typically given in integral form such as in (51) and (55), *solution* refers to a strong solution, unless specifically cited as a weak solution. The following are easily obtained from existing results; see [18, 22].

Lemma 2 *Let $s \in [0, t)$, $z \in \mathbb{C}^2$ and $(x, y) = \mathscr{V}_0(z)$. There exists a unique solution, $(\xi, \nu) \in \mathscr{X}_s^v$, to (55).*

Lemma 3 *Let $s \in [0, t)$, $z \in \mathbb{C}^2$ and $(x, y) = \mathscr{V}_0(z)$. $\zeta \in \mathscr{X}_s$ is a solution of (51) if and only if $\mathscr{V}(\zeta) \in \mathscr{X}_s^v$ is a solution of (55).*

Lemma 4 *Let $s \in [0, t)$ and $z \in \mathbb{C}^2$. There exists a unique solution, $\zeta \in \mathscr{X}_s$, to (51).*

We remark that one may apply Lemmas 2–4 to the specific case of (49) in order to establish existence and uniqueness. In particular, for the dynamics of (49), the corresponding process $(\xi, \nu) = \mathscr{V}(\zeta)$ satisfies

$$\begin{aligned}\begin{pmatrix}\xi_r\\ \nu_r\end{pmatrix} &= \begin{pmatrix}x\\ y\end{pmatrix} + \int_s^r -\left[\begin{pmatrix}A^r_{>0}(\rho) & -A^i_{>0}(\rho)\\ A^i_{>0}(\rho) & A^r_{>0}(\rho)\end{pmatrix}\begin{pmatrix}\xi_r\\ \nu_r\end{pmatrix} + \begin{pmatrix}b^r_{>0}(\rho)\\ b^i_{>0}(\rho)\end{pmatrix}\right] d\rho\\ &\quad + \sqrt{\tfrac{\hbar}{2m}}\begin{pmatrix}I_{n\times n}\\ I_{n\times n}\end{pmatrix} B_r^\Delta\\ &\doteq \begin{pmatrix}x\\ y\end{pmatrix} + \int_s^r -\bar{A}_{>0}(\rho)\begin{pmatrix}\xi_r\\ \nu_r\end{pmatrix} - \bar{b}_{>0}(\rho)\, d\rho + \sqrt{\tfrac{\hbar}{2m}}\,\bar{\mathscr{I}} B_r^\Delta \quad \forall\, r \in [s, t],\end{aligned} \tag{56}$$

where $A^r_{>0}(\rho) \doteq \mathbf{Re}(A_{>0}(\rho))$, $A^i_{>0}(\rho) \doteq \mathbf{Im}(A_{>0}(\rho))$, $\left((b^r_{>0}(\rho))^T, (b^i_{>0}(\rho))^T\right)^T \doteq \mathscr{V}_0(b_{>0}(\rho))$ for all $\rho \in [0, t)$.

6.2 Itô's Rule

The representation results will rely on a minor generalization of Itô's rule to the specific complex-diffusion dynamics of interest here. It might be worthwhile to note that the complex-valued diffusions considered here belong to a very small subclass of complex-valued diffusions, and this is somehow related to the specific nature of the complex aspect of the Schrödinger equation. The following complex-case Itô rule is similar to existing results (cf., [25]).

Lemma 5 *Let $\bar{g} \in \mathscr{S}_{\mathbb{C}}$ and $(s,z) \in \overline{\mathscr{D}}_{\mathbb{C}}$, and suppose diffusion process $\zeta_\cdot$ is given by (51). Then, for all $r \in [s,t]$,*

$$\bar{g}(r,\zeta_r) = \bar{g}(s,z) + \int_s^r \bar{g}_t(\rho,\zeta_\rho) + \bar{g}_z^T(\rho,\zeta_\rho)\bar{f}(\rho,\zeta_\rho)\,d\rho + \int_s^r \tfrac{1+i}{\sqrt{2}}\bar{g}_z^T(\rho,\zeta_\rho)\sigma\,dB_\rho + \tfrac{1}{2}\int_s^r \mathrm{tr}\big[\bar{g}_{zz}(\rho,\zeta_\rho)(\sigma\sigma^T)\big]\,d\rho. \tag{57}$$

