Michael Griebel *Editor*

Singular Phenomena and Scaling in Mathematical Models

Springer

Singular Phenomena and Scaling in Mathematical Models

Michael Griebel
Editor

Singular Phenomena and Scaling in Mathematical Models

Springer

Editor
Michael Griebel
Institut für Numerische Simulation
Universität Bonn
Bonn
Germany

ISBN 978-3-319-00785-4 ISBN 978-3-319-00786-1 (eBook)
DOI 10.1007/978-3-319-00786-1
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013948877

Math. Subj. Class. (2010): 00B15, 00A69, 00A71, 00A99, 26A30, 35A20, 35B65, 60J60, 65N12, 65Z05

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

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

Preface

The success of quantitative modeling in the rapidly expanding areas of the natural sciences, such as materials science and biology, produces a variety of new mathematical models. The predictive power of these models has to be tested, and methods for their effective treatment have to be developed. This major opportunity for mathematics led to the foundation of the Collaborative Research Center SFB 611 entitled 'Singular Phenomena and Scaling in Mathematical Models' at the University of Bonn in 2002. One of its major goals was the efficient handling of new types of models through the close combination of theoretical and numerical methods.

Funded by the Deutsche Forschungsgesellschaft, we progressively integrated theoretical analysis, numerical simulation and modeling approaches for the treatment of singular phenomena in three consecutive phases until the end of 2012. Our particular projects were focused on actual applied problems, and we developed qualitatively new and mathematically challenging methods for various problems from the natural sciences.

Our Collaborative Research Center was organized in the following three divisions ranging from stochastic and geometric analysis over nonlinear analysis and modeling to numerical analysis and scientific computation:

- Part I: Scaling Limits of Diffusion Processes and Singular Spaces
- Part II: Multiple Scales in Mathematical Models of Materials Science and Biology
- Part III: Numerics for Multiscale Models and Singular Phenomena

All the three divisions addressed, in their specific way, the key aspects of the SFB 611, namely, multiple scales and model hierarchies, singularities and degeneracies and scaling laws and self-similarity. These subjects proved to be timely, challenging and at the forefront of international research. While taking on these topics, the SFB 611 acted as a bridge between analysis, modeling and numerical simulation.

A total number of 19 principal investigators and more than 45 other scientists participated in the research of the last funding period of the SFB from 2009 to 2012. This volume now comprises our final and latest contributions to the exciting field of 'Singular Phenomena and Scaling in Mathematical Models'.

At this place, we want to thank the Deutsche Forschungsgesellschaft and the University of Bonn for their ongoing support.

Bonn, Germany Michael Griebel
January 2013

Contents

Part I Scaling Limits of Diffusion Processes and Singular Spaces

Part II Multiple Scales in Mathematical Models of Materials Science and Biology

Part III Numerics for Multiscale Models and Singular Phenomena

Part I
Scaling Limits of Diffusion Processes and Singular Spaces

Ricci Bounds for Euclidean and Spherical Cones

Kathrin Bacher and Karl-Theodor Sturm

Abstract We prove generalized lower Ricci bounds for Euclidean and spherical cones over complete Riemannian manifolds. These cones are regarded as complete metric measure spaces. In general, they will be neither manifolds nor Alexandrov spaces. We show that the Euclidean cone over an n-dimensional Riemannian manifold whose Ricci curvature is bounded from below by $n-1$ satisfies the curvature-dimension condition $\mathsf{CD}(0, n+1)$ and that the spherical cone over the same manifold fulfills the curvature-dimension condition $\mathsf{CD}(n, n+1)$. More generally, for each $N > 1$ we prove that the condition $\mathsf{CD}(N-1, N)$ for a weighted Riemannian space is equivalent to the condition $\mathsf{CD}(0, N+1)$ for its N-Euclidean cone as well as to the condition $\mathsf{CD}(N, N+1)$ for its N-spherical cone.

1 Introduction

In two similar but independent approaches, the second author [14, 15] and Lott and Villani [8, 9] presented a concept of generalized lower Ricci curvature bounds for metric measure spaces $(\mathsf{M}, \mathsf{d}, \mathsf{m})$. The full strength of this concept appears if the condition $\mathsf{Ric}(\mathsf{M}, \mathsf{d}, \mathsf{m}) \geq K$ is combined with a kind of upper bound N on the dimension. This leads to the so-called curvature-dimension condition $\mathsf{CD}(K, N)$ which can be formulated in terms of optimal transportation for each pair of numbers $K \in \mathbb{R}$ and $N \in [1, \infty)$.

A complete Riemannian manifold satisfies $\mathsf{CD}(K, N)$ if and only if its Ricci curvature is bounded from below by K and its dimension from above by N.

A broad variety of geometric and functional analytic results can be deduced from the curvature-dimension condition $\mathsf{CD}(K, N)$. Among them are the

K. Bacher · K.-T. Sturm (✉)
Institut für Angewandte Mathematik, Rheinische Friedrich-Wilhelms-Universität Bonn,
Endenicher Allee 60, D-53115 Bonn, Germany
e-mail: sturm@uni-bonn.de

M. Griebel (ed.), *Singular Phenomena and Scaling in Mathematical Models*,
DOI 10.1007/978-3-319-00786-1_1,

Brunn-Minkowski inequality and the theorems by Bishop-Gromov, Bonnet-Myers and Lichnerowicz. Moreover, the condition $\mathsf{CD}(K,N)$ is stable under convergence with respect to the L^2-transportation distance d.

1.1 Statement of the Main Results

Let M be a complete n-dimensional Riemannian manifold (with Riemannian distance d and Riemannian volume $d\mathsf{m} = d\mathsf{vol}$). The *Euclidean cone* $\mathsf{Con}(\mathsf{M}) = \mathsf{M}_r\times [0,\infty)$ over M is defined as the quotient of the product $\mathsf{M} \times [0,\infty)$ obtained by identifying all points in the fiber $\mathsf{M} \times \{0\}$. This point is called the origin O of the cone. It is equipped with a metric $\mathsf{d}_{\mathsf{Con}}$ defined by the cosine formula

$$\mathsf{d}_{\mathsf{Con}}((x,s),(y,t)) = \sqrt{s^2 + t^2 - 2st\cos(\mathsf{d}(x,y) \wedge \pi)},$$

and with a measure m_n defined as the product $d\mathsf{m}_n(x,s) := d\mathsf{m}(x) \otimes s^n ds$.

Theorem 1. *The Ricci curvature of* M *is bounded from below by* $n-1$ *and there holds* $\mathsf{diam}(\mathsf{M}) \leq \pi$ *if and only if the metric measure space* $(\mathsf{Con}(\mathsf{M}), \mathsf{d}_{\mathsf{Con}}, \mathsf{m}_n)$ *satisfies the curvature-dimension condition* $\mathsf{CD}(0, n+1)$.

Note that in dimensions $n \neq 1$ the diameter bound $\mathsf{diam}(\mathsf{M}) \leq \pi$ is redundant: it follows from the Ricci bound.

The heuristic interpretation of the assertion in the theorem is that the Euclidean cone – regarded as a metric measure space – has non-negative Ricci curvature in a generalized sense. Note that already in 1982, Cheeger and Taylor [3, 6] observed that the punctured Euclidean cone $\mathsf{Con}(\mathsf{M}) \setminus \{\mathsf{O}\}$ constructed over a compact n-dimensional Riemannian manifold M with $\mathsf{Ric} \geq n-1$ is a $(n+1)$-dimensional Riemannian manifold with $\mathsf{Ric} \geq 0$. Note, however, that the sectional curvature might be unbounded from below (and above). Thus in general $\mathsf{Con}(\mathsf{M})$ will *not* be an Alexandrov space. Moreover, $\mathsf{Con}(\mathsf{M})$ in general is not a manifold and, of course, $\mathsf{Con}(\mathsf{M})\setminus\{\mathsf{O}\}$ is not complete. In particular, the Ricci curvature in the classical sense is not defined in its singularity O.

Actually, we will prove a significantly more general result:

Theorem 2. *For any real number* $N > 1$*, the* $\mathsf{CD}(N-1, N)$ *condition for a weighted Riemannian manifold is equivalent to the* $\mathsf{CD}(0, N+1)$ *condition for the associated* N*-Euclidean cone.*

It is an open question whether analogous assertions hold true with an arbitrary metric measure space $(\mathsf{M}, \mathsf{d}, \mathsf{m})$ in the place of the weighted Riemannian manifold M. A partial result towards this conjecture was derived by Ohta [11] for metric measure spaces satisfying the so-called *measure contraction property* $\mathsf{MCP}(K,N)$, a property being slightly weaker than the curvature-dimension condition $\mathsf{CD}(K,N)$.

Remark 1. If a complete separable metric measure space $(\mathsf{M}, \mathsf{d}, \mathsf{m})$ satisfies the measure contraction property $\mathsf{MCP}(N-1, N)$ for some $N \geq 1$ and if $\mathsf{diam}(\mathsf{M}) \leq \pi$

(which follows from the previous condition if $N \neq 1$) then its N-Euclidean cone $(\mathsf{Con}(\mathsf{M}), \mathsf{d}_{\mathsf{Con}}, \mathsf{m}_N)$ satisfies the measure contraction property $\mathsf{MCP}(0, N+1)$.

As a second main result we deduce a generalized lower Ricci bound for the *spherical cone* $\Sigma(\mathsf{M}) = \mathsf{M}\ {}_{\sin(r)}\!\times [0, \pi]$ over the compact Riemannian manifold M. It can be defined as the quotient of the product space $\mathsf{M} \times [0, \pi]$ obtained by contracting all points in the fiber $\mathsf{M} \times \{0\}$ to the south pole $\mathscr{S}$ and all points in the fiber $\mathsf{M} \times \{\pi\}$ to the north pole $\mathscr{N}$. It is endowed with a metric d_Σ defined via

$$\cos\left(\mathsf{d}_\Sigma(p,q)\right) = \cos s \cos t + \sin s \sin t \cos\left(\mathsf{d}(x,y) \wedge \pi\right)$$

for $p = (x,s), q = (y,t) \in \Sigma(\mathsf{M})$ and with a measure $d\hat{\mathsf{m}}_n(x,s) := d\mathsf{vol}(x) \otimes (\sin^n s ds)$.

Theorem 3. *(i) The Ricci curvature of* M *is bounded from below by* $n-1$ *and* $\mathsf{diam}(\mathsf{M}) \leq \pi$ *if and only if the metric measure space* $(\Sigma(\mathsf{M}), \mathsf{d}_\Sigma, \hat{\mathsf{m}}_n)$ *satisfies the curvature-dimension condition* $\mathsf{CD}(n, n+1)$.

(ii) A weighted Riemannian manifold satisfies the curvature-dimension condition $\mathsf{CD}(N-1, N)$ *for a given real number* $N > 1$ *if and only if the associated* N*-spherical cone satisfies the curvature-dimension condition* $\mathsf{CD}(N, N+1)$.

Note that the analogous results holds true for generalized lower bounds for the sectional curvature.

Remark 2 (see e.g. [2], Theorem 4.7.1, 10.2.3). Let (M, d) be a complete length metric space with $\mathsf{diam}(\mathsf{M}) \leq \pi$.

(i) Then (M, d) has curvature bounded from below by 1 in the sense of Alexandrov if and only if the Euclidean cone $(\mathsf{Con}(\mathsf{M}), \mathsf{d}_{\mathsf{Con}})$ has nonnegative curvature in the sense of Alexandrov.

(ii) Moreover, (M, d) has curvature bounded from below by 1 in the sense of Alexandrov if and only if the spherical cone $(\mathsf{Con}(\mathsf{M}), \mathsf{d}_{\mathsf{Con}})$ has curvature bounded from below by 1 in the sense of Alexandrov.

Note that the diameter bound is redundant if M is not one-dimensional.

Metric cones play an important role in the study of limits of Riemannian manifolds. Assume for instance that (M, d) is the Gromov-Hausdorff limit of a sequence of complete n-dimensional Riemannian manifolds whose Ricci curvature is uniformly bounded from below. Then in the non-collapsed case, every tangent cone $\mathsf{T}_x\mathsf{M}$ is a metric cone $\mathsf{Con}(\mathsf{S}_x\mathsf{M})$ with $\mathsf{diam}(\mathsf{S}_x\mathsf{M}) \leq \pi$ [4, 5]. The latter we would expect from the diameter estimate by Bonnet-Myers if $\mathsf{Ric} \geq n-2$ on $\mathsf{S}_x\mathsf{M}$ which in turn is consistent with the formal assertion '$\mathsf{Ric} \geq 0$ on $\mathsf{T}_x\mathsf{M}$'.

1.2 Basic Definitions and Notations

Throughout this paper, (M, d) always will denote a complete separable metric space (M, d) and m a locally finite measure on $(\mathsf{M}, \mathscr{B}(\mathsf{M}))$ with full support. That is, for all $x \in \mathsf{M}$ and all sufficiently small $r > 0$ the volume $\mathsf{m}(B_r(x))$ of balls centered at x is positive and finite. To avoid pathologies, we assume that M has more than one point. Such a triple $(\mathsf{M}, \mathsf{d}, \mathsf{m})$ will henceforth called *metric measure space*.

The metric space (M, d) is called a *length space* iff $\mathsf{d}(x, y) = \inf \mathsf{Length}(\gamma)$ for all $x, y \in \mathsf{M}$, where the infimum runs over all curves γ in M connecting x and y. (M, d) is called a *geodesic space* if and only if every two points $x, y \in \mathsf{M}$ are connected by a curve γ with $\mathsf{d}(x, y) = \mathsf{Length}(\gamma)$. Distance minimizing curves of constant speed are called *geodesics*. The space of all geodesics $\gamma : [0, 1] \to \mathsf{M}$ will be denoted by $\Gamma(\mathsf{M})$.

(M, d) is called *non-branching* if for every tuple (z, x_0, x_1, x_2) of points in M for which z is a midpoint of x_0 and x_1 as well as of x_0 and x_2, it follows that $x_1 = x_2$. $\mathscr{P}_2(\mathsf{M}, \mathsf{d})$ denotes the L^2-Wasserstein space of probability measures μ on $(\mathsf{M}, \mathscr{B}(\mathsf{M}))$ with finite second moments which means that $\int_{\mathsf{M}} \mathsf{d}^2(x_0, x) d\mu(x) < \infty$ for some (hence all) $x_0 \in \mathsf{M}$. The L^2-Wasserstein distance $\mathsf{d}_{\mathsf{W}}(\mu_0, \mu_1)$ between two probability measures $\mu_0, \mu_1 \in \mathscr{P}_2(\mathsf{M}, \mathsf{d})$ is defined as

$$\mathsf{d}_{\mathsf{W}}(\mu_0, \mu_1) = \inf \left\{ \left(\int_{\mathsf{M}\times\mathsf{M}} \mathsf{d}^2(x, y)\, d\mathsf{q}(x, y) \right)^{1/2} : \mathsf{q} \text{ coupling of } \mu_0 \text{ and } \mu_1 \right\}.$$

Here the infimum ranges over all *couplings* of μ_0 and μ_1, i.e. over all probability measures on $\mathsf{M}\times\mathsf{M}$ with marginals μ_0 and μ_1. Equipped with this metric, $\mathscr{P}_2(\mathsf{M}, \mathsf{d})$ is a complete separable metric space. The subspace of m-absolutely continuous measures is denoted by $\mathscr{P}_2(\mathsf{M}, \mathsf{d}, \mathsf{m})$.

Definition 1. (i) A subset $\Xi \subset \mathsf{M} \times \mathsf{M}$ is called d^2-*cyclically monotone* if and only if for any $k \in \mathbb{N}$ and for any family $(x_1, y_1), \ldots, (x_k, y_k)$ of points in Ξ the inequality

$$\sum_{i=1}^{k} \mathsf{d}^2(x_i, y_i) \le \sum_{i=1}^{k} \mathsf{d}^2(x_i, y_{i+1})$$

holds with the convention $y_{k+1} = y_1$.

(ii) Given probability measures μ_0, μ_1 on M, a probability measure q on $\mathsf{M} \times \mathsf{M}$ is called *optimal coupling* of them iff q has marginals μ_0 and μ_1 and

$$\mathsf{d}^2_{\mathsf{W}}(\mu_0, \mu_1) = \int_{\mathsf{M}\times\mathsf{M}} \mathsf{d}^2(x, y)\, d\mathsf{q}(x, y).$$

(iii) A probability measure ν on $\Gamma(\mathsf{M})$ is called *optimal path measure* (or dynamical optimal transference plan) iff the probability measure $(e_0, e_1)_*\nu$ on $\mathsf{M} \times \mathsf{M}$ is an optimal coupling of the probability measures $(e_0)_*\nu$ and $(e_1)_*\nu$ on M.

Here and in the sequel $e_t : \Gamma(\mathsf{M}) \to \mathsf{M}$ for $t \in [0, 1]$ denotes the evaluation map $\gamma \mapsto \gamma_t$. Moreover, for each measurable map $f : \mathsf{M} \to \mathsf{M}'$ and each measure μ on M the push forward (or image measure) of μ under f will be denoted by $f_*\mu$.

From [14, Lemma 2.11], [16, Theorem 5.10] we quote:

Lemma 1. *(i) For each pair $\mu_0, \mu_1 \in \mathscr{P}_2(\mathsf{M}, \mathsf{d})$ there exists an optimal coupling* q.

(ii) The support of any optimal coupling q *is a* d^2*-cyclically monotone set.*

(iii) If M *is geodesic then for each pair $\mu_0, \mu_1 \in \mathscr{P}_2(\mathsf{M}, \mathsf{d})$ there exists an optimal path measure with given initial and terminal distribution: $(e_0)_*\nu = \mu_0$ and $(e_1)_*\nu = \mu_1$.*

(iv) Given any optimal path measure ν as above, a geodesic $(\mu_t)_{t\in[0,1]}$ in $\mathscr{P}_2(\mathsf{M}, \mathsf{d})$ connecting μ_0 and μ_1 is given by

$$\mu_t := (e_t)_*\nu.$$

(v) If (M, d) *is a non-branching space, then for each pair of geodesics γ, γ' in the support of an optimal path measure we have:*

$$\gamma_{1/2} = \gamma'_{1/2} \quad \Longrightarrow \quad \gamma = \gamma'.$$

1.3 *The Curvature-Dimension Condition*

Definition 2. Given $K \in \mathbb{R}$ and $N \in [1, \infty)$, the condition $\mathsf{CD}(K, N)$ states that for each pair $\mu_0, \mu_1 \in \mathscr{P}_2(\mathsf{M}, \mathsf{d}, \mathsf{m})$ there exist an optimal coupling q of $\mu_0 = \rho_0\mathsf{m}$ and $\mu_1 = \rho_1\mathsf{m}$ and a geodesic $\mu_t = \rho_t\, m$ in $\mathscr{P}_2(\mathsf{M}, \mathsf{d}, \mathsf{m})$ connecting them such that

$$\int_{\mathsf{M}} \rho_t^{1-1/N'} d\mathsf{m} \geq \tag{1}$$

$$\int_{\mathsf{M}\times\mathsf{M}} \left[\tau_{K,N'}^{(1-t)}(\mathsf{d}(x_0, x_1))\rho_0^{-1/N'}(x_0) + \tau_{K,N'}^{(t)}(\mathsf{d}(x_0, x_1))\rho_1^{-1/N'}(x_1)\right] d\mathsf{q}(x_0, x_1)$$

for all $t \in (0, 1)$ and all $N' \geq N$.

In the case $K > 0$, the *volume distortion coefficients* $\tau_{K,N}^{(t)}(\cdot)$ for $t \in (0, 1)$ are defined by

$$\tau_{K,N}^{(t)}(\theta) = t^{1/N} \cdot \left[\frac{\sin\left(\sqrt{\frac{K}{N-1}}t\theta\right)}{\sin\left(\sqrt{\frac{K}{N-1}}\theta\right)}\right]^{1-1/N}$$

if $0 \le \theta < \sqrt{\frac{N-1}{K}}\pi$ and by $\tau_{K,N}^{(t)}(\theta) = \infty$ if $\theta \ge \sqrt{\frac{N-1}{K}}\pi$. In the case $K < 0$ an analogous definition applies with $\sin\left(\sqrt{\frac{K}{N-1}}\dots\right)$ replaced by $\sinh\left(\sqrt{\frac{-K}{N-1}}\dots\right)$. In the case $K = 0$ simply

$$\tau_{0,N}^{(t)}(\theta) = t.$$

Therefore, the condition $\mathsf{CD}(0,N)$ just asserts that for each $N' \ge N$ the *Rényi entropy*

$$\mathsf{S}_{N'}(\nu_t|\mathsf{m}) := -\int_{\mathsf{M}} \rho_t^{1-1/N'}\, d\mathsf{m}$$

is convex in $t \in [0,1]$.

Replacing the volume distortion coefficients $\tau_{K,N}^{(t)}(\cdot)$ by slightly smaller coefficients $\sigma_{K,N}^{(t)}(\cdot)$ in the definition of $\mathsf{CD}(K,N)$ leads to the reduced curvature-dimension condition $\mathsf{CD}^*(K,N)$, a condition introduced and studied in [1,7].

The definitions of the condition $\mathsf{CD}(K,N)$ in [15] and [8] slightly differ. We follow the notation of [15]. For non-branching spaces, both concepts coincide. In this case, it suffices to verify (1) for $N' = N$ since this already implies (1) for all $N' \ge N$. Even more, the condition (1) can be formulated as a pointwise inequality.

Lemma 2 ([8, 15, 16]). *A nonbranching metric measure space* $(\mathsf{M},\mathsf{d},\mathsf{m})$ *satisfies the curvature dimension condition* $\mathsf{CD}(K,N)$ *for given numbers K and N if and only if for each pair $\mu_0,\mu_1 \in \mathscr{P}_2(\mathsf{M},\mathsf{d},\mathsf{m})$ there exist an optimal path measure ν with initial and terminal distributions $(e_0)_* = \mu_0$, $(e_1)_* = \mu_1$ such that for ν-a.e. $\gamma \in \Gamma(\mathsf{M})$ and all $t \in (0,1)$*

$$\rho_t^{-1/N}(\gamma_t) \ge \tau_{K,N}^{(1-t)}(\dot\gamma)\cdot\rho_0^{-1/N}(\gamma_0) + \tau_{K,N}^{(t)}(\dot\gamma)\cdot\rho_1^{-1/N}(\gamma_1) \tag{2}$$

where $\dot\gamma := \mathsf{d}(\gamma_0,\gamma_1)$ and ρ_t denotes the Radon-Nikodym density of $(e_t)_*\nu$ with respect to m.

Lemma 3. *Assume that a metric measure space* $(\mathsf{M},\mathsf{d},\mathsf{m})$ *satisfies the curvature dimension condition* $\mathsf{CD}(N-1,N)$ *for some number $N > 1$.*

(i) Then the diameter of M *is bounded by π.*
(ii) Moreover, for every $x \in \mathsf{M}$ the set $\mathsf{M}_x := \{x' \in \mathsf{M} : \mathsf{d}(x,x') = \pi\}$ of antipodes of x consists of at most one point.

Assertion (i), the 'generalized Bonnet-Myers theorem' was proven in [15]. Assertion (ii) is due to Ohta [10, Theorem 4.5].

Now let us have closer look on the curvature-dimension condition in the case of weighted Riemannian spaces. Let be given a complete n-dimensional manifold M equipped with its Riemannian distance d and with a weighted measure

$d\mathsf{m}(x) = e^{-V(x)}d\mathsf{vol}_{\mathsf{M}}(x)$ for some function $V : \mathsf{M} \to \mathbb{R}$. Then for each real number $N > n$ the N-Ricci tensor is defined as

$$\mathsf{Ric}_x^{N,V}(v,v) := \mathsf{Ric}_x(v,v) + \left[\mathsf{Hess}\, V - \frac{1}{N-n}\nabla V \otimes \nabla V\right]_x (v,v).$$

For $N = n$ we define

$$\mathsf{Ric}_x^{N,V}(v,v) := \begin{cases} \mathsf{Ric}_x(v,v) + \mathsf{Hess} V_x(v,v), & \text{if } \nabla V(v) = 0 \\ -\infty & \text{else.} \end{cases}$$

For $1 \le N < n$ we define $\mathsf{Ric}_x^{N,V}(v,v) := -\infty$ for all $v \neq 0$.

Lemma 4 ([8, 15]). *The weighted Riemannian space* $(\mathsf{M}, \mathsf{d}, \mathsf{m})$ *satisfies the condition* $\mathsf{CD}(K,N)$ *if and only if* $\mathsf{Ric}^{N,V} \ge K$ *on* M *in the sense that*

$$\mathsf{Ric}_x^{N,V}(v,v) \ge K \cdot \|v\|_{T_x}^2$$

for all $x \in \mathsf{M}$ *and all* $v \in T_x\mathsf{M}$.

2 Euclidean Cones over Metric Measure Spaces

Definition 3 (N-Euclidean cone). For a metric measure space $(\mathsf{M}, \mathsf{d}, \mathsf{m})$ and any $N \in [1,\infty)$, the *N-Euclidean cone* $(\mathsf{Con}(\mathsf{M}), \mathsf{d}_{\mathsf{Con}}, \mathsf{m}_N)$ is a metric measure space defined as follows:

(i) $\mathsf{Con}(\mathsf{M}) := \mathsf{M} \times [0,\infty) \,/\, \mathsf{M} \times \{0\}$
(ii) For $(x,s), (x',t) \in \mathsf{M} \times [0,\infty)$

$$\mathsf{d}_{\mathsf{Con}}((x,s),(x',t)) := \sqrt{s^2 + t^2 - 2st\cos\left(\mathsf{d}(x,x') \wedge \pi\right)}$$

(iii) $d\mathsf{m}_N(x,s) := d\mathsf{m}(x) \otimes s^N ds.$

The point $\mathsf{O} := \mathsf{M} \times \{0\} \in \mathsf{Con}(\mathsf{M})$ is called origin of the cone.

The most prominent example in this setting is the unit sphere $\mathbb{S}^n \subset \mathbb{R}^{n+1}$, endowed with its intrinsic Riemannian distance and with the Riemannian volume measure on it. In other words, $\mathsf{d}(x,y)$ is the Euclidean angle between the rays from the origin $0 \in \mathbb{R}^{n+1}$ to the points x and y on the unit sphere of $\mathbb{R}^{n+1}$. Each $\xi \in \mathbb{R}^{n+1} \setminus \{0\}$ can be uniquely written as $\xi = (x,r)$ with $r \in (0,\infty)$ and $x \in \mathbb{S}^n$, namely, $r = |\xi|$ and $x = \frac{\xi}{|\xi|}$.

The definition of the metric $\mathsf{d}_{\mathsf{Con}}$ and the measure m_n ensures that the n-Euclidean cone over $\mathbb{S}^n$ is the Euclidean space $\mathbb{R}^{n+1}$ equipped with the Euclidean metric and the Lebesgue measure expressed in spherical coordinates.

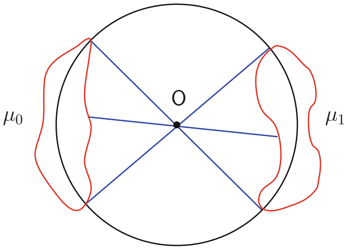


Fig. 1 Mass transport through O

Conjecture 1. A metric measure space $(\mathsf{M},\mathsf{d},\mathsf{m})$ satisfies the curvature-dimension condition $\mathsf{CD}(N-1,N)$ for some real number $N \geq 1$ and $\mathsf{diam}(\mathsf{M}) \leq \pi$ (which follows from the previous condition if $N \neq 1$) if and only if the N-Euclidean cone $(\mathsf{Con}(\mathsf{M}),\mathsf{d}_{\mathsf{Con}},\mathsf{m}_N)$ satisfies the curvature-dimension condition $\mathsf{CD}(0,N+1)$.

Conjecture 1 is true for every weighted Riemannian space. The proof is based on two ingredients:

(a) Optimal transports on the cone never transport mass through the origin – provided the base space M satisfies an appropriate CD condition.
(b) Optimal transports on the punctured cone $\mathsf{Con}_0(\mathsf{M})$ satisfy the CD condition implied by the Ricci bound for the incomplete, weighted Riemannian manifold $\mathsf{Con}_0(\mathsf{M})$. The latter in turn is equivalent to a Ricci bound for the complete weighted Riemannian manifold M.

Property (a) will be proven as a result of independent interest for general metric measure spaces (Fig. 1).

Theorem 4. *Assume that the metric measure space* $(\mathsf{M},\mathsf{d},\mathsf{m})$ *satisfies the curvature-dimension condition* $\mathsf{CD}(N-1,N)$ *for some* $N \geq 1$ *and that* $\mathsf{diam}(\mathsf{M}) \leq \pi$ *(which follows from the previous condition if* $N \neq 1$*). Let* ν *be any optimal path measure on the Euclidean cone* $(\mathsf{Con}(\mathsf{M}),\mathsf{d}_{\mathsf{Con}})$.

(i) For every $t \in (0,1)$ *there exists at most one geodesic* $\gamma \in \mathsf{supp}[\nu]$ *with* $\gamma_t = \mathsf{O}$.
(ii) For every $r > 0$ *there exists at most one* $x \in M$ *such that* $\gamma_0 = (x,r)$ *is the initial point of some geodesic* $\gamma \in \mathsf{supp}[\nu] \cap \Gamma_{\mathsf{O}}$ *where*

$$\Gamma_{\mathsf{O}} := \{\gamma \in \Gamma(\mathsf{Con}(\mathsf{M})) : \ \gamma_t = \mathsf{O} \text{ for some } t \in (0,1)\}.$$

(iii) If $(e_0)_*\nu \ll m_N$ *then* ν *gives no mass to geodesics through* O*:*

$$\nu(\Gamma_{\mathsf{O}}) = 0.$$

Proof. (i) Fix $t \in (0,1)$ and assume that two geodesics $\gamma, \gamma' \in \mathsf{supp}[\nu]$ have the origin as common t-intermediate point, i.e. $\gamma_t = \gamma'_t = \mathsf{O}$. Then $\gamma_0 = (x_0, tr)$, $\gamma_1 = (x_1,(1-t)r)$ for some $x_0, x_1 \in \mathsf{M}$ and with $r = \dot{\gamma} = \mathsf{d}_{Con}(\gamma_0,\gamma_1)$.

Similarly, $\gamma'_0 = (x'_0, tr')$, $\gamma'_1 = (x'_1, (1-t)r')$ for some $x'_0, x'_1 \in \mathsf{M}$ and with $r' = \dot{\gamma}'$. If $r > 0$ then x_0 and x_1 are antipodes of each other (i.e. $\mathsf{d}(x_0, x_1) = \pi$). Similarly, for x'_0 and x'_1. (See e.g. Lemma 6 for a detailed proof in the more sophisticated case of spherical cones.)

Cyclic monotonicity implies

$$0 \le \mathsf{d}^2_{Con}(\gamma_0, \gamma'_1) + \mathsf{d}^2_{Con}(\gamma'_0, \gamma_1) - \mathsf{d}^2_{Con}(\gamma_0, \gamma_1) - \mathsf{d}^2_{Con}(\gamma'_0, \gamma'_1).$$

On the other hand, a simple application of the triangle inequality yields

$$\begin{aligned}\mathsf{d}^2_{Con}(\gamma_0, \gamma'_1) + \mathsf{d}^2_{Con}(\gamma'_0, \gamma_1) &- \mathsf{d}^2_{Con}(\gamma_0, \gamma_1) - \mathsf{d}^2_{Con}(\gamma'_0, \gamma'_1) \\ &\le \left[tr + (1-t)r'\right]^2 + \left[tr' + (1-t)r\right]^2 - r^2 - r'^2 \\ &= -2t(1-t)(r-r')^2.\end{aligned}$$

Hence, $r = r'$.

With this at hand, a more precise calculation yields

$$\begin{aligned}0 &\le \mathsf{d}^2_{Con}(\gamma_0, \gamma'_1) + \mathsf{d}^2_{Con}(\gamma'_0, \gamma_1) - \mathsf{d}^2_{Con}(\gamma_0, \gamma_1) - \mathsf{d}^2_{Con}(\gamma'_0, \gamma'_1) \\ &= 2r^2\left[t^2 + (1-t)^2 - t(1-t)\cos \mathsf{d}(x_0, x'_1) - t(1-t)\cos \mathsf{d}(x'_0, x_1)\right] - 2r^2 \\ &= -2r^2 t(1-t)\left[2 + \cos \mathsf{d}(x_0, x'_1) + \cos \mathsf{d}(x'_0, x_1)\right].\end{aligned}$$

Thus $\mathsf{d}(x_0, x'_1) = \mathsf{d}(x'_0, x_1) = \pi$. That is, x_0 and x'_1 are antipodes (as well as x'_0 and x_1). Since antipodes in M are unique (Lemma 3(ii)) we conclude that $x_0 = x'_0$ and $x_1 = x'_1$. Thus $\gamma_0 = \gamma'_0$ and $\gamma_1 = \gamma'_1$.

In most cases of interest, geodesics are uniquely determined by their initial and terminal points. In these case, we are done. The general case, requires an additional argument. An optimal path measure ν not only induces an optimal coupling $(e_0, e_1)_*\nu$ between its initial and terminal distribution $(e_0)_*\nu$ and $(e_1)_*\nu$. More generally, the measure $(e_\sigma, e_\tau)_*\nu$ will be an optimal coupling of $(e_\sigma)_*\nu$ and $(e_\tau)_*\nu$ for each $0 \le \sigma \le \tau \le 1$. For each $\sigma \in (0, t)$ one can choose $\tau \in (t, 1)$ (and vice versa) such that $(e_t)_*\nu$ is a t-intermediate point of $(e_\sigma)_*\nu$ and $(e_\tau)_*\nu$. Hence, the previous argument will imply that $\gamma_\sigma = \gamma'_\sigma$ and $\gamma_\tau = \gamma'_\tau$. This finishes the proof.

(ii) Assume $\gamma, \gamma' \in \mathsf{supp}[\nu] \cap \Gamma_{\mathsf{O}}$ with $\gamma_0 = (x_0, r)$ and $\gamma'_0 = (x'_0, r)$. The fact that γ passes through the origin implies that $\gamma_1 = (x_1, r_1)$ with $x_1 \in \mathsf{M}$ being an antipode of x_0, i.e. $\mathsf{d}(x_0, x_1) = \pi$. Similarly, $\gamma'_1 = (x'_1, r'_1)$ with $\mathsf{d}(x'_0, x'_1) = \pi$. The radii r_1, r'_1 are arbitrary positive numbers. Cyclic monotonicity implies

$$\begin{aligned}0 &\le \mathsf{d}^2_{Con}(\gamma_0,\gamma'_1)+\mathsf{d}^2_{Con}(\gamma'_0,\gamma_1)-\mathsf{d}^2_{Con}(\gamma_0,\gamma_1)-\mathsf{d}^2_{Con}(\gamma'_0,\gamma'_1)\\ &= r^2+r_1'^2-2rr_1'\cos\mathsf{d}(x_0,x_1')+r^2+r_1^2-2rr_1\cos\mathsf{d}(x_0',x_1)\\ &\qquad -(r+r_1)^2-(r+r_1')^2\\ &= -2rr_1'\left[1+\cos\mathsf{d}(x_0,x_1')\right]-2rr_1\left[1+\cos\mathsf{d}(x_0',x_1)\right].\end{aligned}$$

Hence, $\mathsf{d}(x_0,x_1')=\pi$. That is, x_0 and x_1' are antipodes (as well as x_0' and x_1', which has been observed before). Uniqueness of antipodes in M implies $x_0=x_0'$.

(iii) Let us assume that $\nu(\Gamma_{\mathsf{O}})>0$. Then without restriction we even may assume that ν is supported by Γ_{O}. (Otherwise, replace ν by its restriction onto the set Γ_{O}.) Since $\mathsf{m}_N(\mathsf{O})=0$ we may also assume that $\gamma_0\neq\mathsf{O}$ and $\gamma_1\neq\mathsf{O}$ for ν-a.e. γ.

The previous part (ii) asserts that for each $r>0$ there exists at most one point $x_0=f(r)\in\mathsf{M}$ such that $(f(r),r)$ is the initial point γ_0 of some geodesic $\gamma\in\mathsf{supp}[\nu]\cap\Gamma_{\mathsf{O}}$. Thus the measure $\mu_0:=(e_0)_*\nu$ is concentrated on the set $C_f:=\{(f(r),r)\in\mathsf{Con}(\mathsf{M}):\ r>0\}$.

The curvature-dimension condition for the base space M implies that m has no atoms. Hence,

$$\mathsf{m}_N(C_f)=0$$

and therefore $\mu_0\not\ll\mathsf{m}_N$. □

According to the previous result, we know that – under the given curvature-dimension assumptions – optimal path measures on an Euclidean cone never will transport mass through the origin. It therefore suffices to study optimal transports on the *punctured cone*

$$\mathsf{C}_0:=\mathsf{Con}(\mathsf{M})\setminus\{\mathsf{O}\}.$$

To analyze such transports, we restrict ourselves to base spaces M which are (weighted) Riemannian manifolds. Our results crucially will rely on the fact that in this case the punctured cone C_0 is a incomplete(!) Riemannian manifold and that the Ricci curvature of it can be calculated explicitly. More precisely, the punctured n-Euclidean cone is a Riemannian manifold whereas the punctured N-Euclidean cone is a weighted Riemannian manifold.

Lemma 5. *(i) The punctured Euclidean cone C_0 is an $(n+1)$-dimensional Riemannian manifold. For $(x,r)\in\mathsf{C}_0$ with $x\in\mathsf{M}$ and $r>0$ the tangent space $T_{(x,r)}\mathsf{C}_0$ can be parametrized as $T_x\mathsf{M}\oplus\mathbb{R}$ with $\|(v,t)\|^2_{T_{(x,r)}}=r^2\,\|v\|^2_{T_x}+t^2$. Moreover, for $(v,t)\in T_{(x,r)}\mathsf{C}_0$ with $v\in T_x\mathsf{M}$ and $t\in\mathbb{R}$ we have the identity*

$$\mathsf{Ric}_{(x,r)}((v,t),\ (v,t))=\mathsf{Ric}_x(v,v)-(n-1)\|v\|^2_{T_x}.$$

In particular, $\mathsf{Ric}\ge 0$ *on* C_0 *if and only if* $\mathsf{Ric}\ge n-1$ *on* M.

(ii) *The punctured N-Euclidean cone C_0 is a weighted $(n+1)$-dimensional Riemannian manifold with measure $d\mathsf{m}_N(x,r) = r^N\,dr\,d\mathsf{vol}_\mathsf{M}(x) = e^{-W(r)}\,d\mathsf{m}_n(x,r)$ where $W(r) = -(N-n)\log r$ and $d\mathsf{m}_n(x,r) = d\mathsf{vol}_{\mathsf{C}_0}(x,r) = r^n\,dr\,d\mathsf{vol}_\mathsf{M}(x)$ denotes the Riemannian volume measure on C_0. For each $N \geq n$, the $(N+1)$-Ricci tensor satisfies*

$$\mathsf{Ric}^{N+1,W}_{(x,r)}((v,t),\,(v,t)) = \mathsf{Ric}_x(v,v) - (N-1)\|v\|^2_{T_x}.$$

In particular, $\mathsf{Ric}^{N+1,W} \geq 0$ *on* C_0 *if and only if* $\mathsf{Ric} \geq N-1$ *on* M.

(iii) *More generally, let $N \geq 1$ and let the n-dimensional Riemannian manifold M be equipped with the weighted measure $d\mathsf{m}(x) = e^{-V(x)}\,d\mathsf{vol}_\mathsf{M}(x)$ for some $V : \mathsf{M} \to \mathbb{R}$ and let the punctured cone C_0 be equipped with the measure*

$$d\mathsf{m}_N(x,r) = r^N dr\,d\mathsf{m}(x) = e^{-V(x)-W(r)}\,d\mathsf{vol}_{\mathsf{C}_0}(x,r)$$

with (as before) $W(r) = -(N-n)\log r$ and $d\mathsf{vol}_{\mathsf{C}_0}(x,r) = r^n\,dr\,d\mathsf{vol}_\mathsf{M}(x)$. Then

$$\mathsf{Ric}^{N+1,V+W}_{(x,r)}((v,t),\,(v,t)) = \mathsf{Ric}^{N,V}_x(v,v) - (N-1)\|v\|^2_{T_x}. \tag{3}$$

In particular, $\mathsf{Ric}^{N+1,V+W} \geq 0$ *on* C_0 *if and only if* $\mathsf{Ric}^{N,V} \geq N-1$ *on* M.

Proof. Assertion (i) is a classical result due to Cheeger and Taylor [3,6]. Assertion (ii) is the particular case of (iii) with $V = 0$ and $N \geq n$.

(iii): For arbitrary $V(x,r) = V(x)$ and $W(x,r) = W(r)$ depending only on the radial coordinate $r \in \mathbb{R}$ or on the basic coordinate $x \in M$, respectively, we have

$$\nabla W_{(x,r)}(v,t) = h'(0), \qquad [\mathsf{Hess}\,W]_{(x,r)}\,((v,t),\,(v,t)) = h''(0)$$

for all $(x,r) \in \mathsf{C}_0$ and all $(v,t) \in T_{(x,r)}\mathsf{C}_0$ where $h(s) := W\left(\sqrt{(r+st)^2 + s^2r^2\|v\|^2_{T_x}}\right)$. Moreover,

$$[\nabla V \otimes \nabla W]_{(x,r)}\,((v,t),\,(v,t)) = \nabla V_x(v)\cdot W'(r)\cdot t$$

for all $(x,r) \in \mathsf{C}_0$ and all $(v,t) \in T_{(x,r)}\mathsf{C}_0$ as well as

$$[\mathsf{Hess}\,V]_{(x,r)}\,((v,t),\,(v,t)) = f''(0) = [\mathsf{Hess}\,V]_x\,(v,v) - 2\nabla V_x(v)\cdot\frac{t}{r}.$$

Here the expressions on the LHS always have to be interpreted as quantities on the $(n+1)$-dimensional manifold C_0 whereas the expressions on the RHS are the original data on the basic n-dimensional manifold M and

$$f(s) = V\left(\exp_x\left(\frac{v}{\|v\|_{T_x}} \cdot \arctan \frac{rs\|v\|_{T_x}}{r+st}\right)\right).$$

For the particular choice of $W(x,r) = -(N-n)\log r$, explicit calculations yield

$$\left[\mathsf{Hess}\, W - \frac{1}{N-n}\nabla W \otimes \nabla W\right]_{(x,r)} ((v,t),\ (v,t)) = -(N-n)\,\|v\|^2_{T_x}.$$

Hence, in the case when $N > n$, together with the identity from (i)

$$\begin{aligned}
&\mathsf{Ric}^{N+1,V+W}_{(x,r)}((v,t),\ (v,t))\\
&= \mathsf{Ric}_{(x,r)}((v,t),\ (v,t))\\
&\qquad + \left[\mathsf{Hess}\,(V+W) - \frac{1}{N-n}\nabla(V+W)\otimes\nabla(V+W)\right]_{(x,r)} ((v,t),\ (v,t))\\
&= \mathsf{Ric}_x(v,v) - (n-1)\|v\|^2_{T_x} + \left[\mathsf{Hess}\, V - \frac{1}{N-n}\nabla V \otimes \nabla V\right]_x (v,v) - (N-n)\,\|v\|^2_{T_x}\\
&= \mathsf{Ric}^{N,V}_x(v,v) - (N-1)\|v\|^2_{T_x}.
\end{aligned}$$

The case $N = n$ follows from an analogous computation and our definition of N-Ricci tensor. In the case $N < n$, by definition $\mathsf{Ric}^{N,V}$ as well as $\mathsf{Ric}^{N+1,V+W}$ are $-\infty$. □

Theorem 5. *Let be given a complete n-dimensional manifold M equipped with its Riemannian distance d and with a weighted measure $d\mathsf{m}(x) = e^{-V(x)}d\mathsf{vol}_{\mathsf{M}}(x)$ for some function $V : \mathsf{M} \to \mathbb{R}$. Then for each real number $N > 1$ the following statements are equivalent:*

(i) The weighted Riemannian space $(\mathsf{M}, \mathsf{d}, \mathsf{m})$ satisfies the condition $\mathsf{CD}(N-1, N)$.

(ii) The N-Euclidean cone $(\mathsf{Con}(\mathsf{M}), \mathsf{d}_{\mathsf{Con}}, \mathsf{m}_N)$ satisfies the condition $\mathsf{CD}(0, N+1)$.

Proof. Each of the CD-conditions under consideration will imply that $\mathsf{dim}_M \leq N$. Hence, without restriction $N \geq n$.

$(ii) \Rightarrow (i)$: The condition $\mathsf{CD}(0, N+1)$ for the N-Euclidean cone $(\mathsf{Con}(\mathsf{M}), \mathsf{d}_{\mathsf{Con}}, \mathsf{m}_N)$ implies that this condition holds *locally* on the punctured cone. For this (incomplete) weighted Riemannian manifold, however, the local curvature-dimension condition $\mathsf{CD}_{loc}(0, N+1)$ is equivalent to nonnegativity of the $(N+1)$-Ricci tensor $\mathsf{Ric}^{N+1,V+W}$ on C_0, see Lemma 4. Due to the previous Lemma 5(iii), this implies $\mathsf{Ric}^{N,V}_{\mathsf{M}} \geq N-1$. For the (complete) weighted Riemannian space $(\mathsf{M}, \mathsf{d}, \mathsf{m})$, the latter in turn is equivalent to $\mathsf{CD}(N-1, N)$.

$(i) \Rightarrow (ii)$: Let probability measures μ_0 and μ_1 on $\mathsf{Con}(\mathsf{M})$ be given, absolutely continuous with respect to m_N. According to Theorem 4, any optimal path measure ν with marginal distributions $(e_0)_*\nu = \mu_0$ and $(e_1)_*\nu = \mu_1$ will give no mass to geodesics through the origin. In other words, ν-almost every geodesic will stay within the punctured cone C_0.

According to Lemma 5(iii), assertion (i) implies that the $(N+1)$-Ricci tensor $\mathsf{Ric}^{N+1,V+W}$ on the weighted Riemannian space C_0 is nonnegative. Hence, classical arguments based on Jacobi field calculus – exactly the same as used to deduce Lemma 4 – will imply that (2) holds true with $K = 0$ for ν-a.e. geodesic γ which remains within C_0. That is, $\mathsf{CD}(0, N+1)$ holds true on $\mathsf{Con}(\mathsf{M})$. □

Corollary 1. *Given a complete n-dimensional manifold* M *(equipped with its Riemannian distance* d *and its Riemannian volume* $d\mathsf{m} = d\mathsf{vol}_\mathsf{M}$*) and a real number* $N \geq 1$. *Then the following statements are equivalent:*

(i) $\mathsf{Ric} \geq N-1$ *on* M, $\mathsf{dim}_M \leq N$ *and* $\mathsf{diam}(\mathsf{M}) \leq \pi$ *(the latter follows from the Ricci and dimension bounds if* $N \neq 1$*);*
(ii) The space $(\mathsf{M}, \mathsf{d}, \mathsf{m})$ *satisfies the curvature-dimension condition* $\mathsf{CD}(N-1, N)$ *and* $\mathsf{diam}(\mathsf{M}) \leq \pi$ *(which follows from the* CD *condition if* $N \neq 1$*);*
(iii) The N-Euclidean cone $(\mathsf{Con}(\mathsf{M}), \mathsf{d}_{\mathsf{Con}}, \mathsf{m}_N)$ *satisfies* $\mathsf{CD}(0, N+1)$.

Proof. The equivalence $(i) \Leftrightarrow (ii)$ is well-known. Moreover, it is well-known that for each $N > 1$ the condition $\mathsf{CD}(N-1, N)$ implies $\mathsf{diam}(\mathsf{M}) \leq \pi$. See Lemmas 3 and 4.

In the case $N \neq 1$, the equivalence $(ii) \Leftrightarrow (iii)$ for Riemannian spaces follows from the more general assertion of Theorem 5 for *weighted* Riemannian spaces. Indeed, the arguments there also apply to the case $N = 1$. It only remains to prove that (iii) in the case $N = 1$ implies $\mathsf{diam}(\mathsf{M}) \leq \pi$.

Assume the contrary: i.e. M is a circle or an interval of $\mathsf{diam}(\mathsf{M}) > \pi$. Then there exist non-empty intervals $I, J \subset \mathsf{M}$ of length $R > 0$ such that $d(x, y) > \pi$ for all $x \in I, y \in J$. Thus for all $x \in I, y \in J$ and $r \in (0, \infty)$ the origin O will be the unique midpoint of (x, r) and (y, r) in $\mathsf{Con}(\mathsf{M})$. Moreover, for each pair $(x, s) \in I \times [1-\epsilon, 1]$ and $(y, t) \in J \times [1-\epsilon, 1]$ in $\mathsf{Con}(\mathsf{M})$ the midpoint will lie in the domain $B_\epsilon := ((I \cup J) \times (0, \epsilon]) \cup \{\mathsf{O}\}$.

Let μ_0 and μ_1 be the 'uniform distributions' on $I_\epsilon := I \times [1-\epsilon, 1]$ and let $J_\epsilon := J \times [1-\epsilon, 1]$, resp., i.e. $d\mu_0 = C_\epsilon 1_{I_\epsilon} d\mathsf{m}_N$, $d\mu_1 = C_\epsilon 1_{J_\epsilon} d\mathsf{m}_N$ with suitable $C_\epsilon \geq \frac{1}{R\epsilon}$. Then respectively their Renyi entropy satisfies

$$-S_{N+1}(\mu_0|\mathsf{m}_N) = -S_{N+1}(\mu_1|\mathsf{m}_N) = C_\epsilon{}^{-\frac{1}{N+1}} \leq (R\epsilon)^{\frac{1}{N+1}} = c\,\epsilon^{\frac{1}{N+1}}.$$

On the other hand, the midpoint $\mu_{1/2}$ of μ_0 and μ_1 is supported on B_ϵ. Hence, its Renyi entropy is bounded from below by the Renyi entropy of the uniform distribution on B_ϵ:

$$S_{N+1}(\mu_{1/2}|\mathsf{m}_N) \geq S_{N+1}(C'_\epsilon 1_{B_\epsilon}\, \mathsf{m}_N|\mathsf{m}_N) = -\,C'_\epsilon{}^{-\frac{1}{N+1}} = -c'\epsilon.$$

Note that $C'^{-1}_\epsilon = m_N(B_\epsilon) = 2R \cdot \int_0^\epsilon r^N\, dr = \frac{2R}{N+1}\epsilon^{N+1}$. Thus, choosing ϵ sufficiently small we obtain

$$S_{N+1}(\mu_{1/2}|\mathsf{m}_N) \gg \frac{1}{2}(S_{N+1}(\mu_0|\mathsf{m}_N) + S_{N+1}(\mu_1|\mathsf{m}_N))$$

which contradicts the $\mathsf{CD}(0, N+1)$ condition. □

Example 1. Let $\mathsf{M} = \left(\frac{1}{\sqrt{3}}\mathbb{S}^2\right) \times \left(\frac{1}{\sqrt{3}}\mathbb{S}^2\right)$.

(i) Then the Euclidean cone over M – more precisely, the metric measure space $(\mathsf{Con}(\mathsf{M}), \mathsf{d}_{\mathsf{Con}}, \mathsf{m}_4)$ – satisfies the curvature-dimension condition $\mathsf{CD}(0, 5)$.

(ii) On the other hand, the Euclidean cone over M – more precisely, the metric space $(\mathsf{Con}(\mathsf{M}), \mathsf{d}_{\mathsf{Con}})$ – is not an Alexandrov space: the sectional curvature on the punctured cone C_0 is unbounded from below (and above) in any punctured neighborhood of the origin $\mathsf{0}$.

Proof. (i) Given $x, y \in \frac{1}{\sqrt{3}}\mathbb{S}^2$, let u_1, u_2 be an orthonormal basis of $T_x(\frac{1}{\sqrt{3}}\mathbb{S}^2)$ and v_1, v_2 be an orthonormal basis of $T_y(\frac{1}{\sqrt{3}}\mathbb{S}^2)$. Then an orthonormal basis of $T_{(x,y)}\mathsf{M} = T_x(\frac{1}{\sqrt{3}}\mathbb{S}^2) \oplus T_y(\frac{1}{\sqrt{3}}\mathbb{S}^2)$ is given by $\{\tilde{u}_1, \tilde{u}_2, \tilde{v}_1, \tilde{v}_2\}$ with $\tilde{u}_i = (u_i, 0)$ and $\tilde{v}_i = (0, v_i)$. In this basis

$$\mathsf{Sec}_{(x,y)}(\tilde{u}_1, \tilde{u}_2) = 3, \quad \mathsf{Sec}_{(x,y)}(\tilde{u}_1, \tilde{v}_1) = 0, \quad \mathsf{Sec}_{(x,y)}(\tilde{u}_1, \tilde{v}_2) = 0$$

and analogously for any other basis vector in the place of $\tilde{u}_1$. Hence, in particular,

$$\mathsf{Ric}_{(x,y)}(\xi, \xi) = 3$$

for each $\xi \in \{\tilde{u}_1, \tilde{u}_2, \tilde{v}_1, \tilde{v}_2\}$ and thus for each $\xi \in T_{(x,y)}\mathsf{M}$.

(ii) Thus according to Theorem 5 the Euclidean cone satisfies the $\mathsf{CD}(0, 5)$ condition.

(iii) Given $r > 0$ an orthonormal basis of $T_{(x,y,r)}\mathsf{C}_0 = T_x(\frac{1}{\sqrt{3}}\mathbb{S}^2) \oplus T_y(\frac{1}{\sqrt{3}}\mathbb{S}^2) \oplus \mathbb{R}$ is given by $\{\hat{u}_1, \hat{u}_2, \hat{v}_1, \hat{v}_2, \hat{w}\}$ with $\hat{u}_i = \frac{1}{r}(u_i, 0, 0)$, $\hat{v}_i = \frac{1}{r}(0, v_i, 0)$ and $\hat{w} = (0, 0, 1)$. In this basis

$$\begin{aligned} \mathsf{Sec}_{(x,y,r)}(\hat{u}_1, \hat{u}_2) &= \frac{2}{r^2}, \quad \mathsf{Sec}_{(x,y,r)}(\hat{u}_1, \hat{v}_1) = -\frac{1}{r^2}, \\ \mathsf{Sec}_{(x,y,r)}(\hat{u}_1, \hat{v}_2) &= -\frac{1}{r^2}, \quad \mathsf{Sec}_{(x,y,r)}(\hat{u}_1, \hat{w}) = 0 \end{aligned} \tag{4}$$

and analogously for $\tilde{u}_2, \tilde{v}_1$ or $\tilde{v}_2$ in the place of $\tilde{u}_1$. Of course, this in particular implies $\mathsf{Ric}_{(x,y,r)}(\xi, \xi) = 0$ for each $\xi \in T_{(x,y,r)}\mathsf{C}_0$, see Lemma 5. □

3 Spherical Cones over Metric Measure Spaces

There are further objects with famous Euclidean ancestors – among them is the *spherical cone* or *suspension* over a topological space M. We begin with a familiar example: In order to construct the Euclidean sphere $\mathbb{S}^{n+1}$ out of its equator $\mathbb{S}^n$ we add two poles $\mathscr{S}$ and $\mathscr{N}$ and connect them via semicircles, the *meridians*, through every point in $\mathbb{S}^n$.

In the general case of abstract spaces M, we consider the product $\mathsf{M} \times [0,\pi]$ and contract each of the fibers $\mathscr{S} := \mathsf{M} \times \{0\}$ and $\mathscr{S} := \mathsf{M} \times \{\pi\}$ to a point, the *south* and the *north pole*, respectively. The resulting space is denoted by $\Sigma(\mathsf{M})$ and is called the spherical cone over M.

Definition 4 (*N*-spherical cone). The *N-spherical cone* $(\Sigma(\mathsf{M}), \mathsf{d}_\Sigma, \hat{\mathsf{m}}_N)$ over a metric measure space $(\mathsf{M}, \mathsf{d}, \mathsf{m})$ is the metric measure space defined as follows:

(i) $\Sigma(\mathsf{M}) := \mathsf{M} \times [0,\pi] \Big/ \mathsf{M} \times \{0\}, \mathsf{M} \times \{\pi\}$

(ii) For $(x,s), (x',t) \in \mathsf{M} \times [0,\pi]$

$$\cos\big(\mathsf{d}_\Sigma((x,s),(x',t))\big) := \cos s \cos t + \sin s \sin t \cos\big(\mathsf{d}(x,x') \wedge \pi\big)$$

(iii) $d\hat{\mathsf{m}}_N(x,s) := d\mathsf{m}(x) \otimes (\sin^N s ds)$.

For a nice introduction and detailed information about Euclidean and spherical cones over metric spaces we refer to [2].

Lemma 6. *Assume that* $\operatorname{diam}(\mathsf{M}) \le \pi$. *Let* $\gamma : [0,1] \to \Sigma(\mathsf{M})$ *be a non-constant geodesic with endpoints* $\gamma_0 = (x_0, r_0)$ *and* $\gamma_1 = (x_1, r_1)$ *in* $\Sigma(\mathsf{M})$. *If* $\gamma_t = \mathscr{S}$ *for some* $t \in (0,1)$, *then* x_0 *and* x_1 *are antipodes in* M.

Proof. Due to the definition of d_Σ, it holds that $r_0 = \mathsf{d}_\Sigma(\gamma_0,\gamma_t) = t\mathsf{d}_\Sigma(\gamma_0,\gamma_1)$ as well as $r_1 = \mathsf{d}_\Sigma(\gamma_t,\gamma_1) = (1-t)\mathsf{d}_\Sigma(\gamma_0,\gamma_1)$ and consequently, $r_1 = \frac{1-t}{t} r_0$. Inserting this equality into the expression for $\cos\left(\frac{r_0}{t}\right)$ we obtain

$$\cos\left(\tfrac{r_0}{t}\right) = \cos\left(\mathsf{d}_\Sigma(\gamma_0,\gamma_1)\right) = \cos r_0 \cos\left(\tfrac{1-t}{t} r_0\right) + \sin r_0 \sin\left(\tfrac{1-t}{t} r_0\right) \cos\left(\mathsf{d}(x_0,x_1)\right).$$

Since $\operatorname{diam}(\mathsf{M}) \le \pi$ by assumption, this leads to

$$\begin{aligned}
\cos(\mathsf{d}(x_0,x_1)) &= \frac{\cos\left(\frac{r_0}{t}\right) - \cos r_0 \cos\left(\frac{1-t}{t} r_0\right)}{\sin r_0 \sin\left(\frac{1-t}{t} r_0\right)} \\
&= \frac{\cos\left(\frac{r_0}{t}\right) - \frac{1}{2}\left[\cos\left(\frac{2t-1}{t} r_0\right) + \cos\left(\frac{r_0}{t}\right)\right]}{\frac{1}{2}\left[\cos\left(\frac{2t-1}{t} r_0\right) - \cos\left(\frac{r_0}{t}\right)\right]} \\
&= \frac{\frac{1}{2}\left[\cos\left(\frac{r_0}{t}\right) - \cos\left(\frac{2t-1}{t} r_0\right)\right]}{\frac{1}{2}\left[\cos\left(\frac{2t-1}{t} r_0\right) - \cos\left(\frac{r_0}{t}\right)\right]} = -1.
\end{aligned}$$

That is, $\mathsf{d}(x_0,x_1) = \pi$. □

Theorem 6. *Assume that the metric measure space* $(\mathsf{M},\mathsf{d},\mathsf{m})$ *satisfies the curvature-dimension condition* $\mathsf{CD}(N-1,N)$ *for some* $N \geq 1$ *and that* $\mathsf{diam}(\mathsf{M}) \leq \pi$ *(which follows from the previous condition if* $N \neq 1$*). Let* ν *be any optimal path measure on the spherical cone* $(\Sigma(\mathsf{M}),\mathsf{d}_\Sigma)$ *satisfying* $(e_0)_*\nu \ll m_N$*. Then* ν *gives no mass to geodesics through the poles:*

$$\nu\left(\Gamma_{\mathscr{S}}\right) = \nu\left(\Gamma_{\mathscr{N}}\right) = 0$$

where $\Gamma_{\mathscr{S}} := \{\gamma \in \Gamma(\Sigma(\mathsf{M})) : \gamma_t \in \mathscr{S} \text{ for some } t \in (0,1)\}$ *and analogously* $\Gamma_{\mathscr{N}}$ *with* $\mathscr{N}$ *in the place of* $\mathscr{S}$.

Proof. We follow the argumentation in the proof of assertion (iii) of Theorem 4. Assume that $\nu(\Gamma_{\mathscr{S}}) > 0$. Then without restriction we even may assume that $\nu(\Gamma_{\mathscr{S}}) = 1$. According to Lemma 7 below, for each $r \in (0,\pi)$ there exists at most one point $f(r) \in \mathsf{M}$ such that $(f(r),r) \in \Sigma(\mathsf{M})$ is the initial point γ_0 of some geodesic $\gamma \in \mathsf{supp}[\nu]$. Hence, $\mu_0 := (e_0)_*\nu$ is concentrated on the set $C_f := \{(f(r),r) \in \Sigma(\mathsf{M}) : r \in (0,\pi)\}$.

The curvature-dimension condition for $(\mathsf{M},\mathsf{d},\mathsf{m})$ implies that m has no atoms and thus

$$\hat{\mathsf{m}}_N(C_f) = 0$$

which contradicts the assumption $\mu_0 \ll \hat{\mathsf{m}}_N$. Hence, $\nu(\Gamma_{\mathscr{S}}) = 0$. Analogously, we deduce $\nu(\Gamma_{\mathscr{N}}) = 0$. □

Lemma 7. *Under the assumptions of the previous theorem, for every* $r \in (0,\pi)$ *there exists at most one* $x \in M$ *such that* $\gamma_0 = (x,r) \in \Sigma(\mathsf{M})$ *is the initial point of some geodesic* $\gamma \in \mathsf{supp}[\nu] \cap \Gamma_{\mathscr{S}}$.

Proof. Assume $\gamma,\gamma' \in \mathsf{supp}[\nu] \cap \Gamma_{\mathscr{S}}$ with $\gamma_0 = (x_0,r)$ and $\gamma'_0 = (x'_0,r)$. According to Lemma 6, the fact that γ passes through the south pole implies that $\gamma_1 = (x_1,r_1)$ with $x_1 \in \mathsf{M}$ being an antipode of x_0, i.e. $\mathsf{d}(x_0,x_1) = \pi$. Similarly, $\gamma'_1 = (x'_1,r'_1)$ with $\mathsf{d}(x'_0,x'_1) = \pi$. The radii r_1, r'_1 are arbitrary numbers in $(0,\pi)$.

By the very definition of d_Σ, taking into account that the diameter of M is bounded by π,

$$\begin{aligned}
&\mathsf{d}^2_\Sigma(\gamma_0,\gamma'_1) + \mathsf{d}^2_\Sigma(\gamma'_0,\gamma_1)\\
&\quad= \arccos^2\left[\cos r\cdot\cos r'_1 + \sin r\cdot\sin r'_1\cdot\cos\mathsf{d}(x_0,x'_1)\right]\\
&\qquad\qquad + \arccos^2\left[\cos r\cdot\cos r_1 + \sin r\cdot\sin r_1\cdot\cos\mathsf{d}(x'_0,x_1)\right]\\
&\quad\overset{(*)}{\leq} \arccos^2\left[\cos r\cdot\cos r'_1 - \sin r\cdot\sin r'_1\right] + \arccos^2\left[\cos r\cdot\cos r_1 - \sin r\cdot\sin r_1\right]\\
&\quad= (r+r'_1)^2 + (r+r_1)^2 = \mathsf{d}^2_\Sigma(\gamma'_0,\gamma'_1) + \mathsf{d}^2_\Sigma(\gamma_0,\gamma_1)
\end{aligned}$$

with equality in $(*)$ if and only if $\mathsf{d}(x_0, x_1') = \mathsf{d}(x_0', x_1) = \pi$, that is, if and only if x_0 and x_1' are antipodes and x_0' and x_1 are antipodes. Therefore, d_Σ^2-cyclical monotonicity implies $x_0 = x_0'$. □

From now on, let us again focus on weighted Riemannian spaces, that is, M is a complete, n-dimensional manifold equipped with its Riemannian distance d and with a measure $d\mathsf{m}(x) = e^{-V(x)} d\mathsf{vol}_M(x)$. A crucial fact for our argumentation is that the punctured cone $\Sigma_0 := \Sigma(\mathsf{M}) \setminus \{\mathscr{S}, \mathscr{N}\}$ is given as a warped product $\mathsf{M}\ {}_{\sin(r)}\!\times (0, \pi)$ for which the Ricci curvature can be calculated explicitly.

Lemma 8. *(i) The punctured spherical cone Σ_0 is an incomplete $(n+1)$-dimensional Riemannian manifold whose tangent space $T_{(x,r)}\Sigma_0$ at $(x, r) \in \Sigma_0$ with $x \in \mathsf{M}$ and $0 < r < \pi$ can be parametrized as $T_{(x,r)}\Sigma_0 = T_x\mathsf{M} \oplus \mathbb{R}$ and whose metric tensor is given by $\|(v,t)\|^2_{T_{(x,r)}} = \sin^2 r \cdot \|v\|^2_{T_x} + t^2$ for $(v, t) \in \mathsf{T}_{(x,r)}\Sigma_0$. Furthermore, we have the equality*

$$\mathsf{Ric}_{(x,r)}((v,t),(v,t)) = \mathsf{Ric}_x(v,v) + (1 - n\cos^2 r) \cdot \|v\|^2_{T_x} + n\, t^2.$$

In particular, $\mathsf{Ric} \geq n$ on Σ_0 if and only if $\mathsf{Ric} \geq n-1$ on M.

(ii) Now let us consider the punctured N-spherical cone over the weighted Riemannian manifold M. That is, given any real $N > 1$ put $W(x, r) = -(N-n)\log\sin r$ and $V(x, r) = V(x)$. Then

$$\mathsf{Ric}^{N+1,V+W}_{(x,r)}((v,t),\,(v,t)) - N\|(v,t)\|^2_{T_{x,r}} = \mathsf{Ric}^{N,V}_x(v,v) - (N-1)\|v\|^2_{T_x}. \tag{5}$$

In particular, $\mathsf{Ric}^{N+1,V+W} \geq N$ on Σ_0 if and only if $\mathsf{Ric}^{N,V} \geq N-1$ on M.

Proof. The formula for the Ricci tensor in (i) is well-known, see [12], Corollary 7.43, or e.g. [13]. Note that

$$\begin{aligned}\mathsf{Ric}_{(x,t)}((v,t),(v,t)) - \mathsf{Ric}_x(v,v) &= (1 - n\cos^2 r) \cdot \|v\|^2_{T_x} + n\, t^2 \\ &= n\, \|(v,t)\|^2_{T_{(x,r)}} - (n-1)\, \|v\|^2_{T_x}.\end{aligned}$$

The proof of assertion (ii) follows the lines of argumentation in the previous case of Euclidean cones – with appropriate modifications. For arbitrary $V(x, r) = V(x)$ and $W(x, r) = W(r)$ as above (depending only on the radial coordinate $r \in \mathbb{R}$ or on the basic coordinate $x \in M$, respectively) we have as before

$$[\nabla V \otimes \nabla W]_{(x,r)}\,((v,t),\,(v,t)) = \nabla V_x(v) \cdot W'(r) \cdot t$$

for all $(x, r) \in \Sigma_0$ and all $(v, t) \in T_{(x,r)}\Sigma_0$ and

$$\nabla W_{(x,r)}(v,t) = h'(0), \qquad [\mathsf{Hess}\, W]_{(x,r)}\,((v,t),\,(v,t)) = h''(0)$$

where now

$$h(s) := W\left(\arccos\left[\cos(r+st)\cdot\cos(s\cdot\sin r\cdot\|v\|_{T_x})\right]\right).$$

Moreover,

$$[\mathsf{Hess}\,V]_{(x,r)}\left((v,t),\,(v,t)\right) = f''(0) = [\mathsf{Hess}\,V]_x\,(v,v) - 2\nabla V_x(v)\cdot\cot(r)\cdot t$$

where

$$f(s) = V\left(\exp_x\left(\frac{v}{\|v\|_{T_x}}\cdot\arctan\frac{\tan(\sin(r)\cdot s\|v\|_{T_x})}{\sin(r+st)}\right)\right).$$

For the particular choice of $W(x,r) = -(N-n)\log\sin(r)$, some lengthy calculation yields

$$\begin{aligned}\left[\mathsf{Hess}\,W - \frac{1}{N-n}\nabla W\otimes\nabla W\right]_{(x,r)}\left((v,t),\,(v,t)\right) &= (N-n)\left[t^2-\cos^2(r)\,\|v\|^2_{T_x}\right]\\ &= (N-n)\left[\|(v,t)\|^2_{T_{(x,r)}} - \|v\|^2_{T_x}\right].\end{aligned}$$

Hence, together with the identity from (i)

$$\begin{aligned}&\mathsf{Ric}^{N+1,V+W}_{(x,r)}\left((v,t),\,(v,t)\right)\\ &= \mathsf{Ric}_{(x,r)}\left((v,t),\,(v,t)\right)\\ &\qquad+\left[\mathsf{Hess}\,(V+W) - \frac{1}{N-n}\nabla(V+W)\otimes\nabla(V+W)\right]_{(x,r)}\left((v,t),\,(v,t)\right)\\ &= \mathsf{Ric}_x(v,v) + n\,\|(v,t)\|^2_{T_{(x,r)}} - (n-1)\|v\|^2_{T_x}\\ &\qquad+\left[\mathsf{Hess}\,V - \frac{1}{N-n}\nabla V\otimes\nabla V\right]_x(v,v) + (N-n)\left[\|(v,t)\|^2_{T_{(x,r)}} - \|v\|^2_{T_x}\right]\\ &= \mathsf{Ric}^{N,V}_x(v,v) - (N-1)\|v\|^2_{T_x} + N\,\|(v,t)\|^2_{T_{(x,r)}}.\end{aligned}$$

□

Theorem 7. *Let be given a complete n-dimensional manifold* M *equipped with its Riemannian distance* d *and with a weighted measure* $d\mathsf{m}(x) = e^{-V(x)}d\mathsf{vol}_{\mathsf{M}}(x)$ *for some function* $V:\mathsf{M}\to\mathbb{R}$*. Then for each real number* $N\geq 1$ *the following statements are equivalent:*

- *(i) The weighted Riemannian space* $(\mathsf{M},\mathsf{d},\mathsf{m})$ *has* $\mathsf{diam}(\mathsf{M})\leq\pi$ *and satisfies the condition* $\mathsf{CD}(N-1,N)$*.*
- *(ii) The* N*-spherical cone* $(\Sigma(\mathsf{M}),\mathsf{d}_\Sigma,\mathsf{m}_N)$ *satisfies the condition* $\mathsf{CD}(N,N+1)$*.*

Proof. This is essentially the same argumentation as in the proof of Theorem 5, now with Lemma 8 instead of Lemma 5. Again we may assume without restriction that $N \geq n$.

$(i) \Rightarrow (ii)$: Let probability measures μ_0 and μ_1 on $\Sigma(\mathsf{M})$ be given, absolutely continuous with respect to $\hat{\mathsf{m}}_N$. According to Theorem 6, any optimal path measure ν with marginal distributions $(e_0)_*\nu = \mu_0$ and $(e_1)_*\nu = \mu_1$ will give no mass to geodesics through the poles. In other words, ν-almost every geodesic will stay within the punctured cone Σ_0.

According to Lemma 8, assertion (i) implies that the $(N+1)$-Ricci tensor $\mathsf{Ric}^{N+1,V+W}$ on the weighted Riemannian space Σ_0 is bounded from below by N. Hence, classical arguments based on Jacobi field calculus – exactly the same as used to deduce Lemma 4 – will imply that (2) holds true with $K = N$ for ν-a.e. geodesic γ which remains within Σ_0. That is, $\mathsf{CD}(N, N+1)$ holds true on $\Sigma(\mathsf{M})$.

$(ii) \Rightarrow (i)$: First the curvature-dimension condition $\mathsf{CD}(N, N+1)$ for the N-spherical cone $(\Sigma(\mathsf{M}), \mathsf{d}_\Sigma, \hat{\mathsf{m}}_N)$ implies that this condition holds *locally* on the punctured cone Σ_0. For this (incomplete) weighted Riemannian manifold, however, the local curvature-dimension condition $\mathsf{CD}_{loc}(N, N+1)$ is equivalent to the bound $\mathsf{Ric}^{N+1,V+W} \geq N$ for the $(N+1)$-Ricci tensor on Σ_0, see Lemma 4. Due to Lemma 8, this implies $\mathsf{Ric}^{N,V}_{\mathsf{M}} \geq N-1$. For the (complete) weighted Riemannian space $(\mathsf{M}, \mathsf{d}, \mathsf{m})$, the latter in turn is equivalent to $\mathsf{CD}(N-1, N)$.

Finally, in the case $N = 1$ it remains to prove that (ii) implies the diameter bound $\mathsf{diam}(\mathsf{M}) \leq \pi$. This can be achieved by means of a straightforward adaptation of the argument from the proof of Corollary 1. □

Corollary 2. *The n-spherical cone $(\Sigma(\mathsf{M}), \mathsf{d}_\Sigma, \nu)$ over a complete n-dimensional Riemannian manifold $(\mathsf{M}, \mathsf{d}, \mathsf{vol})$ satisfies $\mathsf{CD}(n, n+1)$ if and only if $\mathsf{Ric} \geq n-1$ on M and $\mathsf{diam}(\mathsf{M}) \leq \pi$.*

Theorem 7 allows to apply the Lichnerowicz theorem [8] in order to obtain a lower bound on the spectral gap of the Laplacian on the spherical cone:

Corollary 3 (Lichnerowicz estimate, Poincaré inequality). *Let $(\Sigma(\mathsf{M}), \mathsf{d}_\Sigma, \hat{\mathsf{m}}_n)$ be the n-spherical cone of a compact n-dimensional Riemannian manifold $(\mathsf{M}, \mathsf{d}, \mathsf{vol})$ with $\mathsf{Ric} \geq n-1$ and $\mathsf{diam}(\mathsf{M}) \leq \pi$. Then for every $f \in \mathsf{Lip}(\Sigma(\mathsf{M}))$ fulfilling $\int_{\Sigma(\mathsf{M})} f \, d\hat{\mathsf{m}}_n = 0$ the following inequality holds true:*

$$\int_{\Sigma(\mathsf{M})} f^2 d\hat{\mathsf{m}}_n \leq \tfrac{1}{n+1} \int_{\Sigma(\mathsf{M})} |\nabla f|^2 d\hat{\mathsf{m}}_n.$$

The Lichnerowicz estimate implies that the Laplacian Δ on the spherical cone $(\Sigma(\mathsf{M}), \mathsf{d}_\Sigma, \hat{\mathsf{m}}_n)$ defined by the identity

$$\int_{\Sigma(\mathsf{M})} f \cdot \Delta g \, d\hat{\mathsf{m}}_n = -\int_{\Sigma(\mathsf{M})} \nabla f \cdot \nabla g \, d\hat{\mathsf{m}}_n$$

admits a spectral gap λ_1 of size at least $n+1$,

$$\lambda_1 \geq n+1.$$

An analogous statement – with N in the place of n – holds true for the Laplacian on the N-spherical cone over a weighted n-dimensional Riemannian manifold satisfying $\mathsf{Ric}^{N,V} \geq N-1$.

Extension to (κ, N)*-Cones*

Let us finally mention that there is a canonical extension of the concept of cones which covers both, the Euclidean cones and the spherical cones.

Definition 5. Given a metric measure space $(\mathsf{M}, \mathsf{d}, \mathsf{m})$ and numbers $\kappa \in \mathbb{R}, N \in (0, \infty)$ we define the (κ, N)-cone over $(\mathsf{M}, \mathsf{d}, \mathsf{m})$ to be the metric measure space $(\overline{\mathsf{M}}, \overline{\mathsf{d}}, \overline{\mathsf{m}})$ with

(i) $\overline{\mathsf{M}} := \mathsf{M} \times [0, \infty)$ if $\kappa \leq 0$ and $\overline{\mathsf{M}} := \mathsf{M} \times [0, \pi/\sqrt{\kappa}]$ if $\kappa > 0$ where all the points $(x, 0)$, $x \in M$, have to be identified as well as – in the case $\kappa > 0$ – all the points $(x, \pi/\sqrt{\kappa})$.

(ii) For $(x, s), (y, t) \in \overline{\mathsf{M}}$

$$\overline{\mathsf{d}}((x,s),(y,t)) := \mathscr{C}_\kappa^{-1}\left(\mathscr{C}_\kappa(s) \cdot \mathscr{C}_\kappa(t) + \kappa \cdot \mathscr{S}_\kappa(s) \cdot \mathscr{S}_\kappa(t) \cdot \cos\left(\mathsf{d}(x,y) \wedge \pi\right)\right) \tag{6}$$

where

$$\mathscr{C}_\kappa(r) = \cos(\sqrt{\kappa}\, r), \ \mathscr{S}_\kappa(r) = \frac{1}{\sqrt{\kappa}} \sin(\sqrt{\kappa}\, r) \text{ if } \kappa > 0 \quad \text{ and}$$

$$\mathscr{C}_\kappa(r) = \cosh(\sqrt{-\kappa}\, r), \ \mathscr{S}_\kappa(r) = \frac{1}{\sqrt{-\kappa}} \sinh(\sqrt{-\kappa}\, r) \text{ if } \kappa < 0.$$

In the case $\kappa = 0$, the metric $\overline{\mathsf{d}}$ will be defined as in Definition 3. Indeed, the formula (6) leads in the limit $\kappa \to 0$ to the definition of $\mathsf{d}_{\mathsf{Con}}$.

(iii) $d\overline{\mathsf{m}}(x, s) := d\mathsf{m}(x) \otimes (\mathscr{S}_\kappa(s)^N ds)$.

The metric space $(\overline{\mathsf{M}}, \overline{\mathsf{d}})$ obtained as such a cone over a metric space (M, d) is discussed in detail in [2]. In the case $\kappa = 0$ it is simply the Euclidean cone and in the case $\kappa = 1$ it is the spherical cone. In the case $\kappa = -1$, the cone is also called *hyperbolic cone* based on (M, d). Without too much effort, our previous results extend to the general case of (κ, N)-cones over weighted Riemannian spaces $(\mathsf{M}, \mathsf{d}, \mathsf{m})$. Indeed, the case $\kappa > 0$ is just a rescaling of the case $\kappa = 1$. Replacing all sin and cos by sinh and cosh (e.g. in Lemma 8) allows to switch from the case $\kappa > 0$ to the case $\kappa < 0$.

Theorem 8. *Given a complete n-dimensional manifold* M *equipped with its Riemannian distance* d *and with a weighted measure* $d\mathsf{m}(x) = e^{-V(x)}d\mathsf{vol}_{\mathsf{M}}(x)$ *for some function* $V : \mathsf{M} \to \mathbb{R}$*. Then for all* $\kappa \in \mathbb{R}$ *and* $N \geq 1$ *the following statements are equivalent:*

(i) The weighted Riemannian space $(\mathsf{M}, \mathsf{d}, \mathsf{m})$ *has* $\mathsf{diam}(\mathsf{M}) \leq \pi$ *and satisfies the condition* $\mathsf{CD}(N-1, N)$*.*
(ii) The (κ, N)*-cone* $(\overline{\mathsf{M}}, \overline{\mathsf{d}}, \overline{\mathsf{m}})$ *satisfies the condition* $\mathsf{CD}(\kappa \cdot N, N+1)$*.*

Acknowledgements The second author would like to thank Jeff Cheeger for stimulating discussions during a visit at the Courant Institute in 2004, in particular, for posing the problem treated in Theorem 1, as well as for valuable remarks on an early draft of this paper.

References

1. Bacher, K., Sturm, K.-T.: Localization and tensorization properties of the curvature-dimension condition for metric measure spaces. J. Funct. Anal. **259**, 28–56 (2010)
2. Burago, D., Burago, Y., Ivanov, S.: A Course in Metric Geometry. Graduate Studies in Mathematics, vol. 33. American Mathematical Society, Providence (2001)
3. Cheeger, J.: Spectral geometry of singular Riemannian spaces. J. Differ. Geom. **18**, 575–657 (1983)
4. Cheeger, J., Colding, T.H.: Almost rigidity of warped products and the structure of spaces with Ricci curvature bounded below. C. R. Acad. Sci. Paris I **320**, 353–357 (1995)
5. Cheeger, J., Colding, T.H.: On the structure of spaces with Ricci curvature bounded below. I/ II/ III. J. Differ. Geom. **46**(3), 406–480 (1997), **54**(1), 13–35 (2000), **54**(1), 37–74 (2000)
6. Cheeger, J., Taylor, M.: On the diffraction of waves by conical singularities. I. Commun. Pure Appl. Math. **XXV**, 275–331 (1982)
7. Deng, Q., Sturm, K.-T.: Localization and tensorization properties of the curvature-dimension condition for metric measure spaces II. J. Funct. Anal. **260**(12), 3718–3725 (2011)
8. Lott, J., Villani, C.: Weak curvature conditions and functional inequalities. J. Funct. Anal. **245**(1), 311–333 (2007)
9. Lott, J., Villani, C.: Ricci curvature for metric measure spaces via optimal transport. Ann. Math. (2) **169**(3), 903–991 (2009)
10. Ohta, S.: On the measure contraction property of metric measure spaces. Comment. Math. Helv. **82**(4), 805–828 (2007)
11. Ohta, S.: Products, cones, and suspensions of spaces with the measure contraction property. J. Lond. Math. Soc.(2) **76**, 225–236 (2007)
12. O'Neill, B.: Semi-Riemannian Geometry: With Applications to Relativity. Pure and Applied Mathematics, vol. 103. Academic, New York (1983)
13. Petean, J.: Isoperimetric regions in spherical cones and Yamabe constants of $\mathsf{M} \times \mathbb{S}^1$. Geom. Dedicata **143**, 37–48 (2009)
14. Sturm, K.-T.: On the geometry of metric measure spaces. I. Acta Math. **196**(1), 65–131 (2006)
15. Sturm, K.-T.: On the geometry of metric measure spaces. II. Acta Math. **196**(1), 133–177 (2006)
16. Villani, C.: Optimal Transport, Old and New. Grundlehren der mathematischen Wissenschaften, vol. 338. Springer, Berlin/Heidelberg (2009)

A Monotone Approximation to the Wasserstein Diffusion

Karl-Theodor Sturm

1 Introduction and Statement of the Main Results

The Wasserstein space $\mathscr{P}(M)$ on an Euclidean or Riemannian space M – i.e. the space of probability measures on M equipped with the L^2-Wasserstein distance d_W – offers a rich geometric structure. This allows to develop a far reaching *first order calculus*, with striking applications for instance to the reformulation of conservative PDEs on M as gradient flows of suitable functionals on $\mathscr{P}(M)$, see e.g. [1, 7, 11]. A *second order calculus* was developed in [12] in the particular case of a one-dimensional state space, say $M = [0, 1]$, based on the construction of a canonical Dirichlet form

$$\mathbb{E}_{\mathscr{P}}(u, v) = \int_{\mathscr{P}} \langle Du(\mu), Dv(\mu)\rangle^2_{L^2(\mu)} d\mathbb{P}^\beta(\mu) \tag{1}$$

with domain $\mathbb{D}_{\mathscr{P}} \subset L^2(\mathscr{P}, \mathbb{P}^\beta)$. Here Du denotes the *Wasserstein gradient* and $\mathbb{P}^\beta$ a suitable measure (*"entropic measure"*). Among others, this leads to a canonical second order differential operator and to a canonical continuous Markov process $(\mu_t)_{t\geq 0}$, called *Wasserstein diffusion*.

The goal of this paper is to derive approximations of these objects – Dirichlet form, semigroup, continuous Markov process – on the infinite dimensional space $\mathscr{P} := \mathscr{P}([0, 1])$ in terms of appropriate objects on finite dimensional spaces. In particular, we will approximate the Wasserstein diffusion in terms of interacting systems of Brownian motions.

K.-T. Sturm (✉)
Institut für Angewandte Mathematik, Rheinische Friedrich-Wilhelms-Universität Bonn, Endenicher Allee 60, D-53115, Bonn, Germany
e-mail: sturm@uni-bonn.de

M. Griebel (ed.), *Singular Phenomena and Scaling in Mathematical Models*,
DOI 10.1007/978-3-319-00786-1_2,

For each $k \in \mathbb{N}$ we consider the strongly local, regular Dirichlet form $(\mathscr{E}_k, \mathscr{D}_k)$ on $L^2(\mathbb{R}^k, \rho_k^\beta\, dx)$ defined on its core $\mathscr{C}^1(\mathbb{R}^k)$ by

$$\mathscr{E}_k(U, V) = k \int_{\mathbb{R}^k} \nabla U(x) \cdot \nabla V(x)\, \rho_k^\beta(x)\, dx. \tag{2}$$

The density

$$\begin{aligned}\rho_k^\beta(x_1, \ldots, x_k) = {} & \frac{\Gamma(\beta) e^\beta \beta^k}{[k\Gamma(\beta/k)]^k} \\ & \cdot \int_{x_{k-1}}^{x_k} \cdots \int_{x_1}^{x_2} \prod_{i=1}^k \left[\int_0^{\frac{x_i - y_{i-1}}{y_i - y_{i-1}}} \left(\frac{x_i - y_{i-1}}{y_i - y_{i-1}} - z_i \right)^{\beta/k-1} \cdot z_i^{-z_i \beta/k} \right. \\ & \cdot (1 - z_i)^{-(1-z_i)\beta/k} \cdot (y_i - y_{i-1})^{\beta/k-2} \\ & \left. \cdot \left(\cos(\pi z_i \beta/k) - \frac{1}{\pi} \sin(\pi z_i \beta/k) \cdot \log \frac{z_i}{1 - z_i} \right) dz_i \right] dy_1 \ldots dy_{k-1}\end{aligned}$$

(where $y_0 := 0$, $y_k := 1$) is continuous, positive and bounded from above by

$$C \cdot [x_1(1 - x_k)]^{\beta/(2k)-1} \cdot \prod_{i=2}^k (x_i - x_{i-1})^{\beta/k-1}$$

on the simplex $\Sigma_k := \{(x_1, \ldots, x_k) : 0 < x_1 < \ldots < x_k < 1\} \subset \mathbb{R}^k$ and vanishes on $\mathbb{R}^k \setminus \Sigma_k$.

The strong Markov process $(X_t^k)_{t \geq 0} = \left(X_t^{k,1}, \ldots, X_t^{k,k} \right)_{t \geq 0}$ associated with the Dirichlet form $(\mathscr{E}_k, \mathscr{D}_k)$ is continuous, reversible and recurrent. At least on those stochastic intervals for which $X_t^k(\omega) \in \Sigma_k$, it can be characterized as the solution to an interacting system of stochastic differential equations

$$dX_t^{k,i} = k \frac{\partial \log \rho_k^\beta}{\partial x_i} \left(X_t^k \right) dt + \sqrt{2k}\, dW_t^i, \quad i = 1, \ldots, k \tag{3}$$

for some k-dimensional Brownian motion $(W_t)_{t \geq 0}$.

In many respects, an alternative representation for (1) is more convenient. The map $\chi : \; g \mapsto g_* \mathrm{Leb}|_{[0,1]}$ establishes an isometry between the set $\mathscr{G}$ of right continuous increasing functions $g : [0, 1) \to [0, 1]$ and $\mathscr{P}$. Here $\mathscr{G}$ will be regarded as a convex subset of the Hilbert space $L^2([0, 1], \mathrm{Leb})$. The image of the form (1) under the map $\chi^{-1} : \mathscr{P} \to \mathscr{G}$ is given by the form $(\mathbb{E}, \mathbb{D})$ on $L^2(\mathscr{G}, \mathbb{Q}^\beta)$ with

$$\mathbb{E}(u, v) = \int_{\mathscr{G}} \langle \mathbf{D}u(g), \mathbf{D}v(g) \rangle \, d\mathbb{Q}^\beta(g) \tag{4}$$

where $\mathbf{D}u$ denotes the Frechet derivative for "smooth" functions $u : \mathscr{G} \to \mathbb{R}$ and $\mathbb{Q}^{\beta}$ is the well-known Dirichlet-Ferguson process with parameter measure $\beta \cdot \mathrm{Leb}|_{[0,1]}$.

Theorem 1. *(i) For each $k \in \mathbb{N}$ the Dirichlet form $(\mathscr{E}_k, \mathscr{D}_k)$ on $L^2(\mathbb{R}^k, \rho_k^{\beta}\, dx)$ is isomorphic to a restriction $(\mathbb{E}, \mathbb{D}_k)$ of the Dirichlet form $(\mathbb{E}, \mathbb{D})$ on the space $L^2(L^2([0,1], Leb), \mathbb{Q}^{\beta})$. The isomorphism is induced by the embedding*

$$\iota : x \mapsto \sum_{i=1}^{k} x_i \cdot 1_{[\frac{i-1}{k}, \frac{i}{k})}$$

of $\mathbb{R}^k$ into $L^2([0,1], Leb)$ (and of Σ_k into $\mathscr{G}$).

(ii) The semigroup $\mathbb{T}_t^k$ associated with $(\mathbb{E}, \mathbb{D}_k)$ is given explicitly in terms of the semigroup T_t^k of the Dirichlet form $(\mathscr{E}_k, \mathscr{D}_k)$. If $g = \iota(x)$ for some $x \in \mathbb{R}^k$ then

$$\mathbb{T}_t^k u(g) = T_t^k U(x)$$

with $U := u \circ \iota$.

(iii) The strong Markov process $(g_t^k)_{t \geq 0}$ on $\mathscr{G}$ associated with $(\mathbb{E}, \mathbb{D}_k)$ is given by

$$g_t^k = \sum_{i=1}^{k} X_t^{k,i} \cdot 1_{[\frac{i-1}{k}, \frac{i}{k})}$$

if $g_0 = \iota(x_0)$ and if $(X_t^k)_{t \geq 0}$ denotes the Markov process on $\mathbb{R}^k$ associated with $(\mathscr{E}_k, \mathscr{D}_k)$ with initial condition $X_0^k = x_0$.

(iv) A strong Markov process $(\mu_t^k)_{t \geq 0}$ on $\mathscr{P}$ (not necessarily normal) is defined by

$$\mu_t^k(\omega) = \left(g_t^k(\omega)\right)_* Leb|_{[0,1]} = \frac{1}{k} \sum_{i=1}^{k} \delta_{X_t^{k,i}(\omega)}$$

that is, as the empirical distribution of the process $(X_t^k)_{t \geq 0}$. It is continuous, recurrent and reversible with invariant distribution $\mathbb{P}_k^{\beta} = (\iota_{\mathscr{P}})_* m_k^{\beta}$ obtained as push forward of the measure $m_k^{\beta}(dx) = \rho_k^{\beta}(x) dx$ under the embedding

$$\iota_{\mathscr{P}} : \overline{\Sigma}_k \to \mathscr{P},\ x \mapsto \frac{1}{k} \sum_{i=1}^{k} x_i.$$

Theorem 2. *(i) The domains $\mathbb{D}_{2^k}$ are increasing in $k \in \mathbb{N}$ with $\mathbb{D} = \overline{\cup_k \mathbb{D}^{2^k}}$. Therefore,*

$$(\mathbb{E}, \mathbb{D}_{2^k}) \to (\mathbb{E}, \mathbb{D}) \quad \textit{in the sense of Mosco}$$

and, hence, for the associated semigroups and resolvents

$$\mathbb{T}_t^{2^k} \to \mathbb{T}_t, \quad \mathbb{G}_\alpha^{2^k} \to \mathbb{G}_\alpha \quad \textit{strongly in } L^2(\mathscr{G}, \mathbb{Q}^\beta) \textit{ as } k \to \infty. \tag{5}$$

(ii) For the associated Markov processes on $\mathscr{P}$ starting from the respective invariant distributions we obtain convergence

$$(\mu_t^{2^k})_{t\geq 0} \to (\mu_t)_{t\geq 0} \quad \textit{as } k \to \infty \tag{6}$$

in distribution weakly on $\mathscr{C}(\mathbb{R}_+, \mathscr{P})$.

A closely related approximation result has been presented by Andres and von Renesse [2]. Their finite dimensional objects are more explicit; the convergence issues in their approximation, however, are quite delicate.

2 Dirichlet-Ferguson Process, Entropic Measure and Wasserstein Diffusion

2.1 The Dirichlet-Ferguson Process

Let $\mathscr{G}$ denote the space of all right continuous nondecreasing maps $g : [0,1] \to [0,1]$ with $g(1) = 1$. We will regard $\mathscr{G}$ as a convex subset of the Hilbert space $L^2([0,1], \text{Leb})$. The scalar product in $L^2([0,1], \text{Leb})$ will always be denoted by $\langle ., . \rangle$.

Proposition 1 ([4]). *For each real number $\beta > 0$ there exists a unique probability measure $\mathbb{Q}^\beta$ on $\mathscr{G}$, called Dirichlet-Ferguson process, with the property that for each $k \in \mathbb{N}$ and each family $0 = t_0 < t_1 < t_2 < \ldots < t_{k-1} < t_k = 1$*

$$\begin{aligned} &\mathbb{Q}^\beta \left(g_{t_1} \in dx_1, \ldots, g_{t_{k-1}} \in dx_{k-1}\right) \\ =&\frac{\Gamma(\beta)}{\prod_{i=1}^k \Gamma(\beta \cdot (t_i - t_{i-1}))} \prod_{i=1}^k (x_i - x_{i-1})^{\beta\cdot(t_i - t_{i-1})-1} dx_1 \ldots dx_{k-1} \end{aligned} \tag{7}$$

with the convention $x_0 = 0$ and $x_k = 1$.

The Dirichlet-Ferguson process can be identified with the normalized distribution of the standard Gamma process $(\gamma_t)_{t\geq 0}$: *For each $\beta > 0$, the law of the process $(\frac{\gamma_{t\cdot\beta}}{\gamma_\beta})_{t\in[0,1]}$ is the Dirichlet-Ferguson process $\mathbb{Q}^\beta$.*

Recall that a right continuous, real valued Markov process $(\gamma_t)_{t\geq 0}$ starting in zero is called standard Gamma process if its increments $\gamma_t - \gamma_s$ are independent and distributed for $0 \leq s < t$ according to $G_{t-s}(dx) = \frac{1}{\Gamma(t-s)} 1_{[0,\infty)}(x) x^{t-s-1} e^{-x} dx$.

In [12] as well as in [13] a change of variable formula (under composition) has been derived for the Dirichlet-Ferguson process.

2.2 The Dirichlet Form on $\mathscr{G}$

Let $\mathfrak{C}^1(\mathscr{G})$ denote the set of all ('cylinder') functions $u : \mathscr{G} \to \mathbb{R}$ which can be written as $u(g) = U(\langle g, \psi_1\rangle, \dots, \langle g, \psi_n\rangle)$ with $n \in \mathbb{N}$, $U \in \mathscr{C}^1(\mathbb{R}^n, \mathbb{R})$ and $\psi_1, \dots, \psi_n \in L^2([0,1], \mathrm{Leb})$. For u of this form the gradient

$$\mathbf{D}u(g) = \sum_{i=1}^{n} \partial_i U(\langle g, \psi_1\rangle, \dots, \langle g, \psi_n\rangle) \cdot \psi_i(.)$$

exists in $L^2([0,1], \mathrm{Leb})$ and

$$\|\mathbf{D}u(g)\|^2 = \int_0^1 \left| \sum_{i=1}^{n} \partial_i U(\langle g, \psi_1\rangle, \dots, \langle g, \psi_n\rangle) \cdot \psi_i(s) \right|^2 ds.$$

For $u, v \in \mathfrak{C}^1(\mathscr{G})$ we define the Dirichlet integral

$$\mathbb{E}(u, v) = \int_{\mathscr{G}} \langle \mathbf{D}u(g), \mathbf{D}v(g)\rangle \, d\mathbb{Q}^\beta(g). \tag{8}$$

Theorem 3 ([12] Theorems 7.5, 7.8, [3]).

(i) $(\mathbb{E}, \mathfrak{C}^1(\mathscr{G}))$ *is closable. Its closure* $(\mathbb{E}, \mathbb{D})$ *is a regular, strongly local, recurrent Dirichlet form on* $L^2(\mathscr{G}, \mathbb{Q}^\beta)$.

(ii) The associated Markov process $(g_t)_{t\geq 0}$ *on* $\mathscr{G}$ *is continuous, reversible and recurrent.*

(iii) The Dirichlet form $(\mathbb{E}, \mathbb{D})$ *satisfies a logarithmic Sobolev inequality with constant* $\frac{1}{\beta}$.

2.3 The Dirichlet Form on the Wasserstein Space

Let $\mathscr{P} = \mathscr{P}([0,1])$ denote the space of probability measures on the unit interval $[0,1]$. The map $\chi : \mathscr{G} \to \mathscr{P}$, $g \mapsto g_* \mathrm{Leb}|_{[0,1]}$ establishes a bijection between $\mathscr{G}$ and $\mathscr{P}$. The inverse map $\chi^{-1} : \mathscr{P} \to \mathscr{G}$, $\mu \mapsto g_\mu$ assigns to each probability measure $\mu \in \mathscr{P}$ its inverse distribution function defined by

$$g_\mu(t) := \inf\{s \in [0,1] : \mu[0,s] > t\}$$

with $\inf \emptyset := 1$. The L^2-Wasserstein distance on $\mathscr{P}$ is characterized by $d_W(\mu, \nu) = \|g_\mu - g_\nu\|_{L^2}$ for all $\mu, \nu \in \mathscr{P}$.

The *entropic measure* $\mathbb{P}^\beta$ on $\mathscr{P} = \mathscr{P}([0,1])$ is defined as the push forward of the Dirichlet process $\mathbb{Q}^\beta$ on $\mathscr{G}$ under the map χ.

Corollary 1 ([12] Theorem 7.17). *The image of the Dirichlet form defined above under the map χ is the regular, strongly local, recurrent Dirichlet form $\mathbb{E}_{\mathscr{P}}$ on $L^2(\mathscr{P}, \mathbb{P}^\beta)$, defined on its core $\mathfrak{Z}^1(\mathscr{P})$ by*

$$\mathbb{E}_{\mathscr{P}}(u, v) = \int_{\mathscr{P}} \langle Du(\mu), Dv(\mu)\rangle^2_{L^2(\mu)} d\mathbb{P}^\beta(\mu). \tag{9}$$

The associated Markov process $(\mu_t)_{t\geq 0}$ on $\mathscr{P}$, called Wasserstein diffusion, is given by

$$\mu_t^{(\omega)} = (g_t^{(\omega)})_* Leb|_{[0,1]}.$$

Here $\mathfrak{Z}^1(\mathscr{P})$ denotes the set of all functions $u : \mathscr{P} \to \mathbb{R}$ which can be written as $u(\mu) = U\left(\int_0^1 \Psi_1 d\mu, \ldots, \int_0^1 \Psi_n d\mu\right)$ with some $n \in \mathbb{N}$, some $U \in \mathscr{C}^1(\mathbb{R}^n)$ and some $\Psi_1, \ldots, \Psi_n \in \mathscr{C}^1([0, 1])$. For u as above we define its 'Wasserstein gradient' $Du(\mu) \in L^2([0, 1], \mu)$ by

$$Du(\mu) = \sum_{i=1}^n \partial_i U(\textstyle\int \Psi_1 d\mu, \ldots, \int \Psi_n d\mu) \cdot \Psi_i'(.)$$

with norm

$$\|Du(\mu)\|_{L^2(\mu)} = \left[\int_0^1 \left|\sum_{i=1}^n \partial_i U(\textstyle\int \Psi_1 d\mu, \ldots, \int \Psi_n d\mu) \cdot \Psi_i'\right|^2 d\mu\right]^{1/2}.$$

Recall that the tangent space at a given point $\mu \in \mathscr{P}$ can be identified with $L^2([0, 1], \mu)$.

The analogue to (9) on multidimensional spaces has been constructed in [10].

3 The Distribution of Random Means

Let $m_1^\beta = \zeta_* \mathbb{P}^\beta$ denote the distribution of the random variable $\zeta : \mu \mapsto \int_0^1 x\, d\mu(x)$ which assigns to each probability measure $\mu \in \mathscr{P}$ its mean value (*random means* of the *random probability measure* $\mathbb{P}^\beta$). Actually, m_1^β coincides with the distribution of the random means of the random probability measure $\mathbb{Q}^\beta$, that is, $m_1^\beta = \tilde{\zeta}_* \mathbb{Q}^\beta$ where $\tilde{\zeta} : g \mapsto \int_0^1 t\, dg(t)$ assigns to each function $g \in \mathscr{G}$ the mean value of the probability measure dg.

Indeed, integration by parts yields $\int_0^1 t\, dg(t) = \int_0^1 (1 - g(t))\, dt = \int_0^1 (1 - x)\, d\mu(x)$ for $\mu = g_*$Leb. Due to the symmetry of the entropic measure under the transformation $x \mapsto 1 - x$ the distribution of $\int_0^1 (1 - x) d\mu(x)$ coincides with m_1^β.

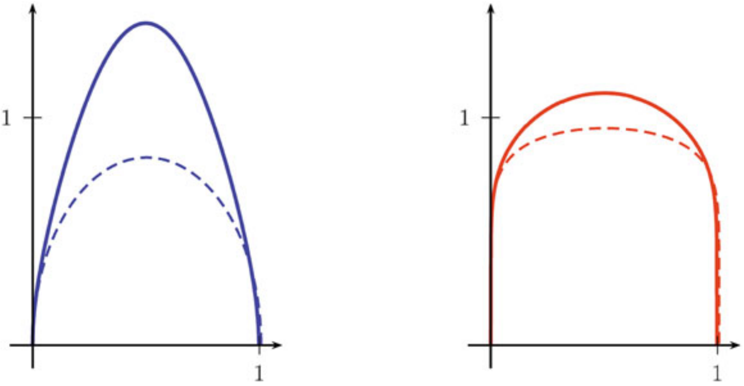


Fig. 1 Graph of $\vartheta_\beta(x)$ for $\beta = \frac{1}{2}$ (in *blue*) and $\frac{1}{8}$ (in *red*). The *dashed lines* represent the graph of $\tilde{\vartheta}(x) = [e \cdot x(1-x)]^\beta$

The law of the random means of the Dirichlet-Ferguson process is a well studied quantity, see e.g. [6]. Let Θ_β be the distribution function of m_1^β. For simplicity, we will restrict ourselves in this section to the case $\beta \in (0, 1)$. The following result can be found e.g. in [9], Propositions 8 and 3.

Lemma 1. *Θ_β admits the following representations*

$$\Theta_\beta(x) = \frac{1}{2} + \frac{1}{\pi}\int_0^\infty \exp\left(-\frac{\beta}{2}\int_0^1 \log\left[1 + t^2(x-y)^2\right] dy\right) \cdot \sin\left(\beta \int_0^1 \arctan\left[t(x-y)\right] dy\right) \frac{dt}{t}$$

and

$$\Theta_\beta(x) = \frac{e^\beta}{\pi}\int_0^x (x-y)^{\beta-1} \cdot y^{-\beta y} \cdot (1-y)^{-\beta(1-y)} \cdot \sin(\pi\beta y)\, dy.$$

Proposition 2. *The measure m_1^β is absolutely continuous with density $\vartheta_\beta = (\Theta_\beta)'$ given by (Fig. 1)*

$$\vartheta_\beta(x) = \beta e^\beta \int_0^x (x-y)^{\beta-1} \cdot y^{-\beta y} \cdot (1-y)^{-\beta(1-y)} \cdot \left[\cos(\pi\beta y) - \frac{1}{\pi}\sin(\pi\beta y) \cdot \log\frac{y}{1-y}\right] dy. \tag{10}$$

Proof. The proof requires some care since we are interested in the case $\beta < 1$. Put

$$\eta(y) = \frac{e^\beta}{\beta\pi} \cdot y^{-\beta y} \cdot (1-y)^{-\beta(1-y)} \cdot \sin(\pi\beta y)$$

in order to obtain

$$\Theta_\beta(x) = \beta \int_0^x (x-y)^{\beta-1} \cdot \eta(y)\, dy = \beta \int_0^x y^{\beta-1} \cdot \eta(x-y)\, dy.$$

Differentiating the latter yields (since $\eta(x-y) \searrow 0$ for $y \nearrow x$)

$$\vartheta_\beta(x) = \beta \int_0^x y^{\beta-1} \cdot \eta'(x-y)\, dy = \beta \int_0^x (x-y)^{\beta-1} \cdot \eta'(y)\, dy.$$

Moreover, calculating η' gives

$$\eta'(y) = e^\beta \cdot y^{-\beta y} \cdot (1-y)^{-\beta(1-y)} \cdot \left[\cos(\pi\beta y) - \frac{1}{\pi}\sin(\pi\beta y) \cdot \log\frac{y}{1-y}\right].$$

This proves the claim. □

Proposition 3. *The density* $\vartheta : [0,1] \to \mathbb{R}$ *has the following properties*

(i) ϑ *is symmetric, i.e.* $\vartheta(x) = \vartheta(1-x)$*;*
(ii) ϑ *is continuous on* $[0,1]$ *and* $\mathscr{C}^\infty$ *on* $(0,1)$*;*
(iii) $\vartheta > 0$ *on* $(0,1)$ *and* $\vartheta(0) = \vartheta(1) = 0$*;*
(iv) $\vartheta(x)/\tilde{\vartheta}(x) \to 1$ *as* $x \to 0$ *or* $x \to 1$ *for* $\tilde{\vartheta}(x) := [e \cdot x(1-x)]^\beta$*;*
(v) $\exists\, C \geq c > 0$*, e.g.* $c = \cos(\pi\beta/2)$ *and* $C = 4^\beta[1+\beta/e]$*, s.t. for all* $x \in [0,1]$

$$c\tilde{\vartheta}(x) \leq \vartheta(x) \leq C\tilde{\vartheta}(x). \tag{11}$$

Proof. (i) Is proven in [9], Proposition 6. It also follows immediately from formula (12).

(ii) The smoothness inside $(0,1)$ follows from the representation formula in the previous Proposition. Continuity at the boundary is a consequence of the estimates in (iv).

(iii) Is a consequence of (v).

(iv) Using the notations from the proof of the previous Proposition and the fact that $\eta'(y) \to e^\beta$ as $y \to 0$ we obtain

$$\frac{\vartheta(x)}{(e\cdot x)^\beta} = \frac{\beta}{(e\cdot x)^\beta}\int_0^x (x-y)^{\beta-1} \cdot \eta'(y)\, dy \quad\to\quad \frac{\beta}{x^\beta}\int_0^x (x-y)^{\beta-1}\, dy = 1$$

as $x \to 0$. Combined with the symmetry (i) this proves the claim.

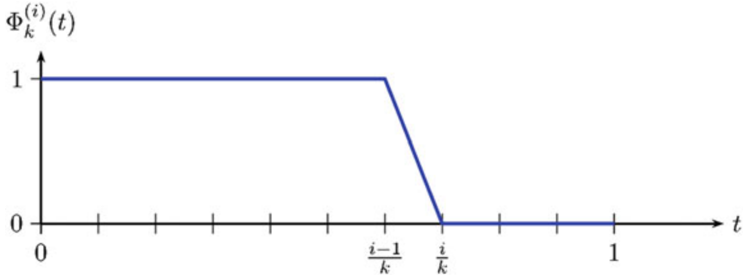


Fig. 2 Graph of $\Phi_k^{(i)}(t)$

(v) A lower estimate of the form

$$\vartheta(x) \geq (e \cdot x)^{\beta} \cdot \cos(\pi\beta/2)$$

for $x \leq 1/2$ follows from the estimate $\eta'(y) \geq e^{\beta} \cdot \cos(\pi\beta/2)$, which is valid for all $y \leq 1/2$. On the other hand, the estimate

$$\eta'(y) \leq (2e)^{\beta} \cdot \left[\cos(\pi\beta y) - \frac{1}{\pi}\sin(\pi\beta y) \cdot \log\frac{y}{1-y}\right] \leq (2e)^{\beta} \cdot \left[1 + \frac{\beta}{e}\right],$$

again valid for $y \leq 1/2$, implies

$$\vartheta(x) \leq (2ex)^{\beta} \cdot \left[1 + \frac{\beta}{e}\right]$$

for all $x \leq 1/2$. Due to the symmetry of ϑ this proves the claim. □

Remark 1. For all $x \in (0,1)$

(i) $\Theta_{\beta}(x) \to x$ and $\vartheta_{\beta}(x) \to 1$ as $\beta \to 0$
(ii) $\Theta_{\beta}(x) \to \frac{1}{2} \cdot 1_{\{\frac{1}{2}\}}(x) + 1_{(\frac{1}{2},1]}(x)$ as $\beta \to \infty$.

4 The Measure m_k^{β} in the Multivariate Case

From a technical point of view, the main result of this paper is the identification of the distribution of the random vector

$$\hat{\mathscr{I}}_k(g) = \left(\int_0^1 \Phi_k^{(1)} dg, \ldots, \int_0^1 \Phi_k^{(k)} dg\right) \tag{12}$$

under $\mathbb{Q}^{\beta}$ where (Fig. 2)

$$\Phi_k^{(i)}(t) := \begin{cases} 1, & \text{for } t \in [0, \frac{i-1}{k}] \\ i - kt, & \text{for } t \in [\frac{i-1}{k}, \frac{i}{k}] \\ 0, & \text{for } t \in [\frac{i}{k}, 1]. \end{cases} \tag{13}$$

Note that integration by parts yields

$$\int_0^1 \Phi_k^{(i)}(t) dg(t) = k \int_{\frac{i-1}{k}}^{\frac{i}{k}} g(t) dt$$

for all $i = 1, \ldots, k$ and all $g \in \mathcal{G}$. Put

$$m_k^\beta := \left(\hat{\mathcal{J}}_k\right)_* \mathbb{Q}^\beta.$$

Theorem 4. *For any $\beta > 0$ and $k \in \mathbb{N}$, $k \geq \beta$, the measure m_k^β on $\mathbb{R}^k$ is absolutely continuous. The density is strictly positive and continuous on the simplex*

$$\Sigma_k := \{(x_1, \ldots, x_k) : 0 < x_1 < \ldots < x_k < 1\} \subset \mathbb{R}^k$$

and vanishes on $\mathbb{R}^k \setminus \Sigma_k$. For $x \in \Sigma_k$ it is given by

$$\begin{aligned} &\rho_k^\beta(x_1, \ldots, x_k) \qquad (14) \\ =&\frac{\Gamma(\beta)}{\Gamma(\beta/k)^k} \int_{x_{k-1}}^{x_k} \cdots \int_{x_1}^{x_2} \prod_{i=1}^k \left[\vartheta_{\beta/k}\left(\frac{x_i - y_{i-1}}{y_i - y_{i-1}}\right) \cdot (y_i - y_{i-1})^{\beta/k-2}\right] dy_1 \ldots dy_{k-1} \end{aligned}$$

(where $y_0 := 0$, $y_k := 1$) with ϑ_β as defined in (10).

Proof. Let us start with the simple observation that

$$\int_0^1 \Phi_k^{(i)} dg = g\left(\frac{i-1}{k}\right) + \left[g\left(\frac{i}{k}\right) - g\left(\frac{i-1}{k}\right)\right] \cdot \int_0^1 (1-t) d\tilde{g}_i(t)$$

with

$$\tilde{g}_i(t) := \frac{g\left(\frac{t+i-1}{k}\right) - g\left(\frac{i-1}{k}\right)}{g\left(\frac{i}{k}\right) - g\left(\frac{i-1}{k}\right)}.$$

Now the crucial fact is that, conditioned on $\left(g\left(\frac{1}{k}\right), \ldots, g\left(\frac{k-1}{k}\right)\right)$, the processes $(\tilde{g}_i(t))_{t \in [0,1]}$ for $i = 1, \ldots, k$ are independent and distributed according to $\mathbb{Q}^{\beta/k}$. (This can be deduced from the explicit representation formula for the finite dimensional distributions (7), see also [12], Proposition 3.15) (Fig. 3).

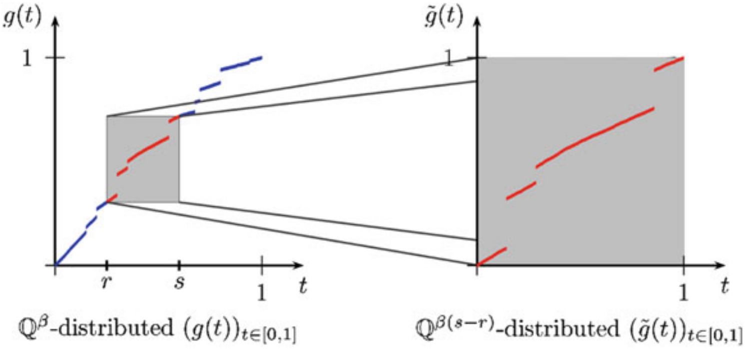

$g(t)$
1
r
s
1
t
$\tilde{g}(t)$
1
1
t
$\mathbb{Q}^{\beta}$-distributed $(g(t))_{t\in[0,1]}$
$\mathbb{Q}^{\beta(s-r)}$-distributed $(\tilde{g}(t))_{t\in[0,1]}$

Fig. 3 Picture of the transformation $\tilde{g}(t) = \dfrac{g(r+t(s-r))-g(r)}{g(s)-g(r)}$

Moreover, according to Proposition 2 the distribution of $\int_0^1(1-t)d\tilde{g}_i(t)$ for $\mathbb{Q}^{\beta/k}$-distributed $(\tilde{g}_i(t))_{t\in[0,1]}$ is given by $dm_1^{\beta/k}(x) = \vartheta_{\beta/k}(x)\,dx$.

Finally, the distribution of the random vector $\left(g\left(\frac{1}{k}\right),\ldots,g\left(\frac{k-1}{k}\right)\right)$ is given explicitly by the Dirichlet distribution, see formula (7).

Putting these informations together we obtain for each bounded Borel function U on $\mathbb{R}^k$

$$\int_{\mathscr{G}} U\left(\left(\int_0^1 \Phi_k^{(i)}dg\right)_{i=1,\ldots,k}\right)d\mathbb{Q}^{\beta}$$

$$= \int_{\mathscr{G}} U\left(\left(g\left(\frac{i-1}{k}\right)+\left[g\left(\frac{i}{k}\right)-g\left(\frac{i-1}{k}\right)\right]\cdot\int_0^1(1-t)d\tilde{g}_i(t)\right)_{i=1,\ldots,k}\right)d\mathbb{Q}^{\beta}$$

$$= \frac{\Gamma(\beta)}{\Gamma(\beta/k)^k}\int_{\Sigma_{k-1}}\left[\int_{\mathscr{G}}\cdots\int_{\mathscr{G}} U\left(\left(y_{i-1}+[y_i-y_{i-1}]\cdot\int_0^1(1-t)d\tilde{g}_i(t)\right)_{i=1,\ldots,k}\right)\right.$$

$$\left. d\mathbb{Q}^{\beta/k}(\tilde{g}_1)\ldots d\mathbb{Q}^{\beta/k}(\tilde{g}_k)\right]\prod_{i=1}^k(y_i-y_{i-1})^{\beta/k-1}\,dy_1\ldots dy_{k-1}$$

$$= \frac{\Gamma(\beta)}{\Gamma(\beta/k)^k}\int_{\Sigma_{k-1}}\left[\int_0^1\cdots\int_0^1 U\left((y_{i-1}+[y_i-y_{i-1}]\cdot z_i)_{i=1,\ldots,k}\right)\right.$$

$$\left.\prod_{i=1}^k\vartheta_{\beta/k}(z_i)\,dz_1\ldots dz_k\right]\prod_{i=1}^k(y_i-y_{i-1})^{\beta/k-1}\,dy_1\ldots dy_{k-1}$$

$$= \frac{\Gamma(\beta)}{\Gamma(\beta/k)^k}\int_{\Sigma_{k-1}}\left[\int_{y_{k-1}}^{y_k}\cdots\int_{y_0}^{y_1} U\left((x_i)_{i=1,\ldots,k}\right)\right.$$

$$\prod_{i=1}^{k}\left[\vartheta_{\beta/k}\left(\frac{x_i-y_{i-1}}{y_i-y_{i-1}}\right)\cdot(y_i-y_{i-1})^{\beta/k-2}\right]dx_1,\dots dx_k\Bigg]\,dy_1\dots dy_{k-1}$$

$$=\frac{\Gamma(\beta)}{\Gamma(\beta/k)^k}\int_{\Sigma_k}\left[\int_{x_{k-1}}^{x_k}\dots\int_{x_1}^{x_2}U\left((x_i)_{i=1,\dots,k}\right)\right.$$

$$\prod_{i=1}^{k}\left[\vartheta_{\beta/k}\left(\frac{x_i-y_{i-1}}{y_i-y_{i-1}}\right)\cdot(y_i-y_{i-1})^{\beta/k-2}\right]dy_1,\dots dy_{k-1}\Bigg]\,dx_1\dots dx_k$$

$$=\int_{\Sigma_k}U\left(x_1,\dots,x_k\right)\cdot\rho_k^\beta\left(x_1,\dots,x_k\right)\,dx_1\dots dx_k$$

with ρ_k^β as defined above (and always with $y_0:=0, y_k:=1$).

The continuity and strict positivity of ρ_k^β on Σ_k follows from the explicit representation formula and from the fact that $\vartheta_{\beta/k}$ is smooth and >0 on $(0,1)$. □

Remark 2. The densities ρ_k^β have the following hierarchical structure:

$$\rho_k^\beta(x_1,x_2,\dots,x_k)=2^k\int_{\mathbb{R}^k}\rho_{2k}^\beta(x_1-\xi_1,x_1+\xi_1,\dots,x_k-\xi_k,x_k+\xi_k)d\xi_1\dots d\xi_k. \tag{15}$$

This is of course a consequence of the fact that they are obtained via projection from the same measure $\mathbb{Q}^\beta$ and that

$$\Phi_k^{(i)}=\frac{1}{2}\left(\Phi_{2k}^{(2i-1)}+\Phi_{2k}^{(2i)}\right)$$

for all $k\in\mathbb{N}$ and all $i=1,\dots,k$. Thus for all U on $\mathbb{R}^k$

$$\int_{\mathbb{R}^k}U\ (x)\rho_k^\beta(x)dx=\int_{\mathbb{R}^{2k}}U\left(\frac{y_1+y_2}{2},\dots,\frac{y_{2k-1}+y_{2k}}{2}\right)\rho_{2k}^\beta(y)dy$$
$$=\int_{\mathbb{R}^k}U(x)\left[2^k\int_{\mathbb{R}^k}\rho_{2k}^\beta(x_1-\xi_1,x_1+\xi_1,\dots,x_k-\xi_k,x_k+\xi_k)d\xi_1\dots d\xi_k\right]dx.$$

Proposition 4. *(i) There exists a constant $C=C_{\beta,k}$ such that for all $x\in\Sigma_k$:*

$$\rho_k^\beta(x_1,\dots,x_k)\le C\cdot[x_1(1-x_k)]^{\beta/(2k)-1}\cdot\prod_{i=2}^{k}(x_i-x_{i-1})^{\beta/k-1}. \tag{16}$$

(ii) For all $l\in\{1,\dots,k-1\}$ there exist continuous functions $\gamma_1>0$ on Σ_l and $\gamma_2>0$ on Σ_{k-l} such that

$$\rho_k^\beta(x) \geq \gamma_1(x_1,\dots,x_l)\cdot\gamma_2(x_{l+1},\dots,x_k)\cdot(x_{l+1}-x_l)^{2\beta/k-1} \tag{17}$$

for all $x \in \Sigma_k$ *with* $|x_{l+1}-x_l| \leq \frac{1}{4}\min\{|x_l-x_{l-1}|,|x_{l+2}-x_{l+1}|\}$.

Proof. (i) Using the fact that $\vartheta_{\beta/k} \leq C$ and the trivial estimate

$$(a+b)^{-p} \leq 2^{-p}\cdot a^{-p/2}\cdot b^{-p/2} (\forall a,b,p>0)$$

we obtain

$$\begin{aligned}
&\rho_k^\beta(x_1,\dots,x_k)\\
&\leq C^k\cdot\frac{\Gamma(\beta)}{\Gamma(\beta/k)^k}\int_{x_{k-1}}^{x_k}\dots\int_{x_1}^{x_2}\prod_{i=1}^k(y_i-y_{i-1})^{\beta/k-2}\,dy_1\dots dy_{k-1}\\
&\leq C^k\cdot\frac{\Gamma(\beta)}{\Gamma(\beta/k)^k}\cdot 2^{\beta-2k}\int_{x_{k-1}}^{x_k}\dots\int_{x_1}^{x_2}\prod_{i=1}^k(y_i-x_i)^{\beta/(2k)-1}\\
&\qquad\qquad\cdot(x_i-y_{i-1})^{\beta/(2k)-1}\,dy_1\dots dy_{k-1}\\
&= C^k\cdot\frac{\Gamma(\beta)}{\Gamma(\beta/k)^k}\left[\frac{\Gamma(\beta/(2k))^2}{\Gamma(\beta/k)}\right]^{k-1}\cdot 2^{\beta-2k}\cdot[x_1(1-x_k)]^{\beta/(2k)-1}\\
&\qquad\qquad\cdot\prod_{i=2}^k(x_i-x_{i-1})^{\beta/k-1}.
\end{aligned}$$

(ii) We assume $k>2\beta$ and $2\leq l\leq k-2$. (The cases $l=1$ and $l=k-1$ require some modifications.) Fix $x\in\Sigma_k$ as above and put $\delta := |x_{l+1}-x_l|$. In the representation formula (4.3) for ρ_k^β, restrict the interval of integration for dy_{l-1} from $[x_{l-1},x_l]$ to $[x_l-2\delta,x_l-\delta]$ and that for dy_{l+1} from $[x_{l+1},x_{l+2}]$ to $[x_{l+1}+\delta,x_{l+1}+2\delta]$. Moreover, use the lower estimate (11) for the $\vartheta_{\beta/k}\left(\frac{x_i-y_{i-1}}{y_i-y_{i-1}}\right)$ for $i\in\{l,l+1\}$ to obtain the estimate

$$\begin{aligned}
&\rho_k^\beta(x_1,\dots,x_k)\\
&\geq C\cdot\int_{x_1}^{x_2}\dots\int_{x_{l-2}}^{x_{l-1}}\int_{x_l-2\delta}^{x_l-\delta}\int_{x_l}^{x_{l+1}}\int_{x_{l+1}+\delta}^{x_{l+1}+2\delta}\int_{x_{l+2}}^{x_{l+3}}\dots\int_{x_{k-1}}^{x_k}\\
&\qquad\prod_{i\in\{1,\dots,l-1\}\cup\{l+2,\dots,k\}}\left[\vartheta_{\beta/k}\left(\frac{x_i-y_{i-1}}{y_i-y_{i-1}}\right)\cdot(y_i-y_{i-1})^{\beta/k-2}\right]\cdot\\
&\qquad\cdot(x_l-y_{l-1})^{\beta/k}\cdot(y_l-x_l)^{\beta/k}\cdot(x_{l+1}-y_l)^{\beta/k}\cdot(y_{l+1}-x_{l+1})^{\beta/k}\,.
\end{aligned}$$

$$\cdot (y_l - y_{l-1})^{-\beta/k-2} \cdot (y_{l+1} - y_l)^{-\beta/k-2}\, dy_1 \dots dy_{k-1}.$$

Here and in the rest of the proof C always denotes a constant > 0 changing from line to line. Now we use the lower estimates

$$\begin{aligned}
(x_l - y_{l-1})^{\beta/k} &\geq \delta^{\beta/k}\ , \quad & (y_{l+1} - x_{l+1})^{\beta/k} &\geq \delta^{\beta/k},\\
(y_l - y_{l-1})^{-\beta/k-2} &\geq (3\delta)^{-\beta/k-2}\ , \quad & (y_{l+1} - y_l)^{-\beta/k-2} &\geq (3\delta)^{-\beta/k-2},\\
(y_{l-1} - y_{l-2})^{\beta/k} &\geq (x_l - y_{l-2})^{\beta/k}\ , \quad & (y_{l+2} - y_{l+1})^{\beta/k} &\geq (y_{l+2} - x_{l+1})^{\beta/k},
\end{aligned}$$

and

$$\begin{aligned}
\vartheta_{\beta/k}\left(\frac{x_{l-1} - y_{l-2}}{y_{l-1} - y_{l-2}}\right) &\geq \vartheta_{\beta/k}\left(\frac{x_{l-1} - y_{l-2}}{x_l - y_{l-2}}\right),\\
\vartheta_{\beta/k}\left(\frac{x_{l+2} - y_{l+1}}{y_{l+2} - y_{l+1}}\right) &\geq \vartheta_{\beta/k}\left(\frac{y_{l+2} - x_{l+2}}{y_{l+2} - x_{l+1}}\right)
\end{aligned}$$

valid for all y_{l-1}, y_l, y_{+1} in the restricted domains of integration. Moreover, we put

$$\begin{aligned}
\gamma_1(x_1, \dots, x_l) := \int_{x_1}^{x_2} \dots \int_{x_{l-2}}^{x_{l-1}} \prod_{i=1}^{l-2} \left[\vartheta_{\beta/k}\left(\frac{x_i - y_{i-1}}{y_i - y_{i-1}}\right) \cdot (y_i - y_{i-1})^{\beta/k-2}\right] \cdot\\
\cdot \vartheta_{\beta/k}\left(\frac{x_{l-1} - y_{l-2}}{x_l - y_{l-2}}\right) \cdot (x_l - y_{l-2})^{\beta/k}\, dy_{l-2} \dots dy_1
\end{aligned}$$

and similarly

$$\begin{aligned}
\gamma_2(x_{l+1}, \dots, x_k) := \int_{x_{l+2}}^{x_{l+3}} \dots \int_{x_{k-1}}^{x_k} \prod_{i=l+3}^{k} \left[\vartheta_{\beta/k}\left(\frac{x_i - y_{i-1}}{y_i - y_{i-1}}\right) \cdot (y_i - y_{i-1})^{\beta/k-2}\right] \cdot\\
\cdot \vartheta_{\beta/k}\left(\frac{y_{l+2} - x_{l+2}}{y_{l+2} - x_{l+1}}\right) \cdot (y_{l+2} - x_{l+1})^{\beta/k}\, dy_{k-1} \dots dy_{l+2}.
\end{aligned}$$

Then we obtain

$$\begin{aligned}
&\rho_k^\beta(x_1, \dots, x_k)\\
&\geq C \cdot \gamma_1(x_1, \dots, x_l) \cdot \gamma_2(x_{l+1}, \dots, x_k) \cdot\\
&\quad \cdot \delta^{-4} \cdot \int_{x_l - 2\delta}^{x_l - \delta} \int_{x_l}^{x_{l+1}} \int_{x_{l+1}+\delta}^{x_{l+1}+2\delta} (y_l - x_l)^{\beta/k} \cdot (x_{l+1} - y_l)^{\beta/k}\, dy_{l-1} dy_l dy_{l+1}
\end{aligned}$$

$$= C \cdot \gamma_1(x_1,\dots,x_l)\cdot \gamma_2(x_{l+1},\dots,x_k)\cdot \delta^{2\beta/k-1}.$$

This proves the claim. □

Remark. We do not know whether the exponent $2\beta/k-1$ in the previous lower estimate can be improved to $\beta/k-1$. In the upper estimate, the exponent $\beta/k-1$ is certainly optimal.

5 Projections, Isomorphisms, Approximations

5.1 Finite Dimensional Projections

For each linear subspace $H \subset L^2([0,1],\mathrm{Leb})$ let $\mathfrak{C}^1_H(\mathscr{G})$ denote the set of all functions $u:\mathscr{G}\to\mathbb{R}$ which can be written as $u(g) = U(\langle g,\psi_1\rangle,\dots,\langle g,\psi_n\rangle)$ with $n\in\mathbb{N}$, $U\in\mathscr{C}^1(\mathbb{R}^n,\mathbb{R})$ and $\psi_1,\dots,\psi_n\in H$. Moreover, let $\mathbb{D}_H$ denote the closure of $\mathfrak{C}^1_H(\mathscr{G})$ in $\mathbb{D} = Dom(\mathbb{E})$ w.r.t. the norm $(\mathbb{E}+\|.\|^2_{L^2(\mathbb{Q}^\beta)})^{1/2}$. Then $(\mathbb{E},\mathbb{D}_H)$ is a – not necessarily densely defined – Dirichlet form on $L^2(\mathscr{G},\mathbb{Q}^\beta)$.

More precisely, let $\mathbb{V}_H$ denote the closure of $\mathbb{D}_H$ in $L^2(\mathscr{G},\mathbb{Q}^\beta)$. Then $(\mathbb{E},\mathbb{D}_H)$ is a closed quadratic form in $\mathbb{V}_H$. As usual, there exist a strongly continuous semigroup $(\mathbb{T}^H_t)_{t\geq 0}$ and a resolvent $(\mathbb{G}^H_\alpha)_{\alpha>0}$, both consisting of Markovian operators on $\mathbb{V}_H$. Let $\pi_H : L^2(\mathscr{G},\mathbb{Q}^\beta)\to\mathbb{V}_H$ be the orthogonal projection onto the closed linear subspace $\mathbb{V}_H$. Then a semigroup on $L^2(\mathscr{G},\mathbb{Q}^\beta)$ – not necessarily strongly continuous, however – can be constructed by

$$\hat{\mathbb{T}}^H_t := \mathbb{T}^H_t \circ \hat{\pi}_H. \tag{18}$$

The projection $\hat{\pi}_H u$ of $u\in L^2(\mathscr{G},\mathbb{Q}^\beta)$ can be characterized as the conditional expectation

$$\hat{\pi}_H u(g) = \int_{\mathscr{G}} u(\tilde{g})\,\mathbb{Q}^\beta\left(d\tilde{g}\,|\{\langle\tilde{g},\varphi\rangle = \langle g,\varphi\rangle \text{ for all } \varphi\in H\}\right)$$

of the random variable $u:\tilde{g}\mapsto u(\tilde{g})$ on $\mathscr{G}$ under the condition

$$\{\langle\tilde{g},\varphi\rangle = \langle g,\varphi\rangle \text{ for all } \varphi\in H\}.$$

5.2 Monotone Convergence

Let $(H(k))_{k\in\mathbb{N}}$ be an increasing family of linear subspaces with $L^2([0,1],\mathrm{Leb}) = \overline{\bigcup_k H(k)}$ and define $\mathbb{D}_{H(k)}$ as above. Then $\mathbb{D}_{H(k)}\nearrow$ with $\overline{\bigcup_k \mathbb{D}_{H(k)}} = \mathbb{D}$. In particular,

$$(\mathbb{E},\mathbb{D}_{H(k)})\to(\mathbb{E},\mathbb{D}) \quad \text{in the sense of Mosco for } k\to\infty.$$

Hence, if $\hat{\mathbb{T}}_t^{H(k)}$ and $\hat{\mathbb{G}}_\alpha^{H(k)}$ denote the semigroup and resolvent operators on $L^2(\mathcal{G}, \mathbb{Q}^\beta)$ associated with $(\mathbb{E}, \mathbb{D}_{H(k)})$ and if $\mathbb{T}_t$ and $\mathbb{G}_\alpha$ denote the corresponding operators associated with $(\mathbb{E}, \mathbb{D})$ then

$$\hat{\mathbb{T}}_t^{H(k)} \to \mathbb{T}_t, \quad \hat{\mathbb{G}}_\alpha^{H(k)} \to \mathbb{G}_\alpha \quad \text{strongly in } L^2(\mathcal{G}, \mathbb{Q}^\beta) \text{ as } k \to \infty,$$

cf. [8].

5.3 Isomorphisms I

Let H be finite dimensional with basis $\mathcal{H} = \{\varphi^{(1)}, \dots, \varphi^{(k)}\}$ and consider the map

$$\hat{\mathcal{J}}_{\mathcal{H}} : L^2([0,1], \text{Leb}) \to \mathbb{R}^k, \quad g \mapsto \left(\langle g, \varphi^{(1)}\rangle, \dots, \langle g, \varphi^{(k)}\rangle\right).$$

Its restriction to H – denoted by $\mathcal{J}_{\mathcal{H}}$ – is a vector space isomorphism with $\mathcal{J}_{\mathcal{H}}^{-1} : \mathbb{R}^k \to H, \quad x \mapsto \sum_{i,j=1}^k x_i a_{ij}^{-1} \varphi^{(j)}$ where (a_{ij}^{-1}) denotes the inverse of the matrix (a_{ij}) defined by $a_{ij} = \langle \varphi^{(i)}, \varphi^{(j)}\rangle$. This map induces an isomorphism between $\mathcal{C}^1(\mathbb{R}^k)$ and $\mathfrak{C}_H^1(\mathcal{G})$:

$$U \in \mathcal{C}^1(\mathbb{R}^k) \overset{U = u \circ \mathcal{J}_{\mathcal{H}}^{-1}}{\longleftrightarrow} u \in \mathfrak{C}_H^1(\mathcal{G}).$$

Let $m_{\mathcal{H}}^\beta$ denote the distribution of the random vector $\left(\langle g, \varphi^{(1)}\rangle, \dots, \langle g, \varphi^{(k)}\rangle\right)$, that is, $m_{\mathcal{H}}^\beta := (\hat{\mathcal{J}}_{\mathcal{H}})_* \mathbb{Q}^\beta$ and define a pre-Dirichlet form on $L^2(\mathbb{R}^k, m_{\mathcal{H}}^\beta) = \left\{u \circ \mathcal{J}_{\mathcal{H}}^{-1} : u \in \mathbb{V}_H\right\}$ by

$$\mathcal{E}_{\mathcal{H}}(U, V) := \sum_{i,j=1}^k a_{ij} \int_{\mathbb{R}^k} \partial_i U(x) \partial_j V(x) \, dm_{\mathcal{H}}^\beta(x) \tag{19}$$

for $U, V \in \mathcal{C}^1(\mathbb{R}^k)$. This form is closable – since the closable form $(\mathbb{E}, \mathfrak{C}_H^1(\mathcal{G}))$ is isomorphic to it – with closure being a strongly local Dirichlet form on $L^2(\mathbb{R}^k, m_{\mathcal{H}}^\beta)$ with domain

$$\mathcal{D}_{\mathcal{H}} = \left\{u \circ \mathcal{J}_{\mathcal{H}}^{-1} : u \in \mathbb{D}_H\right\}$$

and with

$$\mathcal{E}_{\mathcal{H}}(U, V) = \mathbb{E}(U \circ \hat{\mathcal{J}}_{\mathcal{H}}, V \circ \hat{\mathcal{J}}_{\mathcal{H}})$$

for $U, V \in \mathcal{D}_{\mathcal{H}}$, cf. [5].

Let $(T_t^{\mathcal{H}})_{t>0}$ denote the semigroup associated with $(\mathcal{E}_{\mathcal{H}}, \mathcal{D}_{\mathcal{H}})$. Then for all $u \in \mathbb{V}_H \subset L^2(\mathcal{G}, \mathbb{Q}^\beta)$

$$\mathbb{T}_t^H u = \left(T_t^{\mathscr{H}} U\right)\left(\hat{\mathscr{J}}_{\mathscr{H}}\right) \tag{20}$$

with $U \in L^2(\mathbb{R}^k, m^\beta_{\mathscr{H}})$ such that $u = U \circ \hat{\mathscr{J}}_{\mathscr{H}}$.

5.4 Standard Approximations

For each $k \in \mathbb{N}$ let us from now on fix the linear subspace $H(k) \subset L^2([0,1], \text{Leb})$ spanned by the orthogonal system $\mathscr{H}(k) = \{\varphi_k^{(1)}, \dots, \varphi_k^{(k)}\}$ with

$$\varphi_k^{(i)}(t) := k \cdot 1_{(\frac{i-1}{k}, \frac{i}{k}]}(t).$$

To simplify notation, write m_k^β, $\mathscr{J}_k$, $\mathscr{E}_k$, T_t^k etc. instead of $m^\beta_{\mathscr{H}(k)}$, $\mathscr{J}_{\mathscr{H}(k)}$, $\mathscr{E}_{\mathscr{H}(k)}$, $T_t^{\mathscr{H}(k)}$, respectively.

Note that in this case

$$\hat{\mathscr{J}}_k(g) = \left(k \int_0^{\frac{1}{k}} g(t)dt, \dots, k \int_{\frac{k-1}{k}}^1 g(t)dt\right) = \left(\int_0^1 \Phi_k^{(1)} dg, \dots, \int_0^1 \Phi_k^{(k)} dg\right)$$

with $\Phi_k^{(i)}$ as introduced in (13). Hence, the measure $m_k^\beta := (\hat{\mathscr{J}}_k)_* \mathbb{Q}^\beta$ on $\mathbb{R}^k$ coincides with the measure investigated in detail in the previous chapter. In particular,

$$dm_k^\beta(x) = \rho_k^\beta(x)\, dx$$

with ρ_k^β given by formula (14). Recall that ρ_k^β is continuous and > 0 on the open simplex $\Sigma_k \subset \mathbb{R}^k$ and that it vanishes on $\mathbb{R}^k \setminus \Sigma_k$.

The Dirichlet form $(\mathscr{E}_k, \mathscr{D}_k)$ on $L^2(\mathbb{R}^k, \rho_k^\beta)$ is given explicitly on its core $\mathscr{C}^1(\mathbb{R}^k)$ by

$$\mathscr{E}_k(U, V) = k \int_{\mathbb{R}^k} \nabla U(x) \cdot \nabla V(x)\, dm_k^\beta(x) \tag{21}$$

with ∇U denoting the gradient of U on $\mathbb{R}^k$. If we regard it as a Dirichlet form on $L^2(\overline{\Sigma_k}, \rho_k^\beta)$ then it is *regular, strongly local and recurrent.* (Indeed, $\{u|_{\overline{\Sigma_k}} : u \in \mathscr{C}^1(\mathbb{R}^k)\}$ is dense in $\mathscr{C}(\overline{\Sigma_k})$ as well as in $\mathscr{D}_k$. Strong locality and recurrence is inherited from $(\mathbb{E}, \mathbb{D})$.)

The semigroup $(T_t^k)_{t \geq 0}$ associated with $(\mathscr{E}_k, \mathscr{D}_k)$ can be represented as

$$T_t^k u(x) = \mathbb{E}_x \left[u\left(X_t^k \right) \right] \tag{22}$$

(for all Borel functions $u \in L^2(\overline{\Sigma_k}, \rho_k^\beta)$ and a.e. $x \in \overline{\Sigma_k}$) in terms of a strong Markov process

$$(X_t^k)_{t\geq 0} = \left(X_t^{k,1}, \ldots, X_t^{k,k} \right)_{t\geq 0}$$

with state space $\overline{\Sigma_k}$, defined on some probability space $(\Omega, \mathscr{F}, \mathbf{P}_x)_{x\in\overline{\Sigma_k}}$ and canonically associated with $(\mathscr{E}_k, \mathscr{D}_k)$. This process is continuous, recurrent and reversible w.r.t. m_k^β. At least on those stochastic intervals for which $X_t^k(\omega) \in \Sigma_k$ it can be characterized as the solution to an interacting system of stochastic differential equations

$$dX_t^{k,i} = k \frac{\partial \log \rho_k^\beta}{\partial x_i} \left(X_t^k \right) dt + \sqrt{2k}\, dW_t^i, \quad i = 1, \ldots, k \tag{23}$$

for some k-dimensional Brownian motion $(W_t)_{t\geq 0}$.

5.5 Isomorphisms II

Let $\mathscr{G}_k := \mathscr{G} \cap H(k)$ denote the subset of those $g \in \mathscr{G}$ which are constant on each of the intervals $[\frac{i-1}{k}, \frac{i}{k})$ for $i = 1, \ldots, k$. Then

$$\mathscr{J}_k^{-1} : \overline{\Sigma_k} \to \mathscr{G}_k,\ x \mapsto \sum_{i=1}^k x_i \cdot 1_{[\frac{i-1}{k}, \frac{i}{k})}$$

is a bijection. It maps the strong Markov process $(X_t^k)_{t\geq 0}$ on $\overline{\Sigma_k}$ onto a strong Markov process $(g_t^k)_{t\geq 0}$ on $\mathscr{G}_k$ with

$$g_t^k(\omega) := \mathscr{J}_k^{-1} \left(X_t^k(\omega) \right) = \sum_{i=1}^k X_t^{k,i}(\omega) \cdot 1_{[\frac{i-1}{k}, \frac{i}{k})}. \tag{24}$$

Now recall that the Hilbert space $\mathbb{V}_k := \overline{\mathfrak{C}_k^1(\mathscr{G})}^{L^2(\mathscr{G}, \mathbb{Q}^\beta)}$ coincides with

$$\left\{ U \circ \mathscr{J}_k : U \in L^2(\mathbb{R}^k, m_k^\beta) \right\}.$$

Hence, (20) together with (22) and (24) imply

$$\mathbb{T}_t^k u(g) = \mathbf{E}_g\left[u\left(g_t^k\right)\right] = \mathbf{E}_{\mathscr{J}_k(g)}\left[u\left(\sum_{i=1}^k X_t^{k,i}\cdot 1_{[\frac{i-1}{k},\frac{i}{k})}\right)\right] \tag{25}$$

for all Borel functions $u \in \mathbb{V}_k$ and a.e. $g \in \mathscr{G}$. Finally, according to (18)

$$\hat{\mathbb{T}}_t^k u(g) = \mathbf{E}_{\mathscr{J}_k(g)}\left[u_k\left(\sum_{i=1}^k X_t^{k,i}\cdot 1_{[\frac{i-1}{k},\frac{i}{k})}\right)\right] \tag{26}$$

for all Borel functions $u \in L^2(\mathscr{G}, \mathbb{Q}^\beta)$ and a.e. $g \in \mathscr{G}$ with $u_k = \hat{\pi}_k u$ being the projection of u onto $\mathbb{V}_k$ (or, in other words, the conditional expectation of u).

This process canonically extends to a – not necessarily normal – strong Markov process $(g_t^k)_{t\geq 0}$ on $\mathscr{G}$, projecting the initial data by means of the map

$$\pi_k := \mathscr{J}_k^{-1} \circ \hat{\mathscr{J}}_k : \mathscr{G} \to \mathscr{G}_k,\ g \mapsto \frac{1}{k}\sum_{i=1}^k \langle g, \varphi_k^{(i)}\rangle \varphi_k^{(i)}.$$

5.6 Isomorphisms III

Let $\mathscr{P}_k$ denote the subset of $\mu \in \mathscr{P}$ which can be represented as $\mu = \frac{1}{k}\sum_{i=1}^k \delta_{x_i}$ for suitable $x_1, \ldots, x_k \in [0,1]$. The maps $\chi : \mathscr{G}_k \mapsto \mathscr{P}_k$ and $\mathscr{I}_k := \mathscr{J}_k \circ \chi^{-1} : \mathscr{P}_k \to \overline{\Sigma}_k$ establish canonical isomorphisms. The inverse of the latter

$$\mathscr{I}_k^{-1} : x \mapsto \frac{1}{k}\sum_{i=1}^k \delta_{x_i}$$

defines the canonical embedding of $\overline{\Sigma}_k$ into $\mathscr{P}$. On the other hand, the map

$$\hat{\mathscr{I}}_k := \hat{\mathscr{J}}_k \circ \chi^{-1} : \mathscr{P} \to \overline{\Sigma}_k$$

can be characterized as follows: Each $\mu \in \mathscr{P}$ can be represented uniquely as the sum $\mu = \frac{1}{k}\sum_{i=1}^k \mu_i$ with probability measures μ_i supported on $[y_{i-1}, y_i]$ for suitable $0 \leq y_1 \leq \ldots \leq y_k \leq 1$. (Indeed, $y_i = \inf\{t \geq 0 : \mu([0,t]) > \frac{i}{k}\}$ for each $i = 1, \ldots, k$.) Then

$$\hat{\mathscr{I}}_k(\mu) = (x_1, \ldots, x_k)$$

with $x_i = x_i^\mu = \int_0^1 t\, d\mu_i(t)$ being the mean value of the probability measure μ_i.

In particular, the projection $\pi_k = \mathscr{I}_k^{-1} \circ \hat{\mathscr{I}}_k : \mathscr{P} \to \mathscr{P}_k$ is defined by

$$\mu \mapsto \frac{1}{k}\sum_{i=1}^{k}\delta_{x_i^\mu}.$$

Let $(\mu_t^k)_{t\geq 0}$ be the image of the strong Markov process $(g_t^k)_{t\geq 0}$ under the bijection $\chi : g \mapsto g_*\mathrm{Leb}|_{[0,1]}$. Then

$$\mu_t^k(\omega) = \frac{1}{k}\sum_{i=1}^{k}\delta_{X_t^{k,i}(\omega)}.$$

In other words, the strong Markov process $(\mu_t^k)_{t\geq 0}$ on $\mathscr{P}_k$ is the empirical distribution of the strong Markov process $(X_t^k)_{t\geq 0}$ on $\overline{\Sigma_k}$.

Finally, a probabilistic representation – similar to that for $\left(\hat{\mathbb{T}}_t^k\right)_{t\geq 0}$ – also holds true for the semigroup $\left(\hat{\mathbb{T}}_{\mathscr{P},t}^k\right)_{t\geq 0}$ associated with the Dirichlet form $(\mathbb{E}_{\mathscr{P}}, \mathbb{D}_{\mathscr{P}})$ on $L^2(\mathscr{P}, \mathbb{P}^\beta)$:

$$\hat{\mathbb{T}}_{\mathscr{P},t}^k u(\mu) = \mathbf{E}_{x_\mu}\left[u_k\left(\frac{1}{k}\sum_{i=1}^{k}\delta_{X_t^{k,i}}\right)\right] \tag{27}$$

for all Borel functions $u \in L^2(\mathscr{P}, \mathbb{P}^\beta)$ and a.e. $\mu \in \mathscr{P}$ and with $x_\mu := \mathscr{I}_k(\mu)$.

6 Convergence

6.1 Convergence of Finite Dimensional Distributions

Note that $H(2^k) \subset H(2^n)$ for $k, n \in \mathbb{N}$, $k \leq n$, and thus $\mathbb{D}^{2^k} \subset \mathbb{D}^{2^n}$, $\mathbb{V}^{2^k} \subset \mathbb{V}^{2^n}$. According to Sect. 5.1

$$\mathbb{T}_t^{2^k} u \to \mathbb{T}_t u \quad \text{in } L^2(\mathscr{G}, \mathbb{Q}^\beta) \qquad \text{as } k \to \infty \tag{28}$$

for all $u \in \mathbb{V}^\infty := \bigcup_{n\in\mathbb{N}} \mathbb{V}^{2^n}$. The latter is a dense subset in $L^2(\mathscr{G}, \mathbb{Q}^\beta)$. The previous in particular implies

$$\langle u, \mathbb{T}_t^{2^k} v\rangle_{L^2(\mathscr{G},\mathbb{Q}^\beta)} \to \langle u, \mathbb{T}_t v\rangle_{L^2(\mathscr{G},\mathbb{Q}^\beta)} \qquad \text{as } k \to \infty \tag{29}$$

for all $u, v \in \mathbb{V}^\infty$ and thus

$$\mathbf{E}_{\mathbb{Q}_k^\beta}\left[u(g_0^{2^k}) \cdot v(g_t^{2^k})\right] \to \mathbf{E}_{\mathbb{Q}}\left[u(g_0) \cdot v(g_t)\right] \qquad \text{as } k \to \infty \tag{30}$$

for all $u, v \in \mathscr{C}(\mathscr{G})$.

The Markov property of the processes $(g_t)_{t\geq 0}$ and $(g_t^{2^k})_{t\geq 0}$ together with their invariance w.r.t. the measures $\mathbb{Q}^\beta$ and $\mathbb{Q}^\beta_{2^k}$ allows to iterate this argumentation which then yields

$$\begin{aligned}
&\mathbf{E}_{\mathbb{Q}^\beta_{2^k}}\left[u_1(g_{t_1}^{2^k})\cdot u_2(g_{t_2}^{2^k})\cdot\ldots\cdot u_N(g_{t_N}^{2^k})\right]\\
&=\int_{\mathscr{G}} u_1\cdot\mathbb{T}^{2^k}_{t_1-t_0}\left(u_2\cdot T^{2^k}_{t_2-t_1}\left(u_3\cdot\ldots\cdot T^{2^k}_{t_N-t_{N-1}}u_N\right)\ldots\right)\,d\mathbb{Q}^\beta_{2^k}\\
&\qquad\qquad\downarrow\\
&=\int_{\mathscr{G}} u_1\cdot\mathbb{T}_{t_1-t_0}\left(u_2\cdot T_{t_2-t_1}\left(u_3\cdot\ldots\cdot T_{t_N-t_{N-1}}u_N\right)\ldots\right)\,d\mathbb{Q}^\beta\\
&=\mathbf{E}_{\mathbb{Q}^\beta}\left[u_1(g_{t_1})\cdot u_2(g_{t_2})\cdot\ldots\cdot u_N(g_{t_N})\right]
\end{aligned}$$

as $k\to\infty$ for all $N\in\mathbb{N}$, all $0\leq t_1<\ldots<t_N$ and all $u_1,\ldots,u_N\in\mathscr{C}(\mathscr{G})$. Since functions $U\in\mathscr{C}(\mathscr{G}^N)$ can be approximated uniformly by linear combinations of functions of the form $U(g_1,g_2,\ldots,g_n)=\prod_{n=1}^N u_n(g_n)$ it follows that

$$\mathbf{E}_{\mathbb{Q}^\beta_{2^k}}\left[U(g_{t_1}^{2^k},g_{t_2}^{2^k},\ldots,g_{t_N}^{2^k})\right]\to\mathbf{E}_{\mathbb{Q}^\beta}\left[U(g_{t_1},g_{t_2},\ldots,g_{t_N})\right]$$

as $k\to\infty$ for all $N\in\mathbb{N}$, all $0\leq t_1<\ldots<t_N$ and all $U\in\mathscr{C}(\mathscr{G}^N)$. That is, we have proven the convergence

$$(g_t^{2^k})_{t\geq 0}\ \to\ (g_t)_{t\geq 0}\qquad\text{as } k\to\infty\tag{31}$$

in the sense of weak convergence of the finite dimensional distributions of the processes, started with their respective invariant distributions. By means of the various isomorphisms presented before, this can be equivalently restated as convergence

$$(\mu_t^{2^k})_{t\geq 0}\ \to\ (\mu_t)_{t\geq 0}\qquad\text{as } k\to\infty,\tag{32}$$

again in the sense of weak convergence of the finite dimensional distributions of the processes, started with their respective invariant distributions. Here $(\mu_t)_{t\geq 0}$ denotes the Wasserstein diffusion on $\mathscr{P}$ – associated with the Dirichlet form (1) – with the entropic measure $\mathbb{P}^\beta$ as invariant distribution and

$$\mu_t^{2^k}(\omega)=\frac{1}{2^k}\sum_{i=1}^{2^k}\delta_{X_t^{2^k,i}(\omega)}$$

with $\left(X_t^{2^k,i}\right)_{t\geq 0}$ being the continuous Markov process on the simplex $\overline{\Sigma}_{2^k}$ – associated with the Dirichlet form (21) – with invariant distribution $\rho^\beta_{2^k}(x)dx$.

6.2 Convergence of Processes

Convergence of the processes

$$(g_t^{2^k})_{t\geq 0} \;\to\; (g_t)_{t\geq 0} \qquad \text{as } k\to\infty$$

will follow from the convergence (31) of the respective finite dimensional distributions provided we prove tightness of the family $(g_t^{2^k})_{t\geq 0}, k\in\mathbb{N}$ in $\mathscr{C}(\mathbb{R}_+,\mathscr{G})$. The latter is equivalent to tightness of $\left(\langle\psi, g_t^{2^k}\rangle\right)_{t\geq 0}, k\in\mathbb{N}$ in $\mathscr{C}(\mathbb{R}_+,\mathbb{R})$ for all $\psi\in L^2([0,1],\mathrm{Leb})$. It suffices to verify this for a dense subset of ψ, e.g. for all $\psi\in\bigcup_{l=1}^{\infty}\mathscr{H}(2^l)\subset L^2([0,1],\mathrm{Leb})$.

Fix $\psi\in\mathscr{H}(2^l)$ for some $l\in\mathbb{N}$ with $\|\psi\|=1$. For each $k\in\mathbb{N}, k\geq l$ the continuous function $u(g):=\langle\psi,g\rangle$ lies in $\mathbb{V}_{2^k}$ with energy $\mathbb{E}(u)=\|\psi\|^2=1$ and square field operator

$$\Gamma_{\langle u\rangle}(g)=1 \tag{33}$$

for a.e. $g\in\mathscr{G}$.