Proof Let $(g^r(s,x,y), g^i(s,x,y)) \doteq \mathscr{V}_{00}\big(\bar{g}(s,\mathscr{V}_0^{-1}(x,y))\big)$, $(f^r(s,x,y), f^i(s,x,y)) \doteq \mathscr{V}_0\big(\bar{f}(s,\mathscr{V}_0^{-1}(x,y))\big)$ for all $(s,x,y) \in \overline{\mathscr{D}}_2$, and note that it is trivial to show that $\bar{g}_t(r,z) = g_t^r(r,x,y) + ig_t^i(r,x,y)$, for all $(x,y) = \mathscr{V}_0(z)$, $(r,z) \in \mathscr{D}_{\mathbb{C}}$. Also, using the Cauchy-Riemann equations,

$$\bar{g}_z^T(r,z)\bar{f}(r,z) = \big[(g_x^r)^T f^r + (g_y^r)^T f^i\big](r,x,y) + i\big[(g_x^i)^T f^r + (g_y^i)^T f^i\big](r,x,y),$$

for all $(x,y) = \mathscr{V}_0(z)$, $(r,z) \in \overline{\mathscr{D}}_{\mathbb{C}}$. Defining the derivative notation $g_{x^2}^r(s,x,y) \doteq \big((g_x^r)^T, (g_y^r)^T\big)^T(r,x,y)$ and vector notation $\hat{f}(r,x,z) \doteq \big((f^r)^T, (f^i)^T\big)^T(r,x,y)$ for all $(r,x,y) \in \overline{\mathscr{D}}_2$, this becomes

$$\bar{g}_z^T(r,z)\bar{f}(r,z) = \big(g_{x^2}^r(r,x,y)\big)^T\hat{f}(r,x,y) + i\big(g_{x^2}^i(r,x,y)\big)^T\hat{f}(r,x,y), \tag{58}$$

for all $(x,y) = \mathscr{V}_0(z)$, $(r,z) \in \overline{\mathscr{D}}_{\mathbb{C}}$. Similarly, letting $\hat{\sigma} \doteq \big(\sigma^T, \sigma^T\big)^T$,

$$\bar{g}_z^T(r,z)\tfrac{1+i}{\sqrt{2}}\sigma = \tfrac{1}{\sqrt{2}}\big[\big(g_{x^2}^r(r,x,y)\big)^T\hat{\sigma} + i\big(g_{x^2}^i(r,x,y)\big)^T\hat{\sigma}\big], \tag{59}$$

for all $(x,y) = \mathscr{V}_0(z)$, $(r,z) \in \overline{\mathscr{D}}_{\mathbb{C}}$.

Next, let $\bar{a} \doteq \big(\frac{1+i}{\sqrt{2}}\big)^2\sigma\sigma^T$ and $(a_{j,l}^r, a_{j,l}^i) \doteq \mathscr{V}_{00}(\bar{a}_{j,l})$ for all $j,l \in]1,n[$. Using (17), (18), we find

$$\bar{g}_{z_j,z_l}\bar{a}_{j,l} = g_{x_j,y_l}^i a_{j,l}^r + g_{x_j,y_l}^r a_{j,l}^i + i\big[-g_{x_j,y_l}^r a_{j,l}^r + g_{x_j,y_l}^i a_{j,l}^i\big] \quad \forall\, j,l \in]1,n[\,. \tag{60}$$

Also, by the definition of $\bar{a}$, we see that $a^r = 0$ and $a^i = \sigma\sigma^T$. Applying these in (60) yields

$$\bar{g}_{z_j,z_l}\bar{a}_{j,l} = g_{x_j,y_l}^r[\sigma\sigma^T]_{j,l} + ig_{x_j,y_l}^i[\sigma\sigma^T]_{j,l} \quad \forall\, j,l \in]1,n[\,. \tag{61}$$

Considering (58), (59) and (61), and letting $(\xi_r, \nu_r) \doteq \mathscr{V}_0(\zeta_r)$ for all $r \in (0,t]$, a.e. $\omega \in \Omega$ we see that (57) is equivalent to a pair of equations for the real and

imaginary parts, where the real-part equation is

$$g^r(r,\xi_r,\nu_r) = g^r(s,x,y) + \int_s^r g_t^r(\rho,\xi_\rho,\nu_\rho) + (g_{x^2}^r)^T(\rho,\xi_\rho,\nu_\rho)\hat{f}(\rho,\xi_\rho,\nu_\rho) + \tfrac{1}{2}\sum_{j,l=1}^n g_{x_j,y_l}^r(\rho,\xi_\rho,\nu_\rho)[\sigma\sigma^T]_{j,l}\,d\rho + \tfrac{1}{\sqrt{2}}\int_s^r (g_{x^2}^r)^T(\rho,\xi_\rho,\nu_\rho)\hat{\sigma}\,dB_\rho, \tag{62}$$

with an analogous equation corresponding to the imaginary part.

On the other hand, applying Itô's rule to real functions g^r and g^i separately, and then applying (17), (18), we find

$$\begin{aligned} g^r(r,\xi_r,\nu_r) =& g^r(s,x,y) + \int_s^r g_t^r(\rho,\xi_\rho,\nu_\rho) + (g_{x^2}^r)^T(\rho,\xi_\rho,\nu_\rho)\hat{f}(\rho,\xi_\rho,\nu_\rho) \\ &+ \tfrac{1}{4}\sum_{j,l=1}^n \big[g_{x_j,x_l}^r + g_{x_j,y_l}^r + g_{y_j,x_l}^r + g_{y_j,y_l}^r\big](\rho,\xi_\rho,\nu_\rho)[\sigma\sigma^T]_{j,l}\,d\rho \\ &+ \tfrac{1}{\sqrt{2}}\int_s^r (g_{x^2}^r)^T(\rho,\xi_\rho,\nu_\rho)\hat{\sigma}\,dB_\rho, \\ =& g^r(s,x,y) + \int_s^r g_t^r(\rho,\xi_\rho,\nu_\rho) + (g_{x^2}^r)^T(\rho,\xi_\rho,\nu_\rho)\hat{f}(\rho,\xi_\rho,\nu_\rho) \\ &+ \tfrac{1}{2}\sum_{j,l=1}^n g_{x_j,y_l}^r(\rho,\xi_\rho,\nu_\rho)[\sigma\sigma^T]_{j,l}\,d\rho + \tfrac{1}{\sqrt{2}}\int_s^r (g_{x^2}^r)^T(\rho,\xi_\rho,\nu_\rho)\hat{\sigma}\,dB_\rho, \end{aligned} \tag{63}$$

with a similar equation for the imaginary part. Comparing (63) with (62), and similarly for the imaginary parts, one obtains the result. □

We apply this result to the particular case of interest here.

Lemma 6 *Let $\hat{S} \in \mathscr{S}_{\mathbb{C}}$ and $(s,z) \in \overline{\mathscr{D}}_{\mathbb{C}}$, and suppose ζ satisfies (49). Then, for all $r \in (s,t]$,*

$$\hat{S}(r,\zeta_r) = \hat{S}(s,z) + \int_s^r \hat{S}_t(\rho,\zeta_\rho) - \hat{S}_z^T(\rho,\zeta_\rho)\big[A_{>0}(\rho)\zeta_\rho + b_{>0}(\rho)\big] + \tfrac{i\hbar}{2m}\Delta_z\hat{S}(\rho,\zeta_\rho)\,d\rho + \sqrt{\tfrac{\hbar}{m}}\tfrac{1+i}{\sqrt{2}}\int_s^r \hat{S}_z^T(\rho,\zeta_\rho)\,dB_\rho. \tag{64}$$

Proof Dynamics (49) have form (51) with $f(r,z) = A_{>0}(r)z + b_{>0}(r)$ and $\sigma = \sqrt{\frac{\hbar}{m}}\mathscr{I}_{n\times n}$. In this case, $\frac{1}{2}\operatorname{tr}\big[\hat{S}_{zz}(r,z)(\sigma\sigma^T)\big] = \frac{i\hbar}{2m}\Delta_z\hat{S}(r,z)$ for all $(r,z) \in \mathscr{S}_{\mathbb{C}}$, which yields the result. □

Theorem 2 *Let $k \in \mathbb{N}$. Let $\hat{S}^\kappa \in \mathscr{I}^p_{\mathbb{C}}$ satisfy (25)–(26) for all $\kappa \in]1, k[$. Let $(s,z) \in \overline{\mathscr{D}}_{\mathbb{C}}$, and let $\zeta \in \mathscr{X}_s$ satisfy (49). Then,*

$$\hat{S}^k(s,z) = \mathbb{E}\Big\{ \int_s^t -\tfrac{1}{2m} \sum_{\kappa=1}^{k-1} \big[S_z^\kappa(r,\zeta_r)\big]^T S_z^{k-\kappa}(r,\zeta_r) - \hat{V}^k(\zeta_r)\,dr + \phi^k(\zeta_t) \Big\}.$$

Proof Taking expectations in (64), and using the martingale property (cf., [6, 8]), we have

$$\hat{S}^k(s,z) = \mathbb{E}\Big\{ -\int_s^t \hat{S}_t(r,\zeta_r) - \hat{S}_z^T(r,\zeta_r)\big[A_{>0}(r)\zeta_r + b_{>0}(r)\big] + \tfrac{i\hbar}{2m}\Delta_z \hat{S}(r,\zeta_r)\,dr + \hat{S}^k(t,\zeta_t)\Big\}.$$