Given $T>0$, the process

$$\left(u(g_t^{2^k})\right)_{t\in[0,T]}$$

admits a Lyons-Zheng decomposition

$$u(g_t^{2^k})-u(g_0^{2^k})=\frac{1}{2}M_t^{(2^k)}-\frac{1}{2}\left[M_T^{(2^k)}-M_{T-t}^{(2^k)}\right]\circ r_T$$

into a forward martingale and a backward martingale. According to (33) the quadratic variation of the forward martingale – as well as that of the backward martingale – is given by

$$\langle M^{(2^k)}\rangle_t=t,$$

uniformly in $g\in\mathscr{G}$ and in $k\in\mathbb{N}, k\geq l$. Hence, using hitting probabilities of 1-dimensional Brownian motions we deduce for any $R>0$ and uniformly in $k\in\mathbb{N}$, $k\geq l$,

$$\begin{aligned}
&\mathbf{P}_{\mathbb{Q}^{\beta}_{2^k}}\left[\sup_{t\in[0,T]}\left(u(g_t^{2^k})-u(g_0^{2^k})\right)>R\right]\\
&\leq \mathbf{P}_{\mathbb{Q}^{\beta}_{2^k}}\left[\sup_{t\in[0,T]}M_t^{(2^k)}>R\right]+\mathbf{P}_{\mathbb{Q}^{\beta}_{2^k}}\left[\sup_{t\in[0,T]}\left(M_T^{(2^k)}-M_{T-t}^{(2^k)}\right)\circ r_T>R\right]\\
&\leq 2\sqrt{\frac{2}{\pi}}\exp\left(-\frac{(R/2)^2}{2T}\right).
\end{aligned}$$

Since we already know that the 1-dimensional distributions $g_0^{2^k}$ converge, this proves tightness of the family of processes

$$\left(u(g_t^{2^k})\right)_{t\in[0,T]} = \left(\langle\psi, g_t^{2^k}\rangle\right)_{t\in[0,T]}$$

for $k \in \mathbb{N}$. Since this holds for all $\psi \in \bigcup_{l=1}^{\infty} \mathscr{H}(2^l)$ it implies tightness of the family $(g_t^{2^k})_{t\geq 0}, k \in \mathbb{N}$, and thus convergence of the processes

$$(g_t^{2^k})_{t\geq 0} \ \to \ (g_t)_{t\geq 0} \qquad \text{as } k \to \infty.$$

Applying the usual isomorphism, this may be restated as convergence of the processes

$$(\mu_t^{2^k})_{t\geq 0} \ \to \ (\mu_t)_{t\geq 0} \qquad \text{as } k \to \infty$$

in $\mathscr{C}(\mathbb{R}_+\mathscr{P})$.

6.3 Final Remarks

Given $k \in \mathbb{N}$ a mapping $\tilde{\mathscr{J}}_k : \mathscr{G} \to \Sigma_k$ – very similar to our mapping $\hat{\mathscr{J}}_k$ from (12) – is obtained by replacing the functions $\Phi_k^{(i)}$ from (13) by $\tilde{\Phi}_k^{(i)}(x) := 1_{[0,\frac{2i-1}{2k}]}(x)$ which leads to

$$\tilde{\mathscr{J}}_k(g) = \left(\int_0^1 \tilde{\Phi}_k^{(i)}\, dg\right)_{i=1,\dots,k} = \left(g\left(\frac{2i-1}{2k}\right)\right)_{i=1,\dots,k}.$$

In this case, the identification of the push forward measure $\tilde{m}_k^\beta := (\tilde{\mathscr{J}}_k)_* \mathbb{Q}^\beta$ on Σ_k is much easier. Indeed, it is absolutely continuous with density

$$\tilde{\rho}_k(x) = C \cdot [x_1(1-x_k)]^{\beta/(2k)-1} \cdot \prod_{i=2}^{k} (x_i - x_{i-1})^{\beta/k-1}.$$

The strong Markov process on $\overline{\Sigma_k}$ associated with the Dirichlet form $\tilde{\mathscr{E}}_k(U) = k\int_{\Sigma_k} |\nabla U|^2 \tilde{\rho}_k^\beta\, dx$ on $L^2(\Sigma_k, \tilde{\rho}_k^\beta\, dx)$ admits a very explicit characterization: at least on those stochastic intervals on which the process is in the interior of the simplex, it is a weak solution to the coupled system of stochastic differential equations

$$dX_t^{k,i} = \left[\frac{\beta_{i-1} - k}{X_t^{k,i} - X_t^{k,i-1}} - \frac{\beta_i - k}{X_t^{k,i+1} - X_t^{k,i}} \right] dt + \sqrt{2k}\, dW_t^i, \quad i = 1, \ldots, k \tag{34}$$

for some k-dimensional Brownian motion $(W_t)_{t \geq 0}$ and with $X_t^{k,0} := 0, X_t^{k,k+1} := 1$. Here $\beta_0 = \beta_k = \beta/2$ and $\beta_i = \beta$ for $i = 1, \ldots, k-1$. This is essentially the approximation used by Andres and von Renesse [2].

The fundamental disadvantage, however, is that the functions $g \mapsto \int_0^1 \tilde{\Phi}_k^{(i)}\, dg$ are no longer in the domain of the Dirichlet form $\mathbb{E}$. More generally, for any non-constant $U \in \mathscr{C}^1(\mathbb{R}^k)$ the function $u(g) := U(\tilde{\mathscr{J}}_k(g))$ is neither continuous on $\mathscr{G}$ nor does it belong to $\mathbb{D}$.

References

1. Ambrosio, L., Gigli, N., Savaré, G.: Gradient Flows in Metric Spaces and in the Space of Probability Measures. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel (2005)
2. Andres, S., von Renesse, M.: Particle approximation of the wasserstein Diffusion. J. Funct. Anal. **258**(11), 3879–3905 (2010)
3. Doering, M., Stannat, W.: The logarithmic Sobolev inequality for the Wasserstein diffusion. Probab. Theory Relat. Fields **145**, 189–209 (2009)
4. Ferguson, T.: A Bayesian analysis of some nonparametric problems. Ann. Stat. **1**, 209–230 (1973)
5. Fukushima, M., Oshima, Y., Takeda, M.: Dirichlet Forms and Symmetric Markov Processes. de Gruyter Studies in Mathematics, vol. 19. de Gruyter, Berlin (1994)
6. Lijoi, A., Regazzini E.: Means of a Dirichlet process and multiple hypergeometric functions. Ann. Probab. **32**, 1469–1495 (2004)
7. Otto, F.: The geometry of dissipative evolution equations: the porous medium equation. Commun. Partial Differ. Equ. **26**(1–2), 101–174 (2001)
8. Reed, M., Simon, B.: Methods of Modern Mathematical Physics (I). Functional Analysis, 2nd edn. Academic, New York (1980)
9. Regazzini, E., Guglielmi, A., Di Nunno, G.: Theory and numerical analysis for exact distributions of functionals of a Dirichlet process. Ann. Stat. **30**, 1376–1411 (2002)
10. Sturm, K.-T.: Entropic measure on multidimensional spaces. Semiar on stochastic analysis, random fields and applications VI. Prog. Probab. **63**, 261–277 (2011)
11. Villani, C.: Topics in Optimal Transportation. Graduate Studies in Mathematics, vol. 58. American Mathematical Society, Providence (2003)
12. von Renesse, M., Sturm, K.-T.: Entropic measure and Wasserstein diffusion. Ann. Probab. **37**, 1114–1191 (2009)
13. von Renesse, M., Zambotti, L., Yor, M.: Quasi-invariance properties of a class of subordinators. Stoch. Process. Appl. **118**, 2038–2057 (2008)

Adapted Function Spaces for Dispersive Equations

Herbert Koch

Abstract The study of the p variation of functions of one variable has a long history. It has been discussed by Wiener in [21]. Here we define the space of functions of finite p Variation, and the predual space U^q, and we use them to study dispersive equations.

1 Introduction

The study of the p variation of functions of one variable has a long history. It has been discussed by Wiener in [21]. The generalization of the Riemann-Stieltjes integral to functions of bounded p variation against the deriative of a function of bounded q variation $1/p + 1/q > 1$ is due to Young [22]. Much later Lyons developed his theory of rough path [13] and [14], buildung on Young's ideas, but going much further.

In parallel Tataru realized that the spaces of bounded p variation, and their close relatives, the U^p spaces, allow a powerful sharping of Bourgain's technique of function spaces adapted to the dispersive equation at hand. These ideas were applied for the first time in the work of the author and Tataru in [11]. Since then there has been a number of questions in dispersive equations where these function spaces have been used. For example they play a crucial role in [12], but there they could probably be replaced by Bourgain's Fourier restriction spaces $X^{s,b}$. On the other hand, for wellposedness for the Kadomtsev-Petviashvili II equation in a critical function space (see [3]) the $X^{s,b}$ spaces seem to be insufficient. The theory of the spaces U^p and V^p and some of their basic properties like duality and logarithmic interpolation have been worked out in [3], with a focus on what

H. Koch (✉)
Mathematisches Institut, Rheinische Friedrich-Wilhelms-Universität Bonn, Endenicher Allee 60, D-53115 Bonn, Germany
e-mail: koch@math.uni-bonn.de

M. Griebel (ed.), *Singular Phenomena and Scaling in Mathematical Models*,
DOI 10.1007/978-3-319-00786-1_3,

was needed there. Until very recently the developments in stochastic differential equations and dispersive equations were entirely independent. The present treatment considerably extends the theory of [3].

We will introduce the spaces U^p and V^p, study their properties and indicate their role for dispersive equations. After that we turn to wellposedness questions for several dispersive PDEs, where we select a number of relevant and representative problems.

In the sequel $p \in [1,\infty]$. Unless explicitly stated otherwise we consider $p \in (1,\infty)$.

2 The Bounded *p* Variation

Definition 1. Let I be an interval, $1 \leq p < \infty$ and $f : I \to X$. We define

$$\omega_p(v,I) := \sup_{t_i \in I, t_1 < t_2 \dots t_{n+1}} \Big(\sum \|v(t_{i+1}) - v(t_i)\|_X^p\Big)^{1/p} \in [0,\infty].$$

There are obvious properties. The function $t \to \omega_p(v,[a,t))$ is monotonically increasing. The same is true if we consider closed or open intervals. Moreover

$$\omega_p(v,[a,b)) \leq \omega_p(v,[a,c)) \leq 2(\omega_p(v,[a,b]) + \omega_p(v,[b,c))).$$

Finiteness of the p variation implies existence of one sided limits. It is not hard to see that $v \to \omega_p(v,[a,b))$ defines a norm, up to constants. If v is continuous and the p variation is bounded then it is a continuous function of the endpoint. Moreover

$$\omega_p(v,(a,b)) \leq |b-a|^{1/p} \|v\|_{\dot{C}^{1/p}}$$

where $\dot{C}^{1/p}$ denotes the homogeneous Hölder space.

2.1 Step Functions and Ruled Functions

We introduce and study functions from an interval $[a,b)$ to $\mathbb{R}$, $\mathbb{R}^n$, a Hilbert space or a Banach space X, and spaces of such functions which are invariant under continuous monotone reparametrizations of the interval. For the most part of this section there are no more than the obvious modifications when considering Banach space valued functions.

We call a function f a ruled function if at every point (including the endpoints, which may be $\pm\infty$) left and right limits exist. This set is closed with respect to uniform convergence. We denote the Banach space of ruled functions equipped with the supremum norm by $\mathscr{R}$.

A partition τ of $[a,b)$ is a strictly increasing finite sequence

$$a < t_1 < t_2 < \cdots < t_{n+1} < b$$

where we allow $b = \infty$ and $a = -\infty$. A step function is a function f for which there exists a partition so that f is constant on every interval (a,t_1), (t_i,t_{i+1}) and (t_n,b). We do not require that the value at a point coincides with the limit from either side. Step functions are dense in $\mathscr{R}$. We denote the set of step functions by $\mathscr{S}$. Let $\mathscr{R}_{rc}$ be the closed subset of $\mathscr{R}$ of right continuous functions f with $\lim_{t\to a} f(t) = 0$. Similarly, if $X \subset \mathscr{R}$ we denote by X_{rc} the intersection with $\mathscr{R}_{rc}$.

The step functions

$$f_t = \chi_{[t,b)}$$

satisfy

$$\|f_t - f_s\|_{\sup} = 1 \tag{1}$$

for $s \neq t$. We will study Banach spaces Z most of which contain the right continuous step functions $\mathscr{S}_{rc}$, and which embed into $\mathscr{R}$. Moreover we will always have

$$1 \leq \|f_t - f_s\|_Z \leq 2 \tag{2}$$

and hence none of those spaces is separable.

It will be convenient to extend every function on $[a,b)$ by zero to $[a,b]$, i.e. we will always set $f(b) = 0$, even if $a = -\infty$ or $b = \infty$.

Definition 2. For $f \in \mathscr{R}$ and a partition

$$\tau = (t_1, t_2 \dots t_n), \quad a < t_1 < t_2 < t_3 \cdots < t_n < b$$

we define (denoting the limit from the right by $f(t+)$)

$$f_\tau(t) = \begin{cases} f(t) & \text{if } t = t_j \\ f(a+) & \text{if } a < t < t_1 \\ f(t_i+) & \text{if } t_i < t < t_{i+1} \\ f(t_n) & \text{if } t_n < t \end{cases}$$

We observe that f_τ is a step function, and it is right continuous if f is right continuous.

2.2 The Spaces V^p and U^p

In this subsection we consider functions on (a,b) where we allow the cases $a = -\infty$ and $b = \infty$.

Definition 3. Let X be a Banach space, $1 \le p < \infty$ and $v : (a,b) \to X$. We define

$$\|v\|_{V^p((a,b),X)} = \max\{\|v\|_{sup}, \omega_p(v,(a,b))\}.$$

Let $V^p = V^p((a,b)) = V^p(X) = V^p((a,b);X)$ be the set of all functions for which this expression is finite. We often suppress the interval and/or the Banach space in the notation when this seems appropriate.

The interval will usually be of minor importance. We omit it often in the sequel. The following properties are immediate:

1. $V^p(I)$ is closed with respect to this norm and hence $V^p(I)$ is a Banach subspace of $\mathscr{R}$. Moreover $V^p_{rc}(I)$ is a closed subspace.
2. We set $V^\infty = \mathscr{R}$.
3. If $1 \le p \le q \le \infty$ then
$$\|v\|_{V^q} \le \|v\|_{V^p}.$$
4. Let X_i be Banach spaces, $T : X_1 \times X_2 \to X_3$ a bounded bilinear operator, $v \in V^p(X_1)$ and $w \in V^p(X_2)$. Then $T(v,w) \in V^p(X_3)$ and
$$\|T(v,w)\|_{V^p(X_3)} \le 2\|T\|\|v\|_{V^p(X_1)}\|w\|_{V^p(X_2)}.$$
5. We embed $V^p((a,b))$ into $V^p(\mathbb{R})$ by extending v by 0.
6. The space V^1 has some additional structure: Every bounded monotone function is in V^1, and functions in V^1 can be written as the difference of two bounded monotone functions.

The space of bounded p variation is build on the sequence space l^p. We may also replace it by the weak space l^p_w with

$$\|(x_j)\|_{l^p_w} = \sup_\lambda \lambda \#\{|x_j| > \lambda\}^{1/p}.$$

Definition 4. Let $1 \le p < \infty$. The weak V^p_w space consists of all functions such that

$$\|v\|_{V^p_w} = \max\{\sup_{t_1<\dots<t_n} \|(v(t_{i+1}) - v(t_i))_{1\le i\le n-1}\|_{l^p_w}, \|v\|_{sup}\}$$

is finite.

The spaces of bounded p variation are of considerable importance in probability and harmonic analysis. We shall see that V^p is the dual space of a space U^q, $1/p + 1/q = 1$, $1 < p < \infty$, with a duality pairing closely related to the Stieltjes integral, and its variant, the Young integral [22].

Definition 5. A p-atom a is a step function in $\mathscr{S}_{rc}$,

$$a(t) = \sum_{i=1}^{n} \phi_i \chi_{[t_i, t_{i+1})}(t)$$

where $\tau = (t_1 \dots t_n)$ is a partition, $t_{n+1} = b$, with $\sum |\phi_i|^p \leq 1$. A p-atom a is called a strict p atom if

$$\max \|\phi_i\|_X (\#\tau)^{1/p} \leq 1.$$

Let a_j be a sequence of atoms and λ_j a summable sequence. Then

$$u = \sum \lambda_j a_j$$

is a U^p function. The right hand side converges in $\mathscr{R}$. We define

$$\|u\|_{U^p} = \inf\{\sum |\lambda_j| : u = \sum \lambda_j a_j\}.$$

The strict space U^p_{strict} is defined in the same fashion using strict p atoms.

We collect a number of elementary properties.

1. If a is a p-atom then $\|a\|_{U^p} \leq 1$. In general the norm is less than 1.
2. Functions in U^p are continuous from the right. The limit as $t \to a$ vanishes.
3. The expression $\|.\|_{U^p}$ defines a norm on U^p, and U^p is closed with respect to this norm. Hence $U^P \subset \mathscr{R}_{rc}$ is a Banach subspace.
4. If $p < q$ then $U^p \subset U^q$ and

$$\|u\|_{U^q} \leq \|u\|_{U^p}.$$

5. If $1 \leq p < \infty$ then for all $u \in U^p$

$$\|u\|_{V^p} \leq 2^{1/p} \|u\|_{U^p}.$$

6. Let Y be a Banach space, and let the linear operator $T : \mathscr{S}_{rc} \to Y$ satisfy

$$\|Ta\|_Y \leq C \tag{3}$$

for every p atom. Then T has a unique extension to a bounded linear operator from U^p to Y which satisfies

$$\|Tf\|_Y \leq C \|f\|_{U^p}. \tag{4}$$

7. Let X_i be Banach spaces, $T : X_1 \times X_2 \to X_3$ a bounded bilinear operator, $v \in U^p(X_1)$ and $w \in U^p(X_2)$. Then $T(v,w) \in U^p(X_3)$ and

$$\|T(v,w)\|_{U^p(X_3)} \le 2\|T\|\|v\|_{U^p(X_1)}\|w\|_{U^p(X_2)}.$$

8. We consider $U^p([a,b))$ in the same way as subspace of $U^p(\mathbb{R})$ as for V^p.

The following decomposition is crucial for most of the following. It is related to Young's generalization of the Stieltjes integral, and it deals with a crucial point in the theory. A proof is contained in [11].

Lemma 1. *There exists $\delta > 0$ such that for v right continuous with $\|v\|_{V^p_w} = \delta$ there are strict p atoms a_i with*

$$\|a_j(t)\|_{sup} \le 2^{1-j} \qquad and \qquad \#\tau_j \le 2^{jp}$$

such that

$$v = \sum a_j.$$

There are a number of simple interesting and useful consequences.

Lemma 2. *Let $1 < p < q < \infty$. There exists $\kappa > 0$, depending only on p and q, such that for all $v \in V^p_{w,rc}$ and $M \ge 1$ there exist $u \in U^p_{strict}$ and $w \in U^q_{strict}$ with*

$$v = u + w$$

and

$$\frac{\kappa}{M}\|u\|_{V^p_{strict}} + e^M\|w\|_{U^q_{strict}} \le \|v\|_{V^p_w}.$$

Observe that we may replace U^p_{strict} by U^p (since $U^p_{strict} \subset U^p$) and V^p_w by V^p (since $V^p \subset V^p_w$). The proof is simple: We apply Lemma 1 and define u as the sum of the first m a_j. We obtain the following embedding

Lemma 3. *Let $1 < p < q < \infty$. Then*

$$V^p_{rc} \subset V^p_{w,rc} \subset U^q_{strict} \subset U^q.$$

Proof. Apply Lemma 2 with $M = 1$. □

The Riemann-Stieltjes integral defines

$$\int f dg = \int f g_t dt$$

for $f \in \mathscr{R}$ and $g \in V^1$. If $f \in \mathscr{S}_{rc}$ then

$$\int f g_t dt = \sum f(t_i)(g(t_i) - g(t_{i-1})). \tag{5}$$

We take this formula as our starting point for a similar integral for $f \in V^p$ and $g \in U^q$, for $1/p + 1/q = 1$, $q \geq 1$. Results become much cleaner when we use an equivalent norm in V^p,

$$\|v\|_{V^p} = \sup_{a<t_1 \ldots t_n<b} \left(\sum_j |v(t_{j+1}) - v(t_j)|^p + |v(t_n)|^p \right)^{1/p}$$

which we do in the sequel. We also set $v(b) = 0$ and, for any partition, $t_{n+1} = b$.

Theorem 1. *There is a unique continuous bilinear map*

$$B : U^q(X) \times V^p(X^*) \to \mathbb{R}$$

which satisfies (with $t_0 = a$ and $u(t_0) = 0$, and a somewhat sloppy notation for the duality map $X^* \times X \to \mathbb{R}$)

$$B(u, v) = \sum_{i=1}^{n} (u(t_i) - u(t_{i-1}))v(t_i)$$

for $u \in \mathscr{S}_{rc}$ with associated partition $(t_1, \ldots t_n)$ and

$$|B(u, v)| \leq \|u\|_{U^q(X)} \|v\|_{V^p(X^*)}. \tag{6}$$

The map

$$V^p(X^*) \ni v \to (u \to B(u, v)) \in (U^q(X))^*$$

is a surjective isometry. Moreover

$$\|v\|_{V^p(X^*)} = \sup_{u \in U^q(X), \|u\|_{U^q(X)}=1} B(u, v) = \sup_{a \text{ is a } q\text{-atom}} B(a, v). \tag{7}$$

The same statements are true if we replace U^p by U^p_{strict} and V^q by V^q_w.

See [3] for a proof. The previous results show that $U^p \subset V^p_{rc}$, and both spaces are very close. They are, however, not equal. The following example goes back to Young [22] with the same intention, but in a slightly different context. Let with a smooth function ϕ

$$v_p^N = \phi \sum_{j=1}^{N} 2^{-j/p} \sin(2^j x).$$

It is not hard to see that $\sup_N \|v_p^N\|_{C^{1/p}} < \infty$, and hence in $v_p^\infty \in C^{1/p} \subset V^p_{rc}$. Let $u_q = \phi \sum_{j=1}^\infty 2^{-j/q} \cos(2^j x)$. Now, assuming that $u_q \in U^q$,

$$\|u_q\|_{U^p} \|v_p^N\|_{V^q} \geq \left| \int (u_q^\infty)' v_p^N dx \right| = N/2 \int \phi^2 dx + O(1)$$

which is unbounded, hence a contradiction and $V^p_{rc} \ni u_p^\infty \notin U^p$.

Lemma 4. *For all $v \in V^p$ we have (recall Definition 2)*

$$\|v_\tau\|_{V^p(I)} \leq \|v\|_{V^p(I)} \tag{8}$$

and for all $u \in U^p$

$$\|u_\tau\|_{U^p(I)} \leq \|u\|_{U^p(I)}. \tag{9}$$

For $v \in V^p$ and $\varepsilon > 0$ there is a partition τ so that

$$\|v - v_\tau\|_{V^p} < \varepsilon. \tag{10}$$

Given $u \in U^p$ and $\varepsilon > 0$ there exists τ with

$$\|u - u_\tau\|_{U^p} < \varepsilon. \tag{11}$$

In particular $\mathscr{S}$ is dense in V^p and $\mathscr{S}_{rc}$ is dense in U^p.

Proof. When we take the supremum over partitions for v_τ we may restrict to subsets of τ and the first statement becomes obvious. For U^p it suffices to check p atoms a,

$$\|a_\tau\|_{U^p} \leq 1.$$

Density of step functions in U^p follows from the atomic definition of the space: Let $u \in U^p$ and $\varepsilon > 0$. By definition there exists a finite sum of atoms (which is a right continuous step function u_{step}) such that

$$\|u - u_{step}\|_{U^p} < \varepsilon/2.$$

Let τ be the step function associated to u_{step}. Then

$$\begin{aligned} \|u - u_\tau\|_{U^p} &\leq \|u_{step} - u_\tau\|_{U^p} + \|u - u_{step}\|_{U^p} \\ &< \|(u_{step} - u)_\tau\|_{U^p} + \varepsilon/2 < \varepsilon \end{aligned}$$

which is the claim for U^p. Let $\tilde{V}^p$ be the closure of the step functions in V^p. Suppose there exists $v \in V^p$ with distance 1 to $\tilde{V}^p$, and $\|v\|_{V^p} < 1 + \varepsilon$. Such a function exists when $\tilde V^p$ is not V^p. Let $D \subset U^q$ be the subset such $B(u,v)=0$ whenever $u \in D$ and $v \in \tilde V^p$. There exists $u \in D$ with $B(u,v)=1$, and a partition τ so that $\|u-u_\tau\|_{U^p} < \varepsilon$. However

$$0 = B(u,v_\tau) = B(u_\tau, v) = B(u,v) + B(u_\tau - u, v) \geq 1 - \varepsilon(1+\varepsilon)$$

which is a contradiction. Hence the step functions are dense in V^p. We complete the proof as for U^p. □

2.3 Embeddings

The first part of the next result it due to Hardy and Littlewood [4], and the second one follows by duality.

Lemma 5. *If* $1 < p < \infty$,

$$c_p^{-1}\|v\|_{\dot B^{1/p,p}_\infty} \leq \|v\|_{\tilde V^p} \leq 2^{1/p}\|u\|_{U^p} \leq c_p\|u\|_{\dot B_1^{1/p,p}}.$$

Let $\tilde V^p \subset V^p$ be the closed subspace of functions with

$$f(t) = \frac{1}{2}(\lim_{h\to 0}(f(t+h)+f(t-h))).$$

Choose a symmetric function $\phi \in L^1$ with $\int \phi = 1$ and $\phi_h(x) = h^{-1}\phi(x/h)$. The following claims can be easily verified for step functions, which suffices since they are dense.

Lemma 6. *Let* $a=-\infty$, $b=\infty$, $\phi \in L^1$ *symmetric with* $\int \phi dx = 1$. *We denote* $\phi_h(x) = h^{-1}\phi(x/h)$. *Then*

$$\phi_h * v \to v$$

*in the weak * topology for* $v \in \tilde V^p(\mathbb{R})$. *Moreover test functions are weak* dense in* V^p.

There is a second duality statement.

Lemma 7. *The bilinear map* B *defines a surjective isometry*

$$\tilde V^p(X^*) \to (U^q \cap C(X))^*, \frac{1}{p}+\frac{1}{q} = 1, 1 < p,q < \infty.$$

Proof. The kernel of the duality map restricted to $U^p \cap C(X)$ consists exactly of those elements of V^p which are nonzero at most at countably many points. Let $v \in \tilde{V}^p$. Then, by the previous lemma,

$$\phi_h * v \to v$$

in the weak $*$ topology of V^p. Moreover, for atoms

$$B(a, \phi_h * v) = B(\phi_h * a, v)$$

and hence this is true for functions in U^p. Now

$$\|v\|_{V^p} = \sup_{a\ q\text{-atom}} B(a,v) = \sup_a \lim_{h \to 0} B(a, \phi_h * v) = \sup \lim_{h \to 0} B(\phi_h * a, v) = \sup_a B(v,a).$$

It remains to prove surjectivity. Let $L : U^p \cap C(X) \to \mathbb{R}$ be linear. By the theorem of Hahn-Banach there is a extension with the same norm to U^p, and by duality there is $v \in V^q$ with $\|v\|_{V^q} = \|L\|$. Changing v at a countable set does not change the image in $(U^p \cap C(X))^*$, hence we may choose $v \in \tilde{V}^p$. □

We define

$$V_C^q = \{v \in V^q \cap C : \lim_{t \to a} v(t) = \lim_{t \to b} v(t) = 0\}. \tag{12}$$

Lemma 8. *The map*

$$U^p(X^*) \to (V_C^q(X))^*,$$

$$u \to (v \to B(u,v))$$

is a surjective isometry.

Proof. By the duality estimates the duality map is defined, and it is an isometry since the space V_C^q is weak star dense in V^q.

Let $L : V_C^q \to \mathbb{R}$. By Hahn-Banach there is an extension $\tilde{L}$ to V^q. We define (with obvious modifications for Banach space valued maps)

$$\tilde{u}(t) = -\tilde{L}(\chi_{[t,\infty)}).$$

As above we see that $(v \to B(\tilde{u}, v))$ coincides with $\tilde{L}$ on step functions. We define u as the unique right continuous function obtained by modifying $\tilde{u}$ at points of discontinuity. This does not change $B(u,.)$ on V_C^q. Moreover, by the definition of the quadratic form we may assume

$$\tilde{u}(t) \to 0 \qquad \text{as } t \to a.$$

Now $u \in U^{\tilde{p}}$ for all $p \geq 0$. The duality estimate allows to conclude that

$$\|u\|_{U^p} \leq \|L\|.$$

There is an immediate consequence. □

Lemma 9. *Test functions C_0^∞ are weak* dense in U^p.*

3 Dispersive Equations

3.1 Adapted Function Spaces

Here we briefly survey constructions going back to Bourgain, which have become standard. Details can be found in [11] and [3].

The following situation will be of particular interest. Let $t \to S(t)$ be a continuous unitary group on a Hilbert space H. We define U_S^p and V_S^p by

$$\|u\|_{U_S^p} = \|S(-t)u(t)\|_{U^p(H)}.$$

Now atoms are piecewise solutions. By Stone's theorem unitary groups are in one-one correspondence with selfadjoint operators, in the sense that

$$i\partial_t u = AU$$

with a self adjoint operator defines a unitary group $S(t)$ and vice versa. At least formally

$$i\partial_t(S(-t)u(t)) = S(-t)(i\partial_t u - Au)$$

and hence the duality assertion is

$$\|u\|_{U_S^q} = \sup_{\|v\|_{V_S^p} \leq 1} B(S(-t)u(t), S(-t)v(t)).$$

Now suppose that – again formally –

$$i\partial_t u - Au = f$$

then, by Duhamels formula, we obtain the solution

$$u(t) = \int_{-\infty}^{t} S(t-s)f(s)ds.$$

Thus,

$$\|u\|_{U^q_S} = \sup_{\|v\|_{V^p_S}\le 1} |B(S(-t)u(t), S(-t)v(t))| \tag{13}$$

$$= \sup_{\|v\|_{V^p_S}\le 1} \left| \int_{\mathbb{R}} \langle \partial_t S(-t)u(t), S(-t)v(t)\rangle dt \right| \tag{14}$$

$$= \sup_{\|v\|_{V^p_S}\le 1} |-i\langle S(-t)(i\partial_t u - Au), S(-t)v\rangle dt| \tag{15}$$

$$= \sup_{\|v\|_{V^p_S}\le 1} \int_{\mathbb{R}} \langle f, v\rangle dt \tag{16}$$

with a similar statement for V^p_S. This observation will be crucial for nonlinear dispersive equations. It is not hard to justify using our knowledge about weak* dense subspaces.

We want to use this construction for dispersive equations. There A is often defined by a Fourier multiplier, most often even by a partial differential operator with constant coefficients.

In order to be specific we consider the Airy equation – the situation would be similar for many other dispersive equations –

$$v_t + v_{xxx} = 0 \qquad \text{in } [0,\infty)$$

$$v(0) = u_0 \qquad \text{on } \mathbb{R}.$$

Let $v(t) = 0$ for $t < 0$ and the solution otherwise. Then

$$\|v\|_{V^1_{Airy}} = \|u_0\|_{L^2(\mathbb{R}^d)}.$$

There are three types of basic estimates: The *Strichartz estimate*

$$\|v\|_{L^p_t L^q_x} \le \||D|^{-1/p} u_0\|_{L^2} \tag{17}$$

whenever

$$\frac{2}{p} + \frac{1}{q} = \frac{1}{2}, \qquad 2 \le p, q. \tag{18}$$

Here $|D|^s$ is defined by the Fourier multiplier $|\xi|^s$. The Strichartz estimate quantifies the effect of dispersion.

The Strichartz estimate immediately transfers to estimates with respect to U^p_{Airy}:

$$\|v\|_{L^p_t L^q_x} \le c\| |D|^{-1/p} u\|_{U^p_{Airy}}. \tag{19}$$

It suffices to verify this if $S_{Airy}(-t)v$ is an atom with partition $(t_1, t_2 \dots t_n)$. Then, with $t_{n+1} = \infty$, by the Strichartz estimate we can estimate the mixed norm

$$\|v\|_{L^p_t((t_j,t_{j+1});L^q_x(\mathbb{R}))} \le c\| |D|^{-1/p} v(t_j)\|_{L^2(\mathbb{R})}.$$

We raise this to the pth power, and add over j. Then

$$\|v\|_{L^pL^q} \le c \left(\sum \| |D|^{-1/p} v(t_j)\|^p_{L^2} \right)^{1/p} \le c$$

since $S_{Airy}(-t)u$ is a p atom.

Consider now $v(t) = \int_{-\infty}^t S_{Airy}(t-s)f(s)ds$. By duality (13), and with p and q satisfying (18)

$$\begin{aligned}
\|v\|_{V^{p'}_{S_{Airy}}} &= \sup_{\|u\|_{U^p_{Airy}} \le 1} \left| B(S_{Airy}(-t)u, S_{Airy}(-t)v) \right| \\
&= \sup_{\|u\|_{U^p_{Airy}} \le 1} \left| \int u \bar{f} \, dx \, dt \right| \\
&\le \sup_{\|u\|_{U^p_{Airy}} \le 1} \|u\|_{L^pL^q} \|f\|_{L^{p'}L^{q'}} \\
&\le c\|f\|_{L^{p'}L^{q'}}.
\end{aligned}$$

This implies the dual estimate of (17). If $p > 2$ we may combine the estimates with an embedding to obtain the full Strichartz estimate.

Waves with different velocity interact at most in a time interval which is the inverse of the differences of the velocities. Bilinear estimates quantify this fact. For the Airy equation the group velocity is $-3\xi^2$. We define the Fourier projection u_λ by

$$\hat{u}_\lambda = \chi_{1 \le |\xi|/\lambda \le 2}(\xi)\hat{u}$$

where $\hat{u}$ denotes the Fourier transform with respect to x. For solutions to the Airy equation we obtain the estimate

$$\|u_\lambda u_\mu\|_{L^2(\mathbb{R}^2)} \le c\mu^{-1} \|u_\lambda(0)\|_{L^2(R)} \|u_\mu(0)\|_{L^2(\mathbb{R})} \tag{20}$$

provided $\lambda \leq \mu/4$. Again this implies for functions in U^2

$$\|u_\lambda v_\mu\|_{L^2} \leq c\mu^{-1}\|u_\lambda\|_{U^2_{Airy}}\|u_\mu\|_{U^2_{Airy}}. \tag{21}$$

The imbedding estimate (5) immediately implies the high modulation estimate

$$\|u^\Lambda\|_{L^2(\mathbb{R}^2)} \leq c\Lambda^{-1/2}\|u\|_{V^2_{Airy}} \tag{22}$$

where u^Λ is defined by the space-time Fourier multiplier $\chi_{|\tau-\xi^3|>\Lambda}(\tau,\xi)$.

This set of tools is complemented by the interpolation estimate (2).

3.2 The Generalized KdV Equation

For integers $p \geq 1$ we consider the initial value problems

$$u_t + u_{xxx} + (u^p u)_x = 0 \tag{23}$$

$$u(0) = u_0 \tag{24}$$

– the case $p = 1$ is the Korteweg-de-Vries equation, and $p = 2$ the modified Korteweg-de-Vries equation, and

$$u_t + u_{xxx} + (|u|^p u)_x = 0 \tag{25}$$

$$u(0) = u_0 \tag{26}$$

for positive real p.

Both equations have soliton solutions. They are invariant with respect to scaling: $\lambda^{2/p}u(\lambda x, \lambda^3 t)$ is a solution if u satisfies the equation. The mass $\int u^2 dx$ and energy $\int \frac{1}{2}u_x^2 - \frac{1}{p+2}u^{p+2}$ are conserved. The energy however is not bounded from below. The space $\dot{H}^{\frac{1}{2}-\frac{2}{p}}$ (with norm $\|v_0\|_{\dot{H}^s} = \||\xi|^s\hat{v}_0\|_{L^2}$) is invariant with respect to this scaling and it is not hard to see that the generalized *KdV* equation is globally wellposed in H^1 if $p < 4$. For $p \geq 4$ one expects blow-up. This has been proven in series of seminal papers by Martel, Merle and Martel, Merle and Raphael for $p = 4$, see [15–17] and the references therein.

The most prominent equation here is the KdV equation

$$u_t + u_{xxx} + (u^2)_x = 0.$$

The tools described here allow an alternative argument to prove local wellposedness in $H^{-3/4}(\mathbb{R})$. The order of derivatives cannot be improved by contraction arguments. There are however apriori estimates in H^{-1} by different techniques, [1]. For the modified KdV equation

$$u_t + u_{xxx} + (u^3)_x = 0$$

one obtains local wellposedness in $H^{1/4}$, which again is optimal in terms of the number of derivatives.

For the quartic KdV equation

$$u_t + u_{xxx} + (u^4)_x = 0$$

one obtains global existence by a contraction argument in the space

$$\|u\|_X = \sup_\lambda \lambda^{-1/6} \|u_\lambda\|_{U^2_{Airy}}$$

for initial data in a Besov space $\dot{B}^{-1/6}_{2,\infty}(\mathbb{R})$. Statement and proof are contained in [10], where it was one step to prove stability of the soliton in $\dot{B}^{-1/6,2}_{\infty}$, and scattering. This is probably the first stability result of solitons for gKdV which is not based on Weinstein's convexity argument in the energy space.

Wellposedness in a slightly smaller spaces has been proven by Grünrock [2] and Tao [20] based on a modification of the Fourier restriction spaces of Bourgain at the critical level.

The quintic KdV equation

$$u_t + u_{xxx} + u^5_x = 0$$

is of particular interest since it is L^2 critical. Since the work by Kenig, Ponce and Vega it is known to be locally wellposed in L^2. The local existence result has been extended to all equations

$$u_t + u_{xxx} + |u|^p u_x = 0$$

with $p \geq 4$ in [19] in critical function spaces using the techniques above. The case of polynomial (analytic) nonlinearities had been dealt with by Molinet and Ribaud [18] using different techniques.

3.3 The Kadomtsev-Petviashvili II Equation

The Kadomtsev-Petviashvili-II (KP-II) equation

$$\partial_x(\partial_t u + \partial_x^3 u + u\partial_x u) + \partial_y^2 u = \ 0 \qquad \text{in } (0,\infty) \times \mathbb{R}^2 \qquad (27)$$

$$u(0,x,y) = u_0(x,y) \qquad (x,y) \in \mathbb{R}^2 \qquad (28)$$

has been introduced by Kadomtsev and Petviashvili [9] to describe weakly transverse water waves in the long wave regime with small surface tension. It generalizes the Korteweg – de Vries equation, which is spatially one dimensional and thus neglects transversal effects. The KP-II equation has a remarkably rich structure.

Here we describe a setup leading to global wellposedness and scattering for small data. The Hilbert space will be denoted by $\dot{H}^{-1/2,0}$ which is defined by through the norm

$$\|u_0\|_{\dot{H}^{-1/2}} = \| |\xi|^{-1/2} \hat{u}_0 \|_{L^2}$$

where ξ is the Fourier multiplier with respect to x. The Fourier multiplier $|\xi|^{-1/2}$ defines an isomorphism from L^2 to $\dot{H}^{-1/2}$.

For $\lambda > 0$ we define the Fourier projection to the $1 \leq |\xi|/\lambda < 2$ by

$$\hat{u}_\lambda = \chi_{\lambda \leq |\xi| \leq 2\lambda} \hat{u}$$

where $\mathscr{F}$ denotes the Fourier transform and ξ the Fourier variable of x. Usually we choose $\lambda \in 2^{\mathbb{Z}}$, the set of integer powers of 2. We define X by

$$\|u\|_X = \left(\sum_{\lambda \in 2^{\mathbf{Z}}} (\lambda^{-1/6} \|u_\lambda\|_{V^2_{KP}})^2 \right)^{1/2} .$$

The following theorem has been proven in [3] with a proof relying on the space U^2 and V^2.

Theorem 2. *There exists $\delta > 0$ such that for all u_0 with $\|u_0\|_{\dot{H}^{-1/2}}$ there exists a unique solution*

$$u \in X \subset C([0,T]; \dot{H}^{-\frac{1}{2},0}(\mathbb{R}^2))$$

of the KP-II equation (27) on $(0,\infty)$. Moreover, the flow map

$$B_{\delta,R}(0) \ni u_0 \mapsto u \in X$$

is analytic.

A duality argument reduces the proof to an estimate of a trilinear integral. The functions there are expanded according to the Fourier projection, and the key estimate is a bound for

$$\left| \int u_{\lambda_1} v_{\lambda_2} w_{\lambda_3} dx\, dy\, dt \right|$$

Due to symmetry we may assume that $\lambda_1 \leq \lambda_2 \leq \lambda_3$. Since there is only a contribution to the integral if there a point in the support of the Fourier transforms adding up to zero we can only get a contribution if $\lambda_2 \sim \lambda_3$, and only λ_1 may be smaller. Since

$$\begin{aligned}\tau_1 - \xi_1^3 + \eta_1^2/\xi_1 + \tau_2 - \xi_2^3 + \eta_2^2/\xi_2 & \\ -[(\tau_1+\tau_2) - (\xi_1+\xi_2)^3 + (\eta_1+\eta_2)^2/(\xi_1+\xi_2)] & \\ = -3\xi_1\xi_2(\xi_1+\xi_2) - \frac{(\xi_1\eta_2 - \xi_2\eta_1)^2}{\xi_1\xi_2(\xi_1+\xi_2)} &\end{aligned}$$

there is only a contribution if at least for one $j \in \{1,2,3\}$

$$|\tau_j - \xi_j^3 + \eta_j^2/\xi_j| \ge |\xi_1\xi_2(\xi_1+\xi_2)|,$$

$j = 1,2,3$ and $\tau_3 = -\tau_1 - \tau_2, \xi_3 = -\xi_1 - \xi_2, \eta_3 = -\eta_1 - \eta_2$ in the support of the Fourier transforms. We set $\Lambda = \lambda_1\lambda_2\lambda_3/10$ and expand $u_{\lambda_1} = u^{\Lambda}_{\lambda_1} + u^{low}_{\lambda_1}$, where u^Λ is defined by the space-time Fourier multiplier

$$\chi_{|\tau-\xi^3+\eta^2/\xi|\ge\Lambda},$$

and similarly we decompose the other factors. We expand the trilinear integral. There is only a nontrivial contribution if at least one of the terms $u^\Lambda_{\lambda_1}$, $v^\Lambda_{\lambda_2}$ or $w^\Lambda_{\lambda_3}$ occurs. We apply Cauchy-Schwartz and estimate the corresponding term in L^2. For the other product we apply an L^4 space-time Strichartz estimate, or a bilinear estimate.

We obtain

$$\left|\int u_{\lambda_1} v_{\lambda_2} w^\Lambda_{\lambda_3}\, dx\, dy\, dt\right| \le c(\lambda_1\lambda_2\lambda_3)^{-1/2}(\lambda_{min}/\lambda_{max})^{\frac{1}{4}} \|u_{\lambda_1}\|_{V^2_{KP}} \|v_{\lambda_2}\|_{V^2_{KP}} \|w_{\lambda_3}\|_{V^2_{KP}}$$

and

$$\left|\sum_{\lambda_1\le\lambda_2}\int u_{\lambda_1} v_{\lambda_2} w^\Lambda_{\lambda_3}\, dx\, dy\, dt\right| \le c\lambda_{max}^{-1}\|u\|_X \|v_{\lambda_2}\|_{V^2} \|w_{\lambda_2}\|_{V^2}$$

which suffices to conclude the proof.

3.4 The Energy Critical Nonlinear Schrrödinger Equation on Compact Manifolds

We consider the quintic nonlinear Schrödinger equation on the three dimensional torus $\mathbb{T}^3$, either focusing or defocusing

$$i\partial_t u + \Delta u = \pm|u|^4 u. \qquad (29)$$

On $\mathbb{R}^3$ the space $\dot{H}^1$ is critical. We consider solutions on a unit time interval with small initial data in H^1. We define the function space X by

$$\|u\|_X = \left\| (1+k_1^2+k_2^2+k_3^3)^{1/2} \|e^{-it\pi^2(k_1^2+k_2^2+k_3^3)}\hat{u}\|_{V^2(0,1)} \right\|_{l^2_{\mathbf{k}}(\mathbb{Z}^3)} .$$

The following depends on a mix of previous arguments, and estimates for Gaussian sums.

Theorem 3 ([6]). *There exists $\delta > 0$ such that given $u_0 \in H^1$ with $\|u_0\|_{H^1} < \delta$ there exists a unique solution $u \in X$. This solution can be extended to a global solution in time. The map from initial data to solution is real analytic. If $u_0 \in H^1$ there is a local in time solution.*

This result has been extended to global wellposedness on T^3 for large data in H^1 by Ionescu and Pausader [8].

A similar mix of adapted function spaces, eigenfunction estimates and bounds on Gaussian sums has been applied by the same authors to energy critical partial periodic domains in $\mathbb{R}^4$ [7], and by Herr to the quintic Schrödinger equation on the three dimensional sphere [5].

References

1. Buckmaster, T., Koch, H.: The Korteweg-de-Vries equation at H^{-1} regularity (2012). arXiv:1112.4657v2
2. Grünrock, A.: A bilinear airy-estimate with application to gKdV-3. Differ. Integral Equ. **18**(12), 1333–1339 (2005)
3. Hadac, M., Herr, S., Koch, H.: Well-posedness and scattering for the KP-II equation in a critical space. Ann. Inst. H. Poincaré Anal. Non Linéaire **26**(3), 917–941 (2009)
4. Hardy, G.H., Littlewood, J.E.: A convergence criterion for Fourier series. Math. Z. **28**(1), 612–634 (1928)
5. Herr, S.: The quintic nonlinear Schrödinger equation on three-dimensional Zoll manifolds (2011). arXiv:1101.4565
6. Herr, S., Tataru, D., Tzvetkov N.: Global well-posedness of the energy-critical nonlinear Schrödinger equation with small initial data in $H^1(\mathbb{T}^3)$. Duke Math. J. **159**(2), 329–349 (2011)
7. Herr, S., Tataru, D., Tzvetkov N.: Strichartz estimates for partially periodic solutions to Schrödinger equations in 4d and applications (2011). arXiv:1101.0591
8. Ionescu, A.D., Pausader, B.: Global well-posedness of the energy-critical defocusing NLS on $\mathbb{R} \times \mathbb{T}^3$. Commun. Math. Phys. **312**(3), 781–831 (2012)
9. Kadomtsev, B.B., Petviashvili, V.I.: On the stability of solitary waves in weakly dispersing media. Sov. Phys. Dokl. **15**, 539–541 (1970) (English. Russian original)
10. Koch, H., Marzuola, J.L.: Small data scattering and soliton stability in $\dot{H}^{-1/6}$ for the quartic KdV equation. Anal. PDE **5**(1), 145–198 (2012)
11. Koch, H., Tataru D.: Dispersive estimates for principally normal pseudodifferential operators. Commun. Pure Appl. Math. **58**(2), 217–284 (2005)
12. Koch, H., Tataru, D.: A priori bounds for the 1D cubic NLS in negative Sobolev spaces. Int. Math. Res. Not. **36**(16), 1–36 (2007) Art. ID rnm053

13. Lyons, T.: Differential equations driven by rough signals. I. An extension of an inequality of L. C. Young. Math. Res. Lett. **1**(4), 451–464 (1994)
14. Lyons, T.J.: Differential equations driven by rough signals. Rev. Mat. Iberoam. **14**(2), 215–310 (1998)
15. Martel, Y., Merle, F., Raphael, P.: Blow up for the critical gKdV equation I: dynamics near the soliton (2012). arXiv:1204.4625
16. Martel, Y., Merle, F., Raphael, P.: Blow up for the critical gKdV equation II: minimal mass dynamics (2012). arXiv:1204.4625
17. Martel, Y., Merle, F., Raphael, P.: Blow up for the critical gKdV equation III: exotic regimes (2012). arXiv:1209.2510
18. Molinet, L., Ribaud, F.: On the Cauchy problem for the generalized Korteweg-de Vries equation. Commun. Partial Differ. Equ. **28**(11–12), 2065–2091 (2003)
19. Strunk, N.: Well-posedness for the supercritical gKdV equation (2012). arXiv:1209.5206
20. Tao, T.: Scattering for the quartic generalised Korteweg-de Vries equation. J. Differ. Equ. **232**(2), 623–651 (2007)
21. Wiener, N.: The quadratic variation of a function and its Fourier coefficients. J. Mass. Inst. Technol. **3**, 73–94 (1924)
22. Young, L.C.: An inequality of the Hölder type, connected with Stieltjes integration. Acta Math. **67**(1), 251–282 (1936)

A Note on Metastable Behaviour in the Zero-Range Process

Anton Bovier and Rebecca Neukirch

Abstract The zero-range process in the high density phase is known to show condensation behaviour, i.e., a macroscopic fraction of particles is localised on a single site under the canonical equilibrium measure. Recently, Beltrán and Landim (Probab Theory Relat Fields 152(3–4):781–807, 2012) analysed some aspects of the metastable behaviour of this process in one dimension for finite systems in the limit of infinite density. In this note we show that the potential theoretic approach to metastability initiated in Bovier et al. (Commun Math Phys 228(2):219–255, 2002) applies easily to this model and yields more detailed results.

1 Introduction

In a recent paper [2], Beltrán and Landim studied the metastable behaviour of the zero-range process [13] in one dimension on a finite box, $S = \{1, \dots, L\}$, in the limit where the number of particles tends to infinity. In this short note, we improve their results using the methods of the so-called *potential theoretic approach* to metastability, put forward in [8] also in the case that S is infinite. In particular, we show that the model, considered in the limit, fits perfectly into the (simplest instance) of that approach, that the definition of metastability given there applies, and that the abstract results of that paper provide the usual sharp estimates on mean metastable exit times and their exponential distribution. In addition, we show

A. Bovier (✉) · R. Neukirch
Institut für Angewandte Mathematik, Rheinische Friedrich-Wilhelms-Universität Bonn, Endenicher Allee 60, D-53115 Bonn, Germany
e-mail: bovier@uni-bonn.de; rebecca.neukirch@iam.uni-bonn.de

M. Griebel (ed.), *Singular Phenomena and Scaling in Mathematical Models*,
DOI 10.1007/978-3-319-00786-1_4,

that some of the results can be extended to the case when $L = L(N) \uparrow \infty$ but $L(N)/N \downarrow 0$.

1.1 The Model and Basic Properties

The zero-range process $\eta(t)$ is a continuous time Markov process on the state space

$$E_{N,S} = \left\{ \eta \in \mathbb{N}^S : \sum_{x=1}^{L} \eta_x = N \right\}, \tag{1}$$

where $\eta_x \in \mathbb{N}$ represents the number of particles at site $x \in S$. At any time a particle at site x jumps to a site y with a rate of $g(\eta_x(t))r(x,y)$, where $r(\cdot,\cdot)$ is an irreducible transition probability of a reversible random walk $X(t)$ on S. Here, g is chosen as

$$g(0) = 0, \quad g(1) = 1 \quad \text{and} \quad g(n) = \frac{a(n)}{a(n-1)}, \quad n \geq 2, \tag{2}$$

with $a(0) = 1$ and $a(n) = n^\alpha$, for $\alpha > 1$. Formally, the process can be defined through its generator, L_N, that acts on functions, $F \in C(E_{N,S}, \mathbb{R})$, via

$$L_N(F)(\eta) = \sum_{x=1}^{L} \sum_{y \in S} g(\eta_x) r(x,y) \big[F(\eta^{x,y}) - F(\eta) \big]. \tag{3}$$

Here $\eta^{x,y}$ is the configuration obtained from η by moving a particle on x to the position y.

The zero-range process is irreducible and reversible with respect to its unique invariant probability measure (see [1, 9, 11]) given by

$$\mu_N(\eta) = \frac{N^\alpha}{Z_{N,S}} \prod_{x \in S} \frac{m_*(x)^{\eta_x}}{a(\eta_x)} \equiv \frac{N^\alpha}{Z_{N,S}} \frac{m_*^\eta}{a(\eta)}, \qquad \eta \in E_{N,S} \tag{4}$$

where $m_*(x) = \frac{m(x)}{M_*}$ with $M_* = \max\{m(x) \mid x \in S\}$. Furthermore, m denotes here the invariant measure of the random walk $X(t)$. Note that $m_*(x) = 1$, for all $x \in S_* \equiv \{x \in S \mid m(x) = M_*\}$. $Z_{N,S}$ is the normalizing partition function,

$$Z_{N,S} = N^\alpha \sum_{\zeta \in E_{N,S}} \frac{m_*^\zeta}{a(\zeta)}. \tag{5}$$

A first observation is the following lemma [2, 11].

Lemma 1. *If $L < \infty$ is independent of N, then*

$$\lim_{N\uparrow\infty} Z_{N,S} = Z_S = \frac{|S_*|}{\Gamma(\alpha)} \prod_{z\in S} \Gamma_z = |S_*|\Gamma(\alpha)^{|S_*|-1} \prod_{y\notin S_*} \Gamma_y, \tag{6}$$

where

$$\Gamma_x \equiv \sum_{j\geq 0} \frac{m_*(x)^j}{a(j)} \quad \textit{and} \quad \Gamma(\alpha) \equiv \sum_{j\geq 0} \frac{1}{a(j)}. \tag{7}$$

The interesting feature of this model is that it exhibits a *condensation phenomenon* [10–12], namely the zero-range process shows condensation in the sense that asymptotically, the invariant measure concentrates on disjoint sets of configurations, $\mathscr{E}_N^x$, described as follows: Fix a sequence $\{\ell_N : N \geq 1\}$, where $1 \ll \ell_N \ll N$ such that

$$\lim_{N\uparrow\infty} \ell_N = \infty \qquad \text{and} \qquad \lim_{N\uparrow\infty} \frac{\ell_N}{N} = 0. \tag{8}$$

We say that a configuration has a condensate at site $x \in S_*$ if it belongs to the set

$$\mathscr{E}_N^x \equiv \{\eta \in E_N : \eta_x \geq N - \ell_N\}. \tag{9}$$

A result of Großkinsky et al. ([11], Theorem 2) implies the following:

Theorem 1. *Assume that $L(N)/N \downarrow 0$. Then one can choose ℓ_N such that*

$$\lim_{N\uparrow\infty} \mu_N \left(\bigcup_{x\in S_*} \mathscr{E}_N^x \right) = 1. \tag{10}$$

Note in particular that in the case when L is independent of N, the configurations η^x where $\eta_x^x = N$ have positive measure, and indeed they are the configuration with maximal measure.

The question addressed in [2] and here is how to describe the motion of the systems between different condensate configurations.

2 Metastability

In this section we recall some basic facts about the potential theoretic approach to metastability [5,6,8] and show in the next section how to apply this in the zero-range process.

2.1 Definitions and Results

Consider a reversible Markov process $X(t)$ with discrete state space E, generator L, and reversible measure μ. For $A, B \subset E$ we define the *equilibrium potential* $h_{A,B}$ as the solution of the Dirichlet problem

$$\begin{aligned} -(Lh)(\eta) &= 0, \quad \eta \in (A \cup B)^c \\ h(\eta) &= 1, \quad \eta \in A \\ h(\eta) &= 0, \quad \eta \in B. \end{aligned} \tag{11}$$

The *equilibrium measure* $e_{A,B}$ on A is given by

$$e_{A,B}(\eta) \equiv -(Lh_{A,B})(\eta), \quad \eta \in A. \tag{12}$$

The *capacity* of the pair (A, B) is then defined as

$$\text{cap}(A, B) \equiv \sum_{\eta \in A} \mu(\eta) e_{A,B}(\eta). \tag{13}$$

These objects have probabilistic content. Namely,

$$h_{A,B}(\eta) = \mathbb{P}_\eta (\tau_A < \tau_B), \quad \eta \in (A \cup B)^c \tag{14}$$

and

$$e_{A,B}(\eta) = \mathbb{P}_\eta (\tau_B < \tau_A), \quad \eta \in A. \tag{15}$$

Here τ_A denotes the stopping times $\tau_A = \inf\{t > 0 : X(t) \in A\}$.

The importance of capacities for the analysis of metastability is due to the representation of mean hitting times (see e.g. [5])

$$\mathbb{E}_{\nu_{A,B}}[\tau_B] = \sum_{\xi \in E \setminus B} \frac{\mu(\xi)}{\text{cap}(A, B)} h_{A,B}(\xi), \tag{16}$$

for $\nu_{A,B}$, the so-called *last exit measure* on A, namely for $\eta \in A$,

$$\nu_{A,B}(\eta) \equiv \frac{\mu(\eta) e_{A,B}(\eta)}{\text{cap}(A, B)}. \tag{17}$$

In the case when $A = \{\eta\}$ is a single point, this simplifies to

$$\mathbb{E}_\eta[\tau_B] = \sum_{\xi \in E \setminus B} \frac{\mu(\xi)}{\text{cap}(\eta, B)} h_{\eta,B}(\xi). \tag{18}$$

This latter formula is useful in the case when L is independent of N.

Dirichlet's principle gives a variational characterisation of capacities. For two disjoint subsets $A, B \subset E_{N,S}$, define the set of functions

$$\mathscr{H}(A, B) = \{F : E \uparrow \mathbb{R}_+ \mid F(\eta) = 1,\ \forall \eta \in A,\ F(\eta) = 0,\ \forall \eta \in B\}. \tag{19}$$

Then

$$\operatorname{cap}(A, B) = \inf_{F \in \mathscr{H}(A,B)} \Phi(F), \tag{20}$$

where $\Phi(F) \equiv (F, LF)_\mu$ is the Dirichlet form of the process.

In [8], metastability with respect to a set of points was characterized as follows:

Definition 1. A Markov process $X(t)$ is metastable with respect to a set $\mathscr{M}$, if

$$\frac{\sup_{\eta \in \mathscr{M}} \operatorname{cap}(\eta, \mathscr{M} \backslash \eta)/\mu(\eta)}{\inf_{\xi \in \mathscr{M}^c} \operatorname{cap}(\xi, \mathscr{M})/\mu(\xi)} \le \rho \ll 1. \tag{21}$$

Definition 2. For all $\eta \in \mathscr{M}$, the *valley* $A(\eta)$ of the attractor η is given by

$$A(\eta) = \{\xi \in E : \mathbb{P}_\xi(\tau_\eta = \tau_{\mathscr{M}}) = \sup_{\zeta \in \mathscr{M}} \mathbb{P}_\xi(\tau_\zeta = \tau_{\mathscr{M}})\}. \tag{22}$$

From [8] we know:

Theorem 2. *Consider a Markov process $X(t)$ which is metastable with respect to $\mathscr{M}$. For every $\eta \in \mathscr{M}$ we have that*

(i)

$$\mathbb{E}_\eta[\tau_{\mathscr{M} \backslash \eta}] = \frac{\mu(A(\eta))}{\operatorname{cap}(\eta, \mathscr{M} \backslash \eta)}(1 + o(1)), \tag{23}$$

(ii) For $t > 0$

$$\mathbb{P}_\eta\left[\tau_{\mathscr{M} \backslash \eta} > t\mathbb{E}_\eta[\tau_{\mathscr{M} \backslash \eta}]\right] = e^{-t(1+o(1))}(1 + o(1)). \tag{24}$$

3 Results for the Zero-Range Process

Finite L. First, we state the results for the zero-range process when $L < \infty$ is fixed and $N \uparrow \infty$.

Proposition 1. *The zero-range process $\{\eta(t) : t \ge 0\}$ is metastable with respect to $\mathscr{M} = \bigcup_{x \in S_*} \eta^x$ and $\rho = \mathscr{O}(r'^{-1} L(L^2 + N)N^{-\alpha-1})$, where $r' \equiv \inf_{u \in S} r(u, u \pm 1)$.*

To analyse the metastable behavior of the zero-range process we only need to compute capacities between configurations (cf. Definition 1).

For $0 \leq s \leq \frac{1}{2}$ define

$$I_\alpha(s) \equiv \int_s^{1-s} u^\alpha (1-u)^\alpha du, \tag{25}$$

and for $S'_* \subseteq S_*$, $S'_* \neq \emptyset$ set

$$\eta^{S'_*} \equiv \bigcup_{x \in S'_*} \eta^x. \tag{26}$$

Theorem 3. *Assume $L < \infty$ independent of N. Let $S^1_* \subsetneq S_*$ and $S^2_* \subseteq S_* \backslash S^1_*$ be two nonempty sets. We have that*

$$\begin{aligned} \mathrm{cap}_N(\eta^{S^1_*}, \eta^{S^2_*}) = & \frac{N^{-\alpha-1}}{2M_*|S_*|\Gamma(\alpha)I_\alpha(0)}(1+o(1)) \\ & \times \inf_{W \in \mathscr{W}(S^1_*, S^2_*)} \sum_{x,y \in S_*} \mathrm{cap}_S(x,y)[W_y - W_x]^2, \end{aligned} \tag{27}$$

where $\mathscr{W}(S^1_*, S^2_*) = \{W \in [0,1]^{S_*} | W_z = 1, \forall z \in S^1_* \text{ and } W_z = 0, \forall z \in S^2_*\}$.

Remark 1. Note that the second line in (27) is the conductance between S^1_* and S^2_* of a resistor network on S_* with conductances $\mathrm{cap}_S(x,y)$ between nodes in S_*.

These results allow to use Theorem 2 to obtain the following corollary for the metastable exit times:

Corollary 1. *For a zero-range process $\{\eta(t) : t \geq 0\}$ with metastable set $\mathscr{M}$ we have for every $\eta^x \in \mathscr{M}$*

(i) The metastable mean exit time is given by

$$\mathbb{E}_{\eta^x}\left[\tau_{\mathscr{M}\backslash\eta^x}\right] = \frac{N^{\alpha+1}M_* I_\alpha(0)\Gamma(\alpha)}{\sum_{y \in S_* \backslash \{x\}} \mathrm{cap}_S(x,y)}(1+o(1)), \tag{28}$$

(ii) For $t > 0$ the metastable exit time is exponentially distributed

$$\mathbb{P}_{\eta^x}\left[\tau_{\mathscr{M}\backslash\eta^x} > t\mathbb{E}_{\eta^x}[\tau_{\mathscr{M}\backslash\eta^x}]\right] = e^{-t(1+o(1))}(1+o(1)). \tag{29}$$

Remark 2. Combining the remark above with this corollary, one sees that in the limit of large N, on the time-scale $N^{1+\alpha}$, the zero range process observed on the set $\{\eta^x, x \in S_*\}$ behaves like a continuous time random walk with transition rates $M_* I_\alpha(0)\Gamma(\alpha)\mathrm{cap}_S(x,y)/\sum_{z \in S_* \backslash \{x\}} \mathrm{cap}_S(x,z)$. This is a different version of a similar statement in [2].

Diverging L. In the case where $L = L(N) \uparrow \infty$ in such a way that $L(N)/N \downarrow 0$ we get weaker results. Note that in this case, the transitions rates $r(x, y)$ will in general depend on L, and hence on N. Also, the sets S_* will typically depend on L. We suppress these dependences to simplify the notation. We define for $S'_* \subseteq S_*$, $S'_* \neq \emptyset$

$$\mathscr{E}_N(S'_*) = \bigcup_{x \in S'_*} \mathscr{E}_N^x \quad \text{and} \quad \mathscr{E}_N = \mathscr{E}_N(S_*). \tag{30}$$

In the general case, we can only show that the lower bound of capacity between sets $\mathscr{E}_N(S^1_*)$ and $\mathscr{E}_N(S^2_*)$ coincides with the upper bound up to a constant. But for disjoint partitions of S_* we get the following theorem:

Theorem 4. *Assume that $|S_*| \geq 2$ and $L(N) \uparrow \infty$ such that $L(N)/N \downarrow 0$. Let $S^1_* \subsetneq S_*$ and $S^2_* = S_* \backslash S^1_*$ be two nonempty sets. Then*

$$\mathrm{cap}_N(\mathscr{E}_N(S^1_*), \mathscr{E}_N(S^2_*)) = \frac{N^{-\alpha-1}}{M_*|S_*|\Gamma(\alpha)I_\alpha(0)} \sum_{x \in S^1_*, y \in S^2_*} \mathrm{cap}_S(x, y)(1 + o(1)). \tag{31}$$

In view of the representation of mean hitting times (16), we get:

Corollary 2. *Let $L(N) \uparrow \infty$ such that $L(N)/N \downarrow 0$. The metastable exit time for the zero-range process from a set $\mathscr{E}_N^x$, where $x \in S_*$, is given by*

$$\mathbb{E}_{\nu_{\mathscr{E}_N^x, \mathscr{E}_N \backslash \mathscr{E}_N^x}} \left[\tau_{\mathscr{E}_N \backslash \mathscr{E}_N^x} \right] = \frac{N^{\alpha+1} M_* \Gamma(\alpha) I_\alpha(0)}{\sum_{y \in S_* \backslash \{x\}} \mathrm{cap}_S(x, y)} (1 + o(1)). \tag{32}$$

Remark 3. One would like to show that the assertion of the corollary also holds for the process starting in a single configuration η^x, and that the law of the exit time is exponential. Such a result has been obtained in a different model [3] using coupling techniques, but this seems difficult in the present case.

4 Proofs of the Results

4.1 Capacity

We start with the proof of Theorem 3. For the proof we calculate lower and upper bounds for capacities which coincide in the limit $N \uparrow \infty$ and L fixed.

4.1.1 Lower Bound

We begin by proving an a priori bound on the equilibrium potential that shows that it is almost constant on the sets $\mathscr{E}_N^x$, $x \in S_*$.

Lemma 2. *Let $S_*^i \subset S_*$ be disjoint for $i \in \{1,2\}$ and let W denote the equilibrium potential for the capacitor $\eta^{S_*^1}, \eta^{S_*^2}$. Then there is a constant K_α, such that for any $z \in S_*$, and $\xi, \xi' \in \mathscr{E}_N^z$,*

$$|W(\xi) - W(\xi')| \leq K_\alpha \frac{L(L^2+N)}{N^{1+\alpha} r'}, \tag{33}$$

where $r' \equiv \inf_{u \in S} r(u, u \pm 1)$.

Proof. Clearly

$$\begin{aligned} W(\xi) &= \mathbb{P}_\xi[\tau_{\eta S_*^1} < \tau_{\eta S_*^2}] \\ &= \mathbb{P}_\xi[\tau_{\xi'} < \tau_{\eta S_*^1}, \tau_{\eta S_*^1} < \tau_{\eta S_*^2}] + \mathbb{P}_\xi[\tau_{\eta S_*^1} < \tau_{\eta S_*^2}, \tau_{\xi'} > \tau_{\eta S_*^1}] \\ &= \mathbb{P}_\xi[\tau_{\xi'} < \tau_{\eta S_*^1}]\mathbb{P}_{\xi'}[\tau_{\eta S_*^1} < \tau_{\eta S_*^2}] + \mathbb{P}_\xi[\tau_{\eta S_*^1} < \tau_{\xi' \cup \eta S_*^2}] \\ &= (1 - \mathbb{P}_\xi[\tau_{\eta S_*^1} < \tau_{\xi'}])W(\xi') + \mathbb{P}_\xi[\tau_{\eta S_*^1} < \tau_{\xi' \cup \eta S_*^2}]. \end{aligned} \tag{34}$$

Using the renewal equation (Lemma 4.1 in [4]), this yields

$$\left(1 - \frac{\mathbb{P}_\xi[\tau_{\eta S_*^1} < \tau_\xi]}{\mathbb{P}_\xi[\tau_{\xi'} < \tau_\xi]}\right) W(\xi') \leq W(\xi) \leq W(\xi') + \frac{\mathbb{P}_\xi[\tau_{\eta S_*^1} < \tau_\xi]}{\mathbb{P}_\xi[\tau_{\xi'} < \tau_\xi]}. \tag{35}$$

In Sect. 4, Proposition 4 below, we show that

$$\frac{\mathbb{P}_\xi[\tau_{\eta S_*^1} < \tau_\xi]}{\mathbb{P}_\xi[\tau_{\xi'} < \tau_\xi]} \leq K_\alpha r'^{-1} L(L^2+N) N^{-1-\alpha}, \tag{36}$$

which implies the assertion of the lemma. □

Remark 4. This result is most useful if L is fixed and N tends to infinity, but it also allows to push these results to cases of slowly growing $L = L(N)$.

We first prove a lower bound for given values of N and L.

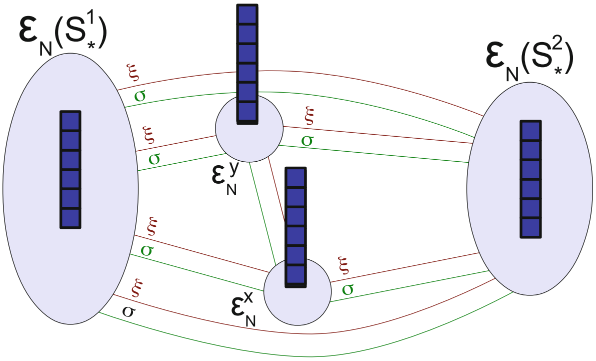


Fig. 1 Restriction of the transition rates

Proposition 2. *Assume that* $|S_*| \geq 2$. *Let* $S_*^1 \subsetneq S_*$ *and* $S_*^2 \subseteq S_* \backslash S_*^1$ *be two nonempty sets. Set* $\delta \equiv r'^{-1} L(L^2 + N) N^{-1-\alpha}$ *and assume that* $\delta \ll 1$. *Then*

$$\operatorname{cap}_N(\mathscr{E}_N(S_*^1), \mathscr{E}_N(S_*^2)) \geq \inf_{W \in \mathscr{W}(S_*^1, S_*^2)} \sum_{x,y \in S_*} \operatorname{cap}_S(x,y)[W_y - W_x]^2$$

$$\times \frac{N^{-\alpha-1}}{2 M_* I_\alpha(0) Z_{N,S}} \sum_{k=0}^{\ell_N} \sum_{\xi \in E_{k,S_0}} \frac{m_*^\xi}{a(\xi)} \left(1 - \mathscr{O}(\delta)\right), \quad (37)$$

where S_0 is the set S without two sites of S_*.

Proof. As usual (see e.g. [4]), a lower bound is obtained using the monotonicity of the Dirichlet form of the zero-range process

$$\Phi_N(F) = \frac{1}{2} \sum_{z,w \in S} \sum_{\eta \in E_N} \mu_N(\eta) r(z,w) g(\eta_z) \big[F(\eta^{z,w}) - F(\eta)\big]^2, \quad (38)$$

for a fixed function $F \in \mathscr{H}_N(\mathscr{E}_N(S_*^1), \mathscr{E}_N(S_*^2))$. The strategy is to set rates to zero until one obtains one-dimensional disjoint paths. Then the sum over the Dirichlet forms of the one-dimensional paths yields a lower bound for the Dirichlet form in (38) and hence for the capacity (see Fig. 1).

For each $\xi \in E_{k,S}$, $k \in \{0, \ldots, \ell_N\}$, we get exactly one one-dimensional path. For each pair $x, y \in S_*$ it consists of the path-segments $\{\xi, p_{x,y}\}$, where first the remaining $N-k$ particles are on site x and than jump one by one until they reach site y. This means that only one particle is jumping at any time (see Fig. 2). Let $\{\xi, p_{x,y}\}$ be the configuration $\{\xi, p_{x,y}\}_z = \xi_z$, for all $z \in S \backslash \{x, y\}$, $\{\xi, p_{x,y}\}_x = \xi_x + p$ and $\{\xi, p_{x,y}\}_y = \xi_y + N - k - p - 1$ for each $p \in \{0, \ldots, N-k-1\}$.

The path-segments are disjoint for the following reason. Let $\{\xi, p_{x,y}\}$, $\{\xi', p_{x,y}\} \in E_{N-1,S}$ be two different path-segments. Assume that at some time t these paths-segments coincide in one configuration due to the jump of the one

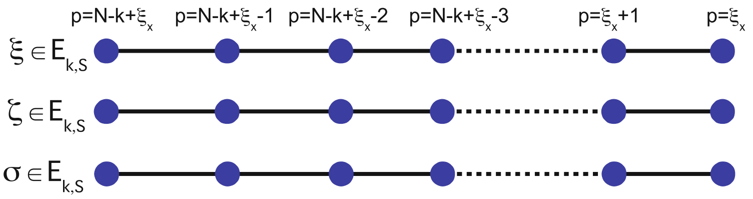


Fig. 2 Jump of the one particle in a path-segment

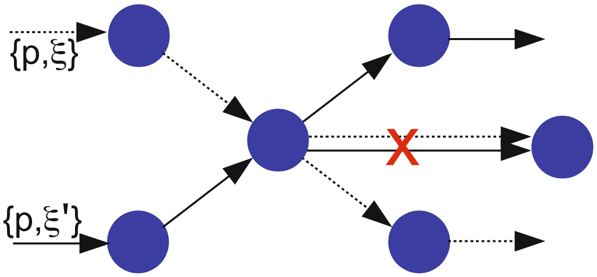


Fig. 3 Disjoint paths

particle. Since $\{\xi, p_{x,y}\}$ and $\{\xi', p_{x,y}\}$ are different, the sites differ in which the one particle is at time t. Thus, in the next step, particles from different sites jump such that the resulting configurations are different and hence the paths-segments do not merge (see Fig. 3).

With the strategy described above we obtain one-dimensional paths which consist of a Dirichlet form of a zero-range process on two sites multiplied with a term which we can estimate by the capacity of the underlying random walk.

Let $\mathfrak{d}_z \in E_{1,S}$ be the configuration with exactly one particle on the site $z \in S$ (the one jumping particle). Thus, we estimate (38) from below by

$$\Phi_N^{\mathrm{red}}(h^*) = \frac{1}{4}\sum_{k=0}^{\ell_N}\sum_{\xi\in E_{k,S}}\sum_{x,y\in S_*}\sum_{p=0}^{N-k-1}\sum_{z,w\in S}\mu_N(\{\xi,p_{x,y}\}+\mathfrak{d}_z)g(\{\xi,p_{x,y}\}_z+1)r(z,w)$$
$$\times\,[h^*(\{\xi,p_{x,y}\}+\mathfrak{d}_w)-h^*(\{\xi,p_{x,y}\}+\mathfrak{d}_z)]^2, \tag{39}$$

where h^* is a function in $\mathscr{H}_N(\mathscr{E}_N(S_*^1),\mathscr{E}_N(S_*^2))$. Inserting the definition of μ_N and g, (39) equals

$$\frac{N^\alpha}{4Z_{N,S}M_*}\sum_{k=0}^{\ell_N}\sum_{\xi\in E_{k,S}}\sum_{x,y\in S_*}\frac{m_*^\xi}{a(\xi\backslash\{x,y\})}\sum_{p=0}^{N-k-1}\frac{1}{a(\xi_x+p)a(\xi_y+N-k-p-1)}$$
$$\times\sum_{z,w\in S}m(z)r(z,w)[h^*(\{\xi,p_{x,y}\}+\mathfrak{d}_w)-h^*(\{\xi,p_{x,y}\}+\mathfrak{d}_z)]^2, \tag{40}$$

where $\xi\backslash\{x,y\}$ is the configuration without the sites x, y. Now we fix $x, y \in S_*$ and a configuration $\xi \in E_{k,S}$ and let the function $f_{x,y} : S \uparrow \mathbb{R}$ be given by

$$f_{x,y}(v) = \frac{h^*(\{\xi, p_{x,y}\} + \mathfrak{d}_v) - h^*(\{\xi, p_{x,y}\} + \mathfrak{d}_y)}{h^*(\{\xi, p_{x,y}\} + \mathfrak{d}_x) - h^*(\{\xi, p_{x,y}\} + \mathfrak{d}_y)}.$$

Obviously, $f_{x,y}$ is a function in $\mathscr{B}(x,y) = \{f : S \uparrow \mathbb{R}_+ | f(x) = 1, f(y) = 0\}$. Inserting $f_{x,y}$ in (40), the sum over $z, w \in S$ equals

$$2D_S(f_{x,y})\big[h^*(\{\xi, p_{x,y}\} + \mathfrak{d}_x) - h^*(\{\xi, p_{x,y}\} + \mathfrak{d}_y)\big]^2. \tag{41}$$

Since $f_{x,y} \in \mathscr{B}(x,y)$ and $D_S(f_{x,y}) \geq \mathrm{cap}_S(x,y)$ we get for (40) the lower bound

$$\begin{aligned}
&\frac{N^\alpha}{2Z_{N,S}M_*}\sum_{k=0}^{\ell_N}\sum_{\xi\in E_{k,S}}\sum_{x,y\in S_*}\mathrm{cap}_S(x,y)\frac{m_*^\xi}{a(\xi\backslash\{x,y\})}\\
&\times\sum_{p=0}^{N-k-1}\frac{1}{a(\xi_x+p)a(\xi_y+N-k-p-1)}[h^*(\{\xi,p_{x,y}\}+\mathfrak{d}_x)-h^*(\{\xi,p_{x,y}\}+\mathfrak{d}_y)]^2\\
&\geq \frac{N^\alpha}{2Z_{N,S}M_*}\sum_{k=0}^{\ell_N}\sum_{\xi\in E_{k,S}}\inf_{W(\xi)\in\mathscr{W}(S_*^1,S_*^2)}\sum_{x,y\in S_*}\mathrm{cap}_S(x,y)\frac{m_*^\xi}{a(\xi\backslash\{x,y\})}\\
&\times\inf_{\substack{h_{W(\xi)}\in\mathscr{H}_N(\mathscr{E}_N(S_*^1),\mathscr{E}_N(S_*^2))\\ h_{W(\xi)}(\eta)=W_x(\xi),\forall\eta\in\mathscr{E}_N^x\\ eta_{W(\xi)}(\eta)=W_y(\xi),\forall\eta\in\mathscr{E}_N^y}}\sum_{p=0}^{N-k-1}\frac{[h_{W(\xi)}(\{\xi,p_{x,y}\}+\mathfrak{d}_x)-h_{W(\xi)}(\{\xi,p_{x,y}\}+\mathfrak{d}_y)]^2}{a(\xi_x+p)a(\xi_y+N-k-p-1)}\\
&=\frac{N^\alpha}{2Z_{N,S}M_*}\sum_{k=0}^{\ell_N}\sum_{\xi\in E_{k,S}}\inf_{W(\xi)\in\mathscr{W}(S_*^1,S_*^2)}\sum_{x,y\in S_*}\mathrm{cap}_S(x,y)\frac{m_*^\xi}{a(\xi\backslash\{x,y\})}[W_x(\xi)-W_y(\xi)]^2\\
&\times\inf_{h_\xi\in\mathscr{H}_N(\mathscr{E}_N(S_*^1\cup x),\mathscr{E}_N(S_*^2\cup y))}\sum_{p=0}^{N-k-1}\frac{[h_\xi(\xi_x+p+1)-h_\xi(\xi_x+p)]^2}{a(\xi_x+p)a(\xi_y+N-k-p-1)}.
\end{aligned}\tag{42}$$

Due to the boundary conditions of the function h_ξ the last sum of (42) reduces to

$$\inf_{h_\xi\in\mathscr{H}_N(\mathscr{E}_N(S_*^1\cup x),\mathscr{E}_N(S_*^2\cup y))}\sum_{p=\ell_N-k+\xi_y}^{N-\ell_N-\xi_x-1}\frac{[h_\xi(\xi_x+p+1)-h_\xi(\xi_x+p)]^2}{a(\xi_x+p)a(\xi_y+N-k-p-1)}. \tag{43}$$

This is just a Dirichlet form of a zero-range process on the two sites x and y and it is minimized by the function (see [4])

$$H(x) = \frac{\sum_{q=\ell_N-k+\xi_x+\xi_y+1}^{x} a(q-1)a(N-k+\xi_x+\xi_y-q)}{\sum_{q=\ell_N-k+\xi_x+\xi_y+1}^{N-\ell_N-\xi_x} a(q-1)a(N-k+\xi_x+\xi_y-q)}. \tag{44}$$

Thus for $p \in [\ell_N - k + \xi_y, N - \ell_N - \xi_x - 1]$ we obtain

$$H(\xi_x + p + 1) - H(\xi_x + p) \geq \frac{a(\xi_x+p)a(\xi_y+N-k-p-1)}{\sum_{p=\ell_N-k+\xi_y}^{N-\ell_N-\xi_x-1} a(\xi_x+p)a(\xi_y+N-k-p-1)}. \tag{45}$$

Inserting (45) in (43) yields the lower bound

$$= \frac{1}{\sum_{p=\ell_N-k+\xi_y}^{N-\ell_N-\xi_x-1} a(\xi_x + p)a(\xi_y + N - k - p - 1)}. \tag{46}$$

Since this lower bound depends on the configuration ξ only through the number of particles k, we get for fixed k that this is bounded from below by $\left(N^{2\alpha+1} I_\alpha(0)\right)^{-1}$.

Combining (42) with Lemma 2 yields the lower bound

$$\begin{aligned} &\frac{N^{-\alpha-1}}{2Z_{N,S} M_* I_\alpha(0)} \sum_{k=0}^{\ell_N} \sum_{\xi \in E_{k,S}} \inf_{W \in \mathscr{W}(S_*^1, S_*^2)} \sum_{x,y \in S_*} \frac{m_*^\xi}{a(\xi \backslash \{x,y\})} \mathrm{cap}_S(x,y)[W_x - W_y + \mathscr{O}(\delta)]^2 \\ &= \frac{N^{-\alpha-1}}{2Z_{N,S} M_* I_\alpha(0)} \inf_{W \in \mathscr{W}(S_*^1, S_*^2)} \sum_{x,y \in S_*} \mathrm{cap}_S(x,y)[W_x - W_y]^2 \sum_{k=0}^{\ell_N} \sum_{\xi \in E_{k,S_0}} \tfrac{m_*^\xi}{a(\xi)} \left(1 - \mathscr{O}(\delta)\right), \end{aligned} \tag{47}$$

which proves the proposition. □

Note that in case that $S_*^2 = S_* \backslash S_*^1$, Lemma 2 is not needed in the proof and thus the error term disappears. This implies the lower bound for Theorem 4.

If $\delta \downarrow 0$, and hence in particular when L is independent of N, it is easy to see that due to the fact Lemma 2 implies that the equilibrium potential W in the sets $\mathscr{E}_N^x$, $x \in S_*^1$ and $x \in S_*^2$ are close to 1 and 0, respectively. Then it is straightforward to see that the same bound as in (37) holds for $\mathrm{cap}_N\left(\eta^{S_*^1}, \eta^{S_*^2}\right)$, provided $\delta \downarrow 0$. This provides the lower bound for Theorem 3.

4.1.2 Upper Bound

To prove an upper bound we follow the methods in [8]. Großkinsky et al. [12] have shown that above a critical particle density ρ_c the condensation phenomenon occurs.

Proposition 3. *Let $\epsilon > 0$, C_ϵ an ϵ-dependent constant and $|S_*| \geq 2$. Let $S_*^1 \subsetneq S_*$ and $S_*^2 \subseteq S_* \backslash S_*^1$ be two nonempty sets. We have*

$$\begin{aligned} \mathrm{cap}_N(\eta^{S_*^1}, \eta^{S_*^2}) \leq \mathrm{cap}_N(\mathscr{E}_N(S_*^1), \mathscr{E}_N(S_*^2)) \leq &\inf_{W \in \mathscr{W}(S_*^1, S_*^2)} \sum_{x,y \in S_*} \mathrm{cap}_S(x,y)[W_y - W_x]^2 \\ &\times \frac{N^{-\alpha-1}}{2Z_{N,S} M_* I_\alpha(3\epsilon)} \sum_{m=0}^{\ell_N} \sum_{\xi \in E_{m,S_0}} \frac{m_*^\xi}{a(\xi)} \left(1 + \mathscr{O}\left(\frac{\ell_N}{\epsilon N}\right)\right) \\ &+ \frac{L C_\epsilon N^{-\alpha-1}}{(\ell_N - L\rho_c)^\alpha}. \end{aligned} \tag{48}$$

Proof. The first inequality in (48) is obvious. We start by choosing a test function out of the set $\mathscr{H}_N(\mathscr{E}_N^x, \mathscr{E}_N^y)$, with $x, y \in S_*$. As in [2] we choose a test function depending on the function which solves the variational problem for the capacity of the underlying random walk and on the harmonic function of a zero-range process on two sites:

$$G^{x,y}(\eta) = \sum_{k=1}^{L-1} [f_{xy}(z_k) - f_{xy}(z_{k+1})] H\left(\frac{\eta_x}{N} + \min\left\{\frac{1}{N}\sum_{n=2}^{k} \eta_{z_n}, \epsilon\right\}\right), \tag{49}$$

where $x = z_1, z_2, \ldots, z_L = y$ is an enumeration of S such that $f_{xy}(z_i) \geq f_{xy}(z_j)$, for all $i < j$ with $i, j \in \{1, .., L\}$ and f_{xy} is the harmonic function in $\mathscr{B}(x, y)$. The function $H : \{0, \ldots, \frac{N-m}{N}\} \uparrow \mathbb{R}_+$ is the harmonic function of the zero-range process on two sites

$$H(z) = \frac{\sum_{q=\lfloor 3\epsilon N \rfloor + 1}^{\lfloor zN \rfloor} a(q-1)a(N-m-q)}{\sum_{q=\lfloor 3\epsilon N \rfloor}^{N-\lfloor 3\epsilon N \rfloor - 1} a(q)a(N-m-q-1)}, \tag{50}$$

with boundary value conditions

$$\begin{aligned} H(z) &= 0, \qquad \forall z \in \{0, .., \lfloor 3\epsilon N \rfloor\}, \\ H(z) &= 1, \qquad \forall z \in \{N - \lfloor 3\epsilon N \rfloor, \ldots, N\}. \end{aligned} \tag{51}$$

Observe that $G^{x,y}$ belongs to the set $\mathscr{H}_N(\mathscr{E}_N^x, \mathscr{E}_N^y)$.
For $x, y \in S_*$ we estimate the Dirichlet form on the set of configurations

$$F_N^{x,y} = \{\eta \in E_N : \eta_x + \eta_y \geq N - \ell_N\} :$$

$$\begin{aligned} \Phi_N(G^{x,y} | F_N^{x,y}) &= \frac{1}{2} \sum_{1 \leq i,j \leq L} \sum_{\eta \in F_N^{x,y}} \mu_N(\eta) g(\eta_{z_i}) r(z_i, z_j) [G^{x,y}(\eta^{z_i,z_j}) - G^{x,y}(\eta)]^2 \\ &= \frac{N^\alpha}{2Z_{N,S}} \sum_{1 \leq i,j \leq L} \sum_{\eta \in F_N^{x,y}} \tfrac{m_*^\eta}{a(\eta)} \tfrac{a(\eta_{z_i})}{a(\eta_{z_i} - 1)} r(z_i, z_j) [G^{x,y}(\eta^{z_i,z_j}) - G^{x,y}(\eta)]^2 \\ &\leq \tfrac{N^\alpha}{2Z_{N,S} M_*} \sum_{1 \leq i,j \leq L} m(z_i) r(z_i, z_j) \\ &\qquad \sum_{\substack{\xi \in E_{N-1,S} \\ \xi_x + \xi_y \geq N - \ell_N - 1}} \tfrac{m_*^\xi}{a(\xi)} [G^{x,y}(\xi + \mathfrak{d}_{z_j}) - G^{x,y}(\xi + \mathfrak{d}_{z_i})]^2. \end{aligned} \tag{52}$$

By the definition of $G^{x,y}$ we obtain

$$\Phi_N(G^{x,y}|F_N^{x,y}) \leq \tfrac{N^\alpha}{2Z_{N,S}M_*} \sum_{1\leq i,j\leq L} m(z_i)r(z_i,z_j) \sum_{\substack{\xi\in E_{N-1,S}\\ \xi_x+\xi_y\geq N-\ell_N-1}} \tfrac{m_*^\xi}{a(\xi)} \Bigg[\sum_{k=1}^{L-1} \big(f_{xy}(z_k)-f_{xy}(z_{k+1})\big)$$

$$\times\left(H\left(\frac{\xi_x}{N}+\sum_{n=2}^{k}\frac{\xi_{z_n}}{N}+\frac{\mathbb{1}_{\{k\geq i\}}}{N}\right)-H\left(\frac{\xi_x}{N}+\sum_{n=2}^{k}\frac{\xi_{z_n}}{N}+\frac{\mathbb{1}_{\{k\geq j\}}}{N}\right)\right)\Bigg]^2. \tag{53}$$

Fix two sites $z_i \neq z_j \in S$ with $i < j$. Since $m_*(x) = m_*(y) = 1$ and setting $m_k \equiv \sum_{n=2}^{k} \xi_{z_n}$, we get for the sum over the configurations ξ in (53) the upper bound

$$\sum_{m=0}^{\ell_N} \sum_{\zeta\in E_{m,S\setminus\{x,y\}}} \frac{m_*^\zeta}{a(\zeta)} \sum_{p=\lfloor 2\epsilon N\rfloor}^{N-\lfloor 3\epsilon N\rfloor-1} \frac{1}{a(p)a(N-m-p-1)}$$

$$\times\left[\sum_{k=i}^{j-1}\big(f_{xy}(z_k)-f_{xy}(z_{k+1})\big)\left(H\left(\frac{p+m_k+1}{N}\right)-H\left(\frac{p+m_k}{N}\right)\right)\right]^2. \tag{54}$$

The sum over p only runs from $\lfloor 2\epsilon N\rfloor$ to $N-\lfloor 3\epsilon N\rfloor - 1$ due to the boundary conditions (51). Inserting the explicit form (50) yields

$$\sum_{m=0}^{\ell_N} \sum_{\zeta\in E_{m,S\setminus\{x,y\}}} \frac{m_*^\zeta}{a(\zeta)} \sum_{p=\lfloor 2\epsilon N\rfloor}^{N-\lfloor 3\epsilon N\rfloor-1} \frac{1}{a(p)a(N-m-p-1)}$$

$$\times\left[\sum_{k=i}^{j-1}\big(f_{xy}(z_k)-f_{xy}(z_{k+1})\big)\frac{a(p+m_k)a(N-m-m_k-p-1)}{\sum_{q=\lfloor 3\epsilon N\rfloor}^{N-\lfloor 3\epsilon N\rfloor-1} a(q)a(N-m-q-1)}\right]^2. \tag{55}$$

Since $m_k \leq \ell_N$ and $p \geq \lfloor 2\epsilon N\rfloor$, we can estimate

$$a(p+m_k) = a(p)\left(1+\tfrac{m_k}{p}\right)^\alpha \leq a(p)\left(1+\tfrac{\ell_N}{\lfloor 2\epsilon N\rfloor}\right)^\alpha \leq a(p)\left(1+\mathscr{O}\left(\tfrac{\ell_N}{\epsilon N}\right)\right),$$

and $a(N-m-m_k-p-1) \leq a(N-m-p-1)$. Inserting these estimates into (55) yields the upper bound

$$\sum_{m=0}^{\ell_N} \sum_{\zeta\in E_{m,S\setminus\{x,y\}}} \frac{m_*^\zeta}{a(\zeta)} \sum_{p=\lfloor 2\epsilon N\rfloor}^{N-\lfloor 3\epsilon N\rfloor-1} \frac{1}{a(p)a(N-m-p-1)}$$

$$\times\left[\sum_{k=i}^{j-1}\big(f_{xy}(z_k)-f_{xy}(z_{k+1})\big)\frac{a(p)a(N-m-p-1)}{\sum_{q=\lfloor 3\epsilon N\rfloor}^{N-\lfloor 3\epsilon N\rfloor-1} a(q)a(N-m-q-1)}\left(1+\mathscr{O}\left(\tfrac{\ell_N}{\epsilon N}\right)\right)\right]^2$$

$$\leq [f_{xy}(z_i) - f_{xy}(z_j)]^2 \quad \sum_{m=0}^{\ell_N} \sum_{\zeta \in E_{m,S\setminus\{x,y\}}} \frac{m_*^{\zeta}}{a(\zeta)} \frac{\sum_{p=\lfloor 2\epsilon N\rfloor}^{N-\lfloor 3\epsilon N\rfloor -1} a(p)a(N-m-p-1)}{\left(\sum_{q=\lfloor 3\epsilon N\rfloor}^{N-\lfloor 3\epsilon N\rfloor -1} a(q)a(N-m-q-1)\right)^2}$$
$$\times \left(1 + \mathscr{O}\left(\tfrac{\ell_N}{\epsilon N}\right)\right). \tag{56}$$

For $j < i$ we get the same bound. Now note that

$$\frac{\sum_{p=\lfloor 2\epsilon N\rfloor}^{N-\lfloor 3\epsilon N\rfloor -1} a(p)a(N-m-p-1)}{\left(\sum_{q=\lfloor 3\epsilon N\rfloor}^{N-\lfloor 3\epsilon N\rfloor -1} a(q)a(N-m-q-1)\right)^2} = \frac{1+\mathscr{R}}{\sum_{p=\lfloor 3\epsilon N\rfloor}^{N-\lfloor 3\epsilon N\rfloor -1} a(p)a(N-m-p-1)} \tag{57}$$

with

$$\mathscr{R} = \frac{\sum_{p=\lfloor 2\epsilon N\rfloor}^{\lfloor 3\epsilon N\rfloor -1} a(p)a(N-m-p-1)}{\sum_{p=\lfloor 3\epsilon N\rfloor}^{N-\lfloor 3\epsilon N\rfloor -1} a(p)a(N-m-p-1)}. \tag{58}$$

It is easy to see that

$$\mathscr{R} = \mathscr{O}\left(\tfrac{\lfloor \epsilon N\rfloor}{N-\lfloor \epsilon N\rfloor}\right). \tag{59}$$

Thus we obtain for (53) the upper bound

$$\frac{N^{\alpha}}{2Z_{N,S}M_*} \sum_{1\leq i,j\leq L} m(z_i)r(z_i,z_j)[f_{xy}(z_i) - f_{xy}(z_j)]^2 \tag{60}$$
$$\times \sum_{m=0}^{\ell_N} \sum_{\zeta \in E_{m,S\setminus\{x,y\}}} \frac{m_*^{\zeta}}{a(\zeta)} \frac{1}{\sum_{p=\lfloor 3\epsilon N\rfloor}^{N-\lfloor 3\epsilon N\rfloor -1} a(p)a(N-m-p-1)} \left(1 + \mathscr{O}\left(\tfrac{\ell_N}{\epsilon N}\right)\right).$$

The sum over i, j in (60) is just the capacity of the underlying random walk between the two sites x and y. Since

$$\sum_{p=\lfloor 3\epsilon N\rfloor}^{N-\lfloor 3\epsilon N\rfloor -1} a(p)a(N-m-p-1) \geq N^{2\alpha+1} I_{\alpha}(3\epsilon)\left(1 - \mathscr{O}\left(\tfrac{\ell_N}{N}\right)\right), \tag{61}$$

we get for (60)

$$\Phi_N(G^{x,y}|F_N^{x,y}) \leq \frac{N^{-\alpha-1}\,\mathrm{cap}_S(x,y)}{Z_{N,S}M_* I_{\alpha}(3\epsilon)} \sum_{m=0}^{\ell_N} \sum_{\zeta \in E_{m,S_0}} \frac{m_*^{\zeta}}{a(\zeta)} \left(1 + \mathscr{O}\left(\tfrac{\ell_N}{\epsilon N}\right)\right). \tag{62}$$

Now we can calculate an upper bound for the desired capacity:

$$\begin{aligned}\mathrm{cap}_N(\mathscr{E}_N(S_*^1),\mathscr{E}_N(S_*^2)) &= \inf_{G\in\mathscr{H}_N(\mathscr{E}_N(S_*^1),\mathscr{E}_N(S_*^2))} \Phi_N(G)\\ &\leq \inf_{W\in\mathscr{W}(S_*^1,S_*^2)} \Phi_N(G_W^S), \qquad (63)\end{aligned}$$

where the test function $G_W^S \in \mathscr{H}_N(\mathscr{E}_N(S_*^1),\mathscr{E}_N(S_*^2))$ is an appropriate combination of an interpolation and convex combinations of the functions $G^{x,y}$, $x,y\in S_*$, also defined on the set of configurations $E_{N,S}\setminus\bigcup_{x,y\in S_*}F_N^{x,y}$ (the detailed construction of G_W^S can be found in [7]). For (63) (cf. [7]) we get the upper bound

$$= \inf_{W\in\mathscr{W}(S_*^1,S_*^2)} \frac{1}{2}\sum_{x,y\in S_*} \Phi_N(G^{x,y}|F_N^{x,y})[W_y-W_x]^2 + \frac{LC_\epsilon N^{-\alpha-1}}{(\ell_N - L\rho_c)^\alpha}, \qquad (64)$$

where the last term comes from the calculation of the Dirichlet form on the set of configurations $E_{N,S}\setminus\bigcup_{x,y\in S_*}F_N^{x,y}$ (cf. [7]). Inserting (62) yields the desired upper bound

$$\begin{aligned}\mathrm{cap}_N(\mathscr{E}_N(S_*^1),\mathscr{E}_N(S_*^2)) \leq \inf_{W\in\mathscr{W}(S_*^1,S_*^2)} &\sum_{x,y\in S_*} \mathrm{cap}_S(x,y)[W_y-W_x]^2 \qquad (65)\\ &\times \frac{N^{-\alpha-1}}{2Z_{N,S}M_*I_\alpha(3\epsilon)} \sum_{m=0}^{\ell_N}\sum_{\xi\in E_{m,S_0}} \frac{m_*^\xi}{a(\xi)}\left(1+\mathscr{O}\left(\frac{\ell_N}{\epsilon N}\right)\right)\\ &+\frac{LC_\epsilon N^{-\alpha-1}}{(\ell_N-L\rho_c)^\alpha}.\end{aligned}$$

□

4.2 *Proofs of Theorems 3 and 4*

Proof (of Theorem 3). By the remark after the proof of Proposition 2, and using Lemma 1, we get in the case of fixed L the lower bound

$$\mathrm{cap}_N\left(\eta^{S_*^1},\eta^{S_*^2}\right) \geq \tfrac{N^{-\alpha-1}(1+o(1))}{2Z_SM_*I_\alpha(0)} \inf_{W\in\mathscr{W}(S_*^1,S_*^2)} \sum_{x,y\in S_*} \mathrm{cap}_S(x,y)[W_x-W_y]^2 \sum_{k=0}^{\ell_N}\sum_{\xi\in E_{k,S_0}} \tfrac{m_*^\xi}{a(\xi)}. \qquad (66)$$

With the definitions (7) and (6), the sum over k in (66) converges for $N\uparrow\infty$ to

$$\sum_{k\geq 0}\sum_{\zeta\in E_{k,S_0}} \frac{m_*^\zeta}{a(\zeta)} = \prod_{z\in S_0}\sum_{j\geq 0}\frac{m_*(z)^j}{a(j)} = \prod_{z\in S_0}\Gamma_z = \frac{Z_S}{|S_*|\Gamma(\alpha)}. \qquad (67)$$

This gives the lower bound for (27). For the upper bound we insert the bound

$$\sum_{m=0}^{\ell_N} \sum_{\xi \in E_{m,S_0}} \frac{m_*^{\xi}}{a(\xi)} \leq \frac{Z_S}{|S_*|\Gamma(\alpha)} \tag{68}$$

into (65) and use Lemma 1. We obtain the upper bound for (27). This concludes the proof of Theorem 3. □

Proof (Theorem 4). Assume $L(N) \uparrow \infty$, such that $L(N)/N \uparrow 0$. In the case where S_*^1 and S_*^2 are a partition of S_* the lower bound takes the form

$$\mathrm{cap}_N(\mathscr{E}_N(S_*^1), \mathscr{E}_N(S_*^2)) \geq \frac{N^{-\alpha-1}}{M_* I_\alpha(0)} \sum_{x \in S_*^1, y \in S_*^2} \mathrm{cap}_S(x,y) \frac{1}{Z_{N,S}} \sum_{k=0}^{\ell_N} \sum_{\xi \in E_{k,S_0}} \frac{m_*^{\xi}}{a(\xi)}. \tag{69}$$

We can estimate the last expression of (69) by the measure of the set $\mathscr{E}_N^x$, for $x \in S_*$ and get

$$\mathrm{cap}_N(\mathscr{E}_N(S_*^1), \mathscr{E}_N(S_*^2)) \geq \frac{N^{-\alpha-1}(1+o(1))}{M_* |S_*| I_\alpha(0) \Gamma(\alpha)} \sum_{x \in S_*^1, y \in S_*^2} \mathrm{cap}_S(x,y). \tag{70}$$

In the case of a partition we get from (65) with the estimation (68) the upper bound

$$\mathrm{cap}_N(\mathscr{E}_N(S_*^1), \mathscr{E}_N(S_*^2)) \leq \frac{N^{-\alpha-1} Z_S}{M_* |S_*| I_\alpha(0) \Gamma(\alpha) Z_{N,S}} \sum_{x \in S_*^1, y \in S_*^2} \mathrm{cap}_S(x,y)\,(1+o(1))\,. \tag{71}$$

Since $\frac{Z_S}{Z_{N,S}} = 1 + o(1)$ we obtain

$$\mathrm{cap}_N(\mathscr{E}_N(S_*^1), \mathscr{E}_N(S_*^2)) \leq \frac{N^{-\alpha-1}(1+o(1))}{M_* |S_*| I_\alpha(0) \Gamma(\alpha)} \sum_{x \in S_*^1, y \in S_*^2} \mathrm{cap}_S(x,y). \tag{72}$$

Combining (70) and (72) yields Theorem 4. □

4.3 Metastability of the Zero-Range Process

We now prove Proposition 1 for $L < \infty$.

Proposition 4.

$$\frac{\sup_{\eta \in \mathscr{M}} \mathrm{cap}_N(\eta, \mathscr{M} \backslash \eta)/\mu_N(\eta)}{\inf_{\xi \in \mathscr{M}^c} \mathrm{cap}_N(\xi, \mathscr{M})/\mu_N(\xi)} \leq \mathscr{O}(r'^{-1} L(L^2+N) N^{-\alpha-1}). \tag{73}$$

Proof. We already have shown an upper bound for the numerator. The following Lemma bounds the denominator from below.

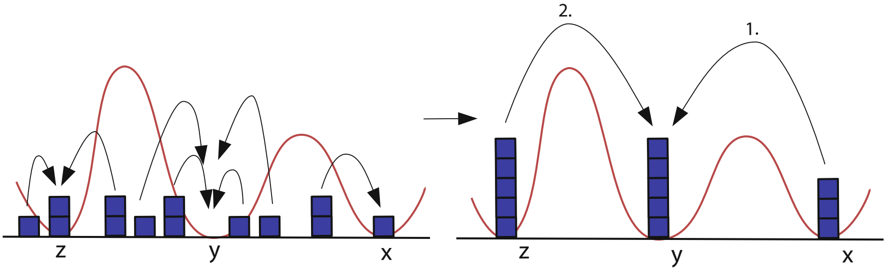


Fig. 4 Way of the particles in the chosen path

Lemma 3. *Let* $r' = \min_{u \in S} r(u, u \pm 1)) > 0$, $\xi \in \mathscr{M}^c$ *and* $K(\alpha)$ *an* α*-dependent constant. We have that*

$$\frac{\text{cap}_N(\xi, \mathscr{M})}{\mu_N(\xi)} \geq \frac{r' L^{-1}}{(L^2 + N)|S_*|K(\alpha)}. \tag{74}$$

First we proof Lemma 3 for a zero-range process with only three attractors, i.e., $|S_*| = 3$, and afterwards we show that it holds also for any $|S_*|$ based on an algorithm.

Proof (The Case $|S_*| = 3$*).* Let $S_* = \{x, y, z\}$ and $r(x, y) \geq r(x, z) \geq r(y, z)$. For estimating the capacity from below we only consider one path from ξ to $\mathscr{M}$. We choose the path where firstly all particles of a valley jump on its attractor. The resulting configuration, where only the three attractors are occupied, is called η. Then we let all particles of the lower occupied attractor in the big valley $A(x) \cup A(y)$ jump on the higher occupied attractor. Thus we get the configuration σ^{xy} with only two occupied sites. Finally all particles from the lower occupied site of σ^{xy} jump to the higher occupied site (see Fig. 4). Since

$$\text{cap}_N(\xi, \mathscr{M}) \geq \frac{1}{\text{cap}_N^{-1}(\xi, \eta) + \text{cap}_N^{-1}(\eta, \sigma^{xy}) + \text{cap}_N^{-1}(\sigma^{xy}, \mathscr{M})}, \tag{75}$$

we have to calculate a lower bound for each capacity on the right hand side of (75). For any $w \in S_*$ let $(w)_n$ be a distance to w increasing enumeration of $A(w) \backslash \{w\}$ and $\xi_w^i \equiv \xi_w + \sum_{j=1}^{i-1} \xi_{w_j}$, the number of particles on site w before the particles of site w_i jump on the attractor w. Let be $|w| = \max\{\text{dist}(w, w_i) | w \in S_*, w_i \in (w)_n\}$. For each transition of the particles from site w_i to site w, via nearest neighbor jumps, we estimate the explicit formula for the capacity of the one dimensional chain (see Chap. 8.1 in [4]). Thus we get for the capacity between ξ and η the lower bound

$$\text{cap}_N^{-1}(\xi, \eta) \leq \sum_{w \in S_*} \sum_{i=1}^{|(w)_n|} \sum_{k=0}^{\xi_{w_i} - 1} \frac{|w|(N-1)^\alpha}{\mu_N(\{\xi_{w_i} - k, \xi_w^i + k\}) r' N^\alpha}, \tag{76}$$

where $\{\xi_{w_i} - k, \xi_w^i + k\}$ is the configuration:

$$\{\xi_{w_i} - k, \xi_w^i + k\}_v = \begin{cases} \xi_v, & \text{for all } v \in S\setminus A(w) \text{ and } v \in \{w_n : n > i\} \\ 0, & \text{for all } v \in \{w_n : n < i\} \\ \xi_{w_i} - k, & \text{for } v = w_i \\ \xi_w^i + k, & \text{for } v = w. \end{cases}$$

Note that the μ_N-measure of the configuration increases after each transition of a particle, because there are more particles on condensate-sites. Equation (76) equals

$$\begin{aligned} &\sum_{w\in S_*}\sum_{i=1}^{|(w)_n|}\sum_{k=0}^{\xi_{w_i}-1} \frac{Z_{N,S}|w|}{N^\alpha}\,\frac{a(\{\xi_{w_i}-k,\xi_w^i+k\}\setminus\{w_i,w\})}{m_*^{\{\xi_{w_i}-k,\xi_w^i+k\}\setminus\{w_i\}}}\,\frac{a(\xi_{w_i}-k)a(\xi_w^i+k)(N-1)^\alpha}{m_*(w_i)^{\xi_{w_i}-k}N^\alpha} \\ &= \sum_{w\in S_*}\sum_{i=1}^{|(w)_n|} \frac{Z_{N,S}|w|}{N^\alpha}\,\frac{a(\{\xi_{w_i}-k,\xi_w^i+k\}\setminus\{w_i,w\})a(\xi_w^i)a(\xi_{w_i})}{m_*^{\{\xi_{w_i}-k,\xi_w^i+k\}\setminus\{w_i\}}m_*(w_i)^{\xi_{w_i}}\,r'} \\ &\times \sum_{k=0}^{\xi_{w_i}-1}\left(1-\tfrac{k}{\xi_{w_i}}\right)^\alpha\left(1+\tfrac{k}{\xi_w^i}\right)^\alpha\left(1-\tfrac{1}{N}\right)^\alpha m_*(w_i)^k. \end{aligned} \tag{77}$$

The sum over k in (77) is bounded by L times an α-dependent constant $K'(\alpha)$. Observe that in the case where ξ_w^1 is zero, $a(\xi_w^1) = 1$ and we can use the same estimation. We obtain

$$\text{cap}_N^{-1}(\xi,\eta) \le \sum_{w\in S_*}\sum_{i=1}^{|(w)_n|}\frac{|w|LK'(\alpha)}{\mu_N(\{\xi_{w_i},\xi_w^i\})r'} \le \sum_{w\in S_*}\sum_{i=1}^{|(w)_n|}\frac{L^2K'(\alpha)}{\mu_N(\{\xi_{w_i},\xi_w^i\})r'}. \tag{78}$$

If $\eta_x = N$ or $\eta_y = N$ we can stop here, because we already have a configuration in $\mathscr{M}$. Otherwise we continue the estimation of the capacities. Without loss of generality let $\eta_x \le \eta_y$. A similar estimation of the formula of the one-dimensional chain yields

$$\begin{aligned} \text{cap}_N^{-1}(\eta,\sigma^{xy}) &\le \sum_{k=0}^{\eta_x-1}\frac{\text{dist}(x,y)(N-1)^\alpha}{\mu_N(\{\eta_x-k,\eta_y+k\})r'N^\alpha} \\ &= \frac{Z_{N,S}\,\text{dist}(x,y)}{N^\alpha}\,\frac{a(\eta\setminus\{x,y\})}{r'm_*^\eta}\sum_{k=0}^{\eta_x-1}\frac{a(\eta_x-k)a(\eta_y+k)(N-1)^\alpha}{N^\alpha} \\ &= \frac{\text{dist}(x,y)}{\mu_N(\eta)r'}\sum_{k=0}^{\eta_x-1}\left(1-\frac{k}{\eta_x}\right)^\alpha\left(1+\frac{k}{\eta_y}\right)^\alpha\left(1-\frac{1}{N}\right)^\alpha, \end{aligned} \tag{79}$$

where $\{\eta_x - k, \eta_y + k\} \in E_{N,S}$ is the configuration where all sites are empty except for site x with $\eta_x - k$ particles, site y with $\eta_y + k$ particles and the site z with η_z particles on it. Observe that $\frac{\eta_x}{\eta_y} \leq 1$. For all $k \in \{1, \dots, \eta_x - 1\}$ we have that

$$\left(1 - \frac{1}{N}\right)^{\alpha} \leq 1, \quad \left(1 - \frac{k}{\eta_x}\right)^{\alpha} \leq 1 \quad \text{and} \quad \left(1 + \frac{k}{\eta_y}\right)^{\alpha} \leq 2^{\alpha}. \tag{80}$$

Since $\eta_x, \eta_y < N$, we can estimate the sum in (79) by $2^{\alpha} N$ from above and get

$$\operatorname{cap}_N^{-1}(\eta, \sigma^{xy}) \leq \frac{2^{\alpha} N \operatorname{dist}(x,y)}{\mu_N(\eta) r'} \leq \frac{2^{\alpha} N L}{\mu_N(\eta) r'}. \tag{81}$$

If $\sigma_y^{xy} = N$ we can stop here. Otherwise we continue the estimation of the capacities. Without loss of generality let $\sigma_z^{xy} \leq \sigma_y^{xy}$. A similar estimation yields

$$\operatorname{cap}_N^{-1}(\sigma^{xy}, \mathscr{M}) \leq \frac{2^{\alpha} N \operatorname{dist}(y,z)}{\mu_N(\sigma^{xy}) r'} \leq \frac{2^{\alpha} N L}{\mu_N(\sigma^{xy}) r'}. \tag{82}$$

Combining (78), (81) and (82) we get for the capacity $\operatorname{cap}_N(\xi, \mathscr{M})/\mu_N(\xi)$ the lower bound

$$\left[\frac{L^2 K'(\alpha)}{r'} \sum_{w \in S_*} \sum_{i=1}^{|(w)_n|} \frac{\mu_N(\xi)}{\mu_N(\{\xi_{w_i}, \xi_w^i\})} + \frac{2^{\alpha} N L}{r'} \left(\frac{\mu_N(\xi)}{\mu_N(\eta)} + \frac{\mu_N(\xi)}{\mu_N(\sigma^{xy})}\right)\right]^{-1}. \tag{83}$$

Since there are more particles on the sites of S_* in the configurations $\{\xi_{w_i}, \xi_w^i\}$, η and σ^{xy} than in the configuration ξ, we have that $\frac{\mu_N(\xi)}{\mu_N(\{\xi_{w_i}, \xi_w^i\})}, \frac{\mu_N(\xi)}{\mu_N(\eta)}, \frac{\mu_N(\xi)}{\mu_N(\sigma^{xy})} \leq 1$. Thus, we can continue

$$\begin{aligned} \frac{\operatorname{cap}_N(\xi, \mathscr{M})}{\mu_N(\xi)} &\geq \left[\frac{L^3 K'(\alpha) |S_*|}{r'} + \frac{2^{\alpha}(|S_*| - 1) N L}{r'}\right]^{-1} \\ &\geq \frac{r' L^{-1}}{(L^2 + N)|S_*| K(\alpha)}. \end{aligned} \tag{84}$$

□

Proof (The Case $|S_*| > 3$). The following algorithm generalizes the case to $|S_*| > 3$. Fix a configuration $\xi \notin \mathscr{M}$. First we let all particles in a valley jump on its attractor. From the resulting configuration we construct a labeled tree. Each attractor corresponds to a leaf of the tree which is labeled with its occupation number. The local maxima of the potential of the random walk on S are the vertices of the tree where the biggest one is the root of the tree. The root is connected with the vertices of the next two biggest local maxima and so on (see Fig. 5).

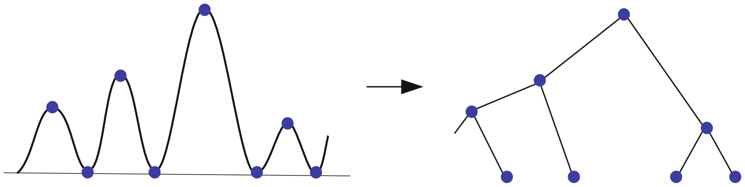


Fig. 5 Construction of the tree

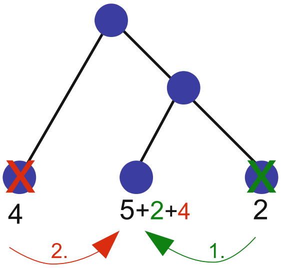


Fig. 6 Algorithm for $|S_*| = 3$

The algorithm works as follows: For each pair of leaves we calculate the length of the shortest path between them and choose the pair of leaves with the shortest path. If there are multiple shortest paths of the same length, we choose the one with the lowest labeled leaf. Next we increase the label of the higher labeled leaf in the pair by the value of the label of the lower labeled leaf and delete this one. We continue this procedure until we obtain a tree with only one leaf. This algorithm describes one path from the configuration $\xi \notin \mathscr{M}$ to a configuration in $\mathscr{M}$, because the final tree corresponds to such a configuration in $\mathscr{M}$. Thus for the general case we have to calculate at most $|S_*| - 1$ transitions between condensate-sites, i.e., we have to estimate at most $|S_*| - 1$ capacities. Hence (84) also holds in the general case. Figure 6 illustrates the algorithm for the case $|S_*| = 3$. □

We now conclude the proof of the proposition. Using Theorem 3 and Lemma 1 yields

$$\sup_{w\in S_*} \frac{\mathrm{cap}_N(\eta^w, \mathscr{M}\backslash\eta^w)}{\mu_N(\eta^w)} = \sup_{w\in S_*} \frac{N^{-\alpha-1} Z_{N,S}(1+o(1))}{M_*|S_*|I_\alpha(0)\Gamma(\alpha)} \sum_{v\in S_*\backslash\{w\}} \mathrm{cap}_S(w,v) \quad (85)$$

$$= \sup_{w\in S_*} \frac{N^{-\alpha-1} Z_S}{|S_*|M_* I_\alpha(0)\Gamma(\alpha)} \sum_{v\in S_*\backslash\{w\}} \mathrm{cap}_S(w,v)(1+o(1)),$$

where $\mu_N(\eta^w) = \frac{1}{Z_{N,S}}$ is the configuration with all N particles at the site w. For the denominator we use Lemma 3 and get the desired result

$$\begin{aligned}\frac{\sup_{\eta\in\mathscr{M}} \mathrm{cap}_N(\eta,\mathscr{M}\backslash\eta)/\mu_N(\eta)}{\inf_{\xi\in\mathscr{M}^c}\mathrm{cap}_N(\xi,\mathscr{M})/\mu_N(\xi)} &\le \tfrac{N^{-\alpha-1}Z_S L(L^2+N)K(\alpha)(1+o(1))}{M_* I_\alpha(0)\Gamma(\alpha)r'} \sup_{w\in S_*}\sum_{v\in S_*\backslash\{w\}} \mathrm{cap}_S(w,v)\\ &= \mathscr{O}(r'^{-1}L(L^2+N)N^{-\alpha-1}). \qquad (86)\end{aligned}$$

□

Acknowledgements Financial support from the German Research Council (DFG) through SFB 611 is gratefully acknowledged.

References

1. Andjel, E.D.: Invariant measures for the zero range processes. Ann. Probab. **10**(3), 525–547 (1982)
2. Beltrán, J., Landim, C.: Metastability of reversible condensed zero range processes on a finite set. Probab. Theory Relat. Fields **152**(3–4), 781–807 (2012)
3. Bianchi, A., Bovier, A., Ioffe, D.: Pointwise estimates and exponential laws in metastable systems via coupling methods. Ann. Probab. **40**, 339–379 (2012)
4. Bovier, A.: Metastability and ageing in stochastic dynamics. In: Maass, A., Martinez, S., San Martin, J. (eds.) Dynamics and Randomness II. Volume 10 of Nonlinear Phenom. Complex Systems, pp. 17–79. Kluwer Academic, Dordrecht (2004)
5. Bovier, A.: Metastability: a potential theoretic approach. In: International Congress of Mathematicians, Madrid, vol. III, pp. 499–518. European Mathematical Society, Zürich (2006)
6. Bovier, A.: Metastability. In: Biskup, M., Kotecký, R., et al. (eds.) Methods of Contemporary Mathematical Statistical Physics. Volume 1970 of Lecture Notes in Mathematics, pp. 177–221. Springer, Berlin (2009)
7. Bovier, A., Neukirch, R.: Metastability in the Zero-Range process (Submitted)
8. Bovier, A., Eckhoff, M., Gayrard, V., Klein, M.: Metastability and low lying spectra in reversible Markov chains. Commun. Math. Phys. **228**(2), 219–255 (2002)
9. Evans, M.: Phase transitions in one-dimensional nonequilibrium systems. Braz. J. Phys. **30**(1), 42–57 (2000)
10. Großkinsky, S., Schütz, G.M.: Discontinuous condensation transition and nonequivalence of ensembles in a zero-range process. J. Stat. Phys. **132**(1), 77–108 (2008)
11. Großkinsky, S., Spohn, H.: Stationary measures and hydrodynamics of zero range processes with several species of particles. Bull. Braz. Math. Soc. (N.S.) **34**(3), 489–507 (2003)
12. Großkinsky, S., Schütz, G.M., Spohn, H.: Condensation in the zero range process: stationary and dynamical properties. J. Stat. Phys. **113**(3–4), 389–410 (2003)
13. Spitzer, F.: Interaction of Markov processes. Adv. Math. **5**, 246–290 (1970)

Convergence of the Two-Point Function of the Stationary TASEP

Jinho Baik, Patrik Lino Ferrari, and Sandrine Péché

Abstract We consider the two-point function of the totally asymmetric simple exclusion process with stationary initial conditions. The two-point function can be expressed as the discrete Laplacian of the variance of the associated height function. The limit of the distribution function of the appropriately scaled height function was obtained previously by Ferrari and Spohn. In this paper we show that the convergence can be improved to the convergence of moments. This implies the convergence of the two-point function in a weak sense along the near-characteristic direction as time tends to infinity, thereby confirming the conjecture in the paper of Ferrari and Spohn.

1 Introduction and Result

The totally asymmetric simple exclusion process (TASEP) is arguably the simplest non-reversible interacting stochastic particle system, and it is also one of the most studied. Particles live on $\mathbb{Z}$ and they satisfy the exclusion constraint: each site can be occupied by at most one particle. Therefore a particle configuration can be denoted by $\eta \in \{0,1\}^{\mathbb{Z}}$, where $\eta_j = 0$ means that site j is empty while $\eta_j = 1$ means that

J. Baik (✉)
Department of Mathematics, University of Michigan, Ann Arbor, MI, 48109, USA
e-mail: baik@umich.edu

P.L. Ferrari
Institut für Angewandte Mathematik, Rheinische Friedrich-Wilhelms-Universität Bonn, Endenicher Allee 60, D-53115 Bonn, Germany
e-mail: ferrari@uni-bonn.de

S. Péché
U.F.R. de Mathématiques, Université Paris Diderot, Case 7012 (Site Chevaleret), F-75205 Paris, France
e-mail: peche@math.univ-paris-diderot.fr

M. Griebel (ed.), *Singular Phenomena and Scaling in Mathematical Models*,
DOI 10.1007/978-3-319-00786-1_5,

the site is occupied. The dynamics of the TASEP is then defined as follows: every particle tries to jump to its right neighbor with rate one. The jumps occurs only if the exclusion constraint is satisfied.

It is known [12] that the only translation-invariant stationary measures of the TASEP are Bernoulli product measures with parameter $\rho \in [0,1]$, namely,

$$\mathbb{P}(\eta_j = 1) = \rho \quad \text{for all } j \in \mathbb{Z}. \tag{1}$$

Here ρ is the average density of particles. The cases $\rho = 0$ and $\rho = 1$ are trivial and in the following we fix $\rho \in (0,1)$. This system is referred as *stationary TASEP*.

The two-point function is defined as

$$S(j,t) := \mathbb{E}\left(\eta_j(t)\eta_0(0)\right) - \rho^2. \tag{2}$$

Note that this equals the covariance of $\eta_j(t)$ and $\eta_0(0)$. Hence the two-point function carries the information on how site j at time t is correlated with site 0 at time 0. It is known that

$$\sum_{j\in\mathbb{Z}} S(j,t) = \rho(1-\rho) =: \chi \tag{3}$$

and also $S(j,t) \geq 0$. This implies that $\frac{1}{\chi}S(j,t)$ can be thought of as a probability mass function in $j \in \mathbb{Z}$. Indeed this equals the probability that a second class particle, which was at site 0 at time 0, is at site j at time t [9]. It is also known that the expectation of j with respect to the probability mass function $\frac{1}{\chi}S(j,t)$ satisfies

$$\sum_{j\in\mathbb{Z}} j\,\frac{S(j,t)}{\chi} = (1-2\rho)t, \tag{4}$$

and the variance scales as [16, 18]

$$\sum_{j\in\mathbb{Z}} j^2\,\frac{S(j,t)}{\chi} - ((1-2\rho)t)^2 = O(t^{4/3}). \tag{5}$$

as $t \to \infty$. Therefore, for large time t, one expects the scaling form for S as[1]

$$S(j,t) \simeq \frac{\chi}{4} g''_{\mathrm{sc}}\left(\frac{j-(1-2\rho)t}{2\chi^{1/3}t^{2/3}}\right)\frac{1}{2\chi^{1/3}t^{2/3}} \tag{6}$$

for some non-random function g_{sc}. The precise expression of g_{sc} was first conjectured in [15] based on the work [6]:

[1] The multiplicative factor $\frac{\chi}{4}$ was incorrectly written as $\frac{\chi}{2}$ in [14]. This is a typographical error.

$$g_{\mathrm{sc}}(w) = \int_{\mathbb{R}} s^2 dF_w(s) \tag{7}$$

where $F_w(s)$ is the distribution function defined in (17) below.

In order to understand the presence of the second derivative in (6) and the second moment formula (7), we recall that TASEP can also be seen as a stochastic growth interface model, whose discrete gradient of the height equals $1-2\eta$. The dynamical rule is that when a particle jumps to the right, a valley ╲╱ changes to a mountain ╱╲. More precisely, let $N_t(j)$ denote the number of particles which have jumped from site j to $j+1$ during the time interval $[0,t]$, and define the height function

$$h_t(j) = \begin{cases} 2N_t(0) + \sum_{i=1}^{j}(1-2\eta_i(t)) & \text{for } j \geq 1, \\ 2N_t(0) & \text{for } j = 0, \\ 2N_t(0) - \sum_{i=j+1}^{0}(1-2\eta_i(t)) & \text{for } j \leq -1. \end{cases} \tag{8}$$

Then initially $h_0(0) = 0$ and $h_0(j) - h_0(j-1) = 1 - 2\eta_j(0)$, and at the instance a particle jumps from site j to $j+1$, the height function at position j increases by two. Note that $h_t(j) - h_0(j) = 2N_t(j)$. It was shown in [14] that the two-point function can be expressed as

$$S(j,t) = \tfrac{1}{8}\big(\varDelta \mathrm{Var}(h_t(\cdot))\big)(j) \tag{9}$$

with $\varDelta$ being the discrete Laplacian, $(\varDelta f)(j) = f(j-1) - 2f(j) + f(j+1)$. Since it is known that $F_w(s)$ has mean, see 0 [6], this explains the presence of the second derivative in the conjectured formula (6) and the second moment formula (7).

Define the probability distribution functions of the location-rescaled height function,

$$\begin{aligned} F_w(s,t) := \mathbb{P}\Big(& h_t([(1-2\rho)t + 2w\chi^{1/3}t^{2/3}]) \\ & \geq (1-2\chi)t + 2w(1-2\rho)\chi^{1/3}t^{2/3} - 2s\chi^{2/3}t^{1/3}\Big). \end{aligned} \tag{10}$$

The function F_w in (7) (which is defined in (17) below) was conjectured in [15] to be the limit

$$\lim_{t\to\infty} F_w(s,t) = F_w(s). \tag{11}$$

The convergence (11) for each s was later proved in [10]. This strongly indicates the validity of (6). A missing part in concluding (6) is the convergence of the moments of $F_w(s,t)$ which is a stronger statement than (11). Our main result is that the moments indeed converge.

Theorem 1. *For all $\ell \in \mathbb{N}$,*

$$\lim_{t\to\infty} \int_{\mathbb{R}} s^{\ell} dF_w(s,t) = \int_{\mathbb{R}} s^{\ell} dF_w(s) \tag{12}$$

uniformly for w in a compact subset of $\mathbb{R}$.

As a consequence we obtain the convergence of the two-point function is a weak sense.

Corollary 2. *We have, with $\chi := \rho(1-\rho)$,*

$$\lim_{t\to\infty} 2\chi^{1/3}t^{2/3} S\left([(1-2\rho)t + 2w\chi^{1/3}t^{2/3}], t\right) = \frac{\chi}{4} g''_{\rm sc}(w) \tag{13}$$

if integrated over smooth functions in w with compact support.

The proof of this corollary is given in Sect. 5. An improvement of the analysis in this paper can yield the convergence in the point-wise sense in (13). However, we do not consider this direction in this paper.

For completeness, let us state a formula of the limiting distribution function $F_w(s)$ explicitly. Let P_u be the orthogonal projector on the interval $[u, +\infty)$. Set

$$\begin{aligned} K_{\mathrm{Ai},s}(x,y) &:= \int_{\mathbb{R}_+} \mathrm{Ai}(x+s+\lambda)\mathrm{Ai}(y+s+\lambda)d\lambda, \\ F_{\rm GUE}(s) &:= \det(\mathbb{1} - P_0 K_{\mathrm{Ai},s} P_0). \end{aligned} \tag{14}$$

$F_{\rm GUE}$ is the GUE Tracy-Widom distribution function [17]. We also define the function

$$g(s,w) := e^{-\frac{w^3}{3}} \left(\int_{\mathbb{R}^2_-} e^{w(x+y)} \mathrm{Ai}(x+y+s)dx\,dy + \int_{\mathbb{R}^2_+} \hat{\Psi}_{w,s}(x)\rho_s(x,y)\hat{\Phi}_{w,s}(y)dx\,dy \right), \tag{15}$$

where

$$\hat{\Phi}_{w,s}(x) := \int_{\mathbb{R}_-} e^{wz+ws} K_{\mathrm{Ai},s}(z,x)dz, \quad \hat{\Psi}_{w,s}(x) := \int_{\mathbb{R}_-} e^{wz}\mathrm{Ai}(x+z+s)dz, \tag{16}$$

and $\rho_s(x,y) := (\mathbb{1} - P_0 K_{\mathrm{Ai},s} P_0)^{-1}(x,y)$. Now

$$F_w(s) := \frac{\partial}{\partial s}\left(F_{\rm GUE}(s+w^2)g(s+w^2,w)\right). \tag{17}$$

There is an alternative formula expressed in terms the Lax pair equations of the Painlevé II equation obtained in [6]. But we will only use the formula (17) in this paper. One can also consider the joint distributions for different values of w and a formula can be found in [5].

2 Setting and Strategy of the Proof

The height function $h_t(j)$ associated to a TASEP with any initial condition can be related to the last passage time of a directed last passage percolation (DLPP) model. Over the last decade or so, the so-called solvable, or determinantal DLPP models [8, 11, 13] were studied extensively. These are the models for which the probability distribution of the last passage time can be expressed explicitly in terms of Fredholm determinants. The DLPP model corresponding to the stationary TASEP is not one of solvable models but can be related to one after suitable analytic continuation of the parameters. This yields the following formula of $F_w(s,t)$.

Fix $w \in \mathbb{R}$. Let us set[2] (recall that $\chi = \rho(1-\rho)$)

$$2m = (1-2\chi)t + 2w(1-2\rho)\chi^{1/3}t^{2/3}, \quad 2d = (1-2\rho)t + 2w\chi^{1/3}t^{2/3}, \tag{18}$$

and define the functions[3]

$$\begin{aligned} L(x,y) &= \frac{-e^{a(x-y)}}{2\pi \mathrm{i}} \oint_{\Gamma_{1-\rho}} e^{-z(x-y)} \frac{(z+\rho)^{m-d}}{(1-\rho-z)^{m+d}} dz \quad \text{for } x > y, \\ R(x,y) &= \frac{e^{a(x-y)}}{2\pi \mathrm{i}} \oint_{\Gamma_{-\rho}} e^{z(y-x)} \frac{(1-\rho-z)^{m+d}}{(\rho+z)^{m-d}} dz \quad \text{for } x < y, \end{aligned} \tag{19}$$

with

$$a := \frac{1}{2} - \rho. \tag{20}$$

We define the kernel

$$K_{m,d}(x,y) = \int_{\mathbb{R}_-} L(x,z)R(z,y)dz \tag{21}$$

and the distribution function

$$F(u) := \det(\mathbb{1} - P_u K_{m,d} P_u). \tag{22}$$

[2] To be precise, we need to take the integer parts of the formulas since m and d need to be integers. Since the error between the formula above and the integer parts is $O(1)$, this does not result in any significant changes in the estimates and hence for convenience we define m and d as in (18) without restricting them to be integers in this paper. However, we remark that if we restrict m and d to be integers, one occasionally needs to be careful in the precise formulation of the estimates and the exposition becomes more involved. We do not discuss these subtleties in this paper.

[3] For any set of points S, the notation $\oint_{\Gamma_S} f(z)\, dz$ denotes the integral over a simple closed contour which encloses the points S but excludes any other poles of the function f. The contour is oriented counter-clockwise.

Finally, we set

$$G_0(u) = g_1(u) + g_2(u) + g_3(u), \tag{23}$$

where

$$\begin{aligned} g_1(u) &= u + \frac{2ad - m}{1/4 - a^2}, \\ g_2(u) &= \langle \psi_a, P_u K_{m,d} \psi_{-a} \rangle, \\ g_3(u) &= \langle K^*_{m,d}(\mathbb{1} - P_u)\psi_a, P_u(\mathbb{1} - P_u K_{m,d} P_u)^{-1} P_u(\mathbb{1} - K_{m,d})\psi_{-a} \rangle, \end{aligned} \tag{24}$$

with $\psi_a(x) = e^{-ax}$. Then it was shown in [10] that[4]

$$F_w(s,t) = \frac{1}{(t/\chi)^{1/3}} \frac{d}{ds} \left(F(u(s,t)) G_0(u(s,t)) \right) \tag{25}$$

where

$$u = u(s,t) := t + s\chi^{-1/3} t^{1/3}. \tag{26}$$

Set

$$\hat{G}_0(s,t) := G_0(u(s,t)), \qquad \hat{F}(s,t) := F(u(s,t)). \tag{27}$$

The main technical part of this paper is on the following estimates[5]:

Proposition 1 (Uniform upper tail estimates). *There exist positive constants s_0, t_0, c and C such that*

$$\left| s - \hat{F}(s,t)\hat{G}_0(s,t)/(t/\chi)^{1/3} \right| \le Ce^{-c|s|}, \qquad s \ge s_0, \quad t \ge t_0. \tag{28}$$

The bound holds uniformly for w in a compact subset of $\mathbb{R}$.

[4] The formula (25) is the formula (4.10) of [10] when $b = -a$ if we take into account (26) . See (5.21) of [10] for the formula of the function $G_0(u) = G^{a,-a}(u)$.

[5] The exponents of the bounds are not optimal. The bound in (28) and (29) can be improved to $Ce^{-c|s|^{3/2}}$ and $Ce^{-c|s|^3}$, respectively. The improved bound for (28) can be achieved if we keep track of a slightly better estimate in the analysis presented in this paper. On the other hand, in order to improve the bound (29), we need a different approach such as Riemann-Hilbert analysis as in [3, 4].

Proposition 2 (Uniform lower tail estimates). *There exist s_0, t_0, c and C such that*

$$|F_w(s,t)| \le Ce^{-c|s|^{3/2}}, \qquad s \le -s_0, \quad t \ge t_0. \tag{29}$$

The bound holds uniformly for w in a compact subset of $\mathbb{R}$.

Theorem 1 now follows.

Proof of Theorem 1. We only consider $\ell \ge 2$. The case $\ell = 1$ follows easily. We first write the integral on the left-hand-side of (12) as the sum of the integral over $\mathbb{R}_+$ and the integral over $\mathbb{R}_-$. For the integral over $\mathbb{R}_+$, integrating by parts twice and using the fact that $F_w(\cdot, t)$ is a cumulative distribution function,

$$\int_{\mathbb{R}_+} s^\ell dF_w(s,t) = -\ell(\ell-1)\int_{\mathbb{R}_+} s^{\ell-2}\Big(s - \hat{F}(s,t)\frac{\hat{G}_0(s,t)}{(t/\chi)^{1/3}}\Big)ds \tag{30}$$

for $\ell \ge 2$. It was in [10] that in addition to (11) we also have the limit

$$\hat{F}(s,t)\frac{\hat{G}_0(s,t)}{(t/\chi)^{1/3}} \to F_{\mathrm{GUE}}(s+w^2)g(s+w^2,w) \tag{31}$$

for each s as $t \to \infty$. Thus, due to Proposition 1, the Lebesgue dominated convergence theorem can be applied and we find that (30) converges to

$$-\ell(\ell-1)\int_{\mathbb{R}_+} s^{\ell-2}\Big(s - F_{\mathrm{GUE}}(s+w^2)g(s+w^2,w)\Big)ds. \tag{32}$$

On the other hand, integrating by parts once gives

$$\int_{\mathbb{R}_-} s^\ell dF_w(s,t) = -\ell\int_{\mathbb{R}_-} s^{\ell-1}F_w(s,t)ds. \tag{33}$$

Thus, again, the Lebesgue dominated convergence theorem can be applied due to Proposition 2 and from (11) we find that (33) converges to

$$-\ell\int_{\mathbb{R}_-} s^{\ell-1}F_w(s)ds. \tag{34}$$

Integrating (32) and (34) by parts backwards and using the fact that F_w is a cumulative distribution function, we find that the sum of these two integrals is the right-hand-side of (12). □

The estimate (28) for the upper tail is obtained by analyzing the formulas (22) and (23) asymptotically using the saddle-point analysis. This asymptotic analysis is very close to that of many previous papers, for example [2, 10, 11]. We use some of the results directly or improve upon them. See Sect. 3.

For the estimate (29) on the lower tail, we note the following. Consider the TASEP with step-initial condition i.e. $\eta_j(0) = 1$ for $j \le 0$ and $\eta_j(0) = 0$ for $j \ge 1$. Then the associated height function $h_t^{\text{step}}(j)$ satisfies $h_0^{\text{step}}(j) = |j|$. This means that initially h_0 is bounded from above by h_0^{step}. Since the initial condition of the stationary TASEP is independent of the dynamics, we find that h_t is stochastically bounded above[6] by h_t^{step}. Hence[7]

$$\mathbb{P}(h_t(j) \ge u) \le \mathbb{P}(h_t^{\text{step}}(j) \ge u). \tag{35}$$

But $\mathbb{P}(h_t^{\text{step}}(j) \ge u)$ is known to be precisely $F(u)$ of (22) [11]. Therefore we have

$$F_w(s,t) \le \hat{F}(s,t) = \det(1 - P_u K_{m,d} P_u). \tag{36}$$

Thus, the estimate (29) follows if we show that $\hat{F}(s,t)$ is bounded from above by $Ce^{-c|s|^{3/2}}$ for negative large enough s. This in turn follows if we show the same bound for the Fredholm determinant (22). For this purpose we follow the idea of Widom [19] which seems not as well-known as it should be. See Sect. 4.

3 Proof of Proposition 1: Upper Tail

The proposition follows from (40), (38) and (52), see below.

3.1 Asymptotics for $\hat{F}$

The function $F(u) = \det(1 - P_u K_{m,d} P_u)$ is the distribution function of the last passage time of the directed last passage model with i.i.d. exponential random variables. It is well-known [11] that this also equals the distribution function of the largest eigenvalue of the Laguerre unitary ensemble (LUE) which is defined as $M_{m,d} = \frac{1}{m-d} XX^*$ where X is a $(m-d) \times (m+d)$ random matrix with i.i.d. standard complex Gaussian entries. This equality can also be seen explicitly in Appendix C of [10] where $K_{m,d}$ was shown to be same as the correlation kernel of the LUE up to a conjugation by a multiplication. The asymptotics of LUE and

[6] This can also be seen easily from the corresponding directed last passage percolation (DLPP) models. The DLPP model for the stationary TASEP is the DLPP model for the TASEP with the step initial condition plus an extra row and an extra column with non-zero weights.

[7] We would like to thank Ivan Corwin and Eric Cator for communicating this observation. This observation simplified the proof of the lower tail estimate which we originally obtained by estimating $F_w(s,t)$ directly.

$\hat{F}(s,t) = F(u(s,t))$ were considered in several papers, especially in [2, 10, 11]. We have:

Lemma 1. *Fix $s_0 \in \mathbb{R}$. Then*

$$\lim_{t\to\infty} \hat{F}(s,t) = F_{\text{GUE}}(s+w^2) \tag{37}$$

uniformly for $s \in [s_0,\infty)$ and w in a compact subset of $\mathbb{R}$. Furthermore, for given $s_0 \in \mathbb{R}$ and $t_0 > 0$, there exist positive constants C and c such that

$$|1-\hat{F}(s,t)| \le Ce^{-cs} \tag{38}$$

for $s \ge s_0$ and $t \ge t_0$.

The bound (38) can be found in, for example, Sect. 3.1 of [2].[8]

3.2 Evaluation of g_1

A direct computation using (18), (20), and (26) shows that[9]

$$g_1(u) = u + \frac{2ad-m}{1/4-a^2} = s(t/\chi)^{1/3}. \tag{39}$$

This implies that

$$s - \frac{\hat{F}(s,t)\hat{G}_0(s,t)}{(t/\chi)^{1/3}} = -\hat{F}(s,t)\frac{g_2(u)+g_3(u)}{(t/\chi)^{1/3}} + s(1-\hat{F}(s,t)). \tag{40}$$

The term $1-\hat{F}(s,t)$ can be estimated using (38) and $\hat{F}(s,t)$ is bounded by 1 since it is a distribution function. We now show that $g_2(u)/(t/\chi)^{1/3}$ and $g_3(u)/(t/\chi)^{1/3}$ are uniformly (in t) bounded by exponentially decaying functions in s.

In the rest of this section, we only consider the case when $w > 0$. If $w < 0$, we need to start with a different decomposition of $G_0(u)$ ((5.22) instead of (5.21) of [10]). After this change, the analysis is completely analogous. For the case when

[8] The exponent of the upper bound is not optimal: the optimal exponent is $e^{-c|s|^{3/2}}$. But we do not consider such an issue in this paper.

[9] The formula becomes $s(t/\chi)^{1/3} + O(1)$ where $O(1)$ is independent of s if we take the integer parts in the definition of m and d in (18). This is an example of the subtleties mentioned in the Footnote 2. This results in the additional term $O(t^{-1/3})$ in (40). Since this is not a function in s, we cannot obtain the bound (C1). However, this issue can be fixed by shifting s to $s - O(1)/(t/\chi)^{1/3}$. In other words, the centering and scaling $u = t + s(t/\chi)^{1/3}$ needs to be changed slightly to reflect the difference of the formula of (18) and their integer counter-parts.

$w = 0$, we can proceed as in the case when $w > 0$ but with a yet slight modification: see (6.31)–(6.34) of [10]. We skip the detail when $w < 0$ and $w = 0$, and assume from now on that $w > 0$.

3.3 Estimations on g_2 and g_3

Recall the definition (24) of $g_2(u)$. It is a direct calculation to show that (see (3.15) of [10])

$$\int_x^\infty R(x,y)\psi_{-a}(y)dy = Z(\rho)\psi_{-a}(x), \qquad Z(\rho) := \frac{(1-\rho)^{m+d}}{\rho^{m-d}}, \tag{41}$$

for $x \in \mathbb{R}$, for $a \in (-1/2, 1/2)$. Using this, $g_2(u) = \langle \psi_a, P_u L(\mathbb{1} - P_0)\psi_{-a}\rangle$. Inserting the formula ψ_a and $L(x,y)$, we obtain

$$g_2(u) = \int_{\mathbb{R}_+^2} \mathscr{H}_t(x+y)dx\,dy, \tag{42}$$

where

$$\mathscr{H}_t(x) := \frac{-Z(\rho)}{2\pi \mathrm{i}} \oint_{\Gamma_{1-\rho}} e^{-z(u+x)} \frac{(z+\rho)^{m-d}}{(1-\rho-z)^{m+d}} dz. \tag{43}$$

Thus (see (6.19) of [10])

$$(t/\chi)^{-1/3} g_2(u) = \int_{\mathbb{R}_+^2} H_t(x+y)dx\,dy, \quad H_t(y) := (t/\chi)^{1/3}\mathscr{H}_t(y(t/\chi)^{1/3}). \tag{44}$$

Similarly, recall the definition (24) of $g_3(u)$. Using (41), an argument similar to that for (42) implies that

$$(\mathbb{1} - K_{m,d})\psi_{-a}(x) = e^{ax}\left[1 - \int_{\mathbb{R}_+} \mathscr{H}_t(-u+x+y)dy\right]. \tag{45}$$

We also note that, similar to (41), we have (see (3.15) of [10])

$$\int_x^\infty \psi_a(y)L(y,x)dy = \frac{1}{Z(\rho)}\psi_a(x) \tag{46}$$

for $x \in \mathbb{R}$, for $a \in (-1/2, 1/2)$. Using this, we find that

$$K^*_{m,d}(1-P_u)\psi_a(x) = e^{-ax}\Big[\int_{\mathbb{R}_+}\tilde{\mathscr{H}}_t(-u+x+y)dy - \int_{\mathbb{R}^2_+}\tilde{\mathscr{H}}_t(-u+x+y)\mathscr{H}_t(z+y)dz\,dy\Big] \tag{47}$$

where

$$\tilde{\mathscr{H}}_t(x) := \frac{1}{2\pi\mathrm{i}\,Z(\rho)}\oint_{\Gamma_{-\rho}} e^{z(u+x)}\frac{(1-\rho-z)^{m+d}}{(z+\rho)^{m-d}}dz. \tag{48}$$

This implies that we can express (see (6.26)–(6.28) in [10])

$$(t/\chi)^{-1/3}g_3(u) = \langle\Phi_t, A_t\Psi_t\rangle \tag{49}$$

where

$$\begin{aligned}\Phi_t(\xi) &= e^{w\xi}\Big[\int_{\mathbb{R}_+}\tilde{H}_t(y+\xi)dy - \int_{\mathbb{R}^2_+}H_t(x+y)\tilde{H}_t(y+\xi)dx\,dy\Big],\\ \Psi_t(\xi) &= e^{-w\xi}\Big[1-\int_{\mathbb{R}_+}H_t(y+\xi)dy\Big],\end{aligned} \tag{50}$$

with $H_t(y) = (t/\chi)^{1/3}\tilde{\mathscr{H}}(y(t/\chi)^{1/3})$ and $\tilde{H}_t(y) = (t/\chi)^{1/3}\tilde{\mathscr{H}}_t(y(t/\chi)^{1/3})$. Here the operator A_t is defined by $A_t = P_0(\mathbb{1}-K_t)^{-1}P_0$ where the kernel of K_t is

$$K_t(\xi_1,\xi_2) = e^{w(\xi_2-\xi_1)}\int_{\mathbb{R}_+}H_t(x+\xi_1)\tilde{H}_t(x+\xi_2)dx, \qquad \xi_1,\xi_2 \geq 0, \tag{51}$$

and $K_t(\xi_1,\xi_2) = 0$ otherwise.

We obtain the following estimates for g_2 and g_3.

Lemma 2. *There are positive constants c and C such that*

$$\Big|(t/\chi)^{-1/3}g_2(u)\Big| \leq Ce^{-cs}, \qquad \Big|(t/\chi)^{-1/3}g_3(u)\Big| \leq Ce^{-cs} \tag{52}$$

for all $s \geq 0$ and $t \geq 0$.

Proof of Lemma 2. Note from the formula (43) that $\mathscr{H}_t(x) = \mathscr{H}_t(x;u)$ is a function of $x+u$. Hence $H_t(y) = H_t(y;s)$ is a function of $y+s$. Thus, $H_t(y;s) = H_t(y+s;0)$. The same holds for $\tilde{H}_t(y) = \tilde{H}_t(y;s)$.

Basic bounds for the functions $H_t(y)$ and $\tilde{H}_t(y)$ were obtained in (6.15) of [10]: for any $\beta > 0$ there exist positive constants C_β and C'_β such that

$$|H_t(y;s)| \le C_\beta e^{-\beta y} \quad \text{and} \quad |\tilde{H}_t(y;s)| \le C'_\beta e^{-\beta y} \tag{53}$$

uniformly for $t \ge 0$, $y \ge 0$, and $s \ge 0$. In particular, the bound holds for $H_t(y;0)$ and $\tilde{H}_t(y;0)$ when $s = 0$, for $t \ge 0$ and $y \ge 0$. Thus using $H_t(y;s) = H_t(y+s;0)$ and inserting $y+s$ in place of y in (53), we find that for any $\beta > 0$ there are positive constants C_β and C'_β such that

$$|H_t(y;s)| \le C_\beta e^{-\beta(s+y)} \quad \text{and} \quad |\tilde{H}_t(y;s)| \le C'_\beta e^{-\beta(s+y)} \tag{54}$$

uniformly in $t \ge 0$, $y \ge 0$ and $s \ge 0$.

The bound for $(t/\chi)^{-1/3} g_2(u)$ follows from (44) and (54).

We now estimate $|(t/\chi)^{-1/3} g_3(u)|$. Choosing $\beta > |w|$, (54) implies that

$$|\Phi_t(\xi)| \le C e^{-\beta s} e^{-(\beta - w)\xi}$$

for a positive constant C. Thus,

$$||\Phi_t||_{L^2(\mathbb{R}_+)} \le C' e^{-\beta s}, \tag{55}$$

for a constant C' uniformly in $t \ge 0$ and $s \ge 0$. On the other hand, (54) implies that $|\Psi_t(\xi) e^{w\xi}|$ is bounded by a constant. Since we assume $w > 0$ (see Sect. 3.2), we find that $||\Psi_t||^2_{L^2(\mathbb{R}_+)}$ is uniformly bounded in $t \ge 0$ and $s \ge 0$. Finally, the inequality

$$||A_t|| \le ||(\mathbb{1} - K_{\mathrm{Ai}, w^2+s})^{-1}|| + ||(\mathbb{1} - K_{\mathrm{Ai}, w^2+s})^{-1} - (\mathbb{1} - K_t)^{-1}|| \tag{56}$$

where K_{Ai, w^2+s} is the Airy kernel restricted on (w^2+s, ∞), and the fact (see (6.36) of [10]) that $||(\mathbb{1} - K_{\mathrm{Ai}})^{-1} - (\mathbb{1} - K_t)^{-1}|| \to 0$ as $t \to \infty$ imply that $||A_t||$ is uniformly bounded in $t \ge 0$ and $s \ge 0$. Therefore, the bound for $(t/\chi)^{-1/3} g_3(u)$ follows from $|\langle \Phi_t, A_t \Psi_t \rangle| \le ||\Phi_t|| \, ||A_t|| \, ||\Psi_t||$. □

4 Proof of Proposition 2: Lower Tail

Recall from Sect. 3.1 that $K_{m,d}$ is a similarity transform of the correlation kernel of the LUE $M_{m,d}$. Since the correlation kernel of the LUE is a positive projection, all the eigenvalues, which we denote by μ_j, $j = 0, 1, 2, \cdots$, of $P_u K_{md} P_u$ are real and $\mu_j \in [0,1]$. It was shown in Appendix B.3 of [10] that $\mu_j \in [0,1)$ if $u > 0$. From this we find that $\det(1 - P_u K_{m,d} P_u) = \prod_{j\ge 0}(1-\mu_j) \le \prod_{j \ge 0} e^{-\mu_j} = e^{-\mathrm{Tr}\,(P_u K_{m,d} P_u)}$. Therefore,

$$\hat{F}(s,t) \le \exp(-\mathrm{Tr}\,(P_u K_{m,d} P_u)). \tag{57}$$

This trick is due to Widom [19].

The trace has the following lower bound:

Proposition 3. *There exist positive constants t_0, s_0, c such that*

$$\mathrm{Tr}\,(P_u K_{m,d} P_u) \geq c|s|^{3/2} \tag{58}$$

for all $s \leq -s_0$ and $t \geq t_0$.

The same estimate was obtained in the context of random permutations and an oriented digital boiling model by Widom [19]. We follow the paper [19] to prove the Proposition, and as such we only sketch the main ideas and do not provide all the details of the proof. Once this proposition is proven, then Proposition 2 follows from (36) and (57).

Proof of Proposition 3. Since the operator K_t is trace class with continuous kernel, we have

$$\begin{aligned}\mathrm{Tr}\,(K_t) = \int_{\mathbb{R}_+} K_t(x,x)dx &= \frac{-1}{(2\pi \mathrm{i})^2} \oint_{\Gamma_1} \oint_{\Gamma_0} \frac{e^{wu}}{e^{zu}} \frac{(1-w)^{m+d}}{(1-z)^{m+d}} \frac{z^{m-d}}{w^{m-d}} \frac{dw\,dz}{(w-z)^2} \\ &= \frac{-1}{(2\pi \mathrm{i})^2} \oint_{\Gamma_1} \oint_{\Gamma_0} \frac{e^{MF_u(w)}}{e^{MF_u(z)}} \frac{dw\,dz}{(w-z)^2}\end{aligned} \tag{59}$$

where

$$F_u(z) := u'z - \ln z + \gamma \ln(1-z), \qquad u' := \frac{u}{M}. \tag{60}$$

Here $M := m-d$ and $\gamma := \frac{m+d}{m-d} = (\frac{1-\rho}{\rho})^2 + O(t^{-1/3})$. Note that $u' = \frac{1}{\rho^2} + O(t^{-1/3})$ if s is in a bounded set, and $u' \geq \frac{1}{\rho^2}$ for all $s \leq 0$. We analyze (59) asymptotically using the saddle-point analysis. Note the presence of the singularity $\frac{1}{(w-z)^2}$ in the integrand.

We first consider the case where

$$(1-\sqrt{\gamma})^2 + \epsilon \leq u' < (1+\sqrt{\gamma})^2 - s_0 t^{-2/3}, \tag{61}$$

for some $\epsilon > 0$ (small, but fixed) and $s_0 \gg 1$ also fixed. The critical points are F_u are

$$z_c^{\pm}(u') = \frac{u'+1-\gamma}{2u'} \pm \frac{1}{2u'}\sqrt{(u'-(1+\sqrt{\gamma})^2)(u'-(1-\sqrt{\gamma})^2)}. \tag{62}$$

The two critical points are non-real and $|z_c^{\pm}(u')| = \frac{1}{\sqrt{u'}} \leq \rho < 1$. Consider the following two contours:

$$w = |z_c^{+}|e^{\mathrm{i}\theta}, \qquad 0 \leq \theta < 2\pi, \tag{63}$$

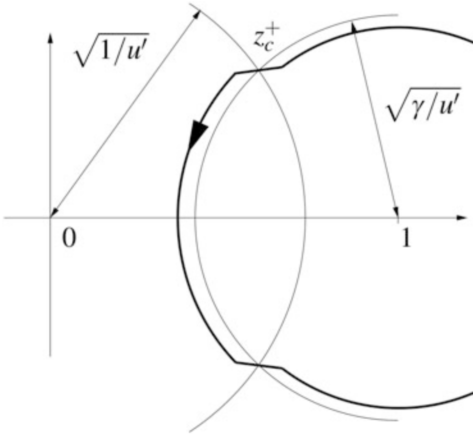


Fig. 1 The bold path Γ is the deformation of Γ_1 that locally follows the steepest descent path

and

$$z = 1 + |z_c^+ - 1|e^{\mathrm{i}\theta}, \qquad 0 \le \theta < 2\pi, \tag{64}$$

respectively. Then

$$\mathrm{Re}\left(\frac{d}{d\theta}F_u(w)\right) = -\mathrm{Im}(w)\left(u' - \frac{\gamma}{|w-1|^2}\right) = -\mathrm{Im}(w)\left(\frac{\gamma}{|z_c^+ - 1|^2} - \frac{\gamma}{|w-1|^2}\right),$$

$$\mathrm{Re}\left(\frac{d}{d\theta}F_u(z)\right) = -\mathrm{Im}(z)\left(u' - \frac{\gamma}{|z|^2}\right) = -\mathrm{Im}(z)\left(\frac{\gamma}{|z_c^+|^2} - \frac{\gamma}{|z|^2}\right). \tag{65}$$

Thus, along these contours, Re (F_u) achieves its relative maximum (resp. minimum) at $z_c^{\pm}$. Hence these paths are of steep-ascent and steep-descent for F_u. We chose to work with these explicit contours instead of the contours of steepest-ascent and steepest-descent for convenience. Due to this reason, we need to modify the contours locally near the critical points if u' is close to $(1 + \sqrt{\gamma})^2$. Namely, in this case, the contours above become almost tangential and are almost parallel to the direction along which $\mathrm{Re}(F_u)$ is constant. Then we cannot apply the saddle-point method. In this case, we simply modify the contours locally near the critical points so that they pass through the critical points along the steepest descent direction as pictured in Fig. 1 for the z-contour. A similar modification is needed for the w-contour. This small modification does not yield any significant changes in the estimation. For the convenience of presentation, we work with the above explicit contours and skip the details on how the formulas changes after the modifications. The same procedure was also explained in Sect. 6.2 of [7] for the similar estimations.