Combining this with (25)–(26) yields the result. □

6.3 Moments and Iteration

Note that Theorem 2 yields an expression for the kth term in our expansion for $\bar{S}$, $\hat{S}^k$, from the previous terms, $\hat{S}^\kappa$ for $\kappa < k$. We now examine how this generates a computationally tractable scheme. It is heuristically helpful to examine the first two iterates. For $(s,z) \in \overline{\mathscr{D}}_{\mathbb{C}}$, we have

$$\begin{aligned}\hat{S}^1(s,z) &= \mathbb{E}\Big\{ \int_s^t -\hat{V}^1(\zeta_r)\,dr + \phi^1(\zeta_t) \Big\} \\ &= \mathbb{E}\Big\{ \int_s^t m\omega^2\Big(-[\zeta_r]_1^3 + (3/2)[\zeta_r]_1[\zeta_r]_2^2 \Big)\,dr + \sum_{l=0}^{3}\sum_{j=0}^{l} b^\phi_{3,l,j}[\zeta_t]_1^j[\zeta_t]_2^{l-j} \Big\},\end{aligned} \tag{65}$$

$$\begin{aligned}\hat{S}^2(s,z) &= \mathbb{E}\Big\{ \int_s^t -\tfrac{1}{2m}\big|\hat{S}_z^1(r,\zeta_r)\big|_c^2 - \hat{V}^2(\zeta_r)\,dr + \phi^2(\zeta_t) \Big\} \\ &= \mathbb{E}\Big\{ \int_s^t -\tfrac{1}{2m}\big|\hat{S}_z^1(r,\zeta_r)\big|_c^2 + m\omega^2\Big([\zeta_r]_1^4 - 3[\zeta_r]_1^2[\zeta_r]_2^2 + (3/8)[\zeta_r]_2^4 \Big)\,dr \\ &\qquad + \sum_{l=0}^{4}\sum_{j=0}^{l} b^\phi_{4,l,j}[\zeta_t]_1^j[\zeta_t]_2^{l-j} \Big\}.\end{aligned} \tag{66}$$

Note that the right-hand side of (65) consists of an expectation of a polynomial in ζ_t and an integral of a polynomial in $\zeta_\cdot$, and further, that the dynamics of ζ are linear in the state variable. Thus, we may anticipate that $\hat{S}^1(s,\cdot)$ may also be

polynomial. Applying this anticipated form on the right-hand side of (66) suggests that the polynomial form will be inherited in each $\hat{S}^k$. This will form the basis of our computational scheme.

The computation of the expectations that generate the $\hat{S}^k$ for $k \geq 1$ will be obtained through the moments of the underlying diffusion process. Thus, the first step is solution of (49). We let the state transition matrices for deterministic linear systems $\dot{y}_r = -A_{>0}(r)y_r$ and $\dot{y}_r^{(2)} = -\bar{A}_{>0}(r)y_r^{(2)}$ be denoted by $\Phi(r,s)$ and $\Phi^{(2)}(r,s)$, respectively. More specifically, with initial (or terminal) conditions, $y_s = \bar{y}$ and $y_s^{(2)} = \bar{y}^{(2)}$, the solutions at time r are given by $y_r = \Phi(r,s)\bar{y}$ and $y_r^{(2)} = \Phi^{(2)}(r,s)\bar{y}^{(2)}$, respectively. The solutions of our SDEs are given by the following.

Lemma 7 *Linear SDE (49) has solution given by $\zeta_r = \mu_r + \Delta_r$, where*

$$\mu_r = \Phi(r,s)z + \int_s^r \Phi(r,\rho)\big(-b_{>0}(\rho)\big)\,d\rho, \quad \Delta_r = \sqrt{\tfrac{\hbar}{m}}\tfrac{1+i}{\sqrt{2}}\int_s^r \Phi(r,\rho)\,dB_\rho$$

for all $r \in [s,t]$. Linear SDE (56) has solution given by $X_r^{(2)} = \mu_r^{(2)} + \Delta_r^{(2)}$, where

$$\mu_r^{(2)} = \Phi^{(2)}(r,s)x^{(2)} + \int_s^r \Phi^{(2)}(r,\rho)\big(-\bar{b}_{>0}(\rho)\big)\,d\rho,$$

$$\Delta_r^{(2)} = \sqrt{\tfrac{\hbar}{2m}}\int_s^r \Phi^{(2)}(r,\rho)\bar{\mathscr{I}}\,dB_\rho$$

for all $r \in [s,t]$, where $x^{(2)} \doteq (x^T, y^T)^T$.

Proof The case of (56) is standard, cf., [12]. We sketch the proof in the minor variant case of (49), where this uses the Itô-rule approach, but for the complex-valued diffusion case. For $0 \leq s \leq r \leq t$, let $\alpha_r \doteq \Phi(s,r)\zeta_r = \Phi^{-1}(r,s)\zeta_r$. By Lemma 5,

$$\alpha_r = \int_s^r \Phi^{-1}(\rho,s)[-b_{>0}(\rho)]\,d\rho + \sqrt{\tfrac{\hbar}{m}}\tfrac{1+i}{\sqrt{2}}\int_s^r \Phi^{-1}(\rho,s)\,dB_\rho,$$

which implies $\zeta_r = \int_s^r \Phi(r,\rho)[-b_{>0}(\rho)]\,d\rho + \sqrt{\frac{\hbar}{m}}\frac{1+i}{\sqrt{2}}\int_s^r \Phi(r,\rho)\,dB_\rho$. □