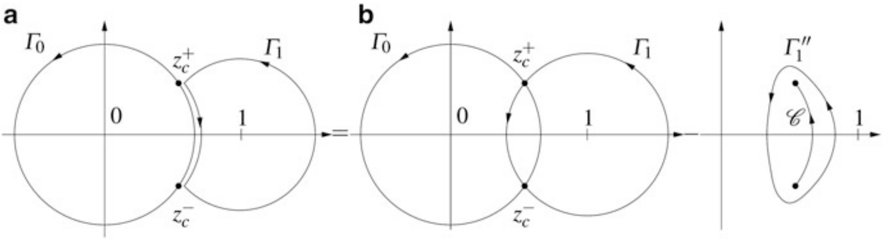


Fig. 2 The subdivision of the integration from (**a**) the ones in (59) to (**b**) the ones in (66)

We now deform the original contours in (59) to the new contours of steepest-ascent and steepest-descent, which we call by the same names, Γ_0 and Γ_1. We first deform the original contours to those in (a) of Fig. 2 where Γ_0 is the contour of steepest-ascent and the part of Γ_1 except for the segment from z_c^- to z_c^+ is the part of the contour of steepest-descent. These contours can be divided as in (b) of Fig. 2 and we have

$$(59) = \frac{-1}{(2\pi \mathrm{i})^2} P.V. \oint_{\Gamma_0} \oint_{\Gamma_1} \frac{e^{MF_u(w)}}{e^{MF_u(z)}} \frac{dz\, dw}{(w-z)^2} + \frac{1}{(2\pi \mathrm{i})^2} \int_{\mathscr{C}} \oint_{\Gamma_1''} \frac{e^{MF_u(w)}}{e^{MF_u(z)}} \frac{dz\, dw}{(w-z)^2}. \tag{66}$$

Here the first integral needs to be interpreted as the Principal Value due to the divergent terms in the integrand. The second integral is from the contributions of the pole in the deformation of the contours. The contours in the second double integral are defined as follows. The w-contour, $\mathscr{C}$, is a segment from z_c^- to z_c^+ to the left of 1 and to the right of 0. The z-contour, Γ_1'', encircles the whole segment $\mathscr{C}$ but not 1, see Fig. 2.

Setting $Q(z) := \exp(MF_u(z))$, the Cauchy's integral formula implies that the second integral of (66) equals

$$\frac{-1}{2\pi \mathrm{i}} \int_{\mathscr{C}} \frac{Q'(w)}{Q(w)} dw = \frac{-M\left(F_u(z_c^+) - F_u(z_c^-)\right)}{2\pi \mathrm{i}}. \tag{67}$$

Noting that $F_u(z_c^+) = \overline{F_u(z_c^-)}$, we have

$$\frac{1}{(2\pi \mathrm{i})^2} \int_{\mathscr{C}} \oint_{\Gamma_1''} \frac{e^{MF_u(w) - MF_u(z)}}{(w-z)^2} dz\, dw = \frac{-M\, \mathrm{Im}(F_u(z_c^+))}{\pi}. \tag{68}$$

Observe that when $u' = (1 + \sqrt{\gamma})^2$, the two critical points coincide and we have $z_c := z_c^\pm = \frac{1}{1+\sqrt{\gamma}}$. In addition, $F_{M(1+\sqrt{\gamma})^2}(z_c) \in \mathbb{R}$. Thus

$$
\begin{aligned}
-\mathrm{Im}(F_u(z_c^+)) &= \mathrm{Im}(F_{M(1+\sqrt{\gamma})^2}(z_c)) - \mathrm{Im}(F_u(z_c^+)) \\
&= \int_u^{M(1+\sqrt{\gamma})^2} \mathrm{Im}\frac{d}{dv}F_v(z_c^+(v))dv.
\end{aligned}
\tag{69}
$$

Using the definition (60) of F_u and the fact that $z_c^+(\nu)$ is a critical value, we find that $\frac{d}{dv}F_v(z_c^+(v)) = \frac{1}{M}z_c^+(v)$. From the formula (62) of z_c^+,

$$
\begin{aligned}
\mathrm{Im}(z_c^+(v'M)) &= \frac{\sqrt{\nu' - (1-\sqrt{\gamma})^2}}{2\nu'}((1+\sqrt{\gamma})^2 - \nu')^{1/2} \\
&\geq \frac{\epsilon}{2}((1+\sqrt{\gamma})^2 - \nu')^{1/2}
\end{aligned}
\tag{70}
$$

since ν' satisfies the condition (61). Therefore, (69) implies that

$$
(68) \geq M\frac{\epsilon\pi}{3}((1+\sqrt{\gamma})^2 - u')^{3/2}. \tag{71}
$$

Recall that $\sqrt{\gamma} = (1-\rho)/\rho + O(t^{-1/3})$, that $M = \rho^2 t(1+O(t^{-1/3}))$, and that $u' = u/M$ with $u = t + s(t/\chi)^{1/3}$. Then, we can choose a $s_0 > 0$ large enough (but fixed independently of t) such that for all $s \leq -s_0$ it holds $(1+\sqrt{\gamma})^2 - u' \geq -c_1 s t^{-2/3}$ for some $c_1 > 0$. Therefore for u' satisfying (61), there is a positive constant c such that

$$
\frac{1}{(2\pi\mathrm{i})^2}\int_{\mathscr{C}}\oint_{\Gamma_1''}\frac{e^{MF_u(w)-MF_u(z)}}{(w-z)^2}dz\,dw \geq c(-s)^{3/2} \tag{72}
$$

uniformly in t.

We now show that the contribution of the Principal Value integral in (66) is much smaller than (72). Indeed we will show that this is $O(1)$. This proves (58) by taking the constant c smaller than one in (72).

A direct computation shows that

$$
F_u''(z_c^+) = (1-\gamma)\frac{(z_c^+ - \frac{1}{1+\sqrt{\gamma}})(z_c^+ - \frac{1}{1-\sqrt{\gamma}})}{(z_c^+)^2(z_c^+-1)^2}. \tag{73}
$$

This implies

$$
|F_u''(z_c^+)| \sim ((1+\sqrt{\gamma})^2 - u')^{1/2} \sim s^{1/2}t^{-1/3} \tag{74}
$$

as $u' \to (1+\sqrt{\gamma})^2 - st^{-2/3}$, while for $(1-\sqrt{\gamma})^2 + \epsilon \leq u' < (1+\sqrt{\gamma})^2 - \epsilon$ we have $|F_u''(z_c^+)| = O(1)$. Thus, for the general u' satisfying (61), $c_1 t^{-1/3} \leq |F_u''(z_c^+)| \leq c_2$ for some positive constants c_1 and c_2. Hence $O(t^{1/3}) \leq \sqrt{MF_u''(z_c^+)} \leq O(t^{1/2})$.

Let us choose the parts $V_0(z_c^\pm)$ and $V_1(z_c^\pm)$ of the paths $\varGamma_0$ and $\varGamma_1$, respectively, whose size are $B(MF_u''(z_c^+))^{-1/2}$. Then both parts become smaller as $t \to \infty$. Because the paths $\varGamma_0$ and $\varGamma_1$ are chosen to be steep-descent, the contribution coming from $\varGamma_0 \times \varGamma_1 \setminus \{V_0(z_c^+) \cup V_0(z_c^-)\} \times \{V_1(z_c^+) \cup V_1(z_c^-)\}$ is at most of order $O(1)$ if B is chosen large enough (but fixed). Let us first consider the contributions from the intersecting contours $V_0(z_c^+) \times V_1(z_c^+)$ and $V_0(z_c^-) \times V_1(z_c^-)$. Due to the symmetry, it is enough to consider the contribution of $V_0(z_c^+) \times V_1(z_c^+)$, given by

$$I(z_c^+) := \frac{-1}{(2\pi \mathrm{i})^2} P.V. \int_{V_0(z_c^+)} \int_{V_1(z_c^+)} e^{MF_u(w) - MF_u(z)} \frac{1}{(w-z)^2} dz\, dw. \tag{75}$$

Now we have to see if this integral is bounded by a constant. Since z converges to z_c^+, we use the Taylor's series of F_u in $z - z_c^+$. Since z_c^+ is a critical point, the function $F_u(z)$ in the exponent may be approximated as $F_u(z_c^+) + \frac{1}{2}F_u''(z_c)(z - z_c^+)^2$. It can be checked that the contributions from the higher order terms are negligible. Changing the variables as $z = z_c^+ + z'(MF_u''(z_c^+))^{-1/2}$, $w = z_c^+ + w'(MF_u''(z_c^+))^{-1/2}$, we obtain

$$I(z_c^+) \approx \frac{-1}{(2\pi \mathrm{i})^2} P.V. \int_{\mathrm{i}\mathbb{R}} \int_{\mathbb{R}} \frac{e^{\frac{1}{2}(w'^2 - z'^2)}}{(w'-z')^2} dz'\, dw' \tag{76}$$

which is finite.

Let us now show that the contribution of the non-intersecting contours

$$V_0(z_c^\pm) \times V_1(z_c^\mp)$$

are also bounded from above by some constant. To that aim, B being fixed, we assume that s_0 is chosen large enough so that $s_0 \gg B$. This time the singularity term $1/|w-z|^2$ is bounded from above and one can easily deduce that for all $u' \in ((1-\sqrt{\gamma})^2 + \epsilon, (1+\sqrt{\gamma})^2 - s_0 t^{-2/3})$,

$$\left| \frac{-1}{(2\pi \mathrm{i})^2} P.V. \int_{V_0(z_c^\pm)} \int_{V_1(z_c^\mp)} e^{MF_u(w) - MF_u(z)} \frac{1}{(w-z)^2} dz\, dw \right| \leq O(1).$$

Combining the whole, we have shown that the contributions from the first integral in (66) is $O(1)$ and (58) is proved for $u' \in [(1-\sqrt{\gamma})^2 + \epsilon, (1+\sqrt{\gamma})^2 - s_0 t^{-2/3})$.

We now consider the case where

$$u' \in (0, (1-\sqrt{\gamma})^2 + \epsilon). \tag{77}$$

In (61), we could have chosen $\epsilon > 0$ small enough so that

$$(1-\sqrt{\gamma})^2 + \epsilon < \frac{1}{2}\big((1+\sqrt{\gamma})^2 + (1-\sqrt{\gamma})^2\big). \tag{78}$$

Consider the Laguerre Unitary Ensemble $\frac{1}{m-d}XX^*$ where X is a $(m-d)\times(m+d)$ random matrix with i.i.d. complex standard Gaussian entries. Denote by $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{m-d}$ its ordered eigenvalues. By the definition of the correlation kernel K_t, we have

$$\operatorname{Tr}(K_t) = \mathbb{E}\left(\sum_{i=1}^{m-d} \mathbb{1}_I(\lambda_i)\right), \tag{79}$$

where $I = (u', +\infty)$. This can be bounded below as

$$\mathbb{E}\left(\sum_{i=1}^{m-d} \mathbb{1}_I(\lambda_i)\right) \geq \mathbb{E}\left(\sum_{i=1}^{m-d} \mathbb{1}_{I_\epsilon}(\lambda_i)\right), \tag{80}$$

where $I_\epsilon = ((1-\sqrt{\gamma})^2 + \epsilon, +\infty)$. Now, we call on the results of [1], giving convergence rates for the spectral distribution of random sample covariance matrices. Let $\mathbf{F}_{m-d}$ denote the empirical probability distribution function associated to the spectral measure:

$$\mathbf{F}_{m-d}(x) = \frac{1}{m-d}\sum_{i=1}^{m-d} \mathbb{1}_{\lambda_i \leq x}. \tag{81}$$

Let also $\mathbf{F}$ be the cumulative distribution function of the Marchenko-Pastur distribution ϱ defined by the density

$$\frac{d\varrho}{dx} = \frac{\sqrt{(u_+^c - x)(x - u_-^c)}}{2\pi x}\mathbb{1}_{[u_-^c, u_+^c]}(x), \tag{82}$$

where $\gamma = \frac{m+d}{m-d}$ and $u_c^\pm = (1 \pm \sqrt{\gamma})^2$. It is well known that $\mathbf{F}_{m-d}(x) \to \mathbf{F}(x)$ a.s. for all x. In [1] it is proven that

$$\max_{x>0} |\mathbb{E}(\mathbf{F}_{m-d}(x)) - \mathbf{F}(x)| \leq (m-d)^{-1/2}. \tag{83}$$

Then (79) and (80) imply that

$$\operatorname{Tr}(K_t) \geq (m-d)(1 - \mathbf{F}((1-\sqrt{\gamma})^2 + \epsilon)) - (m-d)^{1/2}. \tag{84}$$

With the condition (78) on ϵ, $\mathbf{F}((1-\sqrt{\gamma})^2 + \epsilon) < 1$ uniformly and since $m - d = \rho^2 t + O(t^{2/3}) \to \infty$, we find that there exists a positive constant $C = C(\epsilon)$ such that

$$\operatorname{Tr}(K_t) \geq C(m-d) \tag{85}$$

uniformly in t and for u' satisfying (77). Now as $u = t - s(t/\chi)^{1/3}$ and $u \geq 0$, we have $(-s)^{3/2} = (t-u)^{3/2}(\chi/t)^{1/2} \leq \chi^{1/2}t$. Thus since $m - d = \rho^2 t + O(t^{2/3})$, (85) implies that there exists a positive constant c such that

$$\mathrm{Tr}\,(K_t) \geq c(-s)^{3/2} \tag{86}$$

uniformly in t and for u' satisfying (77). Thus (58) is proved for u' satisfying (77). This completes the proof of Proposition 3. □

5 Proof of Corollary 2

Let us consider the rescaled height function

$$H_t(w) := \frac{h_t(j(w)) - [(1-2\chi)t + 2w(1-2\rho)\chi^{1/3}t^{2/3}]}{-2\chi^{2/3}t^{1/3}}, \tag{87}$$

with $j(w) = (1-2\rho)t + 2w\chi^{1/3}t^{2/3}$. By (10), $F_w(s,t) = \mathbb{P}(H_t(w) \leq s)$. We have:

$$G_t(w) := \mathrm{Var}(H_t(w)) = \int_{\mathbb{R}} s^2 dF_w(s,t) - \left(\int_{\mathbb{R}} s dF_w(s,t)\right)^2, \tag{88}$$

and, in the original variables,

$$\mathrm{Var}(h_t(j(w))) = (2\chi^{2/3}t^{1/3})^2 G_t(w). \tag{89}$$

Using the notation $\delta := (2\chi^{1/3}t^{2/3})^{-1}$, by (9)

$$\begin{aligned}\int_{\mathbb{R}} 2\chi^{1/3}t^{2/3}S(j(w),t)f(w)dw &= \frac{\chi}{4}\int_{\mathbb{R}} \frac{G_t(w+\delta) - 2G_t(w) + G_t(w-\delta)}{\delta^2} f(w)dw \\ &= \frac{\chi}{4}\int_{\mathbb{R}} G_t(w)\frac{f(w+\delta) - 2f(w) + f(w-\delta)}{\delta^2} dw.\end{aligned} \tag{90}$$

By Theorem 1 and the fact that $\int_{\mathbb{R}} s\, dF_w(s) = 0$ (see [6]), we have that $G_t(w)$ converges to $g_{\mathrm{sc}}(w)$ uniformly for w in a compact set of $\mathbb{R}$. Therefore, for smooth test functions f with compact support, as $t \to \infty$ this expression converges to

$$\frac{\chi}{4}\int_{\mathbb{R}} g_{\mathrm{sc}}(w)f''(w)dw = \frac{\chi}{4}\int_{\mathbb{R}} g''_{\mathrm{sc}}(w)f(w)dw. \tag{91}$$

□

Acknowledgements We would like to thank Ivan Corwin and Eric Cator for a communication which helped us simplify the proof of Proposition 2. The work of Jinho Baik was supported in part by NSF grants DMS1068646. Patrik Lino Ferrari was supported by the German Research Foundation via the SFB611–A12 project.

References

1. Bai, Z.D., Miao, B., Yao, J.: Convergence rates of spectral distributions of large sample covariance matrices. SIAM J. Matrix Anal. Appl. **25**, 105–127 (2003)
2. Baik, J., Ben Arous, G., Péché, S.: Phase transition of the largest eigenvalue for non-null complex sample covariance matrices. Ann. Probab. **33**, 1643–1697 (2006)
3. Baik, J., Deift, P., Johansson, K.: On the distribution of the length of the longest increasing subsequence of random permutations. J. Am. Math. Soc. **12**, 1119–1178 (1999)
4. Baik, J., Deift, P., McLaughlin, K., Miller, P., Johansson, K.: Optimal tail estimates for directed last passage site percolation with geometric random variables. Adv. Theor. Math. Phys. **5**, 1207–1250 (2002)
5. Baik, J., Ferrari, P.L., Péché, S.: Limit process of stationary TASEP near the characteristic line. Commun. Pure Appl. Math. **63**, 1017–1070 (2010)
6. Baik, J. Rains, E.M.: Limiting distributions for a polynuclear growth model with external sources. J. Stat. Phys. **100**, 523–542 (2000)
7. Borodin, A., Ferrari, P.L.: Anisotropic growth of random surfaces in $2 + 1$ dimensions. Commun. Math. Phys. (to appear)
8. Borodin, A., Péché, S.: Airy kernel with two sets of parameters in directed percolation and random matrix theory. J. Stat. Phys. **132**, 275–290 (2008)
9. Ferrari, P.A.: Shock fluctuations in asymmetric simple exclusion. Probab. Theory Relat. F. **91** (1992), 81–101
10. Ferrari, P.L., Spohn, H.: Scaling limit for the space-time covariance of the stationary totally asymmetric simple exclusion process. Commun. Math. Phys. **265**, 1–44 (2006)
11. Johansson, K.: Shape fluctuations and random matrices. Commun. Math. Phys. **209**, 437–476 (2000)
12. Liggett, T.M.: Coupling the simple exclusion process. Ann. Probab. **4**, 339–356 (1976)
13. Okounkov, A.: Infinite wedge and random partitions. Selecta Math. (N.S.) **7**, 57–81 (2002)
14. Prähofer, M., Spohn, H.: Current fluctuations for the totally asymmetric simple exclusion process. In: Sidoravicius, V. (ed.) In and Out of Equilibrium. Progress in Probability. Birkhäuser, Boston/Basel (2002)
15. Prähofer, M., Spohn, H.: Exact scaling function for one-dimensional stationary KPZ growth. J. Stat. Phys. **115**, 255–279 (2004)
16. Spohn, H.: Excess noise for a lattice gas model of a resistor. Z. Phys. B **57**, 255–261 (1984)
17. Tracy, C.A., Widom, H.: Level-spacing distributions and the Airy kernel. Commun. Math. Phys. **159**, 151–174 (1994)
18. van Beijeren, H., Kutner, R., Spohn, H.: Excess noise for driven diffusive systems. Phys. Rev. Lett. **54**, 2026–2029 (1985)
19. Widom, H.: On convergence of moments for random young tableaux and a random growth model. Int. Math. Res. Not. **9**, 455–464 (2002)

Part II
Multiple Scales in Mathematical Models of Materials Science and Biology

Vortex Motion for the Landau-Lifshitz-Gilbert Equation with Applied Magnetic Field

Matthias Kurzke, Christof Melcher, and Roger Moser

1 Introduction

In micromagnetics, the fundamental evolution law for the magnetization $\mathbf{m}$ in a solid is given by the Landau-Lifshitz-Gilbert equation

$$\frac{\partial \mathbf{m}}{\partial t} = \mathbf{m} \times \left(\alpha \frac{\partial \mathbf{m}}{\partial t} - \gamma \, \boldsymbol{h}_{\text{eff}} \right), \tag{1}$$

which is used to describe the dynamics of a great variety of magnetic microstructures, in particularly the motion of domain walls and vortices in thin films, see e.g. [3]. Here $\boldsymbol{h}_{\text{eff}}$ is the *effective field*, essentially the L^2 gradient of the micromagnetic energy.

A collective coordinate ansatz $\mathbf{m} = \mathbf{m}(x - a(t))$, where $\mathbf{m}$ is the profile of the static problem and $a = a(t)$ describes its translation at time t, has been proposed by Thiele in [24] in order to drastically reduce the complexity of (1). Thiele's approach has been adapted by Huber [8] to the situation of a vortex system, giving rise to a system of ODEs typically called Thiele's equation of motion. More precisely, the

M. Kurzke (✉)
Institut für Angewandte Mathematik, Rheinische Friedrich-Wilhelms-Universität Bonn,
Endenicher Allee 60, D-53115 Bonn, Germany
e-mail: kurzke@iam.uni-bonn.de

C. Melcher
Department of Mathematics I & JARA – Fundamentals of Future Information Technology,
RWTH Aachen University, Wüllnerstraße 5b, D-52056 Aachen, Germany
e-mail: melcher@rwth-aachen.de

R. Moser
Department of Mathematical Sciences, University of Bath, Claverton Down, Bath,
BA2 7AY, UK
e-mail: r.moser@bath.ac.uk

M. Griebel (ed.), *Singular Phenomena and Scaling in Mathematical Models*,
DOI 10.1007/978-3-319-00786-1_6,

resulting system for vortices with trajectories $t \mapsto a_j(t) \in \mathbb{R}^2 \times \{0\}$ $(j = 1, \ldots d)$ takes the form

$$F_j(a) + G_j \times \dot{a}_j + D\dot{a}_j = 0.$$

Here $F_j = F_j(a_1, \ldots, a_d)$ are interaction forces, $G_j = 4\pi q_j \hat{\mathbf{e}}_3$ is the gyro-vector of the jth vortex, which depends only on the topological index $q_j = \pm\frac{1}{2}$ of the vortex (which is half of the product of winding number and polarity), and D is an effective damping constant. In previous joint work with Spirn [14, 15] we have rigorously derived a Thiele equation from (1) in the limit of small vortex size, for an exchange-dominated model energy. In [13] we have generalized the result to an extended version of (1), modeling the influence of an in-plane spin-polarized current $v = v(t)$. More precisely, we have shown that the corresponding spin-torque terms give rise to an additive extension of Thiele's equation

$$F_j(a) + G_j \times (\dot{a}_j - v) + D(\dot{a}_j - \kappa v) = 0$$

where κ is a non-negative constant. The aim of the present work is to derive a Thiele equation from (1) under the influence of a (possibly time-dependent) applied field $\boldsymbol{h} \in \mathbb{R}^3$. Unlike the result for an external current, the effect of the magnetic field is visible only in the interaction force term. The precise result will be given in Theorem 3.

As our model energy we use

$$E_\epsilon(\boldsymbol{h}, \mathbf{m}) = \int_\Omega \left(\frac{1}{2}|\nabla \mathbf{m}|^2 + \frac{m_3^2}{\epsilon^2} - \boldsymbol{h} \cdot \mathbf{m} \right) dx, \tag{2}$$

where $\Omega \subset \mathbb{R}^2$ is a bounded and simply connected domain, with a Dirichlet boundary condition $\mathbf{m} = \boldsymbol{g}$. The most physical choice of $\boldsymbol{g}$ is to use a unit tangent to $\partial\Omega$. We refer to [14] for a justification of this model.

2 Jacobian, Vorticity and Renormalized Energy

Suppose that we have a map $\mathbf{m} : \Omega \to \mathbb{S}^2$ in the Sobolev space H^1. It is convenient to consider the decomposition

$$\mathbf{m} = (m, m_3).$$

Recall that the Jacobian of $m : \Omega \to \mathbb{R}^2$ is defined as

$$J(m) = \det \nabla m.$$

Note that the Jacobian, considered as a differential 2-form, is exact. More precisely, $J(m) = \frac{1}{2}\,\mathrm{curl}\, j(m)$, where $j(m) = m \wedge \nabla m$ is the current, and we write $a \wedge b = a_1 b_2 - a_2 b_1$ for $a, b \in \mathbb{R}^2$. Observe that current and Jacobian are well-defined as distributions for maps $m \in L^\infty \cap W^{1,1}(\Omega; \mathbb{R}^2)$. Moreover, they carry topological information about the $\mathbb{S}^1$-degree of the map m. More precisely, if B is a ball, $m \in C^1(\overline{B}; \mathbb{R}^2)$ is such that $m|_{\partial B} \neq 0$ and $u = m/|m|$, then

$$\int_{\partial B} j(u) \cdot ds = 2\pi \deg(u, \partial B).$$

For $\mathbb{S}^2$-valued maps $\mathbf{m}$, the counterpart of the Jacobian is the vorticity

$$\omega(\mathbf{m}) = \left\langle \mathbf{m}, \frac{\partial \mathbf{m}}{\partial x_1} \times \frac{\partial \mathbf{m}}{\partial x_2} \right\rangle,$$

which is, considered as a differential 2-form, the pull-back of the standard volume form on $\mathbb{S}^2$ with respect to $\mathbf{m}$. Thus, if B is a ball, $\mathbf{m} \in C^1(\overline{B}; \mathbb{S}^2)$ is such that $\mathbf{m}|_{\partial B}$ is an equator map, then

$$\int_B \omega(\mathbf{m})\, dx = 4\pi q,$$

where q is the $\mathbb{S}^2$-degree of, i.e. the oriented number of covers of $\mathbb{S}^2$ by the map $\mathbf{m}$. Thus q is a half-integer if the winding number of $\deg(m, \partial B)$ is odd. In contrast to the Jacobian, however, $\omega(\mathbf{m})$ is not exact, i.e., $\omega(\mathbf{m})$ is not a null-Lagrangian.

2.1 Compactness

We have good compactness results for the Jacobian and, under assumptions on the energy excess, also on the maps themselves. The compactness properties of the vorticity are not as good as those for the Jacobians, and we will not discuss them here in general.

Proposition 1. *Assume that* $(\mathbf{m}_\epsilon)$ *is a sequence of maps* $m_\epsilon \in H^1(\Omega; \mathbb{S}^2)$ *with* $\mathbf{m}_\epsilon = (g, 0)$ *on* $\partial\Omega$ *and* $E_\epsilon(\boldsymbol{h}, \mathbf{m}_\epsilon) \leq C \log \frac{1}{\epsilon}$ *for some fixed* $\mathbf{h} \in \mathbb{R}^3$.

Then we can extract a subsequence (not relabeled) such that

$$J(m_\epsilon) \to \pi \sum_j^d \delta_{a_j} \tag{3}$$

in the dual of $C_0^{0,1}(\Omega)$.

Proof. As $\mathbf{h}$ is independent of ϵ, it follows that $E_\epsilon(0, \mathbf{m}_\epsilon) \leq C \log \frac{1}{\epsilon}$. Hence the 2D Ginzburg-Landau energy of m_ϵ satisfies the same bound, and we can now apply standard compactness results [10]. □

Proposition 2. *Suppose that the sequence* $(\mathbf{m}_\epsilon)$ *satisfies the assumptions of Proposition 1 and suppose that* d *and* $a_1, \ldots, a_d$ *are as in* (3). *If additionally the sequence* $E_\epsilon(\boldsymbol{h}, \mathbf{m}_\epsilon) \leq d\pi \log \frac{1}{\epsilon} + C$ *then* $\mathbf{m}_\epsilon$ *is bounded in* $W^{1,p}(\Omega; \mathbb{S}^2)$ *for* $1 \leq p < 2$ *and in* $H^1_{\mathrm{loc}}(\Omega \backslash \{a_1, \ldots, a_d\})$. *In particular, a subsequence converges strongly in* $L^q(\Omega; \mathbb{S}^2)$ *for every* $q < \infty$ *to a map* $\mathbf{m}_0 = (m_0, 0)$ *with* $|m_0| = 1$.

Proof. From the convergence of the Jacobians for a subsequence ϵ_n and lower bounds near the singularities [9, 19], we obtain for every $r > 0$

$$\limsup_{n \to \infty} \int_{\Omega_r(a)} |\nabla m_{\epsilon_n}|^2 dx \leq 2\pi d \log \frac{1}{r} + C,$$

which shows the H^1_{loc} bound. Using an argument of Struwe [23] and appropriate diagonal subsequences, one can show by Hölder's inequality and summing a series that

$$\limsup_{\epsilon \searrow 0} \int_{\Omega} |\nabla m_\epsilon|^p dx \leq C(p)$$

for all $p \in [1, 2)$. Alternatively, one can obtain the $W^{1,p}$ boundedness from the global bounds on ∇m_ϵ in the Lorentz space $L^{2,\infty}$ given in [21]. Rellich-Kondrachov embedding finally yields strong convergence. □

2.2 The Renormalized Energy

We introduce some notation. We fix a boundary condition $g \in C^\infty(\partial\Omega; \mathbb{S}^1)$ with $\deg(g) = d > 0$. For $a \in \Omega^d$, we set

$$\rho_a := \min(\min_i \operatorname{dist}(a_i, \partial\Omega), \frac{1}{2} \min_{i \neq j} |a_i - a_j|).$$

For $r \in [0, \rho_a)$ we define

$$\Omega_r(a) = \{x \in \Omega : |x - a_j| > r \text{ for } j = 1, \ldots, d\}$$

and we write $\Omega^d_* = \{a \in \Omega_d : \rho_a > 0\}$. As in [2], for $a \in \Omega^d_*$, there exists a corresponding canonical harmonic map $M_* = M_*(\cdot, a)$ with vortex locations a and all local winding numbers equal to 1, i.e.

$$M_*(x;a) = \prod_{j=1}^{d} \frac{x-a_j}{|x-a_j|} e^{i\psi},$$

where ψ is a harmonic function chosen such that $M_*(x;a) = g$ on $\partial\Omega$. Recall that $M_*(\cdot, a) \in W_g^{1,p}(\Omega, \mathbb{S}^1)$ for all $p \in [1,2)$. We also verify by virtue of the explicit representation of $M_*(\cdot, a)$ that the mapping

$$\Omega_*^d \ni a \mapsto M_*(\cdot, a) \in L^p(\Omega; \mathbb{C}) \tag{4}$$

is continuously differentiable for $p \in [1,2)$. For $h \in \mathbb{R}^2$ sufficiently small, we consider

$$W(h,a) = W_0(a) + V(h,a)$$

where $W_0 = W_0(a)$ is the unperturbed renormalized energy as introduced by Bethuel, Brezis and Hélein [2]. The perturbation $V = V(h,a)$ is defined as the following energy minimum

$$V(h,a) = \min_{\theta \in H_0^1(\Omega)} \mathscr{G}(h,a;\theta),$$

where

$$\mathscr{G}(h,a;\theta) = \int_\Omega \frac{1}{2}|\nabla\theta|^2 - h \cdot \left(e^{i\theta} M_*(x;a)\right) dx. \tag{5}$$

Observe that, for h sufficiently small, $\mathscr{G}(h,a;\cdot)$ is a strictly convex functional on $H_0^1(\Omega)$, and hence there exists a unique minimizer $\theta = \theta(h,a) \in H_0^1(\Omega)$. Since $\mathscr{G}$ is a smooth function of h and $\theta = \theta(h,a)$ a critical point, it follows that

$$\frac{\partial W}{\partial h} = \frac{\partial V}{\partial h} = \frac{\partial \mathscr{G}}{\partial h}\Big|_{\theta=\theta(h,a)} = -\int_\Omega m_*(x;a)\, dx, \tag{6}$$

where

$$m_*(x;a) = e^{i\theta(x;h,a)} M_*(x;a).$$

Note that $m_*(\cdot, a) \in W_g^{1,p}(\Omega, \mathbb{S}^1)$ for all $p \in [1,2)$ with

$$J(m_*) = J(M_*) = \pi \sum_{j=1}^{d} \delta_{a_j}$$

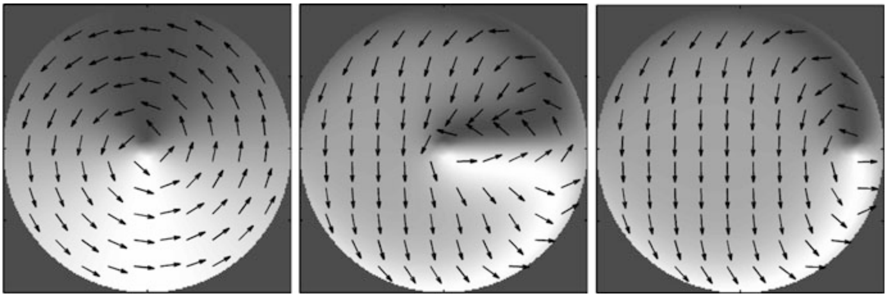


Fig. 1 Numerical plot of $M_*(\cdot;0)$ (*left*), $m_*(\cdot;0)$ (*center*) and $m_*(\cdot;a_{\min})$ (*right*) for $a_{\min}$ minimizing $W(h,a)$. The applied field is $h=(0,-40)$. Note that in the situation presented here, the external field exerts a force on the vortex that is perpendicular to the field. Numerical simulation by Jutta Steiner (using Matlab) based on Newton iteration for minimization of $\mathscr{G}(h,a;\theta)$ for fixed a and h

and that the Euler-Lagrange equation for (5) expressed in terms of m_* reads

$$\nabla\cdot j(m_*) = h\wedge m_*, \tag{7}$$

i.e., $m_* = m_*(\cdot,a)$ is the canonical h-harmonic map corresponding to g and $a\in\Omega_*^d$. We have the following characterization of the renormalized energy:

Lemma 1. *The renormalized energy can be calculated as*

$$W(h,a) = \lim_{r\to 0}\left(\int_{\Omega_r(a)}\frac{1}{2}|\nabla m_*|^2 - \mathbf{h}\cdot m_*\,dx - \pi d\log\frac{1}{r}\right). \tag{8}$$

Proof. As in [2], we can set $\Phi = 2\pi\sum_{j=1}^d \log|x-a_j|$. Then Φ is locally the conjugate harmonic map of the phase of $\prod_{j=1}^d \frac{x-a_j}{|x-a_j|}$. Using that $|m_*| = |M_*| = 1$, we can now write $|\nabla M_*| = |\nabla^\perp\Phi + \nabla\psi|$ and $|\nabla m_*| = |\nabla^\perp\Phi + \nabla\psi + \nabla\theta|$, where ψ is the harmonic function and $\theta = \theta(\cdot;h,a)$ as above. It follows that

$$|\nabla m_*|^2 - |\nabla M_*|^2 = |\nabla\theta|^2 + 2(\nabla^\perp\Phi + \nabla\psi)\cdot\nabla\theta.$$

Integrating this expression over $\Omega_r(a)$ and using that ψ is harmonic, we obtain for $r\to 0$ the claimed result. □

We deduce from (4) a local Lipschitz condition for m_* as a mapping in a, which will be useful for identifying effective motion laws.

Lemma 2. *Suppose* $p\in[1,2)$, $a^0\in\Omega_*^d$. *Then there exists* $c>0$ *such that*

$$\|m_*(\cdot,a) - m_*(\cdot,\hat{a})\|_{L^p} \le c\,|a-\hat{a}|$$

for all $a,\hat{a}\in\Omega_*^d$ *such that* $\max\{|a-a_0|,|\hat{a}-a_0|\} < \rho(a_0)/2$.

Lemma 3. *Suppose $\Phi \in C_0^\infty(\Omega;\mathbb{R}^2)$ and $\rho \in (0,\rho_a)$ such that $\Phi|B_\rho(a_\ell) =$ const. and $\Phi|B_\rho(a_k) = 0$ for all $k \neq \ell$. Then, with $m_* = m_*(\cdot, a)$, we have*

$$\Phi(a_\ell)\cdot\frac{\partial W}{\partial a_\ell}(h,a) = \int_\Omega \nabla\Phi : \left(\left(\frac{1}{2}|\nabla m_*|^2 - h\cdot m_*\right)\mathbf{1} - \nabla m_*\otimes\nabla m_*\right)dx.$$

Proof. The claim of the lemma is in fact a singular version of Noether's formula for the Lagrangian $\frac{1}{2}|\nabla m_*|^2 - h\cdot m_*$ with respect to inner variations $s \mapsto m_*(x - s\,\Phi(x))$. Based on this observation, the argument in [12] for the case $h = 0$ carries over literally. □

We will need the following notion of energy excess for a map $\mathbf{m}$ and a configuration of points $a \in \Omega_*^d$:

$$D_\epsilon^{\mathbf{h}}(\mathbf{m};a) := E_\epsilon(\mathbf{h},\mathbf{m}) - \left(\pi d\log\frac{1}{\epsilon} + d\gamma + W(h,a)\right),$$

where γ is defined as $\lim_{\epsilon\searrow 0}(I_\epsilon - \pi\log\frac{1}{\epsilon})$, and

$$I_\epsilon = \inf\left\{\int_{B_1(0)} e_\epsilon(\mathbf{m})\,dx : \mathbf{m}(x) = (x,0) \text{ on } \partial B_1(0)\right\}.$$

To show that the name "energy excess" is justified, and to relate the micromagnetic energy to the renormalized energy, we have

Proposition 3. *If $J(m_\epsilon) \to \pi\sum_{k=1}^d \delta_{a_k}$ then $\liminf_{\epsilon\searrow 0} D_\epsilon^{\mathbf{h}}(\mathbf{m}_\epsilon;a) \geq 0$.*

Proof. Let $\epsilon_k \to 0$ be a sequence such that

$$A = \liminf_{\epsilon\searrow 0} D_\epsilon^{\mathbf{h}}(\mathbf{m}_\epsilon;a) = \lim_{k\to\infty} D_{\epsilon_k}^{\mathbf{h}}(\mathbf{m}_{\epsilon_k};a).$$

We can assume that $A < \infty$ (otherwise there is nothing to prove). By Proposition 2, we have (for a subsequence) that $\mathbf{m}_{\epsilon_k} \to \mathbf{m}_0 = (m_0, 0)$ weakly in $H^1_{\mathrm{loc}}(\Omega_0(a);\mathbb{R}^3)$ and strongly in all $L^p(\Omega)$, $1 \leq p < \infty$. It follows that $|m_0| = 1$, i.e. $\mathbf{m}_0$ has values in $S^1\times\{0\}$.

Now $D_{\epsilon_k}^{\mathbf{h}}(\mathbf{m}_{\epsilon_k};a) = D_{\epsilon_k}^0(\mathbf{m}_{\epsilon_k};a) - \int_\Omega \mathbf{h}\cdot(\mathbf{m}_{\epsilon_k} - m_*)\,dx$. As in the proof of Theorem 5.3 of [14], for r sufficiently small we have

$$\begin{aligned} D_{\epsilon_k}^{\mathbf{h}}(\mathbf{m}_{\epsilon_k};a) \geq &\int_{\Omega_r(a)} (e_{\epsilon_k}(\mathbf{m}_{\epsilon_k}) - \frac{1}{2}|\nabla M_*|^2)\,dx \\ &+ \sum_{\ell=1}^d\left(\int_{B_r(a_\ell)} e_{\epsilon_k}(\mathbf{m}_{\epsilon_k})\,dx - I_{\epsilon_k/r}\right) - Cr^2 \\ &- \int_\Omega \mathbf{h}\cdot(\mathbf{m}_{\epsilon_k} - m_*). \end{aligned}$$

Using the convergence of $\mathbf{m}_{\epsilon_k}$ and Lemma 5.1 of [14] we obtain

$$\liminf_{k\to\infty} D^{\mathbf{h}}_{\epsilon_k}(\mathbf{m}_{\epsilon_k};a) \geq \int_{\Omega_r(a)} \left(\frac{1}{2}|\nabla m_0|^2 - \frac{1}{2}|\nabla m_*|^2\right) dx \\ - \int_{\Omega} \mathbf{h}\cdot(m_0 - m_*) - Cr^2$$

We decompose $m_0 = e^{i\beta} M_*$. As in the derivation of (8), it is not difficult to see that

$$\int_{\Omega_r(a)} \left(\frac{1}{2}|\nabla m_0|^2 - \frac{1}{2}|\nabla M_*|^2\right) dx \to \int_{\Omega} \frac{1}{2}|\nabla\beta|^2\, dx$$

as $r \to 0$, and now we can use the minimality of θ to conclude the proof of the proposition. □

Now we show that the phase excess in $\Omega_r(a)$ (which measures the distance of $\mathbf{m}_\epsilon$ from an optimal map) can be bounded by the energy excess, up to errors that are small as $\epsilon \to 0$ and $r \to 0$. Unlike the quantitative theory of [11], our proof follows the idea of Lemma 3.7 in [20] and uses weak convergence.

We define

$$\tilde{e}_\epsilon(\mathbf{m}) = \frac{1}{2}\left(\big|\nabla|m|\big|^2 + |\nabla m_3|^2 + \frac{m_3^2}{\epsilon^2}\right)$$

and note the decomposition

$$e_\epsilon(\mathbf{m}) = \tilde{e}_\epsilon(\mathbf{m}) + \frac{1}{2}\left|\frac{j(m)}{|m|}\right|^2.$$

Proposition 4. *Assume $D_\epsilon = D^{\mathbf{h}}_\epsilon(\mathbf{m}_\epsilon;a) \leq C$. Then we have the following estimates for any $\rho < \rho_a$, $\ell = 1,\ldots,d$:*

$$\left|\int_{B_\rho(a_\ell)} |\nabla \mathbf{m}_\epsilon|^2\, dx - \pi|\log\epsilon|\right| \leq C, \tag{9}$$

$$\int_{\Omega_\rho(a)} \tilde{e}_\epsilon(\mathbf{m}_\epsilon)\, dx \leq D_\epsilon + o_\epsilon(1), \tag{10}$$

$$\int_{\Omega_\rho(a)} \frac{1}{2}\left|\frac{j(m_\epsilon)}{|m_\epsilon|} - j(m_*(\cdot;a))\right|^2 dx \leq \frac{1}{1-C|\mathbf{h}|} D_\epsilon + o_\epsilon(1). \tag{11}$$

Proof. As in the proof of Proposition 3, we have for a subsequence that $\mathbf{m}_\epsilon$ converges to $\mathbf{m_0} = (m_0, 0)$ weakly in $H^1_{\mathrm{loc}}(\Omega_0(a))$ and strongly in $L^p(\Omega)$, with

$m_0 = e^{i\beta} M_*$, where $\beta \in H_0^1(\Omega)$. The proof of Proposition 3 also gives for any small $r > 0$

$$\liminf_{\epsilon \searrow 0} \left(\int_{\bigcup B_r(a_\ell)} e_\epsilon(\mathbf{m}_\epsilon)\, dx - \pi d \log \frac{r}{\epsilon} + d\gamma \right) \geq -Cr^2. \tag{12}$$

Furthermore,

$$\liminf_{\epsilon \searrow 0} \int_{\Omega_r(a)} \frac{1}{2} \left| \frac{j(m_\epsilon)}{|m_\epsilon|} \right|^2 dx \geq \int_{\Omega_r(a)} \frac{1}{2} |\nabla m_0|^2\, dx$$

and since $\mathbf{m}_\epsilon \to \mathbf{m}_0$ in $L^1(\Omega)$, we obtain

$$\liminf_{\epsilon \searrow 0} \left(\int_{\Omega_r(a)} \frac{1}{2} \frac{|j(m_\epsilon)|^2}{|m_\epsilon|^2}\, dx - \int_\Omega \mathbf{h} \cdot \mathbf{m}_\epsilon\, dx \right) \geq \int_{\Omega_r(a)} \frac{1}{2} |\nabla m_0|^2\, dx - \int_\Omega \mathbf{h} \cdot \mathbf{m}_0\, dx.$$

From (8) we obtain

$$\int_{\Omega_r(a)} \frac{1}{2} |\nabla m_*|^2\, dx - \int_\Omega \mathbf{h} \cdot m_*\, dx \geq W(h, a) + \pi d \log \frac{1}{r} - o_r(1)$$

so adding this to (12) we obtain

$$\liminf_{\epsilon \searrow 0} \left(D_\epsilon - \int_{\Omega_r(a)} \tilde{e}_\epsilon(\mathbf{m}_\epsilon)\, dx \right) \geq -o_r(1).$$

Since the right-hand side of the previous inequality tends to zero as $r \to 0$, we obtain by monotonicity of the left-hand side for any $\rho > 0$

$$\liminf_{\epsilon \searrow 0} \left(D_\epsilon - \int_{\Omega_\rho(a)} \tilde{e}_\epsilon(\mathbf{m}_\epsilon)\, dx \right) \geq 0.$$

This is (10). From $D_\epsilon \leq C$, we obtain that also (9) must hold.

From the definition of energy excess it follows that

$$\limsup_{\epsilon \searrow 0} \left(\int_{\Omega_r(a)} \frac{1}{2} |\nabla \mathbf{m}_\epsilon|^2 - \frac{1}{2} |\nabla m_*|^2\, dx - \int_\Omega \mathbf{h} \cdot (\mathbf{m}_\epsilon - m_*)\, dx - D_\epsilon \right) \leq -o_r(1)$$

so a fortiori

$$\limsup_{\epsilon \searrow 0} \left(\int_{\Omega_r(a)} \frac{1}{2} \frac{|j(m_\epsilon)|^2}{|m_\epsilon|^2} - \frac{1}{2} |j(m_*)|^2\, dx - \int_\Omega \mathbf{h} \cdot (\mathbf{m}_\epsilon - m_*)\, dx - D_\epsilon \right) \leq -o_r(1).$$

We calculate

$$\frac{1}{2}\left|\frac{j(m_\epsilon)}{|m_\epsilon|}-j(m_*)\right|^2=\frac{1}{2}\frac{|j(m_\epsilon)|^2}{|m_\epsilon|^2}-\frac{1}{2}|j(m_*)|^2-j(m_*)\cdot\left(\frac{j(m_\epsilon)}{|m_\epsilon|}-j(m_*)\right).$$

Using that $j(m_*)=\nabla^\perp\Phi+\nabla\psi+\nabla\theta$ and $j(m_0)=\nabla^\perp\Phi+\nabla\psi+\nabla\beta$, we have that

$$\lim_{\epsilon\searrow 0}\int\limits_{\Omega_r(a)} j(m_*)\cdot\left(\frac{j(m_\epsilon)}{|m_\epsilon|}-j(m_*)\right)dx=\int\limits_{\Omega_r(a)}\left(\nabla^\perp\Phi+\nabla\psi+\nabla\theta\right)\cdot(\nabla\beta-\nabla\theta)\,dx$$

For $r\to 0$, this expression converges using the harmonicity of ψ to

$$\int_\Omega\nabla\theta\cdot(\nabla\beta-\nabla\theta)\,dx.$$

We obtain

$$\limsup_{\epsilon\searrow 0}\left(\int_{\Omega_r(a)}\frac{1}{2}\left|\frac{j(m_\epsilon)}{|m_\epsilon|}-j(m_*)\right|^2-D_\epsilon\right)$$
$$\leq -o_r(1)+\int_\Omega\mathbf{h}\cdot M_*(e^{i\beta}-e^{i\theta})+\nabla\theta\cdot(\nabla\beta-\nabla\theta)\,dx.$$

The Euler-Lagrange for θ in weak form reads as

$$\int_\Omega\nabla\theta\cdot(\nabla\beta-\nabla\theta)\,dx=\int_\Omega\mathbf{h}\cdot(iM_*e^{i\theta})(\beta-\theta)\,dx.$$

We study the expression

$$\mathbf{h}\cdot\left(M_*e^{i\theta}(e^{i(\beta-\theta)}-1-i(\beta-\theta))\right)$$

and note that it can be written using an application of Taylor's theorem to the function $f(t)=\mathbf{h}\cdot(M_*e^{i(\theta+t(\beta-\theta))})$. In fact, we have

$$f(1)=f(0)+f'(0)+\int_0^1 f''(t)(1-t)\,dt,$$

where $f''(t)=\mathbf{h}\cdot M_*e^{i(\theta+t(\beta-\theta))}(\beta-\theta)^2$. Taking absolute values and integrating, it follows that

$$\left|\int_\Omega\mathbf{h}\cdot\left(M_*e^{i\theta}(e^{i(\beta-\theta)}-1-i(\beta-\theta))\right)dx\right|\leq|\mathbf{h}|\int_\Omega(\beta-\theta)^2\,dx.$$

By weak convergence,

$$\int_{\Omega_r(a)} \frac{1}{2}|\nabla(\beta - \theta)|^2 \, dx \le \liminf_{\epsilon \searrow 0} \int_{\Omega_r(a)} \frac{1}{2} \left| \frac{j(m_\epsilon)}{|m_\epsilon|} - j(m_*) \right|^2 dx.$$

Using Poincaré's inequality, we obtain that

$$\limsup_{\epsilon \searrow 0} \left((1 - C\mathbf{h}) \int_{\Omega_r(a)} \frac{1}{2} \left| \frac{j(m_\epsilon)}{|m_\epsilon|} - j(m_*) \right|^2 dx - D_\epsilon \right) \le -o_r(1),$$

and letting $r \to 0$ on the right as before we obtain (11). □

2.3 The Thiele Equation

For $h \in W^{1,1}(0,T;\mathbb{R}^2)$, which is small enough so that $W = W(h(t),\cdot)$ corresponds to a unique minimizer $\theta = \theta(h(t),\cdot)$ for all $t \in [0,T]$, we consider the equation

$$(4\pi q_\ell i + \alpha_0 \pi)\dot{a}_\ell(t) + \frac{\partial W}{\partial a_\ell}(h(t),a(t)) = 0 \quad (\ell = 1,\ldots,d). \tag{13}$$

Lemma 4. *For initial data $a(0) = a_0 \in \Omega_*^d$ the Cauchy problem for (13) has a unique solution $a \in C^1([0,T];\Omega^d)$, which satisfies the energy identity*

$$W(h(t_1),a(t_1)) - W(h(t_2),a(t_2)) = \alpha_0 \, \pi \int_{t_1}^{t_2} |\dot{a}(s)|^2 ds + \int_{t_1}^{t_2} \int_\Omega \dot{h}(s) \cdot m_* \, dx \, ds$$

for all $0 \le t_1 < t_2 \le T$, where $m_*(x;a) = e^{i\theta(x)} M_*(x;a)$.

Proof. Using (6) and (13) we compute for the (unique) local solution $a = a(t)$

$$\begin{aligned}\frac{d}{dt} W(h(t),a(t)) &= \frac{\partial W}{\partial h}(h(t),a(t))\dot{h}(t) + \sum_{j=1}^d \frac{\partial W}{\partial a_j}(h(t),a(t)) \cdot \dot{a}_j(t) \\ &= -\alpha_0 \, \pi \, |\dot{a}(t)|^2 - \int_\Omega \dot{h}(t) \cdot m_*(x;a(t)) \, dx.\end{aligned}$$

The energy identity follows, and the local solution $a = a(t)$ extends to $[0,T]$. □

3 LLG Equation with External Fields

Let us now consider the Landau-Lifshitz-Gilbert equation

$$\frac{\partial \mathbf{m}}{\partial t} = \mathbf{m} \times \left(\alpha_\epsilon \frac{\partial \mathbf{m}}{\partial t} - \boldsymbol{h}_{\text{eff}} \right), \tag{14}$$

where, for an external field $\boldsymbol{h} \in W^{1,1}(0, T; \mathbb{R}^3)$, the effective field is given by

$$\boldsymbol{h}_{\text{eff}} = \Delta \mathbf{m} - \frac{m_3}{\epsilon^2} \hat{\boldsymbol{e}}_3 + \boldsymbol{h}.$$

We consider a specific asymptotic behavior for α_ϵ such that $\alpha_\epsilon \log \frac{1}{\epsilon} \to \alpha_0 \in (0, \infty)$ as $\epsilon \to 0$. The effective field corresponds to minus the L^2 gradient of

$$E_\epsilon(\boldsymbol{h}, \mathbf{m}) = \int_\Omega e_\epsilon(\mathbf{m}) - \boldsymbol{h} \cdot \mathbf{m} \, dx.$$

where, as usual,

$$e_\epsilon(\mathbf{m}) = \frac{1}{2} |\nabla \mathbf{m}|^2 + \frac{m_3^2}{2\epsilon^2}$$

is the energy density of the Ginzburg-Landau type energy $E_\epsilon(\mathbf{m}) = E_\epsilon(0, \mathbf{m})$, which we have considered in [13–15]. In this section we study the equation for a fixed $\epsilon \in (0, 1)$. We impose Dirichlet boundary data given by a smooth map $\mathbf{g} = (g, 0)$ where $g : \partial\Omega \to \mathbb{S}^1$ with $\deg(g) = d$ and initial data $\mathbf{m}^0 \in H^1_{\mathbf{g}}(\Omega; \mathbb{S}^2)$ with

$$E_\epsilon(\mathbf{m}^0) \leq d\pi \log \frac{1}{\epsilon} + C_0. \tag{15}$$

3.1 Conservation Laws

Let us assume $\mathbf{m}$ is a smooth solution of (1) in a space-time cylinder. The vorticity $\omega(\mathbf{m})$ makes contact to the LLG equation through the identity

$$\frac{\partial}{\partial t} \omega(\mathbf{m}) = \operatorname{curl} \left\langle \mathbf{m} \times \frac{\partial \mathbf{m}}{\partial t}, \nabla \mathbf{m} \right\rangle$$

leading to

$$\frac{\partial}{\partial t} \omega(\mathbf{m}) + \alpha_\epsilon \operatorname{curl} \left\langle \frac{\partial \mathbf{m}}{\partial t}, \nabla \mathbf{m} \right\rangle = \operatorname{curl} \operatorname{div} (\nabla \mathbf{m} \otimes \nabla \mathbf{m}). \tag{16}$$

This conservation law for the vorticity will be crucial when identifying motion laws for vortices, which are the concentration points of $\omega(\mathbf{m})$ in the singular limit $\epsilon \searrow 0$.

Moreover, the energy identity for (14) reads

$$\frac{\partial}{\partial t}\Big(e_\epsilon(\mathbf{m}) - \boldsymbol{h}(t)\cdot\mathbf{m}\Big) + \alpha_\epsilon \left|\frac{\partial \mathbf{m}}{\partial t}\right|^2 = \operatorname{div}\left\langle \frac{\partial \mathbf{m}}{\partial t}, \nabla \mathbf{m}\right\rangle + \left\langle \dot{\boldsymbol{h}}(t), \frac{\partial \mathbf{m}}{\partial t}\right\rangle \tag{17}$$

Finally, we have conservation of spin

$$\frac{\partial m_3}{\partial t} + \operatorname{div} j(m) = \alpha_\epsilon\, m \wedge \frac{\partial m}{\partial t} + h \wedge m, \tag{18}$$

which is just the third component of (14), will imply that in the singular limit $\epsilon \searrow 0$, $\mathbf{m}$ will converge to an $h(t)$-harmonic map.

3.2 Weak Solutions and Bubbling

The LLG equation (14), for $\epsilon > 0$ fixed, is a lower order perturbation of the conformally invariant LLG equation $\mathbf{m}_t = \mathbf{m} \times (\alpha\, \mathbf{m}_t - \Delta \mathbf{m})$ which is traditionally studied in mathematical analysis. In dimension two, this equation is critical with respect to the natural energy estimate, and the formation of singularities in finite time must be expected, [1]. On the other hand, a well-known construction of what is called energy decreasing weak solutions, which has been introduced by Struwe [22] for the harmonic map heat flow, see also [4] and [6, 7] for LLG, can be carried out. In this framework, the possible blow-up scenario is precisely characterized through the formation of bubbles at the energy concentration points.

This is in fact the new fundamental difficulty compared with the corresponding problem for the complex Ginzburg-Landau theory, where at the finite ϵ level, evolution equations admit smooth solutions for all times, [12]. Since vortex trajectories are retraced in terms of concentration sets of the energy density $e_\epsilon(\mathbf{m})$ and the vorticity $\omega(\mathbf{m})$, precise information about their behavior near the singular points is a crucial ingredient to our analysis. This information can be obtained from the well-developed bubbling analysis for harmonic maps and flows, established e.g. in [5, 16–18, 25]. Applied to (14) we obtain the following result (cf. [14, Sect. 4] for more information):

Theorem 1. *For initial data* $\mathbf{m}^0 \in C^\infty(\overline{\Omega}; \mathbb{S}^2)$ *there exists a weak solution* $\mathbf{m}$ *of* (14) *which satisfies the energy inequality*

$$\alpha_\epsilon \int_0^{t_0}\int_\Omega \left|\frac{\partial \mathbf{m}}{\partial t}\right|^2 dx\,dt + E_\epsilon(\boldsymbol{h}(t_0), \mathbf{m}^0) \le E_\epsilon(\boldsymbol{h}(0), \mathbf{m}^0) - \int_0^{t_0}\int_\Omega \dot{\boldsymbol{h}} \cdot \mathbf{m}\, dx\,dt$$

for all $0 \le t_0 \le T$ *and is smooth away from a finite number of points* (x_i, t_i) *in space time. Moreover, there exists, for every* i*, an integer* q_i *such that for every sufficiently small* $r > 0$

$$\int_{B_r(x^i)\times\{t_i\}} e_\epsilon(\mathbf{m})\,dx + 4\pi|q_i| \le \liminf_{t\nearrow t^i} \int_{B_r(x^i)\times\{t_i\}} e_\epsilon(\mathbf{m})\,dx$$

and

$$\int_{B_r(x^i)\times\{t_i\}} \omega(\mathbf{m})\,dx + 4\pi q_i = \lim_{t\nearrow t^i} \int_{B_r(x^i)\times\{t_i\}} \omega(\mathbf{m})\,dx.$$

Finally, the (energy decreasing) solution $\mathbf{m}$ *is unique in its class.*

Form the energy inequality we deduce that for $E_\epsilon(\mathbf{m}^0) \le d\pi\log(1/\epsilon) + C_0$,

$$\alpha_\epsilon \int_0^{t_0}\int_\Omega \left|\frac{\partial \mathbf{m}}{\partial t}\right|^2 dx\,dt + E_\epsilon(\mathbf{m}(t_0)) \le d\pi\log(1/\epsilon) + C_1, \tag{19}$$

where $0 \le C_1 - C_0$ can be bounded above by a multiple of $\displaystyle\int_0^T |\dot{\boldsymbol{h}}(t)|dt + |\boldsymbol{h}(0)|$.

4 Convergence and Vortex Trajectories

Now we consider a sequence of initial data $\mathbf{m}_\epsilon^0 \in H_g^1(\Omega;\mathbb{S}^2)$ such that

$$\alpha_\epsilon e_\epsilon(\mathbf{m}_\epsilon^0) \to \alpha_0\pi\delta_{a^0}, \quad \omega_0(\mathbf{m}_\epsilon^0) \to 4\pi\sum_{\ell=1}^d q_\ell\delta_{a_\ell^0} \quad \text{and} \quad \lim_{\epsilon\searrow 0} D_\epsilon(\mathbf{m}_\epsilon^0; a^0) = 0$$

for a certain $a_0 \in \Omega_*^d$ and $q_1, \dots, q_d = \pm\frac{1}{2}$ and the corresponding weak solution $\mathbf{m}_\epsilon$ from Theorem 1. As in [14, Theorem 4.1] (see [13] for more details) and in view of Proposition 2 we obtain the following convergence result.

Theorem 2. *There exist a time* $T_0 \in (0, T]$*, a sequence* $\epsilon_k \searrow 0$*, and a curve*

$$a \in H^1(0, T_0; \Omega^d) \quad \textit{with} \quad a(0) = a^0 \quad \textit{and} \quad \inf_{t\in(0,T_0)} \rho(a(t)) > 0$$

such that for every $t \in [0, T_0]$ *and* $1 \le p < 2$

$$m_{\epsilon_k}(\cdot, t) \rightharpoonup m_*(\cdot, a(t)) \quad \textit{weakly in} \quad W^{1,p}(\Omega;\mathbb{R}^2),$$

$$\alpha_{\epsilon_k} e_{\epsilon_k}(\mathbf{m}_{\epsilon_k}(\,\cdot\,, t)) \overset{*}{\rightharpoonup} \alpha_0\pi\delta_{a(t)} \quad \textit{weakly* in} \quad (C_0^0(\Omega))^*,$$

$$J(m_{\epsilon_k}(\,\cdot\,, t)) \to \pi\delta_{a(t)}, \quad \omega(\mathbf{m}_{\epsilon_k}(\,\cdot\,, t)) \to 4\pi\sum_{\ell=1}^d q_\ell\delta_{a_\ell(t)} \quad \textit{in} \quad (C_0^{0,1}(\Omega))^*.$$

Moreover, for all $t_1, t_2 \in [0, T_0]$ *with* $t_1 \leq t_2$ *and* $\eta \in C^1(\overline{\Omega})$

$$\alpha_0 \pi \sum_{\ell=1}^{d} \eta(a_\ell(t))\Big|_{t=t_1}^{t_2} = \lim_{k\to\infty} \left(\alpha_{\epsilon_k} \int_{t_1}^{t_2} \int_{\Omega} \nabla\eta \cdot \left\langle \frac{\partial \mathbf{m}_{\epsilon_k}}{\partial t}, \nabla \mathbf{m}_{\epsilon_k} \right\rangle dx\, dt \right)$$

and

$$\alpha_0 \pi \int_{t_1}^{t_2} |\dot{a}|^2\, dt \leq \liminf_{k\to\infty} \left(\alpha_{\epsilon_k} \int_{t_1}^{t_2} \int_{\Omega} \left| \frac{\partial \mathbf{m}_{\epsilon_k}}{\partial t} \right|^2 dx\, dt \right).$$

From the energy inequality in Theorem 1, the convergence of m_{ϵ_k} in Theorem 2 and conservation of spin identity (18) we deduce in particular that

$$j(m_{\epsilon_k}(t,\cdot)) \rightharpoonup j(m_*(t,\cdot)) \quad \text{weakly in } L^p(\Omega;\mathbb{R}^2) \tag{20}$$

for every $t \in [0, T_0)$, where

$$\operatorname{div} j(m_*(t,\cdot) = h(t) \wedge m_* \quad \text{and} \quad \operatorname{curl} j(m_*(t,\cdot)) = 2\pi\delta_a(t).$$

5 Motion Law

Theorem 3. *There exist positive numbers* h_0 *and* ϵ_0 *with the following property: For every* $\epsilon \in (0, \epsilon_0)$ *and every smooth* $\boldsymbol{h} : [0, T] \to \mathbb{R}^3$ *with*

$$\int_0^T |\dot{\boldsymbol{h}}(t)| dt + |\boldsymbol{h}(0)| < h_0,$$

there exists a smooth solution $\mathbf{m}_\epsilon \in C^\infty(\overline{\Omega} \times [0, T]; \mathbb{S}^2)$ *of the Landau-Lifshitz-Gilbert equation* (14) *with* $\mathbf{m}_\epsilon(\cdot, 0) = \mathbf{m}_\epsilon^0$ *and* $\mathbf{m}_\epsilon(\cdot, t)|_{\partial\Omega} = \mathbf{g}$ *for every* $t \geq 0$. *Moreover, for every* $t \in [0, T]$,

$$\alpha_\epsilon e_\epsilon(\mathbf{m}_\epsilon(\cdot, t)) \to \pi\alpha_0 \sum_{\ell=1}^{d} \delta_{a_\ell(t)} \quad \textit{and} \quad \omega(\mathbf{m}_\epsilon(\cdot, t)) \to 4\pi \sum_{\ell=1}^{d} q_\ell \delta_{a_\ell(t)}$$

as $\epsilon \searrow 0$, *in the sense of distributions, where* $a \in C^\infty([0, T]; \Omega^d)$ *is the solution of Thiele's equation*

$$G_\ell \times \dot{a}_\ell + D\, \dot{a}_\ell + \frac{\partial W(h, a)}{\partial a_\ell} = 0 \quad (\ell = 1, \ldots, d) \tag{21}$$

with $a(0) = a^0$ *and where* $G_\ell = 4\pi q_\ell \hat{\boldsymbol{e}}_3$ *and* $D = \pi\,\alpha_0$ *with* $\alpha_0 = \lim_{\epsilon\searrow 0} \alpha_\epsilon \log \frac{1}{\epsilon}$.

The rest of this section is devoted to the proof of the Theorem. Let $\hat{a} \in C^\infty([0,\infty); \Omega^d)$ be the unique solution of the initial value problem for (21) with initial values $\hat{a}(0) = a^0 \in \Omega^d$. We choose $T_0 > 0$ and a sequence $\epsilon_k \searrow 0$ that satisfy the conclusions of Theorem 2, and let a be the corresponding curve in Ω^d. We recall that solutions remain smooth in $(0, T_0)$ for small ϵ as shown in [14, Theorem 3], so we can concentrate on the verification of the motion law.

We fix a radius $r \in (0, \rho(a^0)/2]$ and adapt the terminal time T_0 such that the trajectories of a_ℓ and $\hat{a}_\ell$ do not exit $B_{r/2}(a_\ell^0)$ before time T_0 for all $\ell = 1, \ldots, d$. As in [14] we choose $\phi, \psi \in C_0^\infty(\Omega)$ such that for every ℓ, both ϕ and ψ are affine with $\nabla\psi = \nabla^\perp\phi$ in $B_r(a_\ell^0)$. We define

$$\xi_k(t) = \int_{\Omega\times\{t\}} \left(\alpha_{\epsilon_k} \psi\, e_{\epsilon_k}(\mathbf{m}_{\epsilon_k}) + \phi\, \omega_0(\mathbf{m}_{\epsilon_k})\right) dx - \pi \sum_{\ell=1}^{d} (\alpha_0 \psi(\hat{a}_\ell(t)) + 4q_\ell \phi(\hat{a}_\ell(t))),$$

converging, for every $t \in [0, T)$, to

$$\xi(t) = \pi \sum_{\ell=1}^{d} \Big(\alpha_0 \Big(\psi(a_\ell(t)) - \psi(\hat{a}_\ell(t))\Big) + 4q_\ell \Big(\phi(a_\ell(t)) - \phi(\hat{a}_\ell(t))\Big)\Big).$$

In order to apply Proposition 4 we fix h_0 sufficiently small.

Lemma 5. *There exists a constant C such that for all $t_1, t_2 \in [0, T_0]$ with $t_1 \le t_2$ and every $k \in \mathbb{N}$,*

$$\xi_k(t_2) - \xi_k(t_1) \le C \int_{t_1}^{t_2} \left(D_{\epsilon_k}^{h(t)}(\mathbf{m}_{\epsilon_k}; \hat{a}(t)) + |a(t) - \hat{a}(t)| \right) dt + o_{\epsilon_k}(1).$$

Proof. From (13) we obtain

$$\pi \sum_{\ell=1}^{d} \frac{d}{dt} (\alpha_0 \psi(\hat{a}_\ell(t)) + 4q_\ell \phi(\hat{a}_\ell(t))) = -\frac{\partial W(h, \hat{a})}{\partial a_\ell} \cdot \nabla\psi(\hat{a}_\ell(t))$$

while from Lemma 3 with $\hat{m}_* := m_*(\cdot\,; \hat{a})$ and $\Phi = \nabla^\perp\phi$

$$-\sum_{\ell=1}^{d} \nabla\psi(\hat{a}_\ell(t)) \cdot \frac{\partial W(h, \hat{a})}{\partial a_\ell} = \int_{\Omega\times\{t\}} \nabla^\perp\nabla\phi : (\nabla\hat{m}_* \otimes \nabla\hat{m}_*)\, dx.$$

Using conservation of vorticity (16), we find after integration by parts in space and integration in time

$$\int_\Omega \phi\, \omega(\mathbf{m}_{\epsilon_k}(t_2)) - \phi\, \omega(\mathbf{m}_{\epsilon_k}(t_1))\, dx = \alpha_{\epsilon_k} \int_{t_1}^{t_2} \int_\Omega \nabla^\perp\phi \cdot \left\langle \frac{\partial \mathbf{m}_{\epsilon_k}}{\partial t}, \nabla\mathbf{m}_{\epsilon_k} \right\rangle dx\, dt$$
$$+ \int_{t_1}^{t_2} \int_\Omega \nabla^\perp\nabla\phi : \left(\nabla\mathbf{m}_{\epsilon_k} \otimes \nabla\mathbf{m}_{\epsilon_k}\right) dx\, dt.$$

For the terms on the left we use convergence of the vorticity provided by Theorem 2. Concerning the first term on the right we deduce from the energy estimate in Theorem 1

$$\left(\alpha_{\epsilon_k}\int_{t_1}^{t_2}\int_\Omega (\nabla^\perp\phi-\nabla\psi)\cdot\left\langle \nabla\mathbf{m}_{\epsilon_k},\frac{\partial\mathbf{m}_{\epsilon_k}}{\partial t}\right\rangle dx\,dt\right)^2 \\ \le c\int_{t_1}^{t_2}\int_\Omega |\nabla^\perp\phi-\nabla\psi|^2\,\alpha_{\epsilon_k} e_{\epsilon_k}(\mathbf{m}_{\epsilon_k})\,dx\,dt\to 0$$

while by convergence of the kinetic term in Theorem 2

$$\alpha_{\epsilon_k}\int_{t_1}^{t_2}\int_\Omega \nabla\psi\cdot\left\langle \nabla\mathbf{m}_{\epsilon_k},\frac{\partial\mathbf{m}_{\epsilon_k}}{\partial t}\right\rangle dx\,dt\to -\pi\alpha_0\sum_{\ell=1}^{d}\Big(\psi(a_\ell(t_2))-\psi(a_\ell(t_1))\Big)$$

as $\epsilon_k \searrow 0$. Therefore, it suffices to estimate the integrals

$$\int_{t_1}^{t_2}\int_\Omega \nabla^\perp\nabla\phi : (\nabla\mathbf{m}_{\epsilon_k}\otimes\nabla\mathbf{m}_{\epsilon_k}-\nabla\hat{m}_*\otimes\nabla\hat{m}_*)\,dx\,dt,$$

which, by virtue of the usual decomposition argument and Proposition 4 (see [14, Sect. 6]), reduces to the estimation of

$$\int_{t_1}^{t_2}\int_\Omega \nabla^\perp\nabla\phi : ((j(m_{\epsilon_k})-j(\hat{m}_*))\otimes j(\hat{m}_*))\,dx\,dt$$

and

$$\int_{t_1}^{t_2}\int_\Omega \nabla^\perp\nabla\phi : (j(\hat{m}_*)\otimes (j(m_{\epsilon_k})-j(\hat{m}_*)))\,dx\,dt.$$

Taking into account that both integrands are products of the form

$$\sigma\cdot(j(m_{\epsilon_k})-j(\hat{m}_*))$$

for smooth vector fields $\sigma\in C^\infty(\overline{\Omega}\times[0,T_0];\mathbb{R}^2)$ independent of k, we obtain from (20) with $m_*=m_*(\cdot,a(t))$ and $\hat{m}_*=m_*(\cdot,\hat{a}(t))$

$$\int_{t_1}^{t_2}\int_\Omega \sigma\cdot(j(m_{\epsilon_k})-j(\hat{m}_*))\,dx\,dt=\int_{t_1}^{t_2}\int_\Omega \sigma\cdot(j(m_*)-j(\hat{m}_*))\,dx\,dt+o_{\epsilon_k}(1).$$

Next we adopt the Hodge decomposition argument used in [14, Lemma 7]. Writing $-\sigma=\nabla u+\nabla^\perp v$, where $u,v\in C^\infty(\overline{\Omega}\times[0,T_0])$ with $u=0$ on $\partial\Omega\times[0,T_0]$ we infer, also taking into account Lemma 2,

$$\int_{t_1}^{t_2}\int_{\Omega}\sigma\cdot(j(m_*)-j(\hat{m}_*))\,dx\,dt=$$
$$\int_{t_1}^{t_2}\int_{\Omega}h\wedge(m_*-\hat{m}_*)\,u\,dx\,dt+2\pi\sum_{\ell=1}^{d}\int_{t_1}^{t_2}\left(v(a_\ell)-v(\hat{a}_\ell)\right)\,dt$$
$$\leq c\int_{t_1}^{t_2}|a(t)-\hat{a}(t)|\,dt.\qquad\square$$

Proof (Theorem 3). The proof follows by the usual Gronwall argument. For $t\in[0,T_0]$, we consider the functions

$$\zeta_k(t)=D_{\epsilon_k}^{\boldsymbol{h}(t)}(\mathbf{m}_{\epsilon_k}(t);\hat{a}(t))\quad\text{and}\quad\chi(t)=|\hat{a}_\ell(t)-a_\ell(t)|.$$

First we show $\zeta_k\to\zeta$ in $L^1(0,T_0)$ for a function $\zeta\in BV(0,T_0)$ with

$$\dot{\zeta}\leq c\big(|\dot{\hat{a}}-\dot{a}|+\chi\big).\tag{22}$$

In fact, we obtain from Lemma 4

$$W(h(t_1),\hat{a}(t_1))-W(h(t_2),\hat{a}(t_2))=\pi\alpha_0\int_{t_1}^{t_2}|\dot{\hat{a}}|^2\,dt-\int_{t_1}^{t_2}\int_{\Omega}\dot{h}(t)\cdot m_*(\cdot,\hat{a}(t))\,dx\,dt$$

and from (19)

$$E_{\epsilon_k}(\boldsymbol{h}(t_2),\mathbf{m}_{\epsilon_k}(t_2))-E_{\epsilon_k}(\boldsymbol{h}(t_1),\mathbf{m}_{\epsilon_k}(t_1))=$$
$$-\int_{t_1}^{t_2}\int_{\Omega}\left(\alpha_{\epsilon_k}\left|\frac{\partial\mathbf{m}_{\epsilon_k}}{\partial t}\right|^2-\dot{\boldsymbol{h}}(t)\cdot\mathbf{m}_{\epsilon_k}\right)\,dx\,dt,$$

respectively, for $0\leq t_1\leq t_2\leq T_0$, while

$$\left|\int_{t_1}^{t_2}\int_{\Omega}\left(\dot{\boldsymbol{h}}(t)\cdot m_{\epsilon_k}-\dot{h}(t)\cdot m_*(\cdot,\hat{a}(t))\right)\,dx\,dt\right|\leq c\int_{t_1}^{t_2}\chi(t)\,dt+o_{\epsilon_k}(1).$$

In view of Theorem 2 we can select a subsequence such that $\zeta_k(t)\to\zeta(t)$ almost everywhere for a bounded function $\zeta:[0,T_0]\to\mathbb{R}$ with

$$\zeta(t_2)-\zeta(t_1)\leq\int_{t_1}^{t_2}\pi\alpha_0\left(|\dot{\hat{a}}|^2-|\dot{a}|^2\right)+c\,\chi(t)\,dt\leq c\int_{t_1}^{t_2}|\dot{\hat{a}}-\dot{a}|+\chi(t)\,dt$$

for almost all $t_1\leq t_2$, which implies (22). Now Lemma 5 implies, by virtue of (22), for $0\leq t_1\leq t_2\leq T_0$,

$$\xi(t_2) - \xi(t_1) \leq c \int_{t_1}^{t_2} (\zeta(t) + \chi(t))\, dt.$$

With an appropriate choice of ϕ and ψ we obtain the desired inequality

$$|\dot{\hat{a}}(t) - \dot{a}(t)| \leq c \int_0^t |\dot{\hat{a}}(\tau) - \dot{a}(\tau)|\, d\tau.$$

As $\hat{a}(0) = a(0)$, Gronwall's lemma implies $\hat{a} = a$ in $[0, T_0]$. Moreover,

$$\limsup_{k\to\infty} D_{\epsilon_k}^{h(T_0)}(\mathbf{m}_{\epsilon_k}(T_0); a(T_0)) \leq 0,$$

which enables us to iterate the argument for new initial times T_0, and we eventually obtain the motion law for all times before T_0. Note that by uniqueness of energy decreasing solutions, solutions $\mathbf{m}_\epsilon$ extend, for small ϵ, smoothly to $[0, T]$. Finally, thanks to the unique solvability of the limiting ODE, the convergence result for energy density and vorticity can be seen to hold without taking subsequences, as any subsequence of $\epsilon \searrow 0$ will have a further subsequence converging to the same limit. □

References

1. Bartels, S., Ko, J., Prohl, A.: Numerical analysis of an explicit approximation scheme for the Landau-Lifshitz-Gilbert equation. Math. Comput. **77**(262), 773–788 (2008)
2. Bethuel, F., Brezis, H., Hélein, F.: Ginzburg-Landau Vortices. Progress in Nonlinear Differential Equations and their Applications, vol. 13. Birkhäuser Boston Inc., Boston (1994)
3. Capella, A., Melcher, C., Otto, F.: Wave-type dynamics in ferromagnetic thin films and the motion of Néel walls. Nonlinearity **20**(11), 2519–2537 (2007)
4. Chang, K.-C.: Heat flow and boundary value problem for harmonic maps. Ann. Inst. H. Poincaré Anal. Non Linéaire **6**(5), 363–395 (1989)
5. Ding, W., Tian, G.: Energy identity for a class of approximate harmonic maps from surfaces. Commun. Anal. Geom. **3**(3–4), 543–554 (1995)
6. Guo, B.L., Hong, M.C.: The Landau-Lifshitz equation of the ferromagnetic spin chain and harmonic maps. Calc. Var. Partial Differ. Equ. **1**(3), 311–334 (1993)
7. Harpes, P.: Uniqueness and bubbling of the 2-dimensional Landau-Lifshitz flow. Calc. Var. Partial Differ. Equ. **20**(2), 213–229 (2004)
8. Huber, D.L.: Dynamics of spin vortices in two-dimensional planar magnets. Phys. Rev. B **26**(7), 3758–3765 (1982)
9. Jerrard, R.L.: Lower bounds for generalized Ginzburg-Landau functionals. SIAM J. Math. Anal. **30**(4), 721–746 (1999) (electronic)
10. Jerrard, R.L., Soner, H.M.: The Jacobian and the Ginzburg-Landau energy. Calc. Var. Partial Differ. Equ. **14**(2), 151–191 (2002)
11. Jerrard, R.L., Spirn, D.: Refined Jacobian estimates and Gross-Pitaevsky vortex dynamics. Arch. Ration. Mech. Anal. **190**(3), 425–475 (2008)
12. Kurzke, M., Melcher, C., Moser, R., Spirn, D.: Dynamics for Ginzburg-Landau vortices under a mixed flow. Indiana Univ. Math. J. **58**(6), 2597–2621 (2009)

13. Kurzke, M., Melcher, C., Moser, R.: Vortex motion for the Landau–Lifshitz–Gilbert equation with spin-transfer torque. SIAM J. Math. Anal. **43**(3), 1099–1121 (2011)
14. Kurzke, M., Melcher, C., Moser, R., Spirn, D.: Ginzburg-Landau vortices driven by the Landau-Lifshitz-Gilbert equation. Arch. Ration. Mech. Anal. **199**(3), 843–888 (2011)
15. Kurzke, M., Melcher, C., Moser, R., Spirn, D.: Vortex dynamics in the presence of excess energy for the Landau-Lifshitz-Gilbert equation. (Calc. Var. PDE SFB 611 Preprint 526) (2012, to appear)
16. Lin, F., Wang, C.: Energy identity of harmonic map flows from surfaces at finite singular time. Calc. Var. Partial Differ. Equ. **6**(4), 369–380 (1998)
17. Qing, J.: On singularities of the heat flow for harmonic maps from surfaces into spheres. Commun. Anal. Geom. **3**, 297–315 (1995)
18. Qing, J., Tian, G.: Bubbling of the heat flows for harmonic maps from surfaces. Commun. Pure Appl. Math. **50**(4), 295–310 (1997)
19. Sandier, E.: Lower bounds for the energy of unit vector fields and applications. J. Funct. Anal. **152**(2), 379–403 (1998)
20. Sandier, E., Serfaty, S.: Gamma-convergence of gradient flows with applications to Ginzburg-Landau. Commun. Pure Appl. Math. **57**(12), 1627–1672 (2004)
21. Serfaty, S., Tice, I.: Lorentz space estimates for the Ginzburg-Landau energy. J. Funct. Anal. **254**(3), 773–825 (2008)
22. Struwe, M.: On the evolution of harmonic mappings of Riemannian surfaces. Comment. Math. Helv. **60**(4), 558–581 (1985)
23. Struwe, M.: On the asymptotic behavior of minimizers of the Ginzburg-Landau model in 2 dimensions. Differ. Integral Equ. **7**(5–6), 1613–1624 (1994)
24. Thiele, A.A.: Steady-state motion of magnetic domains. Phys. Rev. Lett. **30**(6), 230–233 (1973)
25. Wang, C.: Bubble phenomena of certain Palais-Smale sequences from surfaces to general targets. Houston J. Math. **22**(3), 559–590 (1996)

On Prandtl-Reuss Mixtures

Jens Frehse and Josef Málek

Abstract We study mathematical properties of the model that has been proposed to explain the phenomenon of hardening due to cyclic loading. The model considers two elastic plastic materials, soft and hard, that co-exist while the soft material can be transformed into the hard material. Regarding elastic responses we remain in a simplified framework of linearized elasticity. Incorporating tools such as variational inequalities, penalty approximations and Sobolev spaces, we prove the existence of weak solution to the corresponding boundary-value problem and investigate its uniqueness and regularity.

1 Introduction

In the article [10], Kratochvíl, Rajagopal, Srinivasa and the second author of the present contribution developed a thermodynamically consistent model within the framework of finite elastic plasticity that is capable of 'explaining' the phenomenon of hardening of the material due to cyclic loadings. They consider the mixture of two elastic plastic materials, soft and hard, that coexist. The material that can be thought to be originally almost consisting of soft region builds the hard regions by a process of 'recruitment' of the soft material and its conversion into a hard material. The study in [10] carries on some ideas from Kratochvíl [9]. The authors then also consider a simplified model that is obtained by assuming that the gradient of the

J. Frehse (✉)
Institut für Angewandte Mathematik, Rheinische Friedrich-Wilhelms-Universität Bonn,
Wegelerstr. 6, D-53115 Bonn, Germany
e-mail: erdbeere@iam.uni-bonn.de

J. Málek
Faculty of Mathematics and Physics, Mathematical Institute, Charles University in Prague,
Prague, Czech Republic
e-mail: malek@karlin.mff.cuni.cz

M. Griebel (ed.), *Singular Phenomena and Scaling in Mathematical Models*,
DOI 10.1007/978-3-319-00786-1_7,

displacement is small. This results in a model that can be viewed as the mixture of two (soft and hard) Prandtl-Reuss models of the linearized elastic perfect plasticity where the conversion of soft regions to hard regions is modeled through the variation of the volume fraction α of the soft material within the mixture. Note that $(1-\alpha)$ is the volume fraction of the hard material. The present paper intends to elaborate a rigorous mathematical treatment for this simplified model using the framework of variational inequalities, penalty approximations and Sobolev spaces. We call the materials that are based on the coexistence of the two (or more) Prandtl-Reuss elastic plastic materials 'Prandtl-Reuss mixtures'.

Although a mathematical treatment is very similar to the one of the classical Prandtl-Reuss-problems, several complications arise. A first study of this model has been performed in the thesis of Khasina, see [8].

The paper is organized in the following way. Section 2 starts with a basic mathematical setting, and contains the formulation of the mixture problem via a variational inequality for the convex combination of the soft and hard material, with the side condition that the convex combination $\alpha\sigma_s + (1-\alpha)\sigma_h$ satisfies the balance of linear momentum, and σ_s and σ_h satisfy the relevant yield conditions. The main theorem states that, under appropriate conditions on the data, in particular, under a safe load condition for the mixture, the considered problem for the Prandtl-Reuss mixture has unique solution σ_s, σ_h in the spaces $L^\infty(L^2)$ and $H^{1,2}(L^2)$.

Furthermore, from the formulation via a variational inequality one concludes, see Sect. 3, the existence of partial velocity gradients (or more precisely their symmetric parts) $\frac{1}{2}(\nabla\dot{u}_s + \nabla\dot{u}_s^T)$ and $\frac{1}{2}(\nabla\dot{u}_h + \nabla\dot{u}_h^T)$ so that

$$\frac{1}{2}(\nabla\dot{u}_s + \nabla\dot{u}_s^T) = A_s\frac{\partial}{\partial t}(\alpha\sigma_s) + \dot{e}_{ps},\ \frac{1}{2}(\nabla\dot{u}_h + \nabla\dot{u}_h^T) = A_h\frac{\partial}{\partial t}(\alpha\sigma_h) + \dot{e}_{ph}, \tag{1}$$

where $\dot{e}_{ps}$ and $\dot{e}_{ph}$ are the rates of plastic strains for the soft and hard material, A_s and A_h are the inverse fourth order elastic tensors, so that $A_s\sigma_s$ and $A_h\sigma_h$ correspond to the elastic strains for the soft and hard materials. Then we conclude that

$$\frac{1}{2}(\nabla\dot{u}_s + \nabla\dot{u}_s{}^T) = \frac{1}{2}(\nabla\dot{u}_h + \nabla\dot{u}_h{}^T) =: \frac{1}{2}(\nabla\dot{u} + \nabla\dot{u}^T). \tag{2}$$

Here and below, for any quantity w

$$\dot{w} = \frac{\partial w}{\partial t}, \tag{3}$$

we shall use both notation in what follows. We confine ourselves to the von Mises yield conditions

$$|\sigma_{sD}| \le \kappa_s, |\sigma_{hD}| \le \kappa_h, \tag{4}$$

where κ_s and κ_h may depend on t and x and $B_D = B - (\operatorname{tr} B/3)\, I$ for any $B \in \mathbb{R}^{n\times n}$. The plastic strains are nontrivial only if $|\sigma_{sD}| = \kappa_s$ and $|\sigma_{hD}| = \kappa_h$, and then they are proportional to the outer normal 'vectors' associated with the surfaces $|\sigma_{sD}| = \kappa_s$ and $|\sigma_{hD}| = \kappa_h$, it means that

$$\dot{e}_{ps} = \lambda_s \frac{\sigma_{sD}}{|\sigma_{sD}|}, \dot{e}_{ph} = \lambda_h \frac{\sigma_{hD}}{|\sigma_{hD}|} \text{ with } \lambda_s \text{ and } \lambda_h > 0. \tag{5}$$

These conditions can be rewritten in the compact Kuhn-Tucker forms

$$\lambda_s, \lambda_h \geq 0, \lambda_s(|\sigma_{sD}| - \kappa_s) = 0, \lambda_h(|\sigma_{hD}| - \kappa_h) = 0. \tag{6}$$

Unfortunately, in the rigorous mathematical treatment, similar as for the analysis of the classical Prandtl-Reuss model, the quantities $\frac{1}{2}(\nabla\dot{u}_s + \nabla\dot{u}_s^T)$, $\frac{1}{2}(\nabla\dot{u}_h + \nabla\dot{u}_h^T)$ are only elements of C^*, i.e. they are not functions. This holds also for $\dot{e}_{ps}$ and $\dot{e}_{ph}$, so the above Kuhn-Tucker rule has to be interpreted correctly, see Sect. 8. Fortunately, due to Temam's imbedding theorem, the quantity $\dot{u}$ is an element of $L^\infty(L^{\frac{n}{n-1}})$, i.e. it is a 'function'.

The proof of the main theorem starts in Sect. 3 by introducing a penalty approximation where the yield conditions (4) are penalized. The penalty approximation, in turn, is discretized by a Rothe approximation, and in Sect. 4 up to 6 we establish uniform estimates for Rothe approximations and take the limit in the Rothe approximation in order to obtain the solvability of the penalty approximation. In Sect. 7 we establish uniform $L^\infty(L^2)$-estimates for the stress velocities $\dot{\sigma}_{s\mu}, \dot{\sigma}_{h\mu}$, where $\mu \to 0$ is the penalty parameter.

Finally, in Sect. 8 we pass to the limit with respect to the penalty parameter and complete the proof of the main theorem. As mentioned above we also discuss how the Kuhn-Tucker forms have to be formulated rigorously. In Sect. 9, we consider a generalized model (derived in [10]) in which α, κ_s and κ_h may depend on history of the rate of the plastic strain of the soft material. We discuss how to treat this in the framework of the present paper. In a continuation of this study we intend to focus on the regularity properties of the solution.

We refer to [3] for a detailed survey of the results concerning the mathematical analysis of relevant results concerning initial and boundary value problems for classical Prandtl-Reuss model of the linearized elastic perfect plasticity.

2 Mathematical Formulation of the Problem

2.1 Basic Setting

Let Ω be a bounded domain of $\mathbb{R}^n$ occupied by a body which is supposed to be a mixture of a soft and a hard linearized elastic-perfect-plastic materials in the sense defined below.

We imagine the following deformation process with (slow) cyclic loading in which the mixture with a large portion of soft material is gradually deformed and transforms into a mixture with a large portion of hard material.

If $t \in [0, T]$ is the loading parameter, the interior stresses of the soft or hard material are denoted by $\sigma_s(t, x)$ and $\sigma_h(t, x)$, respectively. Let M^n_{sym} be a set of symmetric $n \times n$ matrices. We require the σ's to be symmetric, i.e.

$$\sigma_s, \sigma_h : [0,T] \times \Omega \to M^n_{sym}. \tag{7}$$

Let $\alpha : [0,T] \times \Omega \to [0,1]$ describe the fraction of the soft material in the mixture such that the stress of the mixture σ is given by

$$\sigma(t,x) = \alpha(t,x)\sigma_s(t,x) + (1-\alpha(t,x))\sigma_h(t,x). \tag{8}$$

Assumption 1. *(a) α is Lipschitz continuous and decreasing.*
(b) $0 < \alpha_0 < \alpha(t,x) < 1 - \alpha_0 < 1$ with a constant α_0 (for all $t \in [0,T]$ and $x \in \Omega$).

Remark 1. In general, α depends on the history of σ_s, but readability is better if we start the theory with the above assumption, see also Sect. 9.

Remark 2. We are not able to treat the case $\alpha_0 = 0$, i.e. an analysis starting with a 'pure' soft material is not possible, up to now. This corresponds also to the results of numerical experiments performed and presented in [8].

The notion 'hard' or 'soft' material is given by the yield condition. We confine the presentation to the von Mises-yield condition:

Condition 1. *Let κ_s, κ_h :$[0,T] \times \Omega \to \mathbb{R}$ be Lipschitz continuous functions such that $0 < \kappa_0 < \kappa_s \le \kappa$, $0 < \kappa_0 < \kappa_h \le \kappa$ with some κ_0, $\kappa \in \mathbb{R}$. We say that σ_s, σ_h satisfy the von Mises-yield condition if $|\sigma_{sD}| \le \kappa_s$, $|\sigma_{hD}| \le \kappa_h$. If $\kappa_s < \kappa_h$ then σ_s is said to be the 'soft' material and σ_h the hard material.*

This means that the modulus of the deviator $\sigma_{sD} = \sigma_s - (\operatorname{tr}\sigma_s/n)I$ and $\sigma_{hD} = \sigma_h - (\operatorname{tr}\sigma_h/n)I$ may not exceed the yield boundary.

Remark 3. The theory presented here works quite similar for other yield functions of the type $F(\sigma_D) \le \kappa$; the function F has to be Lipschitz continuous, convex and coercive. We believe that the readability improves if we confine ourselves to the above case given in Condition 1.

Remark 4. For $n = 2$, in applications, σ_D might be defined in a different way, namely $\sigma_D = \sigma - (\operatorname{tr}\sigma/3)\,\boldsymbol{I}$, see [2].

2.2 Balance of Linear Momentum

The mixture is supposed to satisfy the balance of linear momentum, it means that we have

$$-\,div(\alpha(t,x)\sigma_s(t,x) + (1-\alpha(t,x))\sigma_h(t,x)) = f(t,x), \tag{9}$$

where $f : [0,1] \times \Omega \to \mathbb{R}^n$ is a given volume force (density).

The mixture underlies a mixed boundary condition

$$\begin{aligned} \nu(x)[\alpha(t,x)\sigma_s(t,x) + (1-\alpha(t,x))\sigma_h(t,x)] &= p_0(t,x) \quad \text{on } (0,T) \times \partial\Omega \setminus \Gamma, \\ u &= 0 \quad \text{on } (0,T) \times \Gamma. \end{aligned} \tag{10}$$

Here Γ is a portion of the boundary of Ω, possibly empty,[1] $p_0(t,x) : [0,T]\times(\partial\Omega\setminus\Gamma)\to\mathbb{R}^n$ is a boundary force and $\nu(x)$ is the outer unit vector at $x\in\partial\Omega$, normal to $\partial\Omega$. We extend the definition of p_0 to the whole boundary by setting $p_0=0$ on $(0,T)\times\Gamma$.

The precise version of the weak formulation to (9) and (10) reads as:

$$(\alpha(t,\cdot)\sigma_s(t,\cdot)+(1-\alpha(t,\cdot))\sigma_h(t,\cdot),\nabla\phi)=\int_{\partial\Omega}p_0(t,.)\phi\,\mathrm{d}o+(f(t,.),\phi),$$
$$t\in[0,T],\quad\forall\phi\in H^1_\Gamma(\Omega,\mathbb{R}^n).\qquad(11)$$

The brackets (.,.) denote the usual $L^2(\Omega)$ scalar product, for scalar, vector or tensor valued functions as well. H^1_Γ denotes the subspace of the Sobolev space $H^{1,2}$ whose elements vanish on Γ in the sense of traces.

The weak form of the balance equation (11) is well defined if we assume

$$\begin{aligned}f&\in L^\infty((0,T);L^2(\Omega;\mathbb{R}^n))=:L^\infty(L^2),\\ p_0&\in L^\infty((0,T);L^2(\partial\Omega;\mathbb{R}^n)).\end{aligned}\qquad(12)$$

Furthermore, $\partial\Omega$ and Γ are $(n-1)$-dimensional Lipschitz-manifolds. As follows from above, we use the shortened notation for the Sobolev and Bochener spaces.

2.3 Elasticity

Let

$$\tau=(\tau_s,\tau_h)\,,\quad\hat\tau=(\hat\tau_s,\hat\tau_h):[0,T]\times\Omega\to M^n_{sym}\times M^n_{sym},\qquad(13)$$

and

$$Q\begin{pmatrix}\hat\tau_s\,\tau_s\\ \hat\tau_h\,\tau_h\end{pmatrix}=\int_\Omega[A_s\hat\tau_s:\tau_s+A_h\hat\tau_h:\tau_h]\,\mathrm{d}x.\qquad(14)$$

Here A_s and A_h are inverse elasticity tensors, say of the same structure as in the Lame-Navier linearized elasticity (with possibly different material coefficients). They model the elastic interaction within the soft and hard material. (It is possible to treat additional interaction terms $A_{sh}\hat\tau_h:\tau_s$.)

Assumption 2. *For simplicity we assume that A_s, A_h do not depend on $x\in\Omega$, $t\in[0,T]$, we assume that A_s and A_h are positively definite.*

[1] Here for simplicity (in order to avoid the compatibility condition on the data), we assume that $\Gamma\neq\emptyset$.

Note that the matrix Q represents the total elastic energy of the mixture corresponding to the stresses τ_s and τ_h of a hard and soft materials at the loading 'time' $t \in [0, T]$. It is possible to carry out the theory presented here also for nonlinear convex potential with quadratic growth, but we restrict ourselves to the simplified case described above in order not to be overburdened.

2.4 The Variational Inequality for Elastic-Perfect-Plastic Mixtures

With these preparations, we are able to rigorously formulate the loading process of an elastic-perfect-plastic mixture by a variational inequality. The approach is similar to the standard 'Prandtl-Reuss model', see [5,7,11,12]. For the formulation, we need the following convex set $\mathbb{K}$ of pairs (τ_s, τ_h) of functions $\tau_s, \tau_h \colon [0, T] \times \Omega \to M^n_{sym}$ such that the following holds:

Integrability:

$$\tau_s, \tau_h \in L^\infty(L^2), \quad \dot{\tau}_s, \dot{\tau}_h \in L^\infty(L^2); \tag{15}$$

Initial condition:

$$\tau_s(0) = \sigma_{s0}, \quad \tau_h(0) = \sigma_{h0}; \tag{16}$$

Balance of linear momentum:

$$(\alpha\tau_s + (1-\alpha)\tau_h, \nabla\phi) = (f, \phi) + \int\limits_{\partial\Omega} p_0\phi \, \mathrm{d}o, \ \ \forall\phi \in H^1_\Gamma(\Omega, \mathbb{R}^3), t \in (0, T); \tag{17}$$

Yield conditions:

$$|\tau_{sD}| \le \kappa_s, |\tau_{hD}| \le \kappa_h. \tag{18}$$

Then the variational inequality, i.e. the problem for the Prandtl-Reuss mixture, reads as follows: *Find a pair* $(\sigma_s, \sigma_h) \in \mathbb{K}$ *such that*

$$\int_0^T Q \begin{pmatrix} \frac{\partial}{\partial t}(\alpha\sigma_s) & \alpha(\sigma_s - \tau_s) \\ \frac{\partial}{\partial t}((1-\alpha)\sigma_h) & (1-\alpha)(\sigma_h - \tau_h) \end{pmatrix} \mathrm{d}t \le 0 \quad \text{for all } (\tau_s, \tau_h) \in \mathbb{K}. \tag{19}$$

Remark 5. The function α is Lipschitz in x and t. In the model investigated in [10], α depends also on σ_s. Then we have a quasi variational inequality. For the mathematical treatment of this more complicated case we first have to analyze α's that are σ-independent in order to apply a fixed point theorem for more general case. This is why we restrict ourselves to the simpler case in this study.

For the proof of the main theorem, performed via several levels of approximation and estimates, the following condition seems to be crucial.

Condition 2 (Safe load condition for mixtures). *There exists a pair* $(\tilde{\sigma}_s, \tilde{\sigma}_h) \in \mathbb{K}$ *and a number* $s_0 > 0$ *such that*

$$|\tilde{\sigma}_{sD}| \leq \kappa_s - s_0, \quad |\tilde{\sigma}_{hD}| \leq \kappa_h - s_0. \tag{20}$$

In addition we will deal with differentiability assumptions with respect to the loading parameter t.

Assumption 3. *We assume the following differentiability properties of the data:*

$$\dot{\alpha}, \, \dot{\tilde{\sigma}}_s, \, \dot{\tilde{\sigma}}_h, \, \dot{\kappa}_s, \, \dot{\kappa}_h \in L^\infty(L^\infty) \tag{21}$$

and, for refined regularity estimates,

$$\ddot{\alpha}, \, \ddot{\tilde{\sigma}}_s, \, \ddot{\tilde{\sigma}}_h, \, \ddot{\kappa}_s, \, \ddot{\kappa}_h \in L^\infty(L^\infty). \tag{22}$$

Now we may state our main result.

Theorem 1 (Main theorem). *Let* α *satisfy Assumption 1, let* f *and* p_0 *satisfy* (12) *and let* κ_s, κ_h *satisfy Condition 1. Assume the safe load Condition 2 with the regularity* (21). *Furthermore let* A_s *and* A_h *be positively definite. Then there exists a unique solution of the variational inequality* (19).

Proof. (i) The uniqueness in the case $\alpha = \alpha(x,t)$ is a simple consequence of Assumption 2. Indeed, if σ_s, σ_h and $\hat{\sigma}_s, \hat{\sigma}_h$ are solutions to (19), choose $\tau_s = \hat{\sigma}_s$, $\tau_h = \hat{\sigma}_h$ in the equation for σ_s, σ_h, and use a similar argument with σ_s, σ_h and $\hat{\sigma}_s, \hat{\sigma}_h$ interchanged. Then one concludes that for $w_s = \sigma_s - \hat{\sigma}_s$, $w_h = \sigma_h - \hat{\sigma}_h$

$$\frac{1}{2}\int_0^T \int_\Omega A_s \frac{\partial}{\partial t}(\alpha w_s) : (\alpha w_s) + A_h \frac{\partial}{\partial t}((1-\alpha)w_h) : ((1-\alpha)w_h)\,\mathrm{d}x\,\mathrm{d}t \leq 0, \tag{23}$$

which implies that $w_s = w_h = 0$.

(ii) The existence result is established in the following Sects. 3–6 and 8. In Sect. 3 the approximation of the variational inequality by a penalty method is presented. The variational inequality is approximated by an equation with the balance of linear momentum free functions as test functions and a penalty term where the yield condition is penalized. This is a familiar approach in the framework of classical models, like those of Hencky or Prandtl-Reuss, or hardening models. In Sect. 4 the penalty approximation is discretized via a Rothe method. There we also derive uniform discrete $L^\infty(L^2)$-estimates for the approximate stresses for the soft and hard material, as well as discrete uniform $L^1(L^1)$-estimates for the approximate strain velocities. In Sect. 5 we prove uniform discrete $L^2(L^2)$-estimates for first difference quotients of the

Rothe Approximation. This allows us, in Sect. 6, to pass to the limit in the Rothe Approximation and we obtain, via weak compactness and monotonicity arguments, a solution $\sigma_{s\mu}$, $\sigma_{h\mu}$ of the penalty equation.

Finally, in Sect. 8 we pass to the limit $\mu \to +0$ and obtain a solution of the variational inequality. Again, the proof runs via weak convergence and monotonicity, since the $L^2(L^2)$-estimates for the stress velocities in Sects. 4–6 turned out to be uniform also with respect to μ. The estimates of the stresses and their velocities depend on the $L^\infty(L^\infty)$ norms of $\dot{\alpha}$, $\dot{\kappa}_s, \dot{\kappa}_h, \dot{\bar{\sigma}}_s, \dot{\bar{\sigma}}_h$. For $L^\infty(L^2)$-estimates for $\dot{\sigma}_s, \dot{\sigma}_h$ (rather than $L^2(L^2)$-estimates) we need to assume that $\ddot{\alpha}, \ddot{\kappa}_s, \ddot{\kappa}_h, \ddot{\bar{\sigma}}_s, \ddot{\bar{\sigma}}_h \in L^\infty(L^\infty)$. But this additional derivative in the assumption would be very restrictive to the considered class of nonlinear models, like the one in [10], where α, κ_s, κ_h depend on the history of the stress of the soft material, see the discussion in Sect. 9. For this reason, we arranged the existence theory in the $L^2(L^2)$ setting for $\dot{\sigma}_s$, $\dot{\sigma}_h$. □

Remark 6. The uniqueness needs not hold if α is σ-dependent.

In the classical theory of the Prandtl-Reuss model the inclusion $\dot{\sigma} \in L^\infty(L^2)$ follows in a natural way. A similar theorem is also possible in the present setting.

Theorem 2. *Under the assumptions of the main theorem and (22), the solution couple σ_s, σ_h for the Prandtl-Reuss mixture satisfies*

$$\dot{\sigma}_s, \dot{\sigma}_h \in L^\infty(L^\infty((0,T) \times \Omega; M^n_{sym})) \tag{24}$$

with corresponding $L^\infty(L^\infty)$ bounds depending additionally on the $L^\infty(L^\infty)$ norms of $\ddot{\alpha}, \ddot{\kappa}_s, \ddot{\kappa}_h, \ddot{\bar{\sigma}}_s, \ddot{\bar{\sigma}}_h$.

The proof is done in Sect. 7, where a corresponding bound for the solution of the penalty approximation is established.

The variational inequality (19) is a complete dual formulation of the mechanical problem, i. e. the strains and strain velocities do not appear a priori. However, due to uniqueness and the construction of a solution via the penalty method we conclude the following theorem.

Theorem 3. *Under the assumptions of the main theorem, there exist a Riesz measure $\dot{u} \in C^*((0,T) \times \Omega; \mathbb{R}^n)$ such that $\frac{1}{2}(\nabla \dot{u} + \nabla \dot{u}^T)$ is also a Riesz measure and*

$$\int_0^T \left(A_s \frac{\partial}{\partial t}(\alpha \sigma_s) - \frac{1}{2}(\nabla \dot{u} + \nabla \dot{u}^T), \sigma_s - \tau_s \right) \mathrm{d}t \le 0$$

$$\int_0^T \left(A_s \frac{\partial}{\partial t}((1-\alpha)\sigma_h) - \frac{1}{2}(\nabla \dot{u} + \nabla \dot{u}^T), \sigma_h - \tau_h \right) \mathrm{d}t \le 0 \tag{25}$$

for all τ_s, $\tau_h \in C((0,T) \times \bar{\Omega}; M^n_{sym})$ such that $|\tau_{sD}| \le \kappa_s$, $|\tau_{hD}| \le \kappa_s$. If the assumptions of Theorem 2 are satisfied, the

$$\dot{u} \in L^\infty(L^{\frac{n}{n-1}}). \tag{26}$$

In Sect. 3, from the penalty equation, we derive partial approximate strains and interpret the penalty terms as approximate plastic strain velocities. It turns out that the rates of partial plastic strain for hard and soft materials are equal. In Sects. 4–7, the corresponding $L^1(L^1)$ and $L^\infty(L^1)$-estimates for the strain velocities are proved. This works analogously as in the classical Prandtl-Reuss case via the safe load condition. For the $L^\infty(L^{\frac{n}{n-1}})$ inclusion the tools from Sect. 7 are needed.

There is a lot of further analogy between the problem for the Prandtl-Reuss mixture model considered here and the classical Prandtl-Reuss model. For example, if $n = 2$ one can prove an $L^\infty(L^{2+\delta})$ estimate for the strains, based on the reverse Hölder-inequality and Gehring's lemma. Furthermore the H^1_{loc}-differentiability of the stresses can be done similarly as in the classical Prandtl-Reuss case, see [1, 4]. However, in our case, we need extra differentiability properties and corresponding estimates for the volume fraction α and the yield quantities. This decreases the possibilities to establish the same result if α depends nonlinearly on $\dot{e}_{ps}$.

In the next sections we establish the existence theory needed to complete the proof of Theorem 1.

3 The Penalty Equation

Analogously to the classical Prandtl-Reuss problem we approximate the variational inequality by penalizing the yield conditions. The approximation we use reads:

Find a pair $(\sigma_s, \sigma_h) = (\sigma_{s\mu}, \sigma_{s\mu})$ *such that the properties* (15)–(18) *are satisfied and the following penalty equation holds a.e. with respect to* $t \in [0, T]$

$$Q\begin{pmatrix} \frac{\partial}{\partial t}(\alpha\sigma_s) & \alpha\tau_s \\ \frac{\partial}{\partial t}((1-\alpha)\sigma_h) & (1-\alpha)\tau_h \end{pmatrix}$$
$$+ (\mu^{-1}[|\sigma_{sD}| - \kappa_s]_+ \frac{\sigma_{sD}}{|\sigma_{sD}|}, \alpha\tau_s) + (\mu^{-1}[|\sigma_{hD}| - \kappa_h]_+ \frac{\sigma_{hD}}{|\sigma_{hD}|}, (1-\alpha)\tau_h) = 0 \tag{27}$$

for all $\tau_s, \tau_h : \Omega \to M^n_{sym}$, $\tau_s, \tau_h \in L^2$ *satisfying the balance of linear momentum with force zero*

$$(\alpha\tau_s + (1-\alpha)\tau_h, \nabla\phi) = 0, \qquad \forall\phi \in H^1_\Gamma(\Omega, \mathbb{R}^n). \tag{28}$$

μ *is here the penalty-parameter,* $\mu > 0$.

The penalty equation (27) has a solution:

Theorem 4. *Under the assumptions of Theorem 1, Eq.* (27) *has a unique solution* $\sigma_{s\mu}, \sigma_{h\mu} \in L^\infty(L^2)$ *such that* $\dot{\sigma}_{s\mu}, \dot{\sigma}_{h\mu} \in L^2(L^2)$, *with corresponding uniform bounds as* $\mu \to +0$.

As $\mu \to +0$ *the solutions* $\sigma_{s\mu}, \sigma_{h\mu}$ *converge strongly in* $L^2(L^2)$ *to the solution* σ_s, σ_h *of the variational inequality* (19). *Furthermore we have the uniform*

$L^1(L^1)$-bound for the penalty part

$$\mu^{-1}\int_0^T\int_\Omega [|\sigma_{s\mu D}|-\kappa_s]_+(|\sigma_{s\mu D}|+1)+[|\sigma_{h\mu D}|-\kappa_h]_+(\sigma_{h\mu D}+1)\,\mathrm{d}x\,\mathrm{d}t$$
$$\leq K(\dot{\alpha},\dot{\kappa}_s,\dot{\kappa}_h) \qquad (29)$$

and the bound

$$\operatorname*{ess\,sup}_t \quad \mu^{-1}\int_\Omega [|\sigma_{s\mu D}|-\kappa_s]_+^2+[|\sigma_{h\mu D}|-\kappa_h]_+^2\,\mathrm{d}x \leq K(\dot{\alpha},\dot{\kappa}_s,\dot{\kappa}_h) \qquad (30)$$

uniformly as $\mu \to +0$.

The proof of Theorem 4 is established in Sects. 4–6.

3.1 Reconstruction of Partial Strains

In Eq. (27), we choose

$$\tau_s = \alpha^{-1}\tau_0, \quad \tau_h = 0,$$

or vice versa

$$\tau_s = 0, \quad \tau_h = (1-\alpha)^{-1}\tau_0,$$

where $(\tau_0, \nabla\phi) = 0$ for all $\phi \in H^1_\Gamma$. These pairs (τ_s, τ_h) of test functions are admissible since they satisfy (28). Thus we obtain two equations

$$\left(A_s\frac{\partial}{\partial t}\left(\alpha\sigma_{s\mu}\right), \tau_0\right)+\left(\mu^{-1}[|\sigma_{s\mu D}|-\kappa_s]_+\frac{\sigma_{s\mu D}}{|\sigma_{s\mu D}|}, \tau_0\right)=0 \qquad (31)$$

$$\left(A_h\frac{\partial}{\partial t}\left((1-\alpha)\sigma_{h\mu}\right), \tau_0\right)+\left(\mu^{-1}[|\sigma_{h\mu D}|-\kappa_h]_+\frac{\sigma_{h\mu D}}{|\sigma_{h\mu D}|}, \tau_0\right)=0 \qquad (32)$$

a.e. in $[0,T]$, for all $\tau_0 \in L^2(\Omega, M^n_{sym})$ fulfilling (11). Conversely, from (31), (32) we reach (27).

Now, we use the symmetric Helmholtz decomposition in L^2 to conclude that there exists $v_{s\mu}, v_{h\mu} \in L^2(0,T;H^1_\Gamma(\Omega,\mathbb{R}^n))$ such that $v_{s\mu} = \dot{u}_{s\mu}$, $v_{h\mu} = \dot{u}_{h\mu}$ and

$$A_s\frac{\partial}{\partial t}\left(\alpha\sigma_{s\mu}\right)+\mu^{-1}[|\sigma_{s\mu D}|-\kappa_s]_+\frac{\sigma_{s\mu D}}{|\sigma_{s\mu D}|}=\frac{1}{2}(\nabla\dot{u}_{s\mu}+\nabla\dot{u}_{s\mu}{}^T), \qquad (33)$$

$$A_h \frac{\partial}{\partial t}\big((1-\alpha)\sigma_{h\mu}\big) + \mu^{-1}[|\sigma_{h\mu D}| - \kappa_h]_+ \frac{\sigma_{h\mu D}}{|\sigma_{h\mu D}|} = \frac{1}{2}(\nabla \dot{u}_{h\mu} + \nabla \dot{u}_{h\mu}{}^T). \tag{34}$$

Interestingly, there is a relation between $\dot{u}_{s\mu}$ and $\dot{u}_{h\mu}$ which follows from the penalty equation, namely

$$\big(\dot{u}_{s\mu}, \operatorname{div}(\alpha\tau_s)\big) + \big(\dot{u}_{h\mu}, \operatorname{div}((1-\alpha)\tau_h)\big) = 0, \tag{35}$$

that is valid for all τ_s, $\tau_h \in L^2(\Omega; M^n_{sym})$ such that $\alpha\tau_s + (1-\alpha)\tau_h$ satisfies the balance of linear momentum with zero force.

Theorem 5. *Let $\dot{u}_{s\mu}$ and $\dot{u}_{h\mu}$ be the partial strain velocities arising in* (33) *and* (34)*. Then we have*

$$\frac{1}{2}(\nabla \dot{u}_{s\mu} + \nabla \dot{u}_{s\mu}{}^T) = \frac{1}{2}(\nabla \dot{u}_{h\mu} + \nabla \dot{u}_{h\mu}{}^T). \tag{36}$$

Proof. From (28) and (35) we conclude that

$$\left(\frac{1}{2}(\nabla \dot{u}_{s\mu} + \nabla \dot{u}_{s\mu}{}^T), \alpha\tau_s\right) - \left(\frac{1}{2}(\nabla \dot{u}_{h\mu} + \nabla \dot{u}_{h\mu}{}^T), \alpha\tau_s\right) = 0 \tag{37}$$

for all τ_s, τ_h such that $(\alpha\tau_h + (1-\alpha)\tau_s, \nabla\phi) = 0$, $\phi \in H^1_\Gamma$. For arbitrary $\tau_s^0 \in L^2$ we define

$$\tau_h^0 = -\frac{\alpha}{1-\alpha}\tau_s^0. \tag{38}$$

Then obviously (τ_s^0, τ_h^0) satisfies the balance of linear momentum with zero force and we conclude that

$$\frac{1}{2}(\nabla \dot{u}_{s\mu} + \nabla \dot{u}_{s\mu}{}^T) = \frac{1}{2}(\nabla \dot{u}_{h\mu} + \nabla \dot{u}_{h\mu}{}^T) \qquad \text{in } L^2. \tag{39}$$

Remark 7. Clearly, (36) extends to the limit $\mu \to +0$ in the space $C^*([0,T] \times \Omega; \mathbb{R}^n)$ of Riesz measures.

From the mathematical point of view, we believe that (33) and (34) are the 'best' equations to understand the analysis of Prandtl-Reuss mixtures. The functions $v_{s\mu}$ and $v_{h\mu}$ can be interpreted as the approximate total strain velocities for the soft and the hard material. We have written $v_{s\mu} = \dot{u}_{s\mu}$, $v_{h\mu} = \dot{u}_{h\mu}$, assuming some initial condition for $u_{s\mu}, u_{h\mu}$. The penalty terms correspond to the (approximate) velocities of plastic deformation and the terms $A_s \frac{\partial}{\partial t}\big(\alpha\sigma_{s\mu}\big)$, $A_h \frac{\partial}{\partial t}\big(\alpha\sigma_{h\mu}\big)$ model the elastic deformation of the hard and soft material. Note that (33) and (34) are equivalent to (27).

From the estimates of the penalty term, proved in the following sections, we state the following theorem.

Theorem 6. *Under the assumptions of Theorem 4 we have the uniform estimate*

$$\sup_{\mu}\left\{||\nabla\dot{u}_{s\mu}+\nabla\dot{u}_{s\mu}{}^{T}||_{L^1(L^1)}+||\nabla\dot{u}_{h\mu}+\nabla\dot{u}_{h\mu}{}^{T}||_{L^1(L^1)}\right\}\le K(\dot{\alpha},\dot{\kappa}_s,\dot{\kappa}_h). \quad (40)$$

Due to Temam's imbedding theorem we derive

Corollary 1.

$$\sup_{\mu}\left\{||\dot{u}_{s\mu}||_{L^1(L^{\frac{n}{n-1}})}+||\dot{u}_{h\mu}||_{L^1(L^{\frac{n}{n-1}})}\right\}\le K(\dot{\alpha},\dot{\kappa}_s,\dot{\kappa}_h) \quad \textit{as} \quad \mu\to+0. \quad (41)$$

From the $L^1(L^1)$-estimates for the strain velocities and the penalty terms we have, for a subsequence

$$\frac{1}{2}(\nabla\dot{u}_{s\mu}+\nabla\dot{u}_{s\mu}{}^{T})\rightharpoonup\frac{1}{2}(\nabla\dot{u}_{s}+\nabla\dot{u}_{s}{}^{T}) \quad \text{weakly in } C^*([0,T]\times\bar{\Omega};\mathbb{R}^n), \quad (42)$$

$$\frac{1}{2}(\nabla\dot{u}_{h\mu}+\nabla\dot{u}_{h\mu}{}^{T})\rightharpoonup\frac{1}{2}(\nabla\dot{u}_{h}+\nabla\dot{u}_{h}{}^{T}) \quad \text{weakly in } C^*([0,T]\times\bar{\Omega};\mathbb{R}^n), \quad (43)$$

i.e. the limiting strains are only Riesz-measures. If more regularity is assumed in the safe load condition (see further theorems) we have that $\dot{u}_s,\dot{u}_h\in L^\infty(L^{\frac{n}{n-1}})$, i.e. the velocities and the displacements, are at least functions.

For the penalty terms we have

$$\mu^{-1}[|\sigma_{s\mu D}|-\kappa_s]_+\frac{\sigma_{s\mu D}}{|\sigma_{s\mu D}|}\rightharpoonup\dot{e}_{ps}, \qquad \mu^{-1}[|\sigma_{h\mu D}|-\kappa_h]_+\frac{\sigma_{h\mu D}}{|\sigma_{h\mu D}|}\rightharpoonup\dot{e}_{ph} \quad (44)$$

both weakly in $C^*([0,T]\times\bar{\Omega})$, as $\mu\to+0$.

If we would know that $\sigma_{s\mu}\to\sigma_s$, $\sigma_{h\mu}\to\sigma_h$ in C (= space of continuous functions), we could prove the representation

$$\dot{e}_{ps}=\lambda_s\frac{\sigma_{sD}}{|\sigma_{sD}|}, \quad \dot{e}_{ph}=\lambda_h\frac{\sigma_{hD}}{|\sigma_{hD}|}, \quad (45)$$

where λ_s, λ_h is the weak C^*-limit of $\mu^{-1}[|\sigma_{s\mu D}|-\kappa_s]_+$, $\mu^{-1}[|\sigma_{h\mu D}|-\kappa_h]_+$. The support of λ_s and λ_h is on the set $|\sigma_{sD}|\ge\kappa_s$ and $|\sigma_{hD}|\ge\kappa_h$, respectively. In the case of two dimensions there is a substitute of the argument, taking into account that $\sigma_s,\sigma_h\in C$ is not known, see the discussion in Sect. 9. With the above convergences in C^* the solution of the variational inequality satisfies the equations

$$\frac{1}{2}(\nabla\dot{u}_s+\nabla\dot{u}_s{}^{T})=A_s\frac{\partial}{\partial t}(\alpha\sigma)+\dot{e}_{ps}, \quad (46)$$

$$\frac{1}{2}(\nabla\dot{u}_h+\nabla\dot{u}_h{}^{T})=A_h\frac{\partial}{\partial t}((1-\alpha)\sigma)+\dot{e}_{ph}. \quad (47)$$

With a more restrictive assumption we gain $L^\infty(L^2)$-bounds for $\dot{\sigma}_{s\mu}$, $\dot{\sigma}_{h\mu}$ and $L^\infty(L^1)$-bounds for the partial strains. This is proved in Sect. 7.

Theorem 7. *Under the assumption of Theorem 4 and the additional requirement that the safe loads and the data α, κ_s, κ_h satisfy*

$$\ddot{\tilde{\sigma}}_s, \ddot{\tilde{\sigma}}_h, \ddot{\alpha}, \ddot{\kappa}_s, \ddot{\kappa}_h \in L^\infty(L^\infty) \tag{48}$$

there holds the uniform bound

$$\sup_\mu \left\{ \left\| \dot{\sigma}_{s\mu} \right\|_{L^\infty(L^2)} + \left\| \dot{\sigma}_{h\mu} \right\|_{L^\infty(L^2)} \right\} \leq K(\ddot{\alpha}, \ddot{\kappa}_s, \ddot{\kappa}_h) \tag{49}$$

and

$$\left\| \nabla \dot{u}_{s\mu} + \nabla \dot{u}_{s\mu}{}^T \right\|_{L^\infty(L^1)} + \left\| \nabla \dot{u}_{h\mu} + \nabla \dot{u}_{h\mu}{}^T \right\|_{L^\infty(L^1)} \leq K(\ddot{\alpha}, \ddot{\kappa}_s, \ddot{\kappa}_h). \tag{50}$$

Remark 8. In the above estimates we indicate how the bounds depend on the derivative of α, κ_s, κ_h. This is relevant, later, for the treatment of the quasivariational inequality, where α, κ_s, κ_h depend on σ_s (and also σ_h). A dependence of $\dot{\alpha}$, $\dot{\kappa}_s$, $\dot{\kappa}_h$ does not give problems modelling α, κ_s, κ_h, but a dependence of $\ddot{\alpha}$, $\ddot{\kappa}_s$, $\ddot{\kappa}_h$ leads to restrictions.

It is useful to observe that the solutions of the penalty problems satisfy the variational inequalities

$$\left(A_s \frac{\partial}{\partial t} \left(\alpha \sigma_{s\mu} \right), \sigma_{s\mu} - \omega_s \right) \leq \left(\frac{1}{2} (\nabla \dot{u}_{s\mu} + \nabla \dot{u}_{s\mu}{}^T), \sigma_{s\mu} - \omega_s \right) \tag{51}$$

$$\left(A_h \frac{\partial}{\partial t} \left((1 - \alpha) \sigma_{h\mu} \right), \sigma_{h\mu} - \omega_h \right) \leq \left(\frac{1}{2} (\nabla \dot{u}_{h\mu} + \nabla \dot{u}_{h\mu}{}^T), \sigma_{h\mu} - \omega_h \right) \tag{52}$$

a.e. with respect to t, for all $\omega_s, \omega_h \in L^2(\Omega, M^n_{sym})$ such that $|\omega_{sD}| \leq \kappa_s$, $|\omega_{hD}| \leq \kappa_h$.

This follows from the

$$\begin{aligned} &\left([|\sigma_{sD}| - \kappa_s]_+ \frac{\sigma_{sD}}{|\sigma_{sD}|} : (\sigma_{sD} - \omega_{sD}) \right) = \\ &\left([|\sigma_{sD}| - \kappa_s]_+ \frac{\sigma_{sD}}{|\sigma_{sD}|} - \underbrace{[|\omega_{sD}| - \kappa_s]_+}_{=0} \frac{\omega_{sD}}{|\omega_{sD}|}, \sigma_{sD} - \omega_{sD} \right) \geq 0 \end{aligned} \tag{53}$$

and, correspondingly, for the hard material.

4 The Rothe Approximation

4.1 Definition and Solvability of the Rothe Approximation

We discretize the loading interval [0,T] by a discrete set $I_\delta = \{k\delta | k = 0, \dots, N\}$ with mesh size $\delta = T/N$ and approximate $\frac{\partial}{\partial t} w(t,.)$ by the backward difference quotient

$$D^{-\delta} w(t,.) = \delta^{-1}(w(t,.) - w(t-\delta,.)). \tag{54}$$

Then the **Rothe approximation** of the penalty approximation (27) reads:

Find a pair $(\sigma_s, \sigma_h) = (\sigma_{s\mu\delta}, \sigma_{h\mu\delta}) : I_\delta \times \Omega \to M^n_{sym} \times M^n_{sym}$ *such that* (σ_s, σ_h) *satisfies* (11) *for* $t \in I_\delta$ *and that*

$$Q \begin{pmatrix} D^{-\delta}(\alpha\sigma_s) & \alpha\tau_s \\ D^{-\delta}((1-\alpha)\sigma_h) & (1-\alpha)\tau_h \end{pmatrix} \tag{55}$$

$$+ (\mu^{-1}[|\sigma_{sD}| - \kappa_s]_+ \frac{\sigma_{sD}}{|\sigma_{sD}|}, \alpha\tau_s) + (\mu^{-1}[|\sigma_{hD}| - \kappa_h]_+ \frac{\sigma_{hD}}{|\sigma_{hD}|}, (1-\alpha)\tau_h) = 0$$

for all $(\tau_s, \tau_h) : \Omega \to M^n_{sym} \times M^n_{sym}$, $\tau_s, \tau_h \in L^2$ *such that* (28) *is satisfied.*

Lemma 1. *Let* A_s, A_h *be positively definite and let the set of all* (σ_s, σ_h) *satisfying the balance of linear momentum not be empty. Let the Assumption 1 on* α *be satisfied. Then* (55) *has a unique solution* (σ_s, σ_h).

Proof. We assume that $(\sigma_s, \sigma_h)(t)$ has been constructed for $t = 0, \delta, \dots, (k-1)\delta$ and we want to construct $(\sigma_s, \sigma_h)(t^*)$, $t^* = k\delta$. This is done by minimizing the functional

$$\begin{aligned} J\ (\sigma_s^*, \sigma_h^*) &= \frac{1}{2\delta}(\alpha^2(t^*)A_s\sigma_s^*, \sigma_s^*) + \frac{1}{2\delta}((1-\alpha(t^*))^2 A_h\sigma_h^*, \sigma_h^*) \\ &-\frac{1}{\delta}(\alpha(t^*-\delta)A_s\sigma_s(t^*-\delta), \alpha(t^*)\sigma_s^*) - \frac{1}{\delta}((1-\alpha(t^*-\delta))A_h\sigma_h(t^*-\delta), \sigma_h^*) \\ &+\frac{1}{2\mu}\int_\Omega \alpha(t^*)[|\sigma_{sD}^*| - \kappa_s(t^*)]_+^2 dx + \frac{1}{2\mu}\int_\Omega (1-\alpha(t^*))[|\sigma_{hD}^*| - \kappa_h(t^*)]_+^2 dx \end{aligned} \tag{56}$$

on the set of pairs $(\sigma_s^*, \sigma_h^*) : \Omega \to M^n_{sym} \times M^n_{sym}$, $\sigma_s^*, \sigma_h^* \in L^2$ which satisfies the balance of linear momentum

$$(\alpha(t^*)\sigma_s^* + (1-\alpha(t^*))\sigma_h^*, \nabla\phi) = (f(t^*), \phi) + \int_{\partial\Omega} p_0(t^*)\phi,\ \phi \in H^{1,2}_\Gamma. \tag{57}$$

Since the functional J is strictly convex, coercive, and continuous in the strong topology of L^2, a unique minimizer (σ_s^*, σ_h^*) exists and we define $\sigma_s(t^*) = \sigma_s^*$ and $\sigma_h(t^*) = \sigma_h^*$.

It is easy to see that the Lagrange-Euler equation to the above minimization problem is just the Rothe approximation (55). The uniqueness follows with a monotonicity argument. □

4.2 First Estimates for the Rothe Approximation

In this section, we derive discrete versions of $L^\infty(L^2)$-estimates for the solutions (σ_s, σ_h) of the Rothe equation and also discrete versions of $L^1(L^1)$-estimates for the penalty term. These estimates are uniform as $\delta \to +0$. Since it is convenient to have the uniformity of these estimates also with respect to $\mu \to +0$, we assume a compatibility condition for the yield conditions and the balance of linear momentum.

Condition 3 (Weak safe load condition). *There exists $(\tilde{\sigma}_s, \tilde{\sigma}_h) \in \mathbb{K}$ (cf. Sect. 2.4) such that $\tilde{\sigma}_s, \tilde{\sigma}_h, \dot{\tilde{\sigma}}_s, \dot{\tilde{\sigma}}_h \in L^\infty(L^2)$.*

Theorem 8. *Let $(\sigma_s, \sigma_h) = (\sigma_{s\mu\delta}, \sigma_{h\mu\delta})$ be a solution of the Rothe problem, and let Assumptions 1 and 2 and Condition 3 hold. Then*

$$\max_{t=0,\dots,N\delta} \int_\Omega |\sigma_s|^2 + |\sigma_h|^2 \,\mathrm{d}x$$

$$+\delta \sum_{t=\delta,\dots,N\delta} \int_\Omega [|\sigma_{sD}| - \kappa_s]_+ ||\sigma_{sD}| - |\tilde{\sigma}_{sD}|| + [|\sigma_{hD}| - \kappa_h]_+ ||\sigma_{hD}| - |\tilde{\sigma}_{hD}|| \,\mathrm{d}x$$

$$\le K + K \int_0^t \int_\Omega |\frac{\partial}{\partial t}(\alpha\tilde{\sigma}_s)|^2 + |\frac{\partial}{\partial \alpha}((1-\alpha)\tilde{\sigma}_h)|^2 \,\mathrm{d}x\,\mathrm{d}t, \quad (58)$$

where the constant K does not depend on $\delta \to +0$ and $\mu \to +0$.

Proof. We use the pair $(\sigma_s - \tilde{\sigma}_s, \sigma_h - \tilde{\sigma}_h)$ as a test function in (55) and obtain

$$\delta \sum_{t=\delta,\dots,N\delta} Q \begin{pmatrix} D^{-\delta}(\alpha\sigma_s) & \alpha(\sigma_s - \tilde{\sigma}_s) \\ D^{-\delta}((1-\alpha)\sigma_h) & (1-\alpha)(\sigma_h - \tilde{\sigma}_h) \end{pmatrix}$$

$$+ \mu^{-1}\delta \sum_{t=\delta,\dots,N\delta} \int_\Omega [|\sigma_{sD}| - \kappa_s]_+ \frac{\sigma_s}{|\sigma_s|} : (\sigma_s - \tilde{\sigma}_s) \quad (59)$$

$$+ [|\sigma_{hD}| - \kappa_h]_+ \frac{\sigma_h}{|\sigma_h|} : (\sigma_h - \tilde{\sigma}_h) \,\mathrm{d}x = 0.$$

We abbreviate

$$E_t = Q \begin{pmatrix} \alpha\sigma_s & \alpha\sigma_s \\ (1-\alpha)\sigma_h & (1-\alpha)\sigma_h \end{pmatrix}\Big|_t \tag{60}$$

and we use Hölder's inequality

$$\begin{aligned} &Q \begin{pmatrix} D^{-\delta}(\alpha\sigma_s) & \alpha(\sigma_s-\tilde\sigma_s) \\ D^{-\delta}((1-\alpha)\sigma_h) & (1-\alpha)(\sigma_h-\tilde\sigma_h) \end{pmatrix}\Big|_t \\ &\qquad \geq \frac{1}{2\delta}(E_t - E_{t-\delta}) - Q \begin{pmatrix} D^{-\delta}(\alpha\sigma_s) & \alpha\tilde\sigma_s \\ D^{-\delta}((1-\alpha)\sigma_h) & (1-\alpha)\tilde\sigma_h \end{pmatrix}\Big|_t . \end{aligned} \tag{61}$$

Using the arguments similar to (53) both for the soft and hard material we observe that the third term in (59), which comes from the penalty, is nonnegative and that

$$\begin{aligned} \text{penalty terms in (59)} \geq \mu^{-1}\delta \sum_{t=\delta,\ldots,N\delta} \int_\Omega [|\sigma_{sD}| - \kappa_s]_+(|\sigma_{sD}| - |\tilde\sigma_{sD}|) + \\ + [|\sigma_{hD}| - \kappa_h]_+(|\sigma_{hD}| - |\tilde\sigma_{hD}|)\,\mathrm{d}x \geq 0. \end{aligned} \tag{62}$$

Note that $|\sigma_{sD}| - |\tilde\sigma_{sD}| \geq 0$ on $[|\sigma_{sD}| - \kappa_s]_+$, similar for $(|\sigma_{hD}| - |\tilde\sigma_{hD}|)$, and it also holds that

$$\frac{\sigma_{sD}}{|\sigma_{sD}|} : (\sigma_s - \tilde\sigma_s) \geq |\sigma_{sD}| - |\tilde\sigma_s|. \tag{63}$$

From (59), (61), and (62) we obtain

$$\begin{aligned} &T_1 + T_2 + T_3 := \\ &\quad \delta \sum_{t=\delta,\ldots,N\delta} \Bigg\{ \frac{1}{2\delta}(E_t - E_{t-\delta}) - Q \begin{pmatrix} D^{-\delta}(\alpha\sigma_s) & \alpha\tilde\sigma_s \\ D^{-\delta}((1-\alpha)\sigma_h) & (1-\alpha)\tilde\sigma_h \end{pmatrix} \\ &\qquad + \frac{1}{\mu}\int_\Omega \alpha[|\sigma_{sD}| - \kappa_s]_+(|\sigma_{sD}| - |\hat\sigma_{sD}|) \\ &\qquad + (1-\alpha)[|\sigma_{hD}| - \kappa_h]_+(|\sigma_{hD}| - |\hat\sigma_{hD}|)\,\mathrm{d}x \Bigg\} \leq 0. \end{aligned} \tag{64}$$

Finally, we obtain via partial summation and Hölder's inequality

$$\delta \sum_{t=\delta,\dots,N\delta} Q \begin{pmatrix} D^{-\delta}(\alpha\sigma_s) & \alpha\tilde{\sigma}_s \\ D^{-\delta}((1-\alpha)\sigma_h) & (1-\alpha)\tilde{\sigma}_h \end{pmatrix}$$

$$= -\delta \sum_{t=\delta,\dots,N\delta} Q \begin{pmatrix} \alpha\sigma_s & D^{\delta}\,(\alpha\tilde{\sigma}_s) \\ (1-\alpha)\sigma_h & D^{\delta}\,((1-\alpha)\tilde{\sigma}_h) \end{pmatrix}$$

$$+ \int_\Omega Q \begin{pmatrix} \alpha\sigma_s & \alpha\tilde{\sigma}_s \\ (1-\alpha)\sigma_h & (1-\alpha)\tilde{\sigma}_h \end{pmatrix} \mathrm{d}x \Big|_{t=0}^{t=N\delta}$$

$$\le K\delta \sum_{t=\delta,\dots,N\delta} \int_\Omega |\sigma_s|^2 + |\sigma_h|^2 \,\mathrm{d}x \tag{65}$$

$$+ K \int_0^t \int_\Omega |\frac{\partial}{\partial t}(\alpha\tilde{\sigma}_s)|^2 + |\frac{\partial}{\partial t}((1-\alpha)\tilde{\sigma}_h)|^2 \,\mathrm{d}x\,\mathrm{d}t$$

$$+ \epsilon_0 \int_\Omega |\sigma_s|^2 + |\sigma_h|^2 \,\mathrm{d}x\Big|_{t=N\delta} + K_{\epsilon_0} \int_\Omega |\tilde{\sigma}_s|^2 + |\tilde{\sigma}_h|^2 \,\mathrm{d}x\Big|_{t=N\delta}$$

$$+ K + \int_\Omega |\sigma_{os}|^2 + |\sigma_{oh}|^2 \,\mathrm{d}x.$$

Recall that σ_{os}, σ_{oh} are the initial values of σ_s, σ_h. We use this inequality in (64) for estimating from below. Observe also that

$$\sum_{t=\delta,\dots,N\delta} E_t - E_{t-\delta} = E_{N\delta} - E_0. \tag{66}$$

Since $T_3 \ge 0$ the statement of Theorem 8 then follows by using a discrete version of Gronwall's inequality. □

Corollary 2. *Under the additional assumption of the safe load Condition 2 for $\tilde{\sigma}_s$, $\tilde{\sigma}_h$ we have, for the solutions of the Rothe approximation, the discrete $L^1(L^1)$ estimate*

$$\mu^{-1}\delta \sum_{t=\delta,\dots,N\delta} \int_\Omega [|\sigma_{sD}| - \kappa_s]_+ + [|\sigma_{hD}| - \kappa_h]_+ \,\mathrm{d}x$$

$$+ \mu^{-1}\delta \sum_{t=\delta,\dots,N\delta} \int_\Omega [|\sigma_{sD}| - \kappa_s]_+ |\sigma_{sD}| + [|\sigma_{hD}| - \kappa_h]_+ |\sigma_{hD}| \,\mathrm{d}x$$

$$\le K + K \int_0^t \int_\Omega |\frac{\partial}{\partial t}(\alpha\tilde{\sigma}_s)| + |\frac{\partial}{\partial \alpha}((1-\alpha)\tilde{\sigma}_h)|^2 \,\mathrm{d}x\,\mathrm{d}t \tag{67}$$

holds uniformly as $\delta \to +0$, $\mu \to +0$.

Proof. If we have the safe load condition, the penalty part in (64) can be estimated from below by

$$P_{\delta\mu 1} = \mu^{-1}\delta \sum_{t=\delta,\dots,N\delta} \int_{\Omega} \alpha[|\sigma_{sD}| - \kappa_s]_+ s_0 + (1-\alpha)[|\sigma_{hD}| - \kappa_h]_+ s_0 \,\mathrm{d}x, \tag{68}$$

where $s_0 > 0$ comes from the safe load condition. Thus, this term contributes to the estimates and we obtain

$$|P_{\delta\mu 1}| \leq K \text{ uniformly.} \tag{69}$$

Once knowing this,by inspection of (64), we observe that also

$$P_{\delta\mu 2} = \mu^{-1}\delta \sum_{t=\delta,\dots,N\delta} \int_{\Omega} \alpha[|\sigma_{sD}| - \kappa_s]_+ |\sigma_{sD}| + (1-\alpha)[|\sigma_{hD}| - \kappa_h]_+ |\sigma_{hD}| \,\mathrm{d}x \tag{70}$$

remains bounded as $\mu \to 0$ and $\delta \to +0$. □

5 Estimates for the Rothe Approximation

We finally present a discrete analogue of an $H^{1,2}(L^2)$-estimate for the solutions of the Rothe-equation.

Theorem 9. *Assume the safe load condition with an admissible pair* $(\tilde{\sigma}_s, \tilde{\sigma}_h) \in \mathbb{K}$ *such that* $\dot{\tilde{\sigma}}_s, \dot{\tilde{\sigma}}_h \in L^\infty(L^\infty)$*. Let* $\dot{\alpha}, \dot{\kappa}_s, \dot{\kappa}_h \in L^\infty(L^\infty)$*, and* A_s, A_h *be positively definite and symmetric. Then there is a constant* $C(\dot{\alpha}, \dot{\kappa}_s, \dot{\kappa}_h, \dot{\tilde{\sigma}}_s, \dot{\tilde{\sigma}}_h)$ *such that for the solution* $(\sigma_s, \sigma_h) = (\sigma_{s\mu\delta}, \sigma_{h\mu\delta})$

$$\begin{aligned} &\delta \sum_{t=\delta,\dots,N\delta} \int_{\Omega} |D^{-\delta}\sigma_s|^2 + |D^{-\delta}\sigma_h|^2 \,\mathrm{d}x \\ &+ \mu^{-1} \sup_{t=0,\dots,N\delta} \int_{\Omega} [|\sigma_{sD}| - \kappa_s]_+^2 + [|\sigma_{hD}| - \kappa_h]_+^2 \,\mathrm{d}x \leq C(\dot{\alpha}, \dot{\kappa}_s, \dot{\kappa}_h, \dot{\tilde{\sigma}}_s, \dot{\tilde{\sigma}}_h) \end{aligned} \tag{71}$$

uniformly as $\delta \to 0$, $\mu \to 0$.

Proof. We use the shift operator $S^{-\delta}$ defined by

$$S^{-\delta}(t) = w(t-\delta). \tag{72}$$

It is easy to see that the following pair (τ_s^*, τ_h^*) satisfies the balance of linear momentum with zero force

$$\tau_s^* = \sigma_s - \tilde{\sigma}_s - \alpha^{-1} S^{-\delta}\left(\alpha(\sigma_s - \tilde{\sigma}_s)\right), \tag{73}$$

$$\tau_h^* = \sigma_h - \tilde{\sigma}_h - (1-\alpha)^{-1} S^{-\delta}\left((1-\alpha)(\sigma_h - \tilde{\sigma}_h)\right). \tag{74}$$

Hence we use this pair as test function in the Rothe approximation and we obtain, multiplying with δ^{-1},

$$\int_\Omega A_s D^{-\delta}(\alpha\sigma_s) : D^{-\delta}(\alpha(\sigma_s - \tilde{\sigma}_s)) \\ + A_h D^{-\delta}((1-\alpha)(\sigma_h - \tilde{\sigma}_h)) : D^{-\delta}((1-\alpha)(\sigma_h - \tilde{\sigma}_h))\,\mathrm{d}x + \text{penalty part} = 0. \tag{75}$$

The penalty part consists of a contribution of the soft material, namely

$$P_s = \int_\Omega \mu^{-1}[|\sigma_{sD}| - \kappa_s]_+ \frac{\sigma_{sD}}{|\sigma_{sD}|} D^{-\delta}\left(\alpha(\sigma_s - \tilde{\sigma}_s)\right)\mathrm{d}x, \tag{76}$$

and an analogous term P_h. We rewrite and estimate P_s in view of the discrete Leibniz rule. Thus,

$$P_s = \mu^{-1}\int_\Omega [|\sigma_{sD}| - \kappa_s]_+ \frac{\sigma_{sD}}{|\sigma_{sD}|}(S^{-\delta}\alpha D^{-\delta}\sigma_{sD} + D^{-\delta}\alpha\sigma_{sD})\,\mathrm{d}x \\ - \mu^{-1}\int_\Omega [|\sigma_{sD}| - \kappa_s]_+ \frac{\sigma_{sD}}{|\sigma_{sD}|} D^{-\delta}(\alpha\tilde{\sigma}_{sD})\,\mathrm{d}x = P_{1s} + P_{2s} + P_{3s}. \tag{77}$$

Since

$$\frac{\sigma_{sD}}{|\sigma_{sD}|} D^{-\delta}\sigma_{sD} \geq D^{-\delta}|\sigma_{sD}|, \tag{78}$$

we estimate and rewrite

$$P_{1s} \geq \mu^{-1}\int_\Omega [|\sigma_{sD}| - \kappa_s]_+ S^{-\delta}\alpha D^{-\delta}|\sigma_{sD}|\,\mathrm{d}x \\ = \mu^{-1}\int_\Omega [|\sigma_{sD}| - \kappa_s]_+ S^{-\delta}\alpha D^{-\delta}(|\sigma_{sD}| - \kappa_s)\,\mathrm{d}x \\ + \mu^{-1}\int_\Omega [|\sigma_{sD}| - \kappa_s]_+ S^{-\delta}\alpha D^{-\delta}\kappa_s\,\mathrm{d}x = P_{11s} + P_{12s}. \tag{79}$$

Since

$$[|\sigma_{sD}| - \kappa_s]_+ D^{-\delta}(|\sigma_{sD}| - \kappa_s) \geq \frac{1}{2} D^{-\delta}\left([|\sigma_{sD}| - \kappa_s]_+^2\right), \tag{80}$$

we obtain

$$P_{11s} \geq \frac{1}{2}\mu^{-1}\alpha_0 \int_\Omega D^{-\delta}\left([|\sigma_{sD}| - \kappa_s]_+^2\right) \mathrm{d}x. \tag{81}$$

Furthermore, due to the Lipschitz continuity of κ_s and the $L^1(L^1)$ property stated in Corollary 2, we have

$$\delta \sum_{t=\delta,\ldots,N\delta} P_{12s} \geq -\mu^{-1}\delta \sum_{t=\delta,\ldots,N\delta} \int_\Omega [|\sigma_{sD}| - \kappa_s]_+ |D^{-\delta}\kappa_s| \mathrm{d}x \geq -C_{1s}(\dot{\kappa}_s). \tag{82}$$

With a similar argument, we obtain

$$\delta \sum_{t=\delta,\ldots,N\delta} P_{2s} \geq -\mu^{-1}\delta \sum_{t=\delta,\ldots,N\delta} \int_\Omega [|\sigma_{sD}| - \kappa_s]_+ |D^{-\delta}\alpha||\sigma_{sD}| \,\mathrm{d}x \geq -C_{2s}(\dot{\alpha}) \tag{83}$$

and analogously

$$P_{3s} \geq -C_{3s}(\alpha), \tag{84}$$

where we used the Lipschitz continuity of α and the assumption that $\dot{\tilde{\sigma}}_s \in L^\infty$. From (81) we obtain

$$\begin{aligned} \delta \sum_{t=\delta,\ldots,N\delta} P_{11s} \geq \mu^{-1} & \int_\Omega [|\sigma_{sD}| - \kappa_s]_+^2 \alpha \,\mathrm{d}x \Big|_0^T \\ & -\mu^{-1}\delta \sum_{t=\delta,\ldots,N\delta} \int_\Omega [|\sigma_{sD}| - \kappa_s]_+^2 |D^{\delta}\alpha| \,\mathrm{d}x \\ & = P_{111s} + P_{112s}. \end{aligned} \tag{85}$$

Again

$$P_{112s} \geq -C_{112}(\dot{\alpha}). \tag{86}$$

The constants $C_{1s}, \ldots, C_{112}$ depend on the $L^1(L^1)$ estimate for the penalty term, so, they depend on $\dot{\alpha}, \dot{\kappa}_s, \dot{\kappa}_h$. The penalty parts for the hard material are treated in a similar manner.

We now sum (75) from $t = \delta$ up to $t = N\delta = T$ and obtain, using the estimates for the penalty parts,

$$\frac{1}{2}\sum_{t=\delta}^{N\delta}\int_{\Omega} A_s D^{-\delta}\left(\alpha\sigma_s\right) : D^{-\delta}\left(\alpha\sigma_s\right) + A_h D^{-\delta}\left((1-\alpha)\sigma_h\right) : D^{-\delta}\left((1-\alpha)\sigma_h\right)\,\mathrm{d}x$$

$$\le K(\dot{\alpha}, \dot{\kappa}_s, \dot{\kappa}_h) + \frac{1}{2}\delta \sum_{t=\delta,\ldots,N\delta}\int_{\Omega} A_s D^{-\delta}\left(\alpha\tilde{\sigma}_s\right) : D^{-\delta}\left(\alpha\tilde{\sigma}_s\right) \tag{87}$$

$$+ A_h D^{-\delta}\left((1-\alpha)\sigma_h\right) : D^{-\delta}\left((1-\alpha)\tilde{\sigma}_h\right)\,\mathrm{d}x.$$

The theorem then follows, taking into account that $0 < \alpha_0 < \alpha < 1-\alpha_0$. □

6 Convergence of the Rothe Method

In this section, we extend the solutions $\sigma_{s\mu\delta}, \sigma_{h\mu\delta}$ of the Rothe problem to the step functions $J_\delta\sigma_{s\mu\delta}$ and $J_\delta\sigma_{h\mu\delta}$ and show that they strongly converge in $L^2(L^2)$ to a solution $\sigma_{s\mu}, \sigma_{h\mu}$ of the penalty approximation as $\delta \to 0$. Throughout this section, except in the formulation of the theorem, we drop the index μ. For any function w defined on $I_\delta = \{k\delta | k = 0, \ldots, N\}$ we define the extension as a step function by

$$J_\epsilon w(k\delta + \eta) = w(k\delta), \quad 0 \le \eta \le \delta. \tag{88}$$

If w is not defined in $[0,\delta]$, we define $J_\delta w = 0$ on $[0,\delta]$. In our setting, this acts on the argument of the loading parameter t.

Theorem 10. *Let $\mu > 0$ be fixed, α, κ_s, κ_h be Lipschitz, $0 < \alpha_0 < \alpha < 1-\alpha_0 < 1$, $\kappa_s, \kappa_h \ge \kappa_0 > 0$. Let A_h, A_s be symmetric and positive definite. Assume the safe load condition with an admissible pair $(\tilde{\sigma}_s, \tilde{\sigma}_h) \in \mathbb{K}$ such that $\dot{\tilde{\sigma}}_s, \dot{\tilde{\sigma}}_h \in L^\infty(L^\infty)$. Then the solutions $\sigma_{s\delta\mu}, \sigma_{h\delta\mu}$ converge to a pair $(\sigma_{s\mu}, \sigma_{h\mu})$ which is a solution of the penalty problem in the sense*

$$J_\delta\sigma_{s\mu\delta} \to \sigma_{s\mu}, \quad J_\delta\sigma_{h\mu\delta} \to \sigma_{h\mu} \quad (\delta \to +0) \tag{89}$$

strongly in $L^2(L^2)$, and

$$J_\delta D^{-\delta}\left(\sigma_{s\mu\delta}\right) \to \dot{\sigma}_{s\mu}, \quad J_\delta D^{-\delta}\left(\sigma_{h\mu\delta}\right) \to \dot{\sigma}_{h\mu} \quad (\delta \to +0) \tag{90}$$

weakly in $L^2(L^2)$.

Proof. By the uniform estimates for $\sigma_{s\delta}, \sigma_{h\delta}$ from Theorem 8 we have uniform $L^2(L^2)$-estimates for the functions $J_\delta\sigma_{s\delta}$, $J_\delta\sigma_{h\delta}$, $J_\delta D^{-\delta}\sigma_{s\delta}$, $J_\delta D^{-\delta}\sigma_{h\delta}$ and, by weak compactness in $L^2(L^2)$, any sequence $(\delta_i \to +0)$ has a subsequence such that (89) holds with weak limits σ_s, σ_h $(\delta_i \to 0)$. Furthermore the functions $J_\delta D^{-\delta}\sigma_{s\delta}$, $J_\delta D^{-\delta}\sigma_{h\delta}$ have weak limits which turn out to have the form $\dot{\sigma}_s, \dot{\sigma}_h$, i.e. they are the derivatives of σ_s, σ_h. This is classical and easy to prove. As a consequence, $\sigma_s(t,.)$ and $\sigma_h(t,.)$ are defined for $t \in [0,T]$ as $L^2(\Omega)$ functions. By weak convergence, one sees immediately that σ_s, σ_h satisfy the balance of linear momentum. Due to the representation

$$\sigma_{s0} = -\delta \sum_{t=\delta,\dots,N\delta} D^{-\delta}\sigma_{s\delta} + \sigma_s(T,.) \tag{91}$$

and by averaging with respect to T, we see via weak convergence that $\sigma_s(0,.) = \sigma_{s0}, \sigma_h(0,.) = \sigma_{h0}$, i.e. the weak limit satisfies the initial condition. The main task is to establish strong convergence in order to pass to the limit in the nonlinear penalty term. For this purpose, we define the restriction operator, which assigns to functions $w \in L^2(L^2)$ with $\dot{w} \in L^2(L^2)$ a function $R_\delta w$ on $I_\delta = \{\delta, \dots, N\delta\}$ defined by

$$R_\delta w(k\delta) = w(k\delta). \tag{92}$$

One has

$$J_\delta R_\delta w \to w, \quad J_\delta D^{-\delta} R_\delta w \to \dot{w} \quad \text{strongly in } L^2(L^2) \tag{93}$$

as $\delta \to +0$, provided that $\dot{w} \in L^2(L^2)$. We now turn to the Rothe equation and use the pair $(\sigma_{s\delta} - R_\delta\sigma_s, \sigma_{h\delta} - R_\delta\sigma_h)$ as test function. Note that this test function satisfies the balance of linear momentum with zero force for $t \in I_\delta$. Rewriting the resulting equation and employing the extension operator J_δ we conclude

$$\begin{aligned} \int_0^T & (A_s J_\delta D^{-\delta}(\alpha\sigma_{s\delta}), J_\delta(\alpha(\sigma_{s\delta} - R_\delta\sigma_s))) \\ & + (A_h J_\delta D^{-\delta}((1-\alpha)\sigma_{h\delta}), J_\delta((1-\alpha)(\sigma_{s\delta} - R_\delta\sigma_s))) \\ & + \mu^{-1}([|J_\delta\sigma_{s\delta D}| - J_\delta R_\delta\kappa_s]_+ J_\delta\frac{\sigma_{s\delta D}}{|\sigma_{s\delta D}|}, J_\delta(\alpha(\sigma_{s\delta D} - R_\delta\sigma_{sD}))) \\ & + \mu^{-1}([|J_\delta\sigma_{h\delta D}| - J_\delta R_\delta\kappa_h]_+ J_\delta\frac{\sigma_{h\delta D}}{|\sigma_{h\delta D}|}, J_\delta(\alpha(\sigma_{h\delta D} - R_\delta\sigma_{hD}))) = 0. \end{aligned} \tag{94}$$

In (55) we may add the terms

$$\int_0^T \mu^{-1}\left([J_\delta\sigma_{sD} - J_\delta R_\delta\kappa_s]_+ J_\delta\frac{\sigma_{sD}}{|\sigma_{sD}|}, J_\delta(\alpha\sigma_{s\delta D} - \alpha R_\delta\sigma_{sD})\right) \mathrm{d}t = o(1) \text{ as } \delta \to 0 \tag{95}$$

since the left hand factor in the scalar product is compact in L^2 for μ fixed, and there is a similar term for the hard material. The resulting penalty terms (i.e. summands with factor μ^{-1}) are ≥ 0 due to monotonicity and will be dropped, replacing $=$ by $\leq$. Furthermore, we may add the term

$$-\int_0^T \left(A_s J_\delta D^{-\delta} R_\delta(\alpha\sigma_s), J_\delta(\alpha\sigma_{s\delta} - \alpha R_\delta\sigma_s)\right) \mathrm{d}t = o(1) \tag{96}$$

and a similar term for the hard material due to weak convergence

$$z_{s\delta} := J_\delta(\alpha\sigma_{s\delta} - \alpha R_\delta \sigma_s) \rightharpoonup 0 \quad \text{in } L^2(L^2) \tag{97}$$

and due to strong $L^2(L^2)$ convergence of

$$J_\delta D^{-\delta} R_\delta(\alpha\sigma_s) \to \frac{\partial}{\partial t}(\alpha\sigma_s)\,. \tag{98}$$

Similarly $z_{h\delta} = J_\delta(\alpha\sigma_{h\delta} - \alpha R_\delta\sigma_h) \rightharpoonup 0$. Thus we are left with

$$\int_0^T \left(A_s D^{-\delta} z_{s\delta}, z_{s\delta}\right) + \left(A_h D^{-\delta} z_{h\delta}, z_{h\delta}\right) \mathrm{d}t \le o(1) \tag{99}$$

from which we conclude (see the analogous reasoning in Sect. 5) that

$$\delta^{-1}\int_{T-\delta}^T |z_{s\delta}|^2 + |z_{h\delta}|^2 \,\mathrm{d}t \le o(1). \tag{100}$$

This holds for all $T \in \{k\delta | k = 1, \ldots, N\}$ and we conclude $z_{s\delta} \to 0$, $z_{h\delta} \to 0$ strongly in $L^2(L^2)$ which implies

$$J_\delta\sigma_{s\delta} \to \sigma_s, \quad J_\delta\sigma_{h\delta} \to \sigma_h \text{ strongly in } L^2(L^2). \tag{101}$$

This allows us to pass to the limit in the Rothe equation (employing the extension operator J_δ) and we arrive at the equation

$$\begin{aligned} 0 = \int_{t_1}^{t_2} &\left(A_s \frac{\partial}{\partial t}(\alpha\sigma_s), \tau_s\right) + \left(A_h \frac{\partial}{\partial t}((1-\alpha)\sigma_h), \tau_h\right) \\ &+ \left(\mu^{-1}[|\sigma_{sD}| - \kappa_s]_+ \frac{\sigma_{sD}}{|\sigma_{sD}|}, \alpha\tau_s\right) + \left(\mu^{-1}[|\sigma_{hD}| - \kappa_h]_+ \frac{\sigma_{hD}}{|\sigma_{hD}|}, (1-\alpha)\tau_h\right) \mathrm{d}t \end{aligned} \tag{102}$$

valid for all τ_s, τ_h satisfying the balance of linear momentum with zero force.