Lemma 8 *For all $r \in [s,t]$, $X_r^{(2)}$ and ζ_r have normal distributions.*

Proof The case of $X_r^{(2)}$ is standard, cf. [11], and then one notes $\zeta_r = \mathscr{V}_0(X_r^{(2)})$. □

Lemma 9 *For all $r \in [s,t]$, μ_r is the mean of ζ_r, and Δ_r is a zero-mean normal random variable with covariance given by $\mathbb{E}\big[\Delta_r\Delta_r^T\big] = \frac{i\hbar}{m}\int_s^r \Phi(r,\rho)\Phi^T(r,\rho)\,d\rho$, where further, $\mathbb{E}\big[(\zeta_r - \mu_r)(\zeta_r - \mu_r)^T\big] = \mathbb{E}\big[\Delta_r\Delta_r^T\big]$.*

Proof That Δ_r has zero mean is immediate from its definition. Given Lemmas 7 and 8, it is sufficient to obtain the expression for $\mathbb{E}\big[\Delta_r \Delta_r^T\big]$. By Lemma 7,

$$\mathbb{E}\big[\Delta_r \Delta_r^T\big] = \tfrac{i\hbar}{m}\mathbb{E}\Big\{\big[\textstyle\int_s^r \Phi(r,\rho)\, dB_\rho\big]\big[\int_s^r \Phi(r,\rho)\, dB_\rho\big]^T\Big\},$$

where the term inside the expectation is purely real, and consequently by standard results (cf., [11]), one obtains the asserted representation. □

As noted above, we will perform the computations mainly in the simpler, steady-state case of $\bar{k}_0 = 1$. In this case, we have

$$-A_{>0} = \omega \begin{pmatrix} -i & 0 \\ 2 & 0 \end{pmatrix}, \quad \text{and} \quad -b_{>0} = \tfrac{d}{2m}\begin{pmatrix} -2i \\ 1 \end{pmatrix}. \tag{67}$$

In the case $d = 0$, we have $-b_{>0} = 0$, while in the case $d = -2m\omega\hat{\delta}$, we have $-b_{>0} = \omega\hat{\delta}(2i, -1)^T$.

Theorem 3 *In the case $\bar{k}_0 = 1$, for all $r \in [s,t]$, ζ_r is a normal random variable with mean and covariance given by, with $\hat{d} \doteq d/(m\omega)$,*

$$\mu_r = \begin{pmatrix} \mu_r^1 \\ \mu_r^2 \end{pmatrix} \quad \textit{and} \quad \tilde{\Sigma}_r \doteq \begin{pmatrix} \tilde{\Sigma}_r^{1,1} & \tilde{\Sigma}_r^{1,2} \\ \tilde{\Sigma}_r^{2,1} & \tilde{\Sigma}_r^{2,2} \end{pmatrix}, \quad \textit{where}$$

$$\mu_r^1 = e^{-i\omega(r-s)}z_1 + \hat{d}(e^{-i\omega(r-s)} - 1),$$

$$\mu_r^2 = 2i[e^{-i\omega(r-s)} - 1]z_1 + z_2 + \hat{d}\big[2i((e^{-i\omega(r-s)} - 1) - 3\omega(r-s)/2\big],$$

$$\tilde{\Sigma}_r^{1,1} = \tfrac{\hbar}{m\omega}\tfrac{1}{2}\big(1 - e^{-2i\omega(r-s)}\big),$$

$$\tilde{\Sigma}_r^{1,2} = \tilde{\Sigma}_r^{2,1} = \tfrac{\hbar}{m\omega} i\big[2\big(e^{-i\omega(r-s)} - 1\big) - \big(e^{-2i\omega(r-s)} - 1\big)\big],$$

$$\tilde{\Sigma}_r^{2,2} = \tfrac{\hbar}{m\omega}\big[2\big(e^{-2i\omega(r-s)} - 1\big) - 8\big(e^{-i\omega(r-s)} - 1\big) - 3i\omega(r-s)\big].$$

Proof The expression for μ_r is immediate from Lemma 7. To obtain the expression for the covariance, we evaluate the integral in Lemma 9. Letting $\tilde{\Sigma}_r \doteq \mathbb{E}\big[\Delta_r \Delta_r^T\big]$, component-wise, that integral is

$$\tilde{\Sigma}_r^{1,1} = \tfrac{i\hbar}{m}\int_s^r e^{-2i\omega(r-\rho)}\, d\rho, \qquad \tilde{\Sigma}_r^{1,2} = \tilde{\Sigma}_r^{2,1} = \tfrac{i\hbar}{m}\int_s^r 2i\big[e^{-2i\omega(r-\rho)} - e^{-i\omega(r-\rho)}\big]\, d\rho,$$

$$\tilde{\Sigma}_r^{2,2} = \tfrac{i\hbar}{m}\int_s^r -4\big[e^{-2i\omega(r-\rho)} - e^{-i\omega(r-\rho)}\big]^2 + 1\, d\rho.$$

Evaluating these, one obtains the asserted expression for the covariance. □

Theorem 4 *For $(s,z) \in \overline{\mathscr{D}}_{\mathbb{C}}$, $\hat{S}^1(s,z) = \sum_{l=0}^3 \sum_{j=0}^l \hat{c}_{l,j}^1(s) z_1^j z_2^{l-j}$, where the time-indexed coefficients, $\hat{c}_{l,j}^1(\cdot)$ are obtained by the evaluation of linear*

combinations of moments of up to third-order of the normal random variables ζ_r and closed-form time-integrals. For $k > 1$ and $(s,z) \in \overline{\mathscr{D}}_{\mathbb{C}}$, the $\hat{S}^k$ also take the similar forms, $\hat{S}^k(s,z) = \sum_{l=0}^{k+2} \sum_{j=0}^{l} \hat{c}^k_{l,j}(s) z_1^j z_2^{l-j}$. Given the coefficient functions $\hat{c}^\kappa_{l,j}(s)$ for $\kappa < k$, the time-indexed coefficients $\hat{c}^k_{l,j}(s)$ are obtained by the evaluation of linear combinations of moments of up to $(k+2)^{th}$-order of the normal random variables ζ_r and closed-form time-integrals.

Proof By Fubini's Theorem and Theorem 2,

$$\hat{S}^k(s,z) = \int_s^t -\tfrac{1}{2m} \sum_{\kappa=1}^{k-1} \mathbb{E}\Big\{ \big[S_z^\kappa(r,\zeta_r)\big]^T S_z^{k-\kappa}(r,\zeta_r) \Big\} \tag{68}$$
$$+ m\omega^2 \sum_{j=0}^{k+2} c^V_{k+2,j} \mathbb{E}\big\{ [\zeta_r]_1^j [\zeta_r]_2^{k+2-j} \big\}\, dr + \sum_{l=0}^{k+2} \sum_{j=0}^{l} b^\phi_{k+2,l,j} \mathbb{E}\big\{ [\zeta_t]_1^j [\zeta_t]_2^{l-j} \big\}.$$

In particular,

$$\hat{S}^1(s,z) = \int_s^t m\omega^2 \Big[\mathbb{E}\big\{ -[\zeta_r]_1^3 \big\} + \tfrac{3}{2} \mathbb{E}\big\{ [\zeta_r]_1 [\zeta_r]_2^2 \big\} \Big]\, dr + \sum_{l=0}^{3} \sum_{j=0}^{l} b^\phi_{3,l,j} \mathbb{E}\big\{ [\zeta_t]_1^j [\zeta_t]_2^{l-j} \big\}. \tag{69}$$

We see that (69) immediately yields the assertions regarding $\hat{S}^1$. If for $\kappa < k$, the $\hat{S}^\kappa(s,z)$ are polynomials in z of order at most $\kappa+2$, then the products-of-derivatives, $[S_z^\kappa(r,\zeta_r)]^T S_z^{k-\kappa}(r,\zeta_r)$, in (68) are of order at most $k+2$ in ζ_r, and the asserted form follows. □

7 The $\hat{S}^1$ Term

In Sect. 5, steady-state and periodic solutions for the zeroth-order term in the expansion were computed. Here, we proceed an additional step, computing $\breve{S}^1 \doteq \hat{S}^0 + \frac{1}{\delta} \hat{S}^1$. We perform the actual computations for $\hat{S}^1$ only in the steady-state case of $\bar{k}_0 = 1$. For $(s,z) \in \overline{\mathscr{D}}_{\mathbb{C}}$, we may obtain $\hat{S}^1(s,z)$ from (69), using the expressions for the mean and variance of normal ζ_r given in Theorem 3. We see that we must evaluate integrals of the moments $\mathbb{E}\big\{ [\zeta_r]_1^3 \big\}$ and $\mathbb{E}\big\{ [\zeta_r]_1 [\zeta_r]_2^2 \big\}$ as well as the general moments $\mathbb{E}\big\{ [\zeta_t]_1^j [\zeta_t]_2^{l-j} \big\}$ for $j \in]0, l[$, $l \in]0, 3[$. There are well-known expressions for all moments of normal random variables. In particular,

$$\mathbb{E}\big\{ [\zeta_r]_1^3 \big\} = [\mu_r]_1^3 + 3[\mu_r]_1 \tilde{\Sigma}_r^{1,1},$$
$$\mathbb{E}\big\{ [\zeta_r]_1 [\zeta_r]_2^2 \big\} = [\mu_r]_1 [\mu_r]_2^2 + [\mu_r]_1 \tilde{\Sigma}_r^{2,2} + 2[\mu_r]_2 \tilde{\Sigma}_r^{1,2}.$$