This proves the convergence of the Rothe method and the existence of solutions $\sigma_s, \sigma_h \in L^2(L^2), \dot\sigma_s, \dot\sigma_h \in L^2(L^2)$ of the penalty equation. □

Since the discrete $L^2(L^2)$ norms of $\sigma_{s\mu\delta}$, $\sigma_{h\mu\delta}$, $D^{-\delta}\left(\sigma_{s\mu\delta}\right)$, $D^{-\delta}\left(\sigma_{h\mu\delta}\right)$ are uniformly bounded with respect to $\mu \to +0$ (with error terms converging to zero as $\delta \to 0$, μ fixed) we obtain the corresponding bounds for $\sigma_{s\mu}$, $\sigma_{h\mu}$ as $\mu \to +0$.

With a similar reasoning we have a uniform $L^\infty(L^2)$ bound for the penalty potentials $\mu^{-1}[|\sigma_{s\mu D}| - \kappa_s]_+^2$, $\mu^{-1}[|\sigma_{h\mu D}| - \kappa_h]_+^2$ as well as a uniform $L^1(L^1)$ bound for the terms $\mu^{-1}[|\sigma_{s\mu D}| - \kappa_s]_+(\left|\sigma_{s\mu D}\right| + 1)$, $\mu^{-1}[|\sigma_{h\mu D}| - \kappa_h]_+(\left|\sigma_{h\mu D}\right| + 1)$. This proves the Theorems 4 and 6 of Sect. 3.

We finish the section with a conditional $L^\infty(L^1)$ estimate for the penalty term which is needed for the $L^\infty(L^2)$ estimate of $\dot\sigma_{s\mu}, \dot\sigma_{h\mu}$ in the next section.

Lemma 2 **($L^\infty(L^1)$).** *Under the assumption of the main theorem there is a constant $K(\dot\alpha, \dot{\tilde\sigma}_s, \dot{\tilde\sigma}_h)$ such that, for a.e. $t \in [0, T]$,*

$$\mu^{-1} \int_\Omega [|\sigma_{s\mu D}| - \kappa_s]_+(|\sigma_{s\mu D} + 1|) + [|\sigma_{h\mu D}| - \kappa_h]_+(|\sigma_{h\mu D} + 1|)\,\mathrm{d}x\Big|_t$$
$$\le K(\dot\alpha, \dot{\tilde\sigma}_s, \dot{\tilde\sigma}_h)\ \left(\left\|\dot\sigma_{s\mu}\right\|_{L^2(\Omega)} + \left\|\dot\sigma_{h\mu}\right\|_{L^2(\Omega)}\right)\Big|_t . \tag{103}$$

Proof. Use $\sigma_{s\mu} - \tilde\sigma_{s\mu}$, $\sigma_{h\mu} - \tilde\sigma_{h\mu}$ as test functions and use the safe load condition similar as in the $L^1(L^1)$ estimate for the penalty term before. This implies an estimate for the left hand side of (103) by

$$\left|\left(A_s \frac{\partial}{\partial t}\left(\alpha\sigma_s\right), \sigma_{s\mu} - \tilde\sigma_{s\mu}\right)\right| \tag{104}$$

and a corresponding term for the hard material. Since an $L^\infty(L^2)$ bound is available for $\sigma_{s\mu}, \tilde\sigma_{s\mu}, \sigma_{h\mu}, \tilde\sigma_{h\mu}$ we obtain (103). □

7 $L^\infty(L^2)$-Estimate for the Time Derivatives of the Stresses

In the theory of the classical Prandtl-Reuss-problem there is the well known inclusion $\dot\sigma \in L^\infty(L^2)$ for the stress σ. A similar theorem holds also for Prandtl-Reuss-mixtures, but the proof is a bit involved, although it is obviously motivated by the classical theory.

Theorem 11. *Let $\sigma_{\mu s}, \sigma_{\mu h} \in L^\infty(L^2)$ be the solution to the penalty approximation of the Prandtl-Reuss-mixture problem. Besides the hypotheses of the main theorem let*

$$\ddot\alpha, \ddot\kappa_s, \ddot\kappa_h, \ddot{\tilde\sigma}_s, \ddot{\tilde\sigma}_h \in L^\infty(L^\infty) \tag{105}$$

where $\tilde\sigma_s, \tilde\sigma_h$ are safe loads. Then

$$\left\|\dot\sigma_{s\mu}\right\|_{L^\infty(L^2)} + \left\|\dot\sigma_{h\mu}\right\|_{L^\infty(L^2)} \le K(\ddot\alpha, \ddot\kappa_s, \ddot\kappa_h, \ddot{\tilde\sigma}_s, \ddot{\tilde\sigma}_h) \tag{106}$$

uniformly as $\mu \to +0$.

Proof. Let $D^\eta = \eta^{-1}(S^\eta - I)$, $D^{-\eta} = \eta^{-1}(I - S^{-\eta})$ be the forward and backward difference operators with stepsize η, with respect to the loading variable t; $S^\eta w(t) = w(t + \eta)$. We write σ_s, σ_h rather than $\sigma_{s\mu}, \sigma_{h\mu}$. In the penalty equation we use the test functions

$$\begin{aligned} &-\eta^{-2} \ \left(\alpha^{-1} S^\eta \left(\alpha(\sigma_s - \tilde{\sigma}_s)\right) - 2(\sigma_s - \tilde{\sigma}_s) + \alpha^{-1} S^{-\eta} \left(\alpha(\sigma_s - \tilde{\sigma}_s)\right)\right), \\ &-\eta^{-2} \ \left((1-\alpha)^{-1} S^\eta \left((1-\alpha)(\sigma_h - \tilde{\sigma}_h)\right)\right. \\ &\qquad\qquad \left. -2(\sigma_h - \tilde{\sigma}_h) + (1-\alpha)^{-1} S^{-\eta} \left((1-\alpha)(\sigma_h - \tilde{\sigma}_h)\right)\right). \end{aligned} \tag{107}$$

They obey the balance of linear momentum with zero force. This yields

$$\begin{aligned} &-\left(A_s \frac{\partial}{\partial t} (\alpha\sigma_s), D^\eta D^{-\eta} \left(\alpha(\sigma_s - \tilde{\sigma}_s)\right)\right) \\ &-\left(A_h \frac{\partial}{\partial t} ((1-\alpha)\sigma_h), D^\eta D^{-\eta} \left((1-\alpha)(\sigma_h - \tilde{\sigma}_h)\right)\right) \\ &-\left(\mu^{-1}[|\sigma_{sD}| - \kappa_s]_+ \frac{\sigma_{sD}}{|\sigma_{sD}|}, D^\eta D^{-\eta} \left(\alpha(\sigma_s - \tilde{\sigma}_s)\right)\right) \\ &-\left(\mu^{-1}[|\sigma_{hD}| - \kappa_h]_+ \frac{\sigma_{hD}}{|\sigma_{hD}|}, D^\eta D^{-\eta} \left((1-\alpha)(\sigma_h - \tilde{\sigma}_h)\right)\right) = 0. \end{aligned} \tag{108}$$

We first get rid of the terms where $\tilde{\sigma}_s, \tilde{\sigma}_h$ occurs. We simply estimate (after integration $\int_{t_1}^{t_2} \mathrm{d}t$)

$$\begin{aligned} \left\| \left(A_s \frac{\partial}{\partial t} (\alpha\sigma_s), D^\eta D^{-\eta} (\alpha\tilde{\sigma}_s)\right) \right\| &\leq K \left\| \frac{\partial}{\partial t} (\alpha\sigma_s) \right\|_{L^2(L^2)} \left\| \frac{\partial^2}{\partial t^2} (\alpha\tilde{\sigma}_s) \right\|_{L^2(L^2)} \\ &\leq K_2(\ddot{\alpha}, \ddot{\tilde{\sigma}}_s) \end{aligned} \tag{109}$$

for $\mu \to \infty$, since a uniform $L^2(L^2)$ estimate for

$$\frac{\partial}{\partial t} (\alpha\sigma_s) = \frac{\partial}{\partial t} \left(\alpha\sigma_{s\mu}\right) \tag{110}$$

has been established in Sect. 4 and an appropriate estimate for $\tilde{\sigma}$ has been assumed. A similar reasoning holds for the hard material.

The penalty part where the factor $\tilde{\sigma}$ occurs is simply estimated by a constant $K(\ddot{\alpha}, \ddot{\tilde{\sigma}}_s)$ using the uniform $L^1(L^1)$ estimate for $\mu^{-1}[|\sigma_{sD}| - \kappa_s]_+$ and the $L^\infty(L^\infty)$ estimate for $\frac{\partial^2}{\partial t^2} (a\tilde{\sigma}_s)$ from the assumption. Again, a similar reasoning is done for the hard material. Thus we arrive at

$$\begin{aligned}
&-\int_{t_1}^{t_2}\left(A_s\frac{\partial}{\partial t}\left(\alpha\sigma_s\right),D^{\eta}D^{-\eta}(\alpha\sigma_s)\right)\,\mathrm{d}t\\
&-\int_{t_1}^{t_2}\left(A_h\frac{\partial}{\partial t}\left((1-\alpha)\sigma_h\right),D^{\eta}D^{-\eta}((1-\alpha)\sigma_h)\right)\,\mathrm{d}t \qquad (111)\\
&-\int_{t_1}^{t_2}\mu^{-1}\left([|\sigma_{sD}|-\kappa_s]_+\frac{\sigma_{sD}}{|\sigma_{sD}|},D^{\eta}D^{-\eta}(\alpha\sigma_s)\right)\,\mathrm{d}t\\
&-\int_{t_1}^{t_2}\mu^{-1}\left([|\sigma_{hD}|-\kappa_h]_+\frac{\sigma_{hD}}{|\sigma_{hD}|},D^{\eta}D^{-\eta}((1-\alpha)\sigma_h)\right)\,\mathrm{d}t\\
&\le K(\ddot{\alpha},\ddot{\bar{\sigma}}_s,\ddot{\bar{\sigma}}_h).
\end{aligned}$$

We now move the operator D^{η} to the first function in the scalar products ('partial summation'). It changes into $D^{-\eta}$, we obtain boundary terms $S_{t_1t_2}$ and see that

$$\begin{aligned}
&\int_{t_1}^{t_2}\left(A_sD^{-\eta}\left(\frac{\partial}{\partial t}\left(\alpha\sigma_s\right)\right),D^{-\eta}\left(\alpha\sigma_s\right)\right)\\
&\quad+\left(A_hD^{-\eta}\left(\frac{\partial}{\partial t}\left((1-\alpha)\sigma_h\right)\right),D^{-\eta}\left((1-\alpha)\sigma_h\right)\right)\,\mathrm{d}t\\
&\quad+\int_{t_1}^{t_2}\mu^{-1}\left(D^{-\eta}\left([|\sigma_{sD}|-\kappa_s]_+\frac{\sigma_{sD}}{|\sigma_{sD}|}\right),D^{-\eta}\left(\alpha\sigma_{sD}\right)\right)\,\mathrm{d}t \qquad (112)\\
&\quad+\int_{t_1}^{t_2}\mu^{-1}\left(D^{-\eta}\left([|\sigma_{hD}|-\kappa_h]_+\frac{\sigma_{hD}}{|\sigma_{hD}|}\right),D^{-\eta}\left((1-\alpha)\sigma_{hD}\right)\right)\,\mathrm{d}t+S_{t_1t_2}\\
&\quad=S_A+S_{pen}+S_{t_1t_2}\le K(\ddot{\alpha},\ddot{\bar{\sigma}}_s,\ddot{\bar{\sigma}}_h).
\end{aligned}$$

Then

$$\begin{aligned}
S_A=&\ \frac{1}{2}\left(A_sD^{-\eta}\left(\alpha\sigma_s\right),D^{-\eta}\left(\alpha\sigma_s\right)\right)\Big|_{t_1}^{t_2}\\
&+\frac{1}{2}\left(A_hD^{-\eta}\left((1-\alpha)\sigma_h\right),D^{-\eta}\left((1-\alpha)\sigma_h\right)\right)\Big|_{t_1}^{t_2} \qquad (113)
\end{aligned}$$

and we take the limit $\eta\to 0$.

Since $\dot{\sigma}_s,\dot{\sigma}_h\in L^2(L^2)$, this limit $\eta\to +0$ exists a.e. with respect to t_1, t_2 and we obtain

$$
\begin{aligned}
&\frac{1}{2}\ \left(A_s\frac{\partial}{\partial t}\left(\alpha\sigma_s\right),\frac{\partial}{\partial t}\left(\alpha\sigma_s\right)\right)\Bigg|_{t_1}^{t_2}\\
&\quad+\frac{1}{2}\left(A_h\frac{\partial}{\partial t}\left((1-\alpha)\sigma_h\right),\frac{\partial}{\partial t}\left((1-\alpha)\sigma_h\right)\right)\Bigg|_{t_1}^{t_2}\\
&\quad+\int_{t_1}^{t_2}\mu^{-1}\left(\frac{\partial}{\partial t}\left([|\sigma_{sD}|-\kappa_s]_+\frac{\sigma_{sD}}{|\sigma_{sD}|}\right),\frac{\partial}{\partial t}\left(\alpha\sigma_{sD}\right)\right)\,\mathrm{d}t \qquad (114)\\
&\quad+\int_{t_1}^{t_2}\mu^{-1}\left(\frac{\partial}{\partial t}\left([|\sigma_{hD}|-\kappa_h]_+\frac{\sigma_{hD}}{|\sigma_{hD}|}\right),\frac{\partial}{\partial t}\left((1-\alpha)\sigma_{hD}\right)\right)\,\mathrm{d}t\\
&\quad+\lim_{\eta\to 0}S_{t_1t_2}=A_1+Pen_{soft}+Pen_{hard}+\lim_{\eta\to 0}S_{t_1t_2}\le K(\ddot{\alpha},\ddot{\tilde{\sigma}}_s,\ddot{\tilde{\sigma}}_h).
\end{aligned}
$$

We write

$$
\begin{aligned}
&\frac{\partial}{\partial t}\left([|\sigma_{sD}|-\kappa_s]_+\frac{\sigma_{sD}}{|\sigma_{sD}|}\right):\frac{\partial}{\partial t}\left(\alpha\sigma_{sD}\right) \qquad (115)\\
&=[|\sigma_{sD}|-\kappa_s]_+\alpha|\sigma_{sD}|^{-1}\left(\left|\frac{\partial}{\partial t}\left(\sigma_{sD}\right)\right|^2-\left|\frac{\partial}{\partial t}\,|\sigma_{sD}|\right|^2\right)+\left|\frac{\partial}{\partial t}[|\sigma_{sD}|-\kappa_s]_+\right|^2\alpha\\
&\qquad+\frac{\partial}{\partial t}[|\sigma_{sD}|-\kappa_s]_+\alpha\frac{\partial}{\partial t}\kappa_s+\frac{\partial}{\partial t}[|\sigma_{sD}|-\kappa_s]_+|\sigma_{sD}|\dot{\alpha}\\
&=P_1+P_2+P_3+P_4.
\end{aligned}
$$

We have $P_1\ge 0$, $P_2\ge 0$ and these terms contribute to the final estimate. For P_3 we find via partial integration

$$
\begin{aligned}
\mu^{-1}\int_{t_1}^{t_2}\int_{\Omega}P_3\,\mathrm{d}x\,\mathrm{d}t=&-\mu^{-1}\int_{t_1}^{t_2}\int_{\Omega}[|\sigma_{sD}|-\kappa_s]_+\frac{\partial}{\partial t}\left(\alpha\frac{\partial}{\partial t}\kappa_s\right)\,\mathrm{d}x\,\mathrm{d}t\\
&+\mu^{-1}\int_{\Omega}[|\sigma_{sD}|-\kappa_s]_+\alpha\frac{\partial}{\partial t}\kappa_s\,\mathrm{d}x\Bigg|_{t_1}^{t_2}=\tilde{P}_{31}+\tilde{P}_{32}
\end{aligned} \qquad (116)
$$

The term $\tilde{P}_{31}$ is uniformly bounded due to the $L^1(L^1)$ bound for the penalty term and the hypotheses on α and κ_s. The term $\tilde{P}_{32}$ is estimated via the $L^\infty(L^2)$ lemma for the penalty term. This yields

$$
\left|\mu^{-1}\int_{\Omega}\tilde{P}_{32}\mathrm{d}x\Bigg|_{t_1}^{t_2}\right|\le K(\dot{\kappa}_s)\left(\|\dot{\sigma}_s(t_2,.)\|_{L^2(\Omega)}+\|\dot{\sigma}_h(t_2,.)\|_{L^2(\Omega)}\right)+o(t_1) \qquad (117)
$$

where $o(t_1)$ needs not be uniform in μ.

The term P_4 is treated in a similar manner like P_3. Thus we obtain

$$\begin{aligned} Pen_{soft} &\geq \mu^{-1}\int_{t_1}^{t_2}\int_{\Omega}\left|\frac{\partial}{\partial t}\,|[|\sigma_{sD}|-\kappa_s]_+|\right|^2 \alpha\,\mathrm{d}x\,\mathrm{d}t \\ &= K - \|\dot{\sigma}_s(t_2,.)\|_{L^2(\Omega)} - \|\dot{\sigma}_h(t_2,.)\|_{L^2(\Omega)} - o(t_1) \end{aligned} \tag{118}$$

and a similar inequality for the hard material.

It remains to analyze the boundary terms coming from the partial summation

$$\lim_{\eta\to 0} S_{t_1t_2} = S^1_{t_1t_2} + S^2_{t_1t_2}. \tag{119}$$

From the result of the previous partial summation we obtain

$$\begin{aligned} \lim_{\eta\to 0} S_{t_1t_2} =& -\frac{1}{2}\left(A_s\frac{\partial}{\partial t}\,(\alpha\sigma_s)\,,\,\frac{\partial}{\partial t}\,(\alpha\sigma_s)\right)\Bigg|_{t_1}^{t_2} \\ & -\frac{1}{2}\left(A_h\frac{\partial}{\partial t}\,((1-\alpha)\sigma_h)\,,\,\frac{\partial}{\partial t}\,((1-\alpha)\sigma_h)\right)\Bigg|_{t_1}^{t_2} \\ & -\mu^{-1}\left(\frac{\partial}{\partial t}\left([|\sigma_{sD}|-\kappa_s]_+\frac{\sigma_{sD}}{|\sigma_{sD}|}\right),\,\frac{\partial}{\partial t}\,(\alpha\sigma_{sD})\right)\Bigg|_{t_1}^{t_2} \\ & -\mu^{-1}\left(\frac{\partial}{\partial t}\left([|\sigma_{hD}|-\kappa_h]_+\frac{\sigma_{hD}}{|\sigma_{hD}|}\right),\,\frac{\partial}{\partial t}\,((1-\alpha)\sigma_{hD})\right)\Bigg|_{t_1}^{t_2}. \end{aligned} \tag{120}$$

We have

$$\begin{aligned} S^1_{t_1t_2} &= \left(A_s\tfrac{\partial}{\partial t}\,(\alpha\sigma_s)\,,\,\tfrac{\partial}{\partial t}\,(\alpha\sigma_s)\right)\Big|_{t_1}^{t_2} \\ & \quad + \textit{corresponding term for hard material.} \end{aligned} \tag{121}$$

$$\begin{aligned} S^2_{t_1t_2} &= \mu^{-1}\left([|\sigma_{sD}|-\kappa_s]_+\tfrac{\sigma_{sD}}{|\sigma_{sD}|},\,\tfrac{\partial}{\partial t}\,(\alpha\sigma_s)\right)\Big|_{t_1}^{t_2} \\ & \quad + \textit{corresponding term for hard material.} \end{aligned} \tag{122}$$

Let

$$\begin{aligned} S_t =& \left(A_s\frac{\partial}{\partial t}\,(\alpha\sigma_s)\,,\,\frac{\partial}{\partial t}\,(\alpha\sigma_s)\right)\Bigg|_t + \mu^{-1}\left([|\sigma_{sD}|-\kappa_s]_+\frac{\sigma_{sD}}{|\sigma_{sD}|},\,\frac{\partial}{\partial t}\,(\alpha\sigma_s)\right)\Bigg|_t \\ & + \textit{corresponding term for hard material.} \end{aligned} \tag{123}$$

With this notation $\lim_{\eta\to 0} S_{t_1t_2} = S_{t_1} - S_{t_2}$. Now we present an argument which shows that $\lim_{\eta\to 0} S_t$ remains unchanged if we replace the right hand factors

$\frac{\partial}{\partial t}\left(\alpha\sigma_s\right)$ and $\frac{\partial}{\partial t}\left((1-\alpha)\sigma_h\right)$ in the scalar products by $\frac{\partial}{\partial t}\left(\alpha\tilde{\sigma}_s\right)$ and $\frac{\partial}{\partial t}\left((1-\alpha)\tilde{\sigma}_h\right)$. In fact, this follows if we use the test functions

$$\begin{gathered}\delta_1^{-1}\left(\sigma_s-\tilde{\sigma}_s-\alpha^{-1}S^{-\delta_1}\left(\alpha\sigma_s-\alpha\tilde{\sigma}_s\right)\right),\\ \delta_1^{-1}\left(\sigma_h-\tilde{\sigma}_h-(1-\alpha)^{-1}S^{-\delta_1}\left((1-\alpha)(\sigma_h-\tilde{\sigma}_h)\right)\right)\end{gathered} \tag{124}$$

and pass to the limit in the penalty equation $\delta_1 \to 0$, performing this procedure at t_1 and t_2.

This gives us the equation

$$\left(A_s\frac{\partial}{\partial t}\left(\alpha\sigma_s\right),\frac{\partial}{\partial t}\left(\alpha\sigma_s-\alpha\tilde{\sigma}_s\right)\right)+\mu^{-1}\left([|\sigma_{sD}|-\kappa_s]_+\frac{\sigma_{sD}}{|\sigma_{sD}|},\frac{\partial}{\partial t}\left(\alpha\sigma_s\right)-\alpha\tilde{\sigma}_s\right)$$
$$+\ \textit{a similar term for the hard material}=0 \tag{125}$$

for all $t=t_1$ and a.e. t_2.

We conclude that

$$\begin{aligned}\left|S^1_{t_1t_2}+S^2_{t_1t_2}\right| \le{}& K\left(\left\|\frac{\partial}{\partial t}\left(\alpha\sigma_s\right)\right\|_{L^2(\Omega)}\left\|\frac{\partial}{\partial t}\left(\alpha\tilde{\sigma}_s\right)\right\|_{L^2(\Omega)}\right)\Bigg|_{t_2} \qquad (126)\\ &+K\left(\left\|\mu^{-1}[|\sigma_{sD}|-\kappa_s]_+\right\|_{L^1(L^1)}\left\|\frac{\partial}{\partial t}\left(\alpha\tilde{\sigma}_s\right)\right\|_{L^\infty(L^\infty)}\right)\Bigg|_{t_2}\\ &+\ \textit{a similar term}\big|_{t_1}\\ &+\ \textit{the related summand for hard material at } t_1 \textit{ and } t_2 \textit{ a.e.}\end{aligned}$$

Thus, we see that

$$\begin{aligned}\left|S^1_{t_1t_2}+S^2_{t_1t_2}\right| \le{}& K(\dot{\alpha},\dot{\tilde{\sigma}}_s,\dot{\tilde{\sigma}}_h)(1+\|\sigma_s\|_{L^2(\Omega)}\,|_{t_2}+\|\sigma_s\|_{L^2(\Omega)}\,|_{t_1}\\ &+\|\sigma_h\|_{L^2(\Omega)}\,|_{t_2}+\|\sigma_h\|_{L^2(\Omega)}\,|_{t_1}).\end{aligned} \tag{127}$$

Collecting our results we obtain from (114)

$$\left|\frac{1}{2}\left(A_s\frac{\partial}{\partial t}\left(\alpha\sigma_s\right),\frac{\partial}{\partial t}\left(\alpha\sigma_s\right)\right)\Bigg|_{t_1}^{t_2}+\frac{1}{2}\left(A_h\frac{\partial}{\partial t}\left((1-\alpha)\sigma_h\right),\frac{\partial}{\partial t}\left((1-\alpha)\sigma_h\right)\right)\Bigg|_{t_1}^{t_2}\right|$$
$$\le K\left(\int_\Omega|\dot{\sigma}_s|^2\,\mathrm{d}x\Bigg|_{t_2}+\int_\Omega|\dot{\sigma}_h|^2\,\mathrm{d}x\Bigg|_{t_2}\right)^{\frac{1}{2}} \tag{128}$$

$$+K\left(\int_\Omega |\dot{\sigma}_s|^2\,\mathrm{d}x\bigg|_{t_1} + \int_\Omega |\dot{\sigma}_h|^2\,\mathrm{d}x\bigg|_{t_1}\right)^{\frac{1}{2}}$$

$$+K(\ddot{\alpha}, \ddot{\kappa}_s, \ddot{\kappa}_h, \ddot{\tilde{\sigma}}_s, \ddot{\tilde{\sigma}}_h).$$

We finally get rid of the terms

$$\left(A_s \frac{\partial}{\partial t}(\alpha\sigma_s), \frac{\partial}{\partial t}(\alpha\sigma_s)\right)\bigg|_{t_1}, \qquad \int_\Omega |\dot{\sigma}_s|^2\,\mathrm{d}x\bigg|_{t_1}. \tag{129}$$

In fact, from the penalty equation, with the above reasoning, we obtain

$$\begin{aligned}&\left(A_s \frac{\partial}{\partial t}(\alpha\sigma_s), \frac{\partial}{\partial t}(\alpha\sigma_s - \alpha\tilde{\sigma}_s)\right)\bigg|_{t_1}\\ &\quad + \mu^{-1}\left([|\sigma_{sD}| - \kappa_s]_+ \frac{\sigma_{sD}}{|\sigma_{sD}|}, \frac{\partial}{\partial t}(\alpha\sigma_s - \alpha\tilde{\sigma}_s)\right)\bigg|_{t_1}\\ &\quad + \textit{corresponding term with hard material} = 0.\end{aligned} \tag{130}$$

Now, since, for fixed μ,

$$[|\sigma_{sD}| - \kappa_s]_+ \frac{\sigma_{sD}}{|\sigma_{sD}|} \to 0 \quad \text{in} \quad L^2(L^2) \qquad \text{as } t_1 \to 0, \tag{131}$$

(similarly for the hard material) we conclude that

$$\begin{aligned}&\lim_{t_1\to 0}\left(A_s \frac{\partial}{\partial t}(\alpha\sigma_s), \frac{\partial}{\partial t}(\alpha\sigma_s)\right) + \textit{corresponding term for hard material}\\ &\le \operatorname*{\textit{ess sup}}_{0\le t\le\delta}\left\{\left(A_s \frac{\partial}{\partial t}(\alpha\tilde{\sigma}_s), \frac{\partial}{\partial t}(\alpha\tilde{\sigma}_s)\right) + \textit{corresponding term for hard material}\right\}.\end{aligned} \tag{132}$$

The theorem now follows from (128) and (132). □

Corollary 3.

$$\begin{aligned}&\left\|\mu^{-1}[|\sigma_{s\mu D}| - \kappa_s]_+(|\sigma_{s\mu D}| + 1)\right\|_{L^\infty(L^1)} +\\ &\left\|\mu^{-1}[|\sigma_{h\mu D}| - \kappa_h]_+(|\sigma_{h\mu D}| + 1)\right\|_{L^\infty(L^1)} \le C_0\end{aligned} \tag{133}$$

uniformly as $\mu \to +0$.

8 Passage to the Limit as the Penalty Parameter μ Tends to Zero

Theorem 12. *Under the assumption of the main theorem the solutions* $(\sigma_{s\mu}, \sigma_{h\mu})$ *of the penalty problem converge to the solution* (σ_s, σ_h) *of the variational inequality* (19) *for the Prandtl-Reuss mixture. The convergence is strong in* $L^\infty(L^2)$ *and weak in* $H^1(L^2)$.

Proof. Since $\sigma_{s\mu}, \sigma_{h\mu}, \dot{\sigma}_{s\mu}, \dot{\sigma}_{h\mu}$ are uniformly bounded in $L^2(L^2)$ as $\mu \to +0$ we may subtract a subsequence $\Lambda = \{\mu_m | \mu_m \to +0\}$ such that $\sigma_{s\mu} \rightharpoonup \sigma_s$, $\sigma_{h\mu} \rightharpoonup \sigma_h$, $\dot{\sigma}_{s\mu} \rightharpoonup \dot{\sigma}_s$, $\dot{\sigma}_{h\mu} \rightharpoonup \dot{\sigma}_h$ weakly in $L^2(L^2)$.

We may pass to the limit in the equation of balance of linear momentum and obtain (11) for σ_s, σ_h. Furthermore, the symmetry of σ_s and σ_h is preserved. From the $L^1(L^1)$-estimate (see the corollary to Theorem 9 in Sect. 6) we have that the penalty term is bounded in $L^1(L^1)$ as $\mu \to +0$. This implies $[|\sigma_{s\mu D}| - \kappa_s]_+^2 \leq K\mu$, $[|\sigma_{h\mu D}| - \kappa_h]_+^2 \leq K\mu$ and, since $[|\xi| - \kappa_s]_+^2$ is convex and continuous, we obtain $[|\sigma_{sD}| - \kappa_s]_+^2 \leq 0$, $[|\sigma_{hD}| - \kappa_h]_+^2 \leq 0$ i.e. $|\sigma_{sD}| \leq \kappa_s$, $|\sigma_{hD}| \leq \kappa_h$. □

The variational inequality (19) follows from the penalty equations

$$\begin{aligned}
&\int_{t_1}^{t_2} \left(A_s \frac{\partial}{\partial t} \left(\alpha\sigma_{s\mu}\right), \alpha(\sigma_{s\mu} - \hat{\sigma}_s) \right) \\
&\qquad + \left(A_h \frac{\partial}{\partial t} \left((1-\alpha)\sigma_{h\mu}\right), (1-\alpha)(\sigma_{h\mu} - \hat{\sigma}_h) \right) \mathrm{d}t \\
&\leq -\mu^{-1} \int_{t_1}^{t_2} \left([|\sigma_{s\mu D}| - \kappa_s]_+ \frac{\sigma_{s\mu D}}{|\sigma_{s\mu D}|}, \alpha(\sigma_{s\mu} - \hat{\sigma}_s) \right) \\
&\qquad + \left([|\sigma_{h\mu D}| - \kappa_h]_+ \frac{\sigma_{h\mu D}}{|\sigma_{h\mu D}|}, (1-\alpha)(\sigma_{h\mu} - \hat{\sigma}_h) \right) \mathrm{d}t \\
&\leq 0 \qquad \text{for all } (\hat{\sigma}_s, \hat{\sigma}_h) \in \mathbb{K}.
\end{aligned} \tag{134}$$

The last step concerning that the left hand side is ≤0 follows from the monotonicity property of

$$[|\tau| - \kappa]_+ \frac{\tau_D}{|\tau_D|} \tag{135}$$

and the fact that $[|\hat{\sigma}_s| - \kappa_s]_+ = 0$, $[|\hat{\sigma}_h| - \kappa_h]_+ = 0$ by definition of $\mathbb{K}$. We may pass to the weak limit $\mu \to 0$ in (134), keeping the inequality ≤0 due to lower semicontinuity. This yields (19). The strong convergence

$$\sigma_{s\mu} \to \sigma_s \qquad \sigma_{h\mu} \to \sigma_h \qquad \text{in } L^2(L^2) \tag{136}$$

follows by setting $\hat{\sigma}_s = \sigma_s$, $\hat{\sigma}_h = \sigma_h$ in (134) and adding the terms

$$-\left(A_s\frac{\partial}{\partial t}\left(\alpha\sigma_s\right),\alpha(\sigma_{s\mu}-\sigma_s)\right)$$
$$-\left(A_h\frac{\partial}{\partial t}\left((1-\alpha)\sigma_h\right),(1-\alpha)(\sigma_{h\mu}-\sigma_h)\right)\to 0 \qquad (137)$$

In fact, this yields (for $t_1 = 0$)

$$\limsup_{\mu\to+0}\big(A_s\alpha(\sigma_{s\mu}-\sigma_s),\alpha(\sigma_{s\mu}-\sigma_s)\big)$$
$$+\big(A_h(1-\alpha)(\sigma_{h\mu}-\sigma_h),(1-\alpha)(\sigma_{h\mu}-\sigma_h)\big)\leq 0 \quad (138)$$

which even implies

$$\sigma_{s\mu}\to\sigma_s,\qquad \sigma_{h\mu}\to\sigma_h,\qquad \text{in } L^\infty(L^2). \qquad (139)$$

Since the solution σ_s, σ_h is unique, (136) holds for the full sequence (via the usual contradiction argument). The convergence (139) for the full sequence can be derived with an additional simple $C(L^2)$ argument.

We now want to incorporate the partial strain velocities and the plastic strain velocities into the discussion. Similar, as to the classical Prandtl-Reuss problem, the situation is not quite satisfactory due to the fact that only L^1-estimates are available. Under the assumptions of Theorem (9) (in particular, no assumptions on $\ddot{\alpha}, \ddot{\kappa}_s, \ddot{\kappa}_h$) we have uniform $L^1(L^1)$-bounds for

$$\frac{1}{2}(\nabla\dot{u}_{s\mu}+\nabla\dot{u}_{s\mu}{}^T),\qquad \frac{1}{2}(\nabla\dot{u}_{h\mu}+\nabla\dot{u}_{h\mu}{}^T), \qquad (140)$$

and the corresponding penalty terms.

With Temam's imbedding theorem this implies a uniform $L^1(L^{\frac{n}{n-1}})$-bound for $\dot{u}_{s\mu}$, $\dot{u}_{h\mu}$ and we obtain that, for a subsequence,

$$\dot{u}_{s\mu}\rightharpoonup\dot{u}_s,\qquad \dot{u}_{h\mu}\rightharpoonup\dot{u}_h \qquad (141)$$

and

$$\frac{1}{2}(\nabla\dot{u}_{s\mu}+\nabla\dot{u}_{s\mu}{}^T)\to\frac{1}{2}(\nabla\dot{u}_s+\nabla\dot{u}_s{}^T) \qquad (142)$$

$$\frac{1}{2}(\nabla\dot{u}_{h\mu}+\nabla\dot{u}_{h\mu}{}^T)\to\frac{1}{2}(\nabla\dot{u}_h+\nabla\dot{u}_h{}^T) \qquad (143)$$

weakly in $C^*([0,T]\times\bar{\Omega})$, $\mu\to+0$. This means that the strain velocities need not be functions, they are only Riesz measures.

In case that an $L^\infty(L^1)$-bound is available for the penalty terms, see Theorem 11 , $\dot{u}_{s\mu}$ and $\dot{u}_{h\mu}$ are bounded in $L^\infty(L^{\frac{n}{n-1}})$, the convergence in (141) takes place in $L^{\frac{n}{n-1}}(L^{\frac{n}{n-1}})$, and the limiting deformation velocities are (at least) $L^2(L^{\frac{n}{n-1}})$-functions. We want to derive a variational inequality which takes the strain velocity into account.

From (33) and (34) we conclude

$$\left(A_s\frac{\partial}{\partial t}\left(\alpha\sigma_{s\mu}\right),\sigma_{s\mu}-\tau\right)=\left(\frac{1}{2}(\nabla\dot{u}_{s\mu}+\nabla\dot{u}_{s\mu}{}^T),\sigma_{s\mu}-\tau\right)\le 0 \tag{144}$$

for all $\tau\in C(\bar{\Omega})$ such that $\tau=\tau^T$ and $|\tau_D|\le\kappa_s$, a.e. (no balance of linear momentum for τ is assumed).

The $\le$ inequality in (144) follows from the monotonicity property of the penalty term and the fact that $[|\tau_D|-\kappa_s]_+=0$. An inequality similar to (144) holds for the hard material.

We want to pass to the limit $\mu\to+0$ in (144). For the left hand side this is possible due to weak and strong $L^2(L^2)$ convergence of the functions in the scalar product. For the left hand side, obviously

$$\frac{1}{2}\left(\nabla\dot{u}_{s\mu}+\nabla\dot{u}_{s\mu}{}^T,\tau\right)\to\frac{1}{2}(\nabla\dot{u}_s+\nabla\dot{u}_\mu{}^T,\tau) \tag{145}$$

but for $\frac{1}{2}\left(\nabla\dot{u}_{s\mu}+\nabla\dot{u}_{s\mu}{}^T,\sigma_{s\mu}\right)$ this convergence is not clear, since we do not know that $\sigma_{s\mu}\to\sigma_s$ in $C(\bar{\Omega})$. Thus we assume the hypothesis of Theorem 10 and we use the $L^\infty(L^{\frac{n}{n-1}})$ for $\dot{u}_{s\mu}$, $\dot{u}_{h\mu}$ and write

$$\frac{1}{2}\left(\nabla\dot{u}_{s\mu}+\nabla\dot{u}_{s\mu}{}^T,\sigma_{s\mu}\right)=(\dot{u}_{s\mu},f_s)+\int_{\partial\Omega}p\sigma_{s\mu}\,\mathrm{d}o \tag{146}$$

due to the balance of linear momentum.

In the right hand side we may pass to the limit and obtain as limit

$$(u_s,f)+\int_{\partial\Omega}p\sigma_s\,\mathrm{d}o. \tag{147}$$

So we arrive at the variational inequality

$$\left(A_s\frac{\partial}{\partial t}\left(\alpha\sigma_{s\mu}\right),\sigma_s-\tau\right)\le(f,u_s)+\int_{\partial\Omega}p\sigma_s\,\mathrm{d}o-\frac{1}{2}(\nabla u_s+\nabla u_s{}^T,\tau) \tag{148}$$

for all $\tau\in C(\bar{\Omega},M^n_{sym})$, $|\tau_D|\le\kappa_s$ and a similar inequality for the hard material. Of course, this is not satisfactory.

In the case of the Hencky problem, in two dimension there is an interesting way to overcome the formulation (148). An analogue approach might work also for the Prandtl Reuss problem. For Hencky's problem, and similar for Prandtl-Reuss's problem, via a technique using a reverse Hölder inequality, there are $L^{\frac{n}{n-1}+\delta}$ and $L^\infty(L^{\frac{n}{n-1}+\delta})$-estimates available for the displacements u (see [6]) or the displacement velocities $\dot{u}$, respectively. Thus we have a $(\frac{n}{n-1}-\delta')$ capacity potential ϕ of the set where the $(\frac{n}{n-1}-\delta')$ capacity is small and, testing the penalty equation with $\sigma_s\phi$, we obtain uniform smallness of

$$\mu^{-1}\int_\Omega [|\sigma_{\mu D}|-\kappa_s]_+|\sigma_{\mu D}|\phi\,\mathrm{d}x \tag{149}$$

on sets of small capacity. Hence the limiting Riesz measure shares this property.

Since interior H^1-estimates for σ are available and thus σ_μ is uniformly continuous except on a set of small $2-\delta'$ capacity we may give a meaning to $\left(\frac{1}{2}(\nabla\dot{u}+\nabla\dot{u}^T),\zeta\right)$, $\zeta\in C_0^\infty(\Omega)$ by extending the measure to functions which are $2-\delta'$ quasicontinuous in the sense of capacities.

We do not state the above discussion concerning capacity methods since, for a rigorous discussion, this would take more space than available here. We confine to fix the statement concerning convergence in C^* of the strain velocities.

Theorem 13. *Assume the hypotheses of the main theorem. Then the partial strain velocities constructed in Sect. 3 converge weakly in C^**

$$\begin{aligned}\frac{1}{2}(\nabla\dot{u}_{s\mu}+\nabla\dot{u}_{s\mu}{}^T) &\rightharpoonup \frac{1}{2}(\nabla\dot{u}_s+\nabla\dot{u}_s{}^T)\\ \frac{1}{2}(\nabla\dot{u}_{h\mu}+\nabla\dot{u}_{h\mu}{}^T) &\rightharpoonup \frac{1}{2}(\nabla\dot{u}_h+\nabla\dot{u}_h{}^T)\\ \textit{and}\,\dot{u}_{s\mu}\rightharpoonup\dot{u}_s\quad,\quad \dot{u}_{h\mu}&\rightharpoonup\dot{u}_h\,\textit{weakly in}\,C^*.\end{aligned} \tag{150}$$

If, in addition, the assumption of Theorem 10 are satisfied (150) *($\mu\to+0$) holds in $L^{\frac{n}{n-1}}(L^{\frac{n}{n-1}})$.*

Remark 9. (150) holds strongly in $L^{\frac{n}{n-1}-\delta'}(L^{\frac{n}{n-1}-\delta'})$ due to Temam's imbedding theorem, and in fact in $L^{\frac{n}{n-1}+\delta}(L^{\frac{n}{n-1}+\delta})$, δ small, if the reverse Hölder inequality technique is used (which we did not do here).

Corollary 4. *If the assumptions of Theorem 10 are satisfied, the variational inequality* (148) *holds.*

Concerning the Kuhn Tucker rule there are similar problems like the interpretation of the 'pointwise' inequality (144). The penalty terms converge weakly in C^* as $\mu\to+0$. We have

$$\begin{aligned}\mu^{-1}[|\sigma_{s\mu D}|-\kappa_s]_+\frac{\sigma_{s\mu D}}{|\sigma_{s\mu D}|}&\rightharpoonup\dot{e}_{ps},\\ \mu^{-1}[|\sigma_{h\mu D}|-\kappa_h]_+\frac{\sigma_{h\mu D}}{|\sigma_{h\mu D}|}&\rightharpoonup\dot{e}_{hs}\end{aligned} \tag{151}$$

and one would like to conclude from

$$\left.\begin{array}{l}\mu^{-1}[|\sigma_{s\mu D}| - \kappa_s]_+ \to \lambda_s \\ \mu^{-1}[|\sigma_{h\mu D}| - \kappa_h]_+ \to \lambda_h\end{array}\right\} \text{*-weakly in } C \tag{152}$$

that

$$\dot{e}_{ps} = \lambda_s \frac{\sigma_{sD}}{|\sigma_{sD}|}, \qquad \dot{e}_{ph} = \lambda_h \frac{\sigma_{hD}}{|\sigma_{hD}|}. \tag{153}$$

In the case of the two dimensional Hencky-Model this can be proved by the tools indicated above, using the reverse Hölder inequality for the displacements, smallness of the support of λ_s and λ_h on sets of small $(2+\delta')$ capacity and the fact that $\sigma_{s\mu}, \sigma_{h\mu}$ converges $(2+\delta')$ uniformly for a subsequence.

9 A Model for the Volume Fraction α and the Yield Parameter Depending on the History of the Rate of the Plastic Strain of the Soft Material

In [10], the following model for α, κ_h, κ_s is suggested. Let

$$l(t,x) = \int_0^t |\dot{e}_{ps}(\xi,x)|\, \mathrm{d}\xi \tag{154}$$

where $\dot{e}_{ps}$ is the plastic deformation velocity of the soft material. Then the volume fraction α is defined by

$$\alpha(t,.) = \alpha_0 + (1-\alpha_0)e^{-c_0 l(t,.)} \tag{155}$$

and the yield parameters by

$$\kappa_s = const > 0, \qquad \kappa_h = \kappa_0 + r_0 l \tag{156}$$

with constants $\alpha_0 > 0$, $c_0 > 0$, $\kappa_0 > 0$, $r_0 > 0$. In the rigorous setting $\dot{e}_{ps}(s,x)$ may be a Riesz measure and we have to approach it via the penalty approximations, see below. Since l could be unbounded we apply a simple (from our point of view acceptable) modification by setting

$$l(t,x) = \int_0^t g(|\dot{e}_{ps}(s,x)|)\, \mathrm{d}s + \delta_0, \qquad \delta_0 > 0 \tag{157}$$

with a bounded, non negative function $g \in C^1$.

This also guarantees our condition that $\alpha \geq \delta_1 > 0$ and $(1-\alpha) \leq 1-\delta_1$, and the condition

$$\|\dot{\alpha}\|_{L^\infty(L^\infty)} + \|\dot{\kappa}_s\|_{L^\infty(L^\infty)} + \|\dot{\kappa}_h\|_{L^\infty(L^\infty)} \leq K \tag{158}$$

whatever function $\dot{e}_{ps}$ is chosen.

On the level of the penalty approximation, we have

$$\dot{e}_{ps} = \mu^{-1}[|\sigma_{sD}| - \kappa_s]_+ \frac{\sigma_{sD}}{|\sigma_{sD}|} \text{ and} \tag{159}$$

$$|\dot{e}_{ps}| = \mu^{-1}[|\sigma_{sD}| - \kappa_s]_+ \tag{160}$$

$$\alpha = \alpha_0 + (1-\alpha_0)\exp\left(-c_0 \int_0^t g(\mu^{-1}[|\sigma_{sD}| - \kappa_s]_+)\, \mathrm{d}\xi\right) \tag{161}$$

$$\kappa_h = \kappa_0 + r_0 \int_0^t g(\mu^{-1}[|\sigma_{sD}| - \kappa_s]_+)\, \mathrm{d}\xi. \tag{162}$$

With this definition of α, κ_h, κ_s (which is constant) we assign to every pair (σ_s, σ_h) satisfying the symmetry condition a solution $\bar{\sigma}_{\mu s}$, $\bar{\sigma}_{\mu h}$ to the penalty equation (27), and all the a priori estimates of this paper requiring the L^∞ property on $\dot{\alpha}$, $\dot{\kappa}_h$ (not $\ddot{\alpha}$, $\ddot{\kappa}_h$) are true.

For applying Schauder's fix point theorem to obtain a solution $\bar{\sigma}_{\mu s} = \sigma_{\mu s}, \bar{\sigma}_{\mu h} = \sigma_{\mu h}$ one needs an additional compactness condition in space direction which would be achieved by a non local dependence of α and κ_h in terms of σ_s, say

$$l(t,x) = \int_0^t g(\mu^{-1}[|\int_\Omega K(x,y)\sigma_{sD}(t,y)\, \mathrm{d}y| - \kappa_s]_+)\, \mathrm{d}\xi \tag{163}$$

with a compact singular integral operator $K : L^2 \to L^2$. With this compact dependence it is possible to solve the penalty equation, due to the a priori estimates given in Sects. 4 and 5, and also to prove the convergence of the penalty equation as $\mu \to 0$, which leads to a solution of a quasivariational inequality with the above interpretation.

Acknowledgements Jens Frehse work was supported by SFB611 at University of Bonn, Josef Málek work was supported by the project GAČR 107/12/0121.

References

1. Bensoussan, A., Frehse, J.: Asymptotic behaviour of the time dependent Norton-Hoff law in plasticity theory and H^1 regularity. Commentationes Mathematicae Universitatis Carolinae **37**(2), 285–304 (1996)
2. Blum, H., Frehse, J.: Boundary differentiability for the solution to Hencky's law of elastic plastic plane stress. Preprint 435, SFB 611 University of Bonn (2008)
3. Bulíček, M., Frehse, J., Málek, J.: On boundary regularity for the stress in problems of linearized elasto-plasticity. Int. J. Adv. Eng. Sci. Appl. Math. **1**(4), 141–156 (2009)
4. Demyanov, A.: Regularity of stresses in Prandtl-Reuss perfect plasticity. Calc. Var. Partial Differ. Equ. **34**(1), 23–72 (2009)
5. Duvaut, G., Lions, J.: Inequalities in Mechanics and Physics. Springer, Berlin (1976)

6. Hardt, R., Kinderlehrer, D.: Elastic plastic deformation. Appl. Math. Optim. **10**(1), 203–246, (1983)
7. Hlaváček, I., Haslinger, J., Nečas, J., Lovíšek, J.: Solution of Variational Inequalities in Mechanics. Springer, New York (1988)
8. Khasina, L.: Mathematische Behandlung von Mischungen Elastoplastischer Substanzen. Ph.D. thesis, University of Bonn (2007)
9. Kratochvíl, J.: A theory of non-proportional cyclic plasticity based on micromechanical approach. In: Tokuda, M., Xu, B., Senoo, M. (eds.) Macro/Micro/Meso Mechanical Properties of Materials, Proceedings of IMMM'93, 3–5 Aug 1993, Mie University. Mie Academic Press, Japan (1993)
10. Kratochvíl, J., Málek, J., Rajagopal, K., Srinivasa, A.: Modeling of the response of elastic plastic materials treated as a mixture of hard and soft regions. Zeitschrift für Angewandte Mathematik und Physik (ZAMP) **55**(3), 500–518 (2004)
11. Temam, R.: Mathematical Problems in Plasticity, vol. 15. Gauthier-Villars, Paris (1985)
12. Temam, R.: A generalized Norton-Hoff model and the Prandtl-Reuss law of plasticity. Arch. Ration. Mech. Anal. **95**, 137–183 (1986)

Modeling and Simulation of Lipid Monolayers as Surfactant in Lung Alveoli

Annelene Wittenfeld, Andrey Ryskin, and Wolfgang Alt

1 Biological Function of Surfactant Lipids for the Breathing Cycle

The concave cellular surface inside lung alveoli is covered by a thin water film with surfactant on top, whose continued preservation is essential for the rapid oxygen and carbon dioxide exchange between air and lung tissue. Therefore, in order to guarantee film stability during the regular breathing cycle of inhalation (expansion of lung alveoli) and exhalation (their compression), the surfactant lipid-protein layer must possess strong adaptive properties. This becomes even more important in events of sudden expansion or compression of the lung, for example, while coughing. On one hand, the function of the surfactant layer is to strongly reduce surface tension for minimizing the work of breathing, which is achieved by forming highly ordered monolayers of amphiphilic lipid molecules. On the other hand, the layer has to be fluid enough to cover the expanding surface in a continuous and rapid fashion, which requires stochastic mobility of the lipid molecules.

Synthetic chemical models of efficient pulmonary surfactant, also used for medical treatment, contain specific mixtures of phospholipids (e.g. phosphatidylcholine, DPPC, palmitic acid, PA) and have been experimentally investigated under periodic compression-relaxation conditions [3]. A typical result can be seen in Fig. 1, which shows the quasistatic behavior as an isotherm plot (surface pressure versus inverse density). During the first two compression cycles the pressure-density curve shows an obvious hysteresis effect. The overall pressure decreases more rapidly at the beginning of the expansion periods (to the left), compared to a more steady increase during the compression periods (to the right). This is a strong indication

A. Wittenfeld (✉) · A. Ryskin · W. Alt
Abteilung Theoretische Biologie, Institut für Zelluläre und Molekulare Botanik, Rheinische Friedrich-Wilhelms-Universität Bonn, Kirschallee 1-3, D-53115 Bonn, Germany
e-mail: wolf.alt@uni-bonn.de; a.wittenfeld@gmx.de

M. Griebel (ed.), *Singular Phenomena and Scaling in Mathematical Models*,
DOI 10.1007/978-3-319-00786-1_8,

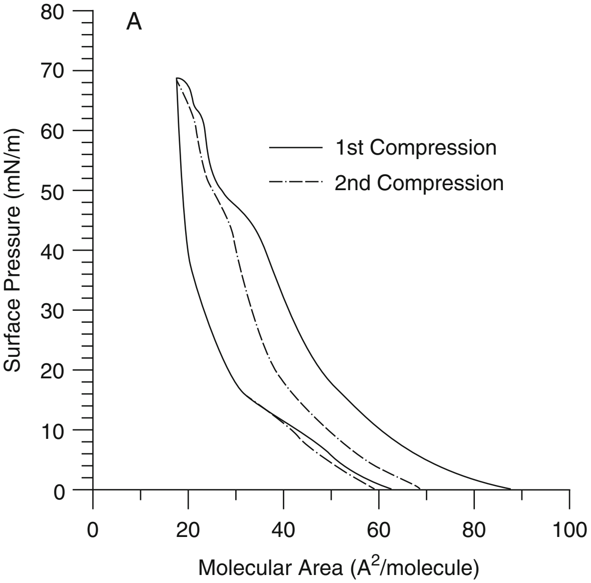


Fig. 1 Quasistatic isotherm plot of surface pressure versus area per lipid molecule (inverse density) for an artificial lipid mixture as a model of surfactant layer on lung alveoli. The two compression-expansion trajectories are cycling counter-clockwise. (From [3], Fig. 1A)

for transformation processes, such as association and dissociation, to occur in the different periods and phases.

Moreover, atomic force microscopy of such surfactant monolayers has revealed remarkable inhomogeneities in height and stiffness (see [1], Fig. 1). One observes larger and mostly rounded patches of thick homogeneous molecular layers, indicating regions of strongly ordered lipid molecules (liquid condensed phase) and, in-between, smaller and more fuzzy patches of thinner layers, indicating regions of non-ordered and diffusing lipid molecules (liquid expanded phase). The interface between these two kinds of patches appears to be quite sharp, suggesting a relatively fast transition zone between the ordered and disordered phase.

These spatial phase separation phenomena together with the observed hysteresis dynamics have important biophysical functions: regions of more fluid disordered monolayers serve as a mechanical buffer of more easily compressible lipid molecules; in addition, they serve as a chemical reservoir of diffusing lipids to be supplied for insertion into the ordered regions of the monolayer. However, in order to describe and understand the full dynamics of these biophysical/chemical processes, one needs detailed mathematical models and their subsequent numerical simulation on various spatio-temporal scales. For a 3-d molecular dynamics simulation see [4], for a general 3-d continuum fluid-diffusion simulation including thin film dynamics on a periodically extending alveolar surface see [5].

Here we present recent modeling and simulation results for two different scaling approaches, showing certain connections with each other. The first is a stochastic multiparticle model of rod-like lipid molecules on a microscopic scale (space: $\sim$ 10–500 nm; time: $\sim$msec) with appropriate classical interaction potentials (Sect. 2). The second describes a deterministic continuum two-phase fluid model on a mesoscopic scale (space: $\gtrsim$ µm; time: $\gtrsim$ sec) with an appropriately chosen free energy function and flow properties (Sects. 3–5).

2 Microscale Stochastic Multi-particle Model

We study a system of N stiff rod-like particles of fixed length L, which are partially submersed in water at a sharp and flat water-air interface. In order to model the amphiphilic properties of lipids, these rods are assumed to have a hydrophilic head for length variable $0 < s < L/4$, and a hydrophobic tail for $L/4 < s < L$ (see Fig. 2). The rods interact by means of a continuous, distance-dependent force density $\mathbf{f}(s)\,ds$ on each rod, accounting for the amphiphilic properties of the rods as well as their location and motion with respect to the water-air interface. For simplicity, we restrict our simulation model to a two-dimensional spatial configuration, so that the rods are not allowed to overlap with each other.

We denote by $\delta = \delta(r, s)$ the distance of two points r, s on distinct rods $i \neq j$. The interactive force between head of rod i $(0 < r < L/4)$ and tail of rod j $(L/4 < s < L)$ is purely repulsive, inversely proportional to the distance δ with strength coefficient α. For head-head and tail-tail interaction we use a Lennard-Jones potential, valid for various kinds of particles and molecules,

$$V_{\varepsilon,\sigma}(\delta) = 4\varepsilon\left(\left(\frac{\sigma}{\delta}\right)^{12} - \left(\frac{\sigma}{\delta}\right)^{6}\right) \tag{1}$$

with different scaling coefficients $\varepsilon > 0$ and $\sigma > 0$.

Since the upper part of each rod is hydrophobic and the lower hydrophilic, there is a kind of capillary force pulling each subpart into its optimal medium by tending to reduce the length of the amphiphilic mismatch g, see Fig. 2.

Here we assume a simple proportionality

$$F_{\text{wat}} = k_s\, g, \tag{2}$$

with effective elasticity coefficient k_s. This water surface force acts onto the center of the amphiphilic mismatch region in the ‘wrong’ medium, pointing towards the ‘right’ medium in direction parallel to the rod. Since a rod has a ‘hydrophilic’ and a ‘hydrophobic’ part, we need to ensure that it does not flip over. Therefore, as angular momentum, we assume an analogous non-linear term, modeled as to be induced by the surface tension acting on both sides of the rod:

$$M_{\text{wat}} = -k_\theta \left(\tan\bigl(\theta - 0.5\pi\bigr) + \tan^3\bigl(\theta - 0.5\pi\bigr)\right). \tag{3}$$

This momentum tends to turn the rod into an upright position ($\theta = 0.5\pi$). The first summand acts as a linear spring in case of small angle differences.

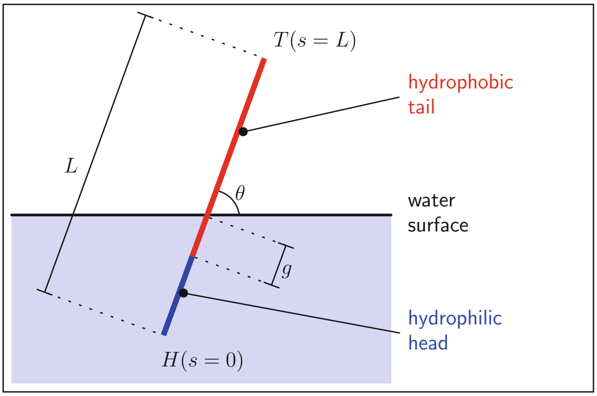


Fig. 2 Schematic picture of an amphiphilic rod and its location at a fixed water-air surface

The second summand represents a strong repulsive spring turning away from any parallel orientation with the water surface ($\theta = 0$ or π) in case of larger angle differences.

Additionally, we model local collision of rods with the surrounding smaller water and air molecules by means of a spatio-temporal stochastic Brownian sheet. On each segment of the discretized rod, we assume small and independent random forces, with amplitude adjusted via a parameter c_w representing the standard deviation of the corresponding Gaussian noise perpendicular to the rod. The perturbation along the rod is twice as large. Additionally, the process depends on the surrounding medium. If the segment of the rod is outside the water, the perturbation strength is one tenth as compared to a segment within the water. Accordingly, the stochastic force density is defined by

$$\mathrm{d}\,\mathbf{f}_{\text{brown}}(s,t) = \mathbb{D}(s)\,\mathrm{d}\mathbf{B}_{s,t} \tag{4}$$

with a corresponding amplitude matrix $\mathbb{D}(s)$. Finally, friction is considered to be proportional to the local velocity of a point on the rod,

$$\mathbf{f}_{\text{friction}}(s) = -\mathbb{F}(s)\mathbf{v}(s). \tag{5}$$

In diagonal representation, i.e. when the coordinate system is aligned with the rod, the friction matrix reads $\mathbb{F} = \operatorname{diag}(\gamma_{\|}, \gamma_{\perp})$. The imposed drag perpendicular to the rod $\propto \gamma_{\perp}$ is twice as large as compared to the friction along the rod $\propto \gamma_{\|}$. Additionally, there are different γ-values for each surrounding medium, namely γ^{A} in air and γ^{L} in liquid, since the viscosity of the fluid is larger than the viscosity of air.

In this model, we treat the rod-like particles as rigid bodies. From the local force densities above, we calculate the effective center-of-mass force and the torque acting on each rod. The resulting Newtonian equations of motion are employed in their over-damped limit with vanishing inertia.

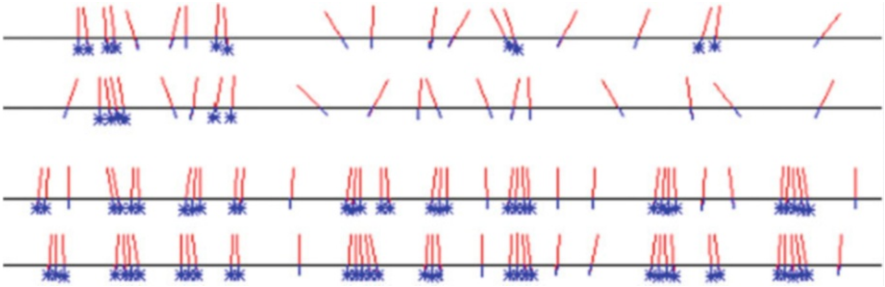


Fig. 3 Snapshots of simulation runs with N = 20 (*first* and *second row*) and N = 40 rods (*third* and *fourth row*), performed on a small interval of length $X = 900\,\text{nm}$, thus with densities $\rho = 0.022\,\text{nm}^{-1}$ and $\rho = 0.044\,\text{nm}^{-1}$, respectively. The *second* and the *fourth row* are about 300–400 ms later than the *first* and the *third*, respectively. Clustered rods are marked by a *star*

The simulation program is conceived in a highly optimized manner, including certain coding in assembler, on several cores at a time. The simulations are performed with different numbers N of rods with fixed length $L = 40$ (in an artificial length unit, comparable to nanometer, nm) initially distributed in random positions at the flat (periodic) waterline (of width $X = 22220\,\text{nm}$) and with randomly chosen angles, $0 < \theta_i < \pi$, but without intersections of the rods. Simulation runs are performed for 20,000,000 timesteps (in an artificial time unit, comparable to microseconds, μsec), thus for about 20 s, see Fig. 3 for typical lipid configurations appearing in two different simulations (with shorter width). After every 1,000 timesteps (thus each millisecond, msec) certain observables are extracted for data evaluation.

As a first evaluation, we quantify proximity of the simulated rods. Two rods i and j are defined to be clustered, if

$$|\theta_i - \theta_j| \leq \beta_{\max} \quad \text{and} \quad |w_i - w_j| \leq b_{\max}. \tag{6}$$

If a rod does not obey these relations with any other, it is said to be unclustered. Here, w_i denotes the horizontal position, at which the rod i crosses the waterline. The positive parameters $\beta_{\max}$ and $b_{\max}$ are suitably chosen thresholds, namely $\beta_{\max} = 10°$, $b_{\max} = 15$. As an observable we define the overall clustering

$$C = \frac{\text{number of clustered rods}}{\text{total number of rods } N} \tag{7}$$

of the simulated rods for every time step. Another observable is the polar order parameter

$$R = \frac{1}{N}\Big|\sum_i \widehat{e_i}\Big|, \tag{8}$$

where $\widehat{e_i} = (\cos\theta_i, \sin\theta_i)$ represents the unit vector along the i-th rod.

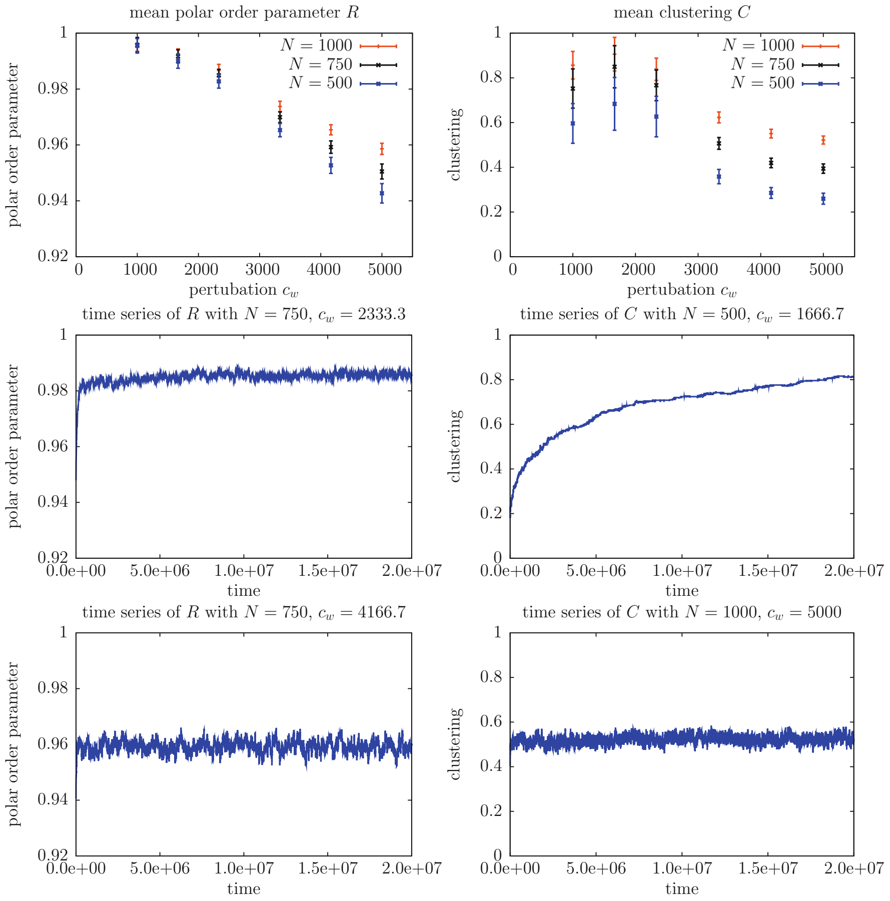


Fig. 4 In the plots above we show the mean polar order parameter R and clustering C depending on the perturbation c_w, for different particle numbers N. The plots below show examples of corresponding time series. Note that the essential parameters c_w and N vary, as separately indicated for each plot. The other simulation parameters are the same for all plots shown. The friction is $\gamma_{\|}^{L} = 300$, $\gamma_{\perp}^{L} = 600$ inside of the liquid and $\gamma_{\|}^{A} = 30$, $\gamma_{\perp}^{A} = 60$ outside. The remaining interactive force parameters are $\sigma_{tt} = 5$, $\varepsilon_{tt} = 100, \sigma_{hh} = 7.5$, $\varepsilon_{hh} = 20$, $\alpha = 1$ and $k_s = k_\theta = 25$

Time behavior of the observables (7) and (8) is presented in the lower part of Fig. 4 for a typical set of simulation parameters. For the polar order index R (plots in the left column), the overall behavior is almost the same for each simulation. Starting from a ‘random’ initial value ∼0.94, within less than 100 ms the index reaches a stationary value, about which it fluctuates, though with a slight adjusting increase in case of the smaller perturbation amplitude $c_w = 2333.3$. In contrast, the clustering index C (plots in the right column) shows a quite different behavior. For $c_w \leq 2333.3$ and every N, the index slowly approaches its potential stationary level with an exponential rate of time scales more than 5 s. This also explains the large variances and the decreasing mean in the corresponding parameter plot (top right). For $c_w \geq 3300$, the stationary level of C is reached almost instantaneously, meaning that the 'clustering rate', i.e. the speed of cluster growth, is strongly dependent on the lipid mobility coefficient c_w. Further simulations could reveal a more detailed dependence.

The polar order index R exhibits a clear monotone dependence on c_w (top left plot in Fig. 4). Moreover, in the two parameter plots it can be seen that a higher density of rods leads to a higher value of both indices, clustering and polar order. The explanation for this feature is obvious: since the rods are generally closer at higher densities, their short-range interactions give rise to enhanced mutual alignment of adjacent rods. Thus, the rate of transition from 'disordered' rods (freely diffusing) to 'ordered' rods (mutually aligned in clusters) depends both on mobility and density of the modeled lipid molecules, amphiphilically embedded into the water surface.

3 Mesoscale Continuum Mixture Model

As shown by the stochastic microscopic model in the previous section, the temporal scale, on which clustering and order indices experience slow changes, lies in the range of 1–5 s. Also, sizes of ordered lipid clusters in this one-dimensional simulation model are in the range of 20–500 nm, which is up to ten times the rod length $L = 40$ nm, whereas the mean distance of lipids in clusters is about 10 nm. Thus, when locally averaging the density of lipids, ρ, and the mean fraction of ordered lipids, φ, on the larger spatial scale $\gtrsim$ μm, these quantities $\rho(t,x)$ and $\varphi(t,x)$ could be considered as continuously varying over the time scale of seconds.

Since the clusters of ordered lipids have quite sharp boundaries in the order of rod length $L = 40$ nm, this spatial transition zone from the disordered to the ordered phase scales with a factor $\varepsilon \lessapprox 0.05$ compared to the μm length scale. Thus, the continuum phase field model defined and treated in [1], seems to be in accordance with the 'empirical' findings from the microscopic model, since the free energy of the continuous lipid monolayer system is defined by

$$f(\rho,\varphi) = f^0(\rho,\varphi) + \frac{\varepsilon^2}{2}|\nabla\varphi|^2. \tag{9}$$

Here $f^0(\rho,\varphi)$ is a suitable 'interpolation' between free energies in the single phases with densities $\rho_\alpha = \rho\,\theta_\alpha$ and volume fractions $\theta_1 = (1-\varphi)$, $\theta_2 = \varphi$, for $0 \leq \varphi \leq 1$. They are chosen as canonical Gibbs energies

$$f_\alpha(\rho) = b_\alpha^0 \rho \left(\log(\rho/\rho_\alpha^0) - 1\right) + c_\alpha \tag{10}$$

with positive model parameters. The free energy f^0 can then generally (even for more than two phases) be written as

$$f^0(\rho, \theta) = \sum_\alpha \chi(\theta_\alpha) \left[f_\alpha(\rho) + b_\alpha^0 \rho \, \Lambda \left(\eta_\alpha (1 - \theta_\alpha) \right) \right] \tag{11}$$

with an interpolating function $\chi : [0, 1] \to [0, 1]$ satisfying $\chi'(0) = \chi''(0) = 0$, $\chi'(1) = \chi''(1) = 0$, and with the monotone function $\Lambda(z) = z + (1 - z) \log(1 - z) :$ $[0, 1] \to [0, 1]$ satisfying $\Lambda'(0) = 0$.

In addition to the χ-interpolation, for each phase α there appears an excess energy term $\eta_\alpha(\vartheta_\alpha)$ evaluated at the complementary volume fraction $\vartheta_\alpha = (1 - \theta_\alpha)$. Here $\eta_\alpha(\vartheta_\alpha)$ represents an increasing 'reaction' function with $\eta_\alpha(0) = 0$, describing the amount of phase associations from other phases towards the α-component. The more 'other' molecules are locally available, the larger is the probability of transition to the α-component. Since 'binding sites' are limited as in usual chemical Michaelis-Menten kinetics, we suppose a corresponding transition kinetics of Monod type:

$$\eta_\alpha(\vartheta) = \frac{s_\alpha \vartheta}{1 + s_\alpha \vartheta / \kappa_\alpha} \, . \tag{12}$$

Here $s_\alpha > 0$ denote the transition strengths and $0 < \kappa_\alpha < 1$ the asymptotic saturation level, so that always $\eta_\alpha(\vartheta) < 1$. In the simulations below we choose $s_1 = 9$ and $s_2 = 4$, meaning that association to ordered lipids is about half as easy as their dissociation.

For a general thermodynamically correct theory of continuum mixture dynamics with corresponding equations for mass and force balances, we refer to the contribution by H.W. Alt and W. Alt in this volume [2]. For example, the total pressure $p^0 = \rho f'^0_\rho - f^0$ of the model above can be computed as

$$p^0(\rho, \theta) = \sum_\alpha \chi(\theta_\alpha) \; (1 + \eta_\alpha(1 - \theta_\alpha)) \; b_\alpha^0 \rho \, . \tag{13}$$

To characterize the special case of a two-phase ordered-disordered lipid monolayer system with total density $\rho(t, x)$ and volume fraction of ordered lipids $\varphi(t, x) :=$ $\theta_2(t, x)$, several particular assumptions have to be implemented.

- There is only one mean transport velocity for both phases, $\mathbf{V}$. We notice that according to the microscopic simulation model (in Sect. 2) all lipids are embedded into the water surface with part of their hydrophilic head region. Then the assumed mean velocity $\mathbf{V}$ can be interpreted as a bulk velocity of the lipid heads together with their surrounding water molecules near the surface. The differences $\mathbf{u}_\alpha = \mathbf{v}_\alpha - \mathbf{V}$ for the two phases are due to possible diffusive flow, which might be derived along the lines indicated in [2], Sects. 6–11.

This would be compatible with the following assumptions that have been made in [1]:

- In addition to transport velocity the mass balance equation for ρ contains a diffusive flux $-d_\varepsilon \nabla f'_\rho$ with inverse diffusion coefficient being interpolated as follows:

$$\frac{1}{d_\varepsilon} = \varphi \frac{1}{\varepsilon d_0} + (1-\varphi)\frac{1}{d_0}. \tag{14}$$

Here ε is the small positive constant defined above, meaning that the ordered lipids diffuse much less than the disordered ones.

- The surfactant monolayer as a fluid-like system on top of the water surface can be regarded as a viscous fluid, where the viscosity coefficients α, β are those of the water surface layer (α_0 and β_0) but increased by lipid shear forces. These are stronger for ordered lipids than for disordered ones. Therefore we define $\zeta = \{\alpha, \beta\}$ as $\zeta = \zeta_0 + (1-\varphi)\,\zeta_1 + \varphi\,\zeta_2$ with suitable positive constants.

Then two mass balance equations are complemented by one quasi-stationary force balance equation in the limit of low Reynolds number (high viscosity), namely

$$\partial_t \rho + \operatorname{div}(\rho \mathbf{V} - d_\varepsilon \nabla f{\prime}^0_\rho) = 0, \tag{15}$$

$$\partial_t \varphi + \mathbf{V}\nabla\varphi + f{\prime}^0_\varphi - \varepsilon^2 \Delta\varphi = 0, \tag{16}$$

$$\partial_j (p^0\,\delta_{ij} - \alpha(\partial_i V_j + \partial_j V_i) - \beta \partial_k V_k \delta_{ij}) = 0. \tag{17}$$

This is the (quasi-stationary version of the) system that has been investigated in [1]. In the following sections we report on some further results that have been obtained by modifying or extending the numerical methods used so far.

4 Sharp Phase Transition Approximation

In the limit of small $\varepsilon \to 0$ we obtain a fast transition layer of width ε, in which the phase transition equation (16) for φ and the mass conservation (15) for ρ in their quasi-steady-state limit for $\varepsilon = 0$ constitute a 3-dimensional ODE system of first order. For given 'density jump' levels $\rho_1 > 0$ and $\rho_2 > 0$ its solutions have to connect the two outer asymptotic states ($\varphi = 0, \rho = \rho_1$) and ($\varphi = 1, \rho = \rho_2$). Moreover, the ODE system contains a free parameter λ representing the difference between the interface velocity and the normal component of $\mathbf{V}$ at the interface. For more details see [1], Sects. 5 and 6.

The solution of this sharp transition problem is equivalent to finding a value λ such that a reduced 2-dimensional ODE system for the phase transition derivative $\psi(\varphi)$ and the density $\rho(\varphi)$ as functions of the phase variable $\varphi \in [0, 1]$, namely

$$\frac{d\psi}{d\varphi} = \frac{B(\rho,\varphi)}{\psi} - \lambda \tag{18}$$

$$\frac{d\rho}{d\varphi} = -\frac{C(\rho,\varphi)}{A(\rho,\varphi)} + \lambda \frac{\varphi(\rho_2 - \rho)}{d_0 A(\rho,\varphi)\psi}\,, \tag{19}$$

has a (unique) solution satisfying $\psi(\varphi) > 0$ for $0 < \varphi < 1$ and the following boundary conditions

$$\psi(0) = \psi(1) = 0 \tag{20}$$

$$\rho(0) = \rho_1, \rho(1) = \rho_2 \tag{21}$$

for given density levels ρ_1 and ρ_2. Definitions of the parameter functions in (18) and (19) can be found in [1], Eq. (7.3). When integrating Eqs. (18) and (19), the boundary conditions above impose two compatibility conditions. One is for the free parameter

$$\lambda = \int_0^1 \frac{B(\rho(\varphi), \varphi)}{\psi(\varphi)} d\varphi\,, \tag{22}$$

representing the relative speed of the sharp transition layer. The other is a condition for the ρ-level difference

$$\rho_2 - \rho_1 = \int_0^1 \frac{1}{A(\rho(\varphi), \varphi)} \left[\lambda \frac{\varphi\,(\rho_2 - \rho(\varphi))}{d_0 \psi(\varphi)} - C(\rho(\varphi), \varphi) \right] d\varphi\,. \tag{23}$$

Below we will show that in cases of positive λ, meaning a 'dissolving' boundary of the ordered domain, this condition determines an additional free integration constant c. However, for negative λ, meaning a 'growing' boundary of the ordered domain, it imposes a compatibility relation between ρ_2 and ρ_1.

Since the boundaries of the unit interval are degenerate points of the ODE system (18) and (19), one has to construct the asymptotics for $\rho(\varphi)$ and $\psi(\varphi)$ in the limit $\varphi \to 0$ and 1, for more details see [1], Sect. 7. In particular, ρ has the following expansion near $\varphi = 1$:

$$\rho(\varphi) = \rho_2 - c(1-\varphi)^\kappa - \frac{\gamma}{2-\kappa}(1-\varphi)^2 + O((1-\varphi)^3) \tag{24}$$

with $\kappa \sim \lambda$ and the integration constant c determined by condition (23). However, it follows that for the case $\lambda < 0$, and consequently $\kappa < 0$, the boundary condition $\rho(1) = \rho_2$ can only be satisfied with $c = 0$.

We now solve the given boundary value problem by the following numerical shooting method, which is slightly different from the method applied in [1] and generally more stable. First, using the asymptotic expressions, particularly (24), we determine the values of the functions ψ and ρ near particular points $\varphi = \delta$ and $1 - \delta$, with a positive distance $\delta \ll 1$. Taking these values as initial data for the ODE system (18) and (19), we solve this system numerically from both sides till the point $\varphi = 1/2$. Thus we get two values for each function. For example, for the function $\psi(\varphi)$, we get $\psi(1/2)_-$ and $\psi(1/2)_+$, where subscripts $-$ and $+$ indicate from which side we have started the integration process.

4.1 Case of Dissolving Ordered Monolayer: $\lambda > 0$

We construct the error functions

$$F_1(\lambda, c, \rho_1, \rho_2) = \psi(1/2)_- - \psi(1/2)_+ \tag{25}$$

$$F_2(\lambda, c, \rho_1, \rho_2) = \rho(1/2)_- - \rho(1/2)_+ \tag{26}$$

where c is the constant of integration in (24). For any given pair of levels $\rho_1 > 0$ and $\rho_2 > 0$ both values λ and c are taken as free parameters. Using a standard recurrent iteration method with suitably chosen relaxation constants, the values converge to a unique pair $\{\hat{\lambda}, \hat{c}\}$ satisfying the zero-error condition

$$F_1(\hat{\lambda}, \hat{c}, \rho_1, \rho_2) = F_2(\hat{\lambda}, \hat{c}, \rho_1, \rho_2) = 0\,. \tag{27}$$

Then we have solved the boundary value problem (18)–(21) and found the corresponding wave speed value $\lambda = \hat{\lambda}$, as long as $\hat{\lambda} \geq 0$. If this fails, we have to switch to the next case.

4.2 Case of Growing Ordered Monolayer: $\lambda < 0$

For the case $\lambda < 0$ the parameter c has to be zero and we cannot solve system (27) for any pair of $\{\rho_1, \rho_2\}$, since we have only one free parameter λ for two equations. In this case a solution is possible only for a special curve in the $\{\rho_1, \rho_2\}$ plane, which then satisfies condition (23). Thus, we perform the same iteration procedure in order to reduce the error functions F_i for $c = 0$ and fixed ρ_1 but with varying free parameters (λ, ρ_2). Then, convergence towards values $\hat{\lambda}$ and $\widehat{\rho_2} = \widehat{\rho_2}(\rho_1)$ satisfying

$$F_1(\hat{\lambda}, 0, \rho_1, \widehat{\rho_2}) = F_2(\hat{\lambda}, 0, \rho_1, \widehat{\rho_2}) = 0 \tag{28}$$

yields a solution with $\hat{\lambda} < 0$.

4.3 Simulation Results

With a chosen parameter set (the same as in [1]), extensive numerical computations reveal that only for values $\rho_1 \lesssim 0.45$ and $\rho_1 \gtrsim 7$ we find $\lambda > 0$, but in-between λ is negative. Figure 5 depicts the points in the (ρ_1, ρ_2) plane allowing for a solution of the transition problem (18–21). Two grey areas on the right and on the left are shown, where solutions exist with positive wave speed λ. In-between the curve $\rho_2 = \widehat{\rho_2}(\rho_1)$ is plotted on which λ is negative. At the endpoints of this curve we have two point singularities representing particular states with $\lambda = 0$.

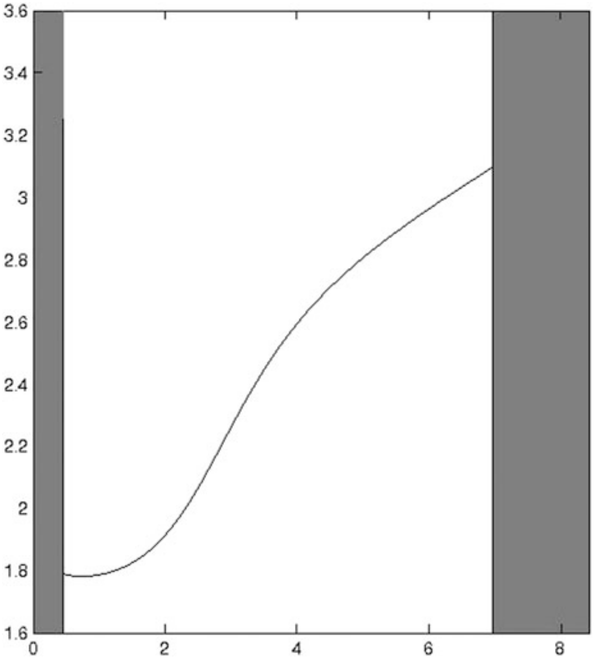


Fig. 5 Diagram of possible lipid phase transitions projected to the (ρ_1, ρ_2) plane of asymptotic boundary values. *Dark grey* areas correspond to positive λ values. In-between the curve $\rho_2 = \widehat{\rho_2}(\rho_1)$ is plotted, on which λ is negative. The two end points of this curve are isolated singularities (ρ_1^*, ρ_2^*)

In [1], Sect. 8, the full two-phase lipid dynamics with sharp phase transition was numerically investigated for the one-dimensional case of a fixed bounded interval [0, X], where no-flux boundary conditions were imposed for the two outer densities $\rho_1(t, x)$, for $x < s(t)$, and $\rho_2(t, x)$, for $x > s(t)$, respectively. Here $x = s(t)$ defines the moving sharp transition interface. If solutions start with an initial sharp transition not too near to the boundaries, they generally converge to the unique steady state with the singular transition levels $\rho_1 = \rho_1^* \approx 0.45$ and $\rho_2 = \rho_2^* = \widehat{\rho_2}(\rho_1^*) \approx 1.79$, representing the singularity in Fig. 5 with lower ρ_1^* value. For initial data to the left of the singularity, with $\rho_1|_{t=0} < \rho_1^*$, both $\lambda(t)$ and $\dot{s}(t)$ become positive, and the (ρ_1, ρ_2) trajectories approach the projected stable manifold of the singular steady state, see [1], Fig. 14. For initial data to the right of the singularity, with $\rho_1|_{t=0} > \rho_1^*$ and $\rho_2|_{t=0} = \widehat{\rho_2}(\rho_1|_{t=0})$, both $\lambda(t)$ and $\dot{s}(t)$ are negative and the (ρ_1, ρ_2) trajectories stay on the singular curve (in Fig. 5), constituting the other side of the projected stable manifold, on which they asymptotically approach the same singular steady state.

However, the so far unsolved question is, what happens with the sharp transition approximation, if the transition levels hit the 'border line', namely $\rho_1(t) = \rho_1^*$, while $\rho_2(t) \neq \rho_2^*$. In Fig. 6a we plot the numerically computed transition profiles of the density function $\rho(\varphi)$ for different boundary values $\rho_1 < \rho_1^*$ and

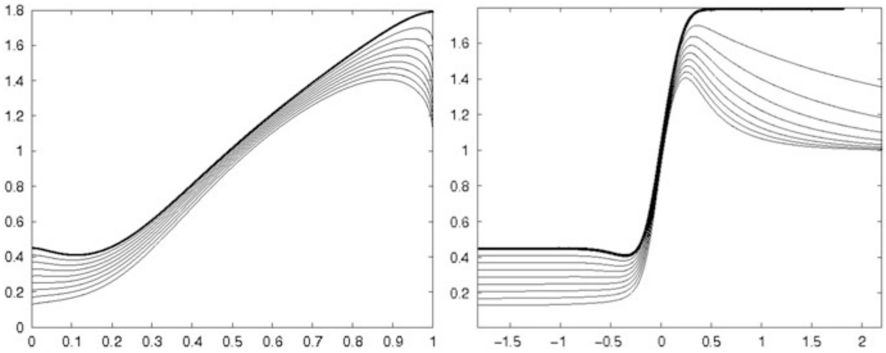


Fig. 6 (**a**) *Left*: Functions $\rho(\varphi)$ for different boundary values $\rho_1 < \rho_1^*$ and for fixed $\rho_2 = 1$. When approaching the critical value $\rho_1 = \rho_1^*$ monotonically from below, the profiles also monotonically increase and approach the thick graph $\rho^*(\varphi)$. Their maximal values $\rho_{max} = \max \rho(\varphi) = \rho(\varphi^*)$ converge to $\rho_2^* > 1$, however, φ^* converges to the right hand boundary $\varphi = 1$. (**b**) *Right*: The profiles $\rho(\xi)$ for different boundary values $\rho(-\infty) = \rho_1 < \rho_1^*$ and for fixed $\rho(\infty) = \rho_2 = 1$. When monotonically approaching the critical value $\rho_1 = \rho_1^*$, the profiles also monotonically increase and approach the thick graph $\rho^*(\xi)$ but with decaying rate of exponential convergence $\rho(\xi) \to \rho_2$ for $\rho_1 \to \rho_1^*$

fixed $\rho_2 = 1 < \rho_2^*$. As we can see, when $\rho_1 \to \rho_1^*$, the profiles $\rho(\varphi)$ approach an asymptotic profile with $\lambda = 0$, namely $\rho^*(\varphi)$, which is plotted as a bold graph. This is the unique singular transition profile connecting ρ_1^* with ρ_2^*, so that the right hand boundary value is strictly larger than the prescribed value $\rho_2 = 1$ of all the other ρ profiles. This means that, the closer we come to the boundary value $\rho_1 = \rho_1^*$, the thinner becomes the boundary layer in the neighborhood of $\varphi \sim 1$.

If we plot the density profiles of the transition layer over the physical space coordinate ξ, we obtain the graphs in Fig. 6b. It is obvious that the decreasing boundary layer near $\varphi = 1$, described in Fig. 6a, corresponds to a decreasing exponential rate of convergence $\rho(\xi) \to 1$ for $\xi \to \infty$. Thus, when approaching the boundary of the left $\{\lambda > 0\}$ region in Fig. 5, the wave speed λ converges to zero, but the convergence towards the singular $\rho^*(\xi)$ profile is non-uniform in space, except when we approach the singularity (ρ_1^*, ρ_2^*). We conclude that the approximate equations (18) and (19) for a thin transition layer loose its validity near the $\{\lambda = 0\}$ borders away from the singular points.

5 One-Dimensional Simulation of the Continuum Phase Field Model

For exploring the quantitative solution behavior around the singular point and for simulating the effects of cyclic breathing, we rely on the original phase field continuum equations (15)–(17) for positive ε with no-flux boundary conditions on a

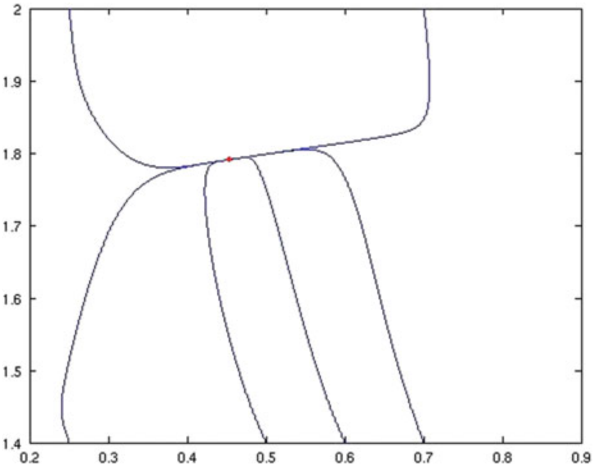


Fig. 7 Trajectories of the (ρ_1, ρ_2) values in simulations of the phase field equations (15)–(17), for six different initial conditions, converging to the asymptotically stable singularity $(\tilde{\rho}_1^*, \tilde{\rho}_2^*)$ (*red dot*)

given interval $[0, X]$. We perform numerical simulations for the qualitatively same situation as in Sect. 4.3, namely a phase separation of the interval into a left hand part with unordered and a right hand part with ordered lipids.

5.1 Smoothed Phase Separation in a Fixed Interval

On the unit interval $\{0 < x < 1\}$ we choose the initial distribution of the phase field φ to be zero in $0 < x < s_0$ and one in $s_0 < x < 1$ for some transition point $0 < s_0 < 1$. Accordingly, for the density ρ we take step distributions with different plateau values $\rho(0) = \rho_1$ and $\rho(1) = \rho_2$. Since the partial differential equations (15)–(17) with positive $\varepsilon(= 0.05)$ constitute a well-posed parabolic-elliptic system with regularity properties, also when discretized by using a standard explicit scheme with finite elements, the initial step data are smoothed instantaneously.

In all simulation runs we observe evolution to the same stationary state $\tilde{\rho}$ with boundary values $\tilde{\rho}_1^* \approx 1.79$ and $\tilde{\rho}_2^* \approx 0.47$, both quite close to the coordinates of the sharp transition singular point (ρ_1^*, ρ_2^*), see the trajectories in Fig. 7. The plots in Fig. 8a show the density profiles of a specific trajectory at different times. One can easily see the convergence towards a stationary profile, which is approximately the same as the asymptotic profile in Fig. 6b. Starting point of the phase transition interface was $s_0 = 0.35$. Depending on the initial data, the position of the interface $s(t)$ moves backward or forward (the latter is to be seen in Fig. 8b) and converges to an asymptotic value (to be seen in Fig. 9a). Simultaneously, the relative speed parameter $\lambda(t)$ converges to zero (see Fig. 9b).

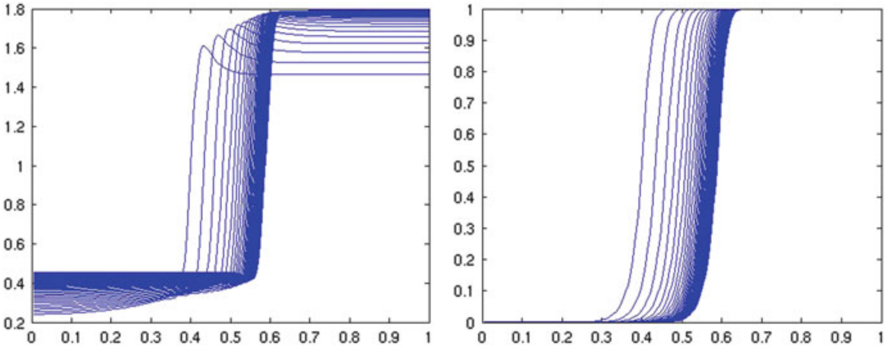


Fig. 8 (**a**) *Left*: time evolution of the density profiles $\rho(t, x)$ for one specific initial condition $(\rho_1, \rho_2) = (0.25, 1.4)$ with $\lambda(t) > 0$. (**b**) *Right*: the corresponding phase transition functions $\varphi(t, x)$

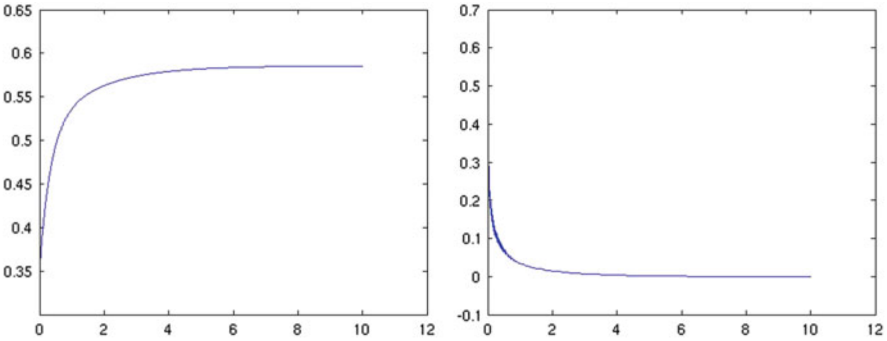


Fig. 9 Simulation as in Fig. 8. (**a**) *Left*: the 'interface' position $s(t)$, where $\varphi(t, s(t)) = 0.5$. (**b**) *Right*: the function $\lambda(t) := \dot{s}(t) - V(t, s(t))$

5.2 Phase Separation Behavior in an Oscillating Domain Simulating "Breathing"

In order to mimic the alveoli expansion and compression during breathing, we make our domain $\{0 < x < X(t)\}$ to change periodically its size, for instance in a sinusoidal manner: $X(t) = X_0(1 + a \sin \omega t)$ with $X_0 = 1$ (comparable to μm), breathing amplitude $a = 0.5$ and breathing frequency $\omega = 2\pi$ (comparable to 1.26/s). As can be observed from the density profile plots of Fig. 10a, it is the 'exhalation' period during which the interval size $X(t)$ shrinks and thereby both densities of disordered and ordered lipids are lifted up while, after a certain time lag, the transition interface starts to move left and leads to a quite fast growth of the ordered lipid monolayer. Then the reverse process can be observed during the

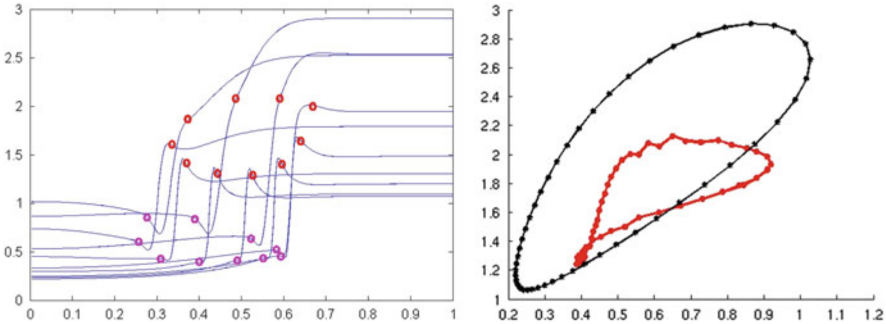


Fig. 10 'Breathing' cycle simulation of the phase field model with $\varepsilon = 0.05$. (**a**) *Left*: time evolution of the profile $\rho(t, \tilde{x})$ over the rescaled variable $\tilde{x} = x/X(t)$ on the fixed unit interval. The $\tilde{\rho}_2(t)$-values just right of the transition layer together with their position values $\tilde{x}_2(t)$ describe a counter-clockwise cycle (*red circles*). A similar but smaller cycle is described by the $\tilde{\rho}_1(t)$-values just left of the transition layer together with their position values $\tilde{x}_1(t)$ (*magenta circles*). (**b**) *Right*: corresponding clockwise trajectories of the level pairs (ρ_1, ρ_2) at the interval boundaries (*black curve*) and $(\tilde{\rho}_1, \tilde{\rho}_2)$ near the transition layer (*red curve*). Notice that the simultaneous decrease of both values during 'inhalation' is much faster than the increase during 'exhalation'

'inhalation' period, when due to the growth of interval size $X(t)$ both ρ_1- and ρ_2-levels rapidly decrease while, again after a time lag, the transition interface moves to the right and the ordered lipid monolayer shrinks.

Due to strong diffusion in the disordered phase, inducing long slopes in the density profile, the left boundary values $\rho_1(t)$ oscillate much stronger than the $\tilde{\rho}_1(t)$-values directly left of the transition zone; the same is true for the corresponding $\rho_2(t)$- and $\tilde{\rho}_2(t)$-levels of the ordered phase to the right, compare the 'circle dots' in Fig. 10a and the corresponding trajectories plotted in Fig. 10b. The inner hysteresis loop (red curve) is not only smaller but also strongly deformed compared to the outer, more ellipsoid hysteresis loop (black curve). This is mainly due to the strong difference $\rho_2(t) - \tilde{\rho}_2(t)$ during the 'exhalation' period, when ρ_2 still increases, while $\tilde{\rho}_2$ already begins to decrease and the transition front moves to the left so rapidly, that there appears a larger secondary transition zone (of size up to 0.25) between the sharp transition layer and the ρ_2-plateau. Simultaneously, also the relative speed $\lambda(t)$ becomes negative, so that the $(\tilde{\rho}_1, \tilde{\rho}_2)$-trajectory enters the 'forbidden region' of Fig. 5, consistent with the result of Sect. 4, that in this region the fast transition approximation breaks down.

Finally, we have simulated more realistic breathing cycles with shorter inhalation and longer exhalation period, where in both cases the expanding or compressing interval length $X(t)$ exponentially tends to a potential steady state: The cycle duration is 0.5 time units (comparable to 5 s) with 0.2 for inhalation and 0.3 for exhalation (see Fig. 11a). This time we used a doubled transition parameter $\varepsilon = 0.1$, resulting in a broader transition layer for the density profiles (see Fig. 12a). Again, as in Fig. 10b, the trajectories in the (ρ_1, ρ_2) state space (Fig. 12b) show a clear hysteresis behavior. Now the inner cycle of $\tilde{\rho}$ levels near the transition zone is

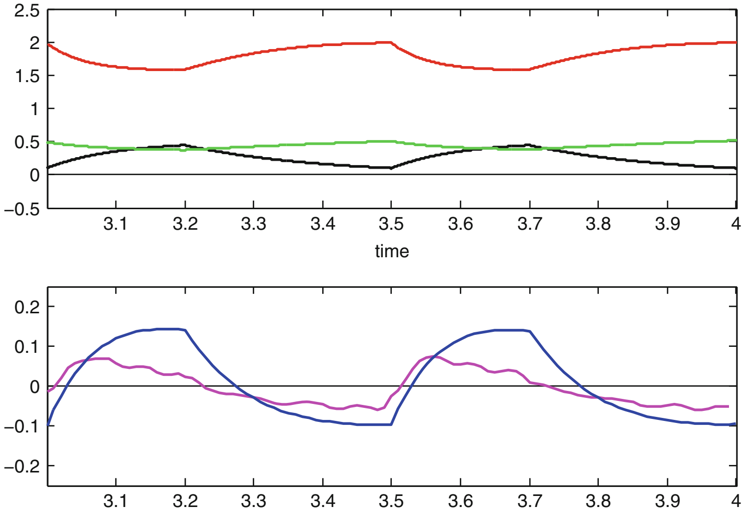


Fig. 11 Two cycles of a more realistic 'breathing' simulation with $\varepsilon = 0.1$. (**a**) *Upper plots*: the interval excess length $X(t) - 1$ (*black*) shows exponential curves with shorter 'inhalation' and longer 'exhalation'; also plotted are the resulting boundary values $\rho_1(t)$ and $\rho_2(t)$ (*green* and *red graphs*, respectively). (**b**) *Lower plots*: the resulting viscous transport velocity $V(t, s(t))$ at the middle point $s(t)$ of the smooth transition (*blue graph*) and the corresponding relative speed parameter $\lambda(t) = \dot{s}(t) - V(t, s(t))$ (*magenta graph*)

not so much reduced in size, but only shifted to higher ρ_1 values in such a way that the transition singularity (compare the small dot in Fig. 7) lies in the center of this hysteresis loop. This again is consistent with the results in Sect. 4, since the upper part of the hysteresis trajectory, roughly above this singularity, shows values $\lambda(t) < 0$, see Fig. 12b, what can also be checked by the time plot in Fig. 11b.

It is worthwhile to have a closer look at the $(\tilde{x}_2, \tilde{\rho}_2)$ cycle in the upper part of Fig. 12a: During 'in- or exhalation' there is first a period of changing $\tilde{\rho}_2$ levels with relatively constant $\tilde{x}_2$, meaning also a constant phase transition zone, followed by a period with changing position of the transition zone, while the $\tilde{\rho}_2$ level stays constant. Obviously this is a characteristic dynamic feature of the ordered monolayer. During a fast change of the interval size, the following two subsequent events take place: first, the lipid density is passively adjusted to the changing pressure, then, as a reactive process at the phase boundary of the monolayer, disordered lipids are associated or dissociated, respectively, while simultaneously the pressure adapts due to viscous flow within the monolayer.

Summarizing the simulation results of this section we can state that both 'breathing cycle' models are able to reproduce the hysteresis property, which has been experimentally observed from plotting the density-pressure isotherm curves (see Fig. 1). Such a curve can be expected also in our simulation, since the total pressure in Eq. (13) explicitly depends on the ρ and φ profiles.

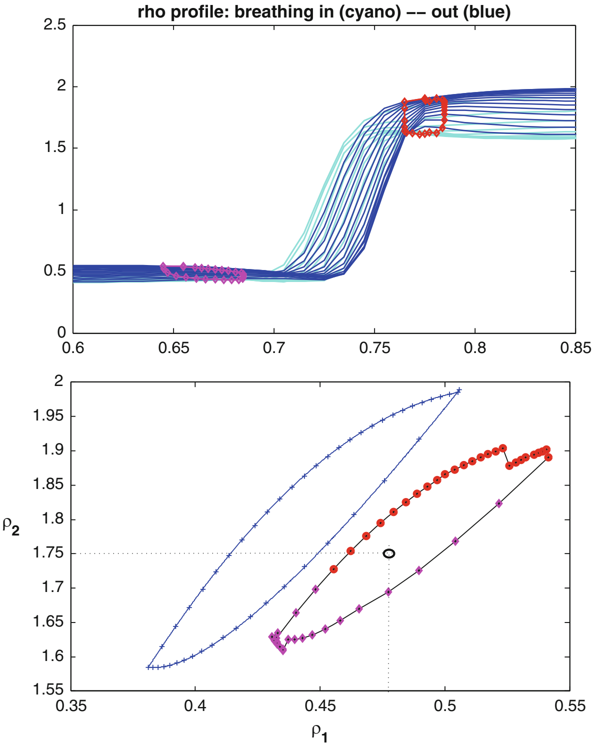


Fig. 12 Simulation of one periodic 'breathing' cycle as in Fig. 11: (**a**) *Top*: zoomed plots of density profiles $\rho(\tilde{x})$ with depicted points $(\tilde{x}_1, \tilde{\rho}_1)$ and $(\tilde{x}_2, \tilde{\rho}_2)$ left and right of the phase transition zone (*magenta* and *red dots*, respectively), cycling in a counter-clockwise sense. (**b**) *Bottom*: corresponding clockwise trajectories of boundary values (ρ_1, ρ_2) (*blue curve*) and of $(\tilde{\rho}_1, \tilde{\rho}_2)$ (*magenta diamonds* when $\lambda > 0$, and *red circles* when $\lambda < 0$), the latter cycling around the approximate steady-state singularity $(\tilde{\rho}_1^*, \tilde{\rho}_2^*)$ (*small ellipse*)

6 Conclusions

In extension of the results in [1] we demonstrated, how the mesoscopic continuum phase transition model can produce realistic simulations of a periodically expanded and compressed one-dimensional water surface with an embedded lipid monolayer on top. The observed phenomenon of hysteresis could be reproduced and further dynamical properties of the 'patchy' lipid monolayer system were revealed, which characterize its suspected functional role as an adaptive surfactant buffering system during breathing of the lung.

Moreover, a simple model for the stochastic motion and interaction of amphiphilic lipids as stiff rods on top of a fixed water surface was introduced and numerically simulated. The essential dynamics of ordered monolayer clustering from a reservoir of disordered diffusing lipids were realized under varying parameter conditions. In spite of its simplicity, this microscopic submodel could serve as an effective tool to study further connections between the detailed molecular interaction parameters and the lumped reaction parameters in the continuum two-phase mixture model.

Acknowledgements The presented results have been elaborated within the research program of the SFB 611 in Project B7. For the effective cooperation during 12 years we express cordial thanks to the co-leader, Hans Wilhelm Alt, moreover, to all coworkers and former members of our project and, in particular, to Martin Rumpf and his coworkers in Project B8. We also thank our former colleague Steven Perry from the Zoological Institute in Bonn for offering the lung alveoli surfactant problem in the starting phase of the project and for mediating the physiological literature.

References

1. Alt, H.W., Alt, W.: Phase boundary dynamics: transition between ordered and disordered lipid monolayers. Interfaces Free Bound. **11**, 1–36 (2009)
2. Alt, H.W., Alt, W.: Fluid mixtures and applications to biological systems. This volume, pp. 191–219 (2013)
3. Ding, J., Takamoto, D.Y., von Nahmen, A., Lipp, M.M., Lee, K.Y.C., Waring, A.J., Zasadzinski, J.A.: Effects of lung surfactant proteins, SP-B and SP-C, and palmitic acid on monolayer stability. Biophys. J. **80**, 2262–2272 (2001)
4. Knecht, V., Bonn, M., Marrink, S.-J., Mark, A.E.: Simulation studies of pore and domain formation in a phospholipid monolayer. J. Chem. Phys. **122**, 024704 (2005)
5. Nemadjieu, S.F.: Finite volume methods for advection diffusion on moving interfaces and application on surfactant driven thin film flow. Dissertation, University of Bonn (2012)

Fluid Mixtures and Applications to Biological Systems

Hans Wilhelm Alt and Wolfgang Alt

Abstract We apply the free energy principle to fluid systems, where the components react with each other. As example we treat the predator-prey system and cyclic reactions. We deal with the polymerization of actin filaments and with the general diffusion limit.

1 Introduction

We consider a mixture of fluids with applications as they widely occur in biology, biophysics and biochemistry. It is assumed that for each fluid a conservation law for the momentum is satisfied. This is true for a mixture of particle systems, where the attraction force for molecules of the same kind is stronger than the attraction between different species of the mixture. For example, this is the case for liquid-solid mixtures, see Rajagopal [10, 3.3 Basic Equations].

In Sect. 3 we consider mass and momentum balances for each component of the mixture, each component having it's own velocity with interaction terms between the different momentum equations, see the system (1).

H.W. Alt (✉)
Zentrum für Mathematik, Technische Universität München, Boltzmannstraße 3,
D-85747 Garching, Germany
e-mail: alt@ma.tum.de

W. Alt
Abteilung Theoretische Biologie, Institut für Zelluläre und Molekulare Botanik, Rheinische
Friedrich-Wilhelms-Universität Bonn, Kirschallee 1-3, D-53115 Bonn, Germany
e-mail: wolf.alt@uni-bonn.de

M. Griebel (ed.), *Singular Phenomena and Scaling in Mathematical Models*,
DOI 10.1007/978-3-319-00786-1_9,

Constitutive equations will be derived with the help of the entropy principle in the version of Müller [9]. That book also contains a treatment of mixtures, but his theory of mixtures of fluids is restricted to the non-viscous case [9, Chap. 6 $(6.18)_2$]. We insert also viscous terms in the momentum equations.

We consider the isothermal case. In this case the entropy principle becomes the free energy inequality. We show how this leads to restrictions on the constitutive equations and end up with an equivalent system (19), which is the basis for further studies.

We deal with several special topics, among them free energies depending on gradients (Sect. 6), the system for total density and fractional densities (Sect. 9), a contribution to quasistatic problems (Sect. 8), which is followed by a consideration of a diffusive limit (Sect. 10). Beside this we give some particular examples from biology (Sects. 5 and 7) as the Lotka-Volterra system and the polymerization of actin filaments. There is a full theory for chemical systems, but a theory for general biological problems goes beyond this and is new. The reason is that the non-negativity for each reaction is a too strong assumption for the non-negativity of the free energy production. In this paper we cannot give a full theory for all cases so this is reserved to considerations in a future paper.

2 Fluid Mixtures

We consider a mixture of compressible fluids, where ρ_α is the mass density and v_α the velocity of the α-th constituent. The balance laws for mass and momentum for each fluid component α are[1]

$$\begin{aligned}&\partial_t \rho_\alpha + \operatorname{div}(\rho_\alpha v_\alpha) = \tau_\alpha,\\&\partial_t(\rho_\alpha v_\alpha) + \operatorname{div}(\rho_\alpha v_\alpha \otimes v_\alpha + \Pi_\alpha) = g_\alpha + \tau_\alpha v_\alpha + \mathbf{f}_\alpha.\end{aligned} \tag{1}$$

This holds for each α. Here τ_α are reaction terms, $\mathbf{f}_\alpha$ are possible external forces and g_α are interaction forces. The matrix Π_α is the pressure tensor containing as part the negative stress tensor, as we will see later in Theorem 1.

Besides these balance laws we have constitutive relations for τ_α, Π_α, and g_α, which couple the equations. These conditions are subject to restrictions coming from the principle of objectivity (see the Remark 1 below). So the mass production τ_α is an objective scalar, the pressure tensor Π_α is an objective tensor, $\mathbf{f}_\alpha$ transform like an external force and the coupling term g_α is an objective vector.

[1] Note that in the second equation the divergence acts on the second index of tensors.

Definition 1 (Constitutive Equations). Using the notation

$$\rho = \left(\rho_\beta\right)_\beta, \quad v = \left(v_\beta\right)_\beta, \tag{2}$$

we assume constitutive relations for

$$\tau_\alpha,\ g_\alpha,\ \Pi_\alpha,$$

in general depending on $(\rho, v, \nabla\rho, \mathrm{D}v)$. We do not specify $\mathbf{f}_\alpha$.

In a single fluid it often happens that dependencies on the velocity drop out by objectivity. For mixtures the situation is quite different, since differences $v_{\alpha_1} - v_{\alpha_2}$ of two velocities are objective vectors (see [10, 4. Constitutive Equations]). Therefore we define in accordance with [7, Chap. XI §2] the following:

Definition 2 (Barycentric Velocity). A mean density and a mean velocity is defined by

$$\bar\rho := \sum\nolimits_\alpha \rho_\alpha, \quad \bar v := \frac{1}{\bar\rho}\sum\nolimits_\alpha \rho_\alpha v_\alpha$$

where α runs from 1 to N, the number of components. We define the relative velocities by

$$u_\alpha := v_\alpha - \bar v, \tag{3}$$

and it follows that these are objective vectors (see Remark 1).

Lemma 1. *Obviously*

$$\sum\nolimits_\alpha \rho_\alpha u_\alpha = 0. \tag{4}$$

As a consequence

$$\mathrm{D}\bar v = \sum\nolimits_\alpha \frac{\rho_\alpha}{\bar\rho}\mathrm{D}v_\alpha + \sum\nolimits_\alpha \frac{1}{\bar\rho} u_\alpha \otimes \nabla\rho_\alpha. \tag{5}$$

Proof. Equation (4) is a direct consequence of the definition of $\bar v$. Computing the derivative of (4) one obtains

$$\begin{aligned} 0 = \mathrm{D}\Big(\sum\nolimits_\alpha \rho_\alpha u_\alpha\Big) &= \sum\nolimits_\alpha u_\alpha \otimes \nabla\rho_\alpha + \sum\nolimits_\alpha \rho_\alpha \mathrm{D}u_\alpha \\ &= \sum\nolimits_\alpha u_\alpha \otimes \nabla\rho_\alpha + \sum\nolimits_\alpha \rho_\alpha \mathrm{D}v_\alpha - \bar\rho\mathrm{D}\bar v \end{aligned}$$

by using that $u_\alpha = v_\alpha - \bar v$. □

Here we consider materials where all velocities v_α are independent variables. Likewise $\bar v$ and the u_α, obviously with the constraint (4), are independent variables. Summing up the Eqs. (1) and using (4) we obtain, that these mean quantities satisfy the total mass and total momentum balance equations

$$\begin{aligned}&\partial_t\bar\rho + \operatorname{div}(\bar\rho\,\bar v) = \sum\nolimits_\alpha \tau_\alpha,\\ &\partial_t(\bar\rho\bar v) + \operatorname{div}\Big(\bar\rho\bar v\otimes\bar v + \sum\nolimits_\alpha \rho_\alpha u_\alpha\otimes u_\alpha + \sum\nolimits_\alpha \Pi_\alpha\Big) = \sum\nolimits_\alpha (g_\alpha + \tau_\alpha v_\alpha + \mathbf{f}_\alpha).\end{aligned} \tag{6}$$

We now interpret the quantities in these common equations.

Proposition 1 (Collective Quantities). *It follows that (6) is equivalent to*

$$\begin{aligned}&\partial_t\bar\rho + \operatorname{div}(\bar\rho\,\bar v) = \bar\tau,\\ &\partial_t(\bar\rho\bar v) + \operatorname{div}\Big(\bar\rho\bar v\otimes\bar v + \bar\Pi\Big) = \bar g + \bar\tau\bar v + \bar{\mathbf{f}},\end{aligned} \tag{7}$$

if the production of the total mass $\bar\rho$ is

$$\bar\tau := \sum\nolimits_\alpha \tau_\alpha$$

and if the pressure tensor $\bar\Pi$ for the total fluid is

$$\bar\Pi := \sum\nolimits_\alpha \rho_\alpha u_\alpha\otimes u_\alpha + \sum\nolimits_\alpha \Pi_\alpha.$$

We assume that $\bar\Pi$ is a symmetric objective tensor (the tensors Π_α do not need to be symmetric). Moreover, the momentum production terms are given by

$$\bar g := \sum\nolimits_\alpha (\tau_\alpha u_\alpha + g_\alpha), \quad \bar{\mathbf{f}} := \sum\nolimits_\alpha \mathbf{f}_\alpha.$$

For a closed system one assumes that $\bar\tau$ and $\bar g$ vanish (see Lemma 2). To include other systems we will proceed with arbitrary values of $\bar\tau$ and $\bar g$, although then one would insist to extend the system to a closed one, or one needs to say on which closed system the model relies. For further considerations it is important how the total kinetic free energy is defined. The sum of the kinetic energy of each fluid

$$f^{kin} := \sum\nolimits_\alpha \frac{\rho_\alpha}{2}|v_\alpha|^2 = \frac{\bar\rho}{2}|\bar v|^2 + \sum\nolimits_\alpha \frac{\rho_\alpha}{2}|u_\alpha|^2 \tag{8}$$

is a basis for it, and a single kinetic energy satisfies the following evolution equation.

Proposition 2. *For each α it follows from the equations in (1) that the following identity*

$$\partial_t\Big(\frac{\rho_\alpha}{2}|v_\alpha|^2\Big) + \mathrm{div}\Big(\frac{\rho_\alpha}{2}|v_\alpha|^2 v_\alpha + \Pi_\alpha^T v_\alpha\Big) = \mathrm{D}v_\alpha \bullet \Pi_\alpha + v_\alpha \bullet (g_\alpha + \mathbf{f}_\alpha) + \frac{\tau_\alpha}{2}|v_\alpha|^2$$

is satisfied.

Proof. For each index α and any linear first order differential operator $L = \beta_0\partial_t + \sum_i \beta_i \partial_i$ we compute

$$L(\frac{\rho_\alpha}{2}|v_\alpha|^2) = \frac{|v_\alpha|^2}{2}L\rho_\alpha + \rho_\alpha v_\alpha \bullet L v_\alpha = -\frac{|v_\alpha|^2}{2}L\rho_\alpha + v_\alpha \bullet L(\rho_\alpha v_\alpha),$$

and for $L = \partial_t + v_\alpha \bullet \nabla$ we write the differential equations (1) as

$$\begin{aligned} L\rho_\alpha &= -\rho_\alpha \mathrm{div} v_\alpha + \tau_\alpha, \\ L(\rho_\alpha v_\alpha) &= -\rho_\alpha(\mathrm{div} v_\alpha) v_\alpha - \mathrm{div}\Pi_\alpha + (g_\alpha + \tau_\alpha v_\alpha + \mathbf{f}_\alpha) \end{aligned}$$

and obtain

$$\begin{aligned} L\Big(\frac{\rho_\alpha|v_\alpha|^2}{2}\Big) &= -\frac{|v_\alpha|^2}{2}L\rho_\alpha + v_\alpha \bullet L(\rho_\alpha v_\alpha) \\ &= \Big(\frac{|v_\alpha|^2}{2} - v_\alpha \bullet v_\alpha\Big)\rho_\alpha \mathrm{div} v_\alpha - \frac{|v_\alpha|^2}{2}\tau_\alpha - v_\alpha \bullet \mathrm{div}\Pi_\alpha + v_\alpha \bullet (g_\alpha + \tau_\alpha v_\alpha + \mathbf{f}_\alpha) \\ &= -\frac{|v_\alpha|^2}{2}\rho_\alpha \mathrm{div} v_\alpha - \mathrm{div}(\Pi_\alpha{}^{\mathrm{T}} v_\alpha) + \mathrm{D}v_\alpha \bullet \Pi_\alpha + v_\alpha \bullet (g_\alpha + \mathbf{f}_\alpha) + \frac{|v_\alpha|^2}{2}\tau_\alpha. \end{aligned}$$

This gives the result

$$\begin{aligned} &\partial_t\Big(\frac{\rho_\alpha}{2}|v_\alpha|^2\Big) + \mathrm{div}\Big(\frac{\rho_\alpha}{2}|v_\alpha|^2 v_\alpha + \Pi_\alpha^T v_\alpha\Big) \\ &= L\Big(\frac{\rho_\alpha}{2}|v_\alpha|^2\Big) + \frac{|v_\alpha|^2}{2}\rho_\alpha \mathrm{div} v_\alpha + \mathrm{div}(\Pi_\alpha^T v_\alpha) \\ &= \mathrm{D}v_\alpha \bullet \Pi_\alpha + v_\alpha \bullet (g_\alpha + \mathbf{f}_\alpha) + \tau_\alpha \frac{|v_\alpha|^2}{2}. \end{aligned}$$

□

The equations have to be supplemented by the entropy principle, which in the here considered isothermal case is equivalent to the free energy inequality. This requires a definition of the (total) free energy, which as a part consists of the kinetic free energy f^{kin}.

Definition 3 (Postulate: Free energy principle). The postulate is that there exist a (total) free energy f^{tot} and a free energy flux ϕ^{tot}, such that for all solutions (ρ, v) of the mixture problem the inequality

$$\partial_t f^{tot} + \mathrm{div}\phi^{tot} - g^{tot} \le 0 \tag{9}$$

holds, and $(f^{tot}, \phi^{tot}, g^{tot})$ satisfy certain constitutive relations. Among this

$$f^{tot} := \sum\nolimits_\alpha \frac{\rho_\alpha}{2}|v_\alpha|^2 + f = f^{kin} + f, \tag{10}$$

where the internal free energy f is an objective scalar. The constitutive assumption on f is a consequence of the materials considered here (see (13) and (25)), in any case it will depend on ρ. For the flux we assume

$$\phi^{tot} := \sum\nolimits_\alpha \frac{\rho_\alpha}{2}|v_\alpha|^2 v_\alpha + f\bar{v} + \sum\nolimits_\alpha \Pi_\alpha^T v_\alpha + \phi,$$

where ϕ is an objective vector, which has to be determined later. The term g^{tot} has in accordance with objectivity the form

$$g^{tot} = \frac{\bar{\tau}}{2}|\bar{v}|^2 + \bar{g} \bullet \bar{v} + \sum\nolimits_\alpha v_\alpha \bullet \mathbf{f}_\alpha.$$

This contains the usual external force terms $\mathbf{f}_\alpha$, but also terms containing the external quantities $\bar{\tau}$ and $\bar{g}$.

It is important to say that the inequality (9) implies that

$$h := \partial_t f^{tot} + \text{div}\phi^{tot} - g^{tot}$$

has to be an objective scalar (see [2, Lemma 10.3]). That is the reason why a term g^{tot} is necessary. The free energy f^{tot} transforms in the same way as f^{kin} does. This determines the transformation formula for the flux ϕ^{tot}, which has an unknown term ϕ and which turns out to be an objective vector. It can be written as

$$\phi^{tot} = f^{tot}\bar{v} + \Big(\sum\nolimits_\alpha \frac{\rho_\alpha}{2}|v_\alpha|^2 u_\alpha + \sum\nolimits_\alpha \Pi_\alpha^T v_\alpha\Big) + \phi.$$

Remark 1 (Objectivity). The objectivity of the system itself can be found in [2, Chap. 8]. In [2] an arbitrary observer transformation is given by a map Y,

$$\begin{bmatrix} t \\ x \end{bmatrix} = Y\Big(\begin{bmatrix} t^* \\ x^* \end{bmatrix}\Big) = \begin{bmatrix} t^* + a \\ X(t^*, x^*) \end{bmatrix}, \quad X(t^*, x^*) = Q(t^*)x^* + b(t^*),$$

where the quantities with respect to the new observer are indicated by a star. Then, besides well known transformations of terms in the α-system (1), in particular the following is true. The velocity transforms like

$$v_\alpha \circ Y = \dot{X} + Qv_\alpha^*,$$

the force of the α-system transforms like

$$\mathbf{f}_\alpha \circ Y = \rho_\alpha^*(\ddot{X} + 2\dot{Q}v_\alpha^*) + Q\mathbf{f}_\alpha^*,$$

and the force for the global system (6) transforms like

$$\bar{\mathbf{f}} \circ Y = \bar{\rho}^*(\ddot{X} + 2\dot{Q}\bar{v}^*) + Q\bar{\mathbf{f}}^*,$$

which is consistent with the definition of the collective $\bar{\mathbf{f}}$. The objectivity about the terms in (9), which assumes (7) can be found in [2, Chap. 10]. As mentioned above, the term g^{tot} has to contain the not objective scalar terms

$$\frac{\bar{\tau}}{2}|\bar{v}|^2 + (\bar{g} + \bar{\mathbf{f}})\bullet\bar{v},$$

and it can be shown that the term $\sum_\alpha u_\alpha \bullet \mathbf{f}_\alpha$ is an objective scalar. (It is not clear, where an objective scalar should be placed, in g^{tot} or h, one has to perform the entropy inequality to clarify this.)

The free energy inequality reduces to the following inequality for the internal free energy.

Proposition 3. *For the free energy production h one computes*

$$0 \geq h = \partial_t f + \mathrm{div}(f\bar{v} + \phi) + \sum\nolimits_\alpha \Big(\mathrm{D}v_\alpha \bullet \Pi_\alpha + u_\alpha \bullet g_\alpha + \frac{\tau_\alpha}{2}|u_\alpha|^2\Big) \tag{11}$$

for every solution of system (1).

Proof. Summing up in Proposition 2 one obtains

$$\partial_t f^{kin} + \mathrm{div}\Big(\sum\nolimits_\alpha (\frac{\rho_\alpha}{2}|v_\alpha|^2 v_\alpha + \Pi_\alpha^T v_\alpha)\Big)$$
$$= \sum\nolimits_\alpha \Big(\mathrm{D}v_\alpha \bullet \Pi_\alpha + v_\alpha \bullet (g_\alpha + \mathbf{f}_\alpha) + \frac{\tau_\alpha}{2}|v_\alpha|^2\Big).$$

Inserting this in the definition of h gives

$$\begin{aligned} h &= \partial_t f^{tot} + \mathrm{div}\phi^{tot} - g^{tot} \\ &= \partial_t (f^{kin} + f) + \mathrm{div}\Big(\sum\nolimits_\alpha \Big(\frac{\rho_\alpha}{2}|v_\alpha|^2 v_\alpha + \Pi_\alpha^T v_\alpha\Big) + f\bar{v} + \phi\Big) - g^{tot} \\ &= \partial_t f + \mathrm{div}(f\bar{v} + \phi) - g^{tot} \\ &+ \sum\nolimits_\alpha \Big(\mathrm{D}v_\alpha \bullet \Pi_\alpha + v_\alpha \bullet (g_\alpha + \mathbf{f}_\alpha) + \frac{\tau_\alpha}{2}|v_\alpha|^2\Big). \end{aligned}$$

Since

$$\begin{aligned} \sum\nolimits_\alpha \Big(v_\alpha \bullet (g_\alpha + \mathbf{f}_\alpha) + \frac{\tau_\alpha}{2}|v_\alpha|^2\Big) &= \sum\nolimits_\alpha \Big((u_\alpha + \bar{v})\bullet g_\alpha + \frac{\tau_\alpha}{2}|u_\alpha + \bar{v}|^2\Big) + \sum\nolimits_\alpha v_\alpha \bullet \mathbf{f}_\alpha \\ &= \sum\nolimits_\alpha \Big(u_\alpha \bullet g_\alpha + \frac{\tau_\alpha}{2}|u_\alpha|^2\Big) + R, \end{aligned}$$

where

$$\begin{aligned} R =& \sum\nolimits_\alpha \Big(\bar{v}\bullet g_\alpha + \frac{\tau_\alpha}{2}(2u_\alpha\bullet\bar{v} + |\bar{v}|^2)\Big) + \sum\nolimits_\alpha v_\alpha\bullet\mathbf{f}_\alpha \\ =& \bar{v}\bullet\sum\nolimits_\alpha \Big(g_\alpha + \tau_\alpha u_\alpha\Big) + |\bar{v}|^2\sum\nolimits_\alpha \frac{\tau_\alpha}{2} + \sum\nolimits_\alpha v_\alpha\bullet\mathbf{f}_\alpha \\ =& \bar{v}\bullet\bar{g} + |\bar{v}|^2\frac{\bar{\tau}}{2} + \sum\nolimits_\alpha v_\alpha\bullet\mathbf{f}_\alpha = g^{tot}, \end{aligned} \tag{12}$$

the assertion follows. □

It is the aim to determine the consequences of the free energy principle. For this it is now essential, what constitutive properties f and ϕ have.

3 Exploiting the Free Energy Inequality

In this section we make use of the assumption that the free energy f depends on all the densities ρ_α, and with this assumption we go into the free energy inequality (11).

Proposition 4. *If the free energy is given by*

$$f \equiv \hat{f}(\rho), \tag{13}$$

then we obtain for the free energy production

$$\begin{aligned} h = \mathrm{div}\phi + \sum\nolimits_\alpha f'_{\rho_\alpha}\tau_\alpha + \sum\nolimits_\alpha u_\alpha\bullet\Big(g_\alpha + \frac{\tau_\alpha}{2}u_\alpha + (\frac{f}{\bar{\rho}} - f'_{\rho_\alpha})\nabla\rho_\alpha\Big) \\ + \sum\nolimits_\alpha \mathrm{D}v_\alpha\bullet\Big(\Pi_\alpha + \rho_\alpha(\frac{f}{\bar{\rho}} - f'_{\rho_\alpha})\mathrm{Id}\Big). \end{aligned}$$

Thus the production term is written in the independent gradient terms $\mathrm{D}v_\alpha$ *and* $\nabla\rho_\alpha$.

Proof. Since $f = \hat{f}((\rho_\alpha)_\alpha)$ we compute, by making use of the mass conservations,

$$\begin{aligned} \partial_t f + \mathrm{div}(f\bar{v}) =& (\partial_t + \bar{v}\bullet\nabla)f + f\mathrm{div}\bar{v} \\ =& \sum\nolimits_\alpha f'_{\rho_\alpha}\cdot\Big(\partial_t\rho_\alpha + \bar{v}\bullet\nabla\rho_\alpha\Big) + f\mathrm{div}\bar{v} \\ =& \sum\nolimits_\alpha f'_{\rho_\alpha}\cdot\Big(\tau_\alpha - \mathrm{div}(\rho_\alpha v_\alpha) + \bar{v}\bullet\nabla\rho_\alpha\Big) + f\mathrm{div}\bar{v} \\ =& \sum\nolimits_\alpha f'_{\rho_\alpha}\tau_\alpha - \sum\nolimits_\alpha f'_{\rho_\alpha}u_\alpha\bullet\nabla\rho_\alpha - \sum\nolimits_\alpha \rho_\alpha f'_{\rho_\alpha}\mathrm{div}v_\alpha + f\mathrm{div}\bar{v}. \end{aligned}$$

We plug this into the expression for h and obtain

$$\begin{aligned} h = \operatorname{div}\phi &+ \textstyle\sum_\alpha f'_{\rho_\alpha}\tau_\alpha - \sum_\alpha f'_{\rho_\alpha} u_\alpha \bullet \nabla \rho_\alpha \\ &- \textstyle\sum_\alpha \rho_\alpha f'_{\rho_\alpha} \operatorname{div} v_\alpha + f \operatorname{div} \bar{v} \\ &+ \textstyle\sum_\alpha \mathrm{D} v_\alpha \bullet \Pi_\alpha + \sum_\alpha u_\alpha \bullet g_\alpha + \sum_\alpha \frac{\tau_\alpha}{2} |u_\alpha|^2. \end{aligned}$$

Now we use (5) to derive

$$\operatorname{div}\bar{v} = \sum_\alpha \frac{\rho_\alpha}{\bar{\rho}} \operatorname{div} v_\alpha + \sum_\alpha \frac{1}{\bar{\rho}} u_\alpha \bullet \nabla \rho_\alpha,$$

which gives

$$\begin{aligned} &- \textstyle\sum_\alpha \rho_\alpha f'_{\rho_\alpha} \operatorname{div} v_\alpha + f \operatorname{div}\bar{v} \\ &= \textstyle\sum_\alpha \big(\frac{\rho_\alpha f}{\bar{\rho}} - \rho_\alpha f'_{\rho_\alpha}\big) \operatorname{div} v_\alpha + \sum_\alpha \frac{f}{\bar{\rho}} u_\alpha \bullet \nabla \rho_\alpha. \end{aligned}$$

Therefore we obtain for the energy production

$$\begin{aligned} h = \operatorname{div}\phi + &\sum_\alpha f'_{\rho_\alpha}\tau_\alpha - \sum_\alpha f'_{\rho_\alpha} u_\alpha \bullet \nabla \rho_\alpha \\ &+ \sum_\alpha u_\alpha \bullet g_\alpha + \sum_\alpha \frac{\tau_\alpha}{2} |u_\alpha|^2 + \sum_\alpha \frac{f}{\bar{\rho}} u_\alpha \bullet \nabla \rho_\alpha \\ &+ \sum_\alpha \mathrm{D} v_\alpha \bullet \Big(\Pi_\alpha + \big(\frac{\rho_\alpha f}{\bar{\rho}} - \rho_\alpha f'_{\rho_\alpha}\big) \mathrm{Id} \Big) \\ = \operatorname{div}\phi + &\sum_\alpha f'_{\rho_\alpha}\tau_\alpha \\ &+ \sum_\alpha u_\alpha \bullet \Big(g_\alpha + \frac{\tau_\alpha}{2} u_\alpha + \big(\frac{f}{\bar{\rho}} - f'_{\rho_\alpha}\big) \nabla \rho_\alpha \Big) \\ &+ \sum_\alpha \mathrm{D} v_\alpha \bullet \Big(\Pi_\alpha + \rho_\alpha \big(\frac{f}{\bar{\rho}} - f'_{\rho_\alpha}\big) \mathrm{Id} \Big). \end{aligned}$$ □

If we now let

$$\Pi_\alpha = p_\alpha \mathrm{Id} - S_\alpha, \quad p_\alpha := \rho_\alpha \big(f'_{\rho_\alpha} - \frac{f}{\bar{\rho}}\big),$$

we can rewrite the $\mathrm{D}v_\alpha$-term in the free energy production h. If in addition we now define the specific free energy and the specific pressures by

$$f^{sp} = \frac{f}{\bar{\rho}}, \quad p^{sp}_\alpha = \frac{p_\alpha}{\rho_\alpha}, \tag{14}$$

where the specific pressure p^{sp}_α is defined with respect to the density ρ_α, then we obtain the following theorem as a consequence.

Theorem 1. *Let be $f \equiv \hat{f}(\rho)$ and let the chemical potential be*

$$\mu_\alpha := f_{,\rho_\alpha}. \tag{15}$$

If in addition to assumption (13) we suppose $\phi = 0$ and

$$\Pi_\alpha = p_\alpha \mathrm{Id} - S_\alpha, \quad p_\alpha = \rho_\alpha p_\alpha^{sp}, \tag{16}$$
$$g_\alpha = p_\alpha^{sp} \nabla \rho_\alpha - \frac{\tau_\alpha}{2} u_\alpha + g_\alpha^{fr} - \rho_\alpha g^{sp},$$

where the objective quantities S_α, τ_α, g_α^{fr}, and g^{sp} are arbitrary constitutive functions, then for solutions of (1) the free energy production h reads

$$0 \geq h = -\sum\nolimits_\alpha \mathrm{D}v_\alpha \bullet S_\alpha + \sum\nolimits_\alpha \tau_\alpha \mu_\alpha + \sum\nolimits_\alpha g_\alpha^{fr} \bullet u_\alpha. \tag{17}$$

Proof. This follows immediately from Proposition 4, where one has to take into account the constraint (4) for the relative velocities u_α. □

We also have the identity

$$f_{,\rho_\alpha} - \frac{f}{\bar{\rho}} = \bar{\rho}\Big(\frac{f}{\bar{\rho}}\Big)_{,\rho_\alpha} = \bar{\rho} \cdot f^{sp}_{,\rho_\alpha}$$

and therefore

$$p_\alpha^{sp} = \bar{\rho} \cdot f^{sp}_{,\rho_\alpha}. \tag{18}$$

The friction terms g_α^{fr} are those terms of the interactive force, which contribute to the free energy production. The vector field g^{sp}, which is independent of α, is due to the constraint (4) and contributes to the external term $\bar{g}$ and to the differential equations, see Proposition 5 and Lemma 3 below.

Lemma 2 (External Quantities). *Define*

$$\bar{g}^{fr} := \sum\nolimits_\alpha g_\alpha^{fr}.$$

Then

$$\bar{\tau} = \sum\nolimits_\alpha \tau_\alpha,$$
$$\bar{g} = \sum\nolimits_\alpha \frac{\tau_\alpha}{2} u_\alpha + \bar{\rho}(\nabla f^{sp} - g^{sp}) + \bar{g}^{fr}.$$

Note here that the external terms $\bar{\tau}$ and $\bar{g}$ vanish for a completely closed model as mentioned in Sect. 2, see Lemma 3 below.

Proof. From (16) we obtain

$$g_\alpha + \tau_\alpha u_\alpha = \frac{\tau_\alpha}{2} u_\alpha + p_\alpha^{sp} \nabla \rho_\alpha + g_\alpha^{fr} - \rho_\alpha g^{sp},$$

hence

$$\bar{g} = \sum\nolimits_\alpha \frac{\tau_\alpha}{2} u_\alpha + \sum\nolimits_\alpha (p_\alpha^{sp} \nabla \rho_\alpha - \rho_\alpha g^{sp}) + \bar{g}^{fr}$$

and from (18)

$$\sum\nolimits_\alpha (p_\alpha^{sp} \nabla \rho_\alpha - \rho_\alpha g^{sp}) = \sum\nolimits_\alpha \bar{\rho} f^{sp}_{,\rho_\alpha} \nabla \rho_\alpha - \big(\sum\nolimits_\alpha \rho_\alpha\big) g^{sp} = \bar{\rho}(\nabla f^{sp} - g^{sp}). \qquad \square$$

In summary, we obtain the following conclusion.

Proposition 5. *Under the assumptions of Theorem 1 the mixture system (1) is equivalent to*

$$\begin{aligned} &\partial_t \rho_\alpha + \operatorname{div}(\rho_\alpha v_\alpha) = \tau_\alpha, \\ &\rho_\alpha \big(\partial_t v_\alpha + (v_\alpha \bullet \nabla) v_\alpha\big) \\ &\quad = \operatorname{div} S_\alpha - \rho_\alpha (\nabla p_\alpha^{sp} + g^{sp}) - \frac{\tau_\alpha}{2} u_\alpha + g_\alpha^{fr} + \mathbf{f}_\alpha \end{aligned} \tag{19}$$

for all α. The free energy inequality (17) is satisfied.

Proof. With the assumptions in Theorem 1 the momentum law in (1) becomes

$$\begin{aligned} &\partial_t (\rho_\alpha v_\alpha) + \operatorname{div}(\rho_\alpha v_\alpha \otimes v_\alpha + p_\alpha \mathrm{Id} - S_\alpha) \\ &= \tau_\alpha v_\alpha + g_\alpha + \mathbf{f}_\alpha \\ &= \tau_\alpha \frac{\bar{v} + v_\alpha}{2} + p_\alpha^{sp} \nabla \rho_\alpha - \rho_\alpha g^{sp} + g_\alpha^{fr} + \mathbf{f}_\alpha, \end{aligned}$$

or if we use the mass equation in (1)

$$\begin{aligned} &\rho_\alpha \big(\partial_t v_\alpha + (v_\alpha \bullet \nabla) v_\alpha\big) + \operatorname{div}(p_\alpha \mathrm{Id} - S_\alpha) \\ &= g_\alpha + \mathbf{f}_\alpha \\ &= -\frac{\tau_\alpha}{2} u_\alpha + p_\alpha^{sp} \nabla \rho_\alpha - \rho_\alpha g^{sp} + g_\alpha^{fr} + \mathbf{f}_\alpha, \end{aligned}$$

or equivalently

$$\begin{aligned} &\rho_\alpha \big(\partial_t v_\alpha + (v_\alpha \bullet \nabla) v_\alpha\big) \\ &= \operatorname{div} S_\alpha - \nabla p_\alpha + g_\alpha + \mathbf{f}_\alpha \end{aligned}$$

$$\begin{aligned}&= \operatorname{div} S_\alpha + p_\alpha^{sp}\nabla\rho_\alpha - \nabla p_\alpha - \rho_\alpha g^{sp} - \tfrac{\tau_\alpha}{2}u_\alpha + g_\alpha^{fr} + \mathbf{f}_\alpha\\ &= \operatorname{div} S_\alpha - \rho_\alpha(\nabla p_\alpha^{sp} + g^{sp}) - \tfrac{\tau_\alpha}{2}u_\alpha + g_\alpha^{fr} + \mathbf{f}_\alpha .\end{aligned}$$

Here we have used the fact that

$$\nabla p_\alpha = \nabla(\rho_\alpha p_\alpha^{sp}) = \rho_\alpha \nabla p_\alpha^{sp} + p_\alpha^{sp}\nabla\rho_\alpha . \qquad \square$$

The additional term g^{sp} can be chosen in a way that $\bar{g} = 0$.

Lemma 3. *It is $\bar{g} = 0$, if we choose g^{sp} as*

$$g^{sp} := \frac{1}{2\bar\rho}\sum\nolimits_\alpha \tau_\alpha u_\alpha + \nabla f^{sp} + \frac{1}{\bar\rho}\bar{g}^{fr}.$$

This follows immediately from the representation in Lemma 2. With this assumption we obtain the following theorem.

Theorem 2. *Under the assumptions of Theorem 1 and if g^{sp} is chosen as in Lemma 3 the system (1) is equivalent to*

$$\begin{aligned}&\partial_t\rho_\alpha + \operatorname{div}(\rho_\alpha v_\alpha) = \tau_\alpha,\\ &\rho_\alpha\big(\partial_t v_\alpha + (v_\alpha \bullet \nabla)v_\alpha\big)\\ &\qquad = \operatorname{div} S_\alpha - \rho_\alpha\nabla\mu_\alpha - \frac{1}{2}\Big(\tau_\alpha u_\alpha + \frac{\rho_\alpha}{\bar\rho}\sum\nolimits_\beta \tau_\beta u_\beta\Big) + g_\alpha^{fr} - \frac{\rho_\alpha}{\bar\rho}\bar{g}^{fr} + \mathbf{f}_\alpha\end{aligned}$$

for all α. Here $\mu_\alpha = f_{'\rho_\alpha}$ are the chemical potentials.

Proof. It is $p_\alpha^{sp} + f^{sp} = f_{'\rho_\alpha} = \mu_\alpha$. With this the assertion follows from the previous Proposition 5. □

If one makes the natural assumption that the friction terms g_α^{fr} sum up to $\bar{g}^{fr} = 0$, the $\bar{g}^{fr}$ term in the momentum equation vanishes. The easiest way to verify that the free energy inequality (17) is satisfied is to assume that all three components of the free energy production have a sign. (We remark that in [3, §4] a different splitting is used.) This is the case in the following lemma.

Lemma 4. *Denote $\mu = (\mu_\alpha)_\alpha$ and $u := (u_\alpha)_\alpha$. If*

$$\begin{aligned}&S_\alpha \equiv \hat{S}_\alpha(\rho, (\mathrm{D}v)^S) := \sum\nolimits_\beta \big(a_{\alpha\beta}(\mathrm{D}v_\beta)^S + b_{\alpha\beta}\cdot \operatorname{div}(v_\beta)\cdot \mathrm{Id}\big),\\ &\tau_\alpha \equiv \hat\tau_\alpha(\rho,\mu), \quad g_\alpha^{fr} \equiv \hat{g}_\alpha^{fr}(\rho, u),\end{aligned}$$

with

$$a_{\alpha\beta} \equiv \hat{a}_{\alpha\beta}(\rho),\ b_{\alpha\beta} \equiv \hat{b}_{\alpha\beta}(\rho) \textit{ positive semidefinite in } (\alpha, \beta),$$
$$\sum\nolimits_{\beta} \mu_\beta \hat{\tau}_\beta(\rho, \mu) \le 0, \quad \sum\nolimits_{\beta} u_\beta \bullet \hat{g}^{fr}_\beta(\rho, u) \le 0,$$

then the free energy inequality (17) *is satisfied.*

4 Remark on Pressure

If we change in the Eqs. (1)

$$\Pi_\alpha = \Pi_\alpha^{new} + \omega_\alpha \mathrm{Id}, \quad g_\alpha = g_\alpha^{new} + \nabla\omega_\alpha, \tag{20}$$

the differential equations stay the same, since $\mathrm{div}(\omega_\alpha \mathrm{Id}) = \nabla\omega_\alpha$. This would only transfer a part of the pressure to the right side of the equations. Exactly this happens if one chooses a nonzero term

$$\phi = \sum\nolimits_\alpha \tilde{\omega}_\alpha u_\alpha$$

in the proof of the free energy inequality (compare [9, $(6.52)_4$]). If we define

$$\bar{\omega} = \sum\nolimits_\alpha \tilde{\omega}_\alpha$$

then, using the representation (5), we have

$$\begin{aligned}
\mathrm{div}\phi &= \sum\nolimits_\alpha u_\alpha \bullet \nabla\tilde{\omega}_\alpha + \mathrm{D}u_\alpha \bullet (\tilde{\omega}_\alpha \mathrm{Id}) \\
&= \sum\nolimits_\alpha u_\alpha \bullet (\nabla\tilde{\omega}_\alpha - \frac{\bar{\omega}}{\bar{\rho}} \nabla\rho_\alpha) + \mathrm{D}v_\alpha \bullet ((\tilde{\omega}_\alpha - \frac{\bar{\omega}}{\bar{\rho}}\rho_\alpha)\mathrm{Id}) \\
&= \sum\nolimits_\alpha u_\alpha \bullet \nabla\omega_\alpha + \mathrm{D}v_\alpha \bullet (\omega_\alpha \mathrm{Id}) \quad \text{if } \omega_\alpha = \tilde{\omega}_\alpha - \frac{\bar{\omega}}{\bar{\rho}}\rho_\alpha,
\end{aligned}$$

since (4) holds. If we would use this in the computation of Sect. 3 we would get the new terms in (20). We remark that then

$$\phi = \sum\nolimits_\alpha \omega_\alpha u_\alpha \text{ with } \sum\nolimits_\alpha \omega_\alpha = 0.$$

5 Examples

We describe three examples, the first is a gradient flow, for which the τ_α are proportional to μ_α, the second is the predator-prey system, for which the τ_α are partially orthogonal to μ_α, and the last one is a cyclic reaction with an intermediate state of τ_α. In all cases the sum of the τ_α values is 0 and the free energy inequality (see (17))

$$\sum_\alpha \tau_\alpha \mu_\alpha \leq 0 \quad \text{where} \quad \mu_\alpha = f'_{\rho_\alpha}(\rho) \tag{21}$$

is satisfied. Besides this we assume for simplicity that the relative velocities u_α are all 0, and that the fluid as a whole is incompressible or, is a rigid body. Then the mass equations reduce to

$$\dot{\rho}_\alpha = \tau_\alpha,$$

where $\dot{\psi} = \partial_t \psi + \bar{v} \bullet \nabla \psi$ for functions ψ. The following considerations generalize to the general case of system (1).

Example 1 (Gradient Flow). For a given free energy function f consider the gradient flow system

$$\dot{\rho}_\alpha = -\lambda \mu_\alpha,$$

where $\lambda \equiv \lambda(\rho) > 0$. Then inequality (21) is satisfied. In this case the sum of the reaction terms does not need to be zero.

In the literature you will find a system consisting of the second and third equation below, the classical Lotka-Volterra system. We refer to [8, Chap. 6] and [11].

Example 2 (Lotka-Volterra System). For the predator-prey model we let $x > 0$ be the number of prey and $y > 0$ the number of predator and consider the system

$$\begin{aligned}
\dot{b} &= -\lambda x, \\
\dot{x} &= x \cdot (\alpha - \beta y), \\
\dot{y} &= -y \cdot (\gamma - \delta x), \\
\dot{z} &= \eta x y, \\
\dot{d} &= \kappa y
\end{aligned}$$

The additional variables are a quantity b proportional to birth of prey, d proportional to death of predator, and z proportional to interactions between predator and prey. This system satisfies the inequality (21), which reduces to

$$\varepsilon \lambda x + \zeta \kappa y + \xi \eta x y \geq 0,$$

if the free energy is given by

$$f \equiv \hat{\hat{f}}(b, x, y, z, d) = -\gamma \log x - \alpha \log y + \delta x + \beta y + \varepsilon b - \zeta d - \xi z,$$

which is a convex function for constants $\gamma > 0$ and $\alpha > 0$. The inequality (21) holds, if in addition the constants ε, ζ, η, λ, κ and ξ satisfy $\varepsilon\lambda > 0$, $\zeta\kappa > 0$, and $\xi\eta > 0$. The remaining quantities β and δ are positive because of biological reasons.

The variables transform into (bio)mass densities by $\rho_b = bm_b$, $\rho_x = xm_x$, $\rho_y = ym_y$, $\rho_d = dm_d$, $\rho_z = zm_z$ with positive mass constants satisfying

$$\lambda = \frac{\alpha m_x}{m_b}, \quad \kappa = \frac{\gamma m_y}{m_d},$$
$$\eta = \frac{\beta m_x - \delta m_y}{m_z}, \tag{22}$$

which implies that the sum of the mass production terms are 0. The parameter η is positive if and only if biomass is lost during transfer from prey to predator.

Proof. It is $f = -\log K + \varepsilon b - \zeta d - \xi z$ with

$$K \equiv \hat{K}(x, y) = \frac{x^{\gamma} y^{\alpha}}{\mathrm{e}^{-\delta x}\mathrm{e}^{-\beta y}}$$

and one computes for solutions of the system that $\dot{K} = 0$, that is, this convex part of f is constant for solutions, and moreover, we see that solutions rotate around the equilibrium

$$x = \frac{\gamma}{\delta}, \quad y = \frac{\alpha}{\beta}.$$

This is the basis for the entire result: For the mass densities the system is

$$\begin{aligned}
\dot{\rho}_b &= \tau_b = m_b \tau_b', \quad \tau_b' := -\lambda x,\\
\dot{\rho}_x &= \tau_x = m_x \tau_x', \quad \tau_x' = x \cdot (\alpha - \beta y),\\
\dot{\rho}_y &= \tau_y = m_y \tau_y', \quad \tau_y' = -y \cdot (\gamma - \delta x),\\
\dot{\rho}_z &= \tau_z = m_z \tau_z', \quad \tau_z' := \eta x y,\\
\dot{\rho}_d &= \tau_d = m_d \tau_d', \quad \tau_d' := \kappa y,
\end{aligned}$$

and, using the identities (22), that is

$$\beta m_x = \delta m_y + \eta m_z, \quad \lambda m_b = \alpha m_x, \quad \kappa m_d = \gamma m_y, \tag{23}$$

we obtain

$$
\begin{aligned}
\tau_b &= -\mathbf{r}_b, \quad \mathbf{r}_b := \lambda m_b x,\\
\tau_x &= \mathbf{r}_b - \mathbf{r}_{xy}, \quad \mathbf{r}_{xy} := cxy, \quad c := \beta m_x,\\
\tau_y &= -\mathbf{r}_d + (1-\omega)\mathbf{r}_{xy}, \quad \omega := \frac{\eta m_z}{c},\\
\tau_z &= \omega \mathbf{r}_{xy}, \quad \big(\omega c = \eta m_z,\, (1-\omega)c = \delta m_y\big)\\
\tau_d &= \mathbf{r}_d, \quad \mathbf{r}_d := \kappa m_d y,
\end{aligned}
$$

hence $\bar{\tau} = 0$. Then one easily computes

$$
\begin{aligned}
\tau_x' \tilde{f}_{\prime x} + \tau_y' \tilde{f}_{\prime y} &= -\frac{1}{K}(\tau_x' K_{\prime x} + \tau_y' K_{\prime y})\\
&= -\frac{1}{K}(\dot{x} K_{\prime x} + \dot{y} K_{\prime y}) = -\frac{1}{K}\dot{K} = 0,
\end{aligned}
$$

and therefore

$$
\begin{aligned}
\textstyle\sum_\beta \tau_\beta \mu_\beta &= \tau_b' \tilde{f}_{\prime b} + \tau_x' \tilde{f}_{\prime x} + \tau_y' \tilde{f}_{\prime y} + \tau_z' \tilde{f}_{\prime z} + \tau_d' \tilde{f}_{\prime d}\\
&= \tau_b' \tilde{f}_{\prime b} + \tau_z' \tilde{f}_{\prime z} + \tau_d' \tilde{f}_{\prime d} = -\varepsilon\lambda x - \zeta\kappa y - \xi\eta x y \le 0. \qquad \square
\end{aligned}
$$

Example 3 (Cyclic Processes). As a last example we consider cyclic reactions, which are important cases and often the basis for biological processes. We have

$$
\begin{aligned}
\dot{\rho}_\alpha &= \tau_\alpha \text{ for } \alpha = 1, \dots, N,\\
\tau_\alpha &:= \eta_{\alpha+1}\rho_{\alpha+1} - \eta_\alpha \rho_\alpha, \qquad (24)
\end{aligned}
$$

with cyclic repetition, $\rho_{N+1} := \rho_1$, $\eta_{N+1} := \eta_1$. Here η_α are positive constants. This system satisfies the inequality (21), if

$$
f \equiv \hat{f}(\rho) = f_0(\bar{\rho}) + b(\bar{\rho}) \textstyle\sum_\alpha \eta_\alpha \rho_\alpha^2
$$

with positive functions $b(\bar{\rho}) > 0$.

The stationary solutions are values $\rho^0 = (\rho_\alpha^0)_\alpha$ with

$$
\eta_{\alpha+1}\rho_{\alpha+1}^0 = \eta_\alpha \rho_\alpha^0 =: \eta^0.
$$

This $\rho^0 \in \mathbb{R}^N$ is a unique point, if the value of $\bar{\rho}^0$ is considered to be given. For general solutions ρ is rotating around the stationary line and converging to a value ρ^0, what can be seen from the free energy. We mention that the sum in the free energy can be written as

$$\sum\nolimits_\alpha \eta_\alpha \rho_\alpha^2 = \sum\nolimits_\alpha \eta_\alpha (\rho_\alpha - \rho_\alpha^0)^2 + 2\eta^0 \bar{\rho} - \sum\nolimits_\alpha \eta_\alpha (\rho_\alpha^0)^2.$$

Moreover, we again obtain overall mass conservation, that is $\bar{\tau} = 0$.

Proof. If $f = f(\bar{\rho}, \rho)$, the derivative with respect to $\bar{\rho}$ has no effect, since the total mass production is zero. Therefore it is enough to consider a free energy

$$f = f(\rho) = \frac{1}{2} \sum\nolimits_\alpha b_\alpha \rho_\alpha^2,$$

so that

$$\mu_\alpha = f'_{\rho_\alpha}(\rho) = b_\alpha \rho_\alpha.$$

Then, with $b_\alpha = \eta_\alpha \tilde{b}_\alpha$ and assuming $\tilde{b}_\alpha > 0$,

$$\begin{aligned}\sum\nolimits_\alpha \tau_\alpha \mu_\alpha &= \sum\nolimits_\alpha (\eta_{\alpha+1}\rho_{\alpha+1} - \eta_\alpha \rho_\alpha) b_\alpha \rho_\alpha \\ &= \sum\nolimits_\alpha \left(\tilde{b}_\alpha (\eta_{\alpha+1}\rho_{\alpha+1})(\eta_\alpha \rho_\alpha) - \tilde{b}_\alpha (\eta_\alpha \rho_\alpha)^2 \right) \\ &= \sum\nolimits_\alpha \left(\sqrt{\frac{\tilde{b}_\alpha}{\tilde{b}_{\alpha+1}}} \cdot \xi_{\alpha+1}\xi_\alpha - \xi_\alpha^2 \right),\end{aligned}$$

where $\xi_\alpha := \eta_\alpha \rho_\alpha \sqrt{\tilde{b}_\alpha}$. Letting

$$c_\alpha := \sqrt{\frac{\tilde{b}_\alpha}{\tilde{b}_{\alpha+1}}}$$

and using $\xi_\alpha \xi_{\alpha+1} \leq \frac{1}{2}(\xi_\alpha^2 + \xi_{\alpha+1}^2)$, this is

$$\begin{aligned}&= \sum\nolimits_\alpha \left(c_\alpha \xi_{\alpha+1}\xi_\alpha - \xi_\alpha^2 \right) \leq \sum\nolimits_\alpha \left(\frac{c_\alpha}{2} \xi_{\alpha+1}^2 + \frac{c_\alpha}{2} \xi_\alpha^2 - \xi_\alpha^2 \right) \\ &= \sum\nolimits_\alpha \left(\frac{c_{\alpha-1}}{2} + \frac{c_\alpha}{2} - 1 \right) \xi_\alpha^2 = 0 \text{ if } c_\alpha = 1,\end{aligned}$$

that is

$$b = \tilde{b}_\alpha = \tilde{b}_{\alpha+1} > 0 \text{ for all } \alpha,$$

or $b_\alpha = \eta_\alpha b$. □

6 Handling Gradient Terms

It is often necessary to consider a gradient dependence of the free energy. For a biological application see for example [6]. In general we consider f to depend on all densities ρ_α and density derivatives $\nabla\rho_\alpha$, that is

$$f \equiv \hat{f}(\rho, \nabla\rho). \tag{25}$$

In particular situations f usually depends only on the gradient of one species or on the gradient of a fraction. Both are special cases of (25). In analogy to Sect. 3 we state a version of Theorem 1 but now with the following chemical potentials

$$\mu_\alpha := \frac{\delta f}{\delta\rho_\alpha} = f_{\prime\rho_\alpha} - \operatorname{div}(f_{\prime\nabla\rho_\alpha}), \tag{26}$$

and the following generalization of the specific pressures

$$\begin{aligned} P_\alpha^{sp} :=& \Big(\frac{\delta f}{\delta\rho_\alpha} - \frac{f}{\bar\rho}\Big)\mathrm{Id} + \sum\nolimits_\beta \frac{1}{\bar\rho}\nabla\rho_\beta \otimes f_{\prime\nabla\rho_\beta} \\ =& \bar\rho\frac{\delta f^{sp}}{\delta\rho_\alpha}\mathrm{Id} + \sum\nolimits_\beta \nabla\rho_\beta \otimes f^{sp}_{\prime\nabla\rho_\beta} \end{aligned} \tag{27}$$

with $f = \bar\rho f^{sp}$ as usual. (If f does not depend on the gradients, the matrix P_α^{sp} will reduce to $p_\alpha^{sp}\mathrm{Id}$.) With these definitions the following holds

Proposition 6. *If the free energy is given by (25) then we obtain for the free energy production*

$$\begin{aligned} h = & \operatorname{div}\Big(\phi + \sum\nolimits_\alpha \dot\rho_\alpha f_{\prime\nabla\rho_\alpha}\Big) \\ & + \sum\nolimits_\alpha \tau_\alpha\mu_\alpha + \sum\nolimits_\alpha u_\alpha\bullet\Big(g_\alpha + \frac{\tau_\alpha}{2}u_\alpha - P_\alpha^{sp}\nabla\rho_\alpha\Big) \\ & + \sum\nolimits_\alpha \mathrm{D}v_\alpha\bullet\Big(\varPi_\alpha - \rho_\alpha P_\alpha^{sp}\Big). \end{aligned}$$

Proof. The proof follows the one of Proposition 4, but now we have to use the identity

$$(\partial_j\rho_\alpha)^{\bullet} = (\partial_t + \bar v\bullet\nabla)\partial_j\rho_\alpha = \partial_j\dot\rho_\alpha - (\partial_j\bar v)\bullet\nabla\rho_\alpha,$$

hence for $z \in \mathbb{R}^n$

$$(\nabla\rho_\alpha)^{\bullet}\bullet z = (\nabla\dot\rho_\alpha)\bullet z - \mathrm{D}\bar v\bullet(\nabla\rho_\alpha \otimes z). \tag{28}$$

We therefore compute

$$
\begin{aligned}
\partial_t f + \operatorname{div}(f\bar{v}) &= \dot{f} + f\operatorname{div}\bar{v}\\
&= \sum\nolimits_\alpha f_{'\rho_\alpha}\dot{\rho}_\alpha + \sum\nolimits_\alpha f_{'\nabla\rho_\alpha}\bullet(\dot{\nabla\rho_\alpha}) + f\operatorname{div}\bar{v}\\
&= \sum\nolimits_\alpha (f_{'\rho_\alpha}\dot{\rho}_\alpha + f_{'\nabla\rho_\alpha}\bullet\nabla\dot{\rho}_\alpha) + \mathrm{D}\bar{v}\bullet\Big(f\mathrm{Id} - \sum\nolimits_\alpha \nabla\rho_\alpha\otimes f_{'\nabla\rho_\alpha}\Big)\\
&= \operatorname{div}\Big(\sum\nolimits_\alpha \dot{\rho}_\alpha f_{'\nabla\rho_\alpha}\Big) + \sum\nolimits_\alpha \frac{\delta f}{\delta\rho_\alpha}\cdot\dot{\rho}_\alpha\\
&\quad + \mathrm{D}\bar{v}\bullet\Big(f\mathrm{Id} - \sum\nolimits_\alpha \nabla\rho_\alpha\otimes f_{'\nabla\rho_\alpha}\Big)
\end{aligned}
$$

and

$$
\begin{aligned}
\sum\nolimits_\alpha \frac{\delta f}{\delta\rho_\alpha}\cdot\dot{\rho}_\alpha &= \sum\nolimits_\alpha \mu_\alpha\cdot\Big(\tau_\alpha - \operatorname{div}(\rho_\alpha v_\alpha) + \bar{v}\bullet\nabla\rho_\alpha\Big)\\
&= \sum\nolimits_\alpha \mu_\alpha\tau_\alpha - \sum\nolimits_\alpha \mu_\alpha u_\alpha\bullet\nabla\rho_\alpha - \sum\nolimits_\alpha \mathrm{D}v_\alpha\bullet(\rho_\alpha\mu_\alpha\mathrm{Id}).
\end{aligned}
$$

We plug this into the expression for h in Proposition 3 and obtain

$$
\begin{aligned}
0 \geq h &= \partial_t f + \operatorname{div}(f\bar{v} + \phi)\\
&\quad + \sum\nolimits_\alpha \Big(\mathrm{D}v_\alpha\bullet\Pi_\alpha + u_\alpha\bullet g_\alpha + \frac{\tau_\alpha}{2}|u_\alpha|^2\Big)\\
&= \operatorname{div}\Big(\phi + \sum\nolimits_\alpha \dot{\rho}_\alpha f_{'\nabla\rho_\alpha}\Big) + \sum\nolimits_\alpha \mu_\alpha\tau_\alpha\\
&\quad + \sum\nolimits_\alpha u_\alpha\bullet g_\alpha + \sum\nolimits_\alpha \frac{\tau_\alpha}{2}|u_\alpha|^2 + R
\end{aligned}
$$

with

$$
\begin{aligned}
R = &-\sum\nolimits_\alpha \mu_\alpha u_\alpha\bullet\nabla\rho_\alpha + \mathrm{D}\bar{v}\bullet\Big(f\mathrm{Id} - \sum\nolimits_\alpha \nabla\rho_\alpha\otimes f_{'\nabla\rho_\alpha}\Big)\\
&+ \sum\nolimits_\alpha \mathrm{D}v_\alpha\bullet(\Pi_\alpha - \rho_\alpha\mu_\alpha\mathrm{Id}).
\end{aligned}
$$

Using formula (5) for $\mathrm{D}\bar{v}$ this equation for R becomes

$$
\begin{aligned}
R &= \sum\nolimits_\alpha (u_\alpha\otimes\nabla\rho_\alpha)\bullet\Big(-\mu_\alpha\mathrm{Id} + \frac{1}{\bar{\rho}}\Big(f\mathrm{Id} - \sum\nolimits_\beta \nabla\rho_\beta\otimes f_{'\nabla\rho_\beta}\Big)\Big)\\
&\quad + \sum\nolimits_\alpha \mathrm{D}v_\alpha\bullet\Big(\Pi_\alpha - \rho_\alpha\mu_\alpha\mathrm{Id} + \frac{\rho_\alpha}{\bar{\rho}}\Big(f\mathrm{Id} - \sum\nolimits_\beta \nabla\rho_\beta\otimes f_{'\nabla\rho_\beta}\Big)\Big)\\
&= \sum\nolimits_\alpha (u_\alpha\otimes\nabla\rho_\alpha)\bullet P_\alpha^{sp} + \sum\nolimits_\alpha \mathrm{D}v_\alpha\bullet(\Pi_\alpha - \rho_\alpha P_\alpha^{sp}),
\end{aligned}
$$

and $(u_\alpha\otimes\nabla\rho_\alpha)\bullet P_\alpha^{sp} = u_\alpha\bullet(P_\alpha^{sp}\nabla\rho_\alpha)$. □

Then we obtain the following version of Theorem 1 as a consequence.

Theorem 3. *Let*

$$f \equiv \hat{f}(\rho, \nabla\rho), \quad \phi := -\sum\nolimits_\alpha \dot{\rho}_\alpha f_{/\nabla\rho_\alpha}. \tag{29}$$

Suppose that

$$\begin{aligned} \Pi_\alpha &= P_\alpha \mathrm{Id} - S_\alpha, \quad P_\alpha = \rho_\alpha P_\alpha^{sp}, \\ g_\alpha &= P_\alpha^{sp} \nabla\rho_\alpha - \frac{\tau_\alpha}{2} u_\alpha + g_\alpha^{fr} - \rho_\alpha g^{sp}, \end{aligned} \tag{30}$$

then for solutions of (1) the free energy production h reads

$$0 \geq h = -\sum\nolimits_\alpha \mathrm{D}v_\alpha \bullet S_\alpha + \sum\nolimits_\alpha \tau_\alpha \mu_\alpha + \sum\nolimits_\alpha g_\alpha^{fr} \bullet u_\alpha. \tag{31}$$

The result is the same as in Sect. 3, where only the scalar p_α is replaced by the matrix P_α. It follows directly from Proposition 6 where now in the free energy inequality the new chemical potentials μ_α from (26) are used. We remark that also now the term $P_\alpha^{sp} \nabla\rho_\alpha$ in the momentum equation cancels since

$$\mathrm{div}(\rho_\alpha P_\alpha^{sp}) = P_\alpha^{sp} \nabla\rho_\alpha + \rho_\alpha \mathrm{div} P_\alpha^{sp}.$$

In summary we arrive at the following conclusion.

Proposition 7. *Under the above assumptions the mixture system (1) is*

$$\begin{aligned} &\partial_t \rho_\alpha + \mathrm{div}(\rho_\alpha v_\alpha) = \tau_\alpha, \\ &\rho_\alpha \big(\partial_t v_\alpha + (v_\alpha \bullet \nabla) v_\alpha\big) \\ &\qquad = \mathrm{div} S_\alpha - \rho_\alpha (\mathrm{div} P_\alpha^{sp} + g^{sp}) - \frac{\tau_\alpha}{2} u_\alpha + g_\alpha^{fr} + \mathbf{f}_\alpha \end{aligned}$$

for all α. The free energy inequality (31) *is satisfied.*

Again a statement like Theorem 2 holds, if in the momentum equation one considers the term $\rho_\alpha(\mathrm{div} P_\alpha^{sp} + \nabla f^{sp})$.

We now go back to the standard case p_α^{sp}.

7 Polymerization of Actin Filaments

We consider a four component system, a reactive polymer-solvent mixture. The mass densities are ρ_m for actin monomers, ρ_a for polymerized actin filaments, ρ_c for cross-linked actin filaments, and the mass density ρ_s for the solvent. We consider the following conservation laws

$$\begin{aligned}
\partial_t \rho_c + \operatorname{div}(\rho_c v_c) &= \tau_c := -\mathbf{r}_c, \\
\partial_t \rho_a + \operatorname{div}(\rho_a v_a) &= \tau_a := \mathbf{r}_c - \mathbf{r}_a, \\
\partial_t \rho_m + \operatorname{div}(\rho_m v_m) &= \tau_m := \mathbf{r}_a, \\
\partial_t \rho_s + \operatorname{div}(\rho_s v_s) &= \tau_s := 0.
\end{aligned} \tag{32}$$

Here the reactions are given by

$$\begin{aligned}
\mathbf{r}_a &= \lambda_a(\rho)\big(\eta_a \rho_a - \nu \rho_m\big), \\
\mathbf{r}_c &= \lambda_c(\rho)\Big(\eta_c \rho_c - \chi \frac{\rho_a^2}{K^2 + \rho_a^2}\Big),
\end{aligned} \tag{33}$$

where η_a, η_c, χ, ν, and K are assumed to be positive and constant, and λ_a and λ_c are positive functions. Obviously the sum of the mass productions

$$\bar{\tau} = \sum\nolimits_{\alpha = m,a,c,s} \tau_\alpha = 0.$$

The following theorem shows the existence of a free energy. We emphasize that there might by a different free energy also satisfying the free energy inequality. It is important for the dynamics of the system which free energy one chooses.

Theorem 4. *With $\rho = (\rho_m, \rho_a, \rho_c, \rho_s)$ a possible free energy is defined by*

$$f(\rho) = \frac{\eta_c}{2}\rho_c^2 + \chi \psi_{K_m}(\rho_m) + \chi \psi_{K_a}(\rho_a) + f_s(\rho_s),$$

where $\nu K_m = \eta_a K_a$ and

$$\psi_K(z) = z - K \arctan\Big(\frac{z}{K}\Big) \text{ for } z \in \mathbb{R}.$$

The function f_s is an arbitrary (convex) function. Then

$$\sum\nolimits_\beta \tau_\beta \mu_\beta \le 0.$$

Therefore this part of the free energy inequality is satisfied (compare Lemma 4).

Proof. It is

$$\psi_K'(z) = 1 - \frac{1}{1 + (\frac{z}{K})^2} = \frac{z^2}{K^2 + z^2},$$

therefore we obtain

$$\begin{aligned}
\mu_c &= \eta_c \rho_c, \\
\mu_a &= \chi \frac{\rho_a^2}{K_a^2 + \rho_a^2} \\
\mu_m &= \chi \frac{\rho_m^2}{K_m^2 + \rho_m^2} = \chi \frac{\left(\frac{\nu}{\eta_a}\rho_m\right)^2}{K_a^2 + \left(\frac{\nu}{\eta_a}\rho_m\right)^2}.
\end{aligned} \tag{34}$$

We then compute

$$\begin{aligned}
&\sum\nolimits_{\beta=c,a,m,s} \tau_\beta \mu_\beta = \tau_c \mu_c + \tau_a \mu_a + \tau_m \mu_m \\
&= -\mathbf{r}_c \mu_c + (\mathbf{r}_c - \mathbf{r}_a)\mu_a + \mathbf{r}_a \mu_m = \mathbf{r}_c \cdot (\mu_a - \mu_c) + \mathbf{r}_a \cdot (\mu_m - \mu_a),
\end{aligned}$$

and obtain

$$\begin{aligned}
&-\mathbf{r}_c \cdot (\mu_a - \mu_c) = \lambda_c(\rho) \cdot \Big(\eta_c \rho_c - \chi \frac{\rho_a^2}{K^2 + \rho_a^2}\Big)^2 \geq 0, \\
&-\mathbf{r}_a \cdot (\mu_m - \mu_a) = \lambda_a(\rho) \eta_a \chi \cdot \big(\frac{\nu}{\eta_a}\rho_m - \rho_a\big)\Big(\frac{\left(\frac{\nu}{\eta_a}\rho_m\right)^2}{K_a^2 + \left(\frac{\nu}{\eta_a}\rho_m\right)^2} - \frac{\rho_a^2}{K_a^2 + \rho_a^2}\Big) \geq 0.
\end{aligned}$$

This shows the result. □

One can also define a different free energy by applying a given monotone function to the definitions of the chemical potentials in (34). There is another point to be mentioned. The proof above shows that each reaction, $\mathbf{r}_c$ and $\mathbf{r}_a$, gives a nonpositive contribution to the free energy production as it is common in chemical processes. But this is in contrast to the proofs of the Examples 2 and 3, indicating that the situation in biological processes is generally more complex.

8 The Quasistatic Problem

The momentum equation of the α-component contains the term Π_α under the divergence and the term g_α on the right side. By (16) these two terms have the representation

$$\begin{aligned}
\Pi_\alpha &= \rho_\alpha p_\alpha^{sp} \mathrm{Id} - S_\alpha, \\
g_\alpha &= p_\alpha^{sp} \nabla \rho_\alpha - \rho_\alpha g^{sp} - \frac{\tau_\alpha}{2} u_\alpha + g_\alpha^{fr}.
\end{aligned} \tag{35}$$

It has been shown in the previous sections that system (1) is equivalent to

$$\begin{aligned}&\partial_t \rho_\alpha + \operatorname{div}(\rho_\alpha v_\alpha) = \tau_\alpha,\\ &\rho_\alpha(\partial_t v_\alpha + (v_\alpha \bullet \nabla) v_\alpha) = \operatorname{div} S_\alpha - \rho_\alpha(\nabla p_\alpha^{sp} + g^{sp}) - \frac{\tau_\alpha}{2} u_\alpha + g_\alpha^{fr} + \mathbf{f}_\alpha.\end{aligned} \tag{36}$$

We suppose that the terms on the right side of these equations have the property in Lemma 4 which is the free energy inequality

$$0 \geq h = -\sum\nolimits_\alpha \mathrm{D}v_\alpha \bullet S_\alpha + \sum\nolimits_\alpha \tau_\alpha \mu_\alpha + \sum\nolimits_\alpha g_\alpha^{fr} \bullet u_\alpha.$$

In biological systems one often has the situation that some of the terms in the differential equation have large coefficients compared to the others, for example

- The stress tensor and the pressure and the friction.
- The pressure and the friction (see Sect. 10).
- The pressure and the friction and the external force (for example in a rotating cylinder).

In the first case, for example as $\varepsilon \searrow 0$,

$$\varepsilon \cdot S_{\varepsilon\alpha} \to S_\alpha, \quad \varepsilon \cdot f_\varepsilon \to f, \quad \varepsilon \cdot g_{\varepsilon\alpha}^{fr} \to g_\alpha^{fr},$$

then it follows also that $\varepsilon \cdot p_{\varepsilon\alpha} \to p_\alpha$ (from (15)) and $\varepsilon \cdot g_\varepsilon^{sp} \to g^{sp}$ (at least for closed systems), whereas the other coefficients stay bounded and have a limit. The solutions $(\rho_\varepsilon, v_\varepsilon)$ of the ε-problem satisfying the system (1)

$$\begin{aligned}&\partial_t \rho_{\varepsilon\alpha} + \operatorname{div}(\rho_{\varepsilon\alpha} v_{\varepsilon\alpha}) = \tau_{\varepsilon\alpha},\\ &\partial_t(\rho_{\varepsilon\alpha} v_{\varepsilon\alpha}) + \operatorname{div}(\rho_{\varepsilon\alpha} v_{\varepsilon\alpha} \otimes v_{\varepsilon\alpha} + \Pi_{\varepsilon\alpha}) = g_{\varepsilon\alpha} + \tau_{\varepsilon\alpha} v_{\varepsilon\alpha} + \mathbf{f}_{\varepsilon\alpha}\end{aligned}$$

converge in the limit $(\rho_\varepsilon, v_\varepsilon) \to (\rho, v)$ and satisfy a reduced problem

$$\begin{aligned}&\partial_t \rho_\alpha + \operatorname{div}(\rho_\alpha v_\alpha) = \tau_\alpha,\\ &\operatorname{div} \Pi_\alpha = g_\alpha\end{aligned} \tag{37}$$

for all α. Alternatively, the equivalent system in Proposition 5 for $(\rho_\varepsilon, v_\varepsilon)$ leads to the reduced problem for the limit (ρ, v)

$$\begin{aligned}&\partial_t \rho_\alpha + \operatorname{div}(\rho_\alpha v_\alpha) = \tau_\alpha,\\ &0 = \operatorname{div} S_\alpha - \rho_\alpha(\nabla p_\alpha^{sp} + g^{sp}) + g_\alpha^{fr}.\end{aligned} \tag{38}$$

One also has in the limit

$$\bar{g} = \bar{\rho}(\nabla f^{sp} - g^{sp}) + \bar{g}^{fr}. \tag{39}$$

In addition one has to consider the limit in the free energy inequality. From the inequality

$$0 \geq \varepsilon h_\varepsilon = -\sum\nolimits_\alpha \mathrm{D}v_{\varepsilon\alpha} \bullet (\varepsilon S_{\varepsilon\alpha}) + \sum\nolimits_\alpha \tau_{\varepsilon\alpha}(\varepsilon\mu_{\varepsilon\alpha}) + \sum\nolimits_\alpha (\varepsilon g^{fr}_{\varepsilon\alpha}) \bullet u_{\varepsilon\alpha}$$

one obtains, that the limit $\varepsilon h_\varepsilon \to h^{red}$ exists with

$$0 \geq h^{red} = -\sum\nolimits_\alpha \mathrm{D}v_\alpha \bullet S_\alpha + \sum\nolimits_\alpha \tau_\alpha \mu_\alpha + \sum\nolimits_\alpha g^{fr}_\alpha \bullet u_\alpha. \tag{40}$$

In this connection we refer to [4] and [5] where a limit entropy inequality is considered. In [1] the second author treats a functional, whose first variation with respect to v are the quasistatic momentum equations.

Remark 2 (Quasistatic functional). Consider a function $J = \hat{J}(\rho, v, \nabla\rho, \mathrm{D}v)$ which satisfies

$$\frac{\delta J}{\delta v_\alpha} = \mathrm{div}\Pi_\alpha - g_\alpha, \tag{41}$$

where $\frac{\delta J}{\delta v_\alpha} := J_{'v_\alpha} - \mathrm{div} J_{'\mathrm{D}v_\alpha}$, which requires some assumptions on Π_α and g_α. Concerning the dependence on $(\mathrm{D}v)^{\mathrm{S}}$, the free energy inequality in the form of Lemma 4 is equivalent to the convexity of J in $(\mathrm{D}v)^{\mathrm{S}}$.

9 Fractional Densities

Often in mixture models one is confronted with the fractional densities

$$\theta_\alpha := \frac{\rho_\alpha}{\bar{\rho}} = \frac{\rho_\alpha}{\rho_1 + \cdots + \rho_N}. \tag{42}$$

Then instead of the variables $(\rho_\alpha)_\alpha$ one can use as new variables $(\bar{\rho}, (\theta_\alpha)_\alpha)$, where one has the side condition

$$\sum\nolimits_\alpha \theta_\alpha = 1. \tag{43}$$

For the mass equations the following lemma holds.

Lemma 5. *The N mass equations in (1) are equivalent to*

$$\begin{aligned} &\partial_t \bar{\rho} + \mathrm{div}(\bar{\rho}\,\bar{v}) = \bar{\tau}, \\ &\bar{\rho}\dot{\theta}_\alpha + \mathrm{div}(\bar{\rho}\theta_\alpha u_\alpha) = \tau_\alpha - \theta_\alpha \bar{\tau} \text{ for all } \alpha, \end{aligned}$$

where $\dot{\psi} := \partial_t \psi + \bar{v} \bullet \nabla \psi$ *for any function* ψ*. These again are* N *independent equations, the sum of the* α*-equations is* 0.

Proof. One obtains the new equations, if one subtracts from the old one θ_α times the equation for the sum. Thus

$$\begin{aligned}
&\tau_\alpha - \theta_\alpha \bar{\tau} \\
&= \partial_t \rho_\alpha - \theta_\alpha \partial_t \bar{\rho} + \operatorname{div}(\rho_\alpha v_\alpha) - \theta_\alpha \operatorname{div}(\bar{\rho}\bar{v}) \\
&= \partial_t (\bar{\rho}\theta_\alpha) - \theta_\alpha \partial_t \bar{\rho} + \operatorname{div}(\bar{\rho}\theta_\alpha (v_\alpha - \bar{v})) + \bar{\rho}\bar{v} \bullet \nabla \theta_\alpha \\
&= \bar{\rho}\partial_t \theta_\alpha + \operatorname{div}(\bar{\rho}\theta_\alpha u_\alpha) + \bar{\rho}\bar{v} \bullet \nabla \theta_\alpha .
\end{aligned}$$

Because of (43) and (4) the sum of these equations is equal to zero. □

One could use the same procedure for the momentum equations. We will do this here for the quasistatic case, that is, for system $\operatorname{div} \Pi_\alpha = g_\alpha$ in (37). We obtain the following theorem.

Theorem 5. *In the quasistatic regime the Eqs. (37) are equivalent to the equations in Lemma 5 and*

$$\begin{aligned}
&\operatorname{div} \bar{\Pi} = \bar{g}, \\
&\operatorname{div}(\Pi_\alpha - \theta_\alpha \bar{\Pi}) = g_\alpha - \theta_\alpha \bar{g} - \bar{\Pi} \nabla \theta_\alpha \text{ for all } \alpha .
\end{aligned}$$

Here we use

$$\begin{aligned}
\Pi_\alpha &= \rho_\alpha p_\alpha^{sp} \mathrm{Id} - S_\alpha , \\
g_\alpha &= p_\alpha^{sp} \nabla \rho_\alpha - \rho_\alpha g^{sp} + g_\alpha^{fr} .
\end{aligned}$$

and the new definitions of $\bar{\Pi}$ *and* $\bar{g}$ *in Eq. (44) below.*

Proof. The procedure for the momentum equation says

$$g_\alpha - \theta_\alpha \bar{g} = \operatorname{div} \Pi_\alpha - \theta_\alpha \operatorname{div} \bar{\Pi},$$

where now

$$\bar{\Pi} = \sum\nolimits_\alpha \Pi_\alpha , \quad \bar{g} = \sum\nolimits_\alpha g_\alpha . \tag{44}$$

With this

$$\operatorname{div} \Pi_\alpha - \theta_\alpha \operatorname{div} \bar{\Pi} = \operatorname{div}(\Pi_\alpha - \theta_\alpha \bar{\Pi}) + \bar{\Pi} \nabla \theta_\alpha .$$

Hence the equation becomes

$$\mathrm{div}(\Pi_\alpha - \theta_\alpha \bar{\Pi}) = g_\alpha - \theta_\alpha \bar{g} - \bar{\Pi}\nabla\theta_\alpha . \qquad \square$$

Since $\bar{\Pi} = \bar{p}\mathrm{Id} - \bar{S}$ and

$$\bar{g} = \bar{\rho}(\nabla f^{sp} - g^{sp}) + \bar{g}^{fr}, \quad \bar{p} = \sum\nolimits_\alpha \rho_\alpha p_\alpha^{sp}, \tag{45}$$

we can write this equation also as

$$\begin{aligned}
&- \mathrm{div}(S_\alpha - \theta_\alpha \bar{S}) \\
&= g_\alpha - \theta_\alpha \bar{g} + \bar{S}\nabla\theta_\alpha - \mathrm{div}(\rho_\alpha p_\alpha^{sp}\mathrm{Id}) + \theta_\alpha \mathrm{div}(\bar{p}\mathrm{Id}) \\
&= p_\alpha^{sp}\nabla\rho_\alpha - \rho_\alpha \bar{g}^{sp} + g_\alpha^{fr} \\
&\quad - \theta_\alpha \sum\nolimits_\beta p_\beta^{sp}\nabla\rho_\beta + \bar{\rho}\theta_\alpha g^{sp} - \theta_\alpha \bar{g}^{fr} \\
&\quad - \nabla(\rho_\alpha p_\alpha^{sp}) + \theta_\alpha \nabla\bar{p} + \bar{S}\nabla\theta_\alpha \\
&= -\rho_\alpha \nabla p_\alpha^{sp} + \theta_\alpha \sum\nolimits_\beta \rho_\beta \nabla p_\beta^{sp} + g_\alpha^{fr} - \theta_\alpha \bar{g}^{fr} + \bar{S}\nabla\theta_\alpha \\
&= -\bar{\rho}\theta_\alpha \big(\nabla p_\alpha^{sp} - \sum\nolimits_\beta \theta_\beta \nabla p_\beta^{sp}\big) + g_\alpha^{fr} - \theta_\alpha \bar{g}^{fr} + \bar{S}\nabla\theta_\alpha .
\end{aligned}$$

Since $p_\alpha^{sp} = \mu_\alpha - f^{sp}$ we can replace

$$\nabla p_\alpha^{sp} - \sum\nolimits_\beta \theta_\beta \nabla p_\beta^{sp} = \nabla\mu_\alpha - \sum\nolimits_\beta \theta_\beta \nabla\mu_\beta .$$

The equation $\mathrm{div}\bar{\Pi} = \bar{g}$ becomes $0 = \mathrm{div}\bar{S} + \bar{g} - \nabla\bar{p}$ and in the quasistatic case $\bar{g}$ is given by (39).

A different equivalent version of the system in Theorem 5 is

$$\begin{aligned}
&\mathrm{div}\bar{S} = \nabla\bar{p} - \bar{\rho}\big(\nabla f^{sp} - g^{sp}\big) - \bar{g}^{fr}, \\
&\mathrm{div}(S_\alpha - \theta_\alpha \bar{S}) = \bar{\rho}\theta_\alpha \big(\nabla\mu_\alpha - \sum\nolimits_\beta \theta_\beta \nabla\mu_\beta\big) - (g_\alpha^{fr} - \theta_\alpha \bar{g}^{fr}) - \bar{S}\nabla\theta_\alpha
\end{aligned} \tag{46}$$

for all α. The sum of the α-equations is zero. Here we mention that the following identity for the entire pressure

$$\bar{\rho}\sum\nolimits_\beta \theta_\beta \nabla\mu_\beta = \nabla\bar{p}$$

holds, if f is a function of ρ alone.

10 Diffusion Limit

Usual biochemical and cell-biological situations are characterized by a relatively low Reynolds number and relatively high friction, so that the quasi-steady-state hypothesis can be assumed (see Sect. 8). Then the corresponding force balance equations yield a generalized system of Darcy type equations, see (51). To derive this we assume the following form of friction forces

$$g_\alpha^{fr} = -\sum\nolimits_\beta \gamma_{\alpha\beta} u_\beta, \quad \gamma_{\alpha\beta} \equiv \hat{\gamma}_{\alpha\beta}(\rho), \tag{47}$$

where the $\gamma_{\alpha\beta}$ are called friction coefficients. Also the free energy f is relatively high, and we assume that there is no mass exchange, that is $\tau_\alpha = 0$, and no viscosity, that is $S_\alpha = 0$. Then the system (38) with g^{sp} as in Lemma 3 is equivalent to

$$\begin{aligned} &\partial_t \rho_\alpha + \operatorname{div}(\rho_\alpha \bar{v} + \rho_\alpha u_\alpha) = 0, \\ &\rho_\alpha \nabla \mu_\alpha = g_\alpha^{fr} - \frac{\rho_\alpha}{\bar{\rho}} \bar{g}^{fr} \end{aligned} \tag{48}$$

for all α. The free energy inequality (40) reduces to

$$0 \geq h^{red} = \sum\nolimits_\alpha g_\alpha^{fr} \bullet u_\alpha = -\sum\nolimits_{\alpha\beta} \gamma_{\alpha\beta} u_\beta \bullet u_\alpha, \tag{49}$$

and this is satisfied, if $\left(\gamma_{\alpha\beta}\right)_{\alpha\beta}$ is positive semidefinite. If

$$\sum\nolimits_{\alpha\beta} \gamma_{\alpha\beta} u_\beta = 0, \tag{50}$$

that is, $\bar{g}^{fr} = 0$, the system (48) is

$$\begin{aligned} &\partial_t \rho_\alpha + \operatorname{div}(\rho_\alpha \bar{v} + \rho_\alpha u_\alpha) = 0, \\ &-\rho_\alpha \nabla \mu_\alpha = \sum\nolimits_\beta \gamma_{\alpha\beta} u_\beta. \end{aligned} \tag{51}$$

These equations are of Darcy's type. Eventually this can be used to compute the relative velocities u_α explicitly, so that by substitution into the mass equations one obtains a system of diffusion equations. This procedure can be used to derive from the general system (1) a single momentum equation for $\bar{v}$ and diffusion equations for the α-components containing the velocity, see (51).

The Eq. (50) is satisfied for the following example. We choose the following form of friction forces

$$g_\alpha^{fr} = -\tilde{\gamma} \rho_\alpha u_\alpha - \sum\nolimits_{\beta \neq \alpha} \tilde{\gamma}_{\alpha\beta} \rho_\alpha \rho_\beta (u_\alpha - u_\beta) \tag{52}$$

with a nonnegative friction coefficient $\tilde{\gamma}$ and nonnegative drag coefficients $\tilde{\gamma}_{\alpha\beta}$ being symmetric in α and β. Then g_α^{fr} has the properties (49) and (50). The property (49) follows from

$$\sum\nolimits_\alpha g_\alpha^{fr} \bullet u_\alpha = -\tilde{\gamma} \sum\nolimits_\alpha \rho_\alpha |u_\alpha|^2 - \frac{1}{2} \sum\nolimits_{\beta \neq \alpha} \tilde{\gamma}_{\alpha\beta} \rho_\alpha \rho_\beta |u_\alpha - u_\beta|^2.$$

For example, such friction forces could appear for a mixture of polymers in a solvent.

11 Polymer Mixtures Including Gradients

Consider general mixtures of polymers being of a similar type but attaining different configuration states. In [3] we have treated a biophysical two-component system of lipid monolayers in lung alveoli. They can consist of ordered lipid clusters, but the lipids can also be in a diffusive unordered phase. In such a model the free energy would typically be a function of the volume fractions of the mixture components, see Sect. 9, and on the partial gradients

$$\nabla\theta_\alpha = \nabla\left(\frac{\rho_\alpha}{\bar{\rho}}\right) = \sum\nolimits_\beta (\rho_\beta \nabla\rho_\alpha - \rho_\alpha \nabla\rho_\beta).$$

For more details compare [12, Sect. 3]. Then, the free energy is of the general type as in (25), namely

$$f \equiv f(\rho, \nabla\rho) = \tilde{f}(\bar{\rho}, \theta, \nabla\theta).$$

Using Sect. 6 for this free energy, we derive the balance equations in Proposition 7 and in the quasistatic case the system (46). However in [3] this free energy was considered in a one momentum system. Section 10 leads to a connection of these two approaches.

12 Conclusion

We have considered a mixture of fluids in the isothermal case, for which we successfully developed a theory based on the free energy inequality. Also a theory with gradients has been presented, but its comparison with various biological problems, see [6], still has to be made. In principle the dependence of the free energy on other quantities is possible.

The theory has been applied to several biological problems, and it turns out that the complexity of biological systems goes beyond the well-known methods for chemical reactions. In biological problems the free energy inequality comprises several reaction terms as the example of Lotka-Volterra system shows (there $\dot{K} = 0$ and K contains two reaction terms).

This paper is a short version of a more detailed elaboration to be published later.

Acknowledgements Wir möchten uns an dieser Stelle bei der DFG bedanken, die in den Sonderforschungsbereichen SBF123 "Stochastische Mathematische Modelle", SFB72 "Approximation und Mathematische Optimierung", SFB256 "Nichtlineare Partielle Differentialgleichungen" und SFB611 "Singuläre Phänomene und Skalierung in mathematischen Modellen" für eine ausgezeichnete Umgebung für Mathematiker gesorgt hat. Dies hat uns das systematische Forschen auf dem Gebiet der Partiellen Differentialgleichungen erst ermöglicht.

References

1. Alt, W.: Nonlinear hyperbolic systems of generalized Navier-Stokes type for interactive motion in biology. In: Hildebrandt, S., Karcher, H. (eds.) Geometric Analysis and Nonlinear Partial Differential Equations, pp. 431–461. Springer, Berlin/New York (2003)
2. Alt, H.W.: The entropy principle for interfaces. Solids and fluids. Adv. Math. Sci. Appl. **19**, 585–663 (2009)
3. Alt, H.W., Alt, W.: Phase boundary dynamics: transition between ordered and disordered lipid monolayers. Interfaces Free Bound. **11**, 1–36 (2009)
4. Alt, H.W., Witterstein, G.: Distributional equation in the limit of phase transition. Interfaces Free Bound. **13**, 531–554 (2011)
5. Alt, H.W., Witterstein, G.: Free energy identity in the limit of phase transitions. Adv. Math. Sci. Appl. (2013, submitted)
6. Chatelain, C., Balois, T., Ciarletta, P., Ben Amar, M.: Emergence of microstructural patterns in skin cancer: a phase separation analysis in a binary mixture. New J. Phys. **13**, 115013 (2011)
7. de Groot, S.R., Mazur, P.: Non-Equilibrium Thermodynamics. North-Holland, Amsterdam (1962)
8. Metzler, W.: Dynamische Systeme in der Ökologie. Mathematische Modelle und Simulationen. Teubner, Stuttgart (1987) (In particular: Kap. 6 Räuber-Beute-Systeme)
9. Müller, I.: Thermodynamics of mixtures of non-viscous fluids (Chap. 6). In: Thermodynamics. Pitman, Boston (1985)
10. Rajagopal, K.R., Johnson, G., Massoudi, M.: Averaged Equations for an Isothermal, Developing Flow of a Fluid-Solid Mixture. DOE/PETC/TR-96/2 (Mar 1996)
11. Wikipedia: Lotka-Volterra equation. http://en.wikipedia.org/wiki/Lotka-Volterra_equation
12. Wittenfeld, A., Ryskin, A., Alt, W.: Modeling and simulation of lipid monolayers as surfactant in lung alveoli. This volume, pp. 171–189 (2013)

A Nested Variational Time Discretization for Parametric Anisotropic Willmore Flow

Ricardo Perl, Paola Pozzi, and Martin Rumpf

Abstract A variational time discretization of anisotropic Willmore flow combined with a spatial discretization via piecewise affine finite elements is presented. Here, both the energy and the metric underlying the gradient flow are anisotropic, which in particular ensures that Wulff shapes are invariant up to scaling under the gradient flow. In each time step of the gradient flow a nested optimization problem has to be solved. Thereby, an outer variational problem reflects the time discretization of the actual Willmore flow and involves an approximate anisotropic L^2-distance between two consecutive time steps and a fully implicit approximation of the anisotropic Willmore energy. The anisotropic mean curvature needed to evaluate the energy integrand is replaced by the time discrete, approximate speed from an inner, fully implicit variational scheme for anisotropic mean curvature motion. To solve the nested optimization problem a Newton method for the associated Lagrangian is applied. Computational results for the evolution of curves underline the robustness of the new scheme, in particular with respect to large time steps.

1 Introduction

This paper generalizes a recently proposed variational time discretization [1] for isotropic Willmore flow to the corresponding anisotropic flow. Thereby, the anisotropic Willmore flow is defined as the gradient flow of the anisotropic Willmore energy with respect to the corresponding anisotropic L^2-metric.

R. Perl (✉) · M. Rumpf
Institut für Numerische Simulation, Rheinische Friedrich-Wilhelms-Universität Bonn,
Endenicher Allee 60, D-53115 Bonn, Germany
e-mail: ricardo.perl@ins.uni-bonn.de; martin.rumpf@ins.uni-bonn.de

P. Pozzi
Fakultät für Mathematik, Universität Duisburg, Forsthausweg 2, D-47057 Duisburg, Germany
e-mail: paola.pozzi@uni-due.de

M. Griebel (ed.), *Singular Phenomena and Scaling in Mathematical Models*,
DOI 10.1007/978-3-319-00786-1_10,

The isotropic Willmore energy is given by $w[x] = \frac{1}{2}\int_{\mathscr{M}} \mathbf{h}^2 da$, where x denotes the identity map and $\mathbf{h}$ the mean curvature on a surface $\mathscr{M}$. The isotropic L^2-metric is given by $(v, v)_{\mathscr{M}} = \int_{\mathscr{M}} |v|^2 da$, which is considered as a squared L^2-distance of the surface $\mathscr{M}$ being displaced with the vector field v from the non displaced surface $\mathscr{M}$. In the hypersurface case Willmore flow leads to a fourth order parabolic evolution problem, which defines for a given initial surface $\mathscr{M}_0$ a family of surfaces $\mathscr{M}(t)$ for $t \geq 0$ with $\mathscr{M}(0) = \mathscr{M}_0$ [30, 47, 49]. Applications of a minimization of the isotropic Willmore energy and the corresponding Willmore flow include the processing of edge sets in imaging [13, 34, 36, 51], geometry processing [8, 9, 48, 50] and the mathematical treatment of biological membranes [24, 29, 46]. Starting with work by Polden [40, 41] existence and regularity of Willmore flow was advanced in the last decade [31, 33, 43].

Now, in the context of Finsler geometry the classical area functional is replaced by the anisotropic area functional $\mathbf{a}_\gamma[x] = \int_{\mathscr{M}} \gamma(n) da$ with a local area weight $\gamma(n)$ depending on the local surface orientation. Here, γ is a positive, 1–homogeneous anisotropy function. In analogy to the isotropic case, the anisotropic mean curvature $\mathbf{h}_\gamma$ is defined as the L^2-representation of the variation of the anisotropic area in the direction of normal variations of the surface and can be evaluated as $\mathbf{h}_\gamma = \mathrm{div}_{\mathscr{M}} (\nabla\gamma(n))$. Hence, a possible first choice for an anisotropic Willmore functional is given by $\frac{1}{2} \int_{\mathscr{M}} \mathbf{h}_\gamma^2 \; da$. Clarenz [15] has shown that Wulff shapes are the only minimizers of this energy. Palmer [39] studied variational problems involving anisotropic bending energies for surfaces with and without boundaries. Unfortunately, this energy definition does not imply the scale invariance property of Wulff shapes known for round spheres under isotropic Willmore flow. Indeed, any round sphere is a stationary point of the isotropic Willmore functional in $\mathbb{R}^3$. In $\mathbb{R}^2$ a circle of radius R_0 evolves under isotropic Willmore flow according to the ordinary differential equation $\dot{R} = \frac{1}{2}R^{-3}$. The counterpart of a round sphere in the anisotropic context is the Wulff shape as the unit ball with respect to the norm associated with the dual γ^* of the anisotropy γ. But there is no such scaling law for the evolution of Wulff shapes under the above anisotropic variant of Willmore flow.

To ensure full consistency with the Finsler geometry, one has to adapt both the anisotropic energy and the anisotropic metric as suggested in [42] (see Sect. 2). Indeed, we make use of the associated anisotropic metric $\int_{\mathscr{M}} \gamma^* (v) (\nabla\gamma^*) (v) \cdot v \, \gamma(n) da$ (here only defined for $v(x) \neq 0$ for all $x \in \mathscr{M}$, cf. Sect. 3 for the general case), acting on a motion field v of the surface $\mathscr{M}$ with normal n. Furthermore, we will use the anisotropic area weight to define the anisotropic Willmore energy, i.e. $w_\gamma(x) = \frac{1}{2} \int_{\mathscr{M}} \mathbf{h}_\gamma^2 \, \gamma(n) da$. Then, it turns out that Wulff shapes in $\mathbb{R}^2$ actually evolve according to the same evolution law for radial parameter valid for the evolution of circles under the isotropic flow. Recently, Bellettini and Mugnai [6] investigated the first variation of this functional in the smooth case. Concerning the proper time and space discretization, this consistent choice of the anisotropic Willmore energy and the anisotropic metric on surface variations perfectly fits to the framework of the natural variational time discretization of geometric gradient flows.

The finite element approximation of Willmore flow was first investigated by Rusu [44] based on a mixed method for the surface parametrization x and the mean curvature vector $\mathbf{h}\, n$ as independent variables, see also [17] for the application to surface restoration. In [23] a level set formulation of Willmore flow was proposed. In the case of graph surfaces Deckelnick and Dziuk [18] were able to prove convergence of a related space discrete and time continuous scheme. Deckelnick and Schieweck established convergence of a conforming finite element approximation for axial symmetric surfaces [20]. In the case of the elastic flow of curves an error analysis was given by Dziuk and Deckelnick in [19]. An alternative scheme, which in particular ensures a better distribution of nodes on the evolving surface was presented by Barrett, Garcke and Nürnberg [2, 4]. Using discrete geometry calculus Bobenko and Schröder [10] suggested a discrete Willmore flow of triangular surfaces. The time discretization of the second order, anisotropic mean curvature flow has been considered by Dziuk already in [27, 28] and he gave convergence results for curves. Diewald [21] has extended the discretization approach for isotropic Willmore flow of Rusu [44] to some anisotropic variant, for which Droske [22] and Nemitz [35] investigated a level set discretization.

Most of the above discretization methods are based on some semi-implicit time discretization, which requires the solution of linear systems of equations at each time step. Thereby, the involved geometric differential operators are assembled on the surface from the previous time step. In the application one observes strong restrictions on the time step size. This shortcoming motivated the development of a new approach for the time discretization of Willmore flow in [1] based on the following general concept for a variational time discretization of gradient flows: The gradient flow on a (in general infinite dimensional) manifold with respect to an energy $e[\cdot]$ and a metric g on the manifold is defined as the evolution problem $\dot{x} = -\mathrm{grad}_g e[x]$ with initial data x^0, where $\mathrm{grad}_g e[x]$ is the representation of the variation $e'[x]$ in the metric g, i.e. $g(\mathrm{grad}_g e[x], \zeta) = e'[x](\zeta)$ for all infinitesimal variations ζ of x. Now, one defines a time discrete family $(x^k)_{k=0,\dots}$ with the desired property $x^k \approx x(k\tau)$ for the given time step size τ. To this end, one successively solves a sequence of variational problems, i.e. in time step k

$$x^{k+1} = \arg\min_x \mathrm{dist}(x^k, x)^2 + 2\tau\, e[x],$$

where $\mathrm{dist}(x^k, x) = \inf_{\gamma \in \Gamma[x^k, x]} \int_0^1 \sqrt{g_{\gamma(s)}(\dot{\gamma}(s), \dot{\gamma}(s))} ds$ denotes the Riemannian distance of x from x^k on the manifold and $\Gamma[x^k, x]$ is the set of smooth curves γ with $\gamma(0) = x^k$ and $\gamma(1) = x$. The striking observation for this abstract scheme is that one immediately obtains an energy estimate, i.e. $e[x^{k+1}] + \frac{1}{2\tau}\mathrm{dist}(x^k, x^{k+1})^2 \leq e[x^k]$. In the context of geometric flows, this approach was studied by Luckhaus and Sturzenhecker [32] leading to a fully implicit variational time discretization for mean curvature motion in BV and by Chambolle [11], who reformulated this scheme in terms of a level set method and generalized it for the approximation of anisotropic mean curvature motion in [7, 12]. The time discretization for Willmore flow proposed in [1] builds upon this general paradigm. In this paper, we will show how to adapt the approach to the time discretization of the anisotropic Willmore flow which is fully consistent with Finsler geometry.

The paper is organized as follows. In Sect. 2 we briefly review the time discretization of isotropic Willmore flow. Building on these prerequisites the generalization to anisotropic Willmore flow is discussed in Sect. 3. Then, in Sect. 4 we discuss a fully discrete numerical scheme based on piecewise affine finite elements on simplicial surface meshes. In Sect. 5 the Lagrangian calculus from PDE constraint optimization is used to develop a suitable algorithm for the solution of the nested optimization problem to be solved in each time step. Finally, in Sect. 6 computational results are presented. An appendix collects essential ingredients of the corresponding algorithm.

2 Review of the Time Discretization of Isotropic Willmore Flow

In this section we will briefly recall the nested time discretization of isotropic Willmore from [1]. We denote a hypersurface in $\mathbb{R}^{d+1}$ by $\mathscr{M} = \mathscr{M}[y]$. Here, y indicates a parametrization of $\mathscr{M}$ and can also be considered as the identity map on $\mathscr{M}$ parametrizing $\mathscr{M}$ over itself. Then, the abstract variational time discretization of isotropic Willmore flow reads as follows:

For a given surface $\mathscr{M}[x^k]$ with parametrization x^k and a time step τ find a mapping $x = x[x^k]$ such that $\text{dist}(\mathscr{M}[x^k], \mathscr{M}[x])^2 + \tau \int_{\mathscr{M}[x]} \mathbf{h}^2 da \longrightarrow \min$, where $\text{dist}(\mathscr{M}[z], \mathscr{M}[v])^2 = \int_{\mathscr{M}[z]} (v-z)^2 da$ is the squared L^2-distance of surfaces $\mathscr{M}[v]$ from the surface $\mathscr{M}[z]$, $\mathbf{h} = \mathbf{h}[x]$ is the mean curvature of $\mathscr{M}[x]$, and $\int_{\mathscr{M}[x]} da$ denotes the surface area of $\mathscr{M}[x]$.

Now, we take into account that the mean curvature $\mathbf{h} = \mathbf{h}[x]$ is the L^2-gradient of the area functional on a surface $\mathscr{M}[x]$ and that mean curvature motion is the corresponding gradient flow. Thus, the mean curvature vector $\mathbf{h}[x]n[x]$ with $n = n[x]$ denoting the normal on $\mathscr{M}[x]$ can be approximated by the discrete time derivative $\frac{y[x]-x}{\tilde{\tau}}$, where $y[x]$ is a suitable approximation of a single time step of the evolution of mean curvature motion with initial data x and time step size $\tilde{\tau}$. This time step itself can again be approximated using an (inner) variational scheme, i.e. we define $y[x]$ to be the minimizer of

$$e_{\text{in}}[x, y] := \int_{\mathscr{M}[x]} (y - x)^2 + \tilde{\tau} |\nabla_{\mathscr{M}[x]} y|^2 da\,. \tag{1}$$

In fact, the corresponding Euler Lagrange equation is identical to the defining equation of the semi-implicit scheme for mean curvature motion proposed by Dziuk [26]:

$$0 = \int_{\mathscr{M}[x]} (y - x)\theta + \tilde{\tau} \nabla_{\mathscr{M}[x]} y \cdot \nabla_{\mathscr{M}[x]} \theta da. \tag{2}$$

Now, given $y[x]$ as the minimizer of (1) for small $\tilde{\tau}$ the functional $\frac{1}{2}\int_{\mathscr{M}[x]} \frac{(y[x]-x)^2}{\tilde{\tau}^2} da$ is an approximation of the Willmore functional on $\mathscr{M}[x]$. This approximation is then used to define a variational scheme for a time step of the actual Willmore flow. To this end, we consider for given surface parametrization x^k the functional

$$e_{\text{out}}[x^k, x, y] := \int_{\mathscr{M}[x^k]} (x - x^k)^2 da + \frac{\tau}{\tilde{\tau}^2} \int_{\mathscr{M}[x]} (y - x)^2 da\,,$$

where we suppose $y = y[x]$ to be the minimizer of (1). To summarize, we obtain the following scheme for the kth time step of Willmore flow:

Given an initial surface $\mathscr{M}[x^0]$ with parametrization x^0 we define a sequence of surfaces $\mathscr{M}[x^k]$ with parametrizations x^k for $k = 1, \ldots$ via the solution of the following sequence of nested variational problems

$$x^{k+1} = \arg\min_x e_{out}[x^k, x, y[x]]\,, \textit{ where} \tag{3}$$

$$y[x] = \arg\min_y e_{in}[x, y]\,. \tag{4}$$

The inner variational problem (4) is quadratic, thus the resulting Euler–Lagrange equation (2) is linear and we end up with a PDE constrained optimization problem to be solved in each time step. For more details we refer to [1].

3 Nested Time Discretization for Anisotropic Willmore Flow

Now, let us investigate the time discretization of anisotropic Willmore flow in the co-dimension one case. Here, we will in particular focus on the proper choice of energy and metric. We assume that $\gamma : \mathbb{R}^{d+1} \to [0, \infty)$ is a positive, 1–homogeneous (i.e. $\gamma(\lambda p) = |\lambda|\gamma(p)$ for all $\lambda \in \mathbb{R}$, $p \in \mathbb{R}^{d+1}$) and sufficiently regular function that satisfies the ellipticity condition

$$\gamma''(p)qq \geq c_0\|q\|^2 \quad \forall\, p, q \in \mathbb{R}^{d+1}, \|p\| = 1, p \cdot q = 0 \tag{5}$$

for some positive constant c_0 and the Euclidean norm $\|\cdot\|$. As already mentioned $\gamma(n)$ represents the anisotropic area weight for a surface normal n. The isotropic case is recovered by choosing $\gamma(\cdot) = \|\cdot\|$. We define the dual function of γ as

$$\gamma^*(x) := \sup\{\langle x, \psi\rangle \mid \psi \in B_\gamma\} \quad \forall\, x \in \mathbb{R}^{d+1}\,,$$

where B_γ denotes the unit Ball in the γ-norm. The ellipticity assumption ensures that $(\mathbb{R}^{d+1}, \gamma)$ and its dual space $(\mathbb{R}^{d+1}, \gamma^*)$ are uniformly convex Banach spaces and the duality map $T : (\mathbb{R}^{d+1}, \gamma^*) \to (\mathbb{R}^{d+1}, \gamma)$, with

$$T(x) = \frac{1}{2}\partial(\gamma^*(x)^2),$$

is an odd single-valued bijective continuous map. More precisely $T(0) = 0$, $T(x) = \gamma^*(x)\nabla\gamma^*(x)$ for $x \neq 0$, and $T^{-1}(\xi) = \gamma(\xi)\nabla\gamma(\xi)$ for $\xi \neq 0$. For details we refer to [42]. The unit ball $\mathscr{F} := \{x \in \mathbb{R}^{d+1} : \gamma(x) \leq 1\}$ in $(\mathbb{R}^{d+1}, \gamma)$ is denoted the Frank diagram, the associated dual unit ball $\mathscr{W} := \{x \in \mathbb{R}^{d+1} : \gamma^*(x) \leq 1\}$ is the corresponding Wulff shape. Wulff shapes are known to be solutions to the isoperimetric problem, that is, $\partial\mathscr{W}$ minimizes the anisotropic area functional

$$\mathbf{a}_\gamma[x] = \int_{\mathscr{M}[x]} \gamma(n[x])da \tag{6}$$

(with $\gamma(n[x])da$ denoting the anisotropic area element) in the class of surfaces enclosing the same volume (cf. [16] and the references therein). Now, based on the anisotropy γ and its dual γ^* we define an anisotropic distance dist_γ of a manifold $\mathscr{M}[y]$ from a manifold $\mathscr{M}[x]$ by

$$\mathrm{dist}_\gamma(\mathscr{M}[x], \mathscr{M}[y])^2 := \int_{\mathscr{M}[x]} \gamma^*(y-x)^2\gamma(n[x])da \tag{7}$$

for sufficiently regular x and y. The choice of the norm γ^* together with the anisotropic area weight $\gamma(n[x])$ in (7) reflects the fact that the anisotropic area of the boundary of a convex body $K \subset \mathbb{R}^{d+1}$ can be interpreted as

$$\mathbf{a}_\gamma(\partial K) = \lim_{\epsilon \to 0} \frac{|K + \epsilon\mathscr{W}| - |K|}{\epsilon},$$

where $|\cdot|$ denotes the usual Lebesgue volume in $\mathbb{R}^{d+1}$. In particular, the underlying metric structure is dictated by the Wulff shape and its norm γ^* (see [5, 42] and references therein).

Based on these considerations let us first consider anisotropic mean curvature motion, which is defined as the gradient flow of the anisotropic surface area with respect to the above anisotropic metric. In this case the variational time discretization is associated with the minimization of

$$\mathrm{dist}_\gamma(\mathscr{M}[x], \mathscr{M}[y])^2 + 2\tilde{\tau}\int_{\mathscr{M}[y]} \gamma(n[y])da \tag{8}$$

with respect to y for a given surface $\mathscr{M}[x]$ and $\tilde{\tau} > 0$. Let us denote by $y[x]$ the minimizer for given surface parameterization x. The Euler Lagrange equation for (8) is given by

$$\begin{aligned} 0 &= \int_{\mathscr{M}[x]} T(y-x)\cdot\theta\,\gamma(n[x])da + \tilde{\tau}\langle a'_\gamma[y],\theta\rangle \\ &= \tilde{\tau}\int_{\mathscr{M}[x]} T\left(\frac{y-x}{\tilde{\tau}}\right)\cdot\theta\,\gamma(n[x])da + \tilde{\tau}\langle a'_\gamma[y],\theta\rangle \end{aligned} \tag{9}$$

for smooth test functions $\theta : \mathscr{M}[x] \to \mathbb{R}^{d+1}$. Together with $\partial_t y(k\tilde{\tau}) \approx \frac{y-x}{\tilde{\tau}}$ this reflects the weak formulation of anisotropic mean curvature motion given by

$$\int_{\mathscr{M}[y]} T(\partial_t y)\cdot\theta\,\gamma(n[y])da = -\langle \mathbf{a}_\gamma{}'[y],\theta\rangle \tag{10}$$

for a parametrization y and smooth test functions θ defined on $\mathscr{M}[y]$ (cf. [42]). Here, the variation of the anisotropic area functional is given by

$$\langle \mathbf{a}_\gamma{}'[y],\theta\rangle = \int_{\mathscr{M}[y]} \mathbf{h}_\gamma[y]\frac{n[y]}{\gamma(n[y])}\cdot\theta\,\gamma(n[y])da\,,$$

where $\mathbf{h}_\gamma[y] = \operatorname{div}_{\mathscr{M}[y]}(n_\gamma[y]) = \operatorname{div}_{\mathscr{M}[y]}(\nabla\gamma(n[y]))$ denotes the anisotropic mean curvature with $n_\gamma[y] = \nabla\gamma(n[y])$ (see [14]). Thus, from (10) we deduce that

$$T(\partial_t y) = -\mathbf{h}_\gamma[y]\frac{n[y]}{\gamma(n[y])}$$

or equivalently we achieve the strong formulation of anisotropic mean curvature motion

$$\partial_t y = \kappa_\gamma[y] := T^{-1}\left(-\mathbf{h}_\gamma[y]\frac{n[y]}{\gamma(n[y])}\right) = -\mathbf{h}_\gamma[y]\nabla\gamma(n[y])\,.$$

Indeed, as pointed out in [42], the last equality holds due to the 1-homogeneity of γ, i.e.

$$\begin{aligned} \gamma\left(-\mathbf{h}_\gamma[y]\frac{n[y]}{\gamma(n[y])}\right)\nabla\gamma\left(-\mathbf{h}_\gamma[y]\frac{n[y]}{\gamma(n[y])}\right) &= -\frac{\mathbf{h}_\gamma[y]}{\gamma(n[y])}\gamma(n[y])\nabla\gamma(n[y]) \\ &= -\mathbf{h}_\gamma[y]\nabla\gamma(n[y])\,. \end{aligned}$$

Next, we deal with the actual anisotropic Willmore flow and consider the anisotropic Willmore functional defined as follows for a parametrization x of $\mathscr{M}[x]$:

$$w_\gamma[x] := \frac{1}{2}\int_{\mathscr{M}[x]} \mathbf{h}_\gamma[x]^2\,\gamma(n[x])da = \frac{1}{2}\int_{\mathscr{M}[x]} \gamma^*(\kappa_\gamma[x])^2\,\gamma(n[x])da\,. \tag{11}$$

Here, we have used that the 1-homogeneity and $\nabla\gamma(\xi) \in \partial\mathscr{W}$ for all $\xi \in \mathbb{R}^{d+1}$ imply

$$\gamma^*(\kappa_\gamma)^2 = \gamma^*\left(-\mathbf{h}_\gamma \nabla\gamma(n)\right)^2 = \mathbf{h}_\gamma^2 \gamma^*\left(-\nabla\gamma(n)\right)^2 = \mathbf{h}_\gamma^2\,.$$

Then the abstract variational time discretization of anisotropic Willmore flow reads as follows:

Given $\mathscr{M}[x^k]$ and time step τ find a mapping $x = x[x^k]$ such that x minimizes

$$\mathrm{dist}_\gamma(\mathscr{M}[x^k], \mathscr{M}[x])^2 + \tau \int_{\mathscr{M}[x]} \gamma^*(\kappa_\gamma[x])^2\, \gamma(n[x])da\,. \tag{12}$$

As in the isotropic case, we will now replace the anisotropic mean curvature vector by the discrete speed extracted from a scheme for a single time step of anisotropic curvature flow (10). Explicitly, $\gamma^*(\frac{y[x]-x}{\tilde{\tau}})^2$ is a suitable approximation of $\mathbf{h}_\gamma^2[x] = \gamma^*(\kappa_\gamma[x])^2$, where $\frac{y[x]-x}{\tilde{\tau}}$ is the time discrete speed extracted from the variational time discretization of anisotropic curvature motion. Furthermore, we use the definition of the anisotropic distance measure in (7). Finally, based on this approximation we derive the actual time discretization of anisotropic Willmore flow. For a given surface parametrization x^k of the surface $\mathscr{M}[x^k]$ at a time step k we define the functionals

$$e_{\text{out}}[x^k, x, y] := \int_{\mathscr{M}[x^k]} \gamma^*(x - x^k)^2\, \gamma(n[x^k])da + \frac{\tau}{\tilde{\tau}^2} \int_{\mathscr{M}[x]} \gamma^*(y - x)^2\, \gamma(n[x])da\,,$$

$$e_{\text{in}}[x, y] := \int_{\mathscr{M}[x]} \gamma^*(y - x)^2\, \gamma(n[x])da + 2\tilde{\tau} \int_{\mathscr{M}[y]} \gamma(n[y])da\,,$$

and in analogy to the isotropic case above, we end up with the following fully nonlinear variational time discretization of anisotropic Willmore flow:

Given an initial surface $\mathscr{M}[x^0]$ with parametrization x^0 we define a sequence of surfaces $\mathscr{M}[x^k]$ with parametrizations x^k for $k = 1, \ldots$ via the solution of the following sequence of nested variational problems

$$x^{k+1} = \arg\min_x e_{out}[x^k, x, y[x]], \qquad \textit{where} \tag{13}$$

$$y[x] = \arg\min_y e_{in}[x, y]\,. \tag{14}$$

Different from the variational scheme for isotropic Willmore flow, the inner variational problem is no longer quadratic. It is worth to mention that this variational time discretization does not involve derivatives of the anisotropy. Nevertheless, as we will discuss below in the context of the actual computation, differentiation is required to run Newton methods for the associated Lagrangian functional. Indeed, for this we will need $\gamma, \gamma^* \in C^3(\mathbb{R}^{d+1} \setminus \{0\})$; moreover, unless $(\gamma^*)^2 \in C^3(\mathbb{R}^{d+1})$

(which holds for $\gamma(p) = \sqrt{Ap \cdot p}$ with a symmetric positive definite matrix A), a regularization will be required (see Sect. 6 below).

Let us conclude this section with a study of boundaries $\partial\mathscr{W}$ of two-dimensional Wulff shapes $\mathscr{W}$ moving under anisotropic Willmore flow in the plane. To this end consider the parametrization $x : (0,T) \times S^1 \to \mathbb{R}^2$, $x(t,\nu) = R(t)\nabla\gamma(\nu)$ of the boundary of a (rescaled) Wulff shape $R(t)\mathscr{W}$. Using the results given in [42] it is easily seen that x moves under anisotropic Willmore flow if $R(t)$ solves the ODE

$$\dot{R(t)} = \frac{1}{2R(t)^3}\,.$$

Hence, we observe that Wulff shapes expand in time like in the isotropic case (cf. [1]) with $R(t) = \sqrt[4]{R(0)^4 + 2t}$. Next let us compare this with the time discrete evolution based on the proposed nested variational time discretization. We write $x, y, x^k : S^1 \to \mathbb{R}^2$, $x(\nu) = R\nabla\gamma(\nu)$, $y(\nu) = \tilde{R}\nabla\gamma(\nu)$, $x^k(\nu) = R^k\nabla\gamma(\nu)$. Since $\gamma^*(\nabla\gamma(\nu)) = 1$ we immediately derive

$$\begin{aligned} e_{\text{out}}[x^k, x, y] &= (R - R^k)^2\mathbf{a}_\gamma(x^k) + \frac{\tau}{\tilde{\tau}^2}(\tilde{R} - R)^2\mathbf{a}_\gamma(x), \\ e_{\text{in}}[x, y] &= (\tilde{R} - R)^2\mathbf{a}_\gamma(x) + 2\tilde{\tau}\mathbf{a}_\gamma(y). \end{aligned}$$

Considering variations $y_\epsilon(\nu) = (\tilde{R} + \epsilon\psi)\nabla\gamma(\nu)$ in direction of the anisotropic normal n_γ we infer from the inner problem that

$$(\tilde{R} - R)\mathbf{a}_\gamma(x) + \frac{\tilde{\tau}}{\tilde{R}}\mathbf{a}_\gamma(y) = 0.$$

More precisely, since $\mathbf{a}_\gamma(y) = \frac{\tilde{R}}{R}\mathbf{a}_\gamma(x)$ due to the homogeneity property of γ, we have that

$$\tilde{R} = R - \frac{\tilde{\tau}}{R}.$$

This, together with $\mathbf{a}_\gamma(x) = \frac{R}{R^k}\mathbf{a}_\gamma(x^k)$, gives

$$e_{\text{out}}[x^k, x, y] = \mathbf{a}_\gamma(x^k)\left((R - R^k)^2 + \frac{\tau}{RR^k}\right),$$

from which we deduce

$$\frac{R - R^k}{\tau} = \frac{1}{2R^kR^2}.$$

Note that this is a slightly different time step scheme than the one reported for the isotropic case ($\gamma(\cdot) = \|\cdot\|$) in [1, § 2.1]. This is due to the fact that we use an implicit formulation of the inner problem as opposed to the linear equation (1) in the scheme for isotropic Willmore flow (cf. Sect. 2).

4 Finite Element Discretization in Space

Following the approach in [1] we now derive a suitable spatial discretization based on piecewise affine finite elements. This is in close correspondence to the surface finite element approach by Dziuk [25]. To this end, we consider simplicial meshes $\mathscr{M}[X]$ as approximations of the hypersurfaces $\mathscr{M}[x]$ in $\mathbb{R}^{d+1}$, i.e. polygonal curves for $d = 1$ and triangular surfaces for $d = 2$. Thereby, X is a parametrization of the simplicial mesh $\mathscr{M}[X]$ which is uniquely described by a vector $\bar{X}$ of vertex positions of the mesh. Here, and in what follows, we will always denote discrete quantities with upper case letters to distinguish them from the corresponding continuous quantities in lower case letters. Furthermore, a bar on top of a discrete function indicates the associated vector of nodal values, i.e. $\bar{X} = (\bar{X}_i)_{i\in I}$, where $\bar{X}_i = (X_i^1, \cdots, X_i^{d+1})$ is the coordinate vector of the ith vertex of the mesh and I denotes the index set of vertices. For $d = 1$ each element T is a line segment with nodes X_0 and X_1 (using local indices) and for $d = 2$ the elements T are planar triangles with vertices X_0, X_1, and X_2 and edge vectors $F_0 = X_2 - X_1$, $F_1 = X_0 - X_2$, and $F_2 = X_1 - X_0$. Given a simplicial surface $\mathscr{M}[X]$, the associated piecewise affine finite element space is given by

$$\mathscr{V}(\mathscr{M}[X]) := \left\{U \in C^0(\mathscr{M}[X]) \,|\, U|_T \in \mathscr{P}_1 \; \forall T \in \mathscr{M}[X]\right\}$$

with the nodal basis denoted by $\{\Phi_i\}_{i\in I}$. Here, $\mathscr{P}_1$ is the space of affine functions on a simplex T. Thus, for $U \in \mathscr{V}(\mathscr{M}[X])$ we obtain $U = \sum_{i\in I} U(X_i)\Phi_i$ and $\bar{U} = (U(X_i))_{i\in I}$. Let us emphasize that the parametrization mapping X itself is considered as an element in $\mathscr{V}(\mathscr{M}[X])^{d+1}$ and we recover the vector of nodes $\bar{X} = (X_i)_{i\in I}$.

With these algorithmic ingredients at hand we now can derive a fully discrete nested time discretization of anisotropic Willmore flow, as the spatially discrete counterpart of (13) and (14):

Given a discrete initial surface $\mathscr{M}[X^0]$ with discrete parametrization X^0 we compute a sequence of surfaces $\mathscr{M}[X^k]$ with parametrizations X^k by solving the nested, finite dimensional variational problems

$$X^{k+1} = \arg\min\nolimits_{X\in\mathscr{V}(\mathscr{M}[X^k])^{d+1}} \mathscr{E}_{out}[X^k, X, Y[X]], \quad \textit{where} \tag{15}$$

$$Y[X] = \arg\min\nolimits_{Y\in\mathscr{V}(\mathscr{M}[X])^{d+1}} \mathscr{E}_{in}[X, Y]\,. \tag{16}$$

Here, the functionals $\mathscr{E}_{\text{in}}$ and $\mathscr{E}_{\text{out}}$ are straightforward spatially discrete counterpart of the functionals $e_{\text{in}}[x, y]$ and $e_{\text{out}}[x^k, x, y]$ and are defined by

$$\mathscr{E}_{\text{in}}[X,Y] := \int_{\mathscr{M}[X]} \mathbf{I}\left(\gamma^*(Y-X)^2\right)\, \gamma(N[X])da + 2\tilde{\tau}\int_{\mathscr{M}[Y]} \gamma(N[Y])da\,,$$

$$\begin{aligned}\mathscr{E}_{\text{out}}[X^k,X,Y] := &\int_{\mathscr{M}[X^k]} \mathbf{I}\left(\gamma^*(X-X^k)^2\right)\, \gamma(N[X^k])da \\ &+\frac{\tau}{\tilde{\tau}^2}\int_{\mathscr{M}[X]} \mathbf{I}\left(\gamma^*(Y-X)^2\right)\, \gamma(N[X])da\,,\end{aligned}$$

where the nodal interpolation operator $\mathbf{I}$ renders the resulting scheme fully practical. To simplify the exposition, we introduce the discrete quadratic form $\mathbf{M}_\gamma[Z,X] = \int_{\mathscr{M}[X]} \mathbf{I}\left(\gamma^*(Z)^2\right)\gamma(N[X])da$ (a nonlinear counterpart of the quadratic form induced by the lumped mass matrix) and the discrete anisotropic area functional $\mathbf{A}_\gamma[Y] = \int_{\mathscr{M}[Y]} \gamma(N[Y])da$, both of which are assembled from local contributions on simplices of the underlying simplicial grid $\mathscr{T}_h$:

$$\mathbf{M}_\gamma[Z,X] = \sum_{T\in\mathscr{T}_h} \frac{1}{(d+1)!}\left(\sum_{i=0,\dots,d} \gamma^*(\bar{Z}_{T,i})^2\right)\gamma(R_T[\bar{X}]), \tag{17}$$

$$\mathbf{A}_\gamma[X] = \sum_{T\in\mathscr{T}_h} \frac{1}{d!}\gamma(R_T[\bar{X}]). \tag{18}$$

Here, $R_T[\bar{X}] = D^{90}(\bar{X}_{T,1} - \bar{X}_{T,0})$ for $d = 1$ and $R_T[\bar{X}] = (\bar{X}_{T,1} - \bar{X}_{T,0}) \wedge (\bar{X}_{T,2} - \bar{X}_{T,0})$ for $d = 2$. Hence, we can rewrite

$$\begin{aligned}\mathscr{E}_{\text{out}}[X^k,X,Y] &= \mathbf{M}_\gamma[X-X^k,X^k] + \frac{\tau}{\tilde{\tau}^2}\mathbf{M}_\gamma[Y-X,X]\,,\\ \mathscr{E}_{\text{in}}[X,Y] &= \mathbf{M}_\gamma[Y-X,X] + 2\tilde{\tau}\mathbf{A}_\gamma[Y]\,.\end{aligned}$$

The necessary condition for $Y[X]$ to be a minimizer of $\mathscr{E}_{\text{in}}[X,\cdot]$ is given by the corresponding discrete Euler Lagrange equation

$$0 = \partial_Y \mathscr{E}_{\text{in}}[X,Y[X]](\Theta) = \partial_Z \mathbf{M}_\gamma[Y-X,X](\Theta) + 2\tilde{\tau}\partial_Y \mathbf{A}_\gamma[Y](\Theta)$$

for all $\Theta \in \mathscr{V}(\mathscr{M}[X])^{d+1}$.

5 Optimization Algorithm for the Time Steps

In this section, the actual optimization algorithm for the nested, fully discrete variational problem derived in Sect. 4 is presented. Thereby, we apply a step size controlled Newton method (cf. [45] Sect. 7) for the corresponding Lagrangian (cf. Nocedal and Wright [37]). In our context the Lagrangian function for problem (15), (16) is given by

$$\mathscr{L}[\bar{X},\bar{Y},\bar{P}] = \mathscr{E}_{\text{out}}[X^k, X, Y] - \partial_Y \mathscr{E}_{\text{in}}[X,Y](P)$$

for independent unknowns $\bar{X},\bar{Y} \in \mathbb{R}^{(d+1)|I|}$ and the Lagrange multiplier $\bar{P} \in \mathbb{R}^{(d+1)|I|}$ (with a slight misuse of notation, we consider these unknowns as finite element function in the spaces $\mathscr{V}(\mathscr{M}[X^k])^{d+1}$ and $\mathscr{V}(\mathscr{M}[X])^{d+1}$, respectively, or as the associated nodal vector in $\mathbb{R}^{(d+1)|I|}$). For an extensive discussion of the Lagrangian ansatz we refer to [38]. Now, we ask for critical points $(\bar{X},\bar{Y},\bar{P})$ of L. Indeed, $0 = \partial_{\bar{P}}\mathscr{L}[\bar{X},\bar{Y},\bar{P}](\bar{\Theta}) = -\partial_Y \mathscr{E}_{\text{in}}[X,Y](\Theta)$ is the Euler Lagrange equation of the inner minimization problem with respect to Y for given X and $0 = \partial_{\bar{Y}}\mathscr{L}[\bar{X},\bar{Y},\bar{P}](\bar{\Theta}) = \partial_Y \mathscr{E}_{\text{out}}[X^k,X,Y](\Theta) - \partial_Y^2 \mathscr{E}_{\text{in}}[X,Y](P,\Theta)$ is the defining equation for the dual solution P given Y as the solution of the above Euler Lagrange equation. Finally, the Euler Lagrange equation for the actual constraint optimization problem coincides with

$$0 = \partial_{\bar{X}}\mathscr{L}[\bar{X},\bar{Y},\bar{P}](\bar{\Theta}) = \partial_X \mathscr{E}_{\text{out}}(X^k,X,Y)(\Theta) - \partial_X\partial_Y \mathscr{E}_{\text{in}}[X,Y](P,\Theta)\,.$$

For the gradient of the Lagrangian $\mathscr{L}$ we obtain

$$\operatorname{grad}\mathscr{L} = \begin{pmatrix} \partial_X \mathscr{E}_{\text{out}} - \partial_X\partial_Y \mathscr{E}_{\text{in}}(P) \\ \partial_Y \mathscr{E}_{\text{out}} - \partial_Y^2 \mathscr{E}_{\text{in}}(P) \\ -\partial_Y \mathscr{E}_{\text{in}} \end{pmatrix}$$

with

$$\begin{aligned} \partial_X \mathscr{E}_{\text{out}}[X^k,X,Y](\Theta) &= \partial_Z \mathbf{M}_\gamma[X-X^k,X^k](\Theta) \\ &\quad + \frac{\tau}{\tilde{\tau}^2}(\partial_X \mathbf{M}_\gamma[Y-X,X](\Theta) - \partial_Z \mathbf{M}_\gamma[Y-X,X](\Theta))\,, \\ \partial_Y \mathscr{E}_{\text{out}}[X^k,X,Y](\Theta) &= \frac{\tau}{\tilde{\tau}^2}\partial_Z \mathbf{M}_\gamma[Y-X,X](\Theta)\,, \\ \partial_X\partial_Y \mathscr{E}_{\text{in}}[X,Y](P,\Theta) &= -\partial_Z^2 \mathbf{M}_\gamma[Y-X,X](P,\Theta) + \partial_X\partial_Z \mathbf{M}_\gamma[Y-X,X](P,\Theta)\,, \\ \partial_Y^2 \mathscr{E}_{\text{in}}[X,Y](P,\Theta) &= \partial_Z^2 \mathbf{M}_\gamma[Y-X,X](P,\Theta) + 2\tilde{\tau}\partial_Y^2 \mathbf{A}_\gamma[Y](P,\Theta)\,. \end{aligned}$$

The Hessian of $\mathscr{L}$, which is required to implement a Newton scheme, is given (in abbreviated form) by

$$\operatorname{Hess}\mathscr{L} = \begin{pmatrix} \partial_X^2 \mathscr{E}_{\text{out}} - \partial_X^2\partial_Y \mathscr{E}_{\text{in}}(P) & \partial_X\partial_Y \mathscr{E}_{\text{out}} - \partial_X\partial_Y^2 \mathscr{E}_{\text{in}}(P) & -\partial_X\partial_Y \mathscr{E}_{\text{in}} \\ \partial_X\partial_Y \mathscr{E}_{\text{out}} - \partial_X\partial_Y^2 \mathscr{E}_{\text{in}}(P) & \partial_Y^2 \mathscr{E}_{\text{out}} - \partial_Y^3 \mathscr{E}_{\text{in}}(P) & -\partial_Y^2 \mathscr{E}_{\text{in}} \\ -\partial_X\partial_Y \mathscr{E}_{\text{in}} & -\partial_Y^2 \mathscr{E}_{\text{in}} & 0 \end{pmatrix}.$$

The different terms in Hess $\mathscr{L}$ are evaluated as follows:

$$\partial_X^2 \mathscr{E}_{\text{out}}(\Theta,\Psi) = \partial_Z^2 \mathbf{M}_\gamma[X-X^k,X^k](\Theta,\Psi) + \frac{\tau}{\tilde{\tau}^2}(\partial_X^2 \mathbf{M}_\gamma[Y-X,X](\Theta,\Psi)$$

$$\begin{aligned}
&\qquad -2\partial_Z \partial_X \mathbf{M}_\gamma[Y-X,X](\Theta,\Psi) + \partial_Z^2 \mathbf{M}_\gamma[Y-X,X](\Theta,\Psi)\big)\,,\\
\partial_Y \partial_X \mathscr{E}_{\text{out}}(\Theta,\Psi) &= \frac{\tau}{\tilde{\tau}^2}\big(\partial_Z \partial_X \mathbf{M}_\gamma[Y-X,X](\Theta,\Psi) - \partial_Z^2 \mathbf{M}_\gamma[Y-X,X](\Theta,\Psi)\big)\,,\\
\partial_Y^2 \mathscr{E}_{\text{out}}(\Theta,\Psi) &= \frac{\tau}{\tilde{\tau}^2}\partial_Z^2 \mathbf{M}_\gamma[Y-X,X](\Theta,\Psi)\,,\\
\partial_X^2 \partial_Y \mathscr{E}_{\text{in}}(\Theta,\Psi,\Xi) &= \partial_Z^3 \mathbf{M}_\gamma[Y-X,X](\Theta,\Psi,\Xi) - \partial_X \partial_Z^2 \mathbf{M}_\gamma[Y-X,X](\Theta,\Psi,\Xi)\\
&\quad -\partial_X \partial_Z^2 \mathbf{M}_\gamma[Y-X,X](\Theta,\Xi,\Psi) + \partial_X^2 \partial_Z \mathbf{M}_\gamma[Y-X,X](\Theta,\Psi,\Xi)\,,\\
\partial_X \partial_Y^2 \mathscr{E}_{\text{in}}(\Theta,\Psi,\Xi) &= -\partial_Z^3 \mathbf{M}_\gamma[Y-X,X](\Theta,\Psi,\Xi) + \partial_X \partial_Z^2 \mathbf{M}_\gamma[Y-X,X](\Theta,\Psi,\Xi)\,,\\
\partial_Y^3 \mathscr{E}_{\text{in}}(\Theta,\Psi,\Xi) &= \partial_Z^3 \mathbf{M}_\gamma[Y-X,X](\Theta,\Psi,\Xi) + 2\tilde{\tau}\partial_Y^3 \mathbf{A}_\gamma[Y](\Theta,\Psi,\Xi)\,.
\end{aligned}$$

In the implementation of the proposed scheme it is convenient to directly treat the squared, dual anisotropy $\gamma^{*,2}(.) := (\gamma^*(.))^2$ in the calculation of derivatives of the anisotropic functionals, which is particularly advantageous for anisotropies of the type $\gamma(p) = \sum_{k=1}^K \sqrt{p \cdot G_k p}$ where the G_k are symmetric and positive definite (cf. Garcke et al. [3]). The different terms of the gradient grad$\mathscr{L}$ and the Hessian Hess$\mathscr{L}$ are in the usual way assembled from local contribution on simplices of the polygonal mesh. The required formulas are given in the Appendix.

6 Numerical Results

In this section, we show applications of the proposed algorithm to the evolution of curves in $\mathbb{R}^2$ under anisotropic Willmore flow. Beside anisotropies with ellipsoidal Wulff shapes we study regularized crystalline anisotropies $\gamma(\cdot) = \|\cdot\|_{\ell^1}$ and $\gamma(\cdot) = \|\cdot\|_{\ell_\infty}$ based on a suitable regularization. A particular emphasis is on the verification of the robustness and stability of the proposed approach in particular for large time steps. Furthermore, we experimentally verify that Wulff shapes grow self-similar in time under the corresponding anisotropic Willmore flow.

At first, we study anisotropies of the type

$$\gamma(z) = \sqrt{a_1^2 z_1^2 + a_2^2 z_2^2}$$

for given $a_1, a_2 > 0$. In that case, the squared dual anisotropy function is given by

$$\gamma^{*,2}(z) = \frac{z_1^2}{a_1^2} + \frac{z_2^2}{a_2^2}.$$

Figure 1 compares the evolution of a circle of radius $R_0 = 1$ under isotropic Willmore flow for $a_1 = a_2 = 1$ with the evolution of an ellipse with half axes $a_1 = 6$ and $a_2 = 1$ under the corresponding anisotropic flow. As discussed in Sect. 2, in both cases the initial curve $\mathscr{M}_0$ expands in a self-similar fashion, i.e. $\mathscr{M}[x(t)] = R(t)\mathscr{M}_0$ with $R(t) = \sqrt[4]{R_0^4 + 2t}$ for $r_0 > 0$. In Fig. 1 we plot the evolution

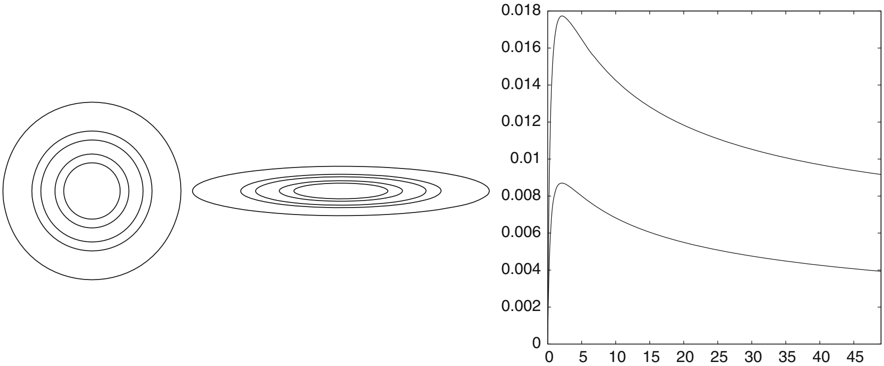


Fig. 1 The evolution of an unit circle under isotropic Willmore flow is plotted on the *top left*. For the computation we used as initial grid size $h = 0.0981$ resulting from 64 vertices. Furthermore, $\tau = h$, $\tilde{\tau} = h^2$ and the resulting discrete *curves* are shown for $t = 0, 10\tau, 50\tau, 100\tau, 500\tau$. In the *bottom left* we display the evolution of an ellipse (with half axes 6 and 1) under anisotropic Willmore flow with 256 elements and $h = 0.0984$. Here, we consider $\tau = h$, $\tilde{\tau} = h^2$ and display the approximate solutions for $t = 0, 10\tau, 50\tau, 100\tau, 500\tau$. Next, the associated L^2-errors are plotted over time on the *right*, where the lower error curve corresponds to the evolution results (*top left*)

Table 1 The L^2-error between the exact solution of the self-similar evolution of circles under Willmore flow and the discrete solution of the fully implicit variational time discretization is plotted at time $t = 0.1542$ for a grid size $h(t)$ (left) and $t = 0.3927$ (right). On the left we consider time step sizes τ and $\tilde{\tau}$ of the order of the squared spatial grid size h_0 at the initial time 0, whereas on the right both time step sizes are taken equal to the grid size. In both cases we have considered 2^n vertices for the polygon, resulting in an initial grid size $h_0 = \frac{2\pi}{2^n}$

		L^2-error			L^2-error	
n	$h(t)$	$(\tau = \tilde{\tau} = h_0^2)$	*eoc*	$h(t)$	$(\tau = \tilde{\tau} = h_0)$	*eoc*
4	4.166e−1	4.830e−3		4.482e−1	1.916e−2	
5	2.096e−1	1.328e−3	1.879	2.258e−1	1.087e−2	0.826
6	1.049e−1	3.403e−4	1.969	1.132e−1	5.804e−3	0.909
7	5.249e−2	8.561e−5	1.992	5.668e−2	3.000e−3	0.954
8	2.625e−2	2.144e−5	1.998	2.836e−2	1.525e−3	0.977

of the error $err(h) := \|\mathscr{I}_h x(t) - x_h(t)\|_{L^2}$ in time. Thereby, the L^2-error is evaluated on the polygonal curve $x_h(t)$ and $\mathscr{I}_h$ denotes the nodal interpolation of $x(t)$ at the projected positions of the nodes of $x_h(t)$ in direction $\nabla\gamma(n[x_h(t)])$. In Tables 1 and 2 we provide results on the experimental order of convergence $eoc := \log(err(h_1)/err(h_2))/\log(h_1/h_2)$ for varying grid and time step size in case of the evolution of the circle and the ellipse.

Now, we want to study crystalline anisotropies $\gamma(\cdot) = \|\cdot\|_{\ell^1}$ and $\gamma(\cdot) = \|\cdot\|_{\ell_\infty}$. As already pointed out, even though the formulation of the scheme itself doesn't explicitly need assumptions on the smoothness of γ, the application of the optimization algorithm requires the computation of derivatives of γ up to order 3. In

Table 2 As in Table 1 experimental orders of convergence are reported, now for the self-similar evolution of the ellipses (with half axis 6 and 1) under anisotropic Willmore flow. Here, again polygons with 2^n vertices are considered, equi-distributed along the initial ellipse with an initial grid size $h_0 = \frac{24.172}{2^n}$. On the left the error is evaluated at time $t = 0.596576$ and on the right at time $t = 0.77238$

		L^2-error			L^2-error	
n	$h(t)$	$(\tau = \tilde{\tau} = h_0^2)$	*eoc*	$h(t)$	$(\tau = \tilde{\tau} = h_0)$	*eoc*
5	1.435e+0	1.648e−1		1.274e+0	1.942e−1	
6	6.487e−1	3.476e−2	1.960	5.875e−1	7.089e−2	1.303
7	3.069e−1	8.762e−3	1.841	2.842e−1	3.424e−2	1.002
8	1.525e−1	2.182e−3	1.987	1.396e−1	1.724e−3	0.966

fact, we use the following regularization: For a small parameter $\varepsilon > 0$ we regularize the ℓ^1-norm by

$$\ell^1_\varepsilon(z) = \sum_{l=1}^{2} \sqrt{\varepsilon |z|^2 + z_l^2}.$$

Since in $\mathbb{R}^2$ the ℓ^∞-norm equals a rotated and scaled ℓ^1-norm we use as regularization of the ℓ^∞-norm

$$\ell^\infty_\varepsilon(z) = \frac{\sqrt{\varepsilon |z|^2 + (z_1 + z_2)^2}}{2} + \frac{\sqrt{\varepsilon |z|^2 + (z_1 - z_2)^2}}{2}.$$

Figure 2 shows the evolution of a sphere with respect to the regularized ℓ^∞-norm under the associated anisotropy Willmore flow with anisotropy $\gamma(\cdot) = \|\cdot\|_{\ell^1_\varepsilon}$ for $\varepsilon = 0.0001$. Results on the self-similar evolution of spheres with respect to the regularized ℓ^1-norm are depicted in Fig. 3. In these simulations, we use the analog regularization for the dual anisotropy γ^* required in the algorithm.

Next, we generalize Willmore flow and replace the Willmore energy by the modified energy

$$e_\gamma[x] := \int_{\mathscr{M}[x]} \left(\frac{1}{2} \mathbf{h}_\gamma^2 + \lambda \right) \gamma(n[x]) da \,, \tag{19}$$

with a second term given by the anisotropic area weighted with a constant $\lambda > 0$. The incorporation of this generalized energy in our computational approach is straightforward. The generalized flow combines expansive forcing with respect to the anisotropic Willmore flow of curves with contractive forcing due to the anisotropic mean curvature motion associated to the anisotropic area functional. Thus, for the generalized model we expect convergence to a limit shape given by a scaled Wulff shape, where the scaling depends on the factor λ. Figure 4 shows the impact of the factor λ on the evolution, whereas in Fig. 5 we compare the evolution of different initial shapes under the generalized anisotropic Willmore flow for different anisotropies.

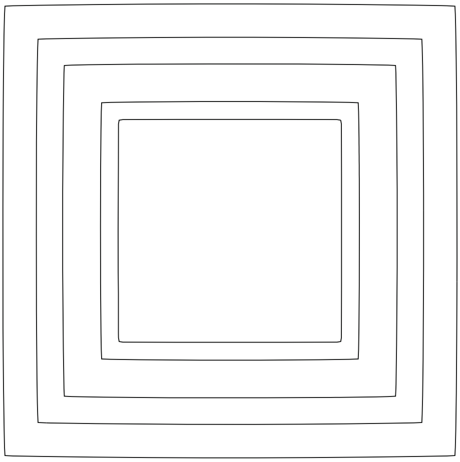


Fig. 2 Evolution of the unit sphere with respect to the regularized ℓ^∞-norm under anisotropic Willmore flow for the anisotropy $\|\cdot\|_{\ell^1_\varepsilon}$ with $\varepsilon = 0.0001$. For this computation we consider 200 vertices leading to an initial grid size $h_0 = 0.04$. Furthermore, $\tau = h_0$ and $\tilde{\tau} = h_0^2$ and the resulting discrete curves are shown for $t = 0, 10\tau, 50\tau, 100\tau, 200\tau$

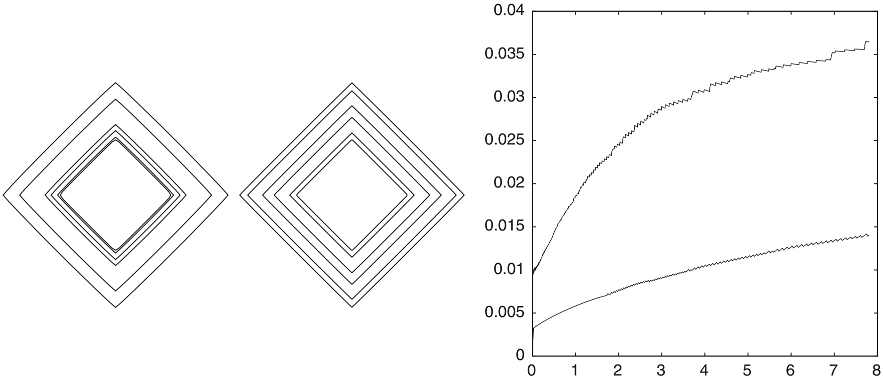


Fig. 3 Evolution of the unit sphere with respect to the regularized ℓ^1-norm under anisotropic Willmore flow for the anisotropy $\gamma(\cdot) = \|\cdot\|_{\ell^\infty_\varepsilon}$. The parameters are $h_0 = 0.0078$, $\varepsilon = 0.001$, $\tau = \tilde{\tau} = h_0^2$ and curves are plotted at times $t = 0, 10\tau, 50\tau, 100\tau, 500\tau, 1{,}000\tau$ on the *top left* and $h_0 = 0.0283$, $\varepsilon = 0.0001$, $\tau = h_0$, $\tilde{\tau} = h_0^2$, $t = 0, 10\tau, 50\tau, 100\tau, 200\tau, 275\tau$ in the *bottom left*. On the *right* the associated L^2-errors are plotted over time, where the *lower error curve* corresponds to the evolution results (*top left*)

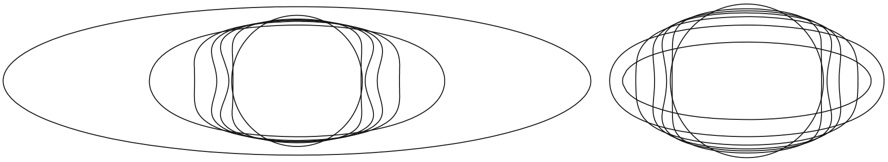


Fig. 4 The impact of the parameter λ is shown for the evolution of a circle to an ellipse with aspect ratio $4:1$ (i.e. $a_1 = 4$ and $a_2 = 1$). We evolve polygons with 160 vertices approximating the unit sphere as initial curve, $h_0 = 0.0393$ and $\tau = \tilde{\tau} = 0.01$, $h = 0.000393$. On the *left* $\lambda = 0.025$ and on the *right* $\lambda = 4$

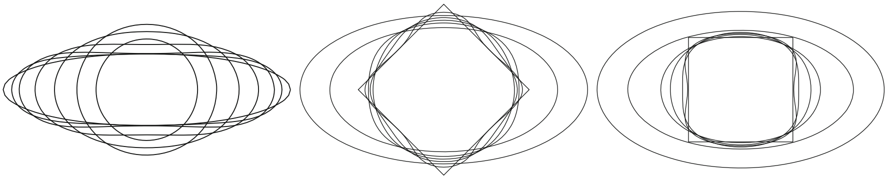


Fig. 5 The evolution of different initial shapes for different anisotropies is displayed. For all computations we use 100 vertices and choose $\lambda = 0.25$. On the *left* we start with an ellipse with aspect ratio $4:1$ under an isotropic flow with $\gamma(\cdot) = \|\cdot\|$ ($h_0 = 0.1739$, $\tau = h_0$, $\tilde{\tau} = h_0^2$) results are shows at $t = 0, 0.1739, 0.5218, 1.739, 3.478, 6.956, 173.9$. In the *middle* and on the *right* an ellipsoidal anisotropy with aspect ratio $2:1$ is used (i.e. $a_1 = 2$, $a_2 = 1$) in the first case (*middle*), we take as initial shape the unit sphere for the l^1-norm ($h_0 = 0.0566$, $\tau = \tilde{\tau} = 0.001\, h_0$) and results are displayed at $t = 0, 0.00017, 0.00085, 0.00169, 0.006, 0.056, 0.251$. In the second example (*right*), the initial shape is the unit sphere for the l^∞-norm ($h_0 = 0.08$ and $\tau = \tilde{\tau} = 0.01\, h_0$) and results are depicted for $t = 0, 0.0024, 0.008, 0.04, 0.08, 0.8, 4.8$

Acknowledgements Ricardo Perl was supported by the DFG project Ru 567/14-1 and Martin Rumpf acknowledges support by the SFB 611.

Appendix

Here, we collect the computational ingredients to evaluate the Lagrangian, its gradient and Hessian based on a standard local assembly procedure. In the following, for vectors $x \in \mathbb{R}^{d+1}$ and functions f, we use the notation $f_{i,j}(x) = \frac{\partial f_i(x)}{\partial x_j}$ and in analogous notation for higher order derivatives. Furthermore, for matrices A we use $f_{k,ij}(A) = \frac{\partial f_k(A)}{\partial A_{ij}}$ and again in analogous notation for higher order derivatives. In fact, we can restrict ourselves to the local functionals

$$\mathbf{M}_{T,\gamma}[Z,X] = \frac{1}{(d+1)!}\left(\sum_{i=0,\dots,d} \gamma^*(\bar{Z}_i)^2\right)\gamma(R[\bar{X}])\,,\ \mathbf{A}_{T,\gamma}[X] = \frac{1}{d!}\gamma(R[\bar{X}])\,, \quad (20)$$

where we denote by $\bar{Z} = (Z_0, \ldots, Z_d)$ and $\bar{X} = (X_0, \ldots, X_d)$ the corresponding vectors of simplex nodes in $\mathbb{R}^{d+1}$ with coordinate representation $Z_j = (Z_{jr})_{r=1,\ldots,d+1}$ and $X_j = (X_{jr})_{r=1,\ldots,d+1}$. Here, R is a mapping from $\mathbb{R}^{(d+1)^2}$ to $\mathbb{R}^{d+1}$ representing the 90° rotated edge vector for $d = 1$ and the cross product of edge vectors for $d = 2$, respectively. For $d = 1$ we obtain for the first derivatives of $R[\bar{X}] = \begin{pmatrix} X_{02} - X_{12} \\ X_{11} - X_{01} \end{pmatrix}$ with respect to the entries (ij) with $i = 0, \ldots, d$ and $j = 1, \ldots, d+1$

$$R_{,01}[\bar{X}] = \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \quad R_{,02}[\bar{X}] = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad R_{,11}[\bar{X}] = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad R_{,12}[\bar{X}] = \begin{pmatrix} -1 \\ 0 \end{pmatrix}.$$

Because of the linearity of R for $d = 1$ all higher derivatives vanish. For $d = 2$ we have

$$R[\bar{X}] = \left(\sum_{u,v=1}^{3} \epsilon_{iuv}(X_{1u} - X_{0u})(X_{2v} - X_{0v}) \right)_{i=1,2,3},$$

where ϵ_{wuv} is the Levi-Civita symbol ($\epsilon_{wuv} = \pm 1$ if (w, u, v) is a even/odd permutation of $(1, 2, 3)$ and 0 else). Thus, for $w = 1, 2, 3$ we have

$$R_{w,js}[\bar{X}] = \sum_{u,v=1}^{3} \epsilon_{wuv} \big((\delta_{1j} - \delta_{0j})\delta_{su}(X_{2v} - X_{0v}) + (\delta_{2j} - \delta_{0j})\delta_{sv}(X_{1u} - X_{0u}) \big),$$

$$R_{w,js\,lt}[\bar{X}] = \sum_{u,v=1}^{3} \epsilon_{wuv} \big((\delta_{1j} - \delta_{0j})(\delta_{2l} - \delta_{0l})\delta_{su}\delta_{tv} + (\delta_{2j} - \delta_{0j})(\delta_{1l} - \delta_{0l})\delta_{sv}\delta_{tu} \big),$$

and all third derivatives $R_{u,ir\,js\,lt}[\bar{X}]$ vanish. Here $j, l \in \{0, 1, 2\}$ refer to the local node and $s, t \in \{1, 2, 3\}$ to the spacial component. Next we derive expressions for the derivatives of (17) and (18) under the assumption that γ, γ^* are sufficiently smooth in $\mathbb{R}^{d+1} \setminus \{0\}$ (thus Z_i, $R[\bar{X}] \neq 0$):

Derivatives of $\mathbf{M}_{T,\gamma}$

$$\partial_{Z_{ir}} \mathbf{M}_{T,\gamma}[Z, X] = \frac{1}{(d+1)!} \gamma^{*,2}_{,r}(Z_i)\gamma(R[\bar{X}]),$$

$$\partial_{Z_{js}} \partial_{Z_{ir}} \mathbf{M}_{T,\gamma}[Z, X] = \frac{\delta_{ij}}{(d+1)!} \gamma^{*,2}_{,rs}(Z_i)\gamma(R[\bar{X}]),$$

$$\partial_{Z_{lt}} \partial_{Z_{js}} \partial_{Z_{ir}} \mathbf{M}_{T,\gamma}[Z, X] = \frac{\delta_{ij}\delta_{il}}{(d+1)!} \gamma^{*,2}_{,rst}(Z_i)\gamma(R[\bar{X}]),$$

$$
\begin{aligned}
\partial_{X_{ir}}\mathbf{M}_{T,\gamma}[Z,X] &= \frac{1}{(d+1)!}\left(\sum_{\alpha=0,\dots,d}\gamma^{*,2}(\bar{Z}_\alpha)\right)\sum_{s=1}^{m}\gamma_{,s}(R[\bar{X}])R_{s,ir}[\bar{X}]\,,\\
\partial_{X_{js}}\partial_{X_{ir}}\mathbf{M}_{T,\gamma}[Z,X] &= \frac{1}{(d+1)!}\left(\sum_{\alpha=0,\dots,d}\gamma^{*,2}(\bar{Z}_\alpha)\right)\\
&\quad\cdot\left(\sum_{t=1}^{m}\gamma_{,t}(R[\bar{X}])R_{t,irjs}[\bar{X}]+\sum_{t,u=1}^{m}\gamma_{,tu}(R[\bar{X}])R_{t,ir}[\bar{X}]R_{u,js}[\bar{X}]\right)\,,\\
\partial_{X_{js}}\partial_{Z_{ir}}\mathbf{M}_{T,\gamma}[Z,X] &= \frac{1}{(d+1)!}\gamma^{*,2}_{,r}(Z_i)\sum_{t=1}^{m}\gamma_{,t}(R[\bar{X}])R_{t,js}[\bar{X}]\,,\\
\partial_{X_{lt}}\partial_{Z_{js}}\partial_{Z_{ir}}\mathbf{M}_{T,\gamma}[Z,X] &= \frac{\delta_{ij}}{(d+1)!}\gamma^{*,2}_{,rs}(Z_i)\sum_{u=1}^{m}\gamma_{,u}(R[\bar{X}])R_{u,lt}[\bar{X}]\,,\\
\partial_{X_{lt}}\partial_{X_{js}}\partial_{Z_{ir}}\mathbf{M}_{T,\gamma}[Z,X] &= \frac{1}{(d+1)!}\gamma^{*,2}_{,r}(Z_i)\\
&\quad\cdot\left(\sum_{v=1}^{m}\gamma_{,v}(R[\bar{X}])R_{v,js\,lt}[\bar{X}]+\sum_{v,u=1}^{m}\gamma_{,vu}(R[\bar{X}])R_{v,js}[\bar{X}]R_{u,lt}[\bar{X}]\right)\,,
\end{aligned}
$$

Derivatives of $\mathbf{A}_{T,\gamma}$

$$
\begin{aligned}
\partial_{Y_{ir}}\mathbf{A}_{T,\gamma}[X] &= \frac{1}{d!}\sum_{s=1}^{m}\gamma_{,s}(R[\bar{X}])R_{s,ir}[\bar{X}]\,,\\
\partial_{Y_{js}}\partial_{Y_{ir}}\mathbf{A}_{T,\gamma}[X] &= \frac{1}{d!}\left(\sum_{t=1}^{m}\gamma_{,t}(R[\bar{X}])R_{t,irjs}[\bar{X}]+\sum_{t,u=1}^{m}\gamma_{,tu}(R[\bar{X}])R_{t,ir}[\bar{X}]R_{u,js}[\bar{X}]\right)\,,\\
\partial_{Y_{lt}}\partial_{Y_{js}}\partial_{Y_{ir}}\mathbf{A}_{T,\gamma}[X] &= \frac{1}{d!}\Big(\sum_{u,v=1}^{m}\gamma_{,vu}(R[\bar{X}])R_{v,irjs}[\bar{X}]R_{u,lt}[\bar{X}]+\sum_{v=1}^{m}\gamma_{,v}(R[\bar{X}])R_{v,irjslt}[\bar{X}]\\
&\quad+\sum_{u,v,w=1}^{m}\gamma_{,vuw}(R[\bar{X}])R_{v,ir}[\bar{X}]R_{u,js}[\bar{X}]R_{w,lt}[\bar{X}]\\
&\quad+\sum_{u,v=1}^{m}\gamma_{,vu}(R[\bar{X}])R_{v,ir\,lt}[\bar{X}]R_{u,js}[\bar{X}]+\sum_{u,v=1}^{m}\gamma_{,vu}(R[\bar{X}])R_{v,ir}[\bar{X}]R_{u,jslt}[\bar{X}]\Big)\,.
\end{aligned}
$$

References

1. Balzani, N., Rumpf, M.: A nested variational time discretization for parametric Willmore flow. Interfaces Free Bound. **14**(4), 431–454 (2012)
2. Barrett, J.W., Garcke, H., Nürnberg, R.: A parametric finite element method for fourth order geometric evolution equations. J. Comp. Phys. **222**, 441–467 (2007)

3. Barrett, J.W., Garcke, H., Nürnberg, R.: Numerical approximation of anisotropic geometric evolution equations in the plane. IMA J. Numer. Anal. **28**(2), 292–330 (2008)
4. Barrett, J.W., Garcke, H., Nürnberg, R.: Parametric approximation of isotropic and anisotropic elastic flow for closed and open curves. Numer. Math. **120**, 489–542 (2012)
5. Bellettini, G.: Anisotropic and crystalline mean curvature flow. In: Bao, D., Bryant, R.L., Chern, S.S., Shen, Z. (eds.) A Sampler of Riemann-Finsler Geometry, vol. 50, pp. 49–82. Cambridge University Press, Cambridge (2004)
6. Bellettini, G., Mugnai, L.: Anisotropic geometric functionals and gradient flows. Banach Cent. Publ. **86**, 21–43 (2009)
7. Bellettini, G., Caselles, V., Chambolle, A., Novaga, M.: Crystalline mean curvature flow of convex sets. Arch. Ration. Mech. Anal. **179**(1), 109–152 (2006)
8. Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: Proceedings of the SIGGRAPH 2000, New Orleans, pp. 417–424 (2000)
9. Bertalmio, M., Bertozzi, A., Sapiro, G.: Navier-Stokes, fluid dynamics, and image and video inpainting. In: IEEE Proceedings of the International Conference on Computer Vision and Pattern Recognition, Kauai, vol. 1, pp. 355–362 (2001)
10. Bobenko, I.A., Schröder, P.: Discrete Willmore Flow, pp. 101–110. ACM (2005)
11. Chambolle, A.: An algorithm for mean curvature motion. Interfaces Free Bound. **6**, 195–218 (2004)
12. Chambolle, A., Novaga, M.: Convergence of an algorithm for anisotropic mean curvature motion. SIAM J. Math. Anal. **37**, 1978–1987 (2006)
13. Chan, T.F., Kang, S.H., Shen, J.: Euler's elastica and curvature-based inpainting. SIAM Appl. Math. **63**(2), 564–592 (2002)
14. Clarenz, U.: Enclosure theorems for extremals of elliptic parametric functionals. Calc. Var. **15**, 313–324 (2002)
15. Clarenz, U.: The Wulff-shape minimizes an anisotropic Willmore functional. Interfaces Free Bound. **6**(3), 351–359 (2004)
16. Clarenz, U., Dziuk, G., Rumpf, M.: On generalized mean curvature flow in surface processing. In: Karcher, H., Hildebrandt, S. (eds.) Geometric Analysis and Nonlinear Partial Differential Equations, pp. 217–248. Springer, Berlin/New York (2003)
17. Clarenz, U., Diewald, U., Dziuk, G., Rumpf, M., Rusu, R.: A finite element method for surface restoration with smooth boundary conditions. Comput. Aided Geom. Des. **21**(5), 427–445 (2004)
18. Deckelnick, K., Dziuk, G.: Error analysis of a finite element method for the Willmore flow of graphs. Interfaces Free Bound. **8**, 21–46 (2006)
19. Deckelnick, K., Dziuk, G.: Error analysis for the elastic flow of parametrized curves. Math. Comp. **78**(266), 645–671 (2009)
20. Deckelnick, K., Schieweck, F.: Error analysis for the approximation of axisymmetric Willmore flow by C1-elements. Interfaces Free Bound. **12**(4), 551–574 (2010)
21. Diewald, U.: Anisotrope Krümmungsflüsse parametrischer Flächen sowie deren Anwendung in der Flächenverarbeitung. Dissertation, University Duisburg (2005)
22. Droske, M.: On variational problems and gradient flows in image processing. Dissertation, University Duisburg (2005)
23. Droske, M., Rumpf, M.: A level set formulation for Willmore flow. Interfaces Free Bound. **6**(3), 361–378 (2004)
24. Du, Q., Liu, C., Wang, X.: Simulating the deformation of vesicle membranes under elastic bending energy in three dimensions. J. Comput. Phys. **212**(2), 757–777 (2006)
25. Dziuk, G.: Finite elements for the Beltrami operator on arbitrary surfaces. In: Hildebrandt, S., Leis, R. (eds.) Partial Differential Equations and Calculus of Variations. Lecture Notes in Mathematics 1357, pp. 142–155. Springer, Berlin/New York (1988)
26. Dziuk, G.: An algorithm for evolutionary surfaces. Numer. Math. **58**, 603–611 (1991)
27. Dziuk, G.: Convergence of a semi-discrete scheme for the curve shortening flow. Math. Models Methods Appl. Sci. **4**, 589–606 (1994)

28. Dziuk, G.: Discrete anisotropic curve shortening flow. Siam J. Numer. Anal. **36**(6), 1808–1830 (1999)
29. Helfrich, W.: Elastic properties of lipid bilayers: theory and possible experiments. Zeitschrift für Naturforschung **28c**, 693–703 (1973)
30. Kuwert, E., Schätzle, R.: The Willmore flow with small initial energy. J. Differ. Geom. **57**(3), 409–441 (2001)
31. Kuwert, E., Schätzle, R.: Gradient flow for the Willmore functional. Commun. Anal. Geom. **10**(5), 1228–1245 (2002). (Electronic)
32. Luckhaus, S., Sturzenhecker, T.: Implicit time discretization for the mean curvature flow equation. Calc. Var. **3**, 253–271 (1995)
33. Mayer, U., Simonett, G.: A numerical scheme for axisymmetric solutions of curvature driven free boundary problems with applications to the Willmore flow. Interfaces Free Bound. **4**(1), 89–109 (2002)
34. Mumford, D.: Elastica and computer vision. In: Bajaj, C. (ed.) Algebraic Geometry and Its Applications, pp. 491–506. Springer, New York (1994)
35. Nemitz, O.: Anisotrope Verfahren in der Bildverarbeitung: Gradientenflüsse, Level-Sets und Narrow Bands. Dissertation, University of Bonn (2008).
36. Nitzberg, M., Mumford, D., Shiota, T.: Filtering, Segmentation and Depth. Lecture Notes in Computer Science, vol. 662. Springer, Berlin/Heidelberg (1993)
37. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer, New York/Berlin (1999)
38. Olischläger, N., Rumpf, M.: Two step time discretization of Willmore flow. Accepted at IMA Conference on the Mathematics of Surfaces (2009)
39. Palmer, B.: Equilibria for anisotropic bending energies. Math. Phys. **50**(2), 023512 (2009)
40. Polden, A.: Closed curves of least total curvature. SFB 382 Tübingen, Preprint **13** (1995)
41. Polden, A.: Curves and surfaces of least total curvature and fourth-order flows. Dissertation, Universität Tübingen (1996)
42. Pozzi, P.: On the gradient flow for the anisotropic area functional. Math. Nachr. **285**, 707–726 (2012)
43. Rivière, T.: Analysis aspects of Willmore surfaces. Invent. Math. **174**(1), 1–45 (2008)
44. Rusu, R.: An algorithm for the elastic flow of surfaces. Interfaces Free Bound. **7**, 229–239 (2005)
45. Schaback, R., Wendland, H.: Numerische Mathematik, 5th edn. Springer, Berlin (2004)
46. Seifert, U.: Configurations of fluid membranes and vesicles. Adv. Phys. **46**, 13–137 (1997)
47. Simonett, G.: The Willmore flow near spheres. Diff. Integral Eq. **14(8)**, 1005–1014 (2001)
48. Welch, W., Witkin, A.: Variational surface modeling. Comput. Graph. **26**(2), 157–166 (1992)
49. Willmore, T.: Riemannian Geometry. Claredon, Oxford (1993)
50. Xu, G., Pan, Q.: G^1 surface modelling using fourth order geometric flows. Comput. Aided Des. **38**(4), 392–403 (2006)
51. Yoshizawa, S., Belyaev, A.G.: Fair triangle mesh generation with discrete elastica. In: Proceedings of the Geometric Modeling and Processing: Theory and Applications (GMP'02), pp. 119–123. IEEE Computer Society, Washington, DC (2002)

Energy Scaling and Domain Branching in Solid-Solid Phase Transitions

Allan Chan and Sergio Conti

Abstract We consider a vectorial model for solid-solid phase transformations, namely,

$$E_\varepsilon[u] = \int_\Omega W(Du) + \varepsilon |D^2 u| \, dx \, ,$$

where $u : \Omega \subset \mathbb{R}^2 \to \mathbb{R}^2$ and W vanishes on a set of the form $K = SO(2)A \cup SO(2)B$, with A, B two rank-one connected matrices representing the eigenstrains of two martensitic variants. We study the scaling of the minimal energy under Dirichlet boundary conditions corresponding to the average of A and B. In the case that A and B have two rank-one connections we show that the minimum of E_ε scales, for small ε, as $\varepsilon^{2/3}$, in agreement with previous results on the scalar version of the model. In the case that the two matrices have a single rank-one connection instead we show that a different scaling appears, with energy proportional to $\varepsilon^{4/5}$. Both results correspond to a self-similar refinement of the microstructure around the boundary, with a different period-doubling pattern. Our results extend to a vectorial, properly frame-indifferent framework previous results on a scalar model by Kohn and Müller.

1 Introduction

Materials undergoing a solid-solid phase transformation, such as shape-memory alloys or iron, develop characteristic microstructures where domain boundaries arrange themselves along a few preferred orientations, which are selected by the

A. Chan (✉) · S. Conti
Institut für Angewandte Mathematik, Rheinische Friedrich-Wilhelms-Universität Bonn,
Endenicher Allee 60, D-53115 Bonn, Germany
e-mail: chan@uni-bonn.de; sergio.conti@uni-bonn.de

M. Griebel (ed.), *Singular Phenomena and Scaling in Mathematical Models*,
DOI 10.1007/978-3-319-00786-1_11,

compatibility conditions among the spontaneous strains. Starting with the works of Ball and James [1, 2] there has been in the last two decades a vast effort in the mathematical community towards understanding such microstructures via the study of appropriate variational models, whose main ingredient is a term of the form

$$E_0[u,\Omega] = \int_\Omega W(Du)dx\,.$$

Here $u : \Omega \subset \mathbb{R}^n \to \mathbb{R}^n$ represents the elastic deformation and $W : \mathbb{R}^{n\times n} \to \mathbb{R}$ the energy density, a typical form being

$$W(F) = \operatorname{dist}^2(F,K) = \inf\{|F-G|^2 : G \in K\}\,. \tag{1}$$

The set of energy-minimizing deformation gradients $K \subset \mathbb{R}^{n\times n}$ depends on the specific phase transformation considered. For cubic-tetragonal phase transitions in three dimensions K equals $K_3 = SO(3)\{U_1,U_2,U_3\}$, where the U_i are the three diagonal matrices with eigenvalues $(\lambda,\lambda,\lambda^{-2})$, $\lambda \neq 1$ being a positive parameter. In two dimensions one uses $K_2 = SO(2)\{U_1,U_2\}$, with eigenvalues λ and $1/\lambda$. If sufficient regularity of u is assumed, then solutions of $Du \in K$ have a very rigid structure. For example, if they are assumed to be piecewise smooth, then they are piecewise affine, and interfaces have prescribed directions. In particular, Du can jump from the value U_i to the value QU_j, with $Q \in SO(n)$, only if the difference is a rank-one matrix, in the sense that

$$U_i - QU_j = a \otimes \nu \tag{2}$$

for some $a,\nu \in \mathbb{R}^n$. The vector ν is then the normal to the interface.

If the regularity requirement is relaxed, then it is possible to find a large class of deformations u such that $W(Du) = 0$ pointwise almost everywhere, using the theory of convex integration [12, 18]. Precisely, for both K_2 and K_3 one can show that there is $\rho > 0$ such that for any bounded Lipschitz set Ω, and $v \in C^{1,\beta}(\Omega;\mathbb{R}^n)$ which obeys $\det v = 1$, $\|Dv - \mathrm{Id}\|_{L^\infty} \leq \rho$, one can find infinitely many $u \in \mathrm{Lip}(\Omega;\mathbb{R}^n)$ such that $u = v$ on $\partial\Omega$ and $\operatorname{dist}(Du,K) = 0$ almost everywhere in Ω. These deformations are, however, unphysical since they have very low regularity. (The maps u are typically Lipschitz but nowhere C^1, and their gradients do not have bounded variation, see [14, 15]).

Therefore, one is led to consider singularly perturbed functionals which include terms penalizing interfaces, of the form

$$E_\varepsilon[u,\Omega] = \int_\Omega W(Du)\,dx + \varepsilon|D^2u|(\Omega)\,. \tag{3}$$

Here ε is a (small) positive parameter, $u \in W^{1,2}(\Omega;\mathbb{R}^n)$ with $Du \in BV(\Omega;\mathbb{R}^{n\times n})$, and $|D^2u|(\Omega)$ denotes the total variation of the measure D^2u, which for smooth functions coincides with $\int_\Omega |D^2u|\,dx$.

A scalar simplification of this problem was considered by Kohn and Müller in 1992–1994, see [16, 17]. Precisely, they proposed the functional

$$J_\varepsilon[v,\Omega] = \begin{cases} \int_\Omega (D_1 v)^2 dx + \varepsilon |D_2 D_2 v|(\Omega) & \text{if } |D_2 v| = 1 \text{ a.e.,} \\ \infty & \text{otherwise}, \end{cases}$$

where $\Omega = (0,L)^2 \subset \mathbb{R}^2$, and $v : \Omega \to \mathbb{R}$. They have shown that there is $c > 0$ such that, for sufficiently small ε,

$$\frac{1}{c}\varepsilon^{2/3} L^{4/3} \le \min\{J_\varepsilon[v,\Omega] : v = 0 \text{ on } \partial\Omega\} \le c\varepsilon^{2/3} L^{4/3} \tag{4}$$

(and a corresponding result on rectangles $(0,L) \times (0,H)$). The functional J_ε is rigid even for $\varepsilon = 0$, in the sense that if we ignore the singular perturbation and require $J_0[v,\Omega] = 0$, which means $Dv \in \{(0,1),(0,-1)\}$ almost everywhere, then only very special deformations v are possible – precisely, those which do not depend on x_1 and obey $D_2 v \in \{1,-1\}$ a.e. This is compatible only with very few boundary conditions. The functional E_0, due to fact that $K = W^{-1}(0)$ is infinite, is instead much softer, and as discussed above possesses a large number of minimizers. One may expect that this difference between J_0 and E_0 could be reflected in a different behavior of the regularized functionals, i.e., in a corresponding difference between J_ε and E_ε for small ε. We prove here that this is not the case.

We consider the full vectorial problem in two dimensions, and obtain a result similar to (4). We focus for simplicity on a domain $\Omega = (0,L)^2$, and assume that the set K has the form

$$K = SO(2)A \cup SO(2)B \tag{5}$$

for two matrices $A, B \in \mathbb{R}^{2\times 2}$ such that $\det A, \det B > 0$ and $\operatorname{rank}(A - B) = 1$. It can be shown that the condition rank $(A - QB) = 1$, $Q \in SO(n)$, has either only the trivial solution $Q = \mathrm{Id}$ or two distinct solutions, corresponding to one or two possible orientations of the interfaces (see (2)). We shall focus on the canonical form

$$A = \begin{pmatrix} 1 & \alpha \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 1 & -\alpha \\ 0 & 1 \end{pmatrix}, \tag{6}$$

for the case of two rank-one connections and on the form

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1+\alpha \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 1 & 0 \\ 0 & 1-\alpha \end{pmatrix}, \tag{7}$$

for the case of a single rank-one connection; in both cases $\alpha \in (0,1)$ is a crystallographic parameter. The reduction to these canonical forms can in general be obtained by composing u with suitable affine deformations, an operation that however changes the shape of the domain. The canonical forms are chosen so that, in the absence of boundary conditions, an essentially one-dimensional structure is possible, with Du depending only on x_2 and taking values A and B.

Our main result is the following.

Theorem 1. *Let E_ε and K be as in (1), (3), and (5) with $n = 2$ and $\Omega = (0, L)^2$. Then there is $c > 0$ such that, for all $\varepsilon \in (0, L)$:*

(i) If A, B are as in (6), then

$$\frac{1}{c}\varepsilon^{2/3}L^{4/3} \le \min\{E_\varepsilon[u,\Omega] : u(x) = x \text{ on } \partial\Omega\} \le c\varepsilon^{2/3}L^{4/3}\,.$$

(ii) If A, B are as in (7), then

$$\frac{1}{c}\varepsilon^{4/5}L^{6/5} \le \min\{E_\varepsilon[u,\Omega] : u(x) = x \text{ on } \partial\Omega\} \le c\varepsilon^{4/5}L^{6/5}\,.$$

The constant c depends only on α.

Part (i) of the statement was first obtained in [5] and announced in [11] for a similar model. The second part was first proven in [6]. The proof is based on giving separately an upper and a lower bound on the energy, building upon the strategy of Kohn and Müller [16, 17] and following refinements [7–10, 19, 20]. We give here a short self-contained proof, discussing the upper bounds in Sect. 2, the lower bounds in Sect. 3.

Proof. Existence of minimizers is immediate by the direct method of the calculus of variations, thanks to the compact embedding of BV into L^1.

The upper bounds follow from Theorems 2 and 3, the lower bounds from Theorems 4 and 5. Notice that by scaling it suffices to prove the theorem for $L = 1$, and that since $\varepsilon \le 1$ the linear terms in the upper bounds can be absorbed into the sublinear one. □

2 Upper Bounds

In this section we show how to prove the upper bounds in Theorem 1, by constructing suitable deformations u. To clearly distinguish the horizontal and vertical directions we shall work on a rectangle $\Omega = (0, L) \times (0, H)$. In the case of two rank-one connections one can use a construction similar to the one developed in the scalar case by Kohn and Müller [17], but the different scaling in the case of one rank-one connection requires a different treatment of the period-doubling step. Further, in both cases one cannot use branching up to the boundary, since the singular perturbation penalizes all derivatives of the strain (at variance with J_ε,

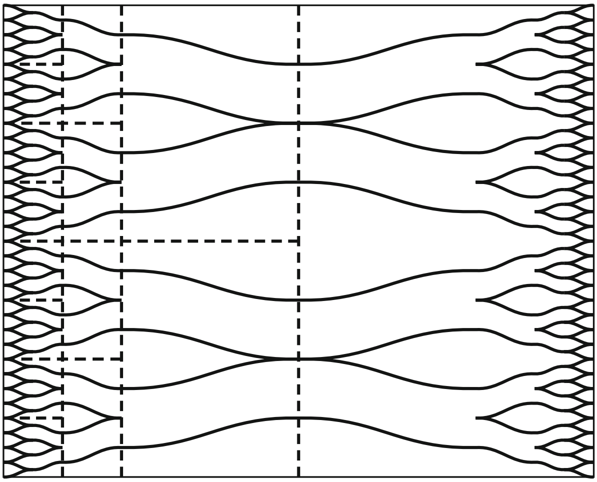


Fig. 1 Sketch of the self-similar construction used for proving the upper bound. The construction in each of the *dotted rectangles* is illustrated in Fig. 2

where only D_2D_2v enters). Therefore, we need to insert interpolation layers close to the lateral boundaries, as it was already done in [5, 19, 20]. In the whole proof of the upper bound we denote by c a positive constant which does not depend on α, ε, H and L. Its value may change from line to line.

2.1 *The Case of Two Rank-One Connections*

Theorem 2. *Let E_ε and K be as in (1), (3), (5), (6) and $\Omega = (0, L) \times (0, H)$. For all $\alpha \in (0, 1)$ and all $\varepsilon \in (0, L)$ there is $u : \Omega \to \mathbb{R}^2$ with $u(x) = x$ on $\partial\Omega$ and*

$$E_\varepsilon[u, \Omega] \leq c\alpha^{4/3}\varepsilon^{2/3}L^{1/3}H + c\alpha\varepsilon(L + H) .$$

The constant c does not depend on α, ε, L and H.

Proof. **Step 1: Decomposition of the domain.** We split Ω vertically into two equal parts Ω_l and Ω_r. We only present the construction on the left part Ω_l since the other one is similar. For a small $\xi_\tau \in (0, L/2)$ we further separate a boundary layer $\Omega_{l,l} = (0, \xi_\tau) \times (0, H)$ from the interior part $\Omega_{l,r} = (\xi_\tau, \frac{L}{2}) \times (0, H)$, where domain branching will be used. On the thin part $\Omega_{l,l}$ the deformation will simply interpolate between the left boundary data and the construction on $\Omega_{l,r}$.

Step 2: Global domain subdivision, domain branching. On $\Omega_{l,r}$ we shall use the technique of global domain branching, see Fig. 1. We shall start from the center of the domain (i.e., from the line $x_1 = L/2$) with N oscillations of period $h_0 = H/N$ in the vertical direction, and refine approaching the lateral

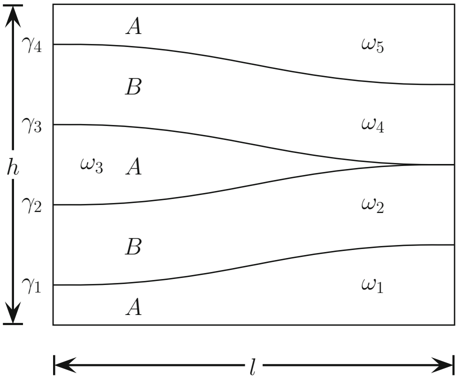


Fig. 2 Details of the period-doubling step in the construction, and subdivision of the reference *rectangle* ω into five subsets

boundaries. At refinement step i, for $i = 1, 2, \ldots, \tau$, the distance to the left boundary is $\xi_i = L\theta^i/2$ and Du has $2^i N$ oscillations, with a period $h_i = 2^{-i}h_0$. The period-doubling transition must then occur on a rectangular box of size $l_i \times h_i$, with $l_i = \xi_{i-1} - \xi_i = \theta^{i-1}(1-\theta)\frac{L}{2}$. The parameter $\theta \in (0, 1/2)$ and the number of branching blocks $N \in \mathbb{N}$ will be chosen below. This will proceed as long as $h_i \leq l_i$ (this condition defines implicitly τ). We observe that the distance to the left boundary $\xi_\tau = L\theta^\tau/2$, the width of the last refinement step l_τ and the last period h_τ differ from each other by factors of order 1. The parameter τ is the number of steps of branching towards the left boundary before coming to the interpolation; in order to have $\tau \geq 1$ we need $h_0 \leq l_0$, which corresponds (up to irrelevant factors) to $H/N \leq L$.

Step 3: Local domain subdivision. For simplicity we work on a reference rectangle $\omega = (0, l) \times (0, h)$, we can assume $h \leq l$. We split ω into five sets as follows (see Fig. 2):

$$\omega_1 = \left\{ x \in \omega \,\middle|\, 0 \leq x_2 \leq \frac{h}{8} + \gamma(x_1) \right\},$$

$$\omega_2 = \left\{ x \in \omega \,\middle|\, \frac{h}{8} + \gamma(x_1) < x_2 < \frac{3h}{8} + \gamma(x_1) \right\},$$

$$\omega_3 = \left\{ x \in \omega \,\middle|\, \frac{3h}{8} + \gamma(x_1) < x_2 < \frac{5h}{8} - \gamma(x_1) \right\},$$

$$\omega_4 = \left\{ x \in \omega \,\middle|\, \frac{5h}{8} - \gamma(x_1) < x_2 < \frac{7h}{8} - \gamma(x_1) \right\},$$

$$\omega_5 = \left\{ x \in \omega \,\middle|\, \frac{7h}{8} - \gamma(x_1) < x_2 < h \right\},$$

where the smooth function $\gamma : [0, l] \to [0, h/8]$ can be chosen so that it obeys the boundary values

$$\gamma(0) = \gamma'(0) = \gamma'(l) = 0, \quad \gamma(l) = \frac{h}{8}$$

and the estimates

$$\|\gamma\|_{L^\infty} \le ch, \quad \|\gamma'\|_{L^\infty} \le c\frac{h}{l}, \quad \|\gamma''\|_{L^\infty} \le c\frac{h}{l^2}\,.$$

In the sets ω_2 and ω_4 the strain Du will be close to B, in the other three close to A.

Step 4: Local construction and local energy. On ω we set $u_2(x) = x_2$ and

$$u_1(x) = \begin{cases} x_1 + \alpha x_2 & \text{in } \omega_1, \\ x_1 - \alpha x_2 + 2\alpha\left(\frac{h}{8} + \gamma(x_1)\right) & \text{in } \omega_2, \\ x_1 + \alpha x_2 - \frac{\alpha h}{2} & \text{in } \omega_3, \\ x_1 - \alpha x_2 + \frac{6}{8}\alpha h - 2\alpha\gamma(x_1) & \text{in } \omega_4, \\ x_1 + \alpha x_2 - \alpha h & \text{in } \omega_5. \end{cases}$$

It is easy to see that u_1 is continuous. In ω_1, ω_3 and ω_5 one has $Du = A$. In ω_2 and ω_4 we have

$$\operatorname{dist}(Du, K) \le |Du - B| \le 2\alpha\|\gamma'\|_{L^\infty}$$

and the elastic energy can be estimated by

$$\int_\omega \operatorname{dist}^2(Du, K)\, dx \le clh\alpha^2\|\gamma'\|^2_{L^\infty}\,.$$

For the surface energy we treat separately the smooth part, inside the ω_j, and the jump part on the boundaries. We obtain

$$|D^2u|(\omega) \le c \cdot \max_j |\text{highest jump of } Du| \cdot \text{length of } \gamma + c\mathscr{L}^2(\omega) \max_j \max_{\omega_j} |D^2u(x)|$$

and therefore

$$\varepsilon|D^2u|(\omega) \le c\varepsilon\alpha(h + l)(1 + \|\gamma'\|_{L^\infty}) + c\varepsilon\alpha hl\|\gamma''\|_{L^\infty}.$$

By the properties of γ stated in Step 3 we obtain

$$\int_\omega \operatorname{dist}^2(Du, K)\, dx + \varepsilon|D^2u|(\omega) \le c\alpha^2\frac{h^3}{l} + c\alpha\varepsilon l\,.$$

Step 5: Global construction and global energy on $\Omega_{l,r}$. We first extend the function u constructed in Step 4 to $(0,l) \times \mathbb{R}$, so that $u(x) - x$ is h-periodic in x_2. Continuity of the resulting function follows from the fact that $u_1(x) = x_1$ for $x_1 \in \{0,h\}$. We denote by u^i the function constructed by this procedure, working with $l = l_i$ and $h = h_i$. We define u globally by setting $u = u^i + \xi_i e_1$ on $(\xi_i, \xi_{i-1}) \times (0,H)$. One can then check that the resulting function is continuous on $\Omega_{l,r}$ (the boundary conditions on γ play a role here). The global energy on $\Omega_{l,r}$ can be computed by

$$E_\varepsilon[u,\Omega_{l,r}] \le \sum_{i=1}^{\tau} \left(c\frac{\alpha^2 H^3}{2^{2i}\theta^i N^2 L(1-\theta)} + c\alpha\varepsilon 2^i N\theta^i(1-\theta)L \right).$$

We now choose $\theta \in (\frac{1}{4}, \frac{1}{2})$, so that both geometric series are convergent, and extend the sum to ∞. Therefore,

$$E_\varepsilon[u,\Omega_{l,r}] \le c\frac{\alpha^2 H^3}{N^2 L} + c\alpha\varepsilon N L\,.$$

It remains to choose for N the integer that makes this expression smallest. Up to irrelevant factors, and recalling the requirements $N \ge 1$, $N \ge H/L$, $N \in \mathbb{N}$, this is

$$N = \left\lceil \frac{\alpha^{1/3} H}{\varepsilon^{1/3} L^{2/3}} + \frac{H}{L} \right\rceil,$$

where $\lceil t \rceil = \min\{z \in Z : z \ge t\} \in [t, t+1)$. Inserting into the previous estimate gives

$$E_\varepsilon[u,\Omega_{l,r}] \le c\alpha^{4/3}\varepsilon^{2/3} H L^{1/3} + c\alpha\varepsilon(H+L)\,.$$

Step 6: Boundary layer $\Omega_{l,l}$. For $x_1 \in (0,\xi_\tau)$ we define $u(x)$ as the affine interpolation in x_1 between the boundary values (i.e., $x_2 e_2$) and $u(\xi_\tau, x_2)$. One obtains $|Du - \mathrm{Id}| \le c\alpha$, $|D^2 u| \le \alpha/\xi_\tau$, and

$$\int_{\Omega_{l,l}} \mathrm{dist}^2(Du,K)\,dx + \varepsilon|D^2 u|(\Omega_{l,l}) \le c\xi_\tau \alpha^2 H + c\varepsilon\alpha H(1 + \frac{\xi_\tau}{h_\tau})\,.$$

Since ξ_τ, l_τ and h_τ are all of the same order, this energy is bounded (up to a factor) by the term with $i = \tau$ in the series above, hence does not modify the sum by more than a factor.

Step 7: Conclusion. The other half of the domain Ω_r is treated in the same way. At the vertical boundaries between the four subdomains interfaces arise, each with a cost of at most $c\alpha\varepsilon H$. Combining all terms the proof is concluded. □

2.2 The Case of One Rank-One Connection

Theorem 3. *Let E_ε and K be as in (1), (3), (5), (7) and $\Omega = (0, L) \times (0, H)$. For all $\alpha \in (0, 1)$ and all $\varepsilon \in (0, L)$ there is $u : \Omega \to \mathbb{R}^2$ with $u(x) = x$ on $\partial\Omega$ and*

$$E_\varepsilon[u, \Omega] \leq c\alpha^{6/5}\varepsilon^{4/5}L^{1/5}H + c\alpha\varepsilon(L + H)\,.$$

The constant c does not depend on α, ε, L and H.

First of all we point out the main difference between the two constructions. The construction in Theorem 2 involves only the u_1 component and makes no use of the rotations. Indeed, the same energy would have been obtained if the distance to $\{A, B\}$ was considered. In that construction, the partial derivative $\partial_2 u_1$ oscillates between α and $-\alpha$, generating error terms of order $\alpha h/l$ in the component $\partial_1 u_1$. These error terms generate the main contribution to the elastic energy.

In the case of a single rank-one connection, we have that $\partial_2 u_2$ oscillates between $1 + \alpha$ and $1 - \alpha$, generating error terms of order $\alpha h/l$ in the component $\partial_1 u_2$. Since this is an off-diagonal component, by $SO(2)$ invariance we can to leading order compensate for this error by using the $\partial_2 u_1$ component which will, therefore, also have oscillations of order $\alpha h/l$ (and opposite sign). In turn, this generates error terms of order $\alpha h^2/l^2$ in the $\partial_1 u_1$ component, which give the main contribution to the elastic energy (see (8) below). To exploit this fact we shall first construct u_2 oscillating as above, and then define u_1 by setting $\partial_2 u_1 = -(1 \pm \alpha)\partial_1 u_2$. Therefore, in this case we need to involve both components, and to track rotations. A similar strategy was used, with $O(2, 3)$ instead of $SO(2)$, in the case of thin-film blistering in [3, 4].

Proof. **Steps 1–3:** are identical to the proof of Theorem 2 and are not repeated.
Step 4: Local construction and local energy. On $\omega = (0, l) \times (0, h)$ we define

$$u(x) = \begin{cases} \begin{pmatrix} x_1 \\ (1+\alpha)x_2 \end{pmatrix} & x \in \omega_1, \\ \begin{pmatrix} x_1 - 2\alpha(1-\alpha)\gamma'(x_1)x_2 + 2\alpha(1-\alpha)\gamma'(x_1)\left(\frac{h}{8} + \gamma(x_1)\right) \\ (1-\alpha)x_2 + 2\alpha\left(\frac{h}{8} + \gamma(x_1)\right) \end{pmatrix} & x \in \omega_2, \\ \begin{pmatrix} x_1 - \frac{1}{2}\alpha(1-\alpha)\gamma'(x_1)h \\ (1+\alpha)x_2 - \frac{1}{2}\alpha h \end{pmatrix} & x \in \omega_3, \\ \begin{pmatrix} x_1 + 2\alpha(1-\alpha)\gamma'(x_1)x_2 - 2\alpha(1-\alpha)\gamma'(x_1)\left(\frac{7}{8}h - \gamma(x_1)\right) \\ (1-\alpha)x_2 + \frac{6}{8}\alpha h - 2\alpha\gamma(x_1) \end{pmatrix} & x \in \omega_4, \\ \begin{pmatrix} x_1 \\ (1+\alpha)x_2 - \alpha h \end{pmatrix} & x \in \omega_5\,. \end{cases}$$

The estimate of $|D^2u|$ is identical to the case of Theorem 2. The estimate for $\operatorname{dist}(Du, K)$ is however different. In ω_1 and ω_5 we have $Du = A \in K$. In ω_2 we use the freedom to choose a rotation in K to estimate

$$\begin{aligned}\operatorname{dist}(Du, K) \le & \min_{Q\in SO(2)} |Du - QB| \\ \le & \min_{\varphi\in\mathbb{R}} |\partial_1 u_1 - \cos\varphi| + |\partial_2 u_2 - (1-\alpha)\cos\varphi| \\ & + |\partial_1 u_2 - \sin\varphi| + |\partial_2 u_1 + (1-\alpha)\sin\varphi| \\ \le & |\partial_1 u_1 - 1| + |\partial_2 u_2 - (1-\alpha)| + |\partial_2 u_1 + (1-\alpha)\partial_1 u_2| + c|\partial_1 u_2|^2 ,\end{aligned} \tag{8}$$

where we have chosen φ such that $\sin\varphi = \partial_1 u_2$ if $\partial_1 u_2 \in [-1, 1]$, $\varphi = 0$ otherwise. This implies

$$\operatorname{dist}(Du, K) \le c\alpha h\|\gamma''\|_{L^\infty} + c\alpha\|\gamma'\|^2_{L^\infty} + c(\alpha\|\gamma'\|_{L^\infty})^2 \le c\alpha\frac{h^2}{l^2} \text{ in } \omega_2 .$$

The set ω_4 is similar, in ω_3 one simply uses $|Du-A| \le c\alpha\|\gamma''\|_{L^\infty} h \le c\alpha h^2/l^2$. Combining all terms, the energy in the rectangle ω can be estimated by

$$E_\varepsilon[u, \omega] \le c\alpha^2\frac{h^5}{l^3} + c\alpha\varepsilon l .$$

Step 5: Global construction and global energy on $\Omega_{l,r}$. The definition of u on $\Omega_{l,r}$ proceeds exactly as in the previous case. This time also the boundary conditions on γ' play a role in checking for continuity.

The global energy on $\Omega_{l,r}$ can be computed as in the previous case,

$$E_\varepsilon[u, \Omega_{l,r}] \le \sum_{i=1}^{\tau}\left(c\frac{\alpha^2 H^5}{2^{4i}N^4\theta^{3i}L^3(1-\theta)^3} + c\alpha\varepsilon 2^i N\theta^i(1-\theta)L\right).$$

Choosing $\theta \in (1/16, 1/2)$ both series converge. We choose

$$N = \left\lceil \frac{\alpha^{1/5}H}{\varepsilon^{1/5}L^{4/5}} + \frac{H}{L} \right\rceil$$

and estimate

$$E_\varepsilon[u, \Omega_{l,r}] \le c\alpha^{6/5}\varepsilon^{4/5}HL^{1/5} + c\varepsilon\alpha(L + H) .$$

Steps 6 and 7: are identical to the proof of Theorem 2 and not repeated. □

3 Lower Bound

Before presenting the proof in detail, we illustrate the main ideas, focusing on the case of two rank-one connections. One main ingredient in the proof of the lower bound is the observation that we can choose inside Ω an appropriate square Q such that u is approximately affine on Q. The optimal length of the sides of Q, called ℓ, will turn out to be proportional to $\varepsilon^{1/3}$. To prove this, we first observe that, since the total energy scales as $\varepsilon^{2/3}$, on a typical square of this size there is an energy of the order $\ell^2\varepsilon^{2/3}$. Since every jump of Du from one to the other well generates an energy proportional to ε times the interfacial area, we obtain that the total area of the interface inside Q is controlled by $\ell^2\varepsilon^{-1/3}$. For $\ell \leq \varepsilon^{1/3}$ this is less than the area needed to cut the cube into two halves, hence for the isoperimetric inequality (or, equivalently, the Poincaré inequality) we obtain that there is one phase which is dominant inside Q. In other words, Du is close to a fixed matrix $F \in K$ inside Q, and u is close to an affine map whose gradient is F.

Having done this, we proceed to show that the stripe connecting Q with the boundary contains high energy. Indeed, inside Q the deformation is close to a map of the form $Fx + c$, whereas on the boundary $u(x) = x$. We observe that $Ae_1 = Be_1 = e_1$ and that the energy controls the component of Du along e_1, in the sense that $|Du\, e_1| \leq |Ae_1| + \mathrm{dist}(Du, K)$. However, the integral of $Du\, e_1$ over a segment parallel to e_1 with endpoints on $\partial\Omega$ can be computed by the boundary values, and equals the integral of e_1. Therefore, the inequality is actually almost an equality. This way we obtain that u is close to x inside the domain. The key step in the proof is to make this quantitative; the difference in scaling between the two cases arises from the fact that we can estimate $u_1 - x_1$ better than $u_2 - x_2$ (see (12) and (19) for details).

3.1 The Case of Two Rank-One Connections

Theorem 4. *Let $K = SO(2)A \cup SO(2)B$, where A and B are as in (6) for some $\alpha \in (0,1)$ and let $\Omega = (0,1)^2$. Then there is $c > 0$ such that for every $u \in W^{1,2}(\Omega;\mathbb{R}^2)$ which obeys $u(x) = x$ on $\partial\Omega$ and every $\varepsilon \in (0,1)$ one has*

$$c\varepsilon^{2/3} \leq \int_\Omega \mathrm{dist}^2(Du, K)\, dx + \varepsilon|D^2u|(\Omega)\,.$$

The constant c may depend only on α.

Proof. We fix $\ell \in (0,1)$, to be chosen later, and subdivide Ω into stripes of width ℓ, of the form $S_i = \{x \in \Omega : i\ell < x_2 < (i+1)\ell\}$, $i = 0, \ldots, \lfloor 1/\ell - 1 \rfloor$. Let S be the one with the least energy, which obeys

$$E_\varepsilon[u, S] \leq 2\ell E \tag{9}$$

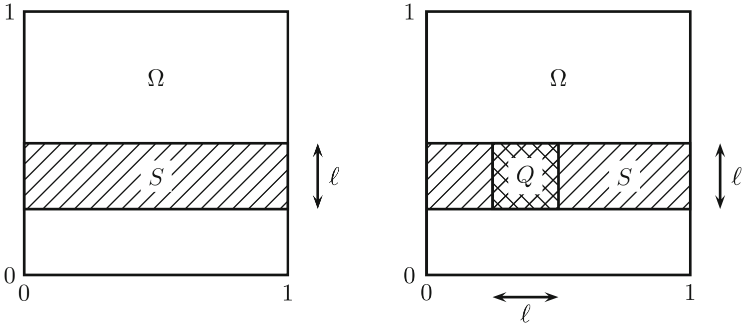


Fig. 3 Sketch of how the stripe S and the square Q are chosen inside the domain Ω

where we set for brevity $E = E_\varepsilon[u, \Omega]$ (see Fig. 3). We proceed analogously in the other direction and choose a square $Q = x_* + (0, \ell)^2 \subset S$ such that

$$E_\varepsilon[u, Q] \le 4\ell^2 E \,. \tag{10}$$

One crucial observation is that $|Fe_1| = 1$ for all $F \in K$. Indeed, $Ae_1 = Be_1 = e_1$, and the length of a vector is not influenced by rotations. This implies that $\operatorname{dist}(Du, K)$ controls the stretching of horizontal lengths,

$$|\partial_1 u_1| \le |\partial_1 u| = |Du\, e_1| \le \min_{F\in K} |Du - F| + |Fe_1| = 1 + \operatorname{dist}(Du, K)\,. \tag{11}$$

Therefore, for any $x \in \Omega$, recalling that $u(x) = x$ on $\partial\Omega$, we have

$$u_1(x) = \int_0^{x_1} \partial_1 u_1(t, x_2) dt \le x_1 + \int_0^{x_1} \operatorname{dist}(Du, K)(t, x_2) dt\,,$$

and analogously

$$u_1(x) = 1 - \int_{x_1}^1 \partial_1 u_1(t, x_2) dt \ge 1 - (1 - x_1) - \int_{x_1}^1 \operatorname{dist}(Du, K)(t, x_2) dt\,.$$

We conclude that

$$|u_1(x) - x_1| \le \int_0^1 \operatorname{dist}(Du(t, x_2), K) dt$$

for all $x \in \Omega$. Integrating over $x \in Q$ this gives

$$\int_Q |u_1(x) - x_1|\, dx \le \ell \int_S \operatorname{dist}(Du, K) dx \le \ell^{3/2} \|\operatorname{dist}(Du, K)\|_{L^2(Q)} \le 2\ell^2 E^{1/2}\,, \tag{12}$$

where we used Fubini, Hölder, and (9), respectively.

The other crucial observation is that if E is sufficiently small, then (10) implies that u is approximately affine on Q, with gradient close to K. This is essentially an isoperimetric fact: (10) implies $|D^2u|(Q) \le 4\ell^2 E/\varepsilon$; if the latter quantity is small compared to ℓ (the side length of Q) then the oscillation of Du has to be small, in an appropriate sense. In turn, if the oscillation is small, and dist(Du, K) is small (in L^2), then Du must be close (in L^2) to a fixed matrix in K. Precisely, by (10) and Poincaré's inequality there is $F \in \mathbb{R}^{2\times 2}$ such that

$$\|Du - F\|_{L^1(Q)} \le c\frac{\ell^3 E}{\varepsilon}\,. \tag{13}$$

Since

$$\|\mathrm{dist}(Du, K)\|_{L^1(Q)} \le \ell \|\mathrm{dist}(Du, K)\|_{L^2(Q)} \le 2\ell(\ell^2 E)^{1/2} = 2\ell^2 E^{1/2}$$

we conclude that

$$\ell^2 \mathrm{dist}(F, K) \le c\frac{\ell^3}{\varepsilon}E + 2\ell^2 E^{1/2}\,.$$

By (13) there is $F_* \in K$ such that $\|Du - F_*\|_{L^1(Q)} \le 2c\frac{\ell^3}{\varepsilon}E + 2\ell^2 E^{1/2}$. Another application of the Poincaré inequality shows that there is $b \in \mathbb{R}^2$ such that

$$\|u - F_* x - b\|_{L^1(Q)} \le c\frac{\ell^4}{\varepsilon}E + c\ell^3 E^{1/2}\,. \tag{14}$$

Equation (12) shows that u_1 is close to x_1, Eq. (14) shows that u is approximately affine. The lower bound comes from the fact that there is no affine function with gradient in K and close to x_1, because there is no matrix in K such that its first row is $(1, 0)$. Indeed, $F \in K$ implies $\det F = 1$ and $|Fe_2|^2 = 1 + \alpha^2$. If we had $F_{11} = 1$ and $F_{12} = 0$ then the condition on the determinant would imply $F_{22} = 1$, contradicting the condition on $|Fe_2|$. Since K is compact this means that

$$\delta_* = \min_{F\in K} |F^T e_1 - e_1| > 0\,.$$

Setting $a = F_*^T e_1 - e_1$ we obtain

$$\begin{aligned}\min_{b\in\mathbb{R}} \int_Q |(F_* x)_1 - x_1 - b|\,dx &= \min_{b\in\mathbb{R}} \int_Q |a\cdot x - b|\,dx \\ &\ge \min_{b\in\mathbb{R}} \int_{B_{\ell/2}} |a\cdot x - b|\,dx = |a| \int_{B_{\ell/2}} |x_1|\,dx = |a|\frac{\ell^3}{6} \\ &\ge \delta_* \frac{\ell^3}{6}\,.\end{aligned}$$

Therefore, with F_* and b_* as above,

$$\frac{1}{6}\delta_*\ell^3 \leq \int_Q |(F_*x)_1 + b_1 - x_1| \leq \int_Q |u_1(x) - x_1| + \int_Q |u(x) - Fx - b|dx\,.$$

Recalling (12) and (14) we conclude that

$$\frac{\delta_*}{6}\ell^3 \leq 2\ell^2 E^{1/2} + c\frac{\ell^4}{\varepsilon}E + c\ell^3 E^{1/2}\,,$$

hence at least one of

$$E \geq c\frac{\varepsilon}{\ell}\,, \qquad E \geq c\ell^2\,, \qquad E \geq c$$

must hold. Choosing $\ell = \varepsilon^{1/3}$ the proof is concluded. □

We remark that the parameter α enters the argument only through δ_*, which is a quantity of order α^2. Taking this into account, we see that the above argument gives

$$E \geq c\frac{\varepsilon\alpha^2}{\ell} \text{ or } E \geq \alpha^4\ell^2$$

which leads to $\ell \sim \varepsilon^{1/3}\alpha^{-2/3}$ (admissible for $\varepsilon < \alpha^2$) and the suboptimal bound $E \geq c\varepsilon^{2/3}\alpha^{8/3}$. The optimal scaling requires a more subtle argument, see [6].

3.2 The Case of One Rank-One Connection

Theorem 5. *Let $K = SO(2)A \cup SO(2)B$, where A and B are as in (7) for some $\alpha \in (0,1)$ and let $\Omega = (0,1)^2$. Then there is $c > 0$ such that for every $u \in W^{1,2}(\Omega;\mathbb{R}^2)$ which obeys $u(x) = x$ on $\partial\Omega$ and every $\varepsilon \in (0,1)$ one has*

$$c\varepsilon^{4/5} \leq \int_\Omega \mathrm{dist}^2(Du, K)\,dx + \varepsilon|D^2u|(\Omega)\,.$$

The constant c may depend only on α.