This implies that for the integral term in (69), we must evaluate the integrals of moments given by $\int_s^t [\mu_r]_1^3\,dr$, $\int_s^t [\mu_r]_1\tilde{\Sigma}_r^{1,1}\,dr$, $\int_s^t [\mu_r]_1[\mu_r]_2^2\,dr$, $\int_s^t [\mu_r]_1\tilde{\Sigma}_r^{2,2}\,dr$, and $\int_s^t [\mu_r]_2\tilde{\Sigma}_r^{1,2}\,dr$. We note that, as our interest is in the solution of the original forward-time problem, it is sufficient to take $s = 0$. Further, as our interest will be in periodic-plus-drift solutions, we take $t = \tau \doteq 2\pi/\omega$. With assiduous effort, one eventually finds

$$\begin{aligned}\mathbb{E}\Big\{\int_0^\tau -\hat{V}^1(\zeta_r)\,dr\Big\} &= \int_0^\tau m\omega^2\Big[\mathbb{E}\big\{-[\zeta_r]_1^3\big\} + \tfrac{3}{2}\mathbb{E}\big\{[\zeta_r]_1[\zeta_r]_2^2\big\}\Big]\,dr\\ &= m\omega^2\Big\{\tfrac{3\pi\hat{d}}{\omega}[z_1^2 + iz_1z_2 - z_2^2] + c_1(\tau)(1,2i)z + c_2(\tau)\Big\},\end{aligned} \tag{70}$$

where

$$\begin{aligned}c_1(\tau) &= (3\pi/\omega)\big[\hat{d}^2(1-3i\pi)/2 - \hbar/(m\omega)\big],\\ c_2(\tau) &= \tfrac{\pi\hat{d}\hbar}{m\omega^2}(18i\pi - 9/2) + \tfrac{3\pi\hat{d}^3}{2\omega}\big((1/3) - 3i\pi - 6\pi^2\big).\end{aligned}$$

From (70), we see that the expected value, $\mathbb{E}\int_0^\tau -\hat{V}^1(\zeta_r)\,dr$ has at most quadratic terms in z. (In contrast, for typical $t \neq \tau$, this integral is cubic in z.) Consequently, it may be of interest to take terminal cost, ϕ^1 to be quadratic rather than the more general hypothesized cubic form. Suppose we specifically take

$$\phi^1(z) \doteq \tfrac{1}{2}z^T Q^1 z, \tag{71}$$

where Q^1 has components $Q^1_{j,k}$. Noting that we are seeking a solution of form $\check{S}^1 = \hat{S}^0 + \frac{1}{\delta}\hat{S}^1$, we find it helpful to now allow general $d \in \mathbb{C}$ with corresponding $\bar{\Lambda}^0$ given by (48). Also, note from Theorem 3 that

$$\mu_\tau = z - \big(0, \tfrac{3\pi d}{m\omega}\big)^T, \qquad \tilde{\Sigma}_\tau^{1,1} = \tilde{\Sigma}_\tau^{1,1} = 0, \qquad \tilde{\Sigma}_\tau^{2,2} = \tfrac{-6i\pi\hbar}{m\omega}. \tag{72}$$

Combining (69) and (70)–(72), we find

$$\hat{S}^1(\tau, z) = \tfrac{1}{2}z^T(Q^1 + Q^\Delta)z + b^T z + \rho^1(\tau), \tag{73}$$

where

$$Q^\Delta = 6\pi d\begin{pmatrix}1 & i/2\\ i/2 & -1\end{pmatrix}, \qquad b = \Big[\tilde{k}_1\big(Q^1 + Q^\Delta\big) + \tilde{k}_2 Q^\Delta\Big]\begin{pmatrix}0\\1\end{pmatrix}, \tag{74}$$

$$\rho^1(\tau) = \tilde{k}_1\big[\tfrac{\tilde{k}_1}{2} + \tfrac{i\hbar}{d}\big]Q^1_{2,2} + \tfrac{\pi d\hbar}{m\omega}(18i\pi - 9/2) + \tfrac{3\pi d^3}{2m^2\omega^2}\big((1/3) - 3i\pi - 6\pi^2\big), \tag{75}$$

$$\tilde{k}_1 = \tfrac{-3\pi d}{m\omega}, \qquad \tilde{k}_2 = \tfrac{i\hbar}{d} + \tfrac{d}{2m\omega}(3\pi - i). \tag{76}$$

Recalling that $\hat{S}^0(\tau, z) = \frac{1}{2} z^T \bar{Q}^0 z + (\bar{\Lambda}^0)^T z + \rho^0(\tau)$, we find

$$\begin{aligned}\check{S}^1(\tau, z) &= \hat{S}^0(\tau, z) + \tfrac{1}{\delta}\hat{S}^1(\tau, z) \\ &= \tfrac{1}{2} z^T \big[\bar{Q}^0 + \tfrac{1}{\delta}\big(Q^1 + Q^{\Delta}\big)\big] z + \big[\bar{\Lambda}^0 + \tfrac{1}{\delta} b\big]^T z + \rho^0(\tau) + \tfrac{1}{\delta}\rho^1(\tau).\end{aligned} \tag{77}$$

In the $\check{S}^1$ solution given by (77), the Q^1 complex matrix coefficient, as well as complex d coefficient are free. Other potentially free parameters include the $\bar{k}_0$ parameter in $\hat{S}^0$ and terms that are not purely quadratic in the possibly cubic ϕ^1.