Proof. The general structure of the proof is similar to the previous case. We set $E = E_\varepsilon[u,\Omega]$, fix $\ell \in (0,1)$, choose a stripe $S = \{x \in \Omega : i\ell < x_2 < (i+1)\ell\}$ with

$$E_\varepsilon[u,S] \leq 2\ell E\,. \tag{15}$$

and a square $Q = x_* + (0,\ell)^2 \subset S$ with

$$E_\varepsilon[u,Q] \leq 4\ell^2 E\,. \tag{16}$$

Since the two matrices differ in the second row, we have to estimate the derivative of u_2. This is, however, not possible with the same method as above, since the length of Du_2 varies inside K. The key idea is to show first that $\partial_1 u_1$ is close to 1, which will imply – since the vector $\partial_1 u$ has length close to 1 - that the other component $\partial_1 u_2$ is small. To make this precise we write

$$\begin{aligned}|\partial_1 u_2|^2 &= |\partial_1 u|^2 - |\partial_1 u_1|^2 = (|\partial_1 u| + |\partial_1 u_1|)(|\partial_1 u| - |\partial_1 u_1|)\\ &\le 2|\partial_1 u|(|\partial_1 u| - \partial_1 u_1)\end{aligned} \tag{17}$$

and estimate, using the boundary conditions,

$$\int_0^1 (|\partial_1 u| - \partial_1 u_1)(t, x_2)\, dt = \int_0^1 (|\partial_1 u| - 1)(t, x_2)\, dt\,.$$

Since $|Fe_1| = 1$ for all $F \in K$ we obtain (dropping the argument (t, x_2) for brevity)

$$\int_0^1 (|\partial_1 u| - 1)\, dt \le \int_0^1 \min_{F\in K}(|Fe_1| + |Du - F| - 1) dt = \int_0^1 \operatorname{dist}(Du, K)\, dt\,.$$

With Hölder we obtain from (17)

$$\int_0^1 |\partial_1 u_2| dt \le \left(2\int_0^1 |\partial_1 u| dt\right)^{1/2} \left(\int_0^1 (|\partial_1 u| - \partial_1 u_1) dt\right)^{1/2},$$

and, using again $|\partial_1 u| \le 1 + \operatorname{dist}(Du, K)$,

$$\int_0^1 |\partial_1 u_2|(t, x_2) dt \le 2\left(\int_0^1 \operatorname{dist}(Du, K) dt\right)^{1/2} + 2\left(\int_0^1 \operatorname{dist}(Du, K) dt\right). \tag{18}$$

Using the boundary conditions and Hölder we deduce, for all $x \in \Omega$,

$$|u_2(x) - x_2| \le 2\left(\int_0^1 \operatorname{dist}^2(Du, K) dt\right)^{1/4} + 2\left(\int_0^1 \operatorname{dist}^2(Du, K) dt\right)^{1/2}.$$

Integrating over $x \in Q$ gives

$$\int_Q |u_2(x) - x_2|\, dx \le 2\ell^{7/4} E[u, S]^{1/4} + 2\ell^{3/2} E[u, S]^{1/2} \le 4\ell^2 E^{1/4} + 4\ell^2 E^{1/2}\,. \tag{19}$$

The same argument leading to (14) shows that there are $F_* \in K$, $b \in \mathbb{R}^2$ with

$$\|u - F_* x - b\|_{L^1(Q)} \le c\frac{\ell^4}{\varepsilon} E + c\ell^3 E^{1/2}\,. \tag{20}$$

Equation (19) shows that u_2 is close to x_2. The lower bound comes from the fact that there is no affine function with gradient in K with this property, since there is no matrix in K such that its second row is $(0, 1)$. Indeed, for all $F \in K$ the two columns are orthogonal and $\det F = 1$. The only matrix with these two properties and $F^T e_2 = e_2$ is the identity matrix, which does not belong to the compact set K. Therefore

$$\delta_* = \min_{F\in K} |F^T e_2 - e_2| > 0 \,.$$

Setting $a = F^T e_2 - e_2$ we obtain

$$\int_Q |(Fx)_2 - x_2 - b_2|\, dx = \int_Q |a\cdot x - b_2|\, dx \geq \int_{B_{\ell/2}} |a||x_1|\, dx = |a|\frac{\ell^3}{6} \geq \delta_* \frac{\ell^3}{6} \,.$$

Then

$$\frac{1}{6}\delta_*\ell^3 \leq \int_Q |(F_*x)_2 - x_2 - b_2| dx \leq \int_Q |u_2(x) - x_2| + \int_Q |u(x) - F_*x - b| dx \,.$$

Recalling (19) and (20) we conclude that

$$\frac{\delta_*}{6}\ell^3 \leq c\frac{\ell^4}{\varepsilon}E + c\ell^3 E^{1/2} + c\ell^2 E^{1/4} + c\ell^2 E^{1/2} \,,$$

hence at least one of

$$E \geq c\frac{\varepsilon}{\ell}\,, \qquad E \geq c\ell^2\,, \qquad E \geq c\ell^4\,, \qquad E \geq c$$

must hold. Choosing $\ell = \varepsilon^{1/5}$ the proof is concluded. □

4 Outlook

The results discussed here can be refined in a variety of directions. One natural question is the scaling with the parameter α. Whereas the upper bounds presented here are optimal, as discussed above, the strategy used here for proving the lower bounds does not deliver the optimal exponents. A more refined strategy is discussed in [6], for both choices of the matrices, and shows that the correct scalings are $\varepsilon^{2/3}\alpha^{4/3}$ and $\varepsilon^{4/5}\alpha^{6/5}$.

A second natural generalization is the extension to domains which are not squares. In [6] rectangles of the form $(0, l) \times (0, h)$ are treated. In the case of two rank-one connections a transition between horizontal and vertical branching appears, if h is large enough. In the case of a single rank-one connection the behavior in h is linear instead.

One may replace the regularization $\varepsilon|D^2u|(\Omega)$ with $\varepsilon^2 \int_\Omega |D^2u|^2 dx$. As in similar problems [19, 20] this only leads to small changes in the argument, which will be discussed elsewhere.

The results discussed here can be naturally extended to higher dimension and possibly to more general domains, building for the upper bound on the strategy used in [3, 4]. A more difficult question, which was recently answered positively in the simpler setting of geometrically linear elasticity by Diermeier [13], concerns the extension of the lower bounds to situations in which boundary conditions are fixed only on part of the boundary.

Acknowledgements This work was partially supported by the Deutsche Forschungsgemeinschaft through SFB 611, Project B9.

References

1. Ball, J.M., James, R.D.: Fine phase mixtures as minimizers of the energy. Arch. Ration. Mech. Anal. **100**, 13–52 (1987)
2. Ball, J.M., James, R.D.: Proposed experimental tests of a theory of fine microstructure and the two-well problem. Philos. Trans. R. Soc. Lond. A **338**, 389–450 (1992)
3. Ben Belgacem, H., Conti, S., DeSimone, A., Müller, S.: Rigorous bounds for the Föppl-von Kármán theory of isotropically compressed plates. J. Nonlinear Sci. **10**, 661–683 (2000)
4. Ben Belgacem, H., Conti, S., DeSimone, A., Müller, S.: Energy scaling of compressed elastic films. Arch. Ration. Mech. Anal. **164**, 1–37 (2002)
5. Chan, A.: Energieskalierung und Domänenverzweigung bei fest-fest Phasenübergängen mit $SO(2)$-Invarianz. Diplomarbeit, Universität Duisburg-Essen (2007)
6. Chan, A.: Energieskalierung, Gebietsverzweigung und $SO(2)$-Invarianz in einem fest-fest Phasenübergangsproblem. Ph.D. thesis, Universität Bonn (2013)
7. Choksi, R., Kohn, R.V., Otto, F.: Energy minimization and flux domain structure in the intermediate state of a type-I superconductor. J. Nonlinear Sci. **14**, 119–171 (2004)
8. Choksi, R., Conti, S., Kohn, R.V., Otto, F.: Ground state energy scaling laws during the onset and destruction of the intermediate state in a type-I superconductor. Commun. Pure Appl. Math. **61**, 595–626 (2008)
9. Conti, S.: Branched microstructures: scaling and asymptotic self-similarity. Commun. Pure Appl. Math. **53**, 1448–1474 (2000)
10. Conti, S.: A lower bound for a variational model for pattern formation in shape-memory alloys. Contin. Mech. Thermodyn. **17**, 469–476 (2006)
11. Conti, S.: Domain structures in solid-solid phase transitions. Oberwolfach Rep. **28**, 1586–1588 (2007)
12. Conti, S., Dolzmann, G., Kirchheim, B.: Existence of Lipschitz minimizers for the three-well problem in solid-solid phase transitions. Ann. I. H. Poincaré (C) Non Linear Anal. **24**, 953–962 (2007)
13. Diermeier, J.: Domain branching in geometrically linear elasticity. Master's thesis, Universität Bonn (2013,in preparation)
14. Dolzmann, G., Müller, S.: Microstructures with finite surface energy: the two-well problem. Arch. Ration. Mech. Anal. **132**, 101–141 (1995)

15. Kirchheim, B.: Lipschitz minimizers of the 3-well problem having gradients of bounded variation. Preprint 12, Max Planck Institute for Mathematics in the Sciences, Leipzig (1998). http://www.mis.mpg.de/publications/preprints/1998/prepr1998-12.html
16. Kohn, R.V., Müller, S.: Branching of twins near an austenite-twinned-martensite interface. Philos. Magazine A **66**, 697–715 (1992)
17. Kohn, R.V., Müller, S.: Surface energy and microstructure in coherent phase transitions. Commun. Pure Appl. Math. **47**, 405–435 (1994)
18. Müller, S., Šverák, V.: Convex integration with constraints and applications to phase transitions and partial differential equations. J. Eur. Math. Soc. (JEMS) **1**, 393–442 (1999)
19. Schreiber, C.: Rapport de stage D.E.A., Universität Freiburg (1994)
20. Zwicknagl, B.: Microstructures in low-hysteresis shape memory alloys: analysis and computation. Preprint 12-CNA-011, Center for Nonlinear Analysis, Carnegie-Mellon University (2012). http://www.math.cmu.edu/CNA/publications/Publications2012/011abs/011abs.html

Part III
Numerics for Multiscale Models and Singular Phenomena

On a Multilevel Preconditioner and its Condition Numbers for the Discretized Laplacian on Full and Sparse Grids in Higher Dimensions

Michael Griebel and Alexander Hullmann

Abstract We first discretize the d-dimensional Laplacian in $(0,1)^d$ for varying d on a full uniform grid and build a new preconditioner that is based on a multilevel generating system. We show that the resulting condition number is bounded by a constant that is independent of both, the level of discretization J and the dimension d. Then, we consider so-called sparse grid spaces, which offer nearly the same accuracy with far less degrees of freedom for function classes that involve bounded mixed derivatives. We introduce an analogous multilevel preconditioner and show that it possesses condition numbers which are at least as good as these of the full grid case. In fact, for sparse grids we even observe falling condition numbers with rising dimension in our numerical experiments. Furthermore, we discuss the cost of the algorithmic implementations. It is linear in the degrees of freedom of the respective multilevel generating system. For completeness, we also consider the case of a sparse grid discretization using prewavelets and compare its properties to those obtained with the generating system approach.

1 Introduction

In this paper, we deal with the preconditioning of finite element system matrices that stem from elliptic partial differential equations (PDEs) of second order. Here, we are especially interested in the higher-dimensional case. For example, high-dimensional Poisson problems and high-dimensional convection diffusion equations result from diffusion approximation techniques or the Fokker–Planck approach. Examples are the description of queueing networks [45,53], reaction mechanisms in molecular biology [54,55], or various models for the pricing of financial derivatives

M. Griebel · A. Hullmann (✉)
Institut für Numerische Simulation, Rheinische Friedrich-Wilhelms-Universität Bonn,
Wegelerstr. 6, D-53115 Bonn, Germany
e-mail: griebel@ins.uni-bonn.de; hullmann@ins.uni-bonn.de

M. Griebel (ed.), *Singular Phenomena and Scaling in Mathematical Models*,
DOI 10.1007/978-3-319-00786-1_12,

[41, 51]. Furthermore, homogenization with multiple scales [1, 17, 38, 43] as well as stochastic uncertainty quantification [2, 3, 18, 36, 39, 42, 47, 48] result in high-dimensional PDEs. Next, we find quite high-dimensional problems in quantum mechanics and particle physics. There, the dimensionality of the Schrödinger equation [44] grows with the number of considered electrons and nuclei. Then, problems in statistical mechanics lead to the Liouville equation or the Langevin equation and related phase space models where the dimension depends on the number of particles [5]. Furthermore, reinforcement learning and stochastic optimal control in continuous time give rise to the Hamilton–Jacobi–Bellman equation in high dimensions [8, 46, 56]. Finally data mining problems involve differential operators as smoothing or regularization terms (priors) whose dimension grows with the number of features of the data [23, 24, 26, 37, 52].

We want to derive multilevel preconditioners with condition numbers that are bounded independently of both, the discretization level J *and* the dimension d. Furthermore, they should possess linear cost complexity with respect to the degrees of freedom.

We will focus on the model problem of the d-dimensional Laplacian, which has been intensively analyzed in numerical analysis, albeit mostly for fixed dimension d. To this end, we first consider the simple case of a discretization based on a uniform grid using, e.g., piecewise d-linear finite elements. The solution of the resulting system of linear equations is computed iteratively. This involves the cost of a matrix-vector multiplication times the number of iterations needed to achieve a given accuracy. Here, a sparse system matrix can usually be applied with a number of floating point operations that is linear in the number of degrees of freedom. An optimal iteration count which is independent of the number of degrees of freedom is typically achieved by multiplicative multigrid methods [11,28,34,58], the additive BPX preconditioner [10, 49, 50] or wavelet-based methods. But even if the overall additive or multiplicative preconditioned matrix-vector product is linear in the number of degrees of freedom and the number of iterations is independent of the mesh width, the involved order constants are in general still dependent on the dimension d, which can be an issue in the higher-dimensional case.

Furthermore, the number of degrees of freedom itself is subject to the curse of dimension [6]. One remedy is the use of so called sparse-grid discretizations. To this end, regular sparse grids, energy sparse grids [12, 14] and general sparse grids [30, 31, 35, 40] have been employed with good success. Furthermore, space- and dimension-adaptive extensions exist [22, 25]. However, the condition number of the resulting system and the cost of a matrix-vector multiplication are now more difficult to reduce than in the regular full-grid case. For example, already for a straightforward regular sparse grid discretization, cf. [32], a simple diagonal scaling similar to the case of the BPX-preconditioner does *not* result in asymptotically bounded condition numbers in dimensions $d \geq 3$. Here, more complicated basis functions like prewavelets offer a solution [33]. Furthermore, the system matrix is not inherently sparse, and a dimension-recursive algorithm based on the so-called unidirectional principle [4,13] is needed to perform the matrix-vector-multiplication in linear time.

In this paper, we present a new additive preconditioner that is based on the multilevel idea and relies on isotropic and anisotropic subspaces. We show for the full grid case that the resulting condition number is bounded independently of the level J of the discretization and that it is also independent of the dimension d. The cost complexity is linear in the number of degrees of freedom of the enlarged generating system with a constant that grows at most polynomially in the dimension. However, it needs to be mentioned that the enlarged generating system has a factor of about 2^d more degrees of freedom than there are on the finest mesh. Our preconditioner is applicable to sparse grid discretizations as well, and the resulting condition number is now also bounded independently of J for $d \geq 3$. Furthermore, it is bounded independently of d and we even observe a falling condition number with rising dimension d. The new preconditioner can also be applied to prewavelet discretizations and then produces exactly the same condition numbers.

In Sect. 2, we introduce a multilevel discretization, and we present a norm equivalence with dimension-independent constants. Then, in Sect. 3, we introduce the full grid preconditioner with dimension-independent condition numbers for our enlarged generating system and discuss its costs. The new approach is extended to sparse grids in Sect. 4. In Sect. 5 we show that the same results can be obtained for prewavelet discretizations as well. In Sect. 6 we give numerical results that support our theory. In fact, for sparse grids, we even observe falling condition numbers with rising dimension d. Final remarks in Sect. 7 conclude the paper.

2 Discretization

We denote the unit interval by $\Omega = (0, 1)$ and its d-fold tensor product by Ω^d. The Poisson problem on Ω^d for a given right-hand side $f : \Omega^d \to \mathbb{R}$ with $\Gamma = \partial\Omega^d$ and homogeneous boundary conditions reads as

$$\begin{aligned} -\Delta u &= f \qquad \text{on} \quad \Omega^d \ , \\ u &= 0 \qquad \text{on} \quad \Gamma \ . \end{aligned} \tag{1}$$

2.1 Discretization by an Isotropic Full-Grid

Our aim is to discretize problem (1) by piecewise polynomials on a uniform grid and to precondition the resulting system of linear equations optimally not only with respect to the number of degrees of freedom, but also with respect to the dimension d. As usual, we define the bilinear form $a : H^1(\Omega^d) \times H^1(\Omega^d) \to \mathbb{R}$ as

$$a(u, v) = \int_{\Omega^d} \nabla u \cdot \nabla v \ \mathrm{d}\mathbf{x}$$

and the right-hand side $b \in H^1(\Omega^d)^*$ as

$$b(v) = \int_{\Omega^d} f \; v \; \mathrm{d}\mathbf{x} \; .$$

The weak formulation then reads: Find a solution $u \in H_0^1(\Omega^d)$ that satisfies

$$a(u, v) = b(v) \quad \text{for all} \;\; v \in H_0^1(\Omega^d) \; . \tag{2}$$

We discretize $H_0^1(\Omega^d)$ by the d-fold tensor product of one-dimensional function spaces. To this end, we first consider a one-dimensional multiresolution scale of subspaces, i.e

$$V_1 \subset V_2 \subset V_3 \subset \dots \; , \tag{3}$$

for which $\overline{V}^{\|\cdot\|_1} = H_0^1([0,1])$ holds with $V = \cup_{l=1}^{\infty} V_l$. Here, we assume

$$V_l = \operatorname{span}\{\phi_{l,i} : 1 \leq i \leq n_l\} \tag{4}$$

with $n_l = \mathscr{O}(2^l)$ locally supported basis functions $\phi_{l,i}$, $1 \leq i \leq n_l$, on level l. We define the d-dimensional tensor product space

$$V^d = V \otimes \dots \otimes V$$

and the spaces

$$V_l^d = V_l \otimes \dots \otimes V_l \; , \tag{5}$$

which are spanned by the functions

$$\phi_{l,\mathbf{i}} = \phi_{l,i_1} \cdots \phi_{l,i_d} \tag{6}$$

for $\mathbf{i} = (i_1, \dots, i_d) \in \mathbb{N}^d$ with $1 \leq i_p \leq n_l$, $p = 1, \dots, d$.

On level J, the weak problem (2) for V_J^d then leads to the system

$$\mathbf{A}_{d,J}\mathbf{x}_{d,J} = \mathbf{b}_{d,J} \tag{7}$$

of $N_{d,J} := (n_J)^d$ linear equations with

$$\mathbf{A}_{d,J} \in \mathbb{R}^{N_{d,J} \times N_{d,J}}, (\mathbf{A}_{d,J})_{\mathbf{i},\mathbf{j}} = a(\phi_{J,\mathbf{i}}, \phi_{J,\mathbf{j}})$$

and

$$\mathbf{x}_{d,J}, \mathbf{b}_{d,J} \in \mathbb{R}^{N_{d,J}} \;, (\mathbf{b}_{d,J})_{\mathbf{i}} = (f, \phi_{J,\mathbf{i}})_{L^2(\Omega^d)}$$

again for $\mathbf{i},\mathbf{j} \in \mathbb{N}^d$ with $1 \leq i_p, j_p \leq n_J, p = 1,\ldots,d$. Note that, with a lexicographic ordering of the degrees of freedom, the system matrix can be expressed as a sum of Kronecker product matrices, i.e.

$$\mathbf{A}_{d,J} = \mathbf{A}_{1,J} \otimes \mathbf{M}_{d-1,J} + \sum_{p=2}^{d-1} \mathbf{M}_{p-1,J} \otimes \mathbf{A}_{1,J} \otimes \mathbf{M}_{d-p,J} + \mathbf{M}_{d-1,J} \otimes \mathbf{A}_{1,J} \ , \tag{8}$$

where $\mathbf{A}_{1,J} \in \mathbb{R}^{n_J \times n_J}$ is the stiffness matrix of the one-dimensional problem

$$(\mathbf{A}_{1,J})_{ij} = \left(\frac{\partial \phi_{J,i}}{\partial x}, \frac{\partial \phi_{J,j}}{\partial x}\right)_{L^2(\Omega)} \quad \text{for} \quad 1 \leq i, j \leq n_J \ ,$$

and $\mathbf{M}_{p,J} \in \mathbb{R}^{(n_J)^p \times (n_J)^p}$ also has Kronecker product structure

$$\mathbf{M}_{p,J} = \bigotimes_{q=1}^{p} \mathbf{M}_{1,J}$$

with $\mathbf{M}_{1,J} \in \mathbb{R}^{n_J \times n_J}$ and

$$(\mathbf{M}_{1,J})_{ij} = (\phi_{J,i}, \phi_{J,j})_{L^2(\Omega)} \quad \text{for} \quad 1 \leq i, j \leq n_J \ . \tag{9}$$

2.2 The Multilevel Approach

The system matrix $\mathbf{A}_{d,J}$ in (7) for, e.g., linear splines, possesses a condition number that is of the order $\mathscr{O}(2^{2J})$. Thus, classical iterative solution methods for (7) like the Jacobi method, the steepest descent approach or the conjugate gradient technique converge successively slower for rising values of J. The same is true for the Gauss-Seidel and the SOR methods. This problem is remedied by a multigrid method or a multilevel preconditioner. Then, the number of iterations necessary to obtain a prescribed accuracy is bounded independently of J, cf. [9, 11, 34, 57]. To this end, besides the grid and the basis functions on the finest scale J, also the grids and basis functions on all coarser isotropic scales are included in the iterative process, i.e. the multiscale generating system

$$\bigcup_{l=1}^{J} \{ \phi_{l,\mathbf{i}} : 1 \leq i_p \leq n_l, p = 1,\ldots,d \}$$

is employed. Note that there is work that relates classical multigrid theory to multiplicative iterative algorithms operating on such a generating system [27, 28]. Furthermore, the BPX-preconditioner [10] can be identified with one step of the additive Jacobi iteration. Both methods guarantee asymptotically optimal convergence rates that are independent of J. However, the corresponding rates still depend on the dimension d.

To overcome this issue, we follow a different approach which relies on *all* coarser isotropic and anisotropic scales. To this end, we define the spaces

$$V_{\mathbf{l}} = V_{l_1} \otimes \cdots \otimes V_{l_d} \tag{10}$$

for the multiindices $\mathbf{l} = (l_1, \ldots, l_d) \in \mathbb{N}^d$. Next, we define the index sets

$$\chi_{\mathbf{l}} = \{1, \ldots, n_{l_1}\} \times \cdots \times \{1, \ldots, n_{l_d}\}$$

and the associated basis functions

$$\phi_{\mathbf{l},\mathbf{i}} = \phi_{l_1,i_1} \cdot \cdots \cdot \phi_{l_d,i_d} \quad \text{for} \quad \mathbf{i} = (i_1, \ldots, i_d) \in \chi_{\mathbf{l}} \,. \tag{11}$$

Obviously, it holds $V_{\mathbf{l}} = \operatorname{span}\{\phi_{\mathbf{l},\mathbf{i}} : \mathbf{i} \in \chi_{\mathbf{l}}\}$. From now on, $n_{\mathbf{l}} := |\chi_{\mathbf{l}}|$ denotes the number of degrees of freedom of the subspace $V_{\mathbf{l}}$. The isotropic spaces (5) can be expressed in this setting by $V_l^d = V_{\mathbf{l}}$, where $\mathbf{l} = (l, \ldots, l)$, and the isotropic functions (6) are given as $\phi_{l,\mathbf{i}} = \phi_{\mathbf{l},\mathbf{i}}$ for $\mathbf{i} \in \chi_{\mathbf{l}}$.

Our enlarged generating system includes all basis functions

$$\bigcup_{\mathbf{l} \in \mathscr{F}_J^d} \{\phi_{\mathbf{l},\mathbf{i}} : \mathbf{i} \in \chi_{\mathbf{l}}\} \,, \tag{12}$$

where the index set

$$\mathscr{F}_J^d = \{\mathbf{l} \in \mathbb{N}^d : |\mathbf{l}|_\infty \leq J\} \tag{13}$$

contains the multiindices $\mathbf{l}$ of all coarser scales, i.e. $V_{\mathbf{l}} \subset V_J^d$ for $\mathbf{l} \in \mathscr{F}_J^d$. Next, the weak problem (2) for V_J^d leads with (12) to the enlarged system

$$\hat{\mathbf{A}}_{d,J} \hat{\mathbf{x}}_{d,J} = \hat{\mathbf{b}}_{d,J} \tag{14}$$

of linear equations, with $\hat{\mathbf{A}}_{d,J} \in \mathbb{R}^{\hat{N}_{d,J} \times \hat{N}_{d,J}}$ and $\hat{\mathbf{x}}_{d,J}, \hat{\mathbf{b}}_{d,J} \in \mathbb{R}^{\hat{N}_{d,J}}$, where $\hat{N}_{d,J} := \left(\sum_{l=1}^J n_l\right)^d$. The matrix $\hat{\mathbf{A}}_{d,J}$ is block-structured with blocks $(\hat{\mathbf{A}}_{d,J})_{\mathbf{l},\mathbf{k}} \in \mathbb{R}^{n_{\mathbf{l}} \times n_{\mathbf{k}}}$ for $\mathbf{l}, \mathbf{k} \in \mathscr{F}_J^d$, where

$$((\hat{\mathbf{A}}_{d,J})_{\mathbf{l},\mathbf{k}})_{\mathbf{i},\mathbf{j}} = a(\phi_{\mathbf{l},\mathbf{i}}, \phi_{\mathbf{k},\mathbf{j}}) \quad \text{for} \quad \mathbf{i} \in \chi_{\mathbf{l}}, \mathbf{j} \in \chi_{\mathbf{k}}$$

and the right-hand side vector $\hat{\mathbf{b}}_{d,J}$ consists of blocks $(\hat{\mathbf{b}}_{d,J})_{\mathbf{l}} \in \mathbb{R}^{n_{\mathbf{l}}}$, $\mathbf{l} \in \mathscr{F}_J^d$, with

$$((\hat{\mathbf{b}}_{d,J})_{\mathbf{l}})_{\mathbf{i}} = (\phi_{\mathbf{l},\mathbf{i}}, f)_{L^2(\Omega^d)} \quad \text{for} \quad \mathbf{i} \in \chi_{\mathbf{l}} \,.$$

Note that the non-unique representation of functions in the enlarged generating system (12) results in a non-trivial kernel of $\hat{\mathbf{A}}_{d,J}$. Thus $\hat{\mathbf{A}}_{d,J}$ is not invertible. But the system (14) is nevertheless solvable since the right-hand side $\hat{\mathbf{b}}_{d,J}$ lies in the range of the system matrix. A solution can be generated by any semi-convergent iterative method [7]. Many convergence results, e.g., for the steepest descent or conjugate gradient method, also apply to the semi-definite case, cf. [28]. There, the usual condition number κ is no longer defined, but the *generalized* condition number $\tilde{\kappa}$, i.e. the ratio of the largest and the smallest *non-zero* eigenvalue, is now decisive for the speed of convergence.

Just like in (8), we can express our enlarged system matrix as the sum of Kronecker product matrices, i.e.

$$\hat{\mathbf{A}}_{d,J} = \hat{\mathbf{A}}_{1,J} \otimes \hat{\mathbf{M}}_{d-1,J} + \sum_{p=2}^{d-1} \hat{\mathbf{M}}_{p-1,J} \otimes \hat{\mathbf{A}}_{1,J} \otimes \hat{\mathbf{M}}_{d-p,J} + \hat{\mathbf{M}}_{d-1,J} \otimes \hat{\mathbf{A}}_{1,J} \ ,$$

where $\hat{\mathbf{A}}_{1,J} \in \mathbb{R}^{(\sum_{l=1}^{J} n_l)\times(\sum_{l=1}^{J} n_l)}$ is the multilevel stiffness matrix of the one-dimensional bilinear form and reads

$$(\hat{\mathbf{A}}_{1,J})_{(l,i),(k,j)} = \Big(\frac{\partial \phi_{l,i}}{\partial x}, \frac{\partial \phi_{k,j}}{\partial x}\Big)_{L^2(\Omega)} \quad \text{for} \quad 1 \le i \le n_l, 1 \le j \le n_k, 1 \le l,k \le J \ .$$

Furthermore, $\hat{\mathbf{M}}_{p,J} \in \mathbb{R}^{(\sum_{l=1}^{J} n_l)^p\times(\sum_{l=1}^{J} n_l)^p}$ also has Kronecker product structure

$$\hat{\mathbf{M}}_{p,J} = \bigotimes_{q=1}^{p} \hat{\mathbf{M}}_{1,J}$$

with $\hat{\mathbf{M}}_{1,J} \in \mathbb{R}^{(\sum_{l=1}^{J} n_l)\times(\sum_{l=1}^{J} n_l)}$ and

$$(\hat{\mathbf{M}}_{1,J})_{(l,i),(k,j)} = (\phi_{l,i}, \phi_{k,j})_{L^2(\Omega)} \quad \text{for} \quad 1 \le i \le n_l, 1 \le j \le n_k, 1 \le l,k \le J \ .$$

Of course, at some point, we need to be able to transform the non-unique solution $\hat{\mathbf{x}}_{d,J}$ of (14) to the unique solution $\mathbf{x}_{d,J}$ of (7). To this end, we assume to have matrices $\mathbf{I}_l^k \in \mathbb{R}^{n_k \times n_l}$, which are one-dimensional restrictions from level l to level k for $l > k$, prolongations from level l to level k for $l < k$ and the identity matrix for $l = k$. Note here that the $\mathbf{I}_l^k, k \neq l \pm 1$ can be expressed as just a product of successive 2-level restrictions and prolongations, respectively, i.e. we have

$$\mathbf{I}_l^k = \mathbf{I}_{k+1}^k \cdots \mathbf{I}_l^{l-1} \quad \text{for} \quad l > k \quad \text{and} \quad \mathbf{I}_l^k = \mathbf{I}_{k-1}^k \cdots \mathbf{I}_l^{l+1} \quad \text{for} \quad l < k \ . \tag{15}$$

Naturally, the multi-dimensional case is obtained by the product construction

$$\mathbf{I}_{\mathbf{l}}^{\mathbf{k}} = \bigotimes_{p=1}^{d} \mathbf{I}_{l_p}^{k_p} \ . \tag{16}$$

Then, for $\mathbf{l}, \mathbf{k} \in \mathscr{F}_J^d$, we can express any block $(\hat{\mathbf{A}}_{d,J})_{\mathbf{l},\mathbf{k}} \in \mathbb{R}^{n_\mathbf{l} \times n_\mathbf{k}}$ and any part $(\hat{\mathbf{b}}_{d,J})_\mathbf{l}$ as

$$(\hat{\mathbf{A}}_{d,J})_{\mathbf{l},\mathbf{k}} = \mathbf{I}_\mathbf{J}^\mathbf{l} \mathbf{A}_{d,J} \mathbf{I}_\mathbf{k}^\mathbf{J} \quad \text{and} \quad (\hat{\mathbf{b}}_{d,J})_\mathbf{l} = \mathbf{I}_\mathbf{J}^\mathbf{l} \mathbf{b}_{d,J} \ , \tag{17}$$

respectively, where $\mathbf{J} = (J, \ldots, J)$ is the multiindex that describes the finest level on the isotropic scale. In the special case of identity matrices, we sometimes abbreviate $\mathbf{I}_l^l$ by $\mathbf{I}_l$ and $\mathbf{I}_\mathbf{l}^\mathbf{l}$ by $\mathbf{I}_\mathbf{l}$, i.e. we drop the superscript if it is equal to the subscript.

Further, let us define the rectangular block-structured matrix

$$\hat{\mathbf{S}}_{1,J} := (\ \mathbf{I}_1^J \mid \ldots \mid \mathbf{I}_J^J\ ) \in \mathbb{R}^{n_J \times (\sum_{l=1}^J n_l)}.$$

Then, we can express the block-structured matrix $\hat{\mathbf{S}}_{d,J}$ as

$$\hat{\mathbf{S}}_{d,J} = \bigotimes_{p=1}^{d} \hat{\mathbf{S}}_{1,J} \ ,$$

and with (17) we obtain

$$\hat{\mathbf{A}}_{d,J} = \hat{\mathbf{S}}_{d,J}^T \mathbf{A}_{d,J} \hat{\mathbf{S}}_{d,J} \quad \text{and} \quad \hat{\mathbf{b}}_{d,J} = \hat{\mathbf{S}}_{d,J}^T \mathbf{b}_{d,J} \ .$$

As a result, we see that $\mathbf{x}_{d,J} = \hat{\mathbf{S}}_{d,J} \hat{\mathbf{x}}_{d,J}$ solves (7), if $\hat{\mathbf{x}}_{d,J}$ is any solution to (14). Note that we will never set up the matrices $\hat{\mathbf{S}}_{d,J}$ and $\hat{\mathbf{S}}_{d,J}^T$ in our implementation, but compute their application to vectors by a straightforward algorithm in $\mathscr{O}(d \cdot \hat{N}_{d,J})$ floating point operations using (15) and (16).

In Sect. 3, we will propose a matrix $\hat{\mathbf{C}}_{d,J}$ that can be applied cheaply to a vector and acts as a preconditioner on the enlarged system (14) with

$$\tilde{\kappa}(\hat{\mathbf{C}}_{d,J} \hat{\mathbf{A}}_{d,J}) = \mathscr{O}(1) \tag{18}$$

independently of the level J and the dimension d. Since

$$\tilde{\kappa}(\hat{\mathbf{C}}_{d,J} \hat{\mathbf{A}}_{d,J}) = \tilde{\kappa}(\hat{\mathbf{C}}_{d,J} \hat{\mathbf{S}}_{d,J}^T \mathbf{A}_{d,J} \hat{\mathbf{S}}_{d,J}) = \kappa(\hat{\mathbf{S}}_{d,J} \hat{\mathbf{C}}_{d,J} \hat{\mathbf{S}}_{d,J}^T \mathbf{A}_{d,J}) \ ,$$

we can deduce that $\mathbf{C}_{d,J} := \hat{\mathbf{S}}_{d,J} \hat{\mathbf{C}}_{d,J} \hat{\mathbf{S}}_{d,J}^T$ is thus a preconditioner for $\mathbf{A}_{d,J}$ with a resulting condition number that is bounded independently of J *and* d. Before we can present this preconditioner in Sect. 3, we need to discuss a specific norm equivalence in the next subsection.

2.3 A Norm Equivalence Based on Orthogonal Subspaces

The multiresolution scale of subspaces (3) induces a sequence of L^2-orthogonal complement spaces $(W_l)_{l=1}^{\infty}$ with

$$V_l = V_{l-1} \oplus_{L^2} W_l \quad \text{for} \quad l \geq 1, \quad \text{and} \quad V_0 := \{0\}\,. \tag{19}$$

A recursive application of (19) then yields $V_l = \oplus_{k=1}^{l} W_k$. Analogously to the anisotropic full-grid subspaces $V_{\mathbf{l}}$ in (10), we can now define anisotropic orthogonal complement spaces by the d-fold tensor products

$$W_{\mathbf{l}} = W_{l_1} \otimes \cdots \otimes W_{l_d}\ , \tag{20}$$

which satisfy $W_{\mathbf{l}} \subset V_{\mathbf{l}}$ and $W_{\mathbf{l}} \perp_{L^2} W_{\mathbf{k}}$ for $\mathbf{l} \neq \mathbf{k}$.

Now, we assume that, due to Jackson- and Bernstein-inequalities [19, 50] for the spaces $(V_l)_{l=1}^{\infty}$, we have a one-dimensional equivalence

$$\lambda_{\min} \sum_{l\in\mathbb{N}} 2^{2l} \|w_l\|^2_{L^2(\Omega)} \leq \left\| \frac{\partial u}{\partial x} \right\|^2_{L^2(\Omega)} \leq \lambda_{\max} \sum_{l\in\mathbb{N}} 2^{2l} \|w_l\|^2_{L^2(\Omega)} \tag{21}$$

for $u \in H_0^1(\Omega)$ with $u = \sum_{l\in\mathbb{N}} w_l$, where $w_l \in W_l, l \in \mathbb{N}$, and $0 < \lambda_{\min} \leq \lambda_{\max} < \infty$. In the following, we will use the symbol $\simeq$ to indicate such an equivalence and call $\lambda_{\min}$ and $\lambda_{\max}$ *norm equivalence constants*. The next theorem shows that a similar equivalence exists in higher dimensions with dimension-independent constants.

Theorem 1. *For $u \in H_0^1(\Omega^d)$, it holds that*

$$a(u,u) \simeq \sum_{\mathbf{l}\in\mathbb{N}^d} \left(\sum_{p=1}^{d} 2^{2l_p} \right) \|w_{\mathbf{l}}\|^2_{L^2(\Omega^d)} \quad \textit{for} \quad u = \sum_{\mathbf{l}\in\mathbb{N}^d} w_{\mathbf{l}} \quad \textit{with} \quad w_{\mathbf{l}} \in W_{\mathbf{l}}, \mathbf{l} \in \mathbb{N}^d\ , \tag{22}$$

where the constants $\lambda_{min}^{(d)}$ and $\lambda_{max}^{(d)}$ associated with (22) are the same as in (21), i.e. $\lambda_{min}^{(d)} = \lambda_{min}$ and $\lambda_{max}^{(d)} = \lambda_{max}$.

Proof. In (4), we have introduced $(\phi_{J,i})_{i=1}^{n_J}$ as a basis for the space V_J. Of course, there also exists a L^2-orthonormal basis $(\psi_{J,i})_{i=1}^{n_J}$ of V_J. Furthermore, we need the orthogonal decomposition $(\omega_{l,i})_{l=1}^{J}$ of $\psi_{J,i} \in V_J$ for all $i = 1, \ldots, n_J$ with $\omega_{l,i} \in W_l, l = 1, \ldots, J$ and

$$\psi_{J,i} = \sum_{l=1}^{J} \omega_{l,i}\ .$$

Next, analogously to (11), we define

$$\psi_{\mathbf{J},\mathbf{i}}(\mathbf{x}) = \psi_{J,i_1}(x_1)\dots\psi_{J,i_d}(x_d) \quad \text{and} \quad \omega_{\mathbf{l},\mathbf{i}}(\mathbf{x}) = \omega_{l_1,i_1}(x_1)\cdots\omega_{l_d,i_d}(x_d)$$

for all $\mathbf{i} \in \chi_{\mathbf{J}}$ and $\mathbf{l} \in \mathscr{F}_J^d$. This opens a direct way to find orthogonal decompositions of functions $u = \sum_{\mathbf{i}\in\chi_{\mathbf{J}}} \alpha_{\mathbf{i}}\psi_{\mathbf{J},\mathbf{i}} \in V_J^d$ by

$$u = \sum_{\mathbf{i}\in\chi_{\mathbf{J}}} \alpha_{\mathbf{i}} \sum_{\mathbf{l}\in\mathscr{F}_J^d} \omega_{\mathbf{l},\mathbf{i}} = \sum_{\mathbf{l}\in\mathscr{F}_J^d} \sum_{\mathbf{i}\in\chi_{\mathbf{J}}} \alpha_{\mathbf{i}}\omega_{\mathbf{l},\mathbf{i}} = \sum_{\mathbf{l}\in\mathscr{F}_J^d} w_{\mathbf{l}}$$

with

$$w_{\mathbf{l}} = \sum_{\mathbf{i}\in\chi_{\mathbf{J}}} \alpha_{\mathbf{i}}\omega_{\mathbf{l},\mathbf{i}} \in W_{\mathbf{l}} \tag{23}$$

for all $\mathbf{l} \in \mathscr{F}_J^d$.

Now, we show that the norm equivalence (22) holds for any $u \in V_J^d$ with the constants $\lambda_{\max}$ and $\lambda_{\min}$ from (21). We have

$$a(u,u) = \sum_{p=1}^{d} \Big(\frac{\partial}{\partial x_p} \sum_{\mathbf{i}\in\chi_{\mathbf{J}}} \alpha_{\mathbf{i}}\psi_{\mathbf{J},\mathbf{i}}, \frac{\partial}{\partial x_p} \sum_{\mathbf{j}\in\chi_{\mathbf{J}}} \alpha_{\mathbf{j}}\psi_{\mathbf{J},\mathbf{j}}\Big)_{L^2(\Omega^d)} \tag{24}$$

$$= \sum_{p=1}^{d} \sum_{\mathbf{i}\in\chi_{\mathbf{J}}} \sum_{\mathbf{j}\in\chi_{\mathbf{J}}} \Big(\frac{\partial}{\partial x_p}\alpha_{\mathbf{i}}\psi_{J,i_p}, \frac{\partial}{\partial x_p}\alpha_{\mathbf{j}}\psi_{J,j_p}\Big)_{L^2(\Omega)} \prod_{\substack{q=1\\ q\neq p}}^{d} (\psi_{J,i_q}, \psi_{J,j_q})_{L^2(\Omega)} \tag{25}$$

$$= \sum_{p=1}^{d} \sum_{\substack{\mathbf{i}'=\mathbf{i}\ominus\{i_p\}\\ \mathbf{i}\in\chi_{\mathbf{J}}}} \Big(\frac{\partial}{\partial x_p} \sum_{i_p=1}^{n_J} \alpha_{\mathbf{i}'\oplus\{i_p\}}\psi_{J,i_p}, \frac{\partial}{\partial x_p} \sum_{j_p=1}^{n_J} \alpha_{\mathbf{i}'\oplus\{j_p\}}\psi_{J,j_p}\Big)_{L^2(\Omega)} . \tag{26}$$

We obtain (25) by repeated application of the distributive law and by using the product structure of the L^2-scalar product. Then, the orthonormal basis property of the $(\psi_{J,i})_{i=1}^{n_J}$ cancels all terms for $i_q \neq j_q, q \neq p$, and we get (26). Note that

$$\mathbf{i}' := \mathbf{i} \ominus \{i_p\} = (i_1,\dots,i_{p-1},i_{p+1},\dots,i_d) \quad \text{and}$$
$$\mathbf{i}' \oplus \{i_p\} = (i_1,\dots,i_{p-1},i_p i_{p+1},\dots,i_d)\,.$$

We can apply the one-dimensional norm equivalence (21) to (26) and get the upper bound

$$\cdots \leq \sum_{p=1}^{d} \lambda_{\max} \sum_{\substack{\mathbf{i}'=\mathbf{i}\ominus\{i_p\} \\ \mathbf{i}\in\chi_{\mathbf{J}}}} \sum_{l_p=1}^{J} 2^{2l_p} \Big(\sum_{i_p=1}^{n_J} \alpha_{\mathbf{i}'\oplus\{i_p\}} \omega_{l_p,i_p}, \sum_{j_p=1}^{n_J} \alpha_{\mathbf{i}'\oplus\{j_p\}} \omega_{l_p,j_p} \Big)_{L^2(\Omega)} \tag{27}$$

$$= \lambda_{\max} \sum_{p=1}^{d} \sum_{\mathbf{i}\in\chi_{\mathbf{J}}} \sum_{\mathbf{j}\in\chi_{\mathbf{J}}} \sum_{l_p=1}^{J} 2^{2l_p} (\alpha_{\mathbf{i}} \omega_{l_p,i_p}, \alpha_{\mathbf{j}} \omega_{l_p,j_p})_{L^2(\Omega)} \cdot \prod_{\substack{q=1 \\ q\neq p}}^{d} (\psi_{J,i_q}, \psi_{J,j_q})_{L^2(\Omega)} \tag{28}$$

$$= \lambda_{\max} \sum_{p=1}^{d} \sum_{\mathbf{i}\in\chi_{\mathbf{J}}} \sum_{\mathbf{j}\in\chi_{\mathbf{J}}} \sum_{\mathbf{l}\in\mathscr{F}_J^d} 2^{2l_p} (\alpha_{\mathbf{i}} \omega_{l_p,i_p}, \alpha_{\mathbf{j}} \omega_{l_p,j_p})_{L^2(\Omega)} \cdot \prod_{\substack{q=1 \\ q\neq p}}^{d} (\omega_{l_q,i_q}, \omega_{l_q,j_q})_{L^2(\Omega)} \tag{29}$$

$$= \lambda_{\max} \sum_{\mathbf{l}\in\mathscr{F}_J^d} \Big(\sum_{p=1}^{d} 2^{2l_p} \Big) \Big(\sum_{\mathbf{i}\in\chi_{\mathbf{J}}} \alpha_{\mathbf{i}} \omega_{\mathbf{l},\mathbf{i}}, \sum_{\mathbf{j}\in\chi_{\mathbf{J}}} \alpha_{\mathbf{j}} \omega_{\mathbf{l},\mathbf{j}} \Big)_{L^2(\Omega^d)} . \tag{30}$$

In (27) and (28), we used the distributive law again and reintroduced the terms we dropped previously. In (29), we replaced the ψ_{J,i_q} and ψ_{J,j_q} by the decompositions $\sum_{l_q=1}^{J} \omega_{l_q,i_q}$ and $\sum_{l_q=1}^{J} \omega_{l_q,j_q}$, respectively. Then, in (30), we recombined the product of d one-dimensional L^2-scalar products to one d-dimensional L^2-scalar product. Note that the lower bound with $\lambda_{\min}$ can be proven in the same way. Now, in combination with (23), we know that (22) is a norm equivalence with constants $\lambda_{\max}^{(d)} \leq \lambda_{\max}$ and $\lambda_{\min}^{(d)} \geq \lambda_{\min}$.

Next, our goal is to prove the sharpness of the estimates, i.e. we will show that indeed $\lambda_{\max}^{(d)} = \lambda_{\max}$ and $\lambda_{\min}^{(d)} = \lambda_{\min}$. Since (21) holds for $\lambda_{\max}$ and $\lambda_{\min}$ on V^d, it also holds on $V_J^d \subset V^d$ with optimal constants $\lambda_{\max}(J) \leq \lambda_{\max}$ and $\lambda_{\min}(J) \geq \lambda_{\min}$. We now choose $u_{\max,J} \in V_J$ associated with the constant $\lambda_{\max}(J)$ of (21), and plug the multivariate function

$$u(\mathbf{x}) = u_{\max,J}(x_1) \cdots u_{\max,J}(x_d)$$

into (24). This results in an equality instead of an upper bound in (27) with the constant $\lambda_{\max}(J)$ instead of $\lambda_{\max}$. Because of

$$\lambda_{\max}^{(d)} \geq \lambda_{\max}(J) \nearrow \lambda_{\max} \quad \text{for} \quad J \to \infty ,$$

we can conclude that $\lambda_{\max}^{(d)} = \lambda_{\max}$. The $\lambda_{\min}^{(d)}$-case can be shown analogously. □

The norm equivalence (22) can be found in, e.g., [31] or in [33] for the $2d$-case, but so far no special attention was paid to the dimension-independence of the equivalence constants. A remark in that direction can also be found in [16].

3 A Dimension-Independent Full Grid Preconditioner

The norm equivalence (22) holds for orthogonal subspaces $(W_{\mathbf{l}})_{\mathbf{l}\in\mathscr{F}_J^d}$. In order to make this result available to our discretization, which is based on the subspaces $(V_{\mathbf{l}})_{\mathbf{l}\in\mathscr{F}_J^d}$, see Sect. 2, we need an orthogonalization operator, which will be defined in the next subsection. Then, in Sect. 3.2, we can finally present our new preconditioner.

So far, we have used d and J as subscripts to indicate the dependence on the dimension and the discretization level. For the following operators and matrices this dependence is still present, but we will omit these subscripts for better readability.

3.1 Orthogonalization Operator

We now consider the whole multivariate sequence of subspaces $V_{\mathbf{l}}, \mathbf{l} \in \mathscr{F}_J^d$, which we denote as

$$\hat{V}_J^d = (V_{\mathbf{l}})_{\mathbf{l}\in\mathscr{F}_J^d} \; .$$

For $\hat{u}, \hat{v} \in (V_{\mathbf{l}})_{\mathbf{l}\in\mathscr{F}_J^d}$, we define the scalar product

$$(\hat{u}, \hat{v})_{\hat{V}_J^d} = \sum_{\mathbf{l}\in\mathscr{F}_J^d} (u_{\mathbf{l}}, v_{\mathbf{l}})_{L^2(\Omega^d)} \quad \text{for} \quad \hat{u} = (u_{\mathbf{l}})_{\mathbf{l}\in\mathscr{F}_J^d} \quad \text{and} \quad \hat{v} = (v_{\mathbf{l}})_{\mathbf{l}\in\mathscr{F}_J^d} \; .$$

Then, we define the operator $\hat{P} : \hat{V}_J^d \to \hat{V}_J^d$ by

$$\hat{P}\hat{u} = (Q_{W_{\mathbf{l}}} u_{\mathbf{l}})_{\mathbf{l}\in\mathscr{F}_J^d} \; ,$$

where $Q_{W_{\mathbf{l}}} : V^d \to W_{\mathbf{l}}$ is the standard L^2-projection into $W_{\mathbf{l}}$, i.e. it holds

$$(Q_{W_{\mathbf{l}}} u, w_{\mathbf{l}})_{L^2(\Omega^d)} = (u, w_{\mathbf{l}})_{L^2(\Omega^d)} \quad \text{for all } \; w_{\mathbf{l}} \in W_{\mathbf{l}} \tag{31}$$

for $u \in V^d$. The following well-known Lemma 1 is the basis for an efficient computation of $Q_{W_{\mathbf{l}}} u, \mathbf{l} \in \mathscr{F}_J^d$, without an explicit discretization of the spaces $W_{\mathbf{l}}$.

Lemma 1. *There holds the identity*

$$Q_{W_{\mathbf{l}}} = (Q_{V_{l_1}} - Q_{V_{l_1-1}}) \otimes \cdots \otimes (Q_{V_{l_d}} - Q_{V_{l_d-1}}) \; ,$$

where $Q_{V_l} : V \to V_l$ denotes the one-dimensional standard L^2-projection into the space V_l.

Proof. We abbreviate $z_{\mathbf{l}} = (Q_{V_{l_1}} - Q_{V_{l_1-1}}) \otimes \cdots \otimes (Q_{V_{l_d}} - Q_{V_{l_d-1}})u$. First, we have to show that $z_{\mathbf{l}} \in W_{\mathbf{l}}$. It is obvious that $z_{\mathbf{l}} \in V_{\mathbf{l}}$, but we also have to establish the orthogonality to all $v_{\mathbf{k}} \in V_{\mathbf{k}}$ for $\mathbf{k} \leq \mathbf{l}, \mathbf{k} \neq \mathbf{l}$. To this end, let us pick an index $i \in \{1, \ldots, d\}$ with $k_i < l_i$. Then, we have

$$\begin{aligned}(z_{\mathbf{l}}, v_{\mathbf{k}})_{L^2(\Omega^d)} &= (\cdots \otimes (Q_{V_{l_i}} - Q_{V_{l_i-1}}) \otimes \ldots u, v_{\mathbf{k}})_{L^2(\Omega^d)} \\ &= (\cdots \otimes (Q_{V_{l_i}} - Q_{V_{l_i}}) \otimes \ldots u, v_{\mathbf{k}})_{L^2(\Omega^d)} = 0 \,. \qquad (32)\end{aligned}$$

In (32), we used the d-dimensional generalization of the equality

$$(Q_{V_{l-1}} u, v_k)_{L^2(\Omega)} = (u, v_k)_{L^2(\Omega)} = (Q_{V_l} u, v_k)_{L^2(\Omega)} \quad \text{for all} \quad v_k \in V_k \text{ with } l-1 \geq k \,,$$

which holds since $V_k \subset V_{l-1} \subset V$. Now, we know that $z_{\mathbf{l}} \in W_{\mathbf{l}}$, but we still need to show (31). Due to the L^2-orthogonality of $w_{\mathbf{l}} \in W_{\mathbf{l}}$ to all functions in $V_{\mathbf{k}}$ with $\mathbf{k} \leq \mathbf{l}, \mathbf{k} \neq \mathbf{l}$, it holds that

$$\begin{aligned}(z_{\mathbf{l}}, w_{\mathbf{l}})_{L^2(\Omega^d)} &= ((Q_{V_{l_1}} - Q_{V_{l_1-1}}) \otimes \cdots \otimes (Q_{V_{l_d}} - Q_{V_{l_d-1}})u, w_{\mathbf{l}})_{L^2(\Omega^d)} \\ &= ((Q_{V_{l_1}} \otimes \cdots \otimes Q_{V_{l_d}})u, w_{\mathbf{l}})_{L^2(\Omega^d)} \\ &= (u, w_{\mathbf{l}})_{L^2(\Omega^d)} \,,\end{aligned}$$

and thus we have proven that $z_{\mathbf{l}} = Q_{W_{\mathbf{l}}} u$. □

The operator $\hat{P}$ can be given in block-diagonal matrix form as $\hat{\mathbf{P}} : \mathbb{R}^{\hat{N}_{d,J} \times \hat{N}_{d,J}}$ with blocks $(\hat{\mathbf{P}})_{\mathbf{l},\mathbf{k}} \in \mathbb{R}^{n_{\mathbf{l}} \times n_{\mathbf{k}}}$ and

$$(\hat{\mathbf{P}})_{\mathbf{l},\mathbf{k}} = \begin{cases} \mathbf{Q}_{W_{\mathbf{l}}} & \text{for} \quad \mathbf{l} = \mathbf{k} \,, \\ 0 & \text{else} \end{cases} \qquad (33)$$

for all $\mathbf{l}, \mathbf{k} \in \mathscr{F}_J^d$, where $\mathbf{Q}_{W_{\mathbf{l}}} \in \mathbb{R}^{n_{\mathbf{l}} \times n_{\mathbf{l}}}$ is the matrix representation of the operator $Q_{W_{\mathbf{l}}}$ restricted to the subspace $V_{\mathbf{l}}$. According to Lemma 1, the matrices $\mathbf{Q}_{W_{\mathbf{l}}}$ can be expressed by

$$\mathbf{Q}_{W_{\mathbf{l}}} = (\mathbf{I}_{l_1} - \mathbf{I}_{l_1-1}^{l_1} (\mathbf{M}_{1,l_1-1})^{-1} \mathbf{I}_{l_1}^{l_1-1} \mathbf{M}_{1,l_1}) \otimes \cdots \otimes (\mathbf{I}_{l_d} - \mathbf{I}_{l_d-1}^{l_d} (\mathbf{M}_{1,l_d-1})^{-1} \mathbf{I}_{l_d}^{l_d-1} \mathbf{M}_{1,l_d}) \,, \qquad (34)$$

where $\mathbf{M}_{1,l}$ are the non-hierarchical isotropic mass matrices from (9) with $J = l$. Note that, besides the simple 2-level restrictions and prolongations, d applications of one-dimensional mass matrices and d applications of the inverse of one-dimensional mass matrices are employed. Both operations can be cheaply executed since only band matrices are involved here, e.g., tridiagonal matrices for linear splines.

Note furthermore that $\hat{\mathbf{P}}$ possesses the overall Kronecker product structure

$$\hat{\mathbf{P}} = \bigotimes_{p=1}^{d} \hat{\mathbf{P}}_1$$

with a block-diagonal $\hat{\mathbf{P}}_1 \in \mathbb{R}^{N_{1,J} \times N_{1,J}}$, where

$$\hat{\mathbf{P}}_1 = \mathrm{diag}(\mathbf{I}_1^1, \mathbf{I}_2^2 - \mathbf{I}_1^2(\mathbf{M}_{1,1})^{-1}\mathbf{I}_2^1\mathbf{M}_{1,2}, \ldots, \mathbf{I}_J^J - \mathbf{I}_{J-1}^J(\mathbf{M}_{1,J-1})^{-1}\mathbf{I}_J^{J-1}\mathbf{M}_{1,J}) \ .$$

The block-diagonal structure of $\hat{\mathbf{P}}$ in combination with the Kronecker product structure (34) allows for an efficient application of $\hat{\mathbf{P}}$ in our generating system, which involves $\mathscr{O}(d \cdot \hat{N}_{d,J})$ floating point operations. A more detailed cost discussion will be given in Sect. 3.3.

Note that even though the matrix $\hat{\mathbf{P}}$ is block-diagonal, it is not symmetric since its blocks on the diagonal are not symmetric. This is even more remarkable as the corresponding operator $\hat{P} : \hat{V}_J^d \to \hat{V}_J^d$ is self-adjoint. In fact, the non-symmetry is a property only of the matrix representation.

For our preconditioner, we also need to apply $\hat{\mathbf{P}}^T$ efficiently. To obtain a favorable representation of $\hat{\mathbf{P}}^T$, we first consider the mapping

$$\hat{Z} : \mathbb{R}^{\hat{N}_{d,J}} \to \hat{V}_J^d$$

that maps a block-structured vector $\hat{\mathbf{x}}_{d,J} = (x_{\mathbf{l},\mathbf{i}})_{\mathbf{i}\in\chi_\mathbf{l},\mathbf{l}\in\mathscr{F}_J^d}$ of the enlarged generating system to a collection of subspaces by

$$\hat{Z} : \hat{\mathbf{x}}_{d,J} \mapsto (\sum_{\mathbf{i}\in\chi_\mathbf{l}} x_{\mathbf{l},\mathbf{i}}\phi_{\mathbf{l},\mathbf{i}})_{\mathbf{l}\in\mathscr{F}_J^d} \ . \tag{35}$$

Note that $\hat{P}$ and $\hat{\mathbf{P}}$ are linked by $\hat{\mathbf{P}} = \hat{Z}^{-1}\hat{P}\hat{Z}$.

Lemma 2. *The adjoint $\hat{Z}^* : \hat{V}_J^d \to \mathbb{R}^{\hat{N}_{d,J}}$ of (35) is given by*

$$\hat{Z}^* : \hat{u} \mapsto \hat{\mathbf{x}}_{d,J} \quad \textit{with} \quad \hat{\mathbf{x}}_{d,J} = ((u_\mathbf{l}, \phi_{\mathbf{l},\mathbf{i}})_{L^2(\Omega^d)})_{\mathbf{i}\in\chi_\mathbf{l},\mathbf{l}\in\mathscr{F}_J^d} \quad \textit{for} \quad \hat{u} = (u_\mathbf{l})_{\mathbf{l}\in\mathscr{F}_J^d} \ .$$

Proof. For any $\hat{v} = (v_\mathbf{l})_{\mathbf{l}\in\mathscr{F}_J^d} \in \hat{V}_J^d$ and $\hat{\mathbf{x}}_{d,J} \in \mathbb{R}^{\hat{N}_{d,J}}$, we have

$$\begin{aligned}(\hat{Z}\hat{\mathbf{x}}_{d,J}, \hat{v})_{\hat{V}_J^d} &= \sum_{\mathbf{l}\in\mathscr{F}_J^d} (\sum_{\mathbf{i}\in\chi_\mathbf{l}} x_{\mathbf{l},\mathbf{i}}\phi_{\mathbf{l},\mathbf{i}}, v_\mathbf{l})_{L^2(\Omega^d)} = \sum_{\mathbf{l}\in\mathscr{F}_J^d} \sum_{\mathbf{i}\in\chi_\mathbf{l}} x_{\mathbf{l},\mathbf{i}}(v_\mathbf{l}, \phi_{\mathbf{l},\mathbf{i}})_{L^2(\Omega^d)} \\ &= (\hat{\mathbf{x}}_{d,J}, \hat{Z}^*\hat{v})_{\ell^2} \ .\end{aligned}$$

□

Now, having $\hat{Z}$ and $\hat{Z}^*$, we are able to give a computationally efficient representation of $\hat{\mathbf{P}}^T$.

Lemma 3. *It holds that*

$$\hat{\mathbf{P}}^T = \hat{\mathbf{G}}\hat{\mathbf{P}}\hat{\mathbf{G}}^{-1} ,$$

where $\hat{\mathbf{G}} : \mathbb{R}^{\hat{N}_{d,J}\times\hat{N}_{d,J}}$ *is a block-diagonal matrix with blocks* $(\hat{\mathbf{G}})_{\mathbf{l},\mathbf{k}} \in \mathbb{R}^{n_\mathbf{l}\times n_\mathbf{k}}$ *and*

$$(\hat{\mathbf{G}})_{\mathbf{l},\mathbf{k}} = \begin{cases} \mathbf{M}_\mathbf{l} & \textit{for} \quad \mathbf{l} = \mathbf{k} , \\ 0 & \textit{else} \end{cases}$$

for all $\mathbf{l}, \mathbf{k} \in \mathscr{F}_J^d$ *with the mass matrices* $\mathbf{M}_\mathbf{l} = \bigotimes_{p=1}^d \mathbf{M}_{1,l_p}$.

Proof. It holds that

$$\begin{aligned}(\hat{Z}^*\hat{Z}\hat{\mathbf{x}}_{d,J}, \hat{\mathbf{y}}_{d,J})_{\ell^2} = (\hat{Z}\hat{\mathbf{x}}_{d,J}, \hat{Z}\hat{\mathbf{y}}_{d,J})_{\hat{V}_J^d} &= \sum_{\mathbf{l}\in\mathscr{F}_J^d} \sum_{\mathbf{i},\mathbf{j}\in\chi_\mathbf{l}} x_{\mathbf{l},\mathbf{i}}(\phi_{\mathbf{l},\mathbf{i}}, \phi_{\mathbf{l},\mathbf{j}})_{L^2(\Omega^d)} y_{\mathbf{l},\mathbf{j}} \\ &= \hat{\mathbf{x}}_{d,J}^T \hat{\mathbf{G}}\hat{\mathbf{y}}_{d,J} ,\end{aligned}$$

and thus $\hat{Z}^*\hat{Z} = \hat{\mathbf{G}}$. Then, we can infer

$$(\hat{\mathbf{P}})^T = (\hat{Z}^{-1}\hat{P}\hat{Z})^* = \hat{Z}^*\hat{P}(\hat{Z}^{-1})^* = \hat{\mathbf{G}}\hat{Z}^{-1}\hat{P}\hat{Z}\hat{\mathbf{G}}^{-1} = \hat{\mathbf{G}}\hat{\mathbf{P}}\hat{\mathbf{G}}^{-1} . \qquad \square$$

Note furthermore that the operator $\hat{P}$ is a projection, i.e. $\hat{P}\hat{P} = \hat{P}$. The same is true for $\hat{\mathbf{P}}$ since

$$\hat{\mathbf{P}}\hat{\mathbf{P}} = \hat{Z}^{-1}\hat{P}\hat{Z}\hat{Z}^{-1}\hat{P}\hat{Z} = \hat{Z}^{-1}\hat{P}\hat{P}\hat{Z} = \hat{Z}^{-1}\hat{P}\hat{Z} = \hat{\mathbf{P}} .$$

Finally, we need the following Lemma.

Lemma 4. *For a block-diagonal scaling matrix* $\hat{\mathbf{D}} \in \mathbb{R}^{\hat{N}_{d,J}\times\hat{N}_{d,J}}$ *with blocks* $(\hat{\mathbf{D}})_{\mathbf{l},\mathbf{k}} \in \mathbb{R}^{n_\mathbf{l}\times n_\mathbf{k}}$ *for* $\mathbf{l}, \mathbf{k} \in \mathscr{F}_J^d$ *and*

$$(\hat{\mathbf{D}})_{\mathbf{l},\mathbf{k}} = \begin{cases} c_\mathbf{l}\mathbf{I}_\mathbf{l} & \textit{for} \quad \mathbf{l} = \mathbf{k} , \\ 0 & \textit{else} , \end{cases}$$

the matrix $\hat{\mathbf{D}}$ *commutes with any other block-diagonal matrix* $\hat{\mathbf{B}} \in \mathbb{R}^{\hat{N}_{d,J}\times\hat{N}_{d,J}}$*, i.e. a block-structured matrix with blocks* $(\hat{\mathbf{B}})_{\mathbf{l},\mathbf{k}} \in \mathbb{R}^{n_\mathbf{l}\times n_\mathbf{k}}$ *for* $\mathbf{l}, \mathbf{k} \in \mathscr{F}_J^d$*, where*

$$(\hat{\mathbf{B}})_{\mathbf{l},\mathbf{k}} = \begin{cases} \mathbf{B}_\mathbf{k} & \textit{for} \quad \mathbf{l} = \mathbf{k} , \\ 0 & \textit{else} , \end{cases}$$

and $\mathbf{B}_\mathbf{k} \in \mathbb{R}^{n_\mathbf{k}\times n_\mathbf{k}}$ *are general matrices.*

Proof. For ease of notation, we use Kronecker's δ in this short proof. It holds that

$$\begin{aligned}(\hat{\mathbf{D}}\hat{\mathbf{B}})_{\mathbf{l},\mathbf{k}} &= \sum_{\mathbf{m}\in\mathscr{I}_J^d} (\hat{\mathbf{D}})_{\mathbf{l},\mathbf{m}}(\hat{\mathbf{B}})_{\mathbf{m},\mathbf{k}} = \sum_{\mathbf{m}\in\mathscr{I}_J^d} \delta_{\mathbf{l},\mathbf{m}} c_{\mathbf{l}} \delta_{\mathbf{m},\mathbf{k}} \mathbf{B}_{\mathbf{k}} = \delta_{\mathbf{l},\mathbf{k}} c_{\mathbf{l}} \mathbf{B}_{\mathbf{k}} \\ &= \delta_{\mathbf{l},\mathbf{k}} \mathbf{B}_{\mathbf{l}} c_{\mathbf{k}} = \sum_{\mathbf{m}\in\mathscr{I}_J^d} \delta_{\mathbf{l},\mathbf{m}} \mathbf{B}_{\mathbf{m}} \delta_{\mathbf{m},\mathbf{k}} c_{\mathbf{m}} = \sum_{\mathbf{m}\in\mathscr{I}_J^d} (\hat{\mathbf{B}})_{\mathbf{l},\mathbf{m}}(\hat{\mathbf{D}})_{\mathbf{m},\mathbf{k}} = (\hat{\mathbf{B}}\hat{\mathbf{D}})_{\mathbf{l},\mathbf{k}} \,,\end{aligned}$$

and thus $\hat{\mathbf{D}}\hat{\mathbf{B}} = \hat{\mathbf{B}}\hat{\mathbf{D}}$. □

Obviously, Lemma 4 can be applied to, e.g., $\hat{\mathbf{B}} = \hat{\mathbf{P}}$ or $\hat{\mathbf{B}} = \hat{\mathbf{G}}$.

3.2 Preconditioner

Now we will present our new preconditioner for the operator matrix $\hat{\mathbf{A}}_{d,J}$. To this end, the most important ingredient is the norm equivalence (22). From Theorem 1 we already know that its constants are independent of the dimension d and bounded independently of J.

Theorem 2. *Let $\hat{\mathbf{D}} \in \mathbb{R}^{\hat{N}_{d,J}\times\hat{N}_{d,J}}$ be a diagonal block-structured scaling matrix with blocks $(\hat{\mathbf{D}})_{\mathbf{l},\mathbf{k}} \in \mathbb{R}^{n_{\mathbf{l}}\times n_{\mathbf{k}}}$ and*

$$(\hat{\mathbf{D}})_{\mathbf{l},\mathbf{k}} = \begin{cases} (\sum_{p=1}^{d} 2^{2l_p})\mathbf{I}_{\mathbf{l}} & \textit{for} \quad \mathbf{l} = \mathbf{k} \,, \\ 0 & \textit{else} \end{cases}$$

for all $\mathbf{l}, \mathbf{k} \in \mathscr{I}_J^d$. Then, the generalized condition number of the symmetric matrix

$$\hat{\mathbf{L}}^{-1}\hat{\mathbf{P}}^T\hat{\mathbf{D}}^{-1/2}\hat{\mathbf{A}}_{d,J}\hat{\mathbf{D}}^{-1/2}\hat{\mathbf{P}}\hat{\mathbf{L}}^{-T} \tag{36}$$

is bounded asymptotically with respect to J and is completely independent of the dimension d. Here, $\hat{\mathbf{L}}$ is the Cholesky-factor of $\hat{\mathbf{G}}$, i.e. $\hat{\mathbf{G}} = \hat{\mathbf{L}}\hat{\mathbf{L}}^T$.

Proof. For any block-structured vector $\hat{\mathbf{x}}_{d,J} \in \operatorname{im}\hat{\mathbf{P}}$, we have

$$\begin{aligned}\hat{\mathbf{x}}_{d,J}^T\hat{\mathbf{P}}^T\hat{\mathbf{A}}_{d,J}\hat{\mathbf{P}}\hat{\mathbf{x}}_{d,J} &= \hat{\mathbf{x}}_{d,J}^T\hat{\mathbf{A}}_{d,J}\hat{\mathbf{x}}_{d,J} \\ &= a(\sum_{\mathbf{l}\in\mathscr{I}_J^d}\sum_{\mathbf{i}\in\chi_{\mathbf{l}}} x_{\mathbf{l},\mathbf{i}}\phi_{\mathbf{l},\mathbf{i}}, \sum_{\mathbf{l}\in\mathscr{I}_J^d}\sum_{\mathbf{i}\in\chi_{\mathbf{l}}} x_{\mathbf{l},\mathbf{i}}\phi_{\mathbf{l},\mathbf{i}})\end{aligned} \tag{37}$$

$$\simeq \sum_{\mathbf{l}\in\mathscr{F}_J^d} \Big(\sum_{p=1}^{d} 2^{2l_p}\Big) \Big\| \sum_{\mathbf{i}\in\chi_\mathbf{l}} x_{\mathbf{l},\mathbf{i}}\phi_{\mathbf{l},\mathbf{i}} \Big\|^2_{L^2(\Omega^d)} \tag{38}$$

$$= \sum_{\mathbf{l}\in\mathscr{F}_J^d} \Big(\sum_{p=1}^{d} 2^{2l_p}\Big) \mathbf{x}_\mathbf{l}^T \mathbf{M}_\mathbf{l} \mathbf{x}_\mathbf{l}$$

$$= \hat{\mathbf{x}}_{d,J}^T \hat{\mathbf{D}}\hat{\mathbf{G}}\hat{\mathbf{x}}_{d,J} \ . \tag{39}$$

In (37), we have used $\hat{\mathbf{P}}\hat{\mathbf{x}}_{d,J} = \hat{\mathbf{x}}_{d,J}$ and in (38), we have applied the norm equivalence (22). The levelwise summation of the mass matrix products was then expressed using the matrix $\hat{\mathbf{G}}$ in (39). In the following, we need the block-diagonal factor $\hat{\mathbf{L}}$ of the Cholesky-decomposition

$$\hat{\mathbf{G}} = \hat{\mathbf{L}}\hat{\mathbf{L}}^T \ .$$

We set $\hat{\mathbf{y}}_{d,J} = \hat{\mathbf{L}}^T\hat{\mathbf{D}}^{1/2}\hat{\mathbf{x}}_{d,J}$ and obtain with $\hat{\mathbf{D}}\hat{\mathbf{G}} = \hat{\mathbf{D}}^{1/2}\hat{\mathbf{G}}\hat{\mathbf{D}}^{1/2}$, see Lemma 4, the equation

$$\hat{\mathbf{x}}_{d,J}^T\hat{\mathbf{D}}\hat{\mathbf{G}}\hat{\mathbf{x}}_{d,J} = \hat{\mathbf{x}}_{d,J}^T\hat{\mathbf{D}}^{1/2}\hat{\mathbf{L}}\hat{\mathbf{L}}^T\hat{\mathbf{D}}^{1/2}\hat{\mathbf{x}}_{d,J} = \hat{\mathbf{y}}_{d,J}^T\hat{\mathbf{y}}_{d,J} \ .$$

Then, using the equivalence of (37) and (39), and $\hat{\mathbf{x}}_{d,J} = \hat{\mathbf{D}}^{-1/2}\hat{\mathbf{L}}^{-T}\hat{\mathbf{y}}_{d,J}$, we obtain the relation

$$\hat{\mathbf{y}}_{d,J}^T\hat{\mathbf{L}}^{-1}\hat{\mathbf{D}}^{-1/2}\hat{\mathbf{P}}^T\hat{\mathbf{A}}_{d,J}\hat{\mathbf{P}}\hat{\mathbf{D}}^{-1/2}\hat{\mathbf{L}}^{-T}\hat{\mathbf{y}}_{d,J} \simeq \hat{\mathbf{y}}_{d,J}^T\hat{\mathbf{y}}_{d,J} \tag{40}$$

for all $\hat{\mathbf{y}}_{d,J} \in \operatorname{im}\hat{\mathbf{L}}^T\hat{\mathbf{D}}^{1/2}\hat{\mathbf{P}}$ with the same favorable constants as in (22). With the commuting of the matrices $\hat{\mathbf{P}}$ and $\hat{\mathbf{D}}^{-1/2}$, see Lemma 4, the left-hand side of (40) leads to (36).

Finally, we have to show that no $\hat{\mathbf{v}}_{d,J} \in \mathbb{R}^{\hat{N}_{d,J}}$ with $\hat{\mathbf{v}}_{d,J} \perp \hat{\mathbf{y}}_{d,J}$ affects the spectrum. From the Fundamental Theorem of Linear Algebra, from $\hat{\mathbf{P}}^T = \hat{\mathbf{G}}\hat{\mathbf{P}}\hat{\mathbf{G}}^{-1}$, see Lemma 3, and from $\hat{\mathbf{P}}^T\hat{\mathbf{D}}^{1/2} = \hat{\mathbf{D}}^{1/2}\hat{\mathbf{P}}^T$ we know that

$$\begin{aligned}\hat{\mathbf{v}}_{d,J} \in \ker\hat{\mathbf{P}}^T\hat{\mathbf{D}}^{1/2}\hat{\mathbf{L}} &= \ker\hat{\mathbf{D}}^{1/2}\hat{\mathbf{G}}\hat{\mathbf{P}}\hat{\mathbf{G}}^{-1}\hat{\mathbf{L}} = \ker\hat{\mathbf{D}}^{1/2}\hat{\mathbf{G}}\hat{\mathbf{P}}\hat{\mathbf{L}}^{-T}\hat{\mathbf{L}}^{-1}\hat{\mathbf{L}} \\ &= \ker\hat{\mathbf{P}}\hat{\mathbf{L}}^{-T} \ . \end{aligned} \tag{41}$$

We dropped the matrix $\hat{\mathbf{D}}^{1/2}\hat{\mathbf{G}}$ from the kernel in the last equality (41), as it is a full-rank matrix and thus has no effect on the kernel. Obviously, if $\hat{\mathbf{v}}_{d,J} \in \ker\hat{\mathbf{P}}\hat{\mathbf{L}}^{-T}$, then $\hat{\mathbf{v}}_{d,J}$ belongs to the kernel of the preconditioned system (36). This finally proves the theorem. □

As a result of Theorem 2, we can express our preconditioner for $\hat{\mathbf{A}}_{d,J}$ as

$$\hat{\mathbf{C}}_{d,J} := \hat{\mathbf{P}}\hat{\mathbf{D}}^{-1}\hat{\mathbf{G}}^{-1}\hat{\mathbf{P}}^T \ .$$

Moreover, this approach also gives us with $\hat{\mathbf{A}}_{d,J} = \hat{\mathbf{S}}_{d,J}^T \mathbf{A}_{d,J} \hat{\mathbf{S}}_{d,J}$ the preconditioner

$$\mathbf{C}_{d,J} := \hat{\mathbf{S}}_{d,J} \hat{\mathbf{P}} \hat{\mathbf{D}}^{-1} \hat{\mathbf{G}}^{-1} \hat{\mathbf{P}}^T \hat{\mathbf{S}}_{d,J}^T \tag{42}$$

for the fine grid system matrix $\mathbf{A}_{d,J}$. The preconditioned system possesses the same condition number, i.e. it is also independent of d and bounded independently of J.

3.3 Cost Discussion for the New Preconditioner

So far, we obtained a preconditioner with condition numbers independent of d and bounded independently of J. Of course, the question is now how high its computational costs are. Remember that a perfect preconditioner would be $\mathbf{A}_{d,J}^{-1}$ anyway, but it involves way too many computations. With (42) we now have a preconditioner $\mathbf{C}_{d,J}$ which comes, up to a d- and J-independent constant, close to $\mathbf{A}_{d,J}^{-1}$, but involves only a number of floating point operations that is linear in the number of degrees of freedom $\hat{N}_{d,J}$ of the enlarged system.

We will now give a short discussion of the required matrix-vector multiplications and their costs, also with respect to the dimension d. As stated earlier, the application of the matrices $\hat{\mathbf{S}}_{d,J}$ and $\hat{\mathbf{S}}_{d,J}^T$ onto a vector can be implemented by a simple algorithm that exploits (15) in $\mathcal{O}(d \cdot \hat{N}_{d,J})$ floating point operations. The application of $\hat{\mathbf{D}}^{-1}$ is obviously possible with the same cost complexity. The matrix $\hat{\mathbf{G}}^{-1}$ needs however a more elaborate discussion. As it is block-diagonal, its action can be implemented with an algorithm that works subspace by subspace. On every $V_{\mathbf{l}}, \mathbf{l} \in \mathscr{F}_J^d$, the mass matrix $\mathbf{M}_{\mathbf{l}} = \bigotimes_{p=1}^d \mathbf{M}_{1,l_p}$ must be inverted. As these matrices have Kronecker product structure, the inversion can be realized by the application of $\mathbf{M}_{1,l_p}^{-1}$ to the dimension p for $p = 1, \dots, d$. We assume the functions $\{\phi_{\mathbf{l},\mathbf{i}}\}_{\mathbf{i} \in \chi_{\mathbf{l}}}$ to be of finite element type (h-version with fixed polynomial degree) having local support. Consequently, the associated one-dimensional matrices $\mathbf{M}_{1,l_p}$ have band matrix structure with constant band size and are thus invertible with linear costs.[1] As a result, we have a cost of $\mathcal{O}(d \cdot n_{\mathbf{l}})$ on each subspace and obtain a cost complexity of $\mathcal{O}(d \cdot \hat{N}_{d,J})$ in total. The same argumentation holds for $\hat{\mathbf{P}}$, which has a somewhat more complicated form, see (33) and (34), but also works subspace by subspace, where we can again exploit a Kronecker product structure. In total, we arrive at costs of $\mathcal{O}(d \cdot \hat{N}_{d,J})$ for our preconditioner. The application of $\mathbf{A}_{d,J}$ is directly

[1] Non-local basis functions (p-version) are likely to result in a Toeplitz-type matrix, which can be inverted in log-linear time.

possible[2] in $\mathscr{O}(d^2 \cdot N_{d,J})$ due to the representation of the system matrix as a sum of Kronecker product matrices (8). In comparison, our preconditioner (42) is slightly more expensive since its costs depend on the *enlarged* system with $\hat{N}_{d,J}$ degrees of freedom. However, a geometric series argument shows that

$$\hat{N}_{d,J} = \mathscr{O}(2^d N_{d,J}) = \mathscr{O}(2^d 2^{Jd}) = \mathscr{O}(2^{(J+1)d}) = \mathscr{O}(N_{d,J+1}) ,$$

and thus the costs for our preconditioner on level J compare simply to the costs for a regular fine grid system on level $J + 1$.

4 Sparse Grids

So far, we have dealt with the preconditioning of an isotropic full grid with $\mathscr{O}(N_{d,J})$ degrees of freedom. They scale exponentially with the dimension d and are thus impossible to deal with for $d > 4$ anyway. Under some additional smoothness requirements, sparse grids [15] remove this curse of dimension to some extent. Then, the multivariate multilevel structure is a fundamental necessity for both, a good preconditioner and the discretization itself. The implementation of a sparse grid multilevel discretization was already dealt with in [15,22]. In the following, we discuss our new preconditioner for the sparse grid case in detail.

4.1 Definition

We can use an index set $\mathscr{I} \subset \mathbb{N}^d$, $|\mathscr{I}| < \infty$, which defines the subspaces included in some discretization by

$$V_{\mathscr{I}} := \sum_{\mathbf{l} \in \mathscr{I}} V_{\mathbf{l}} .$$

A proper choice of $\mathscr{I}$ now depends – besides the error we want to achieve – on the smoothness of the function class[3] for which we want to approximate.

[2] Note that it is even possible to execute this matrix-vector product in $\mathscr{O}(d \cdot N_{d,J})$ operations by the successive multiplications of $\mathbf{M}_{d,J}$ and of $\mathbf{A}_{d,J}\mathbf{M}_{d,J}^{-1} = \sum_{p=1}^{d}(\bigotimes_{q=1}^{p-1} \mathbf{I}_J) \otimes \mathbf{A}_{1,J}\mathbf{M}_{1,J}^{-1} \otimes (\bigotimes_{q=p+1}^{d} \mathbf{I}_J)$.

[3] In this paper, we restrict ourselves to homogeneous boundary conditions and do not introduce functions at the boundary. However, by $u = u_{\Omega^d} + u_{\Gamma}$ with $u_{\Omega^d}|_{\Gamma} = 0$ and $-\Delta u_{\Omega^d} = f + \Delta u_{\Gamma}$, we cover any case with Dirichlet boundary functions $u_{\Gamma} = g$ on Γ.

For example, the full grid space V_J^d from (5) can be described by the index set $\mathscr{F}_J^d$ from (13), i.e. $V_J^d = V_{\mathscr{F}_J^d}$, and has the approximation property[4]

$$\inf_{v \in V_{\mathscr{F}_J^d}} \|u - v\|^2_{H^s(\Omega^d)} \leq c(d) 2^{-2(t-s)J} \|u\|^2_{H^t(\Omega^d)}$$

with rate $t - s$ and $u \in H_0^t(\Omega^d)$. Its number of degrees of freedom is of the order $\mathscr{O}(2^{Jd})$. Thus, the accuracy as function of the degrees of freedom deteriorates exponentially with rising d, which resembles the well-known 'curse of dimensionality', cf. [6, 15].

The sparse grid index set

$$\mathscr{S}_J^d = \{\mathbf{l} \in \mathbb{N}^d : |\mathbf{l}|_1 \leq J + d - 1\} \tag{43}$$

circumvents this problem to some extent provided that additional mixed smoothness $u \in H^t_{0,\mathrm{mix}}(\Omega^d)$ is present. For details, see [15]. An example for the function system associated to $\mathscr{S}_J^d$ is given in Fig. 1 (right) for the two-dimensional case. The associated rate of best approximation

$$\inf_{v \in V_{\mathscr{S}_J^d}} \|u - v\|^2_{H^s(\Omega^d)} \leq c(d) 2^{-2(t-s)J} \|u\|^2_{H^t_{\mathrm{mix}}(\Omega^d)}$$

is the same[5] as for the full grid space, i.e. $t-s$, but the number of degrees of freedom now is only of the order $\mathscr{O}(2^J J^{d-1})$ in J. This is a substantial improvement of the asymptotics in J in comparison to the full grid case. A-priori H^s-optimized sparse grids need, depending on the available smoothness class, even less degrees of freedom. For further details, cf. [15, 31].

It is furthermore possible to adapt the index set $\mathscr{I}$ a-posteriori to a given function by means of a proper error estimation and a successive refinement procedure. This approach results in adaptively refined sparse grids, see e.g. [22, 25]. Note that for both, practical and theoretical reasons, our index set $\mathscr{I}$ needs to satisfy the *admissibility condition*

$$\mathbf{l} \in \mathscr{I}, \mathbf{k} \in \mathbb{N}^d, \mathbf{k} \leq \mathbf{l} \Rightarrow \mathbf{k} \in \mathscr{I} \ . \tag{44}$$

The number of degrees of freedom in the enlarged system for the regular sparse grid space $V_{\mathscr{S}_J^d}$ is $\hat{N}^{\mathrm{SG}}_{d,J} = \sum_{\mathbf{l} \in \mathscr{S}_J^d} n_{\mathbf{l}}$. Note that here again some redundancy is involved, but the asymptotics in J of the number of degrees of freedom remains the

[4] This holds for a range of parameters $0 \leq s < t \leq r$ with r being the order of the spline of the space construction. In our case of linear splines $r = 2$ holds.

[5] Note that an additional logarithmic term appears in the error estimate for $s = 0$, cf. [15].

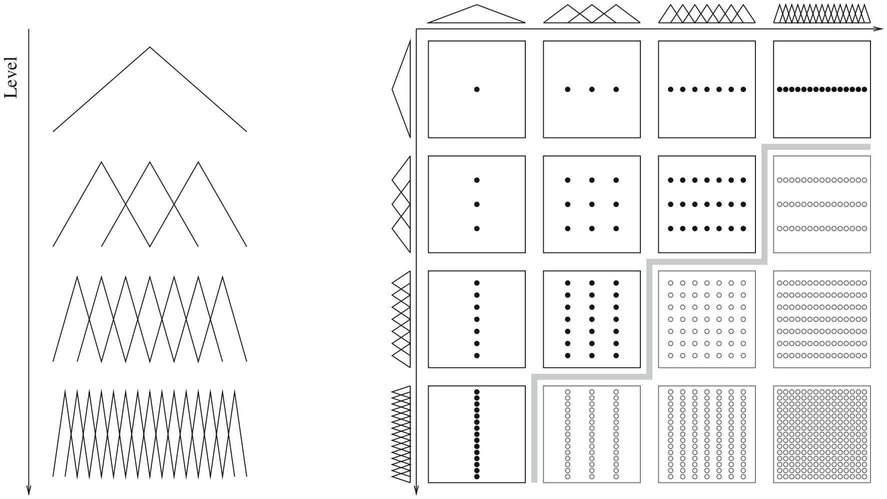


Fig. 1 The first four levels of a one-dimensional multilevel generating system based on linear splines (*left*). Two-dimensional tensorization and the sparse subspace (*right*)

same as for the sparse grid approach based on, e.g., the hierarchical basis [15], that is $\hat{N}^{\mathrm{SG}}_{d,J} = \mathcal{O}(2^J J^{d-1})$ in J.

The weak problem (2) on $V_{\mathcal{S}^d_J}$ with the generating system

$$\bigcup_{\mathbf{l}\in\mathcal{S}^d_J} \{\phi_{\mathbf{l},\mathbf{i}} : \mathbf{i} \in \chi_{\mathbf{l}}\} \tag{45}$$

now leads to the equation

$$\hat{\mathbf{A}}^{\mathrm{SG}}_{d,J}\hat{\mathbf{x}}^{\mathrm{SG}}_{d,J} = \hat{\mathbf{b}}^{\mathrm{SG}}_{d,J} \ . \tag{46}$$

Here, the matrix $\hat{\mathbf{A}}^{\mathrm{SG}}_{d,J}$ is block-structured with blocks $(\hat{\mathbf{A}}^{\mathrm{SG}}_{d,J})_{\mathbf{l},\mathbf{k}} \in \mathbb{R}^{n_{\mathbf{l}}\times n_{\mathbf{k}}}$ for $\mathbf{l},\mathbf{k} \in \mathcal{S}^d_J$, where

$$((\hat{\mathbf{A}}^{\mathrm{SG}}_{d,J})_{\mathbf{l},\mathbf{k}})_{\mathbf{i},\mathbf{j}} = a(\phi_{\mathbf{l},\mathbf{i}}, \phi_{\mathbf{k},\mathbf{j}}) \quad \text{for} \quad \mathbf{i} \in \chi_{\mathbf{l}}, \mathbf{j} \in \chi_{\mathbf{k}}$$

and the right-hand side vector $\hat{\mathbf{b}}^{\mathrm{SG}}_{d,J}$ consists of blocks $(\hat{\mathbf{b}}^{\mathrm{SG}}_{d,J})_{\mathbf{l}} \in \mathbb{R}^{n_{\mathbf{l}}}$, $\mathbf{l} \in \mathcal{S}^d_J$, with

$$((\hat{\mathbf{b}}^{\mathrm{SG}}_{d,J})_{\mathbf{l}})_{\mathbf{i}} = (\phi_{\mathbf{l},\mathbf{i}}, f)_{L^2(\Omega^d)} \quad \text{for} \quad \mathbf{i} \in \chi_{\mathbf{l}} \ .$$

Similar to the full grid case (14), the non-unique representation in an enlarged sparse grid generating system (45) results in a non-trivial kernel of $\hat{\mathbf{A}}^{\mathrm{SG}}_{d,J}$. Thus, the matrix

$\hat{\mathbf{A}}^{\text{SG}}_{d,J}$ is not invertible. But, again, the system (46) is solvable since the right-hand side $\hat{\mathbf{b}}^{\text{SG}}_{d,J}$ lies in the range of the system matrix and a solution can be generated by any semi-convergent iterative method.

We will now describe the enlarged sparse grid system (46) as a submatrix and a subvector of the enlarged full grid system (14). Note that this is done here for theoretical purposes only. In our implementation we of course avoid the full grid system with $\hat{N}_{d,J}$ degrees of freedom. In fact, our computational costs stay proportional to $\hat{N}^{\text{SG}}_{d,J}$, which is substantially smaller, cf. Sect. 4.3.

Like in (17), we can express the blocks of $\hat{\mathbf{A}}^{\text{SG}}_{d,J}$ and $\hat{\mathbf{b}}^{\text{SG}}_{d,J}$ with respect to (7) by

$$(\hat{\mathbf{A}}^{\text{SG}}_{d,J})_{\mathbf{l},\mathbf{k}} = \mathbf{I}^{\mathbf{l}}_{\mathbf{J}}\mathbf{A}_{d,J}\mathbf{I}^{\mathbf{J}}_{\mathbf{k}} \quad \text{and} \quad (\hat{\mathbf{b}}^{\text{SG}}_{d,J})_{\mathbf{l}} = \mathbf{I}^{\mathbf{l}}_{\mathbf{J}}\mathbf{b}_{d,J}$$

for $\mathbf{l}, \mathbf{k} \in \mathscr{S}^d_J$. Now, we can express our sparse grid operator matrix by

$$\hat{\mathbf{A}}^{\text{SG}}_{d,J} = \hat{\mathbf{R}}_{d,J}\hat{\mathbf{A}}_{d,J}\hat{\mathbf{R}}^T_{d,J} \ , \tag{47}$$

and our right-hand side by

$$\hat{\mathbf{b}}^{\text{SG}}_{d,J} = \hat{\mathbf{R}}_{d,J}\hat{\mathbf{b}}_{d,J} \ ,$$

where $\hat{\mathbf{R}}_{d,J} \in \mathbb{R}^{\hat{N}^{\text{SG}}_{d,J} \times \hat{N}_{d,J}}$ is a rectangular block-structured matrix with

$$(\hat{\mathbf{R}}_{d,J})_{\mathbf{l},\mathbf{k}} = \begin{cases} \mathbf{I}_{\mathbf{l}} & \text{for} \quad \mathbf{k} = \mathbf{l} \ , \\ 0 & \text{else} \ , \end{cases}$$

for $\mathbf{l} \in \mathscr{S}^d_J, \mathbf{k} \in \mathscr{F}^d_J$. Note that $\hat{\mathbf{R}}^T_{d,J}\hat{\mathbf{R}}_{d,J} \in \mathbb{R}^{\hat{N}_{d,J} \times \hat{N}_{d,J}}$ is a block-diagonal scaling matrix in the enlarged full grid system which simply sets all vector blocks to zero that belong to $\mathbf{l} \in \mathscr{F}^d_J \setminus \mathscr{S}^d_J$, and $\hat{\mathbf{R}}_{d,J}\hat{\mathbf{R}}^T_{d,J} \in \mathbb{R}^{\hat{N}^{\text{SG}}_{d,J} \times \hat{N}^{\text{SG}}_{d,J}}$ is simply the identity matrix on $\mathbb{R}^{\hat{N}^{\text{SG}}_{d,J}}$.

4.2 Sparse Grid Submatrix and Preconditioner

The proof of Theorem 2 showed that the condition number of $\mathbf{C}_{d,J}\mathbf{A}_{d,J}$ is independent of the dimension d and bounded independently of J. We will now extend this result to the sparse grid case by a submatrix argument. For reasons of simplicity, we stick here to the case of the regular sparse grid space $V_{\mathscr{S}^d_J}$ and the associated matrix $\hat{\mathbf{A}}^{\text{SG}}_{d,J}$, i.e. to the index set $\mathscr{S}^d_J$ of (43). But note that the following proof works with any index set $\mathscr{I} \subset \mathbb{N}^d$ with $\mathscr{I} \subset \mathscr{F}_{d,J}$ for which condition (44) is fulfilled.

Theorem 3. *The generalized condition number of the symmetric matrix*

$$\hat{\mathbf{R}}_{d,J}\hat{\mathbf{L}}^{-1}\hat{\mathbf{P}}^T\hat{\mathbf{D}}^{-1/2}\hat{\mathbf{R}}_{d,J}^T\hat{\mathbf{A}}_{d,J}^{SG}\hat{\mathbf{R}}_{d,J}\hat{\mathbf{D}}^{-1/2}\hat{\mathbf{P}}\hat{\mathbf{L}}^{-T}\hat{\mathbf{R}}_{d,J}^T \in \mathbb{R}^{\hat{N}_{d,J}^{SG}\times\hat{N}_{d,J}^{SG}} \tag{48}$$

is less than or equal to the condition number of the preconditioned system $\mathbf{C}_{d,J}\mathbf{A}_{d,J}$ *of the full grid with same dimension* d *and level* J. *Thus, the generalized condition number of* $\hat{\mathbf{C}}_{d,J}^{SG}\hat{\mathbf{A}}_{d,J}^{SG}$ *with*

$$\hat{\mathbf{C}}_{d,J}^{SG} := \hat{\mathbf{R}}_{d,J}\hat{\mathbf{P}}\hat{\mathbf{D}}^{-1}\hat{\mathbf{G}}^{-1}\hat{\mathbf{P}}^T\hat{\mathbf{R}}_{d,J}^T \tag{49}$$

is bounded asymptotically with respect to J *and* d.

Proof. We recall (40) from the proof of Theorem 2, i.e.

$$\hat{\mathbf{y}}_{d,J}^T\hat{\mathbf{L}}^{-1}\hat{\mathbf{D}}^{-1/2}\hat{\mathbf{P}}^T\hat{\mathbf{A}}_{d,J}\hat{\mathbf{P}}\hat{\mathbf{D}}^{-1/2}\hat{\mathbf{L}}^{-T}\hat{\mathbf{y}}_{d,J} \simeq \hat{\mathbf{y}}_{d,J}^T\hat{\mathbf{y}}_{d,J} \tag{50}$$

for $\hat{\mathbf{y}}_{d,J} \in \operatorname{im}\hat{\mathbf{L}}^T\hat{\mathbf{D}}^{1/2}\hat{\mathbf{P}}$. We obtain equivalence constants, which are at least as good as those in (40), by the stronger condition

$$\hat{\mathbf{y}}_{d,J} \in \operatorname{im}\hat{\mathbf{R}}_{d,J}^T\hat{\mathbf{R}}_{d,J}\hat{\mathbf{L}}^T\hat{\mathbf{D}}^{1/2}\hat{\mathbf{P}} \subset \operatorname{im}\hat{\mathbf{L}}^T\hat{\mathbf{D}}^{1/2}\hat{\mathbf{P}} \subset \mathbb{R}^{\hat{N}_{d,J}} \ .$$

The image of $\hat{\mathbf{R}}_{d,J}^T$ is not enlarged by block-diagonal matrices, and we can safely replace $\hat{\mathbf{A}}_{d,J}$ by $\hat{\mathbf{R}}_{d,J}^T\hat{\mathbf{A}}_{d,J}^{\mathrm{SG}}\hat{\mathbf{R}}_{d,J}$ on $\operatorname{im}\hat{\mathbf{R}}_{d,J}^T$. This gives us

$$\hat{\mathbf{y}}_{d,J}^T\hat{\mathbf{L}}^{-1}\hat{\mathbf{D}}^{-1/2}\hat{\mathbf{P}}^T\hat{\mathbf{R}}_{d,J}^T\hat{\mathbf{A}}_{d,J}^{\mathrm{SG}}\hat{\mathbf{R}}_{d,J}\hat{\mathbf{P}}\hat{\mathbf{D}}^{-1/2}\hat{\mathbf{L}}^{-T}\hat{\mathbf{y}}_{d,J} \simeq \hat{\mathbf{y}}_{d,J}^T\hat{\mathbf{y}}_{d,J}$$

with the same constants as in (50). Setting $\hat{\mathbf{z}}_{d,J}^{\mathrm{SG}} = \hat{\mathbf{R}}_{d,J}\hat{\mathbf{y}}_{d,J}$ results in

$$\hat{\mathbf{y}}_{d,J} = \hat{\mathbf{R}}_{d,J}^T\hat{\mathbf{R}}_{d,J}\hat{\mathbf{y}}_{d,J} = \hat{\mathbf{R}}_{d,J}^T\hat{\mathbf{z}}_{d,J}^{\mathrm{SG}} \ ,$$

and we obtain

$$(\hat{\mathbf{z}}_{d,J}^{\mathrm{SG}})^T\hat{\mathbf{R}}_{d,J}\hat{\mathbf{L}}^{-1}\hat{\mathbf{D}}^{-1/2}\hat{\mathbf{P}}^T\hat{\mathbf{R}}_{d,J}^T\hat{\mathbf{A}}_{d,J}^{\mathrm{SG}}\hat{\mathbf{R}}_{d,J}\hat{\mathbf{P}}\hat{\mathbf{D}}^{-1/2}\hat{\mathbf{L}}^{-T}\hat{\mathbf{R}}_{d,J}^T\hat{\mathbf{z}}_{d,J}^{\mathrm{SG}} \simeq (\hat{\mathbf{z}}_{d,J}^{\mathrm{SG}})^T\hat{\mathbf{z}}_{d,J}^{\mathrm{SG}}$$

on

$$\hat{\mathbf{z}}_{d,J}^{\mathrm{SG}} \in \operatorname{im}\hat{\mathbf{R}}_{d,J}\hat{\mathbf{R}}_{d,J}^T\hat{\mathbf{R}}_{d,J}\hat{\mathbf{L}}^T\hat{\mathbf{D}}^{1/2}\hat{\mathbf{P}} = \operatorname{im}\hat{\mathbf{R}}_{d,J}\hat{\mathbf{L}}^T\hat{\mathbf{D}}^{1/2}\hat{\mathbf{P}} \subset \mathbb{R}^{\hat{N}_{d,J}^{\mathrm{SG}}} \ .$$

It is left to show that vectors $\hat{\mathbf{v}}_{d,J}^{\mathrm{SG}}$ with $\hat{\mathbf{v}}_{d,J}^{\mathrm{SG}} \perp \hat{\mathbf{z}}_{d,J}^{\mathrm{SG}}$ are indeed in the kernel of (48). We obtain this by

$$\begin{aligned}\hat{\mathbf{v}}_{d,J}^{\mathrm{SG}} \in \ker(\hat{\mathbf{R}}_{d,J}\hat{\mathbf{L}}^T\hat{\mathbf{D}}^{1/2}\hat{\mathbf{P}})^T &= \ker\hat{\mathbf{P}}^T\hat{\mathbf{D}}^{1/2}\hat{\mathbf{L}}\hat{\mathbf{R}}_{d,J}^T = \ker\hat{\mathbf{D}}^{1/2}\hat{\mathbf{G}}\hat{\mathbf{P}}\hat{\mathbf{G}}^{-1}\hat{\mathbf{L}}\hat{\mathbf{R}}_{d,J}^T \\ &= \ker\hat{\mathbf{D}}^{1/2}\hat{\mathbf{G}}\hat{\mathbf{P}}\hat{\mathbf{L}}^{-T}\hat{\mathbf{R}}_{d,J}^T = \ker\hat{\mathbf{P}}\hat{\mathbf{L}}^{-T}\hat{\mathbf{R}}_{d,J}^T \ ,\end{aligned}$$

where we have used similar arguments as in the proof of Theorem 2. Altogether, this proves that the matrix (48) has a generalized condition number that is at least as good as that for the full grid case, i.e. that of $\mathbf{C}_{d,J}\mathbf{A}_{d,J}$. Finally, we can rewrite the preconditioner in the form (49) since

$$\begin{aligned}
&\tilde{\kappa}(\hat{\mathbf{R}}_{d,J}\hat{\mathbf{L}}^{-1}\hat{\mathbf{P}}^T\hat{\mathbf{D}}^{-1/2}\hat{\mathbf{R}}_{d,J}^T\hat{\mathbf{A}}_{d,J}^{\mathrm{SG}}\hat{\mathbf{R}}_{d,J}\hat{\mathbf{D}}^{-1/2}\hat{\mathbf{P}}\hat{\mathbf{L}}^{-T}\hat{\mathbf{R}}_{d,J}^T)\\
&= \tilde{\kappa}(\hat{\mathbf{R}}_{d,J}\hat{\mathbf{D}}^{-1/2}\hat{\mathbf{P}}\hat{\mathbf{L}}^{-T}\hat{\mathbf{R}}_{d,J}^T\hat{\mathbf{R}}_{d,J}\hat{\mathbf{L}}^{-1}\hat{\mathbf{P}}^T\hat{\mathbf{D}}^{-1/2}\hat{\mathbf{R}}_{d,J}^T\hat{\mathbf{A}}_{d,J}^{\mathrm{SG}}) \qquad (51)\\
&= \tilde{\kappa}(\hat{\mathbf{R}}_{d,J}\hat{\mathbf{P}}\hat{\mathbf{D}}^{-1}\hat{\mathbf{G}}^{-1}\hat{\mathbf{P}}^T\hat{\mathbf{R}}_{d,J}^T\hat{\mathbf{A}}_{d,J}^{\mathrm{SG}}) = \tilde{\kappa}(\hat{\mathbf{C}}_{d,J}^{\mathrm{SG}}\hat{\mathbf{A}}_{d,J}^{\mathrm{SG}})\,. \qquad (52)
\end{aligned}$$

In (51), we used that $\tilde{\kappa}(\mathbf{EF}) = \tilde{\kappa}(\mathbf{FE})$ for arbitrary square matrices $\mathbf{E}$ and $\mathbf{F}$, and in (52) we used that $\hat{\mathbf{R}}_{d,J}^T\hat{\mathbf{R}}_{d,J}$ is the identity on im $\hat{\mathbf{R}}_{d,J}^T$, that $\hat{\mathbf{L}}^{-T}\hat{\mathbf{L}}^{-1} = \hat{\mathbf{G}}^{-1}$ and, finally, that block-diagonal scaling matrices commute with block-diagonal matrices, see Lemma 4. This proves the theorem. □

4.3 Cost Discussion

It is of course important not to implement the matrix $\hat{\mathbf{A}}_{d,J}^{\mathrm{SG}}$ nor the preconditioner $\hat{\mathbf{C}}_{d,J}^{\mathrm{SG}}$ from (49) naively, if we want to keep their computational costs proportional to the number of degrees of freedom $\hat{N}_{d,J}^{\mathrm{SG}}$ of the sparse grid. First, let us consider $\hat{\mathbf{C}}_{d,J}^{\mathrm{SG}}$ in more detail. In fact, all the matrices $\hat{\mathbf{P}}$, $\hat{\mathbf{D}}^{-1}$, $\hat{\mathbf{G}}^{-1}$ and $\hat{\mathbf{P}}^T$ are block-diagonal, which means that they can be implemented as subspace-wise operations. As for $\hat{\mathbf{x}}_{d,J} \in$ im $\hat{\mathbf{R}}_{d,J}^T$ all vector blocks $(\hat{\mathbf{x}}_{d,J})_{\mathbf{l}}$ with $\mathbf{l} \in \mathscr{F}_J^d \setminus \mathscr{S}_J^d$ are zero and get removed by the final application of $\hat{\mathbf{R}}_{d,J}$ anyway, they do not need to be considered in the implementation. By the same arguments as in Sect. 3.3, we obtain that our preconditioner can indeed be applied in $\mathscr{O}(d \cdot \hat{N}_{d,J}^{\mathrm{SG}})$.

An efficient matrix-vector multiplication with the operator matrix $\hat{\mathbf{A}}_{d,J}^{\mathrm{SG}}$ is however far more complicated than in the full grid case. One reason is that the index $\mathscr{S}_J^d$ has, unlike $\mathscr{F}_J^d$, no representation as a Cartesian product, which means that $\hat{\mathbf{A}}_{d,J}^{\mathrm{SG}}$ has no Kronecker product structure like $\hat{\mathbf{A}}_{d,J}$ does. Of course, we must not use (47), which was given only for theoretical reasons to allow for the submatrix argument of the last subsection. Instead, we resort to a quite sophisticated dimension-recursive algorithm based on the so-called unidirectional principle [4,13] to perform the matrix-vector-multiplication linearly in the number of degrees of freedom $\hat{N}_{d,J}^{\mathrm{SG}}$. Typically the associated dimension-dependent constant in the costs is proportional to 2^d. This factor can however be reduced to d^2 in the case of the Laplacian by exploiting the L^2-orthogonality between subspaces, see [21]. Then, the total algorithmic cost of one application of $\hat{\mathbf{C}}_{d,J}^{\mathrm{SG}}\hat{\mathbf{A}}_{d,J}^{\mathrm{SG}}$ is of the order $\mathscr{O}(d^2 \cdot \hat{N}_{d,J}^{\mathrm{SG}})$.

So far, we have expressed the computational effort with respect to the enlarged sparse grid system with $\hat{N}_{d,J}^{\mathrm{SG}}$ degrees of freedom, which has by a factor of about 2^d more degrees of freedom than the hierarchical basis [15]. We consider this acceptable, because the number of degrees of freedom of regular sparse grids is of the order $\mathscr{O}(2^J J^{d-1})$, which is exponential in d anyway. Moreover, the number of degrees of freedom of energy sparse grids is of the order $\mathscr{O}(2^J)$ in J, but which involves a constant that is also exponential in d, cf. [29].

Note that it is possible to remove the redundancy of our multilevel discretization via the generating system (45) by using a prewavelet discretization. This seems to eliminate the 2^d-factor by construction. It however still appears hidden in the setup of the discrete right-hand side for general functions f. The prewavelet approach and a new improved preconditioner will be discussed in the next section.

5 Prewavelets

The enlarged generating system introduced some additional difficulties like a non-trivial kernel of the operator matrix and the need for an orthogonalization operator $\hat{\mathbf{P}}$. This can be avoided in the first place if a direct discretization of the orthogonal subspaces $W_{\mathbf{l}}$ is available, which is just the case for so-called prewavelets and for wavelets.

Let us first consider the full grid case. To this end, let us assume that we have basis functions $(\psi_{\mathbf{l},\mathbf{i}})_{\mathbf{i}\in\xi_{\mathbf{l}},\mathbf{l}\in\mathscr{F}_J^d}$ with

$$W_{\mathbf{l}} = \mathrm{span}\{\psi_{\mathbf{l},\mathbf{i}} : \mathbf{i} \in \xi_{\mathbf{l}}\} \quad \text{for} \quad \mathbf{l} \in \mathscr{F}_J^d \ , \tag{53}$$

and set $\bar{n}_{\mathbf{l}} := |\xi_{\mathbf{l}}|$. Note here that we have L^2-orthogonality between different levels by definition, but we have not necessarily L^2-orthogonality within one level. The multilevel matrix $\bar{\mathbf{A}}_{d,J} \in \mathbb{R}^{N_{d,J}\times N_{d,J}}$ with

$$(\bar{\mathbf{A}}_{d,J})_{(\mathbf{l},\mathbf{i}),(\mathbf{k},\mathbf{j})} = a(\psi_{\mathbf{l},\mathbf{i}}, \psi_{\mathbf{k},\mathbf{j}}) \quad \text{for} \quad \mathbf{i} \in \xi_{\mathbf{l}}, \mathbf{j} \in \xi_{\mathbf{k}}, \mathbf{l}, \mathbf{k} \in \mathscr{F}_J^d \tag{54}$$

results just from the system matrix $\mathbf{A}_{d,J}$ of (7) by a change of the basis, since

$$\bigoplus_{\mathbf{l}\in\mathscr{F}_J^d} W_{\mathbf{l}} = V_J^d \ .$$

Thus,

$$\bar{\mathbf{A}}_{d,J} = \mathbf{T}^T \mathbf{A}_{d,J} \mathbf{T} \ ,$$

where $\mathbf{T}$ maps from $\{\psi_{\mathbf{l},\mathbf{i}}\}_{\mathbf{i}\in\xi_{\mathbf{l}},\mathbf{l}\in\mathscr{F}_J^d}$ to $\{\phi_{\mathbf{J},\mathbf{i}}\}_{\mathbf{i}\in\chi_{\mathbf{J}}}$. The analogue holds for the right-hand side $\bar{\mathbf{b}}_{d,J} \in \mathbb{R}^{N_{d,J}}$ with

$$(\bar{\mathbf{b}}_{d,J})_{\mathbf{l},\mathbf{i}} = (f, \psi_{\mathbf{l},\mathbf{i}})_{L^2(\Omega^d)}, \mathbf{i} \in \xi_{\mathbf{l}}, \mathbf{l} \in \mathscr{F}_J^d \ , \quad \text{i.e.} \quad \bar{\mathbf{b}}_{d,J} = \mathbf{T}^T \mathbf{b}_{d,J} \ .$$

Note here that, in contrast to $\hat{\mathbf{A}}_{d,J}$, the system matrix $\bar{\mathbf{A}}_{d,J}$ is now invertible, since the functions in (53) form a basis of $V_{\mathscr{F}_J^d}$.

5.1 Preconditioner

Now, we will present our preconditioner for prewavelets and discuss its resulting condition number. To this end, we need a diagonal scaling matrix $\bar{\mathbf{D}} \in \mathbb{R}^{N_{d,J}\times N_{d,J}}$ with blocks $(\bar{\mathbf{D}})_{\mathbf{l},\mathbf{k}} \in \mathbb{R}^{\bar{n}_{\mathbf{l}}\times\bar{n}_{\mathbf{k}}}$ and

$$(\bar{\mathbf{D}})_{\mathbf{l},\mathbf{k}} = \begin{cases} (\sum_{p=1}^{d} 2^{2l_p})\bar{\mathbf{I}}_{\mathbf{l}} & \text{for} \quad \mathbf{l} = \mathbf{k} \ , \\ 0 & \text{else} \ , \end{cases} \tag{55}$$

where the $\bar{\mathbf{I}}_{\mathbf{l}} \in \mathbb{R}^{\bar{n}_{\mathbf{l}}\times\bar{n}_{\mathbf{l}}}$ denote identity matrices on the subspaces. Furthermore, we need the subspace-wise mass matrix $\bar{\mathbf{G}} \in \mathbb{R}^{N_{d,J}\times N_{d,J}}$ with blocks $(\bar{\mathbf{G}})_{\mathbf{l},\mathbf{k}} \in \mathbb{R}^{\bar{n}_{\mathbf{l}}\times\bar{n}_{\mathbf{k}}}$, where

$$(\bar{\mathbf{G}})_{\mathbf{l},\mathbf{k}} = \begin{cases} \bar{\mathbf{M}}_{\mathbf{l}} & \text{for} \quad \mathbf{l} = \mathbf{k} \ , \\ 0 & \text{else} \ . \end{cases}$$

Here, $\bar{\mathbf{M}}_{\mathbf{l}} \in \mathbb{R}^{\bar{n}_{\mathbf{l}}\times\bar{n}_{\mathbf{l}}}$ denotes the mass matrix

$$(\bar{\mathbf{M}}_{\mathbf{l}})_{\mathbf{i},\mathbf{j}} = (\psi_{\mathbf{l},\mathbf{i}}, \psi_{\mathbf{l},\mathbf{j}}) \quad \text{for} \quad \mathbf{i},\mathbf{j} \in \xi_{\mathbf{l}} \ .$$

Then, we have the following theorem:

Theorem 4. *The condition number of*

$$\bar{\mathbf{D}}^{-1}\bar{\mathbf{G}}^{-1}\bar{\mathbf{A}}_{d,J} \tag{56}$$

is bounded asymptotically with respect to J and is completely independent of the dimension d.

Proof. We translate the norm equivalence (22) into the matrix-vector setting for $\bar{\mathbf{x}}_{d,J} \in \mathbb{R}^{N_{d,J}\times N_{d,J}}$ and obtain

$$\begin{aligned}
\bar{\mathbf{x}}_{d,J}^T \bar{\mathbf{A}}_{d,J} \bar{\mathbf{x}}_{d,J} &= a\Big(\sum_{\mathbf{l}\in\mathscr{F}_J^d}\sum_{\mathbf{i}\in\xi_{\mathbf{l}}} \bar{x}_{\mathbf{l},\mathbf{i}}\psi_{\mathbf{l},\mathbf{i}}, \sum_{\mathbf{l}\in\mathscr{F}_J^d}\sum_{\mathbf{i}\in\xi_{\mathbf{l}}} \bar{x}_{\mathbf{l},\mathbf{i}}\psi_{\mathbf{l},\mathbf{i}}\Big) \\
&\simeq \sum_{\mathbf{l}\in\mathscr{F}_J^d}\Big(\sum_{p=1}^{d} 2^{2l_p}\Big)\Big\|\sum_{\mathbf{i}\in\xi_{\mathbf{l}}}\bar{x}_{\mathbf{l},\mathbf{i}}\psi_{\mathbf{l},\mathbf{i}}\Big\|_{L^2(\Omega^d)}^2 \\
&= \sum_{\mathbf{l}\in\mathscr{F}_J^d}\Big(\sum_{p=1}^{d} 2^{2l_p}\Big)\bar{\mathbf{x}}_{\mathbf{l}}^T\bar{\mathbf{M}}_{\mathbf{l}}\bar{\mathbf{x}}_{\mathbf{l}} \\
&= \bar{\mathbf{x}}_{d,J}^T\bar{\mathbf{D}}\bar{\mathbf{G}}\bar{\mathbf{x}}_{d,J}\ .
\end{aligned} \tag{57}$$

From Theorem 1 we know that the constants in (57) are independent of the dimension d and bounded independently of J. This concludes the proof. □

In the case of a regular sparse grid with $\mathscr{I} = \mathscr{S}_J^d$, the equality

$$\bigoplus_{\mathbf{l}\in\mathscr{S}_J^d} W_{\mathbf{l}} = \sum_{\mathbf{l}\in\mathscr{S}_J^d} V_{\mathbf{l}} \tag{58}$$

holds. The analogue is valid for a general sparse grid with any arbitrary index set $\mathscr{I}$ for which the condition (44) is satisfied. Thus, the regular sparse grid space $\sum_{\mathbf{l}\in\mathscr{S}_J^d} V_{\mathbf{l}}$ or the general sparse grid space $\sum_{\mathbf{l}\in\mathscr{I}} V_{\mathbf{l}}$ can both be expressed by the left-hand side of (58), i.e. with the help of $W_{\mathbf{l}}$-subspaces. The resulting linear system matrix $\bar{\mathbf{A}}_{d,J}^{\mathrm{SG}}$ is just a submatrix of the full matrix[6] $\bar{\mathbf{A}}_{d,J}$. Consequently, the condition number for the sparse grid system is at least as good as the one of (56). Analogously the resulting right-hand side $\bar{\mathbf{b}}_{d,J}^{\mathrm{SG}}$ is just a subvector of $\bar{\mathbf{b}}_{d,J}$.

Note here that prewavelets have been frequently used in the past as the basis functions of sparse grid discretizations [22, 33] but mostly no special attention was paid to the dependence of the condition number on the dimension. A simple Jacobi-diagonal scaling of $\bar{\mathbf{A}}_{d,J}$ is equivalent to replacing the subspace-wise inversion of the mass matrices $\bar{\mathbf{G}}^{-1}$ in (56) by the identity and the $\bar{\mathbf{D}}$ from (55) by $\mathrm{diag}(\bar{\mathbf{A}}_{d,J})$. This does not affect the asymptotics in J for L^2-stable basis functions, but the condition number of the operator matrix grows now exponentially with the dimension [21]. Sometimes $(\bar{\mathbf{D}})_{\mathbf{l},\mathbf{l}} = 2^{2|\mathbf{l}|_\infty}\,\mathrm{diag}(\bar{\mathbf{M}}_{\mathbf{l}})$ is chosen, see [31, 32], which also results in condition numbers that grow with the dimension d.

[6] Here, the level J of the full grid is to be equal to the level J of the regular sparse grid or equal to the finest level involved in $\mathscr{I}$ for the general sparse grid.

5.2 Cost Discussion

We now have the preconditioner $\bar{\mathbf{C}}_{d,J} := \bar{\mathbf{D}}^{-1}\bar{\mathbf{G}}^{-1}$ for $\bar{\mathbf{A}}_{d,J}$ in prewavelet discretization. At first sight, this approach looks simpler and more efficient than the more complicated discretizations $\hat{\mathbf{A}}_{d,J}$ and $\hat{\mathbf{A}}^{\mathrm{SG}}_{d,J}$ using the enlarged generating system and the associated preconditioners $\hat{\mathbf{C}}_{d,J}$ and $\hat{\mathbf{C}}^{\mathrm{SG}}_{d,J}$, respectively. This is due to the fact that the prewavelet system $\{\psi_{\mathbf{l},\mathbf{i}} : \mathbf{i} \in \xi_{\mathbf{l}}\}_{\mathbf{l}\in\mathscr{I}}$ forms a basis and therefore exhibits no redundancies. Thus, by a factor of about 2^d less degrees of freedom are involved than for the corresponding generating system.

However, there are additional difficulties to be faced in the prewavelet approach, which should not be underestimated and may give the generating system method a practical advantage. First, prewavelets are less local than, e.g. the corresponding multilevel spline basis. Thus, the mass matrix inversions in $\bar{\mathbf{G}}^{-1}$ become more involved. From a programming perspective, the more complicated basis functions and different types of prewavelet functions near the boundary make the application of the matrix $\bar{\mathbf{A}}_{d,J}$ to a vector more difficult. The efficient application of the action of the sparse grid system matrix $\bar{\mathbf{A}}^{\mathrm{SG}}_{d,J}$ onto a vector is even more involved, since the unidirectional principle strongly relies on the nestedness of the subspaces. If this is no longer the case, the one-dimensional operators have to be tailored to the specific discretization [22] or the algorithm must switch to a generating system anyway [59].

Finally, as already mentioned at the end of Sect. 4.3, the cost complexity of the setup of the right-hand side $\bar{\mathbf{b}}^{\mathrm{SG}}_{d,J}$ also has at least a 2^d-factor if the corresponding integrations are realized by the interpolation of the function f from (1) in our prewavelet sparse grid space and a subsequent multiplication by the mass matrix to account for the necessary numerical quadrature. As stated in [21], for general functions f, this approach requires the inclusion of boundary functions in the interpolation step (even if the solution u of our Poisson problem has homogeneous boundary conditions). Since the d-dimensional hypercube Ω^d has 2^d faces, an additional factor of the order 2^d enters the cost complexity for the setup of the right-hand side. The dependence of the cost complexity on the dimension d of other techniques for the assembly of the right-hand side for wavelets and prewavelets with sufficient accuracy, e.g. by the solution of an eigenvector-moment problem associated with the coefficients of the refinement equation [20], is unknown to us. We however believe that also these methods involve a factor of at least 2^d in the d-dimensional case due to the tensor product construction.

Altogether, the generating system approach from (36) and (48) can be seen as a simple form of implementation of the prewavelet approach and, indeed, both methods give exactly the same condition numbers for the piecewise linear case.

6 Numerical Experiments

Now, we give the results of numerical experiments for our new full and sparse grid preconditioners (42) and (49), respectively. We consider the d-dimensional Laplace operator on the domain $\Omega^d = (0,1)^d$ with vanishing Dirichlet boundary conditions.

Table 1 Degrees of freedom $\hat{N}_{d,J}$ and $\hat{N}_{d,J}^{\mathrm{SG}}$ and condition numbers $\tilde{\kappa}$ of the preconditioned matrices (36) and (48) for the Laplacian on the unit hypercube with a full- and sparse-grid discretization for the generating system approach based on linear splines

		DOFs $\hat{N}_{d,J}$ and $\hat{N}_{d,J}^{\mathrm{SG}}$		Condition number $\tilde{\kappa}$	
	Level J	Full grid	Sparse grid	Full grid	Sparse grid
Dim = 1	2	4	4	3.40	3.40
	3	11	11	4.67	4.67
	4	26	26	5.17	5.17
	5	57	57	5.84	5.84
	6	120	120	6.37	6.37
	7	247	247	6.80	6.80
	8	502	502	7.16	7.16
	9	1,013	1,013	7.47	7.47
	10	2,036	2,036	7.74	7.74
	11	4,083	4,083	7.96	7.96
	12	8,178	8,178	8.16	8.16
	13	16,369	16,369	8.33	8.33
Dim = 2	2	16	7	3.40	2.99
	3	121	30	4.67	4.46
	4	676	102	5.17	5.06
	5	3,249	303	5.84	5.65
	6	14,400	825	6.37	6.20
Dim = 3	2	64	10	3.40	2.71
	3	1,331	58	4.67	4.28
	4	17,576	256	5.17	5.00
Dim = 4	2	256	13	3.40	2.51
	3	14,641	95	4.67	4.12
Dim = 5	2	1,024	16	3.40	2.36

As locally supported basis functions in (4), we choose on level l the $n_l = 2^l - 1$ hat functions

$$\phi_{l,i}(x) = \max(1 - 2^l \left| x - x_{l,i} \right|, 0) ,$$

which are centered at the points of an equidistant mesh

$$x_{l,i} = 2^{-l} i$$

for $i = 1, \ldots, n_l$. The resulting space $\cup_{l=1}^{\infty} V_l$ is indeed equal to the underlying Sobolev space $H_0^1(\Omega)$ up to completion with the H^1-norm, see [15].

Table 1 shows the generalized condition numbers of the preconditioned matrices (36) and (48) of the enlarged generating systems in the full and sparse grid case for different dimensions d and levels J. We clearly observe that the full grid condition numbers are bounded from above by a constant independently of the level J. Moreover, they are perfectly independent of the dimension as our theory suggests. The sparse grid condition numbers are even smaller than the

Table 2 Degrees of freedom and generalized condition numbers of the preconditioned sparse grid system (48) for different dimensions d and levels J

Dim	2	3	4	5	6	7	8	9	10	11	12	13
	Degrees of freedom with respect to the level J											
1	4	11	26	57	120	247	502	1,013	2,036	4,083	8,178	16,369
2	7	30	102	303	825	2,116	5,200	12,381				
3	10	58	256	955	3,178	9,740						
4	13	95	515	2,310	9,078							
5	16	141	906	4,746								
6	19	196	1,456	8,722								
7	22	260	2,192	14,778								
8	25	333	3,141									
9	28	415	4,330									
10	31	506	5,786									
	Condition number with respect to the level J											
1	3.40	4.67	5.17	5.84	6.37	6.80	7.16	7.47	7.74	7.96	8.16	8.33
2	2.99	4.46	5.06	5.65	6.20	6.65	7.04	7.36				
3	2.71	4.28	5.00	5.49	6.06	6.53						
4	2.51	4.12	4.94	5.35	5.95							
5	2.36	3.97	4.88	5.23								
6	2.24	3.83	4.82	5.17								
7	2.15	3.71	4.77	5.15								
8	2.07	3.60	4.71									
9	2.00	3.50	4.66									
10	1.94	3.41	4.61									

corresponding full grid ones for $d > 1$, which is in accordance with our submatrix argument from Theorem 3. In Table 2 (top) we give the number of degrees of freedom $\hat{N}^{\mathrm{SG}}_{d,J}$ for various values of J and d.

Finally, Table 2 (bottom) reveals that the condition numbers of the sparse grid even *decrease* with rising dimension d for a fixed level J. For a sparse grid discretization of the Poisson problem, these numbers clearly show that we are altogether able to efficiently solve the associated linear systems of equations for both, quite large values of J and larger dimensions d.

Finally note that the prewavelet approach from Sect. 5 results in exactly the same condition numbers.

7 Concluding Remarks

We presented preconditioners $\mathbf{C}_{d,J}$ (42) and $\hat{\mathbf{C}}^{\mathrm{SG}}_{d,J}$ (49) for an isotropic full grid and an enlarged sparse grid system, respectively. Both result in condition numbers that are bounded independently of the level J and are constant or, in the sparse grid case, even decreasing for rising dimension d.

The computational costs of the preconditioner remained linear in the number of degrees of freedom. The size of the enlarged systems grows by a factor of about 2^d compared to the corresponding basis. This seems a fair price to pay. Better constants in the respective norm equivalence and associated condition numbers reduce the number of iterations of Krylov methods and also make corresponding residual-based error estimates more reliable.

Our new preconditioners can be applied to differential operators other than the Laplacian. The approach works straightforwardly for constant coefficients or variable coefficients which are separable, i.e. can be written as a product of one-dimensional diffusion functions, or can be well approximated by a low-rank representation. But then the equivalence constants of

$$a(u,u) \simeq \|u\|^2_{H^1(\Omega^d)} \ ,$$

i.e. the ellipticity constants, cf. [16], at least partly enter the condition estimate of the resulting system. If they grow exponentially with the dimension d we may run into problems, though.

In a similar way, equivalences to H^s-norms can be dealt with by using diagonal scaling matrices

$$(\hat{\mathbf{D}})_{\mathbf{l},\mathbf{k}} = \begin{cases} (\sum_{p=1}^{d} 2^{2sl_p})\mathbf{I}_\mathbf{l} & \text{for} \quad \mathbf{l} = \mathbf{k} \ , \\ 0 & \text{else} \ , \end{cases}$$

and the associated $\hat{\mathbf{G}}^{-1}$ if the regularity of the employed basis functions is sufficient.

References

1. Allaire, G.: Homogenization and two-scale convergence. SIAM J. Math. Anal. **23**(6), 1482–1518 (1992)
2. Babuška, I., Nobile, F., Tempone, R.: A stochastic collocation method for elliptic partial differential equations with random input data. SIAM Rev. **52**(2), 317–355 (2010)
3. Bäck, J., Nobile, F., Tamellini, L., Tempone, R.: Stochastic spectral Galerkin and collocation methods for PDEs with random coefficients: a numerical comparison. In: Hesthaven, J., Rönquist, E. (eds.) Spectral and High Order Methods for Partial Differential Equations. Volume 76 of Lecture Notes in Computational Science and Engineering, pp. 43–62. Springer Berlin/Heidelberg (2011)
4. Balder, R., Zenger, C.: The solution of multidimensional real Helmholtz equations on sparse grids. SIAM J. Sci. Comput. **17**, 631–646 (1996)
5. Balescu, R.: Statistical Dynamics: Matter Out of Equilibrium. Imperial College Press, London (1997)
6. Bellman, R.: Adaptive Control Processes: A Guided Tour. Princeton University Press, Princeton (1961)

7. Berman, A., Plemmons, R.: Nonnegative Matrices in the Mathematical Sciences. Society for Industrial and Applied Mathematics, Philadelphia (1994)
8. Bokanowski, O., Garcke, J., Griebel, M., Klompmaker, I.: An adaptive sparse grid semi-Lagrangian scheme for first order Hamilton-Jacobi Bellman equations. J. Sci. Comput. **55**, 575–605 (2012)
9. Braess, D.: Finite Elements: Theory, Fast Solvers, and Applications in Solid Mechanics. Cambridge University Press, Cambridge (2007)
10. Bramble, J., Pasciak, J., Xu, J.: Parallel multilevel preconditioners. Math. Comput. **55**(191), 1–22 (1990)
11. Brandt, A., Livne, O.: Multigrid techniques: 1984 guide with applications to fluid dynamics. In: Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphia (2011)
12. Bungartz, H.: An adaptive Poisson solver using hierarchical bases and sparse grids in iterative methods in linear algebra. In: de Groen, P., Beauwens, R. (eds.) Proceedings of the IMACS International Symposium, 2.-4. 4. 1991, pp. 293–310, Brussels. Elsevier, Amsterdam (1992)
13. Bungartz, H.: Dünne Gitter und deren Anwendung bei der adaptiven Lösung der dreidimensionalen Poisson-Gleichung. Dissertation, Fakultät für Informatik, Technische Universität München (1992)
14. Bungartz, H., Griebel, M.: A note on the complexity of solving Poisson's equation for spaces of bounded mixed derivatives. J. Complex. **15**, 167–199 (1999)
15. Bungartz, H., Griebel, M.: Sparse grids. Acta Numer. **13**, 1–123 (2004)
16. Chegini, N., Stevenson, R.: The adaptive tensor product wavelet scheme: sparse matrices and the application to singularly perturbed problems. IMA J. Numer. Anal. **32**(1), 75–104 (2012)
17. Cioranescu, D., Damlamian, A., Griso, G.: Periodic unfolding and homogenization. Comptes Rendus Mathematique **335**(1), 99–104 (2002)
18. Cohen, A., DeVore, R., Schwab, C.: Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDEs. Anal. Appl. **9**, 11–47 (2011)
19. Dahmen, W.: Stability of multiscale transformations. J. Fourier Anal. Appl. **2**, 341–361 (1996)
20. Dahmen, W., Micchelli, C.: Using the refinement equation for evaluating integrals of wavelets. SIAM J. Numer. Anal. **30**(2), 507–537 (1993)
21. Feuersänger, C.: Dünngitterverfahren für hochdimensionale elliptische partielle Differentialgleichungen. Diplomarbeit, Institut für Numerische Simulation, Universität Bonn (2005)
22. Feuersänger, C.: Sparse grid methods for higher dimensional approximation. Dissertation, Institut für Numerische Simulation, Universität Bonn (2010)
23. Garcke, J.: Maschinelles Lernen durch Funktionsrekonstruktion mit verallgemeinerten dünnen Gittern. Doktorarbeit, Institut für Numerische Simulation, Universität Bonn (2004)
24. Garcke, J., Griebel, M., Thess, M.: Data mining with sparse grids. Computing **67**(3), 225–253 (2001)
25. Gerstner, T., Griebel, M.: Dimension-adaptive tensor-product quadrature. Computing **71**(1), 65–87 (2003)
26. Girosi, F., Jones, M., Poggio, T.: Regularization theory and neural networks architectures. Neural Comput. **7**, 219–269 (1995)
27. Griebel, M.: Multilevel algorithms considered as iterative methods on semidefinite systems. SIAM Int. J. Sci. Stat. Comput. **15**(3), 547–565 (1994)
28. Griebel, M.: Multilevelmethoden als Iterationsverfahren über Erzeugendensystemen. Teubner Skripten zur Numerik. Teubner, Stuttgart (1994)
29. Griebel, M.: Sparse grids and related approximation schemes for higher dimensional problems. In: Pardo, L., Pinkus, A., Suli, E., Todd, M.J. (eds.) Foundations of Computational Mathematics (FoCM05), Santander, pp. 106–161. Cambridge University Press, Cambridge (2006)
30. Griebel, M., Knapek, S.: Optimized tensor-product approximation spaces. Constr. Approx. **16**(4), 525–540 (2000)
31. Griebel, M., Knapek, S.: Optimized general sparse grid approximation spaces for operator equations. Math. Comput. **78**(268), 2223–2257 (2009)

32. Griebel, M., Oswald, P.: On additive Schwarz preconditioners for sparse grid discretizations. Numer. Math. **66**, 449–464 (1994)
33. Griebel, M., Oswald, P.: Tensor product type subspace splitting and multilevel iterative methods for anisotropic problems. Adv. Comput. Math. **4**, 171–206 (1995)
34. Hackbusch, W.: Multi-grid Methods and Applications. Springer Series in Computational Mathematics. Springer, Berlin/New York (1985)
35. Hamaekers, J.: Tensor Product multiscale many–particle spaces with finite–order weights for the electronic schödinger equation. Dissertation, Institut für Numerische Simulation, Universität Bonn (2009)
36. Harbrecht, H., Schneider, R., Schwab, C.: Sparse second moment analysis for elliptic problems in stochastic domains. Numer. Math. **109**, 385–414 (2008)
37. Hegland, M.: Adaptive sparse grids. In: Burrage, K., Roger, B., Sidje, (eds.) Proceedings of 10th Computational Techniques and Applications Conference CTAC-2001, Brisbane, vol. 44, pp. C335–C353. (2003)
38. Hoang, V., Schwab, C.: High-dimensional finite elements for elliptic problems with multiple scales. Multiscale Model. Simul. **3**(1), 168–194 (2005)
39. Jakeman, J., Roberts, S.: Stochastic Galerkin and collocation methods for quantifying uncertainty in differential equations: a review. ANZIAM J. **50**(C), C815–C830 (2008)
40. Knapek, S.: Approximation und Kompression mit Tensorprodukt-Multiskalenräumen. Doktorarbeit, Universität Bonn (2000)
41. Kwok, Y.: Mathematical Models of Financial Derivatives. Springer Finance. Springer, London (2008)
42. Maître, O., Knio, O.: Spectral Methods for Uncertainty Quantification: With Applications to Computational Fluid Dynamics. Scientific Computation. Springer, New York (2010)
43. Matache, A.: Sparse two-scale FEM for homogenization problems. J. Sci. Comput. **17**, 659–669 (2002)
44. Messiah, A.: Quantum Mechanics. North-Holland, Amsterdam (1965)
45. Mitzlaff, U.: Diffusionsapproximation von Warteschlangensystemen. Doktorarbeit, TU Clausthal (1997)
46. Munos, R.: A study of reinforcement learning in the continuous case by the means of viscosity solutions. Mach. Learn. **40**(3), 265–299 (2000)
47. Nobile, F., Tempone, R., Webster, C.: An anisotropic sparse grid stochastic collocation method for partial differential equations with random input data. SIAM J. Numer. Anal. **46**(5), 2411–2442 (2008)
48. Nobile, F., Tempone, R., Webster, C.: A sparse grid stochastic collocation method for partial differential equations with random input data. SIAM J. Numer. Anal. **46**(5), 2309–2345 (2008)
49. Oswald, P.: On discrete norm estimates related to multilevel preconditioners in the finite element method. In: Proceedings of the International Conference on Constructive Theory of Functions, Varna 1991, pp. 203–214. Bulgarian Academy of Sciences, Sofia (1992)
50. Oswald, P.: Multilevel Finite Element Approximation: Theory and Applications. Teubner Skripten zur Numerik, Teubner (1994)
51. Reisinger, C.: Numerische Methoden für hochdimensionale parabolische Gleichungen am Beispiel von Optionspreisaufgaben. Dissertation, Universität Heidelberg (2004)
52. Schölkopf, B., Smola, A.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT, Cambridge (2001)
53. Shen, X., Chen, H., Dai, J., Dai, W.: The finite element method for computing the stationary distribution of an SRBM in a hypercube with applications to finite buffer queueing networks. Queueing Syst. Theory Appl. **42**(1), 33–62 (2002)
54. Sjöberg, P.: Partial approximation of the master equation by the Fokker–Planck equation. In: Kagström, B., Elmroth, E., Dongarra, J., Waśniewski, J. (eds.) Applied Parallel Computing. State of the Art in Scientific Computing. Volume 4699 of Lecture Notes in Computer Science, pp. 637–646. Springer, Berlin/Heidelberg (2007)
55. Sjöberg, P., Lötstedt, P., Elf, J.: Fokker–Planck approximation of the master equation in molecular biology. Comput. Vis. Sci. **12**, 37–50 (2009)

56. Sutton, R., Barto, A.: Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning). MIT, Cambridge (1998)
57. Xu, J.: Iterative methods by space decomposition and subspace correction. SIAM Rev. **34**(4), 581–613 (1992)
58. Yserentant, H.: Old and new convergence proofs for multigrid methods. Acta Numer. **2**, 285–326 (1993)
59. Zeiser, A.: Fast matrix-vector multiplication in the sparse-grid Galerkin method. J. Sci. Comput. **47**(3), 328–346 (2011)

Simulation of Droplet Impact with Dynamic Contact Angle Boundary Conditions

Michael Griebel and Margrit Klitz

Abstract The numerical simulation of dynamic wetting processes is of interest for a vast variety of industrial processes, where practical experiments are costly and time-consuming. In these simulations, the dynamic contact angle is a key parameter, but the modeling of its behavior is poorly understood so far. In this article, we simulate droplet impact on a dry flat surface by using two different contact angle models. Both models show good qualitative and quantitative agreement with experimental results. For our numerical method, we solve the three-dimensional Navier-Stokes equations with finite differences on a staggered grid. The free surface is captured by a level-set method, and the contact angle determines the shape of the level-set function at the boundary. Additionally, we investigate the mass-conservation properties of two volume-correction methods, which are invaluable for the analysis of the droplet behavior.

1 Introduction

The numerical simulation of dynamic wetting processes is of critical importance for a number of industrial applications such as coating, lamination, lubrication or ink- and spray-painting. All these applications have in common that liquid comes into contact with a solid surface and that the phase boundary is in motion. Thereby, a moving contact-line is produced along the substrate, which is the line where the air is replaced by the liquid. The quality of the wetting highly controls the quality of the industrial end products and, therefore, needs to be optimized to reduce wetting defects and instabilities such as air entrainment or ribbing. Here, numerical

M. Griebel (✉) · M. Klitz
Institut für Numerische Simulation, Rheinische Friedrich-Wilhelms-Universität Bonn,
Wegelerstr. 6, D-53115 Bonn, Germany
e-mail: griebel@ins.uni-bonn.de; klitz@ins.uni-bonn.de

M. Griebel (ed.), *Singular Phenomena and Scaling in Mathematical Models*,
DOI 10.1007/978-3-319-00786-1_13,

simulation and optimization is a reliable and cheap alternative to the traditional time-consuming adjustment of machines and the expensive waste of raw materials.

Despite the industrial interest in the numerical simulation of dynamic wetting processes, the existing theoretical and numerical approaches so far often fail to correctly predict the results of practical experiments. This is due to two fundamental difficulties which constitute the so-called 'moving-contact line problem': First, the classical theory of continuum fluid mechanics (i.e. the Navier-Stokes equations with the no-slip condition for the velocity) predicts a shear stress singularity at the moving contact line. The second difficulty is the modeling of the contact angle which is usually required as a boundary condition and determines the shape of the free surface at the contact line.

Numerous mathematical models have been developed to remedy the moving contact-line problem. Most of them remove the stress-singularity, but are unable to describe the contact angle and flow behavior as observed in practical experiments; see [13] and the references therein. One of the few models, which considers the overall physical context of the moving contact line problem, is Shikhmurzaev's interface formation model [13]. This model not only removes the stress-singularity, but is also able to describe a large variety of flow with singularities such as breakup, malescence or casp formation.

In this article we couple a reduced version of Shikhmurzaev's interface formation model with our three-dimensional incompressible two-phase Navier-Stokes solver. For this solver, we employ a standard discretization on uniform Cartesian staggered grids and use Chorin's projection approach. The free surface between the two fluid phases is tracked with the level-set approach. Here, the interface conditions are implicitly incorporated into the momentum equations by the continuum surface force (CSF) [2] method. Surface tension is evaluated using a smoothed delta function and third order interpolation. The parallelization of the code is based on conventional domain decomposition techniques using MPI. This allows us to deal with reasonably fine mesh resolutions in three dimensions.

For the simulation of dynamic wetting problems, our numerical scheme has to be mass conservative. Otherwise, the comparison of, e.g., numerically evaluated droplet diameters with those from experiments is impossible. However, the conventional level-set approach is rather renowned for its lack of mass conservation. Therefore, we investigate two different techniques for a better conservation of mass in this article [4, 20].

A further difficulty stems from the need for a correct implementation of the contact angle, which is necessary as a boundary condition for the level-set function φ for the computation of curvature, the level-set advection step and the reinitialization equation. Here, we present a new Neumann boundary condition for the level-set function, which is a refined version of the approach used in [9, 12].

The contribution of this article is as follows: We present a simple and effective way to include dynamic contact angle models into three-dimensional flow solvers. Furthermore, we extend the droplet impact study by Yokoi et al. [22] to three dimensions and compare their contact angle model to the reduced interface formation model by Shikhmurzaev [13]. Thereby, we obtain droplet shapes and diameters, which compare well with those from practical experiments.

The remainder of this article is organized as follows: The first section is dedicated to the details of the moving contact line problem, i.e. to the difficulties involved in the modeling of the dynamic contact angle and to its numerical implementation. In the second section, we discuss our Navier-Stokes solver and the implemented level-set method. We then explain how the contact angle can be included as a boundary condition for the level-set function. Additionally, the dynamic contact angle has to be modeled: Here, we present the model by Yokoi et al. and the interface formation model by Shikhmurzaev. In the third section, we describe the discretization of our two-phase Navier-Stokes solver. We discuss the discretization of the contact angle boundary condition and the incorporation of the dynamic contact angle models into our flow solver. Moreover, we present two different methods to improve on the mass conservation properties of our approach. In the forth section, we simulate droplet impact behavior on a dry surface and compare the evolving droplet shapes and diameters to those obtained from practical experiments. Furthermore, we compare our two improved mass conservation methods and discuss their convergence behavior. Finally, we give some concluding remarks.

2 The Moving Contact Line Problem

In this section, we describe the moving contact line problem and some of the current mathematical models for its solution. In this discussion, we include the two contact angle models by Yokoi et al. [22] and Shikhmurzaev [13], which we use in this article. Furthermore, tackling the problem from a numerical point of view, we address the question how the dynamic contact angle can be incorporated into our two-phase Navier-Stokes solver.

2.1 Modeling Issues

The key to the solution of the moving contact line problem is twofold: On the one hand, the stress singularity has to be removed and, on the other hand, the contact-angle behavior has to be modeled accurately.

In the framework of the so-called 'slip models' both problems are addressed independently: First, the no-slip condition for the velocity is relaxed and the fluid is allowed to slip at the contact line (see, e.g., [7, 8, 15]), which eliminates the stress singularity. Then, the contact angle θ_d is often chosen as a function

$$\theta_d = f(\mathrm{Ca}, \theta_s, k_1, k_2, \ldots). \tag{1}$$

of the contact line speed with the capillary number Ca, the static contact angle θ_s and material-related parameters k_i, which are used to fit the numerical results to the experiments.

A fundamental difficulty with these models is, for example, that they fail to describe the dependency of the contact angle on the flow rate, which is however, observed in experiments [13]. Thus, the dynamic contact angle is not simply a function of the contact-line speed and the material properties of the contacting media as assumed in Eq. (1). Instead, even for the same contact line speed, the dynamic contact angle can be changed by, e.g., changing the flow field or geometry near the contact line. For example, in the curtain coating context, Blake et al. [1] have shown that for a fixed substrate speed, the observed contact angle varies with the flow rate and curtain height. This effect has been termed the 'hydrodynamic assist of wetting' and it cannot be described by an equation such as (1) – where the contact angle depends foremost on the speed of the contact line.

In the theory of thin films, the moving contact line problem is circumvented by assuming that the surface is already covered by a thin film of fluid [20]. Then, with a scaled lubrication approximation, one can derive Tanner's law

$$U = A\theta^3, \tag{2}$$

where θ is now an apparent contact angle, which can be defined anywhere on the free surface. Here, U is a dimensionful velocity and A depends on the fluid properties. This law is often used for the modeling of the contact angle behavior in the framework of slip models as a variant of Eq. (1) and becomes

$$\mathrm{Ca} = k(\theta_d - \theta_s)^3. \tag{3}$$

This use of Tanner's law is conceptually questionable [13, p. 165], since there is no actual contact angle involved in the derivation of Tanner's law in the first place. In this article, we use a related kind of slip model which has been improved and extended by Yokoi et al. [22].

Instead of an ad-hoc and separable treatment of the moving contact-line problem, Shikhmurzaev [13] considers its overall physical context in the derivation of his model for the formation and disappearance of interfaces. Let us shortly state his idea: We know from experimental observation that the dynamic contact angle differs from the static one and that the Young equation (Fig. 1) holds. Therefore, we can conclude that the surface tensions in the Young equation become dynamic as well, when the phase boundary is in motion. The fluid particles at the different interfaces relax to their new equilibrium values. This process occurs in finite time and is captured by Shikhmurzaev's interface formation model.

In contrast to other approaches, the interface formation model not only removes the singularity, but is able to predict the experimentally observed rolling motion of the interface, as well as the dependence of the contact angle on the flow field. Within the literature, Shikhmurzaev's interface formation model has been considered scarcely so far, which is mostly due to its complexity (see [13] and the references therein and [17]). Therefore, we consider a reduced version of the interface formation model for small capillary numbers in this article. However, even this simplified model has been scarcely used so far [5, 11, 14].

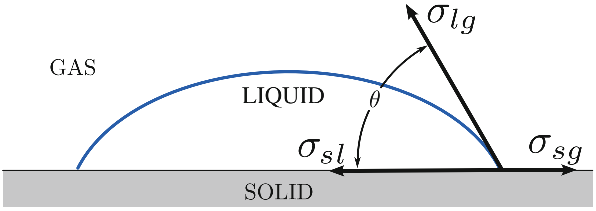


Fig. 1 Young's force diagram: $\sigma_{sg} = \sigma_{sl} + \sigma_{lg} \cos\theta$. Here, σ_{sg} is the solid/gas, σ_{sl} the solid/liquid and σ_{lg} the liquid/gas interfacial tension

Obviously, there are fundamental differences between both contact angle models considered in this article: The contact angle model by Yokoi et al. lacks a thorough underlying mathematical theory. It is developed for and based on a single droplet impact experiment only. A straightforward application of this model to the numerical simulation of other wetting experiments is difficult. Rather, this model is a prescription of contact angle values which fit well with a specific practical experiment. Since the prescribed dynamic contact angle values are very close to the ones observed in the experiment, the modeling error for the specific experiment is very small. Therefore, it gives us the opportunity to test our contact angle implementation as well as our methods for volume conservation and compare them to the results of the experiments.

On the other hand, behind Shikhmurzaev's model, there is a whole theoretical framework to explain the formation and disappearance of interfaces. This model is able to describe a vast variety of dynamic wetting phenomena. However, we use the reduced interface formation model, which is derived for small capillary numbers only. Therefore, we expect to obtain an approximate and smoothed contact line speed-contact angle relationship compared to the practical experiments and Yokoi's results in [22]. Still, the reduced model is an excellent trade-off between the complex full interface formation model and an easily implementable and reasonably accurate dynamic contact angle model. This will be seen in the remainder of this article.

2.2 *Numerical Issues*

Numerically, we have to address the difficulty of the correct implementation of the contact angle, which is needed as a boundary condition for the level-set function φ.

In the literature, a number of different approaches for the contact angle boundary condition of the level-set function can be found: Here, the simplest model is the zero Neumann boundary condition, which effectively fixes the contact angle to be 90°. If θ is variable, the implementation is less clear. One of the main approaches [22, 23] was developed by Sussman [18]. Here, the contact angle is taken into account by extrapolating the liquid interface, represented by the level-set function, into the solid. This approach requires the construction of an appropriate extension velocity, but the exact location of the position of the contact-line is not needed. Moreover, in a method by Spelt [16], contact-line position and contact angle or contact line velocity are determined iteratively.

In [24], the movement of the contact line is induced by diffusion. Instead of using a direct relation between the gradient of the level-set function and the normal of the interface, a regularized normal vector field is constructed to avoid flux of φ over the boundary. Thereby, two additional regularization parameters appear, which influence the shape of the free surface at the contact line.

In [9, 12] a Neumann boundary condition for the level-set function is derived as follows: Let our fluid flow domain Ω be a box and $\mathbf{x} = (x, y, z) \in \Omega$. Then, at, e.g., the wall $y = 0$, the geometric relation

$$\mathbf{n}_l \cdot \mathbf{n}_w = \cos(\theta) \tag{4}$$

holds for the contact angle θ. Here, $\mathbf{n}_l$ is the outward surface normal and $\mathbf{n}_w = (0, -1, 0)^t$ is the outward normal at $y = 0$. Now, a Neumann boundary condition for the level-set function can be prescribed by rewriting relation (4) as

$$\varphi_y = -\cos(\theta) \quad \text{for} \quad \|\varphi\| = 1. \tag{5}$$

However, the condition $\|\varphi\| = 1$ is not always fulfilled. Thus, the recovery of this property of the level-set function is achieved by some reinitialization equation in the first place. In this article, we extend this approach and show that

$$\varphi_y = -\cot(\theta)\sqrt{\varphi_x^2 + \varphi_z^2} \quad \text{for} \quad 0 < \theta < \pi \tag{6}$$

without further assumptions on φ. Similar techniques have already been used by Fang et al. [6] and Mourik [21] within the volume-of-fluid approach. Note here, that our approach is consistent with the extension technique by Sussman [18] for the case that the geometry is a box. Like in his approach, we do not need to locate the exact contact line position. Additionally, relation (6) allows us to set contact angle boundary conditions for complex geometrical objects in our fluid flow domain.[1]

3 Mathematical Model

In this section, we discuss the mathematical model for the three-dimensional flow of two immiscible incompressible fluids. We show how the contact angle can become a boundary condition for the level-set function and present two different models for the dynamic contact angle as given by Yokoi et al. [22] and Shikhmurzaev [13].

[1] Then, contact angles at corner cells of the geometry have to fulfill further restrictions as described in [6].

3.1 The Navier-Stokes Solver

The behavior of the fluids is governed by the incompressible Navier-Stokes equations defined on an open set $\Omega = \Omega_1 \cup \Omega_2 \cup \Gamma_f \subset \mathbb{R}^3$ with Lipschitz boundary $\Gamma := \partial\Omega$. The two fluid domains Ω_1 and Ω_2 and the free interface $\Gamma_f := \partial\Omega_1 \cap \partial\Omega_2$ depend on time. We capture the interface by a level-set formulation, and surface tension effects are included via the CSF method [2]. Thus,

$$\begin{aligned} \rho(\varphi)\left(\partial_t \mathbf{u} + (\mathbf{u}\cdot\nabla\mathbf{u})\right) + \nabla p &= \nabla\cdot(\mu(\varphi)\mathbf{S}) - \sigma\kappa(\varphi)\delta(\varphi)\nabla\varphi + \rho(\varphi)\mathbf{g} \\ \nabla\cdot\mathbf{u} &= 0 \end{aligned} \tag{7}$$

with time $t \in [0, T]$, fluid velocity $\mathbf{u}$, pressure p and volume forces $\mathbf{g}$. Here, μ is the viscosity and ρ the density. The fluid stress tensor is defined by $\mathbf{S} = \nabla\mathbf{u} + (\nabla\mathbf{u})^t$. The curvature of the free surface is given by κ, the surface tension is denoted by σ, and δ is the one-dimensional Dirac-delta functional introduced in the CSF approach.

We choose a level-set function φ as a signed-distance function such that

$$\varphi(\mathbf{x}, t) \begin{cases} < 0 & \text{if } \mathbf{x} \in \Omega_1 \\ = 0 & \text{if } \mathbf{x} \in \Gamma_f \\ > 0 & \text{if } \mathbf{x} \in \Omega_2 \end{cases} \tag{8}$$

holds and the Eikonal equation $\|\nabla\varphi\| = 1$ is fulfilled. The interface between the two fluids is then given by the zero level-set of φ:

$$\Gamma_f(t) = \{\mathbf{x} : \varphi(\mathbf{x}, t) = 0\} \tag{9}$$

for all times $t \in [0, T]$. The level-set function is advected by the pure transport equation

$$\varphi_t + \mathbf{u}\cdot\nabla\varphi = 0 \tag{10}$$

with initial value $\varphi_0(\mathbf{x}) = \varphi(\mathbf{x}, 0)$.

With the help of φ we define the density ρ and the viscosity μ on the whole domain, i.e., on both fluid-phases. To this end, we set

$$\rho(\varphi) := \rho_2 + (\rho_1 - \rho_2)H(\varphi) \quad \text{and} \quad \mu(\varphi) := \mu_2 + (\mu_1 - \mu_2)H(\varphi), \tag{11}$$

where $H(\varphi)$ denotes the Heaviside function which is defined as

$$H(\varphi) := \begin{cases} 0 & \text{if } \varphi < 0 \\ \frac{1}{2} & \text{if } \varphi = 0 \\ 1 & \text{if } \varphi > 0. \end{cases} \tag{12}$$

The Navier-Stokes equations (7) have to be complemented with boundary conditions for the pressure, the velocity and the level-set function.

In the next subsection, we present a boundary condition for the level-set function, which is required for the transport equation (10), the level-set reinitialization and the computation of the curvature κ. This condition determines the shape of the free surface at the contact line and, therefore, depends on the dynamic contact angle θ_d as soon as the contact line is moving.

Finally, in addition to the standard equations of fluid dynamics described above, the dynamic contact angle has to be modeled properly, which will be presented in Sects. 3.3 and 3.4.

3.2 The Contact Angle as a Boundary Condition

As already exemplified in Sect. 2.2, we now formulate a Neumann boundary condition for the level-set function, which incorporates the dynamic contact angle. Thus, at the boundary of Ω the contact angle is defined by the geometric relation

$$\mathbf{n}_l \cdot \mathbf{n}_w = \cos(\theta), \tag{13}$$

where θ is the contact angle (static or dynamic), $\mathbf{n}_w$ is the outward normal drawn from the flow region into the boundary, and $\mathbf{n}_l$ is normal to the level-set function which points from the fluid phase with lower level-set values to the one with higher values, i.e.

$$\mathbf{n}_l = \frac{\nabla\varphi}{\|\nabla\varphi\|}. \tag{14}$$

Then, the boundary condition for the level-set function is given in the following proposition.

Proposition 1. *At any wall of Ω, whose outward normal is given by $\mathbf{n}_w^i = \pm\mathbf{e}^i$ for some $i \in \{1,2,3\}$, the level-set's i-th derivative φ_{x_i} can be related to θ by*

$$\varphi_{x_i} = \pm\cot(\theta)\sqrt{\sum_{j=1,\, j\neq i}^{3} \varphi_{x_j}^2} \tag{15}$$

for any angle $0 < \theta < \pi$.

Proof. We prove this proposition here for two walls with outward normals $\mathbf{n}_w^2 = -\mathbf{e}^2$ and $\mathbf{n}_w^2 = \mathbf{e}^2$, since the cases $\mathbf{n}_w^i = \pm\mathbf{e}^i$ for $i = 1$ or $i = 3$ can be treated in the very same way. For both boundaries, we have to distinguish between the cases $0 < \theta \leq \frac{\pi}{2}$ and $\frac{\pi}{2} < \theta < \pi$.

Let $\mathbf{n}_w^2 = -e^2 = (0, -1, 0)^t$. Then, from Eqs. (13) and (14), we have

$$\mathbf{n}_l \cdot \mathbf{n}_w^2 = \cos(\theta) \Leftrightarrow -\varphi_{x_2} = \cos(\theta)\|\nabla\varphi\|. \tag{16}$$

First, let $0 < \theta \leq \frac{\pi}{2}$. Then, $0 \leq \cos\theta < 1$ and $\sin^2(\theta) = 1 - \cos^2(\theta) > 0$. Since $\cos(\theta) \geq 0$, we conclude that $-\varphi_{x_2} \geq 0$ as well, and define the positive function $\tilde{\varphi} := -\varphi_{x_2}$ with $\tilde{\varphi}^2 = \varphi_{x_2}^2$. Inserting $\tilde{\varphi}$ into Eq. (16), we obtain

$$\tilde{\varphi} = \cos(\theta)\sqrt{\varphi_{x_1}^2 + \tilde{\varphi}^2 + \varphi_{x_3}^2}.$$

Now, only positive variables constitute both sides of the equation. Thus, we are allowed to take the square of both sides and still obtain the equivalent relation

$$\begin{array}{rrcl} & \tilde{\varphi}^2 & = & \cos^2(\theta)(\varphi_{x_1}^2 + \tilde{\varphi}^2 + \varphi_{x_3}^2) \\ \Leftrightarrow & \tilde{\varphi}^2\left(1 - \cos^2(\theta)\right) & = & \cos^2(\theta)(\varphi_{x_1}^2 + \varphi_{x_3}^2) \\ \Leftrightarrow & \tilde{\varphi}^2 & = & \dfrac{\cos^2(\theta)}{\sin^2(\theta)}(\varphi_{x_1}^2 + \varphi_{x_3}^2) \\ \Leftrightarrow & \tilde{\varphi} & = & \dfrac{\cos(\theta)}{\sin(\theta)}\sqrt{\varphi_{x_1}^2 + \varphi_{x_3}^2}. \end{array}$$

Again, taking the root to obtain the last equivalency relation is only allowed since all parts of the equation (including $\sin^2(\theta)$) are greater than or equal to zero. Then, for $0 < \theta \leq \frac{\pi}{2}$, we have $\sin(\theta) = \sqrt{1 - \cos^2(\theta)}$. Resubstituting $\tilde{\varphi} = -\varphi_{x_2}$, we obtain the desired result

$$\varphi_{x_2} = -\cot(\theta)\sqrt{\varphi_{x_1}^2 + \varphi_{x_3}^2}.$$

Now, let $\frac{\pi}{2} < \theta < \pi$. Then, $-1 < \cos\theta < 0$ and $\sin^2(\theta) = 1 - \cos^2(\theta) > 0$. From (16), we know that $\varphi_{x_2} > 0$, since $\cos\theta < 0$. We define the positive function $\tilde{c} := -\cos(\theta)$ with $\tilde{c}^2 = \cos^2(\theta)$ and obtain likewise

$$\varphi_{x_2} = \tilde{c}\sqrt{\varphi_{x_1}^2 + \varphi_{x_2}^2 + \varphi_{x_3}^2} \Leftrightarrow \varphi_{x_2} = -\cot(\theta)\sqrt{\varphi_{x_1}^2 + \varphi_{x_3}^2}.$$

In the second part of this proof, the outward normal of the boundary is given by $\mathbf{n}_w^2 = \mathbf{e}^2$. Then, from (13) and (14), we have

$$\mathbf{n}_l \cdot \mathbf{n}_w^2 = \cos(\theta) \Leftrightarrow \varphi_{x_2} = \cos(\theta)\|\nabla\varphi\|. \tag{17}$$

Again, we consider the two cases $0 < \theta \leq \frac{\pi}{2}$ and $\frac{\pi}{2} < \theta < \pi$. First, let $0 < \theta \leq \frac{\pi}{2}$. Then, $0 \leq \cos\theta < 1$ and $\sin^2(\theta) = 1 - \cos^2(\theta) > 0$. From (17), we know that $\varphi_{x_2} > 0$, since $\cos\theta > 0$. Similar to the first case above, we can then show that

$$\varphi_{x_2} = \cos(\theta)\sqrt{\varphi_{x_1}^2 + \varphi_{x_2}^2 + \varphi_{x_3}^2} \Leftrightarrow \varphi_{x_2} = \cot(\theta)\sqrt{\varphi_{x_1}^2 + \varphi_{x_3}^2}.$$

Now, let $\frac{\pi}{2} < \theta < \pi$. Then, $-1 < \cos\theta < 0$ and $\sin^2(\theta) = 1 - \cos^2(\theta) > 0$. From Eq. (17), we know that $-\varphi_{x_2} > 0$, since $\cos\theta < 0$. We define the positive function $\tilde{c} := -\cos(\theta)$ with $\tilde{c}^2 = \cos^2(\theta)$ and the positive function $\tilde{\varphi} := -\varphi_{x_2}$ with $\tilde{\varphi}^2 = \varphi_{x_2}^2$. With these definitions and similar equivalency relations as in all the other cases, we obtain

$$\tilde{\varphi} = \tilde{c}\sqrt{\varphi_{x_1}^2 + \tilde{\varphi}^2 + \varphi_{x_3}^2} \Leftrightarrow \varphi_{x_2} = \cot(\theta)\sqrt{\varphi_{x_1}^2 + \varphi_{x_3}^2}. \qquad \square$$

The above proposition allows us to write the contact angle θ as a boundary condition of the level-set function. Now, the hard part is a reliable model for the dynamic contact angle $\theta = \theta_d$. In the next two subsections, we will describe two different dynamic contact angle models, which we will later employ in our numerical method.

3.3 The Dynamic Contact Angle Model by Yokoi et al. (C1)

Yokoi et al. [22] propose a dynamic contact angle model, which combines Tanner's law (3) with a static advancing and receding contact angle, since Tanner's law holds for small capillary numbers only. In the combined model, similar to what is observed in experiments, the contact angle tends to both the limit of a maximum advancing angle θ_{mda} as the dimensionful contact line speed u_{cl} increases and the limit of a minimum dynamic receding angle θ_{mdr} as u_{cl} decreases:

$$\theta_d = \begin{cases} \min\{\theta_s + \left(\frac{\mu u_{\mathrm{cl}}}{\sigma k_a}\right)^{\frac{1}{3}}, \theta_{\mathrm{mda}}\} & \text{if} \quad u_{\mathrm{cl}} \geq 0 \\ \max\{\theta_s + \left(\frac{\mu u_{\mathrm{cl}}}{\sigma k_r}\right)^{\frac{1}{3}}, \theta_{\mathrm{mdr}}\} & \text{if} \quad u_{\mathrm{cl}} < 0. \end{cases} \tag{C1}$$

Here, θ_s is the static contact angle, and the material-related parameters k_a and k_r are adjusted to fit the numerical results to the results obtained by measurements. Furthermore, the stress singularity is circumvented by inducing numerical slip for the velocity at the contact line.

3.4 The Dynamic Contact Angle Model by Shikhmurzaev (C2)

The second model for the dynamic contact angle at small capillary number is a reduced version of Shikhmurzaev's interface formation model [13]. The full model accounts for different classes of flows, where interfaces are formed or destroyed. The equations, which capture the surface tension relaxation process and have to be solved on the surface itself, are derived from mass, momentum and energy conversation. For the case of small capillary and Reynolds numbers, we can analyze them as a local problem whose solution can be incorporated into various types of global flow problems. Here, lots of experimental works simplify the verification of numerical results.

We follow the description in [13]. There, the flow domain is split into two asymptotic regions, and in both the limit Ca → 0 is studied analytically. In the inner asymptotic region to leading order in Ca the dynamic contact angle and the dimensionless contact-line speed V are related by

$$\cos(\theta_s) - \cos(\theta_d) = \frac{2V\left[\cos(\theta_s) - \sigma_{sg} + (1-\rho_G^s)^{-1}(1+\rho_G^s u_{(12)}(\theta_d, k_\mu))\right]}{V + \left[V^2 + 1 + (\cos(\theta_s) - \sigma_{sg})(1-\rho_G^s)\right]^{\frac{1}{2}}} \tag{C2}$$

with θ_s the static contact angle, k_μ the gas-to-liquid viscosity ratio, and $\rho_G^s \equiv 1 - \frac{\sigma_{sg}-\sigma_{sl}}{\lambda \cos\theta_d}$, where σ_{sg} and σ_{sl} denote the surface tension in the gas-solid and liquid-solid interface, respectively, and λ is a material parameter. Here, the radial velocity $u_{(12)}(\theta_d, k_\mu)$ must be derived from the solution in the outer region.

In particular, in the outer region to leading order in Ca the free-surface curvature becomes zero and one obtains a flow problem in a wedge [13]. The solution to this problem was given by Moffatt in [10] as

$$u_{(12)}(\theta_d, 0) = \frac{\sin\theta_d - \theta_d\cos\theta_d}{\sin\theta_d\cos\theta_d - \theta_d}. \tag{18}$$

If the viscosity of the gas phase is taken into account, Moffatt's solution becomes

$$u_{(12)}(\theta_d, k_\mu) = \frac{(\sin\theta_d - \theta_d\cos\theta_d)K(\theta_2) - k_\mu(\sin\theta_2 - \theta_2\cos\theta_2)K(\theta_d)}{(\sin\theta_d\cos\theta_d - \theta_d)K(\theta_2) + k_\mu(\sin\theta_2\cos\theta_2 - \theta_2)K(\theta_d)}, \tag{19}$$

with $\theta_2 = \pi - \theta_d$ and $K(\theta) = \theta^2 - \sin^2\theta$ [13].

Alternatively, $u_{(12)}(\theta_d, k_\mu)$ in (C2) can be replaced by the inner limit of the outer solution, i.e. by a numerically computed far field velocity sufficiently close to the contact line. This alters the dynamic contact angle for the same contact line speed and is exactly what is observed in laboratory experiments as the nonlocal influence of the flow field/geometry on the dynamic contact angle.

As described in [13], we introduce the reference velocity U and the scaling factor Sc by

$$U = \sqrt{\frac{\gamma\rho_0^s(1+4\alpha\beta)}{\tau\beta}} \text{ and } \mathrm{Sc} = \sqrt{\frac{\sigma^2\tau\beta}{\mu^2\gamma\rho_0^s(1+4\alpha\beta)}}. \tag{20}$$

Here, σ is the equilibrium surface tension, α and β are phenomenological constants depending on the 'state of the interface', γ is a phenomenological constant

Table 1 Parameters for distilled water impacting on a silicon wafer onto which hydrophobic silane has been grafted

Distilled water:	$\mu_l = 1.0_{-3}\,\mathrm{kg/ms}$, $\rho_l = 1_3\,\mathrm{kg/m^3}$	Surface tension:	$\sigma = 7.2_{-2}\,\mathrm{N/m}$
Air:	$\mu_g = 1.82_{-5}\,\mathrm{kg/ms}$, $\rho_g = 1.25_0\,\mathrm{kg/m^3}$	Boundary conditions:	no-slip
Forces:	$\mathbf{g} = (0\,, -9.81,\, 0)\,\mathrm{m/s^2}$	Droplet diameter	$2.28_{-3}\,\mathrm{m}$
Inital velocity:	$\mathbf{u} = (0\,, -1.0, 0)\,\mathrm{m/s}$	k_a:	9.0_{-9}
θ_{mda}:	114°	k_r:	9.0_{-8}
θ_{mdr}:	52°		

describing the compressibility of the fluid, τ is the surface tension relaxation time and ρ_0^s is the surface density for zero surface tension, both of which can be treated as material constants. Thus, Sc depends on the material properties of the fluid and the interface. Then, the dimensionless contact line velocity is given by

$$V = \frac{u_{\mathrm{cl}}}{U} = \frac{u_{\mathrm{cl}}\mu}{\sigma}\,\mathrm{Sc}, \tag{21}$$

and Sc can be chosen to fit the numerical results to the experimental data.

Let us demonstrate how the dimensionless parameter Sc influences the results of Eq. (C2). As an example, we consider a droplet of distilled water which impacts on a silicon wafer onto which hydrophobic silane has been grafted. The equilibrium contact angle of the substrate with distilled water is 90°, and the relevant physical and numerical parameters of this experiment are listed in Table 1.

Thus, (C2) can be resolved with respect to positive V as given in [13]. Since we are interested in the relationship between the dynamic contact angle and the dimensionful contact line velocity, we also use (21) to plot the dimensionful speed-angle relationship for different values of Sc. The remaining parameters are chosen according to Table 1, $\rho_G^s = 0.9$, $\sigma_{sg} = 0$, and $u_{(12)}(\theta_d\,, 0)$ is determined by Moffatt's solution (18). The effect of the variation of Sc is shown in Fig. 2: We see that the contact angle as a function of u_{cl} increases from its static value θ_s and tends the faster to 180° the more we increase Sc. The horizontal straight red line indicates the maximum dynamic advancing contact angle of 114° determined from the experiments. For $\mathrm{Sc} = 11.5$ this value is reached at a maximum contact line speed of about $u_{\mathrm{cl}} = 0.4\ \mathrm{m/s}$.

4 The Numerical Method

In this section, we describe the discretization of the Navier-Stokes equations (7) in space and time with special emphasis on the level-set method. We discuss the implementation of the contact angle boundary condition and of the two contact angle models. Last, we present two methods for a better conservation of mass within the level-set method.

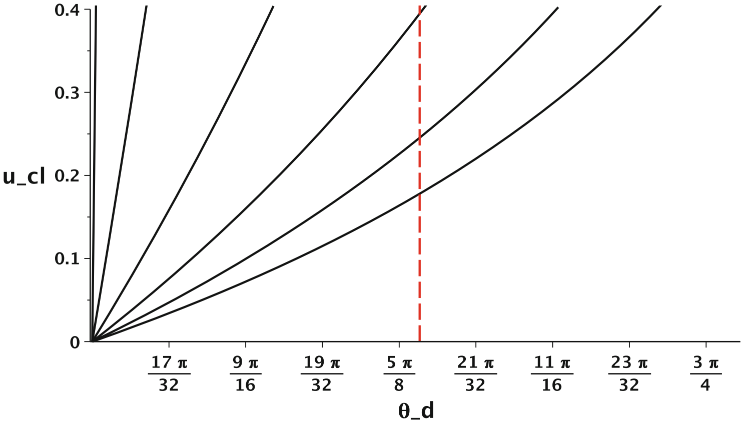


Fig. 2 Dynamic contact angle vs. dimensionful contact-line speed for different values of Sc given in [rad] and [m/s], respectively. From *left* to *right*: Sc = 0.1, 1.5, 5.5, 11.5, 18.5, 25.5. The *straight dashed line* corresponds to the maximum of the dynamic advancing contact angle of 114° observed in the experiments

4.1 Discretization of the Navier-Stokes Equations and the Level-Set Method

We discretize the Navier-Stokes equations with finite differences on a staggered uniform grid and use an explicit second-order Adams-Bashforth time integration scheme. The solution process is based on the well-known projection method: First, an intermediate velocity field $\mathbf{u}^*$, which may not be divergence free, is advanced by the Adams-Bashforth time scheme; second, we compute a correction ∇p^{n+1} of the intermediate velocity field by the pressure Poisson equation which leads to a divergence free velocity field $\mathbf{u}^{n+1}$. Thus, we treat the pressure implicitly and solve the Poisson equation by a Jacobi-preconditioned conjugate gradient method. A fifth-order weighted essentially non-oscillatory (WENO) scheme is used for the discretization of the convective transport of the Navier-Stokes equations (7) as well as for the level-set transport (10). The diffusion term is computed by using second-order central differences.

For the treatment of the free surface between the two fluid phases we employ the level-set approach [3, 4]. Here, the interface conditions are implicitly incorporated into the momentum equations by the continuum surface force (CSF) [2] method.

Note that we have to reinitialize the level-set function φ^* after each transport step to recover its signed distance property $|\nabla\varphi^{n+1}| = 1$ without disturbing the zero level-set. To generate the appropriate signed-distance function $\varphi^{n+1}(\mathbf{x})$ with the same zero level-set as $\varphi^*(\mathbf{x})$, we solve the following pseudo-transient Hamilton-Jacobi problem to steady state

$$\varphi^*_\tau = \text{sign}(\varphi_0)(1 - |\nabla\varphi^*|) \tag{22}$$

with initial value $\varphi_0 = \varphi^*(\mathbf{x})$. Again, we discretize this equation by a fifth order WENO scheme in space and employ a third-order Runge-Kutta for its time-integration.

For reasons of numerical stability, we employ a regularized signum function

$$S(\varphi^*) = \frac{\varphi^*}{\sqrt{(\varphi^*)^2 + |\nabla\varphi^*|^2(\delta x^2)}}. \tag{23}$$

and a smoothed Heaviside and Dirac-delta functional in an ε-environment of the free surface. Then the Hamilton-Jacobi problem reads

$$\varphi^*_\tau = S(\varphi^*)(1 - |\nabla\varphi^*|). \tag{24}$$

For further details on the implementation of our Navier-Stokes solver NaSt3DGPF and of the level-set method see [3, 4].

4.2 Discretization of the Contact Angle Boundary Condition

The discretization of the contact angle boundary condition (15) is very similar to the discretization of the standard Neumann boundary condition for the level-set function. Again, we exemplify this at the wall $y = 0$, where Eq. (15) becomes

$$\varphi_y = -\cot(\theta)\sqrt{\varphi_x^2 + \varphi_z^2}. \tag{25}$$

On the staggered grid (Fig. 3), the level-set values are discretized in the cell center. Then, with grid cells denoted by integers (i, j, k),

$$\frac{\varphi_{i,j,k} - \varphi_{i,j-1,k}}{\delta y_j} = -\cot(\theta)\sqrt{\varphi^2_{x_{i,j,k}} + \varphi^2_{z_{i,j,k}}}, \tag{26}$$

where δy_j is the mesh width. The derivatives $\varphi_{x_{i,j,k}}$ and $\varphi_{z_{i,j,k}}$ can be discretized by central differences. This equation can be solved for the staggered grid's ghost cell value $\varphi_{i,j-1,k}$, which gives the required boundary condition for φ.

The values for the contact angle θ are computed by the discretized dynamic contact angle models of Yokoi et al. (C1) or Shikhmurzaev (C2), which we will discuss subsequently.

4.3 Implementation of the Contact Angle Models

In this subsection, we describe how the contact angle models are incorporated into our two-phase Navier-Stokes solver. For this, both models require the computation

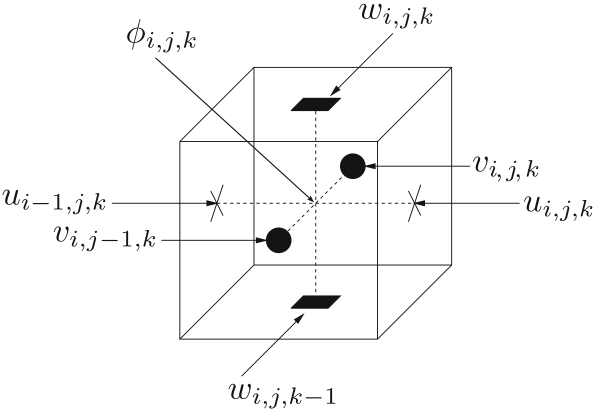


Fig. 3 On the staggered grid, the level-set function φ is discretized at the cell center and the velocity is discretized at the face centers of the grid

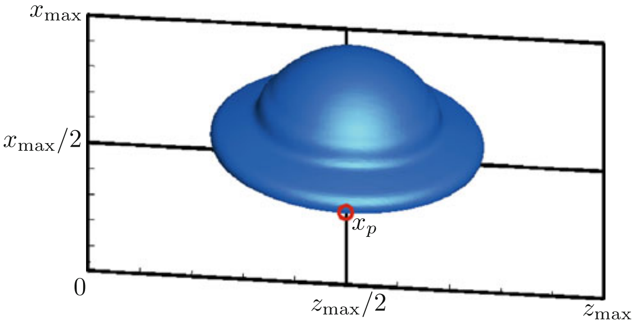


Fig. 4 The contact line velocity is evaluated at the contact point x_p at the intersection with the line $z = z_{\max}/2$

of the contact line velocity u_{cl}. Additionally, for Shikhmurzaev's model, we also need the velocity $u_{(12)}(\theta_d, k_\mu)$.

In the following, we focus on the example of drop impact, and we assume that the drop spreads symmetrically (cf. Fig. 4). Then u_{cl} is taken as the velocity value u in x-direction which is closest to the contact point x_p at the line $z_{\max}/2$ and still lies in the droplet's fluid phase. This simplified computation of the contact line velocity is also done by Yokoi et al. [22] and we stick to it for the sake of comparability.

Furthermore, we compute $u_{(12)}$ either by Moffatt's solution (18) or we use a velocity value of the far field. This far field velocity value is arbitrarily chosen to be about two grid cells away from the contact line. Thus, if e.g. at $y = 0$, $\varphi_{i,1,k} \cdot \varphi_{i+1,1,k} < 0$ and $\varphi_{i,1,k}$ in the liquid phase, $u_{(12)} = u_{i-1,2,k}$. In the following, we will refer to these two options for $u_{(12)}$ as (M1) and (M2), respectively.

For (M1), the contact angle Eq. (C2) becomes nonlinear and we invoke a Newton iteration method to solve for θ_d. For (M2), the equation can be solved directly by evaluating the arccos-function. Here, if the argument of the arccos is not in $[-1, 1]$, we use Moffatt's solution (M1) instead.

All in all, the contact line models fit into our flow solver as follows:

1. Let θ^n be given from the previous time-step.
2. Solve the level-set advection Eq. (10) with the boundary condition (15) and $\theta = \theta^n$.
3. C1: Use a velocity value near x_p^{n+1} for u_{cl}^{n+1} and compute θ^{n+1} from (C1)
   C2: Use a velocity value near x_p^{n+1} for u_{cl}^{n+1}. Compute the radial velocity $u_{(12)}$ by (M1) or (M2) and θ^{n+1} from (C2).
4. Solve the level-set reinitialization (24) with the boundary condition (15) and $\theta = \theta^{n+1}$.

Finally, note that we use the no-slip condition for the velocity for both contact angle models. On the staggered grid, as drawn in Fig. 3, the no-slip condition is never fulfilled exactly, which introduces enough numerical slip to eliminate the stress singularity at the contact line.

4.4 Methods for Mass Conservation

An important issue – especially for level-set methods – is mass conservation. In the reinitialization step the current level-set function is replaced by a smoother, less distorted function which has the same zero level-set. However, this also introduces numerical diffusion to the solution which leads to difficulties with volume conservation. To this end, we use a global and a local volume correction method to remedy this problem.

For the global volume correction, already described and investigated in [4], we employ a Picard iteration after the reinitialization step:

$$\varphi^{n+1} \leftarrow \varphi^{n+1} + \omega(V(\varphi^0) - V(\varphi^{n+1})). \tag{27}$$

Here, $V_i(\varphi^0)$ is the initial volume of Ω_2^0 and $V(\varphi^{n+1}) := \int_\Omega H(\varphi^{n+1})d\mathbf{x}$ denotes the volume of Ω_2^{n+1} at time $t = n + 1$ after the reinitialization procedure. The relaxation parameter ω depends on the specific problem and is chosen to minimize the number of iterations in the relaxation process.

For the local volume correction, we follow [19] in improving the re-distancing algorithm of the level-set function by formulating a constraint which conserves the volume of the domain and prevents the straying of the level-set function from its initial position. We require that

$$\partial_t \int_\Omega H(\varphi^*)d\mathbf{x} = 0 \tag{28}$$

and modify the Hamilton-Jacobi problem by

$$\varphi^*_\tau = \operatorname{sign}(\varphi_0)(1 - |\nabla\varphi^*|) + \lambda f(\varphi^*). \tag{29}$$

Then we determine the time-dependent function λ by

$$\begin{aligned}\partial_\tau \int_\Omega H(\varphi^*) d\mathbf{x} = \int_\Omega H'(\varphi^*)\varphi^*_\tau d\mathbf{x} = \\ \int_\Omega H'(\varphi^*) \left(\operatorname{sign}(\varphi_0)(1 - |\nabla\varphi^*|) + \lambda f(\varphi^*)\right) d\mathbf{x},\end{aligned} \tag{30}$$

i.e.

$$\lambda = \frac{-\int_\Omega H'(\varphi^*) \operatorname{sign}(\varphi_0)(1 - |\nabla\varphi^*|) d\mathbf{x}}{\int_\Omega H'(\varphi^*) f(\varphi^*) d\mathbf{x}}. \tag{31}$$

The choice of

$$f(\varphi^*) = H'(\varphi^*)\, |\nabla\varphi^*| \tag{32}$$

ensures that the correction takes place at the interface only.

The discretization of the local mass correction in two-dimensions is described in [19]. In three dimensions, the numerical integration of some function g over the domain

$$\Omega_{ijk} = \{(x, y, z) \in \Omega : x_{i-\frac{1}{2}} < x < x_{i+\frac{1}{2}}, y_{j-\frac{1}{2}} < y < j_{j+\frac{1}{2}}, z_{k-\frac{1}{2}} < z < z_{k+\frac{1}{2}}\},$$

changes to

$$\int_{\Omega_{ijk}} g_{ijk} \quad d\mathbf{x} \approx \frac{1}{78}\Bigg[52 g_{ijk}(\delta x_i \delta y_j \delta z_k) + \sum_{\substack{p,q,r=-1 \\ (p,q,r)\neq(0,0,0)}}^{1} \left(g_{i+p;j+q;k+r}(\delta x_i \delta y_j \delta z_k)\right)\Bigg].$$

Furthermore, in the original article [19], a non-smooth signum-function is employed. For better conservation properties and numerical stability, we again choose a smooth variant and replace $\operatorname{sign}(\varphi_0)$ by $S(\varphi^*)$ as given in (24).

This volume correction is 'local' since the mass should remain unchanged in any sub-domain of Ω, so that $\int H(\varphi^*) d\mathbf{x}$ is preserved in every grid cell. It is also 'local' in a negative sense, since it only prevents the straying of the level-set function, but does not correct mass errors which occur due to the numerical diffusion introduced when solving the transport equation (10).

5 Numerical Results

In this section we evaluate the mathematical and numerical models for the example of a droplet impact simulation. Specifically, we consider a droplet of distilled water which impacts on a silicon wafer onto which hydrophobic silane has been grafted. The equilibrium contact angle of the substrate with distilled water is 90°, and the relevant physical and numerical parameters of this experiment have already been listed in Table 1.

The numerical simulation of this specific droplet impact scenario is valuable due to various reasons: First, Yokoi et al. [22] provide experimental results for the droplet behavior. Thus, we can compare the droplet shape, droplet diameter and dynamic contact angle from the physical experiments with the numerical results of our dynamic contact angle models (C1) and (C2). Additionally, we do not have to re-adapt the parameters k_a and k_r in (C1), since the contact angle model has been designed for this specific experiment. Second, Yokoi et al. [22] present two-dimensional numeric results in their work, which we can use for comparison with our three-dimensional results as well. Last, in this specific droplet impact experiment, the numerically computed droplet behavior is very sensitive to the applied contact angles: For example, using the static contact angle instead of a dynamic contact angle model causes the drop to rebound; the same happens if only static advancing and receding contact angles are applied; see [22] for further details. Therefore, this specific kind of droplet impact simulation is a very sensitive test case for our Navier-Stokes solver, the implemented contact angle boundary condition and the two contact angle models.

However in the work of Yokoi et al., we also see the difficulty in obtaining accurate experimental results. There, the presented droplet shapes are obtained from a different experiment (E1) than the measured contact angle and diameter (E2). For the latter, only the right hand side of the droplet has been observed to increase the resolution around the contact line. If we compare the experimental droplet shapes in Figs. 5 and 6 with the experimentally measured droplet diameter in Fig. 7, we see that at times $t = 10$ and 15 ms the diameter of the droplet shapes (obtained from E1) is visibly smaller than the one given in Fig. 7 (obtained from E2), which gives us an indication of the involved measuring error.

Let us give an example for the discrepancy between experiments (E1) and (E2). In Figs. 5, 6 and 8 at $t = 4$ ms all numerical methods predict a horizontally wider droplet than the laboratory experiment (E1), which is depicted in the first row of the respective figures. However, if you compare this result with the diameter-time curve in Figs. 7 and 9, the numerically computed droplet diameter at $t = 4$ ms reproduces the droplet diameter from the laboratory experiment (E2) nearly perfectly – at least for the highest numerical resolution, which is also used in the pictures for the droplet shapes.

In the following, we present our results in three steps: First, we take Yokoi's model (C1) and compare the experimentally evaluated droplet shapes and diameters with our three-dimensional simulation computed by the pure level set method and

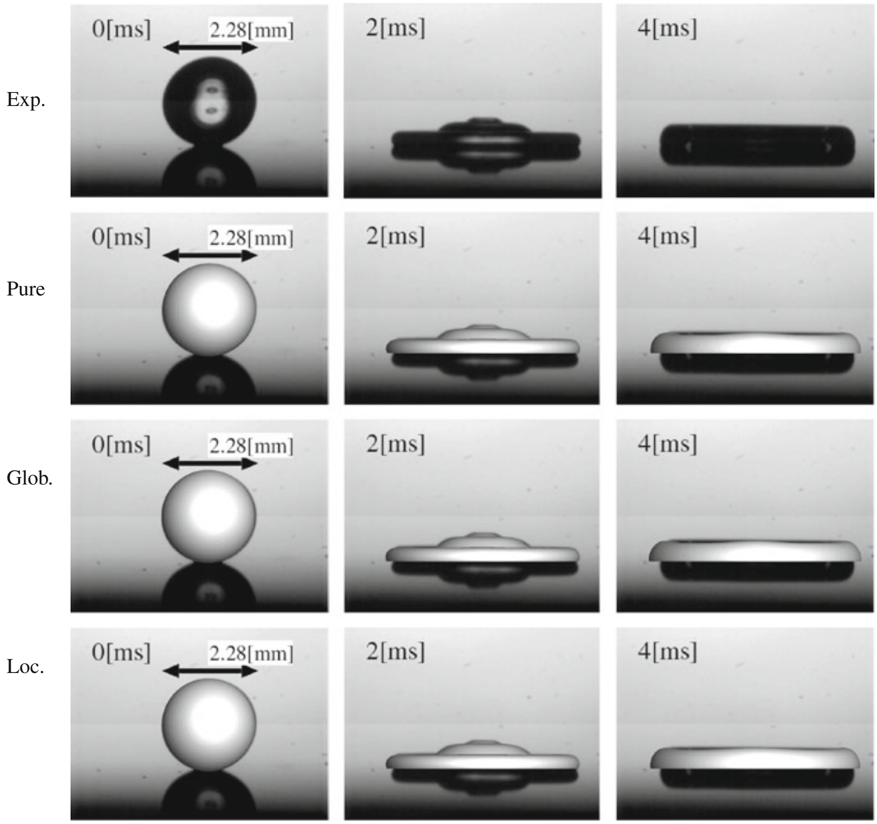


Fig. 5 Droplet impact $t = 0$–4 ms. In descending order: experimental results, pure level set method, level-set with global volume-correction, level-set with local volume-correction

the two volume correction methods. Second, we use two variants of the reduced interface formation model (C2) to simulate the same droplet with global volume correction only. In a last step, we discuss the mass conservation behavior of our numerical methods.

5.1 The Contact Angle Model by Yokoi

In this subsection we present the results of the droplet impact simulation with the Navier-Stokes equations (7) and (C1). The numerically obtained droplet shapes (white) during the impact are shown in Figs. 5 and 6 compared with experiments by Yokoi et al. [22] (black).

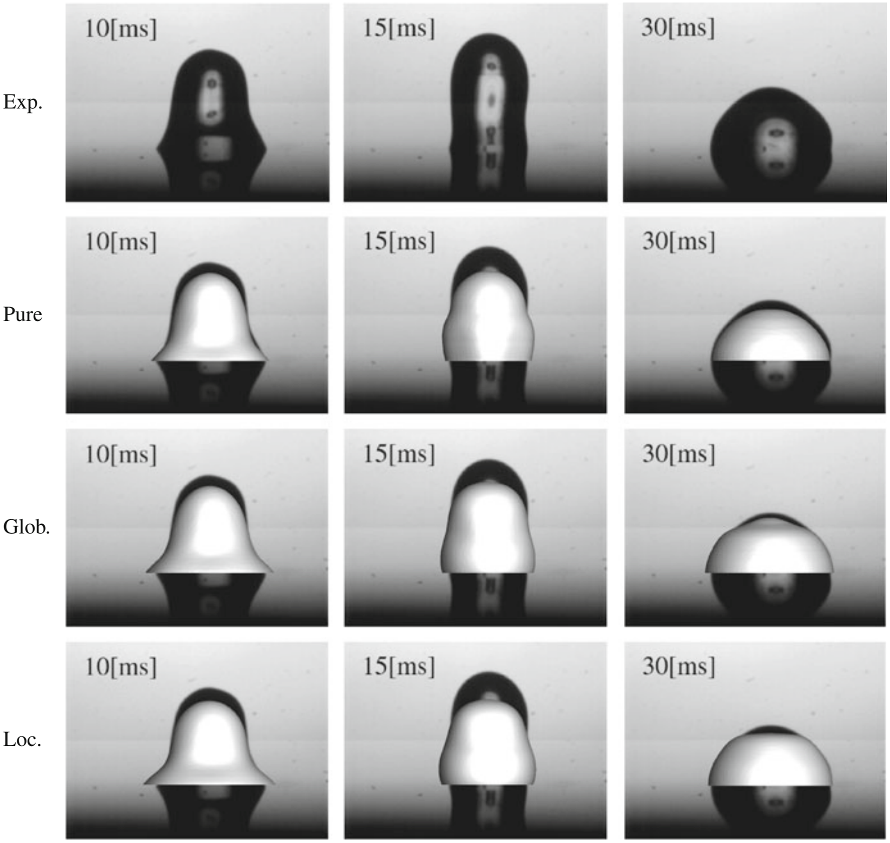


Fig. 6 Droplet impact $t = 10$–30 ms. In descending order: experimental results, pure level set method, level-set with global volume-correction, level-set with local volume-correction

The results of the laboratory experiments (E1) are always shown in the first row of the respective figures. The second row corresponds to the computed numerical solution with the pure level set method, i.e. without any volume-correction methods: We see that during the first three points in time (Fig. 5), when inertia is dominant, these droplet shapes compare very well with the experimental snapshots from (E1). At $t = 10$ ms the simulated droplet shape is still remarkably close to the experiment, while at the next time steps, the numerical droplet fails to reproduce the correct droplet height and width of the experiment (Fig. 6). This is partly due to its obvious loss in mass: Despite the very high resolution, the droplet still looses about 20 % of its volume during the simulation. Thus, the comparison with the droplet diameter measured in the laboratory experiment (E2) becomes difficult as well. In Fig. 7, we see that the pure level set method is unable to produce the final droplet diameter for all applied resolutions. Nevertheless, at about $t = 4$ ms the maximum diameter is well recovered and the overall behavior of the diameter-time curve is close to

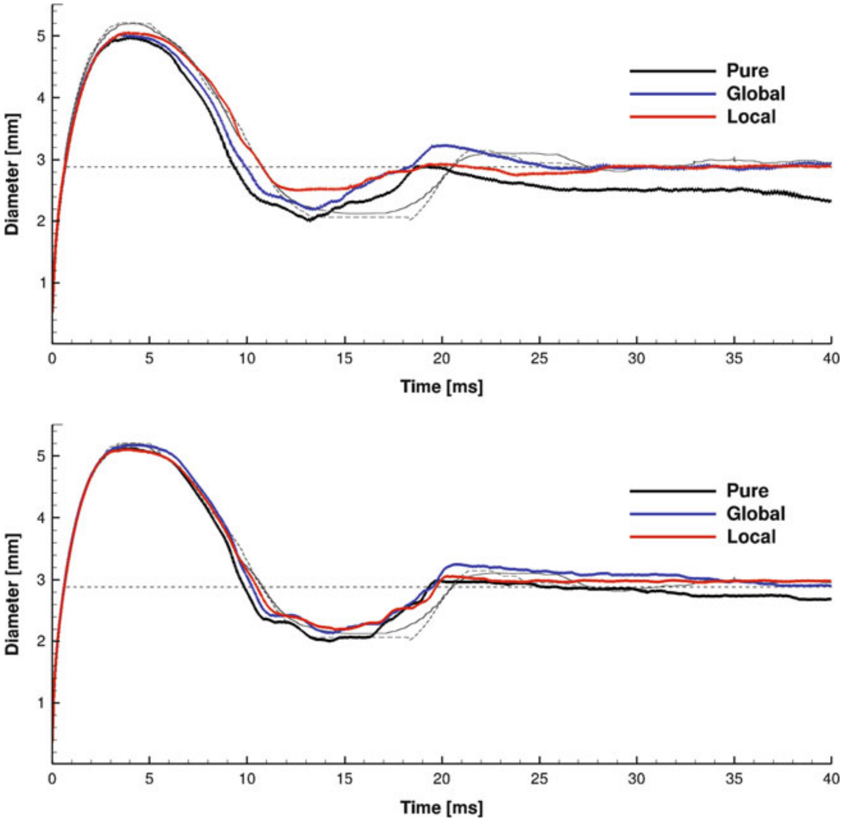


Fig. 7 Comparison of droplet diameter over time with experimental results. Our 3D simulations with the pure level set method (*black*), global (*blue*) and local (*red*) volume correction compared to 2D numerical and experimental results by Yokoi et al. [22]. The theoretical final droplet diameter is given by the *straight dashed line*. The *thin black line* corresponds to the 2D numerical results and is very close to that of the experiments given by the *thin dashed line*. Above, the grid resolution is 121 × 61 × 121 and below it is 241 × 121 × 241

the experiment. We see here that the simulated droplet diameter at times $t = 10$ and 15 ms is closer to the experiment (E2) than to (E1). All in all, despite its obvious loss in volume and within the experimental measuring error, the pure level set method is able to produce simulation results, which correspond well with the experimentally observed droplet behavior.

At this point, we nevertheless see the need for volume-conserving simulation methods. Therefore, we simulate the droplet impact with the global volume correction and the local volume correction. The resulting droplet shapes are presented in the third and forth row of Figs. 5 and 6.

We expect that the results from the global volume correction are very close to those by the pure level set method, since the fix-point iteration tends to simply inflate

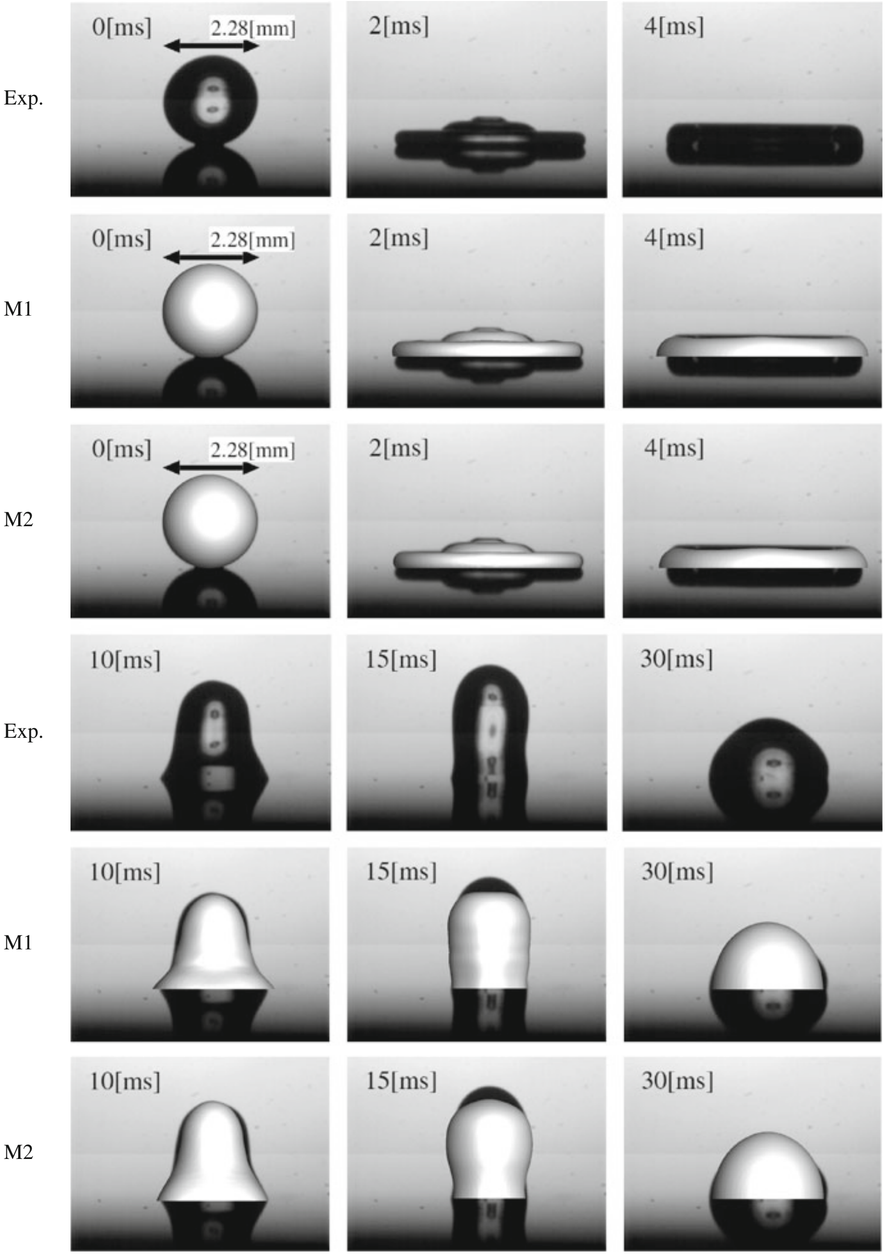


Fig. 8 Droplet impact $t = 0$–30 ms. In descending order: experimental results, reduced interface formation model with Moffatt's solution (M1) and $\mathrm{Sc} = 11.5$, reduced interface formation model with far field velocity (M2) and $\mathrm{Sc} = 5.5$

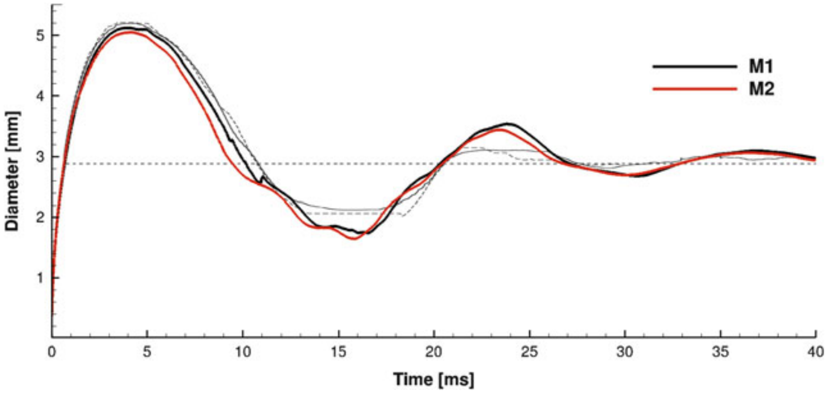


Fig. 9 Comparison of droplet diameter over time with experimental results. Our 3D simulations with M1 (*black*) and M2 (*red*) compared to 2D numerical and experimental results by Yokoi et al. [22]. The theoretical final droplet diameter is given by the *straight dashed line*. The *thin black line* corresponds to the 2D numerical results and is very close to that of the experiments given by the *thin dashed line*. Above, the grid resolution is 241 × 121 × 241

the droplet. This is exactly what we observe in Figs. 5 and 6: The droplet shapes recovered by the global correction method are similar to those by the pure level set method, but the volume of the drop is now preserved up to 100 %. Again, the droplet diameters produced by the global volume correction method are closer to experiment (E2) than to (E1). If we compare the droplet diameter with (E2), we see that its evolution over time is also very similar to the pure level set method's results (Fig. 7). Due to the improved volume conservation, the maximum droplet diameter at $t = 4$ ms and the final droplet diameter are captured excellently by the global volume correction method.

If we apply the local volume correction, we get results which also compare better with (E2) than with (E1): The droplet diameter results agree with the experimental ones, as we can see from Fig. 7. For the first three points in time, the droplet shapes with the local volume correction method are in good agreement with the other simulation results and the experiments (Fig. 5). However, we then get a larger deviation from (E1) at times $t = 10$ and 15 ms (Fig. 6); where the droplet width is more overshot than with the other two methods.

If we compare our two volume correction methods, we first note that both are able to conserve the volume of the droplet up to nearly 100 % for the highest grid resolution. Second, we observe that the global volume correction simply inflates the droplet everywhere, while the local volume correction tends to widen the droplet horizontally. All in all, we conclude that both correction methods are in good agreement with the experimental results and lie well within the scope of the experimental error.

In a next step, we compare our three-dimensional results to the two-dimensional ones by Yokoi et al. obtained by a coupled level set and volume-of-fluid method.

The 2D droplet diameter is given in Fig. 7 and is nearly indistinguishable from experiment (E2). Therefore, we must expect the computed 2D droplet shapes to deviate from (E1). These 2D droplet shapes can be found in Fig. 8 of [22], where they are visualized as three-dimensional results. Contrary to our simulation, the droplet shape at $t = 2\,\mathrm{ms}$ tends more to a pyramid shape and does not convincingly show the three layers obtained in the experiment. This might well be due to the lacking third dimension. Later, the two-dimensional results are comparable with our 3D results, but the width of the droplet at $t = 10$ and $15\,\mathrm{ms}$ is even larger than in our case. Here, we have to remember that the two parameters k_a and k_r in (C1) were used to fit Yokoi's 2D numerical results to the experiment (E2), and we did not adapt these parameters for our 3D simulation. Even so, the 2D and 3D simulations show remarkably good agreement with each other.

5.2 *The Contact Angle Model by Shikhmurzaev*

In this subsection, we present results for the droplet impact simulation with the reduced interface formation model (C2) combined with the global volume correction for the level-set method. We present two different variants of the model: On the one hand we take Moffatt's solution for the radial velocity (M1) and on the other hand we choose a far field velocity value to incorporate the influence of the flow field on the dynamic contact angle (M2). The parameter Sc is chosen to be 11.5 for (M1) and 5.5 for (M2). We set the values $\rho_G^s = 0.9$ and $\sigma_{sg} = 0$ according to [13].

The results of the laboratory experiments (E1) are always shown in the first row of Fig. 8. The droplet shapes computed by (M1) are given in the second row. As with (C1), the first three results are very close to the experimental snapshots, which is to be expected, since inertia dominates capillary effects. In addition, also at the later time steps, the computed droplet shapes agree very well with (E1). At $t = 10\,\mathrm{ms}$ the height and width of the droplet is reproduced very accurately. Further, at $t = 15\,\mathrm{ms}$, the droplet even forms a little dent before it meets the substrate and compares best with the experiment of all simulation results. A look at the droplet diameter evolution (Fig. 9) confirms that the interface formation model, although it is not specifically based on this experiment, produces nearly as good results as Yokoi's model: The maximum and final droplet diameter are captured, and the computed curve is very close to that of the experimental results.

In a next step, we compare the angle-speed curve of the interface formation model (M1) with Yokoi's 2D numerical results and the experiments. We expect that Yokoi's model, since it is fit to this experiment, produces dynamic contact angles which are very close to the ones measured in the experiment. From Shikhmurzaev's model we expect a smooth angle-speed relationship equal to our preliminary computation in Fig. 2. This is exactly what we see in Fig. 10: The contact angles computed by (C1) are nearly identical to the experimental values,

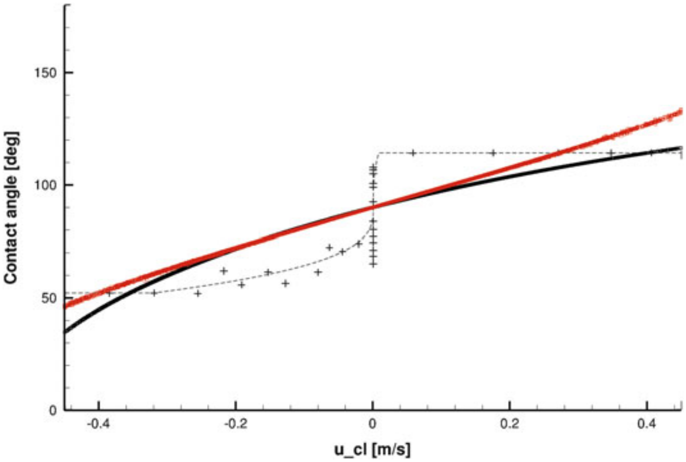


Fig. 10 Angle-speed relationship of a 3D simulation with the reduced interface formation model M1 (*black*) and M2 (*red*) compared to experiments (+) and 2D simulation with Yokoi's model (*dashed*), both taken from [22]

while the ones computed by (C2) develop in a smoother and nearly linear fashion but show a larger deviation from the experiment. For zero contact line velocity, the model predicts the equilibrium contact angle of 90°. Here, the values measured in the experiment are between 52° and 110°. This is due to the space-time resolution of the experiment, where the contact line velocity is considered to be zero, if the liquid interface does not cross any pixel. Interestingly, however, the smooth speed-angle curve predicted by the interface formation model, lies quite well in between the maximum advancing and minimum receding contact angle of the experiment.

The droplet shapes computed by (M2) are close to the results by (M1) (Fig. 8). Additionally, the droplet diameter varies only little between both models (Fig. 9). Furthermore, in Fig. 10, the angle-speed curve shows that both models compute very similar dynamic contact angles in specific regimes of the contact line velocity. For large contact line velocities, (M2) overshoots the maximum dynamic advancing angle determined from the laboratory experiment even more than (M1). However, for small contact line velocities, (M2) tends to be closer to the minimum dynamic receding angle than (M1). The curve for (M2) is scattered, since we use Moffatt's solution, if the contact angle cannot be evaluated directly from the arccos of Eq. (C2). Here, we observe that for similar contact line velocities both Moffatt's solution and an inserted far field velocity value give similar results with a difference of only a few degrees for the computed contact angle.

All in all, the results by Shikhmurzaev's reduced interface formation model are most promising. Although it is not based on this specific droplet experiment, the computed droplet shapes are very close to those observed in the experiment. Also, the evolving droplet diameter and speed-angle relationship support these very good results.

Table 2 Details of the mesh used for the mass convergence study (left) and grid convergence of initial mass towards the analytical solution

Level l	Δx_l	Δy_l	Δz_l	dof_l	m_l^0	ρ_l
1	2.206_{-4}	2.280_{-4}	2.206_{-4}	$31 \times 15 \times 31$	6.511_{-9}	–
2	1.121_{-4}	1.103_{-4}	1.121_{-4}	$61 \times 31 \times 61$	6.281_{-9}	2.020
3	0.565_{-4}	0.561_{-4}	0.565_{-4}	$121 \times 61 \times 121$	6.225_{-9}	1.965
4	0.284_{-4}	0.283_{-4}	0.284_{-4}	$241 \times 121 \times 241$	6.211_{-9}	1.985

5.3 Mass Conservation

In this subsection, we investigate the mass convergence behavior of our numerical schemes for the impact of water on hydrophobic siliane. First, we measure how well the analytical sphere volume can be initially approximated on our computational grids. Then, we investigate how much volume can be conserved with the pure level set method and the two volume correction methods at time $t = 30\,\text{ms}$, which corresponds to the last time-step in Fig. 6. Last, we study the convergence of the mass error. The details of the grids used for our convergence study are given in Table 2.

The analytical volume of the sphere, computed from the parameters in Table 1 is $V = \frac{4}{3}\pi r^3 = 6.2059_{-9}$. We evaluate the initial mass of the droplet as

$$m_l^0 = \sum_{\mathbf{x}_i} H(\varphi_l^0(\mathbf{x}_i))\Delta x_l \Delta y_l \Delta z_l \tag{33}$$

on grid levels $l = 1 \ldots 4$ as given in Table 2. Then, at time $t = 0$, the discrete error norm and convergence rate are evaluated as

$$e^l = \left| m^l - V \right| \text{ and } \rho^{l+1} = \frac{\log \frac{e^l}{e^{l+1}}}{\log 2}, \tag{34}$$

since there holds $2h_l \approx h_{l+1}$ for the discrete mesh width. The convergence results are given in Table 2. We clearly see that initial sphere volume shows second-order convergence towards the analytical value.

In a next step, we quantify the volume loss of our numerical schemes after 30 ms, which corresponds to the last droplet shapes displayed in Fig. 6. We expect that the local volume-correction will perform worse than the global volume correction at least on the coarser grids, since the local volume correction only prevents the straying of the level set function, but does not correct mass errors which occur due to the numerical diffusion introduced when solving the transport equation (10). The global volume correction, on the other hand, employs an absolute stopping criterion of $\varepsilon = 10^{-7}$ in the fixed point iteration and we anticipate to find a very small error in volume for all mesh sizes.

Table 3 Volume conservation in % at $t = 30\,\text{ms}$

Level l	Pure	Local	Global
1	0	64.4	100.0
2	39.5	91.5	100.0
3	62.8	96.4	100.0
4	78.9	99.3	100.0

Table 4 Table of mass convergence for the three level set-methods

	Pure level-set		Local correction		Global correction	
Level	e^l_{pure}	ρ^l_{pure}	e^l_{local}	ρ^l_{local}	e^l_{global}	ρ^l_{global}
1	1.587_{-1}	–	9.572_{-2}	–	3.453_{-8}	–
2	7.152_{-2}	1.150	2.954_{-2}	1.696	2.769_{-8}	0.318
3	2.633_{-2}	1.442	1.114_{-2}	1.408	4.342_{-8}	−0.649
4	8.971_{-3}	1.553	3.897_{-3}	1.515	6.412_{-8}	−0.562

Thus, we measure the volume

$$m_l^t = \sum_{\mathbf{x}_i} H(\varphi_l^t(\mathbf{x}_i))\Delta x_l \Delta y_l \Delta z_l \tag{35}$$

at $t = 30\,\text{ms}$ and compare it to the initial sphere volume at $t = 0$. In Table 3 the percentage of the still remaining volume is given. Mass conservation with the pure level set method is difficult for this particular case: On the coarsest grid, no mass is left after $t = 30\,\text{ms}$ and on the finest grid, we still loose about 20 % of mass. However, both the local and the global volume correction method tend to near 100 % mass conservation on the finest grid. As we expected, the global volume correction is able to conserve 100 % of mass on all meshes, while the local volume correction performs worse on coarser grids.

Last, we distinguish the effects of the pure level set method and the local and global volume correction on the overall convergence behavior in space and time at $t = 2\,\text{ms}$.

We compute the discrete error norm and convergence rate by

$$e^l = \frac{|m_l^t - m_l^0|}{|m_l^0|} \text{ and } \rho^{l+1} = \frac{\log \frac{e^l}{e^{l+1}}}{\log 2}, \tag{36}$$

since there holds $2h_l \approx h_{l+1}$ for the discrete mesh width. Our results are summarized in Table 4, where we see at least first order convergence for the pure level set method and the local volume correction in space and time. As expected, the discrete error for the global volume correction is constant $<10^{-7}$ due to the absolute stopping criterion. Therefore, we obtain a convergence rate $\rho^l_{\text{global}} \approx 0$.

6 Conclusion

In this paper we presented the numerical simulation of droplet impact with two different models for the dynamic contact angle in three space dimensions. First, we used Yokoi's model. The resulting droplet shapes were very close to those of the experiments and the previous two-dimensional results. Furthermore, we measured the droplet diameter over time, which confirmed the validity of the model and our numerical method. Here, we saw that both, the global and the local volume correction, are able to conserve the mass of the droplet, while still giving accurate results.

In a next step, we employed the reduced interface formation model by Shikhmurzaev for the droplet impact simulation, combined with the global volume correction. Here, we used Moffatt's solution for the radial velocity on the one hand, and a far field velocity value near the contact line but within the bulk flow on the other hand. The droplet shapes computed with this model were very close to each other and in good agreement with the experiment, as also confirmed by the computation of the droplet diameter over time.

Additionally, we compared the contact line speed-contact angle curve of Shikhmurzaev's model with the angle-speed curves of the experiments and Yokoi's model. As was to be expected, Shikhmurzaev's reduced model gives an approximate smoothed angle-speed relationship compared to the practical experiments and Yokoi's results. Thus, a future challenge might be to implement the full interface formation model without any restrictions for the capillary number in three dimensions. But still, the reduced model offers an excellent trade-off between the complex and costly full model and an easily implementable and accurate dynamic contact angle model, which is not restricted to a specific wetting experiment like Yokoi's model is.

In a last step, we compared the pure level set method with the two volume correction methods concerning their ability to conserve mass. On the finest grid, both the local and the global volume-correction were able to conserve about 100 % of the droplet's mass, while the pure level set method only retained about 80 %. We are currently implementing a coupled level set and volume-of-fluid method, which should further improve the mass conservation behavior of our flow solver.

Acknowledgements The authors acknowledge the support from the Sonderforschungsbereich 611 "Singular Phenomena and Scaling in Mathematical Models" of the Deutsche Forschungsgemeinschaft DFG. We are grateful to Kensuke Yokoi (Cardiff University) for providing the experimental and numerical results of the droplet impact and the permission to use the corresponding pictures in this article.

References

1. Blake, T.D., Bracke, M., Shikhmurzaev, Y.D.: Experimental evidence of nonlocal hydrodynamic influence on the dynamic contact angle. Phys. Fluids **11**(9), 1995–2007 (1999)
2. Brackbill, J.U., Kothe, D.B., Zemach, C.: A continuum method for modeling surface tension. J. Comput. Phys. **100**, 335–354 (1992)

3. Croce, R., Griebel, M., Schweitzer, M.A.: A parallel level-set approach for two-phase flow problems with surface tension in three space dimensions. Preprint 157, Sonderforschungsbereich 611, Universität Bonn (2004)
4. Croce, R., Griebel, M., Schweitzer, M.A.: Numerical simulation of bubble and droplet deformation by a level set approach with surface tension in three dimensions. Int. J. Numer. Methods Fluids **62**(9), 963–993 (2010)
5. Decent, S.P.: Hydrodynamic assist and the dynamic contact angle in the coalescence of liquid drops. IMA J. Appl. Math. **71**(5), 740–767 (2006)
6. Fang, C., Hidrovo, C., Wang, F., Eaton, J., Goodson, K.: 3-D numerical simulation of contact angle hysteresis for microscale two phase flow. Int. J. Multiph. Flow **34**, 690–705 (2008)
7. Goodwin, R., Homsy, G.M.: Viscous flow down a slope in the vicinity of a contact line. Phys. Fluids A **3**(4), 215–528 (1991)
8. Hocking, L.M.: A moving fluid interface. Part 2. The removal of the force singularity by a slip flow. J. Fluid Mech. **79**(2), 209–229 (1977)
9. Liu, J., Nguyen, N.T., Yap, Y.F.: Numerical studies of sessile droplet shape with moving contact lines. Micro Nanosyst. **3**(1), 56–64 (2011)
10. Moffatt, H.K.: Viscous and resistive eddies near a sharp corner. J. Fluid Mech. **18**(1), 1–18 (1964)
11. Monnier, J., Witomski, P.: Analysis of a local hydrodynamic model with Marangoni effect. J. Sci. Comput. **21**(3), 369–403 (2004)
12. Mukherjee, A., Kandlikar, S.G.: Numerical study of single bubbles with dynamic contact angle during nucleate pool boiling. Int. J. Heat Mass Transf. **50**, 127–138 (2007)
13. Shikhmurzaev, Y.D.: Capillary Flows with Forming Interfaces. Chapman & Hall/CRC, Boca Raton (2008)
14. Sibley, D.N., Savva, N., Kalliadasis, S.: Slip or not slip? A methodical examination of the interface formation model using two-dimensional droplet spreading on a horizontal planar substrate as a prototype system. Phys. Fluids **24**, 082105 (2012)
15. Somalinga, S., Bose, A.: Numerical investigation of boundary conditions for moving contact line problems. Phys. Fluids **12**(3), 499 (2000)
16. Spelt, P.D.M.: A level-set approach for simulations of flows with multiple moving contact lines with hysteresis. J. Comput. Phys. **207**, 389–404 (2005)
17. Sprittles, J.E., Shikhmurzaev, Y.D.: Finite element simulation of dynamic wetting flows as an interface formation process. J. Comput. Phys. **233**, 34–65 (2013)
18. Sussman, M.: An adaptive mesh algorithm for free surface flows in general geometries. In: Vande Wouwer, A., Saucez, P., Schiesser, W.E. (eds.) Adaptive Method of Lines, pp. 207–231. Chapman and Hall/CRC, Boca Raton (2001)
19. Sussman, M., Fatemi, E.: An efficient, interface-preserving level set redistancing algorithm and its application to interfacial incompressible fluid flow. SIAM J. Sci. Comput. **20**(4), 1165–1191 (1999)
20. Tanner, L.H.: The spreading of silicone oil drops on horizontal surfaces. J. Phys. D Appl. Phys. **12**(9), 1473 (1979)
21. van Mourik, S.: Numerical modelling of the dynamic contact angle. Master's thesis, University of Groningen (2002)
22. Yokoi, K., Vadillo, D., Hinch, J., Hutchings, I.: Numerical studies of the influence of the dynamic contact angle on a droplet impacting on a dry surface. Phys. Fluids **21**(7), 072102 (2009)
23. Yun-chao, S., Chun-hai, W., Zhi, N.: Study on wetting model with combined Level Set-VOF method when drop impact onto a dry surface. In: Electronic and Mechanical Engineering and Information Technology (EMEIT), Harbin, pp. 2583–2586 (2011)
24. Zahedi, S., Gustavsson, K., Kreiss, G.: A conservative level set method for contact line dynamics. J. Comput. Phys. **228**, 6361–6375 (2009)

A Parallel Multiscale Simulation Toolbox for Coupling Molecular Dynamics and Finite Elements

Dorian Krause, Konstantin Fackeldey, and Rolf Krause

Abstract It is the ultimate goal of concurrent multiscale methods to provide computational tools that allow to simulation physical processes with the accuracy of micro-scale and the computational speed of macro-scale models. As a matter of fact, the efficient and scalable implementation of concurrent multiscale methods on clusters and supercomputers is a complicated endeavor. In this article we present the parallel multiscale simulation tool MACI which has been designed for efficient coupling between molecular dynamics and finite element codes. We propose a specification for a thin yet versatile interface for the coupling of molecular dynamics and finite element codes in a modular fashion. Further we discuss the parallelization strategy pursued in MACI, in particular, focusing on the parallel assembly of transfer operators and their efficient execution.

1 Introduction

The goal of project C6 of the collaborative research center 611 "Singular Phenomena and Scaling in Mathematical Models" at the University of Bonn, Germany, was the development and implementation of novel computational techniques for the concurrent coupling of different physical models in the numerical simulation of solids. In particular, the project was concerned with multiscale coupling between atomistic and continuum models. Such concurrent multiscale approaches can be used, for example, in the numerical simulation of fracture processes. By using a

D. Krause (✉) · R. Krause
Institute of Computational Science, Università della Svizzera italiana, CH-6900 Lugano, Switzerland
e-mail: dorian.krause@usi.ch; rolf.krause@usi.ch

K. Fackeldey
Zuse Institute Berlin, D-14195 Berlin, Germany
e-mail: fackeldey@zib.de

M. Griebel (ed.), *Singular Phenomena and Scaling in Mathematical Models*,
DOI 10.1007/978-3-319-00786-1_14,

molecular dynamics model to capture the complicated physical processes in the vicinity of the crack tip and a computationally faster but less accurate continuum model for the surrounding material, one can achieve good accuracy at a lower price compared to fully atomistic simulations.

The design of efficient computational tools for such multiscale simulations is itself a challenging task. This is even more so when building parallel simulation tools. In this article we describe the design of the versatile multiscale simulation toolbox MACI and discuss the novel parallelization approach used in MACI. We introduce a thin yet capable interface designed for efficient coupling between molecular dynamics (MD) and finite elements (FE) codes.

1.1 Related Work

While the design of algorithms for concurrent multiscale coupling is an active field of research in the past years, relatively few work has been published about implementation and parallelization of these algorithms. Broughton et al. [8] report on a parallel multiscale simulation using the concurrent coupling of length scales method. This work is limited to one-dimensional domain decompositions for the molecular dynamics domain. Ma et al. [20] have implemented their MD/GIMP method in the SAMRAI framework. In comparison to most multiscale methods for the coupling of MD and finite elements their constraints are local. Xiao et al. [28] describe a parallel implementation of the Bridging Domain method in a grid environment. However, this work is restricted to one-dimensional simulations. Anciaux et al. [2] have implemented the Bridging Domain method in the parallel LIBMULTISCALE. Their approach is closest to our work.

In this article we present a versatile interface for coupling MD and FE codes. The common component architecture (CCA) [3] aims to develop a component model for high performance scientific computing. So far we are not aware of any work using CCA for multiscale coupling between atomistic and continuum models.

It is one of the goals of the European MAPPER project [21] to develop software and services for distributed multiscale simulations. While our work focuses on tightly-coupled simulations on clusters and supercomputers, this work is aimed towards the utilization of distributed resources in the European e-Infrastructure.

1.2 Article Contribution and Outline

The outline of the article is as follows. In Sect. 2 we review the multiscale simulation method implemented in MACI, focusing on the computational aspects. In Sect. 3 we propose and discuss a thin interface allowing for the reusing coupling logic with different molecular dynamics and finite element codes. This work is not limited to the coupling algorithm presented in Sect. 2 but can be applied to a broad range of multiscale coupling methods. In Sect. 4 we discuss the parallelization of MACI focusing on the description of the data and work distribution in the code. In comparison to our previous work [18] the focus of this section is the description of the high-level structure without a detailed discussion of the communication mechanisms employed.

2 Multiscale Simulation Method

In this section we shortly present atomistic (micro-scale) and continuum (macro-scale) models for the simulation of the behavior of a solid $\Omega \subset \mathbb{R}^3$. We then proceed to discuss an approach to concurrent coupling of these models using projection-based constraints.

2.1 Molecular Dynamics

On an atomistic level we can model Ω as a discrete set of N atoms/particles $\mathbb{A} = \{\alpha\}$ with coordinates and momenta $(\mathbf{x}_\alpha, \mathbf{p}_\alpha) \in \mathbb{R}^6$. The motion of these particles is governed by the Hamiltonian equations

$$\begin{aligned} \dot{\mathbf{x}}_\alpha &= \frac{\partial \mathscr{H}}{\partial \mathbf{p}_\alpha} = \frac{1}{m_\alpha}\mathbf{p}_\alpha \\ \dot{\mathbf{p}}_\alpha &= -\frac{\partial \mathscr{H}}{\partial \mathbf{x}_\alpha} = -\nabla_{\mathbf{x}_\alpha} V + \mathbf{F}_\alpha^{\text{ext}} \end{aligned} \tag{1}$$

with the Hamiltonian $\mathscr{H} = K + V$. Here, K denotes the kinetic energy of the atomic system $K = \sum_\alpha \frac{1}{2m_\alpha}\mathbf{p}_\alpha^2$, V is the interaction potential $V = V(\mathbf{x}_1, \dots, \mathbf{x}_N)$ and m_α the particle mass. In this article, we concentrate on short-ranged potential that allow for efficient (i.e., in linear time) computation of energy and forces using a linked cell method [16] or Verlet neighbor lists [1].

As a particle method, MD does not require discretization in space but only in time. Usually, lower order symplectic integrators (such as a second order Störmer-Verlet scheme) are used for their computational efficiency and good long-term stability properties.

2.2 Continuum Mechanics and Finite Elements

In continuum mechanics, the macroscopic deformation of a body $\Omega \subset \mathbb{R}^3$ is described by a volume preserving mapping $\varphi : [0, T] \times \Omega \rightarrow \mathbb{R}^3$, such that $\varphi(\{t\} \times \Omega)$ equals the configuration of the body at time t. The deformation field

$\mathbf{U} = \varphi - \mathbf{1}$ is the solution of the variational problem

$$\int_{\Omega} \rho \ddot{\mathbf{U}} \cdot \mathbf{V} \, \mathrm{d}\mathbf{x} = \int_{\Omega} \rho \mathbf{b} \cdot \mathbf{V} \, \mathrm{d}\mathbf{x} - \int_{\Omega} \underline{\mathbf{P}}(\mathbf{U}) : \nabla \mathbf{U} \, \mathrm{d}\mathbf{x} + \int_{\Gamma_N} \mathbf{f} \cdot \mathbf{V} \, \mathrm{d}S \;, \tag{2}$$
$$\mathbf{U} = \mathbf{U}_D \text{ on } \Gamma_D \quad + \text{ initial conditions for } \mathbf{U} \text{ and } \dot{\mathbf{U}} \;.$$

Here, ρ denotes the density in the undeformed configuration, $\mathbf{b}$ and $\mathbf{f}$ are external body and surface forces (the latter one applied on $\Gamma_N \subset \partial\Omega$) and $\underline{\mathbf{P}}$ denotes the first Piola-Kirchhoff tensor. Dirichlet values $\mathbf{U}_D$ are applied on $\Gamma_D \subset \partial\Omega$. The test function $\mathbf{V}$ is an element of an appropriate subspace of $C^0\big([0,T]; H^1(\Omega)\big)$.

In this article, we concern ourselves with first-order ($\mathbb{P}1$ or $\mathbb{Q}1$) finite elements for the spatial discretization of (2) resulting in a system of coupled partial differential equations

$$\ddot{\mathbf{U}}_A = M_A^{-1}\left(\mathbf{F}_A + \mathbf{F}_A^{\text{ext}}\right) \text{ for each mesh node } A \;,$$

which has the same structure as Eq. (1). Hence, the same temporal discretization methods can be applied.

2.3 *Coupling Method*

The goal of concurrent coupling schemes is to allow for interfacing highly accurate, but expensive, simulation techniques (such as MD) with less accurate, but faster, approximate schemes. For the latter we consider a continuum mechanics model discretized on a finite element mesh of a mesh size that is sufficiently larger than the average atomic distance. In the following we refer to this problem as *MD-FE coupling*.

The challenge in the design of concurrent coupling schemes is implementing appropriate transfer conditions between the *micro-* (MD) and *macro-* (FE) *scales*. Since each scale features a different resolution, not all modes (e.g., pressure waves of high wave number) can be resolved on all scales. The interface conditions need to account for this, in order not to create spurious effects (e.g., wave reflection) that spoil the solution accuracy.

In the following we review the coupling strategy using *projection-based constraints* described in [11, 12].

2.3.1 Coupling with Overlap

We consider an overlapping decomposition of the simulation domain Ω into an MD domain Ω_{MD} and an FE domain Ω_{FE} with *handshake region* $\Omega_{\text{H}} = \Omega_{\text{MD}} \cap \Omega_{\text{FE}}$.

In Ω_{H}, the micro- and macro-scale coexist. Inspired by the Bridging Domain method [29], volumetric constraints

$$\mathbf{0} \overset{!}{=} \mathbf{G}(\mathbf{u}, \mathbf{U}) = \mathbf{O}_1 \mathbf{u} - \mathbf{O}_2 \mathbf{U} \tag{3}$$

are used in [12] to couple the MD displacement field $\mathbf{u}$ and the FE displacement field $\mathbf{U}$. Here, the atomistic displacement field is given by $\mathbf{u}_\alpha(t) = \mathbf{x}_\alpha(t) - \mathbf{x}_\alpha(0)$.

In [12], the operators $\mathbf{O}_1$ and $\mathbf{O}_2$ are chosen to be equal to a projection Π from micro- to macro-scale and the identity $\mathbf{1}$, respectively. The projection Π allows for additively decomposing the micro-scale displacement field $\mathbf{u}$ into a macro-scale field $\bar{\mathbf{u}}$ and *high fluctuation* remainder $\mathbf{u}'$ (cf. [27]):

$$\mathbf{u} = \bar{\mathbf{u}} + \mathbf{u}' = \Pi \mathbf{u} + (\mathbf{1} - \Pi)\, \mathbf{u}\,.$$

Note that $\Pi \mathbf{u}' = \mathbf{0}$. Hence, the constraints $\mathbf{G}$ provide a pointwise coupling between $\mathbf{U}$ and $\bar{\mathbf{u}}$ while not affecting the high fluctuation field $\mathbf{u}'$ which is not representable on the macro-scale.

Inspired by non-conforming domain decomposition theory, in [10], an L^2 projection is proposed for micro-to-macro scale transfer. An embedding of the atomistic displacement space $\left(\mathbb{R}^3\right)^N$ into $L^2(\Omega)$ is constructed using scattered-data approximation methods. Hence, given a vector $(\mathbf{w}_\alpha)_{\alpha \in \mathbb{A}}$ a function $\mathbf{w}^\natural$ is constructed such that $\mathbf{w}^\natural(\mathbf{x}_\alpha(0)) \approx \mathbf{w}_\alpha$. One possible approach for constructing $\mathbf{w}^\natural \in \mathbb{X} \subset L^2(\Omega_{\mathrm{H}})$ is to introduce a set partition of unity basis functions ψ_α (see, for example, [13]) with $\sum_{\alpha \in \mathbb{A}} \psi_\alpha = 1$ and define

$$\mathbf{w}^\natural = \sum_{\alpha \in \mathbb{A}} \mathbf{w}_\alpha \psi_\alpha\,.$$

Given the embedding of $\left(\mathbb{R}^3\right)^N$ into L^2 we can define the projection $\Pi : \left(\mathbb{R}^3\right)^N \to \mathbb{S}_{\mathrm{H}}$ by

$$(\Pi \mathbf{u}, \mathbf{V})_\rho = \left(\mathbf{u}^\natural, \mathbf{V}\right)_\rho \quad \text{for all } \mathbf{V} \in \mathbb{S}_{\mathrm{H}}.$$

Here, $\mathbb{S}_{\mathrm{H}}$ denotes the first-order finite element space on Ω_{H} (we assume that Ω_{H} can be written as the union of a set of elements in the tessellation of Ω_{FE}) and $(-,-)_\rho$ equals the L^2 scalar product weighted by the continuum mass density ρ.

The assembly of the L^2 projection Π requires the computation of (and quadrature on) the cuts between the elements in the tessellation of Ω_{H} and the support of the basis functions ψ_α. Even though, this computation needs to be performed only as part of the simulation setup (i.e., not during the time integration), the assembly can be costly. Alternatively, a least-squares projection

$$\Pi \mathbf{u} = \operatorname{argmin}_{\mathbf{V} \in \mathbb{S}_{\mathrm{H}}} \frac{1}{2} \sum_{\alpha} m_{\alpha} |\mathbf{u}_{\alpha} - \mathbf{V}(\mathbf{x}_{\alpha}(0))|^2$$

has been discussed in [11].

Let us point out that in either case we can write $\Pi = \tilde{\mathbf{M}}^{-1}\mathbf{T}$ with a mass matrix $\tilde{\mathbf{M}}$ and a rectangular matrix $\mathbf{T}$ and hence we can equivalently use the constraints $\mathbf{G} = \mathbf{Tu} - \tilde{\mathbf{M}}\mathbf{U}$.

The coupled equations of motion for the micro- and macro-scale are derived from a weighted Hamiltonian/Lagrangian (cf., [12, 29]) resulting in a system of algebraic differential equations. We use a RATTLE integration scheme, requiring two linear solves per time step.

2.3.2 Damping High Fluctuation Modes

The design of the projection-based constraints $\mathbf{G}$ ensures that the high fluctuation field $\mathbf{u}'$ is not affected by the constraints, irrespective of the resolution of the finite element mesh. To avoid spurious reflections at $\partial\Omega_{\mathrm{MD}}$, a modified perfectly matched boundary layer (PML) method is proposed in [12] which (approximately) removes the high fluctuation field and has only a minor effect on the information transfer between micro- and macro-scale. To this end, an additional force term

$$\mathbf{f}_{\alpha}^{\mathrm{PML}} = -2d(\mathbf{x}_{\alpha}(0)) \Big((\mathbf{qv})_{\alpha} + d(\mathbf{x}_{\alpha}(0))(\mathbf{qu})_{\alpha} \Big)$$

is added to the MD forces $\mathbf{f}_{\alpha}$. Here, $d : \Omega_{\mathrm{MD}} \to [0, \infty)$ is a scalar function with support in $\overline{\Omega_{\mathrm{H}}}$ and $\mathbf{q} = \mathbf{1} - \mathbf{N}\Pi$, $\mathbf{N}$ being the interpolation operator from $\mathbb{S}_{\mathrm{H}} \to \left(\mathbb{R}^3\right)^N$.

2.3.3 Complete Algorithm

In Algorithm 1 the seven most important steps in the RATTLE integration from time t to time $t + \tau$ are explained. As mentioned earlier, two linear systems need to be solved in each timestep to compute the Lagrange multipliers λ and μ. We refer to [12] for the definition of the symmetric positive definite matrix Λ.

Two simulation results for a wave propagation benchmark and mode-I fracture computation using this concurrent coupling technique are shown in Figs. 1 and 2.

3 Multiscale Simulation Toolbox

The development of capable, efficient and scalable molecular dynamics or finite element codes is a complicated and labor-intensive task. Unless the scope of the application is limited, it is therefore often infeasible to develop a multiscale

Algorithm 1: RATTLE time integration scheme

1. Apply standard “Verlet kicks” and “Verlet drifts” to the micro-and macro-scale displacements and velocities yielding trial values $\mathbf{u}^*$, $\mathbf{v}^*$, $\mathbf{U}^*$, $\mathbf{V}^*$:

$$\begin{bmatrix}\mathbf{v}^*\\ \mathbf{V}^*\end{bmatrix} = \begin{bmatrix}\mathbf{v}^n\\ \mathbf{V}^n\end{bmatrix} + \frac{\tau}{2}\begin{bmatrix}\mathbf{m}^{-1}\mathbf{f}^{n+1}\\ \mathbf{M}^{-1}\mathbf{F}^{n+1}\end{bmatrix}, \quad \begin{bmatrix}\mathbf{u}^*\\ \mathbf{U}^*\end{bmatrix} = \begin{bmatrix}\mathbf{u}^n\\ \mathbf{U}^n\end{bmatrix} + \tau\begin{bmatrix}\mathbf{v}^*\\ \mathbf{V}^*\end{bmatrix},$$

   where $\mathbf{f}^n$, $\mathbf{F}^n$ denote the forces computed in step 4 of the previous time step.
2. Evaluate the displacement residual $\mathbf{G}^* = \mathbf{T}\mathbf{u}^* - \tilde{\mathbf{M}}\mathbf{U}^*$ and solve $\mathbf{G}^* = \Lambda\lambda$ for λ.
3. Correct the trial values

$$\begin{bmatrix}\mathbf{v}^{n+\frac{1}{2}}\\ \mathbf{V}^{n+\frac{1}{2}}\end{bmatrix} = \begin{bmatrix}\mathbf{v}^*\\ \mathbf{V}^*\end{bmatrix} + \frac{1}{\tau}\begin{bmatrix}\mathbf{m}^{-1}\mathbf{T}^{\mathrm{T}}\lambda\\ -\mathbf{M}^{-1}\tilde{\mathbf{M}}^{\mathrm{T}}\lambda\end{bmatrix}, \quad \begin{bmatrix}\mathbf{u}^{n+1}\\ \mathbf{U}^{n+1}\end{bmatrix} = \begin{bmatrix}\mathbf{u}^*\\ \mathbf{U}^*\end{bmatrix} + \begin{bmatrix}\mathbf{m}^{-1}\mathbf{T}^{\mathrm{T}}\lambda\\ -\mathbf{M}^{-1}\tilde{\mathbf{M}}^{\mathrm{T}}\lambda\end{bmatrix}.$$

4. Evaluate forces $\mathbf{f}^{n+1}$, $\mathbf{F}^{n+1}$ according to the Hamiltonian equation (without constraints).
5. Add the damping term $\mathbf{f}^{\mathrm{PML}}$ to the MD force $\mathbf{f}^{n+1}$.
6. Compute trial velocity values

$$\begin{bmatrix}\mathbf{v}^*\\ \mathbf{V}^*\end{bmatrix} = \begin{bmatrix}\mathbf{v}^{n+\frac{1}{2}}\\ \mathbf{V}^{n+\frac{1}{2}}\end{bmatrix} + \frac{\tau}{2}\begin{bmatrix}\mathbf{m}^{-1}\mathbf{f}^{n+1}\\ \mathbf{M}^{-1}\mathbf{F}^{n+1}\end{bmatrix}.$$

7. Evaluate the velocity residual $\dot{\mathbf{G}}^* = \mathbf{T}\mathbf{v}^* - \tilde{\mathbf{M}}\mathbf{V}^*$ and solve $\dot{\mathbf{G}}^* = \Lambda\mu$ for μ.
8. Correct the velocities

$$\begin{bmatrix}\mathbf{v}^{n+1}\\ \mathbf{V}^{n+1}\end{bmatrix} = \begin{bmatrix}\mathbf{v}^*\\ \mathbf{V}^*\end{bmatrix} + \begin{bmatrix}\mathbf{m}^{-1}\mathbf{T}^{\mathrm{T}}\mu\\ -\mathbf{M}^{-1}\tilde{\mathbf{M}}^{\mathrm{T}}\mu\end{bmatrix}.$$

simulation tool as a single monolithic code that implements MD and FE functionality along with the coupling logic. Instead we focus on reusing existing, established molecular dynamics and finite element implementations, such as TREMOLO [16], LAMMPS [19, 23] and UG [5].

In this article we are concerned with the design and efficient implementation of concurrent coupling codes for MD-FE coupling that allow for reuse of the coupling logic with different implementations of the molecular dynamics and finite element functionality. In comparison to the on-going research on the *common component architecture* [3], we are restricting ourselves to the scenario of concurrent MD-FE coupling and expose more details (e.g., about the data distribution) to the coupling code to simplify the development of efficient and scalable code. Additionally we impose some restrictions onto the MD and FE codes that we consider (see below). We have verified that our assumptions are fulfilled by the major molecular dynamics and finite element software packages that are discussed in the literature.

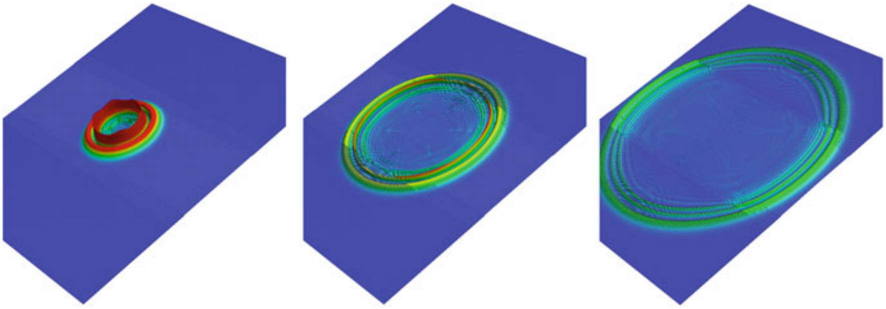


Fig. 1 Results of a two-dimensional wave propagation benchmark at the *beginning*, *middle* and *end* of the simulation. A radial wave propagates from Ω_{MD} into Ω_{FE} on the lower and upper side of the MD domain. The elongation in z-direction equals the (scaled) magnitude of the displacement field

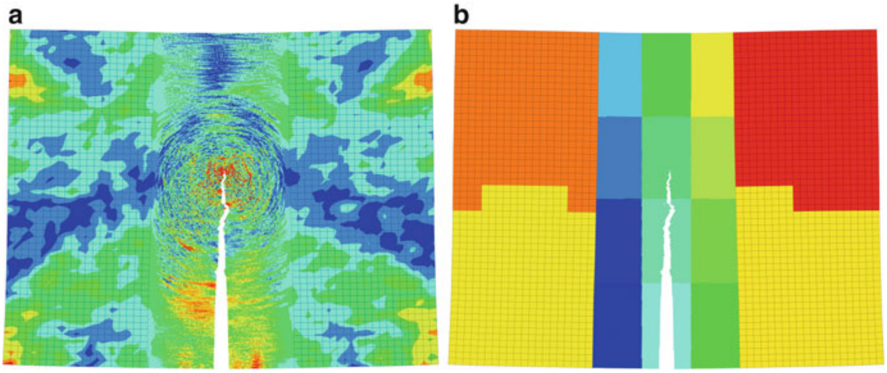


Fig. 2 Results of a mode-I fracture simulation using 2,496 finite elements and 62,390 atoms. Surface forces are applied on the left and right boundary of Ω_{FE}. (**a**) Velocity distribution. The velocities can be seen to fluctuate strongly in Ω_{MD} but to be smooth towards the boundary of the handshake region. (**b**) Distribution of atoms and finite elements over 12 + 4 processing elements

3.1 Interface Design

In this section we propose a simple yet versatile interface for coupling molecular dynamics and finite element simulations. Our work addresses modularity, performance and parallelization. We assume that the FE component is parallelized with a standard domain decomposition approach based on a partition of elements. We do not expose halo or ghost-cells through the presented interface. We moreover assume that the finite element mesh is statically balanced, i.e., that no dynamic load balancing (as used, for example, for adaptive mesh refinement) is performed. For the MD component we also assume a non-overlapping decomposition of the particles, i.e., each particle is stored on exactly one processing element. These assumptions are fulfilled by the majority of the molecular dynamics codes known to the authors, whether they use a domain decomposition (as do most codes such as TREMOLO, LAMMPS, NAMD [22] and DESMOND [7]) or particle-based decomposition (as, e.g., used by DDCMD [26]). Note that the (potentially) dynamic distribution of particles is deliberately exposed to the coupling code for parallel scalability considerations. For further discussions of the parallelization aspects we refer to Sect. 4.

Our approach is based on the following three pillars:

- Use of *opaque handles* for particles, nodes and elements to hide the details of the data layouts used by the MD and FE components.
- Use of *access epochs* to hide differences in data representation and allow to couple codes working in separate address spaces.
- The use of *piggybacking* to manage metadata in a simple and effective manner.

In the following we elaborate on theses three aspects of the coupling interface.

3.1.1 Opaque Handles

In general, the data layouts used by different MD or FE codes will depend strongly on the choice of the algorithms and the scope as well as the intended use case for the application. For example, the data layout used by a MD code based on a linked-cell method will be very different from the data layout used in a code that utilizes Verlet neighbor lists. Similarly, the data layout in a FE code will be different depending on whether the code supports dynamic (e.g., adaptively refined) or only static meshes. To hide these differences, the proposed interface provides opaque handles for the local particles, nodes and elements on a processing element in the form of iterators **`ParticleHandle`**, **`NodeHandle`** and **`ElementHandle`**. These iterators implement increment, comparison and assignment operators.

Since the abstraction of the data layout necessarily incurs a performance penalty, these iterators are intended for use in gather/scatter operations that copy the component data from or to a buffer in a layout suited for the coupling code. Each iterator provides a **`GetLocalId()`** function that can be used to address a contiguous buffer. Moreover, to have a unique local identifier for all mesh nodes in $\overline{\Omega_{\mathrm{H}}}$ we provide **`GetUserChosenId()`** and **`SetUserChosenId()`** functions for **`NodeHandle`** that allow to assign arbitrary (local or global) indices to the mesh nodes (for **`ParticleHandle`** this index can be stored as part of the **`PiggybackType`**, see below). Access to the particle data and dynamic variables (**`ParticleMass`**, **`ParticlePos`**, **`ParticleDispl`**, **`ParticleVel`**, **`ParticleForce`**, **`FeDispl`**, **`FeVel`**, **`FeLumpedMass`**) is possible through static functions taking a **`ParticleHandle`** or **`NodeHandle`** instance as an argument. Note that these functions are application specific (in this case targeted to coupled simulations of solids) but the approach can be generalized easily. Since the data distribution of the MD component can change dynamically in a parallel simulation, the life time of a **`ParticleHandle`** should be limited to the scope of the coupling routine that created it.

To allow for parallelization of the coupling code we also expose node ownership information (for nodes shared by multiple processing elements) and we provide **`ParallelSumup`**, **`ParallelMax`** and **`ParallelCopy`** routines to compute the sum (or max) of values stored at duplicated mesh nodes. These functions can be more efficiently implemented by taking advantage of the communication primitives of the FE component.

3.1.2 Access Epochs

MD and FE components do not always work with compatible data representations or with the same reference frame. For example, some MD codes rescale the simulation domain to the unit cube $[0, 1]^3$. Hence, all coordinates, velocities and forces need to be scaled before being accessed by the coupling code. Similarly, any updated particle position needs to be rescaled. Moreover, updating the particle positions might require a subsequent exchange of particles that have crossed subdomain boundaries (if the MD component uses a domain decomposition approach).

To cope with these difficulties we propose the use of *access epochs*, which work similar to RMA epochs in the MPI standard [14]. The coupling interface provides subroutines **`AccessBegin(int)`**, **`AccessEnd()`** and **`CanAccess()`**. A call to **`AccessBegin`** starts an access epoch. The bit field passed to **`AccessBegin`** specifies which data fields can be accessed in read, write or read-write mode during the epoch. Access to any data field (via the functions **`ParticlePos`**, **`FeDispl`**, etc.) outside of an access epoch is illegal. An access epoch ends with a call to **`AccessEnd`**. The function **`CanAccess`** allows to check whether access is permitted (in particular for debugging). For example, in Algorithm 1, the third step would be wrapped by calls to **`AccessBegin(VEL_RD | VEL_WR | DISPL_RD | DISPL_WR)`** and **`AccessEnd()`** (note that "|" is a bit-wise or operation in C allowing to build bitfields from, e.g., **`enum`** variables). Providing detailed information about read and write accesses to the state variables allows the interface code to optimize the actions performed in **`AccessEnd()`**. While it is likely that in a parallel MD code, **`AccessEnd`** needs to trigger an exchange of particles between processing elements after step three, this usually is not required after step six, since in this step only velocities are modified. The calls to **`AccessBegin`** and **`AccessEnd`** are collective, i.e., all MD or FE processing elements need to call these functions in order to achieve progress. The rationale for this decision is that **`AccessEnd`** might require exchange of particles and hence (global) communication.

Beyond a transparent handling of the differences in data representation between MD and FE components, this epoch-based design also permits for coupling codes storing data in a different address space than the one of the coupling code. For example we have successfully coupled MACI with a CUDA MD code. In this case, the coupling code ran on the CPU while the particle data resided in the graphics card memory. In **`AccessBegin`** and **`AccessEnd`** data is copied between CPU and GPU memory. The use of asynchronous copies is possible but would require modifications of the MD code to ensure that the MD code blocks for the completion of the host-to-device copy started in `AccessEnd` at the appropriate time.

3.1.3 Piggybacking of Metadata

For the efficient implementation of a concurrent coupling scheme such as Algorithm 1, a set of states need to be maintained for each particle. For example, each particle with $\mathbf{x}_\alpha(0) \in \overline{\Omega_{\mathrm{H}}}$ is assigned a local index and needs to store a weight $w \in (0, \infty)$ as well as the value $d\,(\mathbf{x}_\alpha(0))$. Depending on the algorithm and use case, the amount of data and its structure can vary. In order not to impact the scalability of the coupling code this data should be migrated together with the particles. Hence, it appears impracticable to leave the management of it to the coupling code since particle migration is managed by the MD component. Here, we *piggyback* this data onto the particles and use the communication subroutines of the MD code to exchange it along with the other state of the particle (positions, velocities, etc.). This might require modification of the MD code, for example, to add a `PiggybackType` to the `Particle` structure and to ensure that the additional data is communicated correctly. We have done these modifications in a copy of the TREMOLO code in less than 50 lines of code (mainly to add serialization and de-serialization of `PiggybackType`). For other codes, such as LAMMPS even less modifications may be required since serialization and de-serialization routines can be easily added by defining a new `AtomVec` class.

Note that we do not provide a `PiggybackType` for the FE component because we restricted ourselves to statically balanced meshes. However, the same piggyback technique can used in dynamically balanced finite element simulations.

3.2 Description of the MACI *Code*

We have implemented a new concurrent coupling code MACI (**M**ultiscale **a**tomistic **c**ontinuum **i**nterface) based on the interface defined in the previous section. In this section we shortly describe the architecture of MACI, as depicted in Fig. 3.

MACI is written in C/C++ for efficiency and portability. Since C, C++ and Fortran are the predominant programming language in high performance computing, this choice allows us to interface to most MD and FE codes without the need for additional language translation (for example, via BABEL [25]). MACI is scriptable using the Python programming language. The translation from C++ to Python is performed using the SWIG tool [6]. It is worth pointing out that while we believe that scripting capabilities are of great advantage for complicated scientific applications like MACI, the use of Python in this project had some inevitable impact on portability (for example, onto earlier Cray massively parallel systems) and complexity (in particular the handling of dynamically shared objects without circular references).

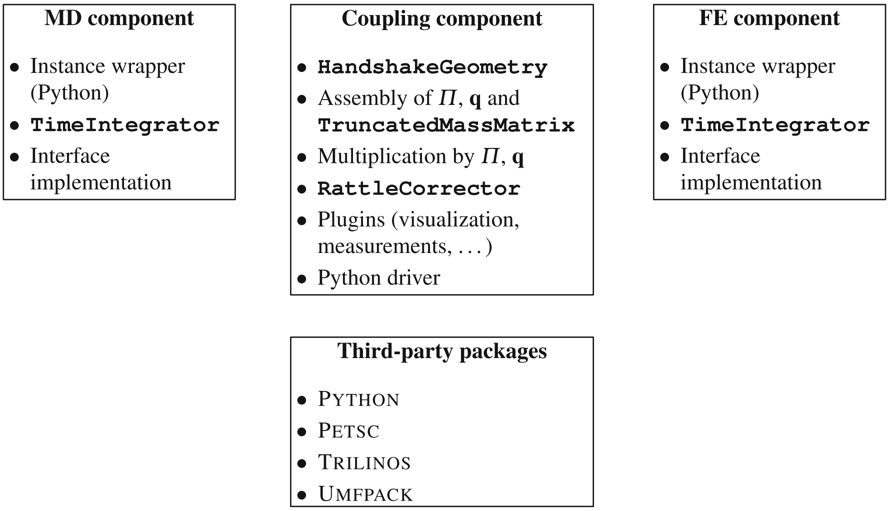


Fig. 3 Overview of the architecture of the MACI multiscale simulation tool

MACI consists of three major components (MD component, FE component, coupling component). The code is designed to run in an SPMD (single program, multiple data) fashion. Each processing element runs the coupling code along with either MD or FE code. Hence, MACI needs to run with at least two processing elements. Each component performs communication using different MPI communicators, effectively shielding the MD and FE code from mutual interference.

The coupling code implements functionality for managing the handshake geometry (to allow, for example, to find all particles with $\mathbf{x}_\alpha(0) \in \overline{\Omega_{\mathrm{H}}}$), for the assembly of the transfer operator Π, the high fluctuation filter $\mathbf{q}$ and Λ; for computing the Lagrange multipliers λ, μ (cf. Algorithm 1) and the corresponding Lagrange accelerations as well as for the computation of $\mathbf{f}^{\mathrm{PML}}$. To solve the linear systems arising in the RATTLE integration scheme, MACI can use iterative solvers from the PETSC [4] and TRILINOS [17] packages as well as the direct solver packages UMFPACK [9] (if the handshake region $\overline{\Omega_{\mathrm{H}}}$ is not distributed over multiple FE processing elements).

4 Parallelization Aspects

Molecular dynamics and finite element workloads each are well parallelizable and highly scalable implementations do exist. To allow for the treatment of interesting problem sizes using concurrent multiscale methods, their parallelization is of high interest. Unfortunately, the coupling of two scalable codes is *not* readily scalable. In fact, the parallelization of the coupled code introduces several challenges related to the data and work distribution and load balancing. In this section we describe how these challenges are approached in MACI.

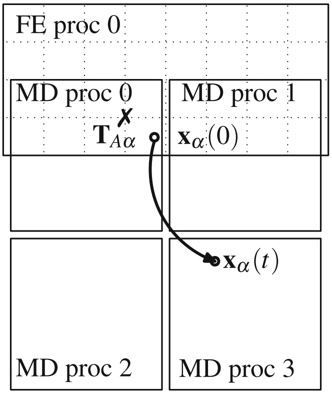


Fig. 4 The challenge of dynamic particle distribution in parallel concurrent MD-FE coupling. Particle migration introduces new edges in the communication graph

4.1 Challenges

Finite elements codes are usually parallelized using an element-wise partitioning of the computational mesh (computed, for example, via graph partitioning algorithms). As mentioned earlier, we restrict ourselves to statically balanced meshes in which this *domain decomposition* is kept fixed over the course of the (time-dependent) simulation.

In contrast to this fixed partition, molecular dynamics codes that support short-ranged interactions usually feature dynamically balanced load since the pair interaction lists (i.e., the set of tuples (α, β) of particles that interact with each others) depends on the current particle positions. Hence, to achieve maximum locality in the expensive force evaluation, particles are migrated between processing elements. One common scheme (found, e.g., in LAMMPS, TREMOLO, NAMD and DESMOND) is to statically decompose the simulation box $B = \bigcup_p B_p$ into subdomains B_p (one for each processing element) and to assign particles to processing element p if $\mathbf{x}_\alpha(t) \in B_p$. Hence, if a particle crosses a subdomain boundary it is assigned to a different processing element.

In the context of our concurrent coupling strategy the dynamic data distribution of particles is a challenge since our displacement-based constraints (3) are *non-local*. In fact, we have $\mathbf{T}_{A\alpha} \neq 0$ if and only if $\operatorname{meas}(\operatorname{supp}\psi_\alpha \cap \operatorname{supp}\theta_A) > 0$, where θ_A is the nodal basis function with $\theta_A(\mathbf{x}_A) = 1$. Since $\operatorname{supp}\psi_\alpha$ is a polygon or sphere centered at the initial particle position $\mathbf{x}_\alpha(0)$ we can have $\mathbf{T}_{A\alpha} \neq 0$ even if $|\mathbf{x}_\alpha(t) - \mathbf{x}_A|$ is very large, cf. Fig. 4. This implies that the *communication graph* (i.e., the graph with processing elements as nodes and edges between pairs of nodes that exchange messages) is dynamically changing. Thus a scalable implementation of the multiplication by the matrices $\mathbf{T}$ (the *scale transfer*) and $\mathbf{q}$ is much complicated compared to "classical" parallel sparse linear algebra, cf. [18].

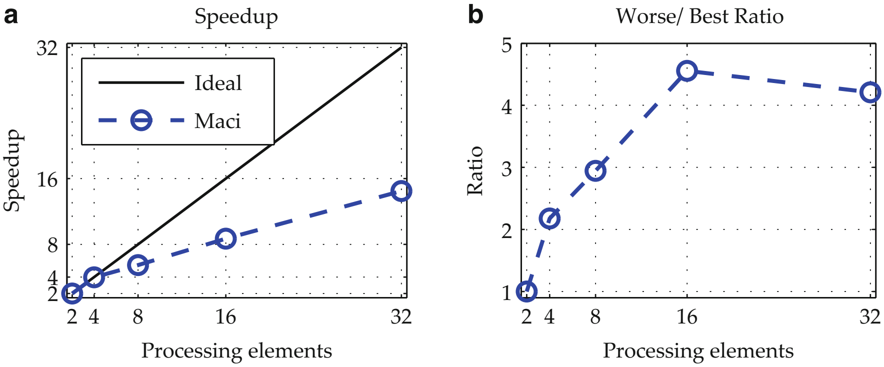


Fig. 5 Speedup plots from a two-dimensional fracture simulation with 142,628 atoms and 28,178 finite elements on up to 32 cores. The simulation was run on an 4× DDR Infiniband cluster with dual-socket quad-core Barcelona Opteron nodes. (**a**) Speedups (for the optimal choice of MD and FE processing element). (**b**) Ratio between time per time step in the worst and best configuration

Additionally, parallel concurrent coupling introduces novel challenges for load balancing. Much research has been devoted to devising and implementing good load balancing schemes for MD and FE algorithms (and hence for Steps 1, 3, 4, 6 and 8 in Algorithm 1). However, Steps 2, 5 and 7 introduce additional load on a subset of processing elements. For example, the matrix Λ is of size $L \times L$ where L is the number of mesh nodes in $\overline{\Omega_{\mathrm{H}}}$. In practice L is much smaller than the total number of mesh nodes or particles and hence the (iterative) solver for $\Lambda\lambda = \mathbf{G}^*$ usually does not scale well enough to distribute this task over all processing elements. Instead only a subset (e.g., all the FE processing elements that own cells intersecting the handshake region) will be responsible for solving the linear system. This introduces a strong load imbalance. Even worse, the synchronous nature of the RATTLE integrator does not permit the other processing elements to overlap the wait time with other computations (since the Lagrange forces need to be available before the algorithm can proceed), resulting in unwanted idle time.

In this article we concentrate on the first challenge. At this point MACI does not provide functionality to optimize the load balancing. This is a strong limiting factor for the parallel efficiency (cf. Fig. 5a). As can be seen in Fig. 5b, for a fixed number of processing elements, the choice of the number of MD and the number of FE processing elements is crucial for the performance even on a moderate number of cores. Here, a priori load models need to be developed to assist users in finding an optimal configuration.

4.2 Data Distribution in the MACI *Code*

MACI is written in an MPMD (multiple programm, multiple data) fashion (even though it is implemented as a single executable), i.e., processing elements that run the MD component (plus coupling code) take a (mostly) disjoint execution path compared to the execution path of the FE processing elements. Any exchange of data between MD and FE processing elements is done via message passing. To decrease communication cost it might be advantages to use threading and let one MD and one FE processing element share one address space. We refrained from this design in MACI since it complicates the coupling code (which in this case must be able to cope with one FE and one MD component) and requires good a priori knowledge about the communication graph including the communication volume per edge. Since the graph is dynamic, it usually is not feasible to do an optimal process mapping statically.

An example of the data distribution used by MACI is shown in Fig. 2b. In this simulation, the MD domain is distributed over 12 processing elements. The FE mesh is distributed over four processing elements. All the datastructures (including the $\mathbf{T}$, $\mathbf{q}$ and Λ matrices) are distributed over the four FE processing elements and the eight MD processing element that own mesh nodes with $\mathbf{x}_A \in \overline{\Omega_\mathrm{H}}$ or owned (at $t = 0$) particles in the handshake region.

The matrices $\mathbf{T}$, $\mathbf{q}$ and Λ are distributed by row. For $\mathbf{T}$ and Λ the Ath row is stored on the processing element that owns the node A (note that the cellwise mesh decomposition results in the duplication of mesh nodes on several processing elements). Also for the matrix $\mathbf{q}$ we use a static distribution: The αth row is stored on the processing element p with $\mathbf{x}_\alpha(0) \in B_p$. This static decomposition of $\mathbf{q}$ implies that the matrix-vector multiplication $\mathbf{y} = \mathbf{q}\mathbf{x}$ requires two communication steps: one gather operation to collect $\mathbf{x}$ values on the processing elements storing rows of the matrix and a second scatter operation, after the local matrix vector multiplication, to send the entry $\mathbf{y}_\alpha$ to the current owner of particle α. On the other hand a dynamic distribution of $\mathbf{q}$ (where the αth row of $\mathbf{q}$ is stored on the processing element owning the particle α) would require MD processing elements to be informed about the particle distribution on other processing elements. In particular if $\mathbf{q}_{\alpha\beta} \neq 0$ and $\beta \in B_q$, the processing element q would need to know on which processing element particle α is located. Maintaining such a mapping from particles to processing elements in a scalable manner is itself very complicated.

The sets $\{A \mid \mathbf{T}_{A\alpha} \neq 0\}$, $\left\{\beta \mid \mathbf{q}_{\beta\alpha} \neq 0\right\}$ are stored as part of the `PiggybackType` structure of the particle α and hence migrate with the particle. This allows MD processing elements to create lists of data item that have to be send to either FE processing elements (multiplication by $\mathbf{T}$) or handshake MD processing elements (multiplication by $\mathbf{q}$). Note however that this information is only available on the sender side. To the receiver side, the particle data distribution is unknown, cf. Sect. 4.4.

4.3 Parallel Assembly

In order to avoid bottlenecks in the computational workflow it is crucial to parallelize the complete application, in particular, the assembly of the transfer

Algorithm 2: Parallel assembly of **T**

```
Exchange tree root bounding boxes between MD and FE processing elements.
On FE processing elements: Build lists of elements $E$ intersecting the bounding boxes of
    the MD processing elements.
Send element lists from FE processing elements to MD processing elements.
forall the elements $E$ received (only on handshake MD processing elements) do
    Query (locally) all particles $\alpha$ with meas (supp $\psi_\alpha \cap E$) $> 0$.
    forall the found $\alpha$ do
        forall the $Q$ as in Equation (4) do
            Compute intersection $Q \cap E$.
            Perform quadrature on the intersection.
Send computed values back to the FE processing element.
On FE processing elements: Merge the received lists and construct sparse matrix
    datastructure.
```

operator **T** and of the interpolation matrix **N**. In MACI, the matrices **q** and Λ are computed from **T**, $\tilde{\mathbf{M}}$ and **N** using sparse matrix-matrix multiplications. Compared to direct assembly of the matrix entries (used initially in MACI), this approach is slower but the additional flexibility allows us to easily implement different transfer operators.

For a given mesh node A, the assembly of **T** and **N** requires the identification of all particles $\alpha \in \mathbb{A}$ such that either meas $(\operatorname{supp}\psi_\alpha \cap \operatorname{supp}\theta_A) > 0$ or $\theta_A(\mathbf{x}_\alpha) \neq 0$. To find all α in (quasi-)optimal time we use (parallel) tree queries.

In a first step, a parallel quad- or octree is build with particles as leaves. The bounding boxes of intermediate nodes are chosen in a leaf-to-root pass such that the bounding box of a parent node contains the union of the bounding boxes of all children. On the leaf level, the bounding box is chosen to be the support of ψ_α. Note that, different from [24], we do not use the tree to construct $\omega_\alpha = \operatorname{supp}\psi_\alpha$, since we require an algorithm producing open covers without adding additional points to the set of particles. Instead we make use of the lattice structure of the MD system to choose the patch size **h** a priori in such a way that

$$\omega_\alpha = \prod_{i=1}^{3} \left((\mathbf{x}_\alpha)_i - \frac{1}{2}\mathbf{h}_i, (\mathbf{x}_\alpha)_i + \frac{1}{2}\mathbf{h}_i \right)$$

yields an open covering of Ω_{H}. When using Shepard's method, the evaluation of the basis function ψ_α requires the detection of overlapping patches. This can be achieved by using again a tree. In MACI, the rectangular domain decomposition used by most MD applications is employed to compute a priori potential remote intersection partners (this is possible since we know the patch size **h**). These lists are exchanged and inserted into the local tree. Once this extended local tree is constructed, all intersection queries can be performed locally. The assembly of the transfer operator **T** is performed on the MD processing elements as shown in Algorithm 2.

Note that basis functions ψ_α computed with Shepard's method are only C^0. To achieve good accuracy in the computation of $\mathbf{T}$ via numerical quadrature, we need to compute a partition

$$\omega_\alpha = \bigcup_Q Q \quad \text{such that} \quad \psi_\alpha|_Q \in C^\infty. \tag{4}$$

This is possible (though not inexpensive) since our patches ω_α are axis parallel quadrilaterals or hexahedra.

4.4 Runtime Support

In this section we consider the implementation of Steps 2, 5 and 7 in Algorithm 1 in MACI. As noted in [18], the computation of the Lagrange forces $\mathbf{T}^{\mathrm{T}}\,\Lambda^{-1}\mathbf{G}^*$ and $\mathbf{T}^{\mathrm{T}}\,\Lambda^{-1}\dot{\mathbf{G}}^*$ can be handled in the same way as the multiplication by $\mathbf{q}$. These operations can be written as

$$\mathtt{Scat}_t \circ \mathtt{Op} \circ \mathtt{Gat}_t \ , \tag{5}$$

where $\mathtt{Scat}_t$ and $\mathtt{Gat}_t$ are time-dependent (since the particle distribution is changing over time) scatter and gather operations and where $\mathtt{Op}$ is some (black box) operation executed on a subset of processing elements (the *workers*). For example, in Step 2 of Algorithm 1, the black box operation is given by

$$\mathtt{Op}(\mathbf{z}) = \mathbf{T}^{\mathrm{T}}\,\Lambda^{-1}\left(\mathbf{Tz} - \tilde{\mathbf{M}}\mathbf{U}^*\right).$$

Finite element processing elements owning handshake mesh nodes (or equivalently, a non-zero row of $\mathbf{T}$) are designated as workers. The computation of the Lagrange multipliers required for the correction of the FE displacement is a *side effect* of the execution of $\mathtt{Op}$ on the worker processing elements.

Note that the choice of the workers and the order of the input and output data to and from $\mathtt{Op}$ is not time-dependent. Hence, $\mathtt{Gat}_t$ and $\mathtt{Gat}_{t'}$ $(t \neq t')$ are required to order the input data for $\mathtt{Op}$ in the same way. Similarly, $\mathtt{Scat}_t$ receives the output of $\mathtt{Op}$ always in the same order.

The advantage of this approach is that the data distribution is *transparent* to the workers. In particular no adaptation of the worker data structures are required when the particle distribution changes (this is to compare with the approach in [2] where the worker datastructures are updated using an event-based notification system).

The implementation of the gather and scatter operations are based on the piggybacked metadata. As noted in Sect. 4.2, the piggyback data allows MD processing elements to build send lists. However, workers/receivers do not know about the particle distribution and therefore cannot post matching receives. To cope

Algorithm 3: Implementation of step two in Algorithm 1

```
1. Pack displacements into contiguous buffer.
2. On MD processing elements: Extract list of workers and corresponding local indices for
   each particle.
3. On FE processing elements: Compute $\tilde{\mathbf{M}}\mathbf{U}^*$.
4. Communicate MD displacements to worker (Mexico).
5. On worker: Compute $\mathbf{G}^*$ using the input buffer containing MD displacements.
6. On worker: Solve $\Lambda\lambda = \mathbf{G}^*$.
7. On worker: Compute $\mathbf{T}^{\mathrm{T}}\lambda$ and store the result in the output buffer.
8. Communicate Lagrange forces to MD processing elements (Mexico).
9. Correct displacements and velocities.
```

with this problem, in [18] the use of *one-sided communication* or *remote memory access* is proposed.

In Maci we use the newly developed communication library Mexico [18] to implement the operation (5). The unique feature of Mexico is that the library provides gather and scatter operations in the described asymmetric setup. All information required by Mexico is provided by the source processing elements (processing elements that provide data to $\mathtt{Gat}_t$) and target processing elements (processing elements that retrieve data from $\mathtt{Scat}_t$). In Maci this information (the list of worker processing elements including local indices in the input and output buffers of $\mathtt{Op}$) are stored in the piggyback data. Mexico can use MPI RMA, MPI Point-to-point, MPI collectives, the Global Arrays library [15] or SHMEM for inter-process communication.

In Algorithm 3, the implementation of the second step in the RATTLE integration scheme (cf. Algorithm 1) is shown. The computational work is performed in Steps 5, 6 and 7. The input and output buffers for these operations are ordered according to an a priori (during the assembly phase) chosen ordering. Hence, the sparse-matrix storage scheme for, e.g., $\mathbf{T}$, can be kept unmodified over time.

Acknowledgements This work was supported by the Deutsche Forschungsgemeinschaft through the SFB 611 “Singular Phenomena and Scaling in Mathematical Models” and by the “Swiss High Performance and Productivity Computing” (HP2C) initiative.

References

1. Allen, M.P., Tildesley, D.J.: Computer Simulation of Liquids. Oxford Science, Oxford (1987)
2. Anciaux, G., Coulaud, O., Roman, J.: High performance multiscale simulation or crack propagation. In: Proceedings of the 2006 International Conference Workshops on Parallel Processing, Columbus, pp. 473–480. IEEE Computer Society (2006)
3. Armstrong, R., Gannon, D., Geist, A., Keahey, K., Kohn, S., McInnes, L., Parker, S., Smolinski, B.: Toward a common component architecture for high-performance scientific computing. In: Proceedings of the Eighth International Symposium on High Performance Distributed Computing, Redondo Beach, pp. 115–124 (1999)

4. Balay, S., Brown, J., Buschelman, K., Gropp, W., Kaushik, D., Knepley, M., McInnes, L., Smith, B., Zhang, H.: Petsc web page. http://www.msc.anl.gov/petsc (2012)
5. Bastian, P., Birken, K., Johannsen, K., Lang, S., Neuss, N., Rentz-Reichert, H., Wieners, C.: UG – a flexible software toolbox for solving partial differential equations. Comput. Vis. Sci. **1**, 27–40 (1997)
6. Beazley, D.M.: Swig: an easy to use tool for integrating scripting languages with C and C++. In: Proceedings of the 4th Conference on USENIX Tcl/Tk Workshop, USENIX Association, TCLTK'96, Berkeley, vol. 4, pp. 15–15 (1996)
7. Bowers, K.J., Chow, E., Xu, H., Dror, R.O., Eastwood, M.P., et al.: Scalable algorithms for molecular dynamics simulations on commodity clusters. In: Proceedings of the ACM/IEEE Conference on Supercomputing (SC06), Tampa, pp. 84–88 (2006)
8. Broughton, J.Q., Abraham, F.F., Bernstein, N., Kaxiras, E.: Concurrent coupling of length scales: methodology and application. Phys. Rev. B **60**, 2391–2403 (1999)
9. Davis, T.A.: Algorithm 832: UMFPACK, an unsymmetric-pattern multifrontal method. ACM Trans. Math. Softw. **30**, 196–199 (2004)
10. Fackeldey, K., Krause, R.: Multiscale coupling in function space – weak coupling between molecular dynamics and continuum mechanics. Int. J. Numer. Method Eng. **79**(12), 1517–1535 (2009)
11. Fackeldey, K., Krause, D., Krause, R.: Numerical validation of a constraints-based multiscale method for solids. In: Meshfree Methods for Partial Differential Equations V. Lecture Notes in Computational Science and Engineering, vol. 79, pp. 141–154. Springer, Berlin (2011)
12. Fackeldey, K., Krause, D., Krause, R., Lenzen, C.: Coupling molecular dynamics and continua with weak constraints. Multiscale Model. Simul. **9**(4), 1459–1494 (2011)
13. Fasshauer, G.E.: Meshfree Approximation Methods with Matlab. World Scientific, Singapore (2007)
14. Forum MPI: MPI: a message-passing interface standard, version 2.2 (2009)
15. Global Array: http://www.emsl.pnl.gov/docs/global (2012)
16. Griebel, M., Knapek, S., Zumbusch, G.: Numerical Simulation in Molecular Dynamics. Springer, Berlin/Heidelberg (2007)
17. Heroux, M.A., Bartlett, R.A., Howle, V.E., Hoekstra, R.J., et al.: An overview of the Trilinos project. ACM Trans. Math. Softw. **31**(3), 397–423
18. Krause, D., Krause, R.: Parallel scale-transfer in multiscale MD-FE coupling using remote memory access. In: Workshop Proceedings of the IEEE 7th International Conference on E-Science, e-Science 2011, Stockholm, pp. 66–73, 5–8 Dec 2011
19. Lammps: http://lammps.sandia.gov (2012)
20. Ma, J., Lu, H., Wang, B., Hornung, R., Wissink, A., Komanduri, R.: Multiscale simulations using generalized interpolation material point (GIMP) method and molecular dynamics (MD). Comput. Model. Eng. Sci. **14**, 101–118 (2006)
21. Mapper Project: http://www.mapper-project.eu (2012)
22. Phillips, J.C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R.D., Kalé, L., Schulten, K.: Scalable molecular dynamics with NAMD. J. Comput. Chem. **26**(16), 1781–1802 (2005)
23. Plimpton, S.J.: Fast parallel algorithms for short-range molecular dynamics. J. Comput. Phys. **117**, 1–19 (1995)
24. Schweitzer, M.A.: A Parallel Multilevel Partition of Unity Method for Elliptic Partial Differential Equations. Lecture Notes in Computational Science and Engineering, vol. 29. Springer, Berlin (2003)
25. Smolinski, B.A., Kohn, S.R., Elliott, N., Dykman, N.: Language interoperability for high-performance parallel scientific components. In: Proceedings of the Third International Symposium on Computing in Object-Oriented Parallel Environments, ISCOPE'99, San Francisco, pp 61–71. Springer (1999)
26. Streitz, F.H., Glosli, J.N., Patel, M.V., Chan, B., Yates, R.K., et al.: 100+ TFlop solidification simulations on BlueGene/L. In: Proceedings of the ACM/IEEE Conference on Supercomputing (SC05), Seattle (2005)

27. Wagner, G.J., Liu, W.K.: Coupling of atomistic and continuum simulations using a bridging scale decomposition. J. Comput. Phys. **190**, 249–274 (2003)
28. Xiao, S., Ni, J., Wang, S.: The bridging domain multiscale method and its high performance computing implementation. J. Comput. Theor. Nanosci. **5**, 1–10 (2008)
29. Xiao, S.P., Belytschko, T.: A bridging domain method for coupling continua with molecular dynamics. Comput. Method Appl. Eng. **193**, 1645–1669 (2004)

A Moving Least Squares Approach to the Construction of Discontinuous Enrichment Functions

Marc Alexander Schweitzer and Sa Wu

Abstract In this paper we are concerned with the construction of a piecewise smooth field from scattered data by a moving least squares approach. This approximation problem arises when so-called enrichment functions for a generalized finite element method are computed by a particle scheme on a finer scale. The presented approach is similar in spirit to the so-called visibility criterion but avoids the explicit reconstruction of the location of the discontinuity.

1 Introduction

Generalized finite element methods (GFEM) were essentially introduced [6] to attain numerical schemes whose convergence properties are not limited by the regularity of the solution u of the considered partial differential equation (PDE), see also [1, 2, 4, 5, 7–9, 11–13, 21–23, 28, 29]. This desirable property is obtained in GFEM via the use of so-called enrichment functions which can represent the (dominant) non-smooth behavior of the solution. Thus, the approximation space V^{GFEM} used in a GFEM comprises classical piecewise-polynomial and problem-dependent discontinuous and singular shape functions; i.e.,

$$V^{\mathrm{GFEM}} = V^{\mathrm{smooth}} + V^{\mathrm{enrichment}} = V^{\mathrm{smooth}} + V^{\mathrm{discontinuous}} + V^{\mathrm{singular}}\,, \tag{1}$$

and we attain an improved convergence behavior essentially if the sought solution $u \in H^t(\Omega)$ e.g. admits a splitting

$$u = u_{\mathrm{s}} + u_{\mathrm{e}} \tag{2}$$

M.A. Schweitzer (✉) · S. Wu
Institut für Parallele und Verteilte Systeme, Universität Stuttgart, Universitätsstraße 38,
D-70569 Stuttgart, Germany
e-mail: marc.alexander.schweitzer@ipvs.uni-stuttgart.de; sa.wu@ipvs.uni-stuttgart.de

M. Griebel (ed.), *Singular Phenomena and Scaling in Mathematical Models*,
DOI 10.1007/978-3-319-00786-1_15,

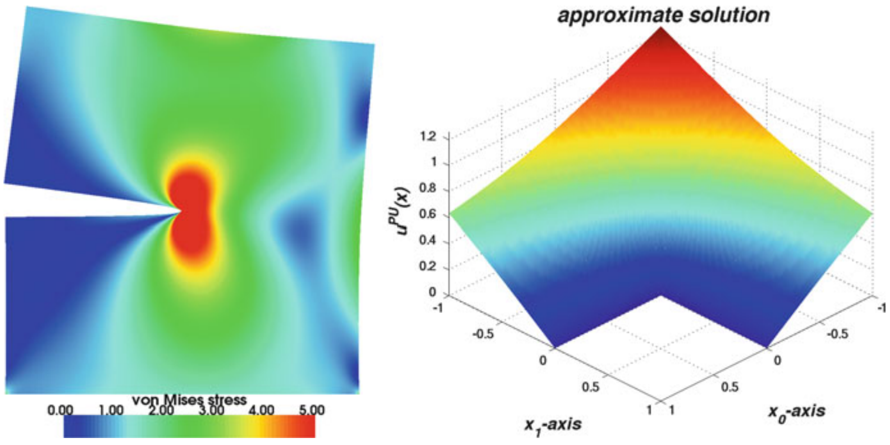


Fig. 1 Contour plot of the Von Mises stress on the deformed material configuration for a crack problem (*left*) and a surface plot of the approximation of a Poisson problem on a L-shaped domain (*right*)

such that $u_{\mathrm{e}} \in V^{\mathrm{enrichment}}$ and $u_{\mathrm{s}} \in H^{\mathrm{r}}(\Omega)$ with $r > t$. Thus, this information must be available a priori to be incorporated in the design of the trial space V^{GFEM}, i.e. in the selection of the basis functions for the enrichment space $V^{\mathrm{enrichment}}$.

If the discontinuities and singularities of the solution are directly induced by geometric information as for instance at re-entrant corners or in crack problems, compare Fig. 1, this a priori knowledge is (often) available in closed form.

For a straight crack in two-dimensional linear fracture mechanics appropriate enrichment basis functions are obtained from an asymptotic expansion of the solution, compare [16, 26]. However, in three dimensions or for a curved crack, the quality of these enrichment functions is limited and the improvement in the convergence behavior is less pronounced. Thus, we must either employ adaptive refinement in the global GFEM or we can try to improve on the quality of the enrichment functions. This may either be achieved in an a priori fashion [30, 31] or a posteriori. Here, one technique is the global-local approach [10, 17] which allows for the correction of the enrichment functions in an iterative fashion. The approach proposed in [3] is aimed directly at the construction of optimal local approximation spaces for the considered PDE problem via the approximate solution of a sequence of local cell problems.

The common feature of all enrichment strategies is that the enrichment space must resolve an intrinsically finer scale or feature of the solution than the resolution of the standard approximation space V^{smooth} admits. Thus, enriched GFEM can be interpreted as a multiscale finite element method where the construction of the enrichment space aims at the resolution of the fine scale features of the solution $u - \Pi_{\mathrm{smooth}} u$, i.e. the components of u that cannot be approximated well in V^{smooth}.

In this paper, we assume that this approximation of the fine scale components is carried out by a particle method and focus on the (re-)construction of a respective field on the continuum scale which can be used as an enrichment function in a GFEM. The assumption that a particle method is employed on the fine scale stems from the fact that many physical processes such as crack formation and crack initiation, which we consider as our reference problem, are naturally included in particle models whereas they require additional assumptions and equations in classical continuum models. Thus, we focus on the construction of enrichment functions with discontinuities and singularities from given particle data.

The remainder of this paper is structured as follows. First we discuss the types of data we assume to get from the particle method used for the construction of enrichment functions. This will turn out to be data admitting a broad class of underlying particle methods with a natural representation of particle adjacency. Next we give a brief review of the Moving Least Squares (MLS) Method for approximating scattered data. Then we discuss an adjacency data based modification of the MLS weights that allows for the construction of discontinuous approximants. Finally we give some application examples for simple standard discontinuities and conclude with some remarks.

2 Particle Data

As basis for the construction of enrichment functions $\eta : \Omega \subseteq \mathbb{R}^d \to \mathbb{R}$ we assume to have access to

$$\text{particle positions} \quad \mathbf{x}_\mathrm{i} \in \Omega\ , \tag{3a}$$

$$\text{function values} \quad u_\mathrm{i} \in \mathbb{R} \ \text{ and} \tag{3b}$$

$$\text{adjacency data} \quad \mathsf{A}_{\mathrm{i,j}} = \begin{cases} 1 & \text{no discontinuity} \\ 0 & \text{discontinuity} \end{cases} \text{ between particles } \mathbf{x}_\mathrm{i}, \mathbf{x}_\mathrm{j} \tag{3c}$$

from the fine scale approximation by a particle method. We first remark that $\eta : \Omega \to \mathbb{R}$ is not a restriction as vector fields can be simply modeled as several such interpolation problems. Except for the $\mathsf{A}_{\mathrm{i,j}}$ this is the basis for a regular textbook interpolation or approximation problem. However, as stated in the introduction, we want to construct fields on the continuum scale from the fine scale particle data which may be discontinuous or may have discontinuous derivatives.

On the particle level an adjacency matrix based representation of known fault lines seems very natural. With molecular dynamics the nucleation and growth of microfractures can be modelled by using bonded potentials and checking for individual bond breakage. This kind of representation also allows the integration of adjacency information strictly based on function values as used in edge detection. Furthermore we require no such thing as level set functions or even an explicit mesh for lines of discontinuity.

Discontinuous interpolation techniques, e.g. [20, 33], often rely on such a direct representation of discontinuities or extract these from function data. On the other hand many algorithms from image analysis deal with discontinuities, e.g. edge detection, but generally do not take more information than just the image into account, i.e. just $(\mathbf{x}_{\mathrm{i}}, u_{\mathrm{i}})$.

Note, however, that we are not concerned with the solution of an arbitrary approximation problem of discontinuous functions (which is usually stated in different function spaces). The fields u we are interested in are piecewise smooth functions e.g. $u|_{\Omega_k} \in H^t(\Omega_k)$ for $\Omega_k \subset \Omega$ and $k = 1, 2, \ldots, M$ with (relatively) small M. Thus, our focus is on the automatic identification of such discontinuities from function values and adjacency information only to attain a piecewise smooth field.

3 The Moving Least Squares Method

The Least Squares Method is a standard meshfree method for approximation of scattered data. By minimizing the (global) error functional

$$J(\eta) = \sum_{\mathrm{k}=1}^{\mathrm{N}} (\eta(\mathbf{x}_{\mathrm{k}}) - u_{\mathrm{k}})^2 \tag{4}$$

over some function space $V(\Omega)$ we obtain at *one* $\eta \in V(\Omega)$ approximating all data in an average sense. This minimization problem is typically posed in some linear, finite dimensional space $V = \mathrm{span}\,(p_1, \cdots, p_{\mathrm{D_V}})$, often the space of polynomial functions $\mathscr{P}_{\mathrm{K}}$ for some K. This leads to a single linear system

$$\left(\sum_{\mathrm{k}=1}^{\mathrm{N}} p_{\mathrm{i}}(\mathbf{x}_{\mathrm{k}}) p_{\mathrm{j}}(\mathbf{x}_{\mathrm{k}})\right)_{\mathrm{i,j}=1}^{D_V} \left(\alpha_j\right)_{\mathrm{j}=1}^{\mathrm{D}_V} = \left(\sum_{\mathrm{k}=1}^{\mathrm{N}} p_{\mathrm{i}}(\mathbf{x}_k) u_{\mathrm{k}}\right)_{\mathrm{i}=1}^{\mathrm{D}_V}, \tag{5}$$

whose solution gives the coefficients of the global solution $u = \sum_{\mathrm{i}=1}^{\mathrm{D}_V} \alpha_{\mathrm{i}} p_{\mathrm{i}}$. With this approach local changes to u_{i} or $\mathbf{x}_{\mathrm{i}}$ generally change the whole global solution u everywhere.

A localized weighted extension to remedy this issue can be obtained by the inclusion of weighting functions $W_{\mathrm{k}} : \Omega \to [0, \infty)$, usually splines or Gaussians. For each $\mathbf{x}$, $W_{\mathrm{k}}(\mathbf{x})$ should signify, how important data u_{k} at $\mathbf{x}_k$ is for the approximation at $\mathbf{x} \in \Omega$. This leads to the Moving Least Squares Method [14, 15, 18, 19, 25, 27, 32]. Taking (4) and adding the weights W_{k} leads to a family of $\mathbf{x}$ dependant functionals

$$J_{\mathbf{x}}(\eta) = \sum_{\mathrm{k}=1}^{N} W_k(\mathbf{x}) \left(\eta(\mathbf{x}_{\mathrm{k}}) - u_{\mathrm{k}}\right)^2 \ . \tag{6}$$

defined on some possibly $\mathbf{x}$ dependant function spaces $V_{\mathbf{x}}$. The minimizer $\eta_{\mathbf{x}} \in V_{\mathbf{x}}$ of $J_{\mathbf{x}}$ now denotes an approximation at a specific location $\mathbf{x} \in \Omega$ only; i.e. just the value $\eta_{\mathbf{x}}(\mathbf{x})$ is relevant from $\eta_{\mathbf{x}}$. The global solution η is then simply the collection of all the pointwise approximations $\eta_{\mathbf{x}}(\mathbf{x})$, i.e.

$$\eta(\mathbf{x}) = \eta_{\mathbf{x}}(\mathbf{x}) \ . \tag{7}$$

As with regular Least Squares, taking $\eta_{\mathbf{x}}$ from some linear space

$$V_{\mathbf{x}} = \operatorname{span}\left(p_{\mathbf{x},1}, \cdots, p_{\mathbf{x},\mathrm{D}_{V_{\mathbf{x}}}}\right),$$

we obtain the minimizer of (6) via the solution of a linear system $\mathsf{G}_{\mathbf{x}}\alpha_{\mathbf{x}} = \mathbf{f}_{\mathbf{x}}$, i.e.

$$\underbrace{\left(\sum_{\mathrm{k}=1}^{\mathrm{N}} W_{\mathrm{k}}(\mathbf{x}) p_{\mathbf{x},\mathrm{i}}(\mathbf{z}_{\mathrm{k}}) p_{\mathbf{x},\mathrm{j}}(\mathbf{z}_{\mathrm{k}})\right)_{\mathrm{i,j}=1}^{\mathrm{D}_{V_{\mathbf{x}}}}}_{=:\mathsf{G}_{\mathbf{x}}} \underbrace{\left(\alpha_{\mathbf{x},j}\right)_{\mathrm{j}=1}^{\mathrm{D}_{V_{\mathbf{x}}}}}_{=:\alpha_{\mathbf{x}}} = \underbrace{\left(\sum_{\mathrm{k}=1}^{\mathrm{N}} W_{\mathrm{k}}(\mathbf{x}) p_{\mathbf{x},\mathrm{i}}(\mathbf{z}_{\mathrm{k}}) u_{\mathrm{k}}\right)_{\mathrm{i}=1}^{\mathrm{D}_{V_{\mathbf{x}}}}}_{=:\mathbf{f}_{\mathbf{x}}}, \tag{8}$$

where $\mathbf{z}_{\mathrm{k}} = \mathbf{x}_{\mathrm{k}} - \mathbf{x}$. This yields

$$\eta(\mathbf{x}) = \eta_{\mathbf{x}}(x) = \sum_{i=1}^{\mathrm{D}_{V_{\mathbf{x}}}} u_{\mathbf{x},\mathrm{i}} p_{\mathbf{x},\mathrm{i}}(0) \ . \tag{9}$$

The formulation is given in terms of shifted coordinates $\mathbf{x}_{\mathrm{k}} - \mathbf{x}$ rather than $\mathbf{x}_{\mathrm{k}}$, i.e. the basis functions are given as $f(\mathbf{y}) = g(\mathbf{y} - \mathbf{x})$. This is useful since with $V = \mathscr{P}_{\mathrm{K}}$ and the usual monomial basis this shifted approach yields $\eta(\mathbf{x}) = \alpha_{\mathbf{x},1}$, compare Chap. 22 of [15]. Additionally shifting and scaling relative to the evaluation point is a standard technique for improvement of stability.

For stability concerns one might also circumvent the direct solution of (8) by "solving"

$$\underbrace{\left(\sqrt{W_{\mathrm{k}}(\mathbf{x})}\, p_{\mathbf{x},\mathrm{j}}(\mathbf{x}_{\mathrm{k}} - \mathbf{x})\right)_{\mathrm{k}=1,\ldots,\mathrm{N};\mathrm{j}=1,\ldots,\mathrm{D}_{V_{\mathbf{x}}}}}_{=:\mathsf{A}_{\mathbf{x}}} \left(\alpha_{\mathbf{x},\mathrm{j}}\right)_{\mathrm{j}=1}^{\mathrm{D}_{V_{\mathbf{x}}}} = \underbrace{\left(\sqrt{W_{\mathrm{k}}(\mathbf{x})}\, u_{\mathrm{k}}\right)_{\mathrm{k}=1}^{\mathrm{N}}}_{=:\mathbf{b}_{\mathbf{x}}} \tag{10}$$

via a pseudoinverse obtained from a singular value decomposition of $\mathsf{A}_{\mathbf{x}}$ to obtain a minimum of (6), as seen in [24], Sect. 2.6.

Note that in the most general case the approximation is defined by single evaluations of different functions $\eta_{\mathbf{x}}$ from different spaces $V_{\mathbf{x}}$. Even if the local approximations $\eta_{\mathbf{x}}$ stem from the same function space $V_{\mathbf{x}} = V$, generally $\eta \notin V$. For instance taking Shepard approximation [27], i.e. $V = \mathscr{P}_0 = \operatorname{span}(\mathbf{x} \mapsto 1)$, η is not a constant function, as it would be with regular least squares with $V = \mathscr{P}_0$.

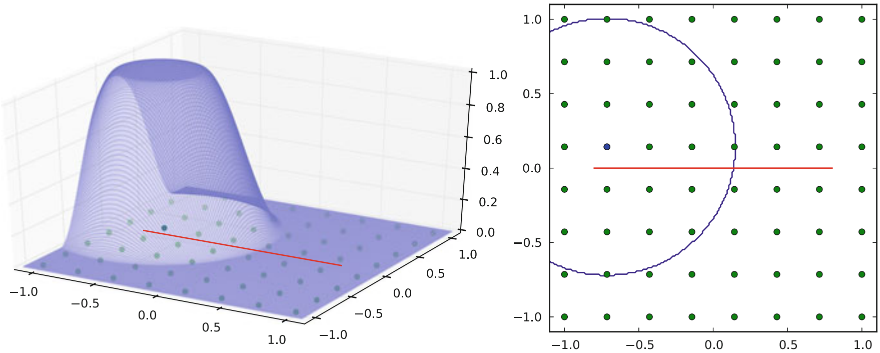


Fig. 2 Surface plot of a particular weight W_k (*left*) and its support (*right*)

4 Adjacency Based Modification

Recall that we are interested in the construction of piecewise smooth functions. The smoothness of η in (9) depends on the smoothness of the weights W_k, the distribution of $\mathbf{x}_\mathrm{k}$ and the smoothness of the functions found in $V_\mathbf{x}$. Let us assume that the employed approximation spaces $V_\mathbf{x} = V$ are identical throughout the domain. This leaves us the particle distribution and the weight functions to control the regularity of η. In our setting the particle distribution comes from a simulation itself and thus is not a free parameter that we may manipulate. Hence, we may affect the regularity of η through the choice of the weight functions W_k only.

To this end, we will modify the classical MLS weights W_k, e.g. see Fig. 2, with the help of the adjacency information A in such a way, that the approximation at a particular location $\mathbf{x} \in \Omega$ employs information of nearby particles $\mathbf{x}_\mathrm{k}$ in a strongly weighted fashion. Consider the particle $\mathbf{x}_\mathrm{k}$ which is closest to the point of evaluation $\mathbf{x} \in \Omega$. Now assume that there is another particle $\mathbf{x}_\mathrm{l}$ in the vicinity of $\mathbf{x}$ that would be used in the classical MLS approximation but is *not* connected to the particle $\mathbf{x}_\mathrm{k}$; i.e., there may be a discontinuity between particles $\mathbf{x}_\mathrm{k}$ and $\mathbf{x}_\mathrm{l}$. Then, we want to limit or reduce the influence of the value u_l at $\mathbf{x}_\mathrm{l}$ on the value η_x at $\mathbf{x}$ whereas the influence of u_k at $\mathbf{x}_\mathrm{k}$ should be increased. In essence, the value of the weight $W_\mathrm{k}(\mathbf{x})$ at $\mathbf{x}$ should be much larger than the value of the weight $W_\mathrm{l}(\mathbf{x})$. Such a modification of the weight functions is attained in the following way.

Assume that each particle is assigned a standard spline or Gaussian weight function W_k such that the diameter of the support of this weight function is e.g.

$$\frac{1}{2}\,\mathrm{diam}(\mathrm{supp}(W_\mathrm{k})) = 3\,h \tag{11}$$

where

$$h_\mathrm{k} = \min_\mathrm{l} \|\mathbf{x}_\mathrm{k} - \mathbf{x}_\mathrm{l}\|\ , \qquad\qquad h = \max_\mathrm{k} h_\mathrm{k}\ . \tag{12}$$

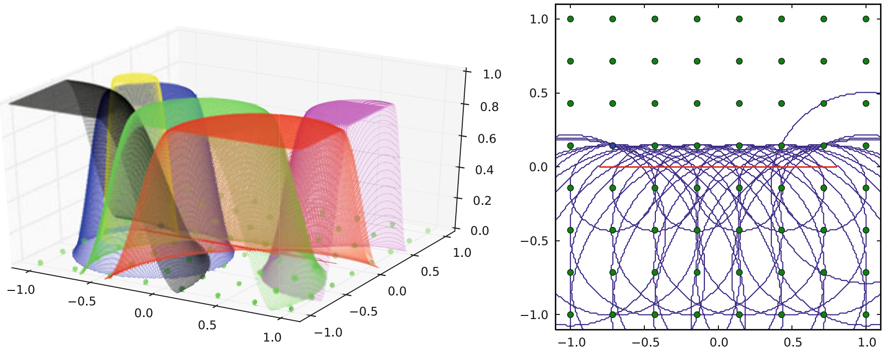


Fig. 3 Supports of all w_l with $\mathsf{A}_{\mathrm{k,l}} = 0$ (*right*) and a surface plot of several w_l (*left*)

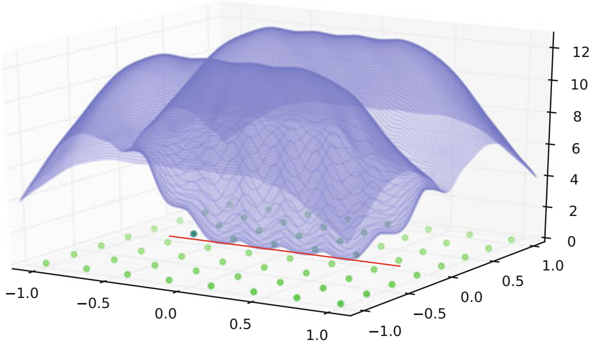


Fig. 4 Surface plot of the sum $\sum \tilde{W}_\mathrm{k}$ of all modified weights $\tilde{W}_\mathrm{k}$ (14) for a uniform particle distribution (*green dots*) and a straight discontinuity (*red line*)

To make use of the adjacency information A we assign an additional weight function w_k to each particle $\mathbf{x}_\mathrm{k}$ which satisfies

$$w_\mathrm{k}(\mathbf{x}_\mathrm{k}) = 1, \qquad\qquad \frac{1}{2}\,\mathrm{diam}(\mathrm{supp}(w_\mathrm{k})) \geq h_\mathrm{k}. \tag{13}$$

With the help of these weights w_l we now correct the original MLS weights W_k such that the flow of information across a potential discontinuity indicated by A is limited; i.e., we define the modified weights

$$\tilde{W}_\mathrm{k}(\mathbf{x}) = W_\mathrm{k}(\mathbf{x}) \prod_{\mathrm{l}:\,\mathsf{A}_{\mathrm{k,l}}=0} (1 - w_\mathrm{l}(\mathbf{x})) \tag{14}$$

to be used in (6), (8), compare Fig. 3. Note that $\tilde{W}_\mathrm{k}(\mathbf{x}_\mathrm{l}) = \tilde{W}_\mathrm{l}(\mathbf{x}_\mathrm{k}) = 0$ if $\mathsf{A}_{\mathrm{k,l}} = 0$, i.e., if the particles $\mathbf{x}_\mathrm{k}$ and $\mathbf{x}_\mathrm{l}$ are not directly connected, see Fig. 4.

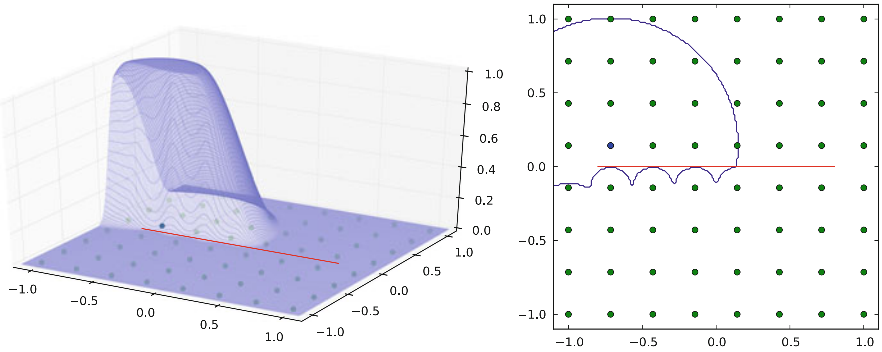


Fig. 5 Surface plot (*left*) of the modified weight $\tilde{W}_\mathrm{k}$ (14) and its support (*right*)

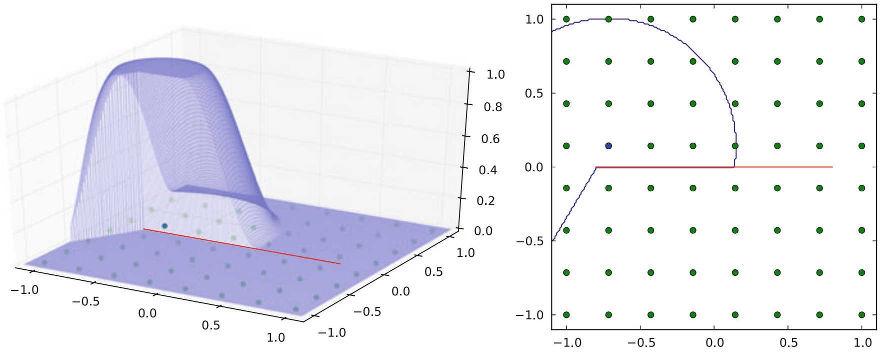


Fig. 6 Surface plot of a weight incorporating the visibility criterion (*left*) and its support (*right*)

Thus, the MLS approximation (6), (8) using the weights (14) in the vicinity of $\mathbf{x}_\mathrm{l}$ is (essentially) not influenced by the value u_k at $\mathbf{x}_\mathrm{k}$ and vice versa. The resulting weight $\tilde{W}_\mathrm{k}$ and its support are depicted in Fig. 5. The formulas for the particular weights used here can be found in Sect. 5.

Note that by multiplication with w_k we essentially shrink the size of the supports of the resulting weight functions. This may lead to a violation of the *unisolvence* condition and thereby to a deterioration of the regularity of $\mathsf{G}_\mathbf{x}$. Hence we may obtain discontinuities wherever our modification makes the system matrix singular. This, however, is not a cause of concern in our setting since the respective enrichment (on the macroscale) will not be evaluated in the immediate vicinity of this discontinuity.

In spirit this approach is similar to various visibility criteria otherwise known from meshfree methods, e.g. EFG, which result in weights as depicted in Fig. 6. These however generally require explicit knowledge about the location of discontinuity prior to the approximation.

5 Proof of Concept

In the following, we present some initial numerical results which are especially concerned with the automatic identification of a known discontinuity by our modified MLS approach. To this end we consider weight functions

$$W_{\mathrm{k}}(\mathbf{x}) = s_{0,\mathrm{R}}(\|\mathbf{x}-\mathbf{x}_{\mathrm{k}}\|) \tag{15a}$$

$$w_{\mathrm{k}}(\mathbf{x}) = s_{\mathrm{r_k},2\mathrm{r_k}}(\|\mathbf{x}-\mathbf{x}_{\mathrm{k}}\|) \tag{15b}$$

where

$$p(x) = 2x^3 - 3x^2 + 1\ , \qquad s_{\mathrm{a,b}}(x) = \begin{cases} 1 & x \le a \\ p(\frac{x-a}{b-a}) & a < x < b \\ 0 & x \ge b \end{cases} \tag{16}$$

with p being the first Hermite polynomial, which satisfies

$$p(0) = 1\ , \qquad p(1) = 0\ , \qquad p'(0) = 0\ , \qquad p'(1) = 0\ ,$$

h is as in (12) and

$$R = 3h\ , \qquad r_{\mathrm{k}} = \min_{\mathrm{l}:\ \mathrm{A_{k,l}}=0} \frac{1}{2}\|\mathbf{x}_{\mathrm{k}}-\mathbf{x}_{\mathrm{l}}\|\ . \tag{17}$$

For the sampling points $\mathbf{x}_{\mathrm{k}}$ we use a regular Cartesian grid.

In the 1D case we employ the presented approach to the approximation of the three reference functions

$$u_1(x) = \chi_{[0,1]}(x) = \begin{cases} 0 & x < 0 \\ 1 & x \ge 0 \end{cases}\ , \quad u_2(x) = |x|\ , \quad u_3(x) = \begin{cases} -p(-x) & x < 0 \\ p(x) & x \ge 0 \end{cases} \tag{18}$$

on $\Omega = [-1, 1]$.

The adjacency information was just initialized to split $[-1, 1]$ at 0, i.e.

$$\mathsf{A}_{\mathrm{i,j}} = \begin{cases} 1 & \big(x_{\mathrm{i}} \ge 0 \wedge x_{\mathrm{j}} \ge 0\big) \vee (x_{\mathrm{i}} < 0 \wedge x_{\mathrm{i}} < 0) \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

for all three functions.

The plots in Figs. 7–9 show the approximations with 20 nodes and $V = \mathscr{P}_3$ using regular weights on the left, our modified approach on the right. The underlying functions $u_{\mathrm{i}}^{(\mathrm{l})}$ are the blue lines, the point evaluations of the MLS approximants red

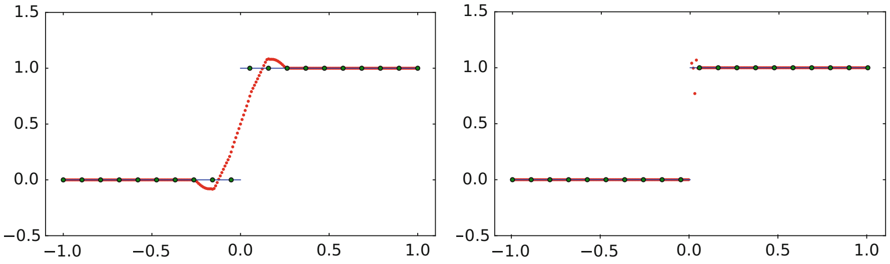


Fig. 7 Approximations of u_1 (18) by a MLS approach using W_k (15) (*left*) and with our modified weights $\widetilde{W}_\mathrm{k}$ (14) (*right*)

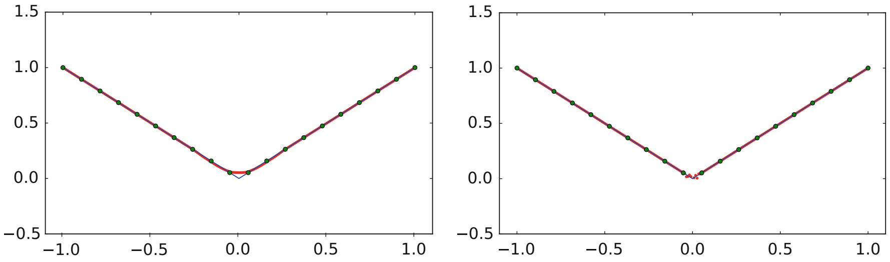


Fig. 8 Approximations of u_2 (18) by a MLS approach using W_k (15) (*left*) and with our modified weights $\widetilde{W}_\mathrm{k}$ (14) (*right*)

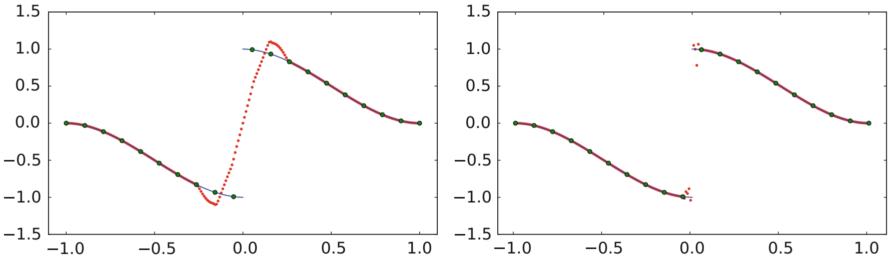


Fig. 9 Approximations of u_2 (18) by a MLS approach using W_k (15) (*left*) and with our modified weights $\widetilde{W}_\mathrm{k}$ (14) (*right*)

dots, and the underlying data nodes $(\mathbf{x}_\mathrm{k}, u_\mathrm{k})$ green dots. From these plots we can clearly observe the improvement due to the modification of the weights. With the standard weights over- and undershoots near the discontinuities are attained whereas our approach yields a perfect agreement with the data at the nodes even next to the discontinuity.

Note that the values of the approximation obtained in the interval $[\mathbf{x}_\mathrm{k}, \mathbf{x}_{\mathrm{k}+1}]$ which contains the discontinuity will never be used in our enrichment approach. An enrichment function will essentially be evaluated only at certain integration points (on the macroscale) which are placed well away from the discontinuities.

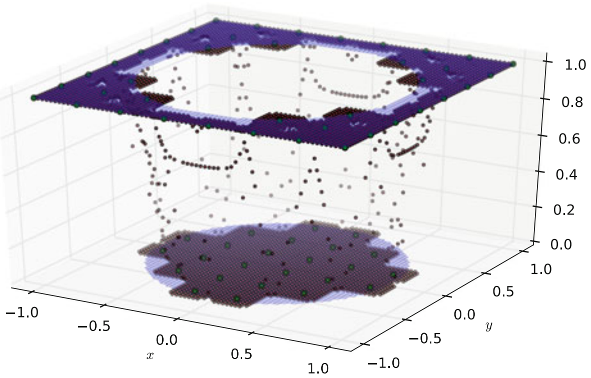


Fig. 10 Approximation of v_1 (20a) by a MLS approach using our modified weights $\widetilde{W}_k$ (14)

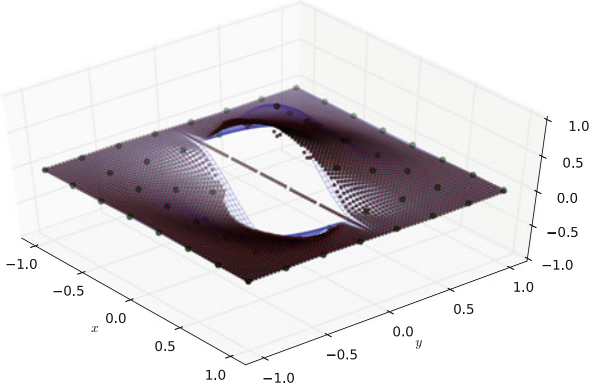


Fig. 11 Approximation of v_2 (20b) by a MLS approach using our modified weights $\widetilde{W}_k$ (14)

Finally, shown in Figs. 10 and 11, we consider the two dimensional examples

$$v_1(\mathbf{x}) = 1 - \chi_{B_{0.8}(0)}(\mathbf{x}) = \begin{cases} 1 & \|x\| \geq 0.8 \\ 0 & \|x\| < 0.8 \end{cases} \tag{20a}$$

$$v_2(\mathbf{x}) = \begin{cases} -s_{0,0.8}(|\mathbf{x}_1|)s_{0,0.8}(|\mathbf{x}_2|) & \mathbf{x}_2 < 0 \\ s_{0,0.8}(|\mathbf{x}_1|)s_{0,0.8}(|\mathbf{x}_2|) & \mathbf{x}_2 \geq 0 \end{cases} \tag{20b}$$

on $\Omega = [-1, 1]^2$ with s from (16) and

$$(\mathsf{A}_1)_{\mathrm{k,l}} = \begin{cases} 1 & (\mathbf{x}_\mathrm{k}, \mathbf{x}_\mathrm{l} \in B_{0.8}(0)) \vee (\mathbf{x}_\mathrm{k}, \mathbf{x}_\mathrm{l} \notin B_{0.8}(0) \wedge [\mathbf{x}_\mathrm{k}, \mathbf{x}_\mathrm{l}] \cap B_{0.8}(0) \neq \emptyset) \\ 0 & \text{otherwise} \end{cases} \tag{21a}$$

$$(\mathsf{A}_2)_{\mathrm{k,l}} = \begin{cases} 1 & [\mathbf{x}_\mathrm{k}, \mathbf{x}_\mathrm{l}] \text{ does not intersect } [(-0.8, 0), (0.8, 0)] \\ 0 & \text{otherwise} \end{cases} \tag{21b}$$

denoting the adjacency information for v_i.

Again, with respect to the (fine scale) data nodes $\mathbf{x}_\mathrm{k}$ we obtain a very good agreement of our approach and observe a highly localized zone near the discontinuity where $\mathsf{G}_\mathbf{x}$ becomes singular.

6 Concluding Remarks

We presented an MLS based approach for the construction of discontinuous approximants from particle data. The overall goal of our research is the construction of appropriate enrichment functions for a generalized finite element method from (local) fine scale particle simulations. So far we have merely considered simple reference problems in one and two dimensions. These results are promising. However a detailed numerical study and comparison with other approaches (visibility, direct representation of the discontinuity) is necessary prior to the use of the computed enrichment functions in a three dimensional application setting.