Acknowledgements Research partially supported by AFOSR Grant FA9550-15-1-0131 and NSF Grant DMS-1312569.

References

1. Azencott, R., Doss, H.: L'équation de Schrödinger quand $\hbar$ tend vers zéro: une approche probabiliste. Stoch. Aspects Classical Quant. Syst. Lect. Notes Math. **1109**, 1–17 (1985)
2. Doss, H.: Sur une resolution stochastique de l'equation de Schrödinger à coefficients analytiques. Commun. Math. Phys. **73**, 247–264 (1980)
3. Doss, H.: On a probabilistic approach to the Schrödinger equation with a time-dependent potential. J. Funct. Anal. **260**, 1824–1835 (2011)
4. Feynman, R.P.: Space-time approach to non-relativistic quantum mechanics. Rev. Mod. Phys. **20**, 367–387 (1948)
5. Fleming, W.H.: Stochastic calculus of variations and mechanics. J. Optim. Theory Appl. **41**, 55–74 (1983)
6. Fleming, W.H., Rishel, R.W.: Deterministic and Stochastic Optimal Control. Springer, New York (1982)
7. Freidlin, M.I.: Functional Integration and Partial Differential Equations. Princeton University Press, Princeton (1985)
8. Gikhman, I.I., Skorohod, A.V.: Introduction to the Theory of Random Processes. Saunders, Philadelphia (1969). (reprint, Dover, 1996)
9. Grabert, H., Hänggi, H., Talkner, P.: Is quantum mechanics equivalent to a classical stochastic process? Phys. Rev. A **19**, 2440–2445 (1979)
10. Kac, M.: On distributions of certain Wiener functionals. Trans. Am. Math. Soc. **65**, 1–13 (1949)
11. Karatzas, I., Shreve, S.E.: Brownian Motion and Stochastic Calculus. Springer, New York (1987)
12. Kloeden, P.E., Platen, E., Schurz, H.: Numerical Solution of SDE Through Computer Experiments. Springer, Berlin (1994)
13. Kolokoltsov, V.N.: Semiclassical Analysis for Diffusions and Stochastic Processes. Springer Lecture Notes in Mathematics, vol. 1724. Springer, Berlin (2000)
14. Krener, A.J.: Reciprocal diffusions in flat space. Prob. Theory Relat. Fields **107**, 243–281 (1997)
15. Litvinov, G.L.: The Maslov dequantization, idempotent and tropical mathematics: a brief introduction. J. Math. Sci. **140**, 426–444 (2007)
16. Maslov, V.P.: A new superposition principle for optimization problem. Russ. Math. Surv. [translation of Uspekhi Mat. Nauk] **42**, 39–48 (1987)

17. McEneaney, W.M.: A stationary-action control representation for the dequantized Schrödinger Equation. In: Proceedings of the 22nd International Symposium on Mathematical Theory of Networks and Systems (2016)
18. McEneaney, W.M.: A stochastic control verification theorem for the dequantized Schrödinger equation not requiring a duration restriction. Appl. Math. Optim. 1–26 (2017). https://doi.org/10.1007/s00245-017-9442-0
19. McEneaney, W.M., Dower, P.M.: Staticization, its dynamic program and solution propagation. Automatica **81**, 56–67 (2017)
20. Nagasawa, M.: Schrödinger Equations and Diffusion Theory. Birkhäuser, Basel (1993)
21. Nelson, E.: Derivation of the Schrödinger equation from Newtonian mechanics. Phys. Rev. **150**, 1079–1085 (1966)
22. Perret, C.: The Stability of Numerical Simulations of Complex Stochastic Differential Equations. ETH, Zurich (2010)
23. Range, R.M.: Holomorphic Functions and Integral Representations in Several Complex Variables. Springer, New York (1986)
24. Takagi, S.: Quantum dynamics and non-inertial frames of reference. I. Generality. Prog. Theor. Phys. **85**, 463–479 (1991)
25. Uboe, J.: Complex valued multiparameter stochastic integrals. J. Theor. Prob. **8**, 601–624 (1995)
26. Zambrini, J.C.: Probability in quantum mechanics according to E. Schrödinger. Phys. B+C **151**, 327–331 (1988)