References

1. Aragón, A.M., Duarte, C.A.M., Geubelle, P.H.: Generalized finite element enrichment functions for discontinuous gradient fields. Int. J. Numer. Methods Eng. **82**(2), 242–268 (2010)
2. Babuška, I., Banerjee, U.: Stable generalized finite element method. Technical report, ICES (2011)
3. Babuška, I., Lipton, R.: Optimal local approximation spaces for generalized finite element methods with application to multiscale problems. SIAM Multiscale Model. Simul. **9**(1), 373–406 (2011)
4. Babuška, I., Melenk, J.M.: The partition of unity finite element method: basic theory and applications. Comput. Methods Appl. Mech. Eng. **139**, 289–314 (1996). Special issue on meshless methods
5. Babuška, I., Melenk, J.M.: The partition of unity method. Int. J. Numer. Methods Eng. **40**, 727–758 (1997)
6. Babuška, I., Caloz, G., Osborn, J.E.: Special finite element methods for a class of second order elliptic problems with rough coefficients. SIAM J. Numer. Anal. **31**, 945–981 (1994)
7. Babuška, I., Banerjee, U., Osborn, J.E.: Meshless and generalized finite element methods: a survey of some major results. In: Griebel, M., Schweitzer M.A. (eds.) Meshfree Methods for Partial Differential Equations. Lecture Notes in Computational Science and Engineering, vol. 26, pp. 1–20. Springer, Berlin/Heidelberg (2003)

8. Belytschko, T., Black, T.: Elastic crack growth in finite elements with minimal remeshing. Int. J. Numer. Methods Eng. **45**, 601–620 (1999)
9. Belytschko, T., Moës, N., Usui, S., Parimi, C.: Arbitrary discontinuities in finite elements. Int. J. Numer. Methods Eng. **50**, 993–1013 (2001)
10. Duarte, C.A., Kim, D.J.: Analysis and applications of a generalized finite element method with global-local enrichment functions. Comput. Mech. **50**, 563–578 (2012)
11. Duarte, C.A.M., Babuška, I., Oden, J.T.: Generalized finite element methods for three dimensional structural mechanics problems. Comput. Struct. **77**, 215–232 (2000)
12. Duarte, C.A.M., Hamzeh, O.N., Liszka, T.J., Tworzydlo, W.W.: A generalized finite element method for the simulation of three-dimensional dynamic crack propagation. Int. J. Numer. Methods Eng. **190**, 2227–2262 (2001)
13. Duarte, C.A., Reno, L.G., Simone, A.: A higher order generalized FEM for through-the-thickness branched cracks. Int. J. Numer. Methods Eng. **72**(3), 325–351 (2007)
14. Farwig, R.: Multivariate interpolation of arbitrarily spaced data by moving least squares methods. J. Comput. Appl. Math. **16**(1), 79–93 (1986)
15. Fasshauer, G.F.: Meshfree Approximation Methods with MATLAB. World Scientific, River Edge (2007)
16. Fries, T.P., Belytschko, T.: The extended/generalized finite element method: an overview of the method and its applications. Int. J. Numer. Methods Eng. **84**(3), 253–304 (2010)
17. Kim, D.J., Duarte, C.A., Proenca, S.P.: Generalized finite element method with global-local enrichments for nonlinear fracture analysis. In: da Costa Mattos, H.S., Alves, M. (eds.) Mechanics of Solids in Brazil 2009, Rio de Janeiro (2009)
18. Levin, D.: The approximation power of moving least-squares. Math. Comput. **67**(224), 1517–1531 (1998)
19. Li, S., Liu, W.K.: Meshfree Particle Methods. Springer, Berlin/New York (2004). Incorporated
20. López de Silanes, M.C., Parra, M.C., Pasadas, M., Torrens, J.J.: Spline approximation of discontinuous multivariate functions from scattered data. J. Comput. Appl. Math. **131**(1–2), 281–298 (2001)
21. Moës, N., Dolbow, J., Belytschko, T.: A finite element method for crack growth without remeshing. Int. J. Numer. Methods Eng. **46**, 131–150 (1999)
22. Mohammadi, S.: Extended Finite Element Method. Blackwell Publishing, Oxford/Malden (2008)
23. Pereira, J.P., Duarte, C.A.M., Guoy, D., Jiao, X.: Hp-generalized fem and crack surface representation for non-planar 3-d cracks. Int. J. Numer. Methods Eng. **77**(5), 601–633 (2009)
24. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipes: The Art of Scientific Computing, 3rd edn. Cambridge University Press, Cambridge/New York (2007)
25. Schweitzer, M.A.: Generalized finite element and meshfree methods. Habilitation, Univeristät Bonn (2008)
26. Schweitzer, M.A.: Generalizations of the finite element method. Cent. Eur. J. Math. **10**, 3–24 (2012)
27. Shepard, D.: A two-dimensional interpolation function for irregularly-spaced data. In: Proceedings of the 1968 23rd ACM national conference, ACM'68, pp. 517–524. ACM, New York (1968)
28. Strouboulis, T., Babuška, I., Copps, K.: The design and analysis of the generalized finite element method. Comput. Methods Appl. Mech. Eng. **181**(I 3), 43–69 (2000)
29. Strouboulis, T., Copps, K., Babuška, I.: The generalized finite element method. Comput. Methods Appl. Mech. Eng. **190**, 4081–4193 (2001)
30. Strouboulis, T., Zhang, L., Babuška, I.: Generalized finite element method using mesh-based handbooks: application to problems in domains with many voids. Comput. Methods Appl. Mech. Eng. **192**, 3109–3161 (2003)
31. Strouboulis, T., Zhang, L., Babuška, I.: p-Version of the generalized FEM using mesh-based handbooks with applications to multiscale problems. Int. J. Numer. Methods Eng. **60**, 1639–1672 (2004)

32. Wendland, H.: Local polynomial reproduction and moving least squares approximation. IMA J. Numer. Anal. **21**, 285–300 (2001)
33. Xu, J., Belytschko, T.: Discontinuous radial basis function approximations for meshfree methods. In: Griebel, M., Schweitzer, M.A. (eds.) Meshfree Methods for Partial Differential Equations II. Lecture Notes in Computational Science and Engineering, vol. 43, pp. 231–253. Springer, Berlin (2005)

Second Moment Analysis for Robin Boundary Value Problems on Random Domains

Helmut Harbrecht

Abstract We consider the numerical solution of Robin boundary value problems on random domains. The proposed method computes the mean and the variance of the random solution with leading order in the amplitude of the random boundary perturbation relative to an unperturbed, nominal domain. The variance is computed from the solution's two-point correlation which satisfies a deterministic boundary value problem on the tensor product of the nominal domain. We solve this moderatly high-dimensional problem by either a low-rank approximation by means of the pivoted Cholesky decomposition or the combination technique. Both approaches are presented and compared by numerical experiments with respect to their efficiency.

1 Introduction

Many problems in physics and engineering sciences lead to boundary value problems for an unknown function. In general, the numerical simulation is well understood provided that the input parameters are given exactly. Since, however, exact input parameters are often not known in engineering, it is of growing interest to model such parameters as random variables.

A principal approach to solve boundary value problems with random input parameters is the Monte Carlo approach, see e.g. [37] and the references therein. However, it is hard and extremely expensive to generate a large number of suitable samples and to solve a deterministic boundary value problem on each sample. Particularly in the present case of random domains, each new sample corresponds to a new domain which needs to be discretized. Thus, we aim here at a direct, deterministic method to compute the random solution.

H. Harbrecht (✉)
Mathematisches Institut, Universität Basel, Rheinsprung 21, CH-4051 Basel, Switzerland
e-mail: helmut.harbrecht@unibas.ch

M. Griebel (ed.), *Singular Phenomena and Scaling in Mathematical Models*,
DOI 10.1007/978-3-319-00786-1_16,

Deterministic approaches to solve stochastic partial differential equations have been proposed in e.g. [1, 11, 13, 14, 28, 32, 38]. Therein, loadings and coefficients have been considered as random input parameters. Recently, in [6,23,26,33,34,43], also the underlying domain has been modeled as a random input parameter $D(\omega)$. For example, this enables the consideration of tolerances in the shape of products fabricated by line production. Other applications arise from blurred interfaces like cell membranes or molecular surfaces.

The present paper is dedicated to the numerical treatment of Robin boundary value problems on random domains which, to the best of our knowledge, is the first time this subject is dealt with in the scientific literature. We assume small random perturbations around a nominal domain $\overline{D}$ with known second order statistics. Then, following [26], we can linearize to derive, with leading order in the amplitude of the perturbation parameter, deterministic equations for the random solution's expectation and two-point correlation

$$\left.\begin{aligned} \mathbb{E}_u(\mathbf{x}) &= \int_\Omega u(\mathbf{x},\omega)\,\mathrm{d}P(\omega) \\ \mathrm{Cor}_u(\mathbf{x},\mathbf{y}) &= \int_\Omega u(\mathbf{x},\omega)u(\mathbf{y},\omega)\,\mathrm{d}P(\omega) \end{aligned}\right\} \quad \mathbf{x},\mathbf{y}\in\overline{D}.$$

From these quantities the variance is derived by

$$\mathbb{V}_u(\mathbf{x}) = \mathrm{Cor}_u(\mathbf{x},\mathbf{y})\big|_{\mathbf{x}=\mathbf{y}} - \mathbb{E}_u^2(\mathbf{x}), \quad \mathbf{x}\in\overline{D}.$$

The solution's two-point correlation is given by a partial differential equation which lives on the tensor product domain $\overline{D}\times\overline{D}$. We solve this moderatly high-dimensional problem by either a low-rank approximation via the pivoted Cholesky decomposition or by the combination technique which is a special variant of a sparse tensor product approximation. This way, we are able to compute both, the expectation and the variance, by standard finite element techniques.

Besides the modeling and the derivation of the underlying equations, we discuss in this paper the implementation of the proposed algorithms. In particular, we compare the low-rank approximation and the sparse tensor product approximation with respect to their cost-complexities by numerical results.

The rest of the paper is organized as follows. In Sect. 2, we model the random domain under consideration. Moreover, for the associated Robin boundary value problem, we derive deterministic boundary value problems for the expectation and two-point correlation of the random solution. In Sect. 3, we introduce the variational formulations of these deterministic boundary value problems. Section 4 is dedicated to an abstract overview on the efficient solution of tensor product-type boundary value problems which arise in the present context. The particular finite element discretization of the problems under consideration is performed in Sect. 5. In Sect. 6, numerical experiments are carried out to validate the theoretical findings and to compare the low-rank approximation with the sparse grid approach. Finally, in Sect. 7, we state concluding remarks.

2 Robin Boundary Value Problems on Random Domains

Let (Ω, Σ, P) be a suitable probability space. We consider the domain $D(\omega)$ as the uncertain input parameter of an elliptic boundary value problem with Robin boundary conditions, i.e.,

$$\left.\begin{aligned} -\Delta u(\mathbf{x},\omega) &= f(\mathbf{x}), && \mathbf{x} \in D(\omega) \\ \alpha(\mathbf{x})u(\mathbf{x},\omega) + \frac{\partial u}{\partial \mathbf{n}}(\mathbf{x},\omega) &= g(\mathbf{x}), && \mathbf{x} \in \partial D(\omega) \end{aligned}\right\} \quad \omega \in \Omega. \tag{1}$$

Here, $\alpha(\mathbf{x}) \geq 0$ is a nonnegative function, where the particular choice $\alpha(\mathbf{x}) \equiv 0$ yields the Neumann boundary condition.

To model the random domain $D(\omega)$, let $\overline{D}$ denote a smooth reference domain and consider random boundary variations in the direction of the outer normal

$$\mathbf{U}(\mathbf{x},\omega) = \varepsilon\kappa(\mathbf{x},\omega)\mathbf{n}(\mathbf{x}) : \partial\overline{D} \to \mathbb{R}^n$$

with

$$\kappa(\omega) \in L_P^2\big(\Omega, C^{2,1}(\partial\overline{D})\big) \quad \text{such that} \quad \|\kappa(\omega)\|_{C^{2,1}(\partial\overline{D})} \leq 1$$

almost surely. Then, the random domain $D(\omega)$ will be described via perturbation of identity

$$\partial D(\omega) = \big\{\big(\mathbf{I} + \varepsilon\mathbf{U}(\omega)\big)(\mathbf{x}) = \mathbf{x} + \varepsilon\kappa(\mathbf{x},\omega)\mathbf{n}(\mathbf{x}) : \mathbf{x} \in \partial\overline{D}\big\}.$$

For what follows we assume that the expectation $\mathbb{E}_\kappa$ and the two-point correlation Cor_κ of the boundary perturbation κ are given. Without loss of generality (otherwise we redefine $\overline{D}$ correspondingly) we further assume that the perturbation field κ is centered, i.e., that $\mathbb{E}_\kappa \equiv 0$.

For a small perturbation amplitude $\varepsilon > 0$, one can linearize (1) by means of shape calculus [12, 40]. This leads to the following stochastic shape-Taylor expansion

$$u(\mathbf{x},\omega) = \overline{u}(\mathbf{x}) + \varepsilon\delta u[\kappa(\omega)](\mathbf{x}) + \mathscr{O}(\varepsilon^2), \quad \mathbf{x} \in K \Subset \overline{D}. \tag{2}$$

Therein, the compact set $K \Subset \overline{D}$ is assumed to satisfy $K \Subset D(\omega)$ almost surely. Moreover, $\overline{u} \in H^1(\overline{D})$ denotes the solution to the deterministic Robin boundary value problem

$$\begin{aligned} -\Delta\overline{u}(\mathbf{x}) &= f(\mathbf{x}), && \mathbf{x} \in \overline{D} \\ \alpha(\mathbf{x})\overline{u}(\mathbf{x}) + \frac{\partial\overline{u}}{\partial\mathbf{n}}(\mathbf{x}) &= g(\mathbf{x}), && \mathbf{x} \in \partial\overline{D} \end{aligned} \tag{3}$$

and the shape derivative $\delta u = \delta u[\kappa] \in H^1(\overline{D})$ satisfies the following Robin boundary value problem with random loading (cf. [31])

$$\begin{aligned} \Delta \delta u(\mathbf{x}) &= 0, && \mathbf{x} \in \overline{D} \\ \alpha(\mathbf{x})\delta u(\mathbf{x}) + \frac{\partial \delta u}{\partial \mathbf{n}}(\mathbf{x}) &= \operatorname{div}_\Gamma \big(\kappa(\mathbf{x})\nabla_\Gamma \overline{u}(\mathbf{x})\big) + \kappa(\mathbf{x})h(\mathbf{x}), && \mathbf{x} \in \partial\overline{D}. \end{aligned} \tag{4}$$

Here, we used the abbreviation

$$h(\mathbf{x}) := f(\mathbf{x}) + \mathscr{H}(\mathbf{x})\big(g(\mathbf{x}) - \alpha(\mathbf{x})\overline{u}(\mathbf{x})\big) + \frac{\partial(g - \alpha\overline{u})}{\partial \mathbf{n}}(\mathbf{x}), \tag{5}$$

where $\mathscr{H} = (n-1)\overline{\mathscr{H}}$ is the additive curvature and $\overline{\mathscr{H}}$ is the mean curvature of the surface Γ.

Theorem 1. *Assume that the compact set $K \Subset \overline{D}$ satisfies $K \Subset D(\omega)$ almost surely. Then, it holds that*

$$\left.\begin{aligned} \mathbb{E}_u(\mathbf{x}) &= \overline{u}(\mathbf{x}) + \mathscr{O}(\varepsilon^2) \\ \mathbb{V}_u(\mathbf{x}) &= \varepsilon^2 \operatorname{Cor}_{\delta u}(\mathbf{x},\mathbf{y})\big|_{\mathbf{x}=\mathbf{y}} + \mathscr{O}(\varepsilon^3) \end{aligned}\right\} \quad \mathbf{x} \in K. \tag{6}$$

Herein, $\overline{u} \in H^1(\overline{D})$ and $\operatorname{Cor}_{\delta u}(\mathbf{x},\mathbf{y}) \in H^1_{mix}(\overline{D} \times \overline{D}) := H^1(\overline{D}) \times H^1(\overline{D})$ satisfy the deterministic boundary value problems (3) and

$$\begin{aligned} (\Delta_\mathbf{x} \otimes \Delta_\mathbf{y}) \operatorname{Cor}_{\delta u}(\mathbf{x},\mathbf{y}) &= 0, \quad \mathbf{x},\mathbf{y} \in \overline{D}, \\ \Delta_\mathbf{x} \operatorname{Cor}_{\delta u}(\mathbf{x},\mathbf{y}) &= 0, \quad \mathbf{x} \in \overline{D},\ \mathbf{y} \in \partial\overline{D}, \\ \Delta_\mathbf{y} \operatorname{Cor}_{\delta u}(\mathbf{x},\mathbf{y}) &= 0, \quad \mathbf{x} \in \partial\overline{D},\ \mathbf{y} \in \overline{D}, \end{aligned}$$

$$\begin{aligned} \left[\left(\alpha(\mathbf{x}) + \frac{\partial}{\partial \mathbf{n}_\mathbf{x}}\right) \otimes \left(\alpha(\mathbf{y}) + \frac{\partial}{\partial \mathbf{n}_\mathbf{y}}\right)\right] \operatorname{Cor}_{\delta u}(\mathbf{x},\mathbf{y}) &= \operatorname{Cor}_\kappa(\mathbf{x},\mathbf{y})\big[h(\mathbf{x}) \otimes h(\mathbf{y})\big] \\ + \operatorname{div}_{\Gamma,\mathbf{x}} \big[\operatorname{Cor}_\kappa(\mathbf{x},\mathbf{y})\big(\nabla_\Gamma \overline{u}(\mathbf{x}) \otimes f(\mathbf{y})\big)\big] &+ \operatorname{div}_{\Gamma,\mathbf{y}} \big[\operatorname{Cor}_\kappa(\mathbf{x},\mathbf{y})\big(h(\mathbf{x}) \otimes \nabla_\Gamma \overline{u}(\mathbf{y})\big)\big] \\ + (\operatorname{div}_{\Gamma,\mathbf{x}} \otimes \operatorname{div}_{\Gamma,\mathbf{y}})\big[\operatorname{Cor}_\kappa(\mathbf{x},\mathbf{y})\big(\nabla_\Gamma \overline{u}(\mathbf{x}) \otimes \nabla_\Gamma \overline{u}(\mathbf{y})\big)\big], &\qquad \mathbf{x},\mathbf{y} \in \partial\overline{D}. \end{aligned} \tag{7}$$

Proof. By using the shape-Taylor expansion (2), we obtain

$$\mathbb{E}_u(\mathbf{x}) = \overline{u}(\mathbf{x}) + \varepsilon\mathbb{E}\big(\delta u[\kappa(\omega)](\mathbf{x})\big) + \mathscr{O}(\varepsilon^2).$$

By the linearity of the expectation operator $\mathbb{E}$, taking the expectation on both sides of (4), and observing that $\mathbb{E}_\kappa(\mathbf{x}) \equiv 0$, we have $\mathbb{E}_{\delta u}(\mathbf{x}) = \mathbb{E}\big(\delta u[\kappa(\omega)](\mathbf{x})\big) \equiv 0$, which yields the first claim.

Now, observe the following estimate

$$\begin{aligned}\mathbb{V}(a+bX+cY) &= b^2\mathbb{V}(X)+2bc\,\mathrm{Cov}(X,Y)+c^2\mathbb{V}(Y)\\ &\le b^2\mathbb{V}(X)+2bc\sqrt{\mathbb{V}(X)\mathbb{V}(Y)}+c^2\mathbb{V}(Y),\end{aligned}$$

where X and Y are two random variables with finite second moments. By combining this estimate with the shape-Taylor expansion (2), we conclude

$$\begin{aligned}\mathbb{V}_u(\mathbf{x}) &= \varepsilon^2\mathbb{V}\big(\delta u[\kappa(\omega)](\mathbf{x})\big)+\sqrt{\mathbb{V}\big(\delta u[\kappa(\omega)](\mathbf{x})\big)}\mathscr{O}(\varepsilon^3)+\mathscr{O}(\varepsilon^4)\\ &= \varepsilon^2\mathbb{V}_{\delta u}(\mathbf{x})+\mathscr{O}(\varepsilon^3).\end{aligned}$$

Due to $\mathbb{E}_{\delta u}(\mathbf{x})\equiv 0$, we arrive at the identity $\mathbb{V}_{\delta u}(\mathbf{x})=\mathrm{Cor}_{\delta u}(\mathbf{x},\mathbf{y})\big|_{\mathbf{x}=\mathbf{y}}$ which proves the second claim. The boundary value problem (7) for $\mathrm{Cor}_{\delta u}$ is finally derived by tensorizing (4) and taking the expectation. This completes the proof. □

Remark 1. The relative error of the expectation is $\mathscr{O}(\varepsilon^2)$ while the relative error of the variance is $\mathscr{O}(\varepsilon)$. According to [7], the first order shape-Taylor expansion (2) is nevertheless sufficient to compute also higher order moments of the random solution with relative accuracy $\mathscr{O}(\varepsilon)$.

3 Variational Formulation

We shall introduce the variational formulations of the boundary value problems under consideration. The approximate expectation $\overline{u}\in H^1(\overline{D})$, satisfying (3), is determined by the variational formulation

$$\text{seek } \overline{u}\in H^1(\overline{D}) \text{ such that } a(\overline{u},v)=\ell_1(v) \text{ for all } v\in H^1(\overline{D}), \tag{8}$$

where the bilinear form $a: H^1(\overline{D})\times H^1(\overline{D})\to\mathbb{R}$ is given by

$$a(u,v):=\int_{\overline{D}}\nabla u(\mathbf{x})\nabla v(\mathbf{x})\,\mathrm{d}\mathbf{x}+\int_{\partial\overline{D}}\alpha(\mathbf{x})u(\mathbf{x})v(\mathbf{x})\,\mathrm{d}\sigma$$

and the linear form $\ell_1: H^1(\overline{D})\to\mathbb{R}$ by

$$\ell_1(v):=\int_{\overline{D}}f(\mathbf{x})v(\mathbf{x})\,\mathrm{d}\mathbf{x}+\int_{\partial\overline{D}}g(\mathbf{x})v(\mathbf{x})\,\mathrm{d}\sigma.$$

The shape derivative $\delta u=\delta u[\kappa]\in H^1(\overline{D})$ in a given direction $\kappa\in C^{2,1}(\partial\overline{D})$ satisfies the boundary value problem (4). The associated variational formulation involves the same bilinear form as (8), but a different linear form on the right hand side. Namely, we find

$$\text{seek } \delta u \in H^1(\overline{D})\text{such that } a(\delta u, v) = \ell_2(v) \text{ for all } v \in H^1(\overline{D}), \tag{9}$$

with the linear form $\ell_2 : H^1(\overline{D}) \to \mathbb{R}$ being defined by

$$\ell_2(v) := \int_{\partial\overline{D}} \kappa(\mathbf{x})\big\{h(\mathbf{x}) - \nabla_\Gamma \overline{u}(\mathbf{x})\nabla_\Gamma\big\} v(\mathbf{x}) \, \mathrm{d}\sigma.$$

Note that we applied integration by parts in the definition of the linear form. Moreover, the function h is defined in (5). Thus, the two-point correlation function $\mathrm{Cor}_{\delta u} \in H^1_{mix}(\overline{D} \times \overline{D})$, which is given by the tensor Robin boundary value problem (7), satisfies the variational formulation

$$\begin{aligned} &\text{seek } \mathrm{Cor}_{\delta u} \in H^1_{mix}(\overline{D} \times \overline{D}) \text{ such that} \\ &\qquad A(\mathrm{Cor}_{\delta u}, v) = L(v) \text{ for all } v \in H^1_{mix}(\overline{D} \times \overline{D}). \end{aligned} \tag{10}$$

Here, the bilinear form $A : H^1_{mix}(\overline{D} \times \overline{D}) \times H^1_{mix}(\overline{D} \times \overline{D}) \to \mathbb{R}$ reads as

$$\begin{aligned} A(u, v) := &\int_{\overline{D}}\int_{\overline{D}} (\nabla_\mathbf{x} \otimes \nabla_\mathbf{y}) u(\mathbf{x}, \mathbf{y}) (\nabla_\mathbf{x} \otimes \nabla_\mathbf{y}) v(\mathbf{x}, \mathbf{y}) \, \mathrm{d}\mathbf{y} \, \mathrm{d}\mathbf{x} \\ &+ \int_{\overline{D}}\int_{\partial\overline{D}} \alpha(\mathbf{y}) \nabla_\mathbf{x} u(\mathbf{x}, \mathbf{y}) \nabla_\mathbf{x} v(\mathbf{x}, \mathbf{y}) \, \mathrm{d}\sigma_\mathbf{y} \, \mathrm{d}\mathbf{x} \\ &+ \int_{\partial\overline{D}}\int_{\overline{D}} \alpha(\mathbf{x}) \nabla_\mathbf{y} u(\mathbf{x}, \mathbf{y}) \nabla_\mathbf{y} v(\mathbf{x}, \mathbf{y}) \, \mathrm{d}\mathbf{y} \, \mathrm{d}\sigma_\mathbf{x} \\ &+ \int_{\partial\overline{D}}\int_{\partial\overline{D}} \alpha(\mathbf{x})\alpha(\mathbf{y}) u(\mathbf{x}, \mathbf{y}) v(\mathbf{x}, \mathbf{y}) \, \mathrm{d}\sigma_\mathbf{y} \, \mathrm{d}\sigma_\mathbf{x} \end{aligned}$$

and the linear form $L : H^1_{mix}(\overline{D} \times \overline{D}) \to \mathbb{R}$ is

$$\begin{aligned} L(v) := \int_{\partial\overline{D}}\int_{\partial\overline{D}} &\mathrm{Cor}_\kappa(\mathbf{x}, \mathbf{y}) \big\{h(\mathbf{x}) - \nabla_\Gamma \overline{u}(\mathbf{x}) \nabla_{\Gamma,\mathbf{x}}\big\} \\ &\cdot \big\{h(\mathbf{y}) - \nabla_\Gamma \overline{u}(\mathbf{y}) \nabla_{\Gamma,\mathbf{y}}\big\} v(\mathbf{x}, \mathbf{y}) \, \mathrm{d}\sigma_\mathbf{y} \, \mathrm{d}\sigma_\mathbf{x}. \end{aligned}$$

Theorem 2. *The variational problems* (8)–(10) *are uniquely solvable provided that* $\alpha(\mathbf{x}) \not\equiv 0$.

Proof. The standard theory of Robin boundary value problems yields the existence of constants $0 < c_E \leq c_S < \infty$ such that it holds

$$c_E \|u\|^2_{H^1(\overline{D})} \leq a(u,u), \quad a(u,v) \leq c_S \|u\|_{H^1(\overline{D})} \|v\|_{H^1(\overline{D})}$$

for all $u, v \in H^1(\overline{D})$. Thus, we conclude

$$c_E^2 \|u\|^2_{H^1_{mix}(\overline{D}\times\overline{D})} \leq A(u,u), \quad A(u,v) \leq c_S^2 \|u\|_{H^1_{mix}(\overline{D}\times\overline{D})} \|v\|_{H^1_{mix}(\overline{D}\times\overline{D})}$$

for all $u, v \in H^1_{mix}(\overline{D} \times \overline{D})$ by a tensor product argument since the bilinear form $A(\cdot,\cdot)$ is derived from $a(\cdot,\cdot)$ via tensorization. The Lax-Milgram theorem implies finally the assertion. □

Remark 2. If $\alpha(\mathbf{x}) \equiv 0$, then we arrive at the Neumann boundary value problem and obtain thus the ellipticity of $a(\cdot,\cdot)$ only in the space $\overline{H}^1(\overline{D}) := H^1(\overline{D}) \setminus \mathbb{R}$ and that of $A(\cdot,\cdot)$ in the space $\overline{H}^1_{mix}(\overline{D} \times \overline{D}) := \overline{H}^1(\overline{D}) \otimes \overline{H}^1(\overline{D})$. Consequently, unique solvability of the variational problems (8)–(10) is obtained in these energy spaces.

4 Solving Tensor Product Boundary Value Problems

4.1 An Abstract View on the Linearization Approach

The linearization of a linear second order elliptic boundary value problem with respect to a given input parameter $\kappa(\omega)$ involves the associated derivative $\delta u(\omega) \in \mathscr{H}(D)$. It is generally given by a boundary value problem

$$\mathscr{A}\delta u(\omega) = f(\omega) \quad \text{on } D,$$

where $\mathscr{A} : \mathscr{H}(D) \to \mathscr{H}'(D)$ denotes a linear, second order elliptic partial differential operator which is defined on a domain $D \subset \mathbb{R}^n$. Typically one might think of $\mathscr{H}(D)$ being a Sobolev space with dual $\mathscr{H}'(D)$. Moreover, the random input parameter linearly enters the right hand side $f(\omega) \in \mathscr{H}'(D)$ since the mapping $\kappa(\omega) \mapsto \delta u(\omega)$ is linear.

The two-point correlation $\mathrm{Cor}_{\delta u} \in \mathscr{H}_{mix}(D \times D) := \mathscr{H}(D) \otimes \mathscr{H}(D)$, which pops up in the asymptotic expansions (6), is given by the tensor product problem

$$(\mathscr{A} \otimes \mathscr{A})\,\mathrm{Cor}_{\delta u} = \mathrm{Cor}_f \quad \text{on } D \times D. \tag{11}$$

Especially it holds $\mathrm{Cor}_f \in \mathscr{H}'_{mix}(D \times D) = \mathscr{H}'(D) \otimes \mathscr{H}'(D)$.

In the following, we give an overview on the efficient solution of partial differential equations with the tensor product operator $\mathscr{A} \otimes \mathscr{A}$ on the product of the physical domain $D \times D$ such as (11). Various concepts are available to overcome the *curse of dimension* which is already observed in this moderatly high-dimensional situation.

4.2 Sparse Tensor Product Spaces

The starting point of the definition of sparse tensor product spaces for the Sobolev space $\mathscr{H}_{mix}(D \times D)$ are traditional and widely used multilevel hierarchies

$$V_0 \subset V_1 \subset V_2 \subset \cdots \subset \mathscr{H}(D), \tag{12}$$

where $\dim(V_j) \sim 2^{jn}$. Then, appropriate complement spaces

$$W_0 := V_0, \qquad W_j := V_j \ominus V_{j-1}, \quad j > 0$$

are chosen to derive the multiscale decomposition

$$V_J = W_0 \oplus W_1 \oplus \cdots \oplus W_J.$$

In general, such complement spaces are defined by hierarchical bases like e.g. wavelet or multilevel bases, see [5] and the references therein. The sparse tensor product space $\hat{V}_J \subset \mathscr{H}_{mix}(D \times D)$ is finally given via the complementary spaces according to

$$\hat{V}_J = \bigoplus_{j+j' \le J} W_j \otimes W_{j'} = \bigoplus_{j=0}^{J} V_j \otimes W_{J-j}. \tag{13}$$

The sparse tensor product space $\hat{V}_J$ possesses only $\mathscr{O}(2^{Jn}J)$ degrees of freedom which is much less than the $\mathscr{O}(2^{2Jn})$ degrees of freedom of the full tensor product space $V_J \otimes V_J$. However, the approximation power of the sparse tensor product space and the full tensor product space are essentially (i.e., except for logarithmic factors) identical if extra smoothness in terms of Sobolev spaces with dominating mixed derivative is given [5].

4.3 Sparse Multilevel Frames

In the meantime, the construction of wavelets on fairly general domains and surfaces is well understood [24, 25, 42]. However, the construction is expensive and the wavelets have large supports, particularly on complicated geometries. Therefore, other sparse tensor product approximations have been developed. In [17, 27], the sparse tensor product approximation has been performed via multilevel frames. The frame construction is based on the BPX-preconditioner (see e.g. [3, 10, 35]) and related generating systems (see e.g. [16, 17, 19, 20]).

By rewriting the sparse tensor product space (13) according to

$$\hat{V}_J = \sum_{j+j' \le J} V_j \otimes V_{j'}$$

it is obvious that the collection of tensor products of the basis functions in $\{V_j\}_{j=0}^J$ can be used to represent the functions in $\hat{V}_J$. It has been shown in [27] that this collection forms a frame for the sparse tensor product space provided that the basis functions are appropriately normalized.

The discretization of boundary value problems by frames and the solution of operator equations in frame coordinates is well understood and quite similar to the basis case, see e.g. [8, 9, 41]. The algorithms developed in [38], especially the applications of tensor product operators, can be extended to multilevel frames. It turns out that, in order to efficiently solve boundary value problems of the type (11), it suffices to provide standard multigrid hierarchies and associated finite elements together with prolongations and restrictions, see [22, 27].

4.4 Combination Technique

Consider the tensor product boundary value problem (11). With respect to the ansatz spaces (12), we define the associated complement spaces by

$$W_j := (P_j - P_{j-1})\mathscr{H}(D) \subset V_j$$

with $P_j : \mathscr{H}(D) \to V_j$ being the Galerkin projection associated with the operator $\mathscr{A}$. Then, the Galerkin system decouples due to Galerkin orthogonality. Namely, it holds

$$\big((\mathscr{A} \otimes \mathscr{A})v_{i,i'}, w_{j,j'}\big)_{L^2(D\times D)} = 0 \text{ for all } v_{i,i'} \in W_i \otimes W_{i'},\ w_{j,j'} \in W_j \otimes W_{j'}$$

provided that $i \neq j$ or $i' \neq j'$. As a consequence, the Galerkin solution $\widehat{\mathrm{Cor}}_{\delta u,J}$ to (11) in the sparse tensor product space (13) can be written as

$$\widehat{\mathrm{Cor}}_{\delta u,J} = \sum_{j=0}^{J}(p_{j,J-j} - p_{j,J-j-1}) \in \bigoplus_{j=0}^{J} V_j \otimes W_{J-j} = \hat{V}_J$$

where $p_{j,j'}$ denotes the Galerkin solution of (11) in the full (but small) tensor product space $V_j \otimes V_{j'}$, cf. [28]. If the differential operator has not the form (11), then the combination technique induces an approximation error. Related error estimates have been derived in [21, 30, 36].

4.5 Low-Rank Approximation

A rank-r approximation of a given function $\mathrm{Cor}_f \in L^2(D \times D)$ is defined by

$$\mathrm{Cor}_f(\mathbf{x},\mathbf{y}) \approx \mathrm{Cor}_{f,r}(\mathbf{x},\mathbf{y}) := \sum_{\ell=1}^{r} a_\ell(\mathbf{x}) b_\ell(\mathbf{y})$$

with certain functions $a_\ell, b_\ell \in L^2(D)$. Inserting such a low-rank approximation in the tensor product boundary value problem (11) leads to the representation

$$\mathrm{Cor}_{\delta u} = \big(\mathscr{A}^{-1} \otimes \mathscr{A}^{-1}\big)\,\mathrm{Cor}_f \approx \big(\mathscr{A}^{-1} \otimes \mathscr{A}^{-1}\big)\,\mathrm{Cor}_{f,r} = \sum_{\ell=1}^{r} \big(\mathscr{A}^{-1} a_\ell\big) \otimes \big(\mathscr{A}^{-1} b_\ell\big),$$

i.e., the tensor product boundary value problem is reduced to $2r$ simple boundary value problems on the domain D.

This approach has firstly been proposed in [15] for m-fold tensor product problems and right hand sides of tensor product type. In the case of the second moment analysis in uncertainty quantification, we find the special situation that Cor_f is symmetric and positive semi-definite. Thus, the pivoted Cholesky decomposition can be used to efficiently compute the low-rank approximation to the right hand side, see [23, 29].

5 Finite Element Discretization

5.1 Parametric Finite Elements

For the application of multilevel techniques, we shall define a nested sequence of finite dimensional trial spaces

$$V_0 \subset V_1 \subset \cdots \subset V_j \subset \cdots \subset H^1(\overline{D}). \tag{14}$$

In general, due to our smoothness assumptions on the domain, we have to deal with non-polygonal domains. To realize the multiresolution analysis (14) we will use parametric finite elements.

Let $\triangle$ denote the reference simplex in $\mathbb{R}^n$. We assume that the domain $\overline{D}$ is partitioned into a finite number of patches

$$\mathrm{clos}(\overline{D}) = \bigcup_k \tau_{0,k}, \quad \tau_{0,k} = \gamma_k(\triangle), \quad k = 1, 2, \ldots, M,$$

where each $\gamma_k : \triangle \to \tau_{0,k}$ defines a diffeomorphism of $\triangle$ onto $\tau_{0,k}$. The intersection $\tau_{0,k} \cap \tau_{0,k'}$, $k \neq k'$, of the patches $\tau_{0,k}$ and $\tau_{0,k'}$ is either $\emptyset$, or a

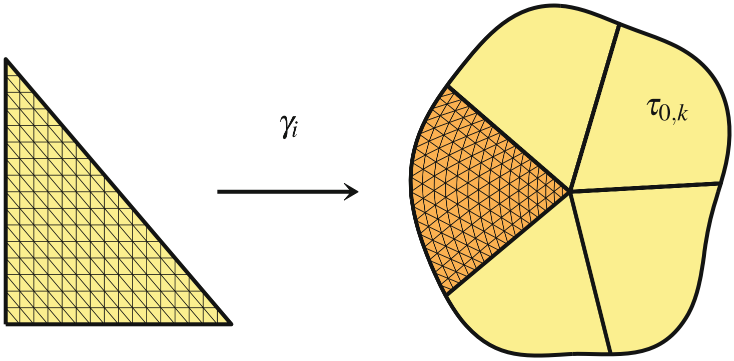


Fig. 1 Construction of parametric finite elements

lower dimensional face. The parametric representation is supposed to be globally continuous which means that the diffeomorphisms γ_i and $\gamma_{i'}$ coincide at common patch interfaces except for orientation. A mesh of level j on $\overline{D}$ is then induced by regular subdivisions of depth j of $\triangle$ into 2^{jn} simplices. This generates the $2^{jn}M$ curved elements $\{\tau_{j,k}\}$. An illustration of such a triangulation is found in Fig. 1.

The ansatz functions $\Phi_j = \{\varphi_{j,k} : k \in \Delta_j\}$ are finally defined via parameterization, lifting continuous piecewise linear Lagrangian finite elements from $\triangle$ to the domain $\overline{D}$ by using the mappings γ_i and gluing across patch boundaries. Setting $V_j = \operatorname{span} \Phi_j$ yields (14), where $\dim V_j \sim 2^{jn}$.

5.2 Galerkin Discretization

We shall be concerned with Galerkin's method for solving the variational problems (8)–(10). To this end, we define first the system matrix

$$\mathbf{A}_j := (\nabla\Phi_j, \nabla\Phi_j)_{L^2(\overline{D})} + (\alpha\Phi_j, \Phi_j)_{L^2(\partial\overline{D})}. \tag{15}$$

Then, the Galerkin solution

$$\overline{u}_j = \sum_{k\in\Delta_j} u_{j,k}\varphi_{j,k} = \Phi_j\mathbf{u}_j \in V_j$$

of the variational formulation (8) is derived from the linear system of equations

$$\mathbf{A}_j\mathbf{u}_j = \mathbf{f}_j, \quad \text{where} \quad \mathbf{f}_j := (f, \Phi_j)_{L^2(\overline{D})} + (g, \Phi_j)_{L^2(\partial\overline{D})}. \tag{16}$$

The solution of this Eq. (16) by multigrid accelerated finite element methods is straightforward and along the lines of the standard literature, see e.g. [2, 4]. Therefore, we will skip all the details here.

The shape derivative $\delta u = \delta u[\kappa]$, given by (9), is approximated in a similar way: we seek

$$\delta u_j = \sum_{k \in \Delta_j} v_{j,k} \varphi_{j,k} = \Phi_j \mathbf{v}_j \in V_j$$

such that

$$\mathbf{A}_j \mathbf{v}_j = \mathbf{g}_j, \quad \text{where} \quad \mathbf{g}_j := (\kappa h, \Phi_j)_{L^2(\partial \overline{D})} + (\kappa \nabla_\Gamma \overline{u}, \nabla_\Gamma \Phi_j)_{L^2(\partial \overline{D})}. \tag{17}$$

Likewise to the mean field equation, the solution of (17) is straightforward.

For the combination technique, we need to compute certain Galerkin approximations

$$p_{j,j'} = \sum_{k \in \Delta_j} \sum_{k' \in \Delta_{j'}} w_{(j,k),(j',k')} (\varphi_{j,k} \otimes \varphi_{j',k'}) = (\Phi_j \otimes \Phi_{j'}) \mathbf{w}_{j,j'}$$

to the two-point correlation $\operatorname{Cor}_{\delta u}$ (10) in full tensor product spaces $V_j \otimes V_{j'}$. They are obtained from the following linear system of equations

$$(\mathbf{A}_j \otimes \mathbf{A}_{j'}) \mathbf{w}_{j,j'} = \mathbf{h}_{j,j'}. \tag{18}$$

Here, the right hand side is given by

$$\begin{aligned} \mathbf{h}_{j,j'} := & \big(\operatorname{Cor}_\kappa(h \otimes h), \Phi_j \otimes \Phi_{j'}\big)_{L^2(\partial\overline{D} \times \partial\overline{D})} \\ & - \big(\operatorname{Cor}_\kappa(\nabla_\Gamma \overline{u} \otimes h), \nabla_\Gamma \Phi_j \otimes \Phi_{j'}\big)_{L^2(\partial\overline{D} \times \partial\overline{D})} \\ & - \big(\operatorname{Cor}_\kappa(h \otimes \nabla_\Gamma \overline{u}), \Phi_j \otimes \nabla_\Gamma \Phi_{j'}\big)_{L^2(\partial\overline{D} \times \partial\overline{D})} \\ & + \big(\operatorname{Cor}_\kappa(\nabla_\Gamma \overline{u} \otimes \nabla_\Gamma \overline{u}), \nabla_\Gamma \Phi_j \otimes \nabla_\Gamma \Phi_{j'}\big)_{L^2(\partial\overline{D} \times \partial\overline{D})}. \end{aligned} \tag{19}$$

The iterative solution of the tensor product problem (18) is of optimal complexity if the tensor product of the BPX-preconditioner [3] is applied.

5.3 Implementation of the Combination Technique

According to Sect. 4.4, the combination technique amounts to solving all the Galerkin systems (18) which are needed to determine the expression

$$\widehat{\operatorname{Cor}}_{\delta u, J} = \sum_{j=0}^{J} (p_{j,J-j} - p_{j,J-j-1}) \in \hat{V}_J.$$

For the implementation of the combination technique, we have thus to explain how to efficiently compute the right hand side (19) to the linear system of equations (18). To this end, we shall introduce some notation first.

Let the index set $\Delta_j^{\partial\overline{D}} \subset \Delta_j$ denote the indices which belong to finite element functions at the boundary $\partial\overline{D}$ and set $\varphi_{j,k}^{\partial\overline{D}} := \varphi_{j,k}|_{\partial\overline{D}}$ for all $k \in \Delta_j^{\partial\overline{D}}$. Then, setting $\nabla_0^{\partial\overline{D}} := \Delta_0^{\partial\overline{D}}$ and $\nabla_j := \Delta_j^{\partial\overline{D}} \setminus \Delta_{j-1}^{\partial\overline{D}}$ for $j > 0$, the hierarchical basis in the trace space $V_J|_{\partial\overline{D}}$ is given by $\bigcup_{j=0}^{J}\{\varphi_{j,k}^{\partial\overline{D}}\}_{k\in\nabla_j^{\partial\overline{D}}}$. We replace the two-point correlation function Cor_κ by its piecewise linear sparse grid interpolant

$$\widehat{\mathrm{Cor}}_{\kappa,J} = \sum_{j+j'\leq J} \sum_{k\in\nabla_j^{\partial\overline{D}}} \sum_{k'\in\nabla_{j'}^{\partial\overline{D}}} \gamma_{(j,k),(j',k')}\big(\varphi_{j,k}^{\partial\overline{D}} \otimes \varphi_{j',k'}^{\partial\overline{D}}\big) \subset \hat{V}_J|_{\partial\overline{D}\times\partial\overline{D}}$$

which can be computed in optimal complexity (see [5]). Thus, the right hand side $\mathbf{h}_{j,j'}$ becomes

$$\mathbf{h}_{j,j'} = \sum_{\ell+\ell'\leq J} (\mathbf{M}_{j,\ell} \otimes \mathbf{M}_{j',\ell'})[\gamma_{(\ell,k),(\ell',k')}]_{k\in\nabla_j^{\partial\overline{D}},k'\in\nabla_{j'}^{\partial\overline{D}}} \tag{20}$$

where the matrices $\mathbf{M}_{j,j'}$, $0 \leq j, j' \leq J$, are given by

$$\mathbf{M}_{j,j'} = \Big[\big(\varphi_{j',k'}^{\partial\overline{D}} h, \varphi_{j,k}\big)_{L^2(\partial\overline{D})} + \big(\varphi_{j',k'}^{\partial\overline{D}} \nabla_\Gamma \bar{u}, \nabla_\Gamma \varphi_{j,k}\big)_{L^2(\partial\overline{D})}\Big]_{k\in\Delta_j,k'\in\nabla_{j'}^{\partial\overline{D}}}.$$

The expression (20) can be evaluated in essentially optimal complexity by applying the matrix-vector multiplication from [27]. In particular, by using prolongations and restrictions, the matrices $\mathbf{M}_{j,j'}$ are needed only in the situation $j = j'$. Thus, the over-all computational complexity of the combination technique is essentially linear in the number $|\Delta_J|$ of finite element functions on $\overline{D}$.

5.4 Implementation of the Low-Rank Approximation

The piecewise linear interpolant of the two-point correlation Cor_κ in the trace space $(V_j \otimes V_j)|_{\partial\overline{D}\times\partial\overline{D}}$ is given by

$$\mathrm{Cor}_{\kappa,j} = \sum_{k,k'\in\Delta_j^{\partial\overline{D}}} \mathrm{Cor}_\kappa(\mathbf{x}_{j,k}, \mathbf{x}_{j',k'})\big(\varphi_{j,k}^{\partial\overline{D}} \otimes \varphi_{j,k'}^{\partial\overline{D}}\big).$$

Here, $\mathbf{x}_{j,k} \in \partial\overline{D}$ denotes the node which belongs to the finite element basis function $\varphi_{j,k}^{\partial\overline{D}} \in V_j|_{\partial\overline{D}}$. We shall thus compute a low-rank approximation of the matrix

$$\mathbf{C} = [\mathrm{Cor}_\kappa(\mathbf{x}_{j,k}, \mathbf{x}_{j',k'})]_{k,k' \in \Delta_j^{\partial \overline{D}}} \approx \mathbf{C}_r = \sum_{i=1}^{r} \kappa_i \kappa_i^T \tag{21}$$

by the pivoted Cholesky decomposition. Afterwards, we just have to compute all the local shape derivatives δu in the directions $\sum_{k \in \Delta_j^{\partial \overline{D}}} \kappa_{i,k} \varphi_{j,k}^{\partial \overline{D}}$ via (17). Thus, having the low-rank approximation (21) at hand, the complexity to compute $\mathrm{Cor}_{\delta u, j}$ is $\mathscr{O}(r|\Delta_j|)$. Note here that, in accordance with [18, 39], the rank r hinges on the smoothness of the underlying two-point correlation Cor_κ.

The pivoted Cholesky decomposition is a purely algebraic approach which is quite simple to implement, see Algorithm 1. It produces a low-rank approximation of $\mathbf{C}$ for any given precision $\varepsilon > 0$ where the approximation error is rigorously controlled in the trace norm. $n = |\Delta_j^{\partial \overline{D}}|$. A rank-$r$ approximation is computed

Algorithm 1: Pivoted Cholesky decomposition

```
Data: matrix $\mathbf{C} = [c_{i,j}] \in \mathbb{R}^{n \times n}$ and error tolerance $\varepsilon > 0$
Result: low-rank approximation $\mathbf{C}_m = \sum_{i=1}^{m} \boldsymbol{\ell}_i \boldsymbol{\ell}_i^T$ such that trace$(\mathbf{C} - \mathbf{C}_m) \leq \varepsilon$
begin
    set $m := 1$;
    set $\mathbf{d} := \mathrm{diag}(\mathbf{C})$ and $error := \|\mathbf{d}\|_1$;
    initialize $\pi := (1, 2, \ldots, n)$;
    while $error > \varepsilon$ do
        set $i := \arg\max\{d_{\pi_j} : j = m, m+1, \ldots, n\}$;
        swap $\pi_m$ and $\pi_i$;
        set $\ell_{m,\pi_m} := \sqrt{d_{\pi_m}}$;
        for $m + 1 \leq i \leq n$ do
            compute $\ell_{m,\pi_i} := \Big(c_{\pi_m,\pi_i} - \sum_{j=1}^{m-1} \ell_{j,\pi_m} \ell_{j,\pi_i}\Big) \Big/ \ell_{m,\pi_m}$;
            update $d_{\pi_i} := d_{\pi_i} - \ell_{m,\pi_i}^2$;
        compute $error := \sum_{i=m+1}^{n} d_{\pi_i}$;
        increase $m := m + 1$;
```

in $\mathscr{O}(r^2 n)$ operations, where n denotes the matrix dimensions, that is Exponential convergence rates in r are proven under the assumption that the eigenvalues of $\mathbf{C}$ exhibit a sufficiently fast exponential decay, see [29]. Numerical experiments given there show that the pivoted Cholesky decomposition in general converges optimally in the sense that the rank r is bounded by the number of terms required for the spectral decomposition of $\mathbf{C}$ to achieve the error ε.

6 Numerical Results

6.1 Model Verification

We present some numerical tests to demonstrate our theoretical predictions. Let $\overline{D} = \{\mathbf{x} \in \mathbb{R}^2 : \|\mathbf{x}\| < 1\}$ be the unit disk. We parametrize the boundary $\partial\overline{D}$ by polar coordinates

$$\overline{\gamma} : [0, 2\pi] \to \partial\overline{D}, \quad s \mapsto \overline{\gamma}(s) := \begin{bmatrix} \cos(s) \\ \sin(s) \end{bmatrix}.$$

Correspondingly, the boundary $\partial D_\varepsilon(\omega)$ of the random domain $D_\varepsilon(\omega)$ can be expressed via the perturbed parametrization

$$\gamma(s, \omega) := \overline{\gamma}(s) + \varepsilon\kappa(s, \omega) \begin{bmatrix} \cos(s) \\ \sin(s) \end{bmatrix}.$$

Herein, we assume that the random perturbation is given by

$$\kappa(s, \omega) := \sum_{k=0}^{5} a_k(\omega)\cos(ks) + b_k(\omega)\sin(ks)$$

with random coefficients $a_k(\omega)$ and $b_k(\omega)$ which are equally distributed in $[-1, 1]$ and mutually stochastically independent. This results in the two-point correlation function

$$\mathrm{Cor}_\kappa(s, t) = \frac{1}{3}\sum_{k=0}^{5} \cos(ks)\cos(kt) + \sin(ks)\sin(kt). \tag{22}$$

For our numerical experiments, we vary $0 \le \varepsilon \le 0.05$. Even though ε is small, the perturbation is considerably large since the norm $\|\kappa(\omega)\|_{C^{2,1}([0,2\pi])}$ might become large.

On the above defined random domain $D_\varepsilon(\omega)$, we consider the Robin boundary value problem (1) with $f(\mathbf{x}) \equiv 1$, $\alpha(\mathbf{x}) \equiv 1$, and $g(\mathbf{x}) \equiv 0$. For a given value of ε, we determine first the expectation and the variance of the random solution by a Monte Carlo method, using $M = 25{,}000$ samples. Note that the triangulation hast to be constructed for each sample in order to resolve the random domain. To evaluate the sample mean and variance, we interpolate each solution to a fixed quadrangular grid on the disk $K = \{\mathbf{x} \in \mathbb{R}^2 : \|\mathbf{x}\| \le 0.7\}$ with radius 0.7 which lies always in the interior of the random domain $D_\varepsilon(\omega)$. The result of the Monte Carlo simulation is then compared with the solution of our deterministic model. Here, we used the pivoted Cholesky decomposition since the two-point correlation (22) is of finite rank $r = 11$.

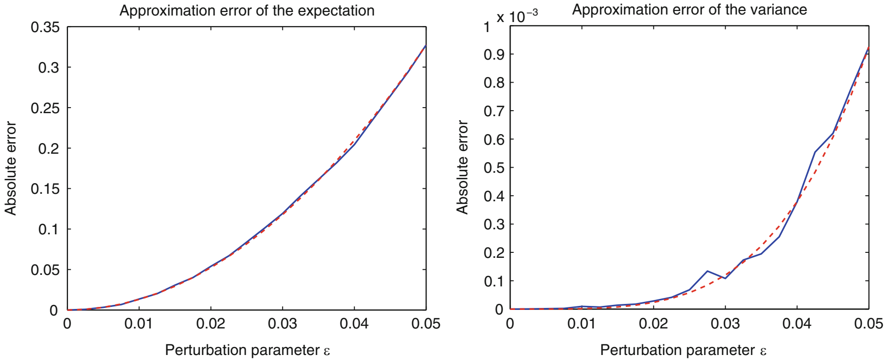


Fig. 2 Asymptotic behaviour with respect to the perturbation parameter ε in the case of the expectation (*left plot*) and in the case of the variance (*right plot*)

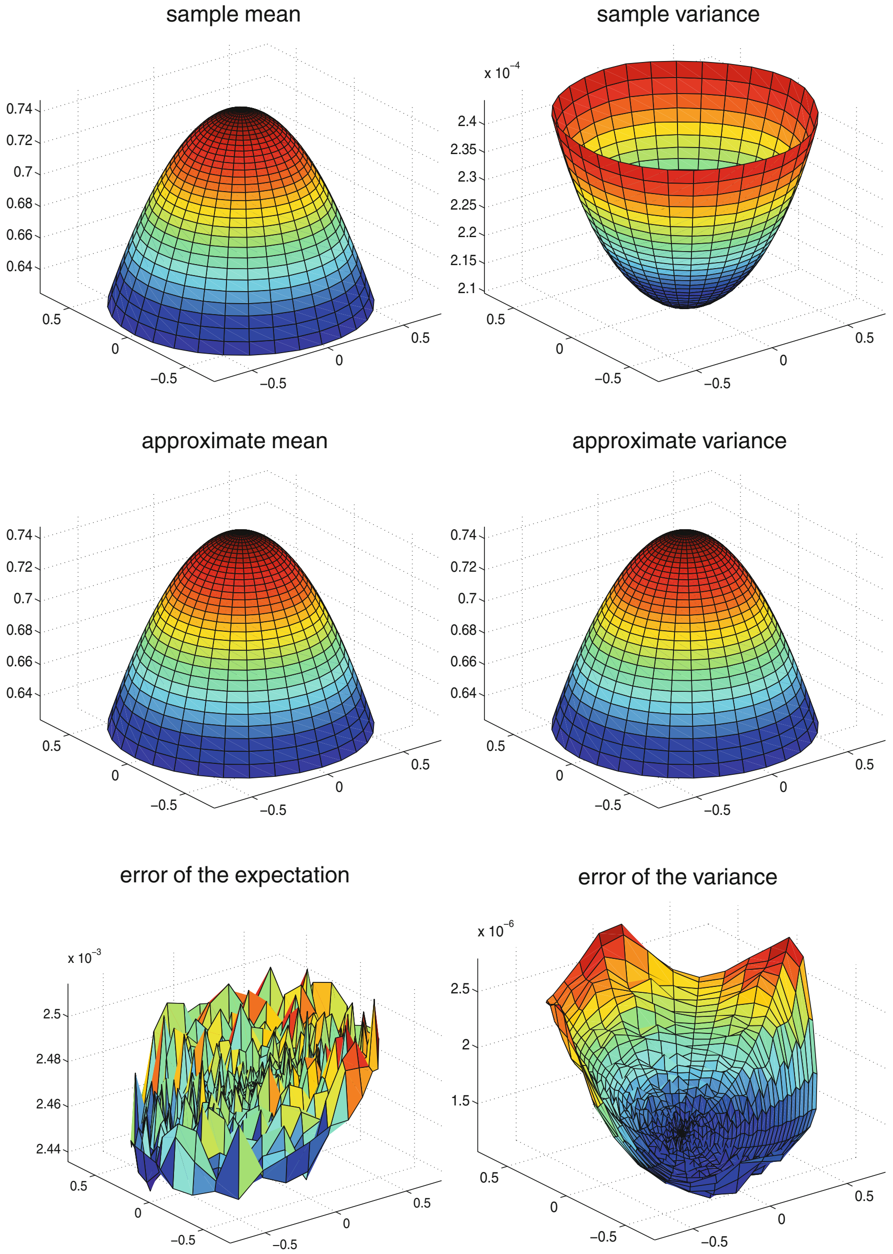
In Fig. 2, one finds the absolute difference between the mean (left plot) and variance (right plot) of the Monte Carlo simulation and the deterministic approach. To be on save ground, we repeated the comparison five times and computed the average of the differences. We observe that the difference behaves like $\mathscr{O}(\varepsilon^2)$ for the expectation (left plot) and like $\mathscr{O}(\varepsilon^4)$ for the variance (right plot) as indicated by the dashed lines. Hence, in this example, the asymptotic behaviour of the expectation with respect to the perturbation parameter ε is as predicted by Theorem 1. But the asymptotic behaviour of the variance with respect to the perturbation parameter ε is even one order better than predicted.

In Fig. 3, we visualized the approximate moments computed by the Monte Carlo simulation (first row of Fig. 3) and by the deterministic approach (second row of Fig. 3) in the specific case $\varepsilon = 0.025$. The difference between both approaches are found in the last row of Fig. 3. The relative difference in the mean has the order of magnitude 10^{-3} while the relative difference in the variance has the order of magnitude 10^{-2}.

6.2 A Correlation Kernel of Arbitrary Smoothness

We shall next compare the low-rank approximation with the combination technique based sparse grid approach. To this end, we choose the same input data as before, but employ the Gaussian kernel

$$k(r) = \exp\left(-\frac{r^2}{\ell^2}\right), \quad r = \|\gamma(s) - \gamma(t)\|$$


sample mean

sample variance

approximate mean

approximate variance

error of the expectation

error of the variance

Fig. 3 Sample mean and variance (*first row*) versus the deterministic mean and variance (*second row*) in the case of $\varepsilon = 0.025$. The differences are found in the *last row*

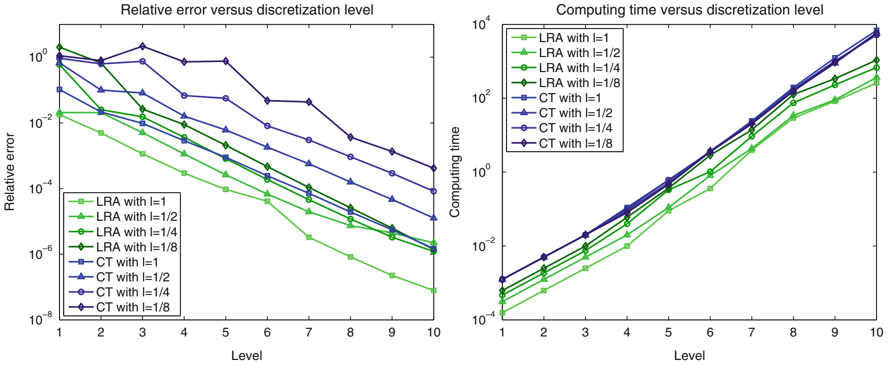


Fig. 4 Accuracy (*left plot*) and computing times (*right plot*) in the case of the Gaussian kernel

instead of the kernel (22). The Gaussian kernel is of arbitrary smoothness for any given correlation length $\ell > 0$. In particular, the eigenvalues of the associated Hilbert-Schmidt operator decay double-exponentially (see e.g. [39]). In our numerical tests, we vary the correlation length according to $\ell = 1, 1/2, 1/4, 1/8$.

We compute a reference solution on a very fine level and compare the solutions of both approaches with respect to lower levels with this reference solution. The results are plotted in Fig. 4, where the left plot shows the relative error of the variance versus the discretization level and the right plot shows the related computing times versus the discretization level. Note that on level 10, there are about two million finite elements.

It is observed that both, the convergence rates (left plot of Fig. 4) and the computing times (right plot of Fig. 4), scale identically for both approaches. The relative errors of both approaches increase when the correlation length decreases. The approximation errors of the low-rank approximation (green lines) are, however, a certain factor lower than the related approximation errors of the sparse grid method (blue lines). Also the computing times of the low-rank approximation (green lines) are a certain factor lower than the related computing times of the sparse grid approach (blue lines). Nevertheless, the computing times with respect to the sparse grid approach are essentially independent of the correlation length ℓ while the computing times of the low-rank approximation increase in ℓ as the rank increases.

6.3 A Correlation Kernel of Finite Smoothness

We finally compare the low-rank approximation with the combination technique in case of the Matérn kernel

$$k_{3/2}(r) = \left(1 + \frac{\sqrt{3}r}{\ell}\right) \exp\left(-\frac{\sqrt{3}r}{\ell}\right), \quad r = \|\gamma(s) - \gamma(t)\|$$

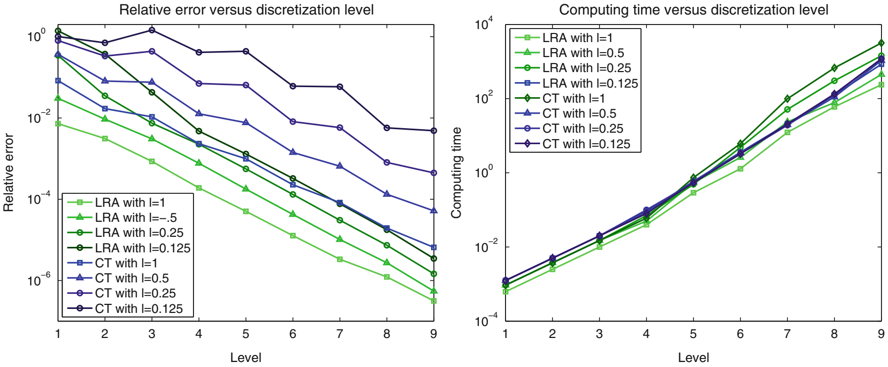


Fig. 5 Accuracy (*left plot*) and computing times (*right plot*) in the case of the Matérn kernel

which is of finite smoothness. The correlation length ℓ is again chosen to be $\ell = 1, 1/2, 1/4, 1/8$. The computational set-up of our comparison is in complete analogy to that of Sect. 6.2.

In the left plot of Fig. 5, we plotted the relative error of the variance versus the discretization level. Again, both approaches seem to produce the same convergence rates but the relative errors of the low-rank approximation (green lines) are again a certain factor lower than relative error of the sparse grid approach (blue lines). Moreover, for a fixed discretization level, the relative error increases as the correlation length decreases.

In the right plot of Fig. 5, the associated computing times are found. The computing times of the low-rank approximation (green lines) clearly depend on the correlation length, whereas, in the case of the sparse grid approach, the computing times are independent of the correlation length. Additionally, one figures out of the plot that the computing times of the low-rank approximation seem to grow with a higher rate compared with the sparse grid approach. This corresponds to the theoretical predictions from [18]. Nevertheless, if one compares accuracy versus computing time, the low-rank approximation is still superior to the sparse grid approach.

7 Concluding Remarks

In this paper, we modeled and solved Robin boundary value problems on random domains. We derived deterministic equations for the expectation and variance of the associated random solution. The variance can be computed by means of a low-rank approximation or by the combination technique. By numerical experiments, we compare these two approaches. It turns out that for our specific examples the low-rank approximation performs better than the combination technique. However, the combination technique has the advantage that the memory requirements are independent of the given two-point correlation function. We emphasize that, in the present case of boundary value problems on random domains, the low-rank approximation needs only to be computed for an $(n-1)$-dimensional function (cf. (21)) whereas the combination technique is an n-dimensional approach. Nevertheless, we expect that, in the case of random coefficients (see [28]) or random loadings (see [38]), the combination technique performs much better in comparison with the low-rank approximation since there the low-rank approximation of an n-dimensional function is required.

References

1. Babuška, I., Nobile, F., Tempone, R.: Worst case scenario analysis for elliptic problems with uncertainty. Numer. Math. **101**, 185–219 (2005)
2. Braess, D.: Finite Elements. Theory, Fast Solvers, and Applications in Solid Mechanics, 2nd edn. Cambridge University Press, Cambridge (2001)
3. Bramble, J., Pasciak, J., Xu, J.: Parallel multilevel preconditioners. Math. Comput. **55**, 1–22 (1990)
4. Brenner, S.C., Scott, L.R.: The Mathematical Theory of Finite Element Methods. Texts in Applied Mathematics, vol. 15. Springer, New York (1994)
5. Bungartz, H.J., Griebel, M.: Sparse grids. Acta Numer. **13**, 147–269 (2004)
6. Canuto, C., Kozubek, T.: A fictitious domain approach to the numerical solution of PDEs in stochastic domains. Numer. Math. **107**, 257–293 (2007)
7. Chernov, A., Schwab, C.: First order k-th moment finite element analysis of nonlinear operator equations with stochastic data. Math. Comput. **82**, 1859–1888 (2013)
8. Christiansen, O.: An Introduction to Frames and Riesz Bases. Applied and Numerical Harmonic Analysis. Birkhäuser, Boston (2002)
9. Dahlke, S., Fornasier, M., Raasch, T., Stevenson, R., Werner, M.: Adaptive frame methods for elliptic operator equations. The steepest descent approach. IMA J. Numer. Math. **27**, 717–740 (2007)
10. Dahmen, W.: Wavelet and multiscale methods for operator equations. Acta Numer. **6**, 55–228 (1997)
11. Deb, M.K., Babuška, I., Oden, J.T.: Solution of stochastic partial differential equations using Galerkin finite element techniques. Comput. Methods Appl. Mech. Eng. **190**, 6359–6372 (2001)
12. Delfour, M., Zolesio, J.-P.: Shapes and Geometries. SIAM, Philadelphia (2001)
13. Frauenfelder, P., Schwab, C., Todor, R.A.: Finite elements for elliptic problems with stochastic coefficients. Comput. Methods Appl. Mech. Eng. **194**, 205–228 (2004)
14. Ghanem, R.G., Spanos, P.D.: Stochastic Finite Elements. A Spectral Approach. Springer, New York (1991)
15. Grasedyck, L.: Existence and computation of a low Kronecker-rank approximant to the solution of a tensor system with tensor right-hand side. Computing **72**, 247–265 (2004)
16. Griebel, M.: Multilevel algorithms considered as iterative methods on semidefinite systems. SIAM J. Sci. Comput. **15**, 547–565 (1994)
17. Griebel, M.: Multilevelmethoden als Iterationsverfahren über Erzeugendensystemen. Teubner Skripten zur Numerik. B.G. Teubner, Stuttgart (1994)
18. Griebel, M., Harbrecht, H.: Approximation of bivariate functions: singular value decomposition versus sparse grids. IMA J. Numer. Anal. (2013, to appear)
19. Griebel, M., Oswald, P.: On additive Schwarz preconditioners for sparse grid discretizations. Numer. Math. **66**, 449–463 (1994)

20. Griebel, M., Oswald, P.: Tensor product type subspace splittings and multilevel iterative methods for anisotropic problems. Adv. Comput. Math. **4**, 171–206 (1995)
21. Griebel, M., Schneider, M., Zenger, C.: A combination technique for the solution of sparse grid problems. In: de Groen, P., Beauwens, R. (eds.) Iterative Methods in Linear Algebra. IMACS, pp. 263–281. Elsevier, North Holland (1992)
22. Harbrecht, H.: A finite element method for elliptic problems with stochastic input data. Appl. Numer. Math. **60**, 227–244 (2010)
23. Harbrecht, H., Li, J.: First order second moment analysis for stochastic interface problems based on low-rank approximation. ESAIM Math. Model. Numer. Anal. (2013, to appear)
24. Harbrecht, H., Schneider, R.: Biorthogonal wavelet bases for the boundary element method. Math. Nachr. **269–270**, 167–188 (2004)
25. Harbrecht, H., Stevenson, R.: Wavelets with patchwise cancellation properties. Math. Comput. **75**, 1871–1889 (2006)
26. Harbrecht, H., Schneider, R., Schwab, C.: Sparse second moment analysis for elliptic problems in stochastic domains. Numer. Math. **109**, 167–188 (2008)
27. Harbrecht, H., Schneider, R., Schwab, C.: Multilevel frames for sparse tensor product spaces. Numer. Math. **110**, 199–220 (2008)
28. Harbrecht, H., Peters, M., Siebenmorgen, M.: Combination technique based k-th moment analysis of elliptic problems with random diffusion. J. Comput. Phys. **252**, 128–141 (2013)
29. Harbrecht, H., Peters, M., Schneider, R.: On the low-rank approximation by the pivoted Cholesky decomposition. Appl. Numer. Math. **62**, 428–440 (2012)
30. Hegland, M., Garcke, J., Challis, V.: The combination technique and some generalisations. Linear Algebra Appl. **420**, 249–275 (2007)
31. Hiptmair, R., Li, J.: Shape derivatives in differential forms I. An intrinsic perspective. Ann. Mat. Pura Appl. (2012, to appear)
32. Matthies, H.G., Keese, A.: Galerkin methods for linear and nonlinear elliptic stochastic partial differential equations. Comput. Methods Appl. Mech. Eng. **194**, 1295–1331 (2005)
33. Mohan, P.S., Nair, P.B., Keane, A.J.: Stochastic projection schemes for deterministic linear elliptic partial differential equations on random domains. Int. J. Numer. Methods Eng. **85**, 874–895 (2011)
34. Nouy, A., Chevreuil, M., Safatly, E.: Fictitious domain method and separated representations for the solution of boundary value problems on uncertain parameterized domains. Comput. Methods Appl. Mech. Eng. **200**, 3066–3082 (2011)
35. Oswald, P.: Multilevel Finite Element Approximation. Theory and Applications. Teubner Skripten zur Numerik. B.G. Teubner, Stuttgart (1994)
36. Pflaum, C., Zhou, A.: Error analysis of the combination technique. Numer. Math. **84**, 327–350 (1999)
37. Protter, P.: Stochastic Integration and Differential Equations. A New Approach. Springer, Berlin (1990)
38. Schwab, C., Todor, R.: Sparse finite elements for elliptic problems with stochastic loading. Numer. Math. **95**, 707–734 (2003)
39. Schwab, C., Todor, R.: Karhunen-Loéve approximation of random fields by generalized fast multipole methods. J. Comput. Phys. **217**, 100–122 (2006)
40. Sokolowski, J., Zolesio, J.-P.: Introduction to Shape Optimization. Springer, Berlin (1992)
41. Stevenson, R.: Adaptive solution of operator equations using wavelet frames. SIAM J. Numer. Anal. **41**, 1074–1100 (2003)
42. Stevenson, R.: Composite wavelet bases with extended stability and cancellation properties. SIAM J. Numer. Anal. **45**, 133–162 (2007)
43. Tartakovsky, D.M., Xiu, D.: Numerical methods for differential equations in random domains. SIAM J. Sci. Comput. **28**, 1167–1185 (2006)

Simulation of Q-Tensor Fields with Constant Orientational Order Parameter in the Theory of Uniaxial Nematic Liquid Crystals

Sören Bartels and Alexander Raisch

Abstract We propose a practical finite element method for the simulation of uniaxial nematic liquid crystals with a constant order parameter. A monotonicity result for Q-tensor fields is derived under the assumption that the underlying triangulation is weakly acute. Using this monotonicity argument we show the stability of a gradient flow type algorithm and prove the convergence outputs to discrete stable configurations as the stopping parameter of the algorithm tends to zero. Numerical experiments with singularities illustrate the performance of the algorithm. Furthermore, we examine numerically the difference of orientable and non-orientable stable configurations of liquid crystals in a planar two dimensional domain and on a curved surface. As an application, we examine tangential line fields on the torus and show that there exist orientable and non-orientable stable states with comparing Landau-de Gennes energy and regions with different tilts of the molecules.

1 Introduction and Derivation of the Mathematical Setting

The modelling of liquid crystals has attracted considerable attention among mathematicians in the last decade [1–4, 10, 12–14, 18]. Starting with the prediction of stationary configurations for the classical Oseen-Frank model we continue right through to studies of the motion of liquid crystals governed by the

S. Bartels (✉)
Abteilung für Angewandte Mathematik, Universität Freiburg, Hermann-Herder-Str. 10,
D-79104 Freiburg, Germany
e-mail: bartels@mathematik.uni-freiburg.de

A. Raisch
Institut für Numerische Simulation, Rheinische Friedrich-Wilhelms-Universität Bonn,
Wegelerstr. 6, D-53115 Bonn, Germany
e-mail: raisch@ins.uni-bonn.de

M. Griebel (ed.), *Singular Phenomena and Scaling in Mathematical Models*,
DOI 10.1007/978-3-319-00786-1_17,

Ericksen-Leslie model. In recent years it has become more popular to use Q-tensors to describe nematic liquid crystals. One of the main features of this theory is that it captures symmetries of the molecules which are not seen by classical models. In [4] this feature is examined analytically and examples are constructed to show that there are settings where the classical theory misses stationary configurations that are energetically more favorable for the liquid crystal. The analysis leads directly to topological issues and the question of the orientability of given line fields. It is the aim of this paper to devise numerical methods for both models and to understand relations between them. Following [4], the molecules of a nematic liquid crystal can be thought of as rod-like molecules with two ends indistinguishable from each other, a center of mass at a position $x \in \Omega$ and a certain direction in space. Here and in the rest of this report $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ is a bounded Lipschitz domain representing the vessel. For a well-defined macroscopic variable describing the crystal it is required to use statistical averages of the molecular orientation of the crystal. We let $\mathscr{L}(\mathbb{S}^2)$ denote the family of Lebesgue measurable subsets of the unit sphere $\mathbb{S}^2$, and assign to every point $x \in \Omega$ a probability measure $\mu(x, \cdot) : \mathscr{L}(\mathbb{S}^2) \to [0, 1]$ so that $\mu(x, \{n\})$ is the probability of the crystal x to point in direction n. Since the molecules admit the so-called head-to-tail symmetry we have that $\mu(x, A) = \mu(x, -A)$ for every $x \in \Omega$ and every $A \subset \mathbb{S}^2$. This property yields that all odd moments of μ must vanish. The lowest order even moment, which is assumed to be the most important quantity for describing liquid crystals, is given by

$$M_{ij}(x) = \int_{\mathbb{S}^2} p_i p_j \, \mathrm{d}\mu(x, p), \quad i, j = 1, 2, 3, \ x \in \Omega.$$

The matrix valued function $M : \Omega \to \mathbb{R}^{3\times 3}$ has the properties

$$M = M^T, \ M \geq 0 \text{ and } \operatorname{tr} M = 1.$$

We define the trace-free *de Gennes* order parameter tensor $Q := M - \frac{1}{3}\mathrm{id}$ and distinguish three different cases: (1) If Q has three equal eigenvalues then $Q = 0$ and we call the liquid crystal isotropic, that means, the orientation of molecules is totally random. (2) If Q has two equal eigenvalues, then the liquid crystal is called uniaxial and Q admits a representation of the form

$$Q = s(n \otimes n - \frac{1}{3}\mathrm{id}),$$

where $n \in \mathbb{S}^2$ is the optical axis and $s \in \mathbb{R}$ is the orientational order parameter. The orientational order s takes values between $s = -\frac{1}{2}$ (molecules are planar oriented and perpendicular to the optical axis) and $s = 1$ (perfect alignment of molecules with the optical axis). The uniaxial case is characteristic for nematics and cholesterics. (3) If Q has three distinct eigenvalues, then the liquid crystal is called biaxial. In practice, however, it is observed that liquid crystals are uniaxial almost everywhere with a constant order parameter s between 0.6 and 0.8. We therefore restrict ourselves to the case $Q(x) = s(n(x) \otimes n(x) - \frac{1}{3}\mathrm{id})$ with s constant and $n(x) \in \mathbb{S}^2$ for $x \in \Omega$. When we talk about the classical Oseen-Frank model we think of the liquid crystal being described simply by a director field $n : \Omega \to \mathbb{S}^2$, the optical axis. As in our simplified Q-tensor model we assume a constant orientational order parameter. More details for a substantial treatment of this derivation and an introduction to the classical model can be found in [13, 18].

A model to predict stable liquid crystal configurations is to compute stationary points of the energy

$$\begin{aligned} E_{OF}(n) &:= \int_\Omega W(n, \nabla n)\,\mathrm{d}x \\ &:= \int_\Omega k_1|\mathrm{div}\,n|^2 + k_2|n\cdot\mathrm{curl}\,n|^2 + k_3|n\times\mathrm{curl}\,n|^2 + \\ &\qquad (k_2+k_4)(|\nabla n|^2 - |\mathrm{div}\,n|^2)\,\mathrm{d}x, \end{aligned}$$

with elastic constants k_1, k_2, k_3, k_4. It is possible to choose a function Ψ, depending on the tensor $Q = s(n\otimes n - \frac{1}{3}\mathrm{id})$, so that the energy density W can be expressed as

$$W(n, \nabla n) = \Psi(Q, \nabla Q),$$

see [4] for details. We will refer to the constrained Landau-de Gennes theory when considering the energy density Ψ and de Gennes order parameter tensors. If $Q = s(n \otimes n - \frac{1}{3}\mathrm{id})$ almost everywhere in Ω with $n : \Omega \to \mathbb{S}^2$ then

$$E_{OF}(n) = \int_\Omega W(n, \nabla n)\,\mathrm{d}x = \int_\Omega \Psi(Q, \nabla Q)\,\mathrm{d}x =: E_{LdG}(Q),$$

and the Landau-de Gennes theory can be interpreted as a generalization of the classical Oseen-Frank model. In the most simple (equal constant) setting E_{OF} reduces to the standard Dirichlet energy for functions with values in $\mathbb{S}^2$ and the fact that

$$|\nabla(n\otimes n)|^2 = 2|\nabla n|^2$$

yields

$$|\nabla Q|^2 = s^2|\nabla(n\otimes n - \frac{1}{3}\mathrm{id})|^2 = s^2|\nabla(n\otimes n)|^2 = 2s^2|\nabla n|^2.$$

Thus, if $Q = s(n\otimes n - \frac{1}{3}\mathrm{id})$ in Ω then

$$\frac{1}{2}\int_\Omega |\nabla n|^2\,\mathrm{d}x = \frac{1}{4s^2}\int_\Omega |\nabla Q|^2\,\mathrm{d}x$$

and E_{LdG} is a multiple of the Dirichlet energy for functions with values in

$$\tilde{\mathbb{L}}^2 := \left\{ A \in \mathbb{R}^{3\times 3} : \exists n \in \mathbb{S}^2, \exists s \in [-1/2, 1] : A = s(n \otimes n - \frac{1}{3}\mathrm{id}) \right\}.$$

Since we are interested only in the Dirichlet energy it is convenient to set $s = 1$ and replace $\tilde{\mathbb{L}}^2$ by the submanifold

$$\mathbb{L}^2 := \left\{ A \in \mathbb{R}^{3\times 3} : \exists n \in \mathbb{S}^2, A = n \otimes n \right\}.$$

We observe that $\mathbb{L}^2$ can be identified with the real projective space $\mathbb{R}P^2 = \mathbb{S}/\pm$ using the map

$$b : \mathbb{L}^2 \to \mathbb{R}P^2, \;\; A = n \otimes n \mapsto \{n, -n\}.$$

It is possible to endow $\mathbb{L}^2$ with a Riemannian structure so that it is a Riemannian manifold. Throughout this work we refer to the Oseen-Frank energy as

$$E_{OF} : W^{1,2}(\Omega, \mathbb{S}^2) \to \mathbb{R}, \;\; n \mapsto \frac{1}{2} \int_\Omega |\nabla n|^2 \,\mathrm{d}x$$

and to the Landau-de Gennes energy as

$$E_{LdG} : W^{1,2}(\Omega, \mathbb{L}^2) \to \mathbb{R}, \;\; Q \mapsto \frac{1}{4} \int_\Omega |\nabla Q|^2 \,\mathrm{d}x.$$

We will call stationary points of E_{OF} *harmonic director fields* and stationary points of E_{LdG} will be called *Q harmonic tensor fields* or *harmonic line fields*.

The molecules of the liquid crystal tend to align themselves parallel to the boundary when they are in contact with other materials. These boundary conditions are often referred to as partial constraint or planar anchoring conditions. When the surface is worked in a special manner the liquid crystal aligns with the treatment and can be specified. In this case one speaks about strong or homeotropic anchoring conditions. In our two-dimensional simulation in Sect. 6 we will also allow for Neumann boundary conditions in parts of $\partial\Omega$. They are not motivated by the physics but simplify the computations and help us to underline the difference of the Oseen-Frank and the Landau-de Gennes theory.

Clearly every $n \in W^{1,2}(\Omega, \mathbb{S}^2)$ defines a map $Q = n \otimes n \in W^{1,2}(\Omega, \mathbb{L}^2)$. The interesting question is whether the converse statement holds in the sense that for $Q \in W^{1,2}(\Omega, \mathbb{L}^2)$ there exists $n \in W^{1,2}(\Omega, \mathbb{S}^2)$ such that $Q = n \otimes n$. This is true in some situations, cf. the left plot in Fig. 1, but in general this is not the case as can be seen in the right plot of Fig. 1. In the latter one we set $G_1 = (-1, 1) \times (-1, 0)$, $G_2 = B_1(0) \cap \{x_2 \geq 0\}$ and $G = G_1 \cup G_2 \setminus B_{1/2}(0)$. We define the field

$$n(x) = \begin{cases} (-x_2, x_1, 0) & \text{if } x \in G \cap \{x_2 \geq 0\} \\ (0, 1, 0) & \text{if } x \in G \cap \{x_2 < 0\}. \end{cases}$$

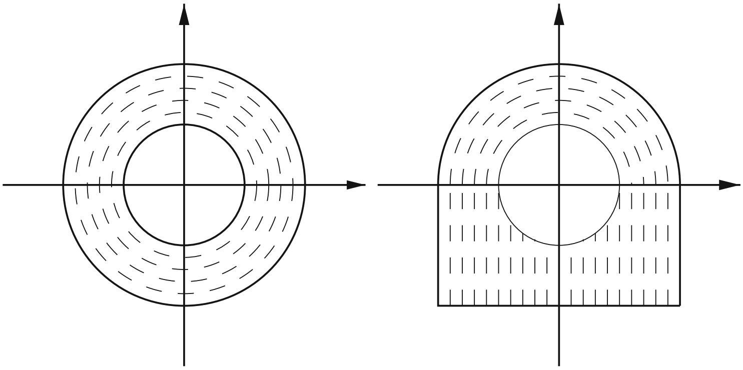


Fig. 1 Orientable and non-orientable line fields in the plane

Then $Q := n \otimes n \in W^{1,2}(\Omega, \mathbb{L}^2)$ and if $Q = \tilde{n} \otimes \tilde{n}$ then $\tilde{n} = \pm n$ almost everywhere but there is no way to construct a vector field $\tilde{n}$ without any jump in Ω and satisfying $\tilde{n}(x) = n(x)$ or $\tilde{n}(x) = -n(x)$ for almost every $x \in \Omega$. These observations motivate the following definition.

Definition 1. We say that a line field $Q \in W^{1,2}(\Omega, \mathbb{L}^2)$ is *orientable* if there exists $n \in W^{1,2}(\Omega, \mathbb{S}^2)$ such that $Q = n \otimes n$ a.e. in Ω. Otherwise Q is called *non-orientable.*

In the discrete setting we work with piecewise affine tensor fields Q_h that satisfy $Q_h(z) \in \mathbb{L}^2$ for all nodes z in the triangulation. Analogously we work with discrete vector fields n_h which are piecewise affine and satisfy $n_h(z) \in \mathbb{S}^2$. Thus, there always exists a discrete director field n_h such that $Q_h(z) = n_h(z) \otimes n_h(z)$ for all nodes z and for $h > 0$ fixed we have that $n_h \in W^{1,2}(G, \mathbb{R}^3)$. From this it follows that we can always assign a discrete director field to a discrete line field and every line field is orientable in a discrete sense but the identity $|\nabla Q_h|^2 = 2|\nabla n_h|^2$ may be violated. Thus, using the energies E_{LdG} and E_{OF} enables us to compare the two models and to introduce the notion of discrete orientable and non-orientable stable configurations. In Fig. 2 we depict a non-orientable line field in G and a possible discrete vector field n_h such that $Q_h(z) = n_h(z) \otimes n_h(z)$ for all nodes z. This discrete effect is reflected in the critical mesh-dependence of E_{OF}. The energies are $E_{LdG} \approx 0.9543$ and $E_{OF} \approx 13.7151$, thus, the jump in n_h contributes dramatically to the energy. The difference becomes even more dramatic when the mesh is refined reflecting the fact that there exists no continuous extension.

The outline of this work is as follows. In Sect. 2 we characterize the manifold $\mathbb{L}^2$ and derive Euler-Lagrange equations for E_{LdG}. In Sect. 3 we deduce a finite element discretization of E_{LdG} with pointwise constraints on the admissible functions. In Sect. 4 we propose an algorithm for the computation of Q harmonic line fields based on a gradient flow approach. Right after that we prove stability and convergence of the algorithm to a discrete Q harmonic line field in Sect. 5. Finally

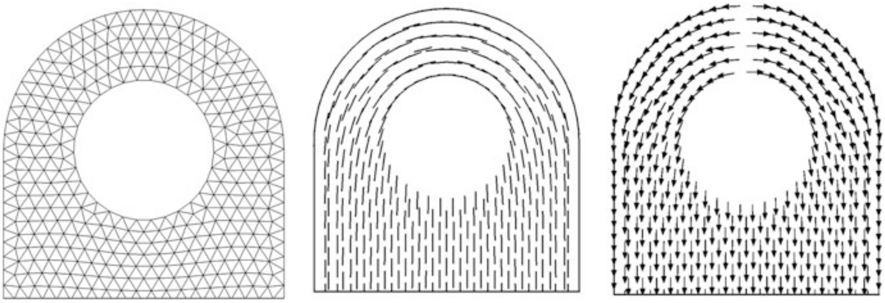


Fig. 2 Discrete line fields and vector fields and triangulation of G (*left*). The line field Q_h (*middle*) and a possible discrete director field n_h that satisfies $Q_h(z) = n_h(z) \otimes n_h(z)$ at the nodes z. The vectors and lines at the nodes are scaled by the factor $1/5$. Note that $E_{LdG}(Q_h) \approx 0.9543 \ll E_{OF}(n_h) \approx 13.7151$

in Sect. 6 we execute some interesting experiments illustrating the performance of our algorithm. Furthermore, we numerically examine orientability issues discussed in [4] in two and three space dimensions.

2 Euler Lagrange Equations for E_{LdG} and E_{OF}

We denote the tangent space of $\mathbb{L}^2$ at a given $Q_0 \in \mathbb{L}^2$ by $T_{Q_0}\mathbb{L}^2$ and for every tangent vector $V \in T_{Q_0}\mathbb{L}^2$ there exists a path $\gamma_0 : (-\delta, \delta) \to \mathbb{L}^2$ $(\delta > 0)$ satisfying $\gamma(0) = Q_0$ and $\frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0}\gamma(t) = V$. The following lemma can be found in [4] and will help us to establish a complete characterization of $T_{Q_0}\mathbb{L}^2$. We include a sketch of the proof to give an idea of how to work with line fields.

Lemma 1 ([4], Lemma 3). *If $-\infty < t_1 < t_2 < \infty$ and $Q : [t_1, t_2] \to \mathbb{L}^2$ is continuous then there exist exactly two continuous maps (liftings) $n^+, n^- : [t_1, t_2] \to \mathbb{S}^2$, so that $Q(t) = n^\pm(t) \otimes n^\pm(t)$ and $n^+ = -n^-$.*

Proof. Let $0 < \varepsilon < \sqrt{2}$. Given $n, \overline{m} \in \mathbb{S}^2$ with $|n \otimes n - \overline{m} \otimes \overline{m}| < \varepsilon$ we have that $2(1 - (n \cdot \overline{m})^2) = |n \otimes n - \overline{m} \otimes \overline{m}|^2 < \varepsilon^2$ and so

$$n \cdot \overline{m} \geq \sqrt{1 - \frac{\varepsilon^2}{2}} > 0 \qquad \text{or} \qquad n \cdot \overline{m} \leq -\sqrt{1 - \frac{\varepsilon^2}{2}} < 0.$$

Thus $n \otimes n = n^+ \otimes n^+ = n^- \otimes n^-$, where $n^+ \cdot \overline{m} > 0$ and $n^- = -n^+$ satisfies $n^- \cdot \overline{m} < 0$. Now let $Q(\tau) = n(\tau) \otimes n(\tau)$ be continuous on $[t_1, t_2]$. Then there exists $\delta > 0$ such that $|n(\tau) \otimes n(\tau) - n(\sigma) \otimes n(\sigma)| \leq \sqrt{2}$. For all $\sigma, \tau \in [t_1, t_2]$ with $|\sigma - \tau| \leq \delta$, and we may suppose that $t_2 - t_1 = M\delta$ for some integer $M \in \mathbb{N}$. First take $\overline{m} := n(t_1)$ and for each $\tau \in [t_1, t_1 + \delta]$ choose $n^+(\tau)$ as above so that $n^+(\tau) \otimes n^+(\tau) = n(\tau) \otimes n(\tau)$ and $n^+(\tau) \cdot \overline{m} > 0$. We claim that $n^+ : [t_1, t_1 + \delta] \to \mathbb{S}^2$ is continuous. Indeed, let $\sigma_j \to \sigma$ in $[t_1, t_1 + \delta]$ and suppose for contradiction that $n(\sigma_j) \not\to n(\sigma)$. Then since $n^+(\sigma_j) \otimes n^+(\sigma_j) \to n^+(\sigma) \otimes n^+(\sigma)$ there is a subsequence σ_{j_k} such that $n^+(\sigma_{j_k}) \to -n^+(\sigma)$. But then $-n^+(\sigma) \cdot \overline{m} \geq 0$ is a contradiction which proves the claim. Repeating this procedure with $\overline{n} := n^+(t_1 + \delta)$ we obtain a continuous lifting $n^+ : [t_1, t_1 + 2\delta] \to \mathbb{S}^2$, and thus inductively a continuous lifting $n^+ : [t_1, t_2] \to \mathbb{S}^2$. Setting $n^- = -n^+$ gives a second continuous lifting. Again, by a standard continuity argument we see that there exist only two continuous liftings. □

Let $n_0 \in \mathbb{S}^2$ satisfy $Q_0 = n_0 \otimes n_0$. According to Lemma 1, there exists for $\delta > 0$ a $\tilde{\gamma} : (-\delta, \delta) \to \mathbb{S}^2$ satisfying $\tilde{\gamma}(0) = n_0$ and $\gamma|_{(-\delta,\delta)} = \tilde{\gamma} \otimes \tilde{\gamma}|_{(-\delta,\delta)}$. We define $\frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} \tilde{\gamma} := v \in T_{n_0}\mathbb{S}^2$ and obtain $V \in T_{Q_0}\mathbb{L}^2$ as

$$\begin{aligned} V = \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} \gamma(t) &= \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} \tilde{\gamma}(t) \otimes \tilde{\gamma}(t) \\ &= n_0 \otimes \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} \tilde{\gamma}(t) + \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} \tilde{\gamma}(t) \otimes n_0 = n_0 \otimes v + v \otimes n_0. \end{aligned}$$

This means that there is a one-to-one correspondence between the tangent space $T_{Q_0}\mathbb{L}^2$ and $T_{n_0}\mathbb{S}^2$.

Let $\partial\Omega = \Gamma_{nor} \cup \Gamma_{tan} \cup \Gamma_N$ be a partition of the boundary of Ω. For line fields and director fields we impose natural Neumann boundary conditions on Γ_N and the essential boundary conditions of homeotropic anchoring and planar anchoring on Γ_{nor} and Γ_{tan}, respectively:

	$Q = n^Q \otimes n^Q \in \mathbb{L}^2$	$n \in \mathbb{S}^2$
$x \in \Gamma_{tan}$	$n^Q(x) \,\|\, \nu_{\partial\Omega}(x)$	$n(x) \,\|\, \nu_{\partial\Omega}(x)$
$x \in \Gamma_{nor}$	$n^Q(x) \perp \nu_{\partial\Omega}(x)$	$n(x) \perp \nu_{\partial\Omega}(x)$

Admissible tensor and director fields for E_{LdG} and E_{OF} are

$$\begin{aligned} \mathscr{A}_{LdG} := \{Q \in W^{1,2}(\Omega, \mathbb{R}^{3\times 3}) :\ & Q \in \mathbb{L}^2 \text{ a.e. in } \Omega, \\ & Q \text{ satisfies the boundary conditions on } \Gamma_{nor} \cup \Gamma_{tan}\} \qquad \text{and} \\ \mathscr{A}_{OF} := \{n \in W^{1,2}(\Omega, \mathbb{R}^3) :\ & n \in \mathbb{S}^2 \text{ a.e. in } \Omega, \\ & n \text{ satisfies the boundary conditions on } \Gamma_{nor} \cup \Gamma_{tan}\}. \end{aligned}$$

Stationary points of E_{LdG} in the set of admissible line fields satisfy the imposed boundary conditions and

$$(\nabla Q, \nabla V) = 0,$$

for all $V \in \mathscr{F}_{\mathbb{L}^2}[Q]$ given by

$$\mathscr{F}_{\mathbb{L}^2}[Q] := \{V \in C_0^\infty(\overline{\Omega} \setminus (\Gamma_{nor} \cup \Gamma_{tan}), \mathbb{R}^{3\times 3}) \,:\, V(x) \in T_{Q(x)}\mathbb{L}^2 \text{ for } x \in \Omega\}.$$

At least locally, there exists always a vector field n satisfying $Q = n^Q \otimes n^Q$ and therefore we can rewrite the Euler Lagrange equation as

$$(\nabla Q, \nabla(n^Q \otimes v + v \otimes n^Q)) = 0$$

for all $v \in \mathscr{F}_{\mathbb{S}^2}[n^Q]$ given by

$$\mathscr{F}_{\mathbb{S}^2}[n^Q] := \{v \in C_0^\infty(\overline{\Omega} \setminus (\Gamma_{nor} \cup \Gamma_{tan}), \mathbb{R}^3) \,:\, v(x) \in T_{n^Q(x)}\mathbb{S}^2 \text{ for } x \in \Omega\}.$$

Stationary points of E_{OF} in the set $\mathscr{A}_{OF}$ satisfy the imposed boundary conditions and the Euler Lagrange equations

$$(\nabla n, \nabla v) = 0$$

for all $v \in \mathscr{F}_{\mathbb{S}^2}[n]$. Clearly, it is only possible to consider E_{OF} if the boundary values are orientable in the sense that there exists an orientable line field realizing the boundary conditions.

3 Discrete Setting

We let $\mathscr{T}_h$ be a regular triangulation into triangles ($d = 2$) or tetrahedra ($d = 3$) of maximal diameter $h > 0$ in the sense of [9]. We denote by $\mathbb{V} = \mathbb{V}(\mathscr{T}_h)$ the space of all continuous functions on Ω that are affine on the elements in the triangulation $\mathscr{T}_h$ and we set $\mathbb{V}_{nor} = \mathbb{V} \cap \{v \in W^{1,2}(\Omega) : v|_{\Gamma_{nor}} = 0\}$. We call a triangulation $\mathscr{T}_h$ *weakly acute* if

$$K_{ij} := \int_\Omega \nabla\varphi_{a_i} \cdot \nabla\varphi_{a_j} \,\mathrm{d}x \le 0 \quad \text{for all } a_i \neq a_j \in \mathscr{N}, \tag{1}$$

where $\mathscr{N} = \{a_1, \dots, a_N\}$ denotes the set of nodes in $\mathscr{T}_h$ and $(\varphi_a)_{a\in\mathscr{N}}$ is the standard nodal basis of $\mathbb{V}$. Note that if $d = 2$ the triangulation $\mathscr{T}_h$ is weakly acute if the sum of every pair of angles opposite to an interior edge is bounded by π and if the angle opposite to every edge on the boundary is less than or equal to $\pi/2$. We denote by $\mathscr{I}_h : C^0(\Omega) \to \mathbb{V}$ the standard nodal interpolant and for a fixed time-step size $\tau > 0$ let $t_j = j\tau$ for all $j \ge 0$.

3.1 Monotonicity Estimates

We include a monotonicity estimate from [6] that is a discrete version of a corresponding statement in [2].

Lemma 2 (Monotonicity I). *Let $\mathscr{T}_h$ be weakly acute, and let $\tilde{n}_h \in \mathbb{V}^3$ be such that $|\tilde{n}_h(a)| \geq 1$ for all $a \in \mathscr{N}$, and define $n_h \in \mathbb{V}^3$ by setting $n_h(a) = \tilde{n}_h(a)/|\tilde{n}_h(a)|$ for all $a \in \mathscr{N}$. Then*

$$\|\nabla n_h\| \leq \|\nabla \tilde{n}_h\|. \tag{2}$$

Proof. Let $(\varphi_{a_i})_{a_i \in \mathscr{N}}$ denote the nodal basis of $\mathbb{V}$. Besides (1), the symmetric matrix $(K_{ij})_{i,j=1}^N$ satisfies $\sum_{j=1}^N K_{ij} = 0$ owing to $\sum_{j=1}^N \varphi_{a_j} = 1$. We observe the relations

$$\begin{aligned}
||\nabla n_h||^2 &= \sum_{i,j=1}^N K_{ij} n_h(a_i) \cdot n_h(a_j) \\
&= \frac{1}{2} \sum_{i,j=1}^N K_{ij} n_h(a_i) \cdot \big(n_h(a_j) - n_h(a_i)\big) + \frac{1}{2} \sum_{i,j=1}^N K_{ij} n_h(a_j) \cdot \big(n_h(a_i) - n_h(a_j)\big) \\
&= -\frac{1}{2} \sum_{i,j=1}^N K_{ij} \big|n_h(a_i) - n_h(a_j)\big|^2.
\end{aligned}$$

The assertion is proved if $|n_h(a_i) - n_h(a_j)|^2 \leq |\tilde{n}_h(a_i) - \tilde{n}_h(a_j)|^2$ for all $i, j = 1, \cdots, N$. Hence, it suffices to show $\left|\frac{a}{|a|} - \frac{b}{|b|}\right| \leq \left|a - b\right|$, for $a, b \in \mathbb{R}^3$ with $|a|, |b| \geq 1$. This follows from the Lipschitz continuity with constant 1 of the map $\pi_{\mathbb{S}^2} : \{x \in \mathbb{R}^3 : |x| \geq 1\} \to \mathbb{S}^2$, $x \mapsto x/|x|$. □

Before we turn to the characterization of discrete harmonic director fields and Q harmonic line fields we state another monotonicity estimate result which is an adoption from the previous argument to the finite element space of functions that have nodal values in $\mathbb{L}^2$. The result is a consequence of the following auxiliary estimate.

Lemma 3 (Tensor Estimate). *Let $\tilde{v}, \tilde{w} \in \mathbb{R}^3$ such that $|\tilde{v}|, |\tilde{w}| \geq 1$. Set $v = \tilde{v}/|\tilde{v}|$ and $w = \tilde{w}/|\tilde{w}|$, then*

$$1 - (v \otimes v) : (w \otimes w) \leq \frac{1}{2}\Big|\tilde{v} \otimes \tilde{v} - \tilde{w} \otimes \tilde{w}\Big|^2, \tag{3}$$

where for $A, B \in \mathbb{R}^3$ $A : B = \mathrm{tr}(A^T B)$ and $\mathrm{tr} : \mathbb{R}^{3\times 3} \to \mathbb{R}$ is the usual trace of a square matrix.

Proof. We deduce

$$1 - (v \otimes v) : (w \otimes w) = 1 - (v \cdot w)^2 = (1 - v \cdot w)(1 + v \cdot w) = \frac{1}{4}|v - w|^2 |v + w|^2.$$

Where we incorporated the identity $1 \pm v \cdot w = \frac{1}{2}|v|^2 + \frac{1}{2}|w|^2 \pm v \cdot w = \frac{1}{2}|v \pm w|^2$ and the fact that

$$(v \otimes v) : (w \otimes w) = \sum_{k,\ell} v_k v_\ell w_k w_\ell = \sum_k v_k w_k \sum_\ell v_\ell w_\ell = (v \cdot w)^2.$$

Since $\pi_{\mathbb{S}^2}$ is point symmetric with respect to the origin and Lipschitz continuous on $\mathbb{R}^3 \setminus B_1(0)$ with Lipschitz constant one we deduce that

$$|v \pm w| = \left| \frac{\tilde{v}}{|\tilde{v}|} \pm \frac{\tilde{w}}{|\tilde{w}|} \right| \leq |\tilde{v} \pm \tilde{w}|.$$

This yields

$$\begin{aligned}
1 - (v \otimes v) : (w \otimes w) &\leq \frac{1}{4}|\tilde{v} - \tilde{w}|^2 |\tilde{v} + \tilde{w}|^2 \\
&= \frac{1}{4}\Big(|\tilde{v}|^2 + |\tilde{w}|^2 - 2\tilde{v} \cdot \tilde{w}\Big)\Big(|\tilde{v}|^2 + |\tilde{w}|^2 + 2\tilde{v} \cdot \tilde{w}\Big) \\
&= \frac{1}{4}\Big[\big(|\tilde{v}|^2 + |\tilde{w}|^2\big)^2 - 4(\tilde{v} \cdot \tilde{w})^2 \Big] \\
&= \frac{1}{4}\Big[\big(|\tilde{v}|^4 - 2(\tilde{v} \cdot \tilde{w})^2 + |\tilde{w}|^4\big) + 2\big(|\tilde{v}|^2|\tilde{w}|^2 - (\tilde{v} \cdot \tilde{w})^2\big) \Big] \\
&\leq \frac{1}{2}\big(|\tilde{v}|^4 - 2(\tilde{v} \cdot \tilde{w})^2 + |\tilde{w}|^4\big) = \frac{1}{2}\Big|\tilde{v} \otimes \tilde{v} - \tilde{w} \otimes \tilde{w}\Big|^2,
\end{aligned}$$

where we used Young's inequality $2|\tilde{v}|^2|\tilde{w}|^2 \leq |\tilde{v}|^4 + |\tilde{w}|^4$ and the identity $|\tilde{v}|^4 = |\tilde{v} \otimes \tilde{v}|^2$. □

Lemma 4 (Monotonicity II). *Let $\mathcal{T}_h$ be weakly acute, and let $\tilde{n}_h \in [\mathbb{V}]^3$ be such that $|\tilde{n}_h(a)| \geq 1$ for all $a \in \mathcal{N}$, and define $n_h \in [\mathbb{V}]^3$ by setting $n_h(a) = \tilde{n}_h(a)/|\tilde{n}_h(a)|$ for all $a \in \mathcal{N}$. Furthermore we define $\tilde{Q}_h, Q_h \in [\mathbb{V}]^{3\times 3}$ by setting*

$$\tilde{Q}_h(a) := \tilde{n}_h(a) \otimes \tilde{n}_h(a) \quad \textit{and} \quad Q_h(a) := \frac{\tilde{n}_h(a)}{|\tilde{n}_h(a)|} \otimes \frac{\tilde{n}_h(a)}{|\tilde{n}_h(a)|} \quad \textit{for all } a \in \mathcal{N}.$$

Then

$$\|\nabla Q_h\| \leq \|\nabla \tilde{Q}_h\|. \tag{4}$$

Proof. We start the proof with the same arguments as in Lemma 2 which yield

$$\|\nabla Q_h\|^2 = -\frac{1}{2}\sum_{i,j}^{N} K_{ij}|Q_h(a_i) - Q_h(a_j)|^2$$

$$= -\frac{1}{2}\sum_{i,j}^{N} K_{ij}\left(|Q_h(a_i)|^2 - 2Q_h(a_i) : Q_h(a_j) + |Q_h(a_j)|^2\right)$$

$$= -\sum_{i,j}^{N} K_{ij}\left(1 - Q_h(a_i) : Q_h(a_j)\right).$$

For $i, j \in \{1, \dots, N\}$ arbitrary we incorporate the estimate (3) from Lemma 3 with $\tilde{v} := \tilde{n}_h(a_i)$, $\tilde{w} := \tilde{n}_h(a_j)$, $v = \tilde{v}/|\tilde{v}|$ and $w = \tilde{w}/|\tilde{w}|$ and arrive at

$$1 - Q_h(a_i) : Q_h(a_j) \leq \frac{1}{2}|\tilde{Q}_h(a_i) - \tilde{Q}_h(a_j)|^2.$$

We conclude

$$||\nabla Q_h||^2 = -\sum_{ij} K_{ij}(1 - Q_h(a_i) : Q_h(a_j)) \leq -\frac{1}{2}\sum_{ij} K_{ij}|\tilde{Q}_h(a_i) - \tilde{Q}_h(a_j)|^2 = ||\nabla \tilde{Q}_h||^2,$$

which proves the lemma. □

3.2 Euler Lagrange Equation in the Discrete Setting

For the discrete version of the boundary conditions we assume that all nodes on the boundary of Ω lie either in Γ_N, Γ_{tan} or Γ_{nor}. For discrete line and director fields we impose natural Neumann boundary conditions on Γ_N and the discrete essential boundary conditions of homeotropic anchoring and planar anchoring on Γ_{nor} and Γ_{tan}, respectively:

	$Q_h(z) = n_h^Q(z) \otimes n_h^Q(z) \in \mathbb{L}^2$ for all $z \in \mathcal{N}_h$	$n_h(z) \in \mathbb{S}^2$ for all $z \in \mathcal{N}_h$
$z \in \Gamma_{tan}$	$n_h^Q(z) \,\|\|\, \nu_{\partial\Omega}(z)$	$n_h(z) \,\|\|\, \nu_{\partial\Omega}(z)$
$z \in \Gamma_{nor}$	$n_h^Q(z) \perp \nu_{\partial\Omega}(z)$	$n_h(z) \perp \nu_{\partial\Omega}(z)$

Thus, we define the discrete admissible line fields and director fields for E_{LdG} and E_{OF}

$$\mathscr{A}_{LdG}^h := \{P_h \in [\mathbb{V}]^{3\times 3} : P_h(a) \in \mathbb{L}^2 \text{ for all } a \in \mathcal{N},$$
$$P_h \text{ satisfies the boundary conditions on } \Gamma_{nor} \cup \Gamma_{tan}\} \qquad \text{and}$$

$$\mathscr{A}_{OF}^h := \{v_h \in [\mathbb{V}]^3 : v_h(a) \in \mathbb{S}^2 \text{ for all } a \in \mathcal{N},$$
$$v_h \text{ satisfies the boundary conditions on } \Gamma_{nor} \cup \Gamma_{tan}\}.$$

Definition 2. (i) A map $Q_h \in \mathscr{A}_{LdG}^h$ is called a *discrete Q harmonic tensor field* into $\mathbb{L}^2$ subject to homeotropic anchoring, planar anchoring and Neumann boundary conditions if Q_h is stationary for E_{LdG} among all $P_h \in \mathscr{A}_{LdG}^h$.

(ii) A vector field $n_h \in \mathscr{A}_{OF}^h$ is called a *discrete harmonic director field* subject to homeotropic anchoring, planar anchoring and Neumann boundary conditions if n_h is stationary for E_{OF} among all $v_h \in \mathscr{A}_{OF}^h$.

Given $n_h \in \mathscr{A}_{OF}^h$ we define the space of tangential updates with respect to the sphere by

$$\begin{aligned}\mathscr{F}_{\mathbb{S}^2}[n_h] = \big\{r_h \in [\mathbb{V}_{nor}]^3 : \quad & r_h(a)\cdot n_h(a) = 0 \ \text{ for all } a \in \mathscr{N}, \\ & \text{and } r_h(a)\cdot \nu_{\partial\Omega}(a) = 0 \ \text{ for all } a \in \mathscr{N} \cap \Gamma_{tan}\big\}. \end{aligned} \tag{5}$$

For computations with line fields we define the space of tangential updates for a given $Q_h \in \mathscr{A}_{LdG}^h$ as

$$\mathscr{F}_{\mathbb{L}^2}[Q_h] = \big\{R_h \in [\mathbb{V}_{nor}]^{3\times 3} : R_h = \mathscr{I}_h[r_h \otimes n_h^Q + n_h^Q \otimes r_h] \,\text{for}\, r_h \in \mathscr{F}_{\mathbb{S}^2}[n_h^Q]\big\}, \tag{6}$$

where $n_h^Q \in [\mathbb{V}]^3$ satisfies $|n_h^Q(a)| = 1$ and $Q_h(a) = n_h^Q(a) \otimes n_h^Q(a)$ for all $a \in \mathscr{N}$. The following proposition is a variation of Lemma 3.1.4 from [5].

Proposition 1. *1. A tensor field $Q_h \in \mathscr{A}_{LdG}^h$ is a discrete Q harmonic tensor field into $\mathbb{L}^2$ according to Definition 2 if and only if there holds*

$$(\nabla Q_h, \nabla V_h) = 0 \tag{7}$$

for all $V_h \in \mathscr{F}_{\mathbb{L}^2}[Q_h]$.

2. A vector field $n_h \in \mathscr{A}_{OF}^h$ is a discrete harmonic director field according to Definition 2 if and only if there holds

$$(\nabla n_h, \nabla v_h) = 0 \tag{8}$$

for all $v_h \in \mathscr{F}_{\mathbb{S}^2}[n_h]$.

Proof. For the proof of (1) we note that a variation of Q_h can be given by the term $\mathscr{I}_h \pi_{\mathbb{L}^2}(Q_h + tP_h)$ for $t > 0$ small enough and $P_h \in [\mathbb{V}_{nor}]^{3\times 3}$. Then

$$\mathscr{I}_h \pi_{\mathbb{L}^2}(Q_h + tP_h)(a) = Q_h(a) + tD\pi_{\mathbb{L}^2}(Q_h(a))P_h(a) + o(t)$$

for all $a \in \mathscr{N}$. If $P_h \in \mathscr{F}_{\mathbb{L}^2}[Q_h]$ then $D\pi_{\mathbb{L}^2}(Q_h(a))P_h(a) = P_h(a)$ for all $a \in \mathscr{N}$ and we obtain that

$$\mathscr{I}_h \pi_{\mathbb{L}^2}(Q_h + tP_h) = Q_h + tP_h + o(t).$$

Thus, (7) follows by computing

$$0 = \lim_{t \to 0} t^{-1} \left(E_{LdG}(\mathscr{I}_h \pi_{\mathbb{L}^2}(Q_h + tP_h)) - E_{LdG}(Q_h) \right) = (\nabla Q_h, \nabla P_h)\,.$$

The proof of (2) follows analogously. □

4 Iterative Algorithms

We compute stationary points of E_{OF} and E_{LdG} via iterative algorithms that are motivated by the corresponding H^1 gradient flows. The continuous H^1 gradient flow for harmonic maps into a submanifold $\Sigma \subset \mathbb{R}^n$ subject to Dirichlet conditions on Γ_D seeks a function $V : (0, \infty) \times \Omega \to \Sigma$ satisfying $V(0, \cdot) = V_0$, $V(t, \cdot)|_{\Gamma_D} = V_D$ and

$$(\nabla \partial_t V, \nabla P) + (\nabla V, \nabla P) = 0 \tag{9}$$

for almost every $t \in (0, \infty)$ and all $P \in W^{1,2}_D(\Omega, \mathbb{R}^n)$ such that $P(x) \in T_{V(x)}\Sigma$ for almost every $x \in \Omega$.

4.1 Fully Discrete Algorithm for Discrete Harmonic Director Fields

For the computation of discrete harmonic director fields a semi-implicit discretization of (9) yields the following algorithm. Well-posedness, unconditional stability for weakly acute triangulations, termination and convergence of the algorithm can be found in [6].

Input Triangulation $\mathscr{T}_h$, stopping criterion $\varepsilon > 0$, time-step size $\tau > 0$ and $n_h^0 \in \mathscr{A}_{OF}^h$. Set $i = 0$.

1. Compute $w_h^i \in \mathscr{F}_{\mathbb{S}^2}[n_h^i]$ such that

$$\left(\nabla w_h^i, \nabla v_h\right) + \left(\nabla (n_h^i + \tau w_h^i), \nabla v_h\right) = 0$$

for all $v_h \in \mathscr{F}_{\mathbb{S}^2}[n_h^i]$.

2. Set

$$n_h^{i+1}(a) := \frac{n_h^i(a) + \tau w_h^i(a)}{|n_h^i(a) + \tau w_h^i(a)|},$$

for all $a \in \mathscr{N}$.

3. Stop, if $||\nabla w_h^i||_{L^2} < \varepsilon$.
4. Set $i = i + 1$ and go to (1).

Output: $n_h^* := n_h^i$.

Remark 1. (i) For $d_t n_h^i := w_h^i$ and $\tilde{n}_h^{i+1} := n_h^i + \tau d_t n_h^i$ the equation in Step 1 reads

$$(\nabla d_t n_h^i, \nabla v_h) + (\nabla \tilde{n}_h^{i+1}, \nabla v_h) = 0$$

which is a discrete version of (9).

(ii) As it was already discussed in [7, 16] the same algorithm without Step 2 yields for the output n_h^*

$$||\mathscr{I}_h[|n_h^*|^2 - 1]||_{L^1} \leq C\tau E_{OF}(n_h^0).$$

Thus, for $\tau > 0$ small enough the projection step can be skipped and the violation of the constraint at the nodes is controlled by the time-step size τ.

4.2 Fully Discrete Algorithm for Discrete Q Harmonic Tensor Fields

For the computation of discrete Q harmonic tensor fields we propose the following algorithm which is a discretization of a variation of the H^1 gradient flow. Locally, we have that $Q = n^Q \otimes n^Q$ and $\partial_t Q = \partial_t n^Q \otimes n^Q + n^Q \otimes \partial_t n^Q$ as well as the relation $V = n^Q \otimes v + v \otimes n^Q$ for $V \in \mathscr{F}_{\mathbb{L}^2}[Q]$ and some $v \in \mathscr{F}_{\mathbb{S}^2}[n^Q]$. We employ the modified H^1 gradient flow as

$$(\nabla \partial_t n^Q, \nabla v) + \lambda(\nabla \partial_t Q, \nabla V) + (\nabla Q, \nabla V) = 0,$$

with a discretization parameter $\lambda > 0$ that coincides with the time-step size. In Sect. 5 we will provide proofs of stability, termination and convergence of the algorithm to a discrete Q harmonic tensor field.

Input Triangulation $\mathscr{T}_h$, stopping criterion $\varepsilon > 0$, time-step size $\tau > 0$ and $Q_h^0 \in \mathscr{A}_{LdG}^h$. Set $i := 0$.

1. Compute $w_h^i \in \mathscr{F}_{\mathbb{S}^2}[n_h^{Q,i}]$ such that

$$\Big(\nabla w_h^i, \nabla v_h\Big) + \Big(\nabla(Q_h^i + \tau \mathscr{I}_h[n_h^{Q,i} \otimes w_h^i + w_h^i \otimes n_h^{Q,i}]), \nabla \mathscr{I}_h[n_h^{Q,i} \otimes v_h + v_h \otimes n_h^{Q,i}]\Big) = 0,$$

for all $v_h \in \mathscr{F}_{\mathbb{S}^2}[n_h^{Q,i}]$.

2. Set

$$Q_h^{i+1}(a) := \frac{n_h^{Q,i}(a) + \tau w_h^i(a)}{|n_h^{Q,i}(a) + \tau w_h^i(a)|} \otimes \frac{n_h^{Q,i}(a) + \tau w_h^i(a)}{|n_h^{Q,i}(a) + \tau w_h^i(a)|},$$

   for all $a \in \mathscr{N}$.
3. Stop, if $||\nabla w_h^i||_{L^2} < \varepsilon$.
4. Set $i = i + 1$ and go to (1).

Output: $Q_h^* = Q_h^i$.

5 Analysis of the Algorithms and *Q* Harmonic Tensor Fields

In the first part of this section we analyze the proposed algorithm for the computation of Q harmonic tensor fields. Related results for the H^1 gradient flow for director fields can be found in [6]. Furthermore, we discuss a weak compactness result for Q harmonic tensor fields on a continuous level which provides the basis for a convergence analysis of the discrete approximations. We refer to [5] for a corresponding analysis in the case of discrete harmonic director fields.

5.1 Stability and Convergence of the Tensor Field Algorithm

We start our analysis by showing well-posedness of the algorithm. A stability result enables us to show termination of the algorithm and convergence to discrete Q harmonic tensor field.

Lemma 5 (Well-posedness). *Given $Q_h \in [\mathbb{V}]^{3\times 3}$ satisfying $Q_h(a) \in \mathbb{L}^2$ for all $a \in \mathscr{N}$ there exists $w_h \in \mathscr{F}_{\mathbb{S}^2}[n_h^Q]$ satisfying*

$$\Big(\nabla w_h, \nabla v_h\Big) + \Big(\nabla(Q_h + \tau \mathscr{I}_h[n_h^Q \otimes w_h + w_h \otimes n_h^Q]), \nabla \mathscr{I}_h[n_h^Q \otimes v_h + v_h \otimes n_h^Q]\Big) = 0 \tag{10}$$

for all $v_h \in \mathscr{F}_{\mathbb{S}^2}[n_h^Q]$. Moreover we have the following estimate

$$\tau \|\nabla w_h\|^2 \leq E_{LdG}(Q_h). \tag{11}$$

Proof. We set

$$T := \Big\{ v_h \in [\mathbb{V}_{nor}]^3 \ : \ v_h(a) \in T_{n_h^Q(a)}\mathbb{S}^2 \text{ for all } a \in \mathscr{N} \Big\}$$

and note that T is a subspace of $[\mathbb{V}_{nor}]^3$. The bilinear form $A_Q : T \times T \to \mathbb{R}$, defined through

$$(w_h, v_h) \quad \mapsto \quad \Big(\nabla w_h, \nabla v_h\Big) + \Big(\nabla \mathscr{I}_h[n_h^Q \otimes w_h + w_h \otimes n_h^Q], \nabla \mathscr{I}_h[n_h^Q \otimes v_h + v_h \otimes n_h^Q]\Big)$$

fulfills the requirements of the Lax-Milgram Lemma. For the unique solution w_h of (10) we obtain by choosing $v_h = \tau w_h$ that

$$\begin{aligned}\tau\|\nabla w_h\|^2 + \tau^2\|\nabla \mathscr{I}_h[n_h^Q \otimes w_h + w_h \otimes n_h^Q]\|^2 &= -\tau\Big(\nabla Q_h, \nabla \mathscr{I}_h[n_h^Q \otimes w_h + w_h \otimes n_h^Q]\Big)\\ &\le \frac{1}{4}\|\nabla Q_h\|^2 + \tau^2\|\nabla \mathscr{I}_h[n_h^Q \otimes w_h + w_h \otimes n_h^Q]\|^2,\end{aligned}$$

which is the asserted estimate. □

Lemma 6 (Stability). *Assume that $\mathscr{T}_h$ is weakly acute. For a given $Q_h^0 \in \mathscr{A}_{LdG}^h$ let $(Q_h^i)_{0 \le i \le J} \subset \mathscr{A}_{LdG}^h$ be the sequence of tensor fields computed in our Q tensor field algorithm and let*

$$C' := 1 - CC_0^{1/2}\tau h^{1-d/2}(\log h_{min}^{-1}) - CC_0\tau^3 h^{2-d}(\log h_{min}^{-1})^2,$$

where $C_0 := E_{LdG}(Q_h^0)$, $h_{min}^2 := \min_{T \in \mathscr{T}_h} \operatorname{diam} T$ and the constant $C > 0$ depends on the geometry of the mesh but is independent of the mesh-size $h > 0$. If the time-step size $\tau > 0$ is small enough, so that $C' > 0$ then for all $J \ge 1$

$$E_{LdG}(Q_h^{J+1}) + C'(\tau/2)\sum_{i=0}^{J}\|\nabla w_h^i\|^2 \le E_{LdG}(Q_h^0),$$

and the Q-field algorithm terminates within a finite number of iterations.

Proof. We recall that $Q_h^i = \mathscr{I}_h[n_h^{Q,i} \otimes n_h^{Q,i}]$ and

$$Q_h^{i+1}(a) = \frac{n_h^{Q,i}(a) + \tau w_h^i(a)}{|n_h^{Q,i}(a) + \tau w_h^i(a)|} \otimes \frac{n_h^{Q,i}(a) + \tau w_h^i(a)}{|n_h^{Q,i}(a) + \tau w_h^i(a)|},$$

for all $a \in \mathscr{N}$. We set

$$\tilde{Q}_h^{i+1} := Q_h^i + \tau \mathscr{I}_h[w_h^i \otimes n_h^{Q,i} + n_h^{Q,i} \otimes w_h^i]$$

and

$$\tilde{\tilde{Q}}_h^{i+1} := \mathscr{I}_h[(n_h^{Q,i} + \tau w_h^i) \otimes (n_h^{Q,i} + \tau w_h^i)] = \tilde{Q}_h^{i+1} + \tau^2 \mathscr{I}_h[w_h^i \otimes w_h^i].$$

Furthermore, we know that according to Lemma 4 $E_{LdG}(\tilde{\tilde{Q}}_h^{i+1}) \geq E_{LdG}(Q_h^{i+1})$, since $\mathscr{T}_h$ is weakly acute. In Step 1 of the algorithm we compute $w_h^i \in \mathscr{F}_{\mathbb{S}^2}[n_h^{Q,i}]$ satisfying

$$(\nabla w_h^i, \nabla v_h) + (\nabla \tilde{Q}_h^i, \nabla \mathscr{I}_h[v_h \otimes n_h^{Q,i} + n_h^{Q,i} \otimes v_h]) = 0,$$

for all $v_h \in \mathscr{F}_{\mathbb{S}^2}[n_h^{Q,i}]$. We test the equation with $v_h = \tau w_h^i$ and obtain

$$\tau \|\nabla w_h^i\|^2 + (\nabla \tilde{Q}_h^{i+1}, \nabla(\tilde{Q}_h^{i+1} - Q_h^i)) = 0.$$

Upon using the binomial identity $2a(a-b) = (a-b)^2 + a^2 - b^2$ we infer that

$$\tau \|\nabla w_h^i\|^2 + 2(E_{LdG}(\tilde{Q}_h^{i+1}) - E_{LdG}(Q_h^i)) + (\tau^2/2)\|\nabla \mathscr{I}_h[w_h^i \otimes n_h^{Q,i} + n_h^{Q,i} \otimes w_h^i]\|^2 = 0.$$

The monotonicity estimate for line fields together with the identity

$$E_{LdG}(\tilde{\tilde{Q}}_h^{i+1}) = E_{LdG}(\tilde{Q}_h^{i+1}) + (\tau^2/2)(\nabla \tilde{Q}_h^{i+1}, \nabla \mathscr{I}_h[w_h^i \otimes w_h^i]) + \tau^4 E_{LdG}(\mathscr{I}_h[w_h^i \otimes w_h^i])$$

yields

$$\begin{aligned} \tau \|\nabla w_h^i\|^2 &+ 2(E_{LdG}(Q_h^{i+1}) - E_{LdG}(Q_h^i)) + (\tau^2/2)\|\nabla \mathscr{I}_h[w_h^i \otimes n_h^{Q,i} + n_h^{Q,i} \otimes w_h^i]\|^2 \\ &- \tau^2(\nabla \tilde{Q}_h^{i+1}, \nabla \mathscr{I}_h[w_h^i \otimes w_h^i]) - 2\tau^4 E_{LdG}(\mathscr{I}_h[w_h^i \otimes w_h^i]) \leq 0. \end{aligned}$$

To bound the first negative term we employ the representation of $\tilde{Q}_h^{i+1}$ and Young's inequality

$$\begin{aligned} \tau^2(\nabla \tilde{Q}_h^{i+1}, \nabla \mathscr{I}_h[w_h^i \otimes w_h^i]) &= \tau^2(\nabla(Q_h^i + \tau \mathscr{I}_h[w_h^i \otimes n_h^{Q,i} + n_h^{Q,i} \otimes w_h^i]), \nabla \mathscr{I}_h[w_h^i \otimes w_h^i]) \\ &\leq 2\tau^2 (E_{LdG}(Q_h^i))^{1/2} \|\nabla \mathscr{I}_h[w_h^i \otimes w_h^i]\| \\ &\quad + 2\tau^4 E_{LdG}(\mathscr{I}_h[w_h^i \otimes w_h^i]) \\ &\quad + (\tau^2/2)\|\nabla \mathscr{I}_h[w_h^i \otimes n_h^{Q,i} + n_h^{Q,i} \otimes w_h^i]\|^2. \end{aligned}$$

Thus, we arrive at

$$\begin{aligned} &\tau \|\nabla w_h^i\|^2 + 2(E_{LdG}(Q_h^{i+1}) - E_{LdG}(Q_h^i)) \\ &\quad - 2\tau^2 (E_{LdG}(Q_h^i))^{1/2} \|\nabla \mathscr{I}_h[w_h^i \otimes w_h^i]\| - 4\tau^4 E_{LdG}(\mathscr{I}_h[w_h^i \otimes w_h^i]) \leq 0. \quad (12) \end{aligned}$$

We argue by induction and assume that $E_{LdG}(Q_h^j) \le C_0$ for $j = 0, \dots, i$. A discrete norm equivalence on every triangle $T \in \mathscr{T}_h$ shows that

$$\|\nabla \mathscr{I}_h[w_h^i \otimes w_h^i]\|_{L^2(T)} \le C\|\nabla(w_h^i \otimes w_h^i)\|_{L^2(T)} \le 2C\|w_h^i\|_{L^\infty(T)}\|\nabla w_h^i\|_{L^2(T)}.$$

We incorporate the inverse estimate $\|w_h^i\|_{L^\infty(T)} \le Ch_T^{1-d/2}\log h_T^{-1}\|\nabla w_h^i\|_{L^2(T)}$, cf., e.g., [9], and sum over all $T \in \mathscr{T}_h$ to arrive at

$$\|\nabla \mathscr{I}_h[w_h^i \otimes w_h^i]\| \le Ch_{min}^{1-d/2}\log h_{min}^{-1}\|\nabla w_h^i\|^2.$$

Furthermore, if we incorporate (11) and the induction hypotheses we obtain the following bound

$$E_{LdG}(\mathscr{I}_h[w_h^i \otimes w_h^i]) \le Ch_{min}^{2-d}(\log h_{min}^{-1})^2\|\nabla w_h^i\|^4 \le CC_0\tau^{-1}h_{min}^{2-d}(\log h_{min}^{-1})^2\|\nabla w_h^i\|^2.$$

We use the derived bounds in (12) and deduce that

$$\begin{aligned}\tau(1 - CC_0^{1/2}\tau h_{min}^{1-d/2}\log h_{min}^{-1} - CC_0\tau^2 h_{min}^{2-d}(\log h_{min}^{-1})^2)\|\nabla w_h^i\|^2 \\ +2(E_{LdG}(Q_h^{i+1}) - E_{LdG}(Q_h^i)) \le 0.\end{aligned}$$

Upon choosing $\tau > 0$ small enough so that

$$C' := 1 - CC_0^{1/2}\tau h_{min}^{1-d/2}\log h_{min}^{-1} - CC_0\tau^2 h_{min}^{2-d}(\log h_{min}^{-1})^2 > 0,$$

we obtain the local energy inequality

$$C'\tau\|\nabla w_h^i\|^2 + 2(E_{LdG}(Q_h^{i+1}) - E_{LdG}(Q_h^i)) \le 0.$$

Therefore, $E_{LdG}(Q_h^{i+1}) \le E_{LdG}(Q_h^i) \le C_0$ and this allows us to proceed by induction. Summing over i from 0 to J yields

$$E_{LdG}(Q_h^{J+1}) + C'(\tau/2)\sum_{i=0}^{J}\|\nabla w_h^i\|^2 \le E_{LdG}(Q_h^0). \qquad \square$$

Theorem 1 (Termination and convergence to a discrete Q harmonic tensor field). *Suppose that the conditions of Lemmas 5 and 6 are satisfied. Then the tensor field algorithm terminates within a finite number of iterations and the output $Q_h^* \in \mathscr{A}_{LdG}^h$ satisfies*

$$\left(\nabla Q_h^*, \nabla \mathscr{I}_h[n_h^{Q,*} \otimes v_h + v_h \otimes n_h^{Q,*}]\right) = \mathscr{R}es(v_h)$$

for all $v_h \in \mathscr{F}_{\mathbb{S}^2}[n_h^{Q,*}]$*, where the linear functional* $\mathscr{R}es : \mathscr{F}_{\mathbb{S}^2}[n_h^{Q,*}] \to \mathbb{R}$ *satisfies* $|\mathscr{R}es(v_h)| \leq \varepsilon \|\nabla v_h\|^2$ *for all* $v_h \in \mathscr{F}_{\mathbb{S}^2}[n_h^{Q,*}]$*. For a sequence* $(\varepsilon_J)_{J\in\mathbb{N}}$ *of positive numbers such that* $\varepsilon_J \to 0$ *as* $J \to \infty$*, every accumulation point of the corresponding bounded sequence of outputs* $(Q_h^{*,J})_{J\in\mathbb{N}} \subset \mathscr{A}_{LdG}^h$ *of the algorithm is a discrete* Q *harmonic line field according to Definition 2.*

Proof. The proof is a direct consequence of Theorem 3.2.7 from [5]. □

5.2 Weak Compactness Result for Q Harmonic Tensor Fields

For ease of presentation we assume homeotropic boundary conditions on the entire boundary and let $(Q_\ell)_\ell \subset W^{1,2}(\Omega;\mathbb{L}^2)$ be a bounded sequence of Q harmonic tensor fields. Then there exists $\pm n_\ell : \Omega \to \mathbb{S}^2$ satisfying $Q_\ell(x) = n_\ell(x) \otimes n_\ell(x)$ for almost every $x \in \Omega$. Note that, on every simply connected $\omega \subset \Omega$, we can chose $\pm n_\ell \in W^{1,2}(\omega, \mathbb{S}^2)$. Moreover, we have that

$$(\nabla Q_\ell, \nabla V) = 0$$

for all $V \in W_0^{1,2}(\Omega;\mathbb{R}^{3\times 3})$ satisfying $V(x) \in T_{Q_\ell(x)}\mathbb{L}^2$ for almost every $x \in \Omega$. We will show convergence on every simply connected $\omega \subset \Omega$. For this let $V \in W_0^{1,2}(\Omega;\mathbb{R}^{3\times 3})$ be such that $\mathrm{supp} V \subset \omega$ and $V(x) \in T_Q(x)\mathbb{L}^2$ for almost every $x \in \omega$. Thus, there exists $v \in W_0^{1,2}(\omega, \mathbb{R}^3)$ satisfying $v(x) \in T_{n_\ell(x)}\mathbb{S}^2$ for almost every $x \in \omega$ and $V = n_\ell \otimes v + v \otimes n_\ell$. Furthermore, we can rewrite $v = n_\ell \times \zeta$ for some function $\zeta \in W_0^{1,2}(\omega, \mathbb{R}^3)$ and the usual cross product ×. From the boundedness of $(Q_\ell)_\ell$ we infer that for (not relabeled) subsequences

$$Q_\ell \rightharpoonup Q \text{ in } W^{1,2}, \quad Q_\ell \to Q \text{ in } L^2 \quad \text{and} \quad Q_\ell \to Q \text{ pointwise almost everywhere in } \Omega.$$

Since $Q_\ell = n_\ell \otimes n_\ell$ almost every, we know that $n_\ell^i n_\ell^j \to n^i n^j$ pointwise almost everywhere for $i, j = 1, \ldots, 3$. We proceed

$$\begin{aligned}
(\nabla Q_\ell, \nabla V) &= (\nabla Q_\ell, \nabla(n_\ell \otimes v + v \otimes n_\ell)) \\
&= (\nabla Q_\ell, \nabla(n_\ell \otimes (n_\ell \times \zeta) + (n_\ell \times \zeta) \otimes n_\ell)) \\
&= \sum_{k=1}^{d} (\partial_k Q_\ell, \partial_k(n_\ell \otimes (n_\ell \times \zeta) + (n_\ell \times \zeta) \otimes n_\ell)) \\
&= \sum_{k=1}^{d} ((\partial_k n_\ell) \otimes n_\ell + n_\ell \otimes (\partial_k n_\ell), \partial_k(n_\ell \otimes (n_\ell \times \zeta) + (n_\ell \times \zeta) \otimes n_\ell)).
\end{aligned}$$

For $a, b, c, d \in \mathbb{R}^3$ we have that $(a \otimes b, c \otimes d) = (a^T d, b^T c)$. Since $n_\ell \perp \partial_k n_\ell$ for $k = 1, \dots, d$ and $n_\ell \perp n_\ell \times \zeta$ we see that terms of the form

$$(\partial_k n_\ell \otimes n_\ell, \partial_k n_\ell \otimes (n_\ell \times \zeta)) = ((\partial_k n_\ell)^T (n_\ell \times \zeta), n_\ell^T (\partial_k n_\ell))$$

vanish and we obtain the identity

$$\begin{aligned}(\nabla Q_\ell, \nabla V) &= \sum_{k=1}^{d} ((\partial_k n_\ell) \otimes n_\ell + n_\ell \otimes (\partial_k n_\ell), n_\ell \otimes (n_\ell \times \partial_k \zeta) + (n_\ell \times \partial_k \zeta) \otimes n_\ell) \\ &= \sum_{k=1}^{d} (\partial_k Q_\ell, n_\ell \otimes (n_\ell \times \partial_k \zeta) + (n_\ell \times \partial_k \zeta) \otimes n_\ell).\end{aligned}$$

The products $n_\ell \otimes (n_\ell \times \partial_k \zeta)$ and $(n_\ell \times \partial_k \zeta) \otimes n_\ell$ are quadratic in the components of n_ℓ and therefore we have that $n_\ell \otimes (n_\ell \times \partial_k \zeta) \to n \otimes (n \times \partial_k \zeta)$ pointwise almost everywhere in Ω. Since $|n_\ell| = 1$ and $\zeta \in L^\infty$ we have by Lebesgue's dominated convergence that $n_\ell \otimes (n_\ell \times \partial_k \zeta) \to n \otimes (n \times \partial_k \zeta)$ strongly in L^2. Together with the weak convergence of $\partial_k Q_\ell$ we infer that

$$\begin{aligned}0 = (\nabla Q_\ell, \nabla V) \longrightarrow &\sum_{k=1}^{d} (\partial_k Q, n \otimes (n \times \partial_k \zeta) + (n \times \partial_k \zeta) \otimes n) \\ &= (\nabla Q, \nabla (n \otimes (n \times \zeta) + (n \times \zeta) \otimes n).\end{aligned}$$

Since this holds for all ζ and the previous arguments are independent of $\omega \subset \Omega$ we have that Q is a harmonic line field.

6 Numerical Experiments

6.1 Extinction of Singularities

We consider a liquid crystal cell $V = (-1, 1)^3 \subset \mathbb{R}^3$ with planar anchoring conditions. In this case defects at the boundary can be observed leading to so called *Schlieren textures*. There are different types of defects (disclinations) and to each type is assigned a number and a sign. Some of them may cancel out each other if they come into contact. We consider the upper boundary of V and simulate the annihilation of opposite degree one-half and degree one singularities in the iteration of the algorithm. The preference of the alignment parallel to the surface $\Omega = (-1, 1)^2 \times \{1\}$ is modelled by the use of a Ginzburg-Landau penalty-term. Thus, we consider

$$E^\varepsilon_{LdG}(Q) = \frac{1}{2} \int_\Omega |\nabla Q|^2 \, \mathrm{d}x + \frac{1}{2\varepsilon^2} \int_\Omega |Q_{33}|^2 \, \mathrm{d}x.$$

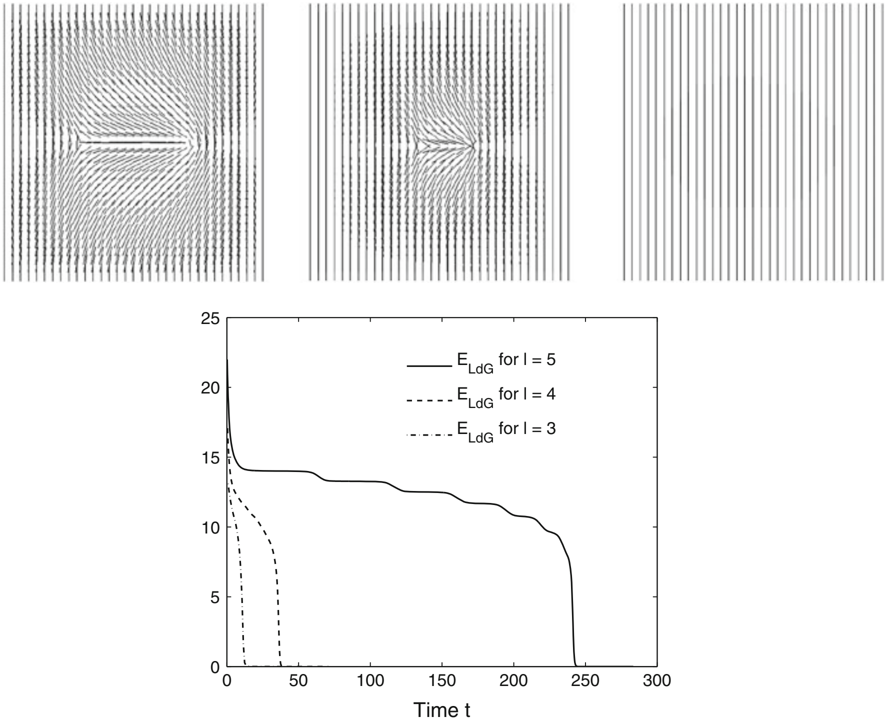


Fig. 3 Annihilation of two opposite degree one-half singularities during the computation and an energy plot demonstrating the decay of energy for different mesh-sizes. The energy shows a strong decay when the attracting defects eventually annihilate

Penalizing the out of plane component is physically consistent since the alignment parallel to Ω is favored but not forced. Mathematically this is crucial since singularities in the plane have infinite energy. Let $\mathscr{T}_{h,0}$ be a triangulation of Ω consisting of two triangles obtained by dividing $(-1,1)^2$ along the diagonal $x_1 = x_2$. The sequence of triangulations $\mathscr{T}_\ell$ is generated by ℓ uniform refinements of $\mathscr{T}_0$ with mesh-size $h_\ell = \sqrt{2}2^{-\ell}$. We use a time-step size $\tau = 5h$ and set $\varepsilon = 10^{-1}$. In our first experiment we examine the extinction of two opposite degree one-half singularities. We place a positive degree one-half singularity at $x_1 = 0.5$ and a negative degree one-half singularity at $x_1 = -0.5$. Boundary values are chosen to be $n_D = [0,1,0]^T$. The unique minimizer of E^{ε}_{LdG} is given by $u = [0,1,0]^T$. For the construction of such initial defect data we refer the reader to [8, 16]. In a second experiment we place a negative degree one singularity at $x_1 = 0$ and two positive degree one-half singularities at $x_1 = -0.3$ and $x_1 = 0.7$. As in the first experiment the boundary values are $n_D = [0,1,0]^T$. Snapshots of the evolution and decay of energy in the two examples can be seen in Figs. 3 and 4.

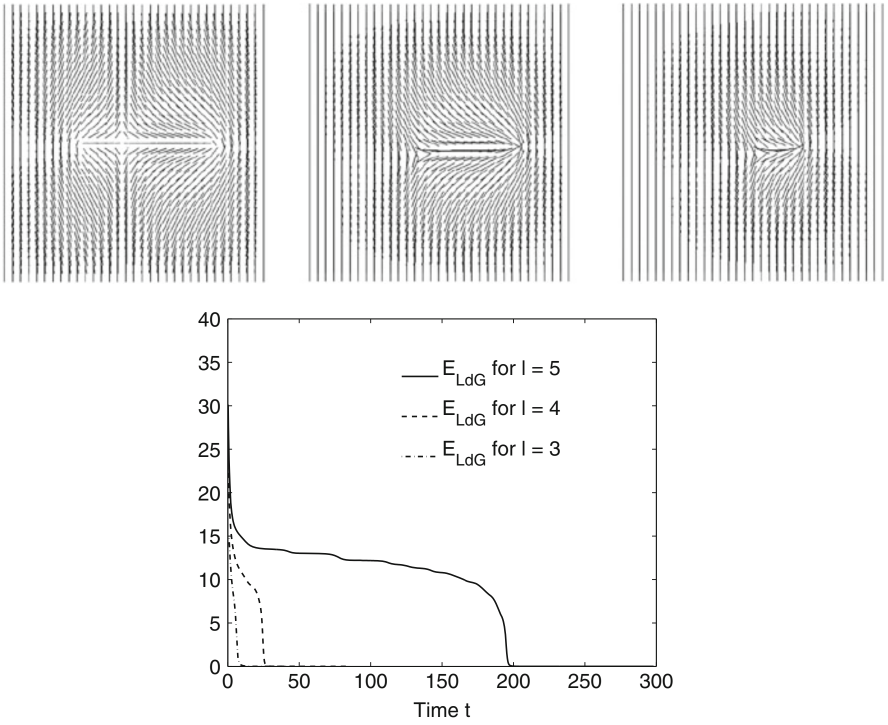


Fig. 4 Extinction of three singularities during the computation and an energy plot demonstrating the decay of energy for different mesh-sizes: The nearby negative degree one and positive degree one-half singularities come together and result in a negative degree one-half singularity. Then, as in the first experiment an annihilation takes place when the remaining singularities meet. The energy shows strong decays when the annihilations take place

6.2 Orientability Versus Non-orientability

Let $D_1 := (-1.5, 1.5) \times (-1, 1)$, let $D_2 := B_1([-1.5, 0]^T) \cup B_1([1.5, 0]^T)$, let $D_3 := B_{1/2}([-1.5, 0]^T) \cup B_{1/2}([1.5, 0]^T)$ and let $D := (D_1 \cup D_2) \setminus D_3$, see Fig. 5. We use the DistMesh package [15] to generate quasi-uniform triangulations of D with arbitrary mesh-size. Thus, the quantities $h = 0.1, 0.075$ and $h = 0.05$ in Fig. 5 are approximate values to the actual mesh-sizes according to the definition in Sect. 3. The two-dimensional domain D was introduced in [4] to point out that there exist settings in which the Q-tensor theory yields stable configurations that cannot be seen by the classical Oseen-Frank model. We impose tangential boundary conditions on the outer part of the boundary and Neumann conditions at the interior. Furthermore we consider the energy

$$E^{\varepsilon}_{LdG}(Q) = \frac{1}{2} \int_{\Omega} |\nabla Q|^2 \, \mathrm{d}x + \frac{1}{2\varepsilon^2} \int_{\Omega} |Q_{33}|^2 \, \mathrm{d}x,$$

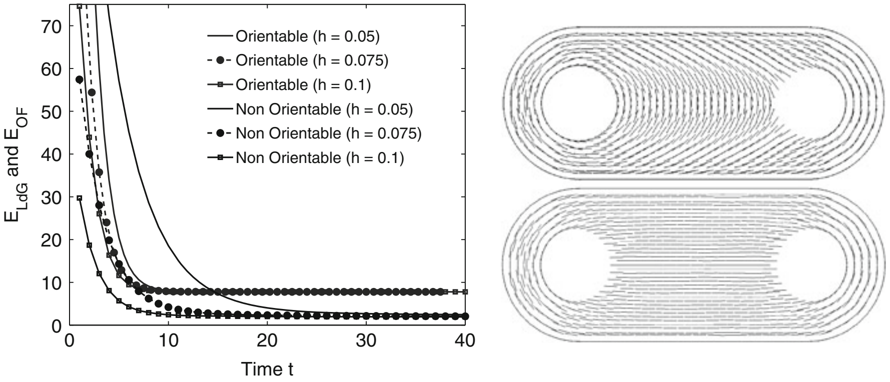


Fig. 5 Energy decay for different mesh-sizes and final states for a triangulation of D with mesh size $h = 0.1$. The energy of the final state in the class of non-orientable line fields is strictly smaller than the energy in the class of orientable line fields. The classical Oseen-Frank theory would not detect the absolute minimum prefered by the liquid crystal

which allows for an out-of-plane component and thereby for singularities in the interior as in Sect. 6.1. We compute stable configurations in the class of orientable and non-orientable line fields with our algorithms from Sect. 4, see Figs. 5 and 6. We observe for a sequence of triangulations with approximate mesh-sizes $h = 0.1, 0.075$ and $h = 0.05$ that the energies in the class of non-orientable line fields are strictly smaller than the energies in the class of orientable line fields. Thus, the classical Oseen-Frank theory fails to detect stable configurations of the liquid crystal with small energy.

6.3 Torus Experiments

We investigate stable configurations of line fields on a vertically stretched torus $\mathbb{T}^2$ which can be parametrized by $X : (0, 2\pi)^2 \to \mathbb{R}^3$,

$$(\varphi, \theta) \mapsto \begin{bmatrix} (R + r\cos\theta)\cos\varphi \\ (R + r\cos\theta)\sin\varphi \\ 2.5r\sin\theta \end{bmatrix}$$

with $R > r > 0$, see Fig. 7. Planar anchoring conditions are imposed everywhere on the surface and we compute the tangent vectors

$$\tau_1 := \frac{\partial_\varphi X}{|\partial_\varphi X|} \quad \text{and} \quad \tau_2 := \frac{\partial_\theta X}{|\partial_\theta X|},$$

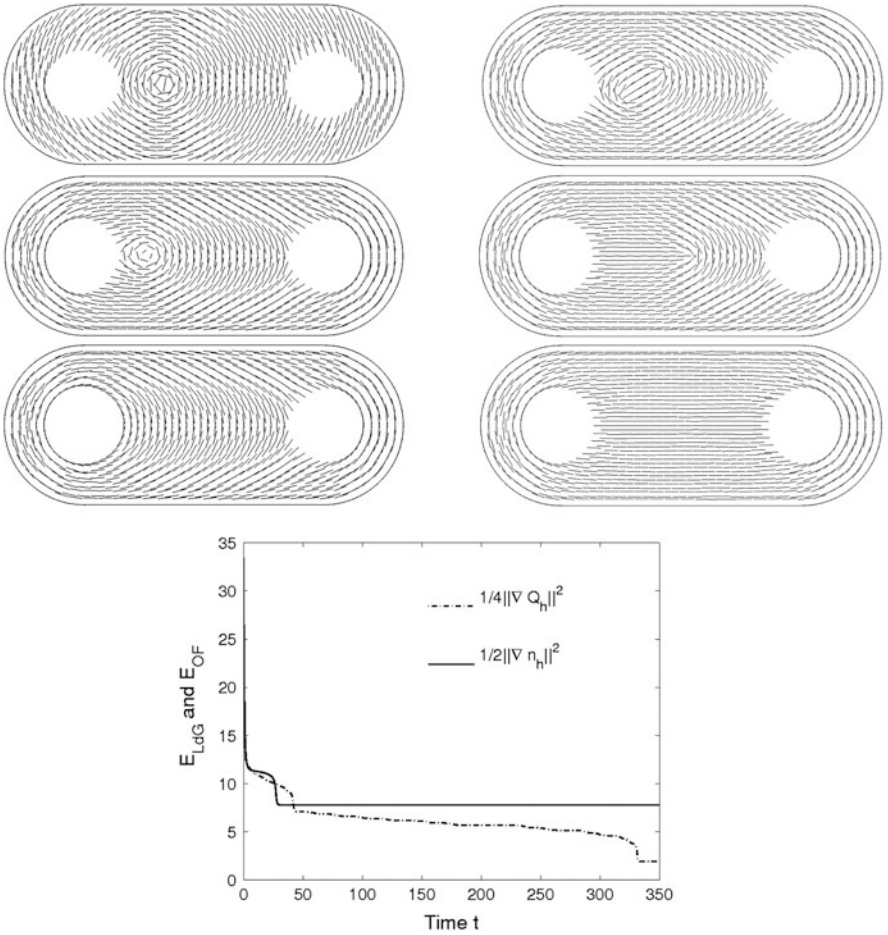


Fig. 6 Snapshots of an evolution under the H^1 gradient flow and decay of energy for director fields (*left*) and line fields (*right*): The initial line field Q_h^0 admits a positive degree one singularity at $x_1 = -0.5$, the holes are located at $x_1 = \pm 1.5$ and have a radius $r = 0.5$. In the orientable case the singularity moves into the hole on the right. In the non-orientable case the singularity splits into two positive degree one-half singularities that repulse from each other and vanish in the holes leading to a strictly smaller energy

as well as the unit normal outer normal $\nu = \frac{\tau_1 \times \tau_2}{|\tau_1 \times \tau_2|}$. Let $\tilde{\mathscr{T}}_0$ be a triangulation of $(0, 2\pi)^2$ consisting of two triangles obtained by dividing $(0, 2\pi)^2$ along the diagonal $x_1 = x_2$. The sequence of triangulations $\tilde{\mathscr{T}}_\ell$ with mesh-size $\tilde{h}_\ell = \sqrt{2}(2\pi)2^{-\ell}$ and nodes $\tilde{\mathscr{N}}_\ell$ is generated by ℓ uniform refinements of $\tilde{\mathscr{T}}_0$. We identify the following nodes

$$[0, \xi_{\tilde{h}_\ell}]^T \longleftrightarrow [2\pi, \xi_{\tilde{h}_\ell}]^T \quad \text{and} \quad [\xi_{\tilde{h}_\ell}, 0]^T \longleftrightarrow [\xi_{\tilde{h}_\ell}, 2\pi]^T$$

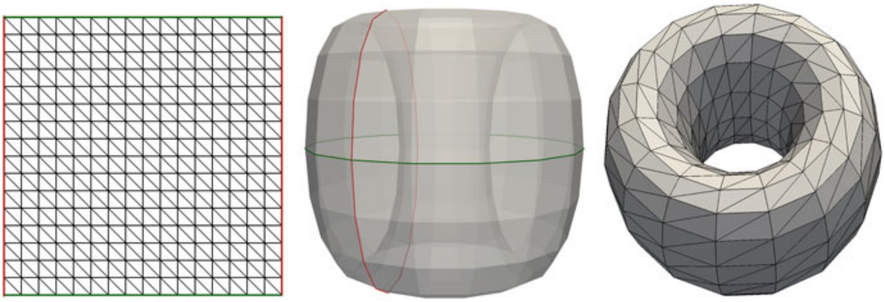


Fig. 7 Generating the triangulation of a torus: On a uniform triangulation of $(0, 2\pi)^2$ we identify the nodes on the left with the ones on the right (*red lines*) and the nodes on the *top* with the ones on the bottom (*green lines*). We plot the obtained stretched torus for $\ell = 4$, $r = 1$ and $R = 2$ and the two identification lines (*middle*) as well as the resulting mesh (*right*)

for $\xi_{\tilde{h}_\ell} = 0, \tilde{h}/\sqrt{2}, 2\tilde{h}/\sqrt{2}, \dots, 2\pi$. By this, we obtain a new triangulation $\mathscr{T}_\ell$ with a new set of nodes $\overline{\mathscr{N}_\ell}$ and define

$$\mathscr{N}_\ell := \left\{X(z) : z \in \overline{\mathscr{N}_\ell}\right\}.$$

This results in a closed triangulated surface $\mathbb{T}^2_h$ approximating $\mathbb{T}^2$ with a new mesh size $h_\ell = ||DX||_{L^\infty}\tilde{h}_\ell > 0$, where $||DX||_{L^\infty} := \max_{ij}||\partial_i X_j||_{L^\infty} = \max\{R + r, 2.5r\}$. On $\mathbb{T}^2_h$ we define the discrete tangent vector fields

$$\begin{aligned}\tilde{n}^0_{h,1} &:= \mathscr{I}_h[\tau_2 + \text{rand}]\\ \tilde{n}^0_{h,2} &:= \mathscr{I}_h[\sin(\varphi/2)\tau_1 + \sin(\varphi/2)\tau_2 + \text{rand}],\end{aligned}$$

where $\text{rand} : \mathbb{T}^2_h \to \mathbb{R}^3$ takes random values in $(-0.1, 0.1)^3$ and φ denotes the horizontal angle in the torus coordinates defined by the parametrization X. To obtain a vector field that is tangential and has unit length at the nodes we define $\overline{n}^0_{h,i} := \mathscr{I}_h[\tilde{n}^0_{h,i} - (\tilde{n}^0_{h,i} \cdot \nu)\nu]$ for $i = 1, 2$ and then the initial data

$$n^0_{h,i} := \mathscr{I}_h\left[\frac{\overline{n}^0_{h,i}}{|\overline{n}^0_{h,i}|}\right] \quad \text{for} \quad i = 1, 2.$$

As can be seen in Fig. 8 the initial line field $Q^0_{h,1} := \mathscr{I}_h[n^0_{h,1} \otimes n^0_{h,1}]$ is orientable while $Q^0_{h,2} := \mathscr{I}_h[n^0_{h,2} \otimes n^0_{h,2}]$ is a Moebius strip rotated around the x_3-axis and, therefore, non-orientable.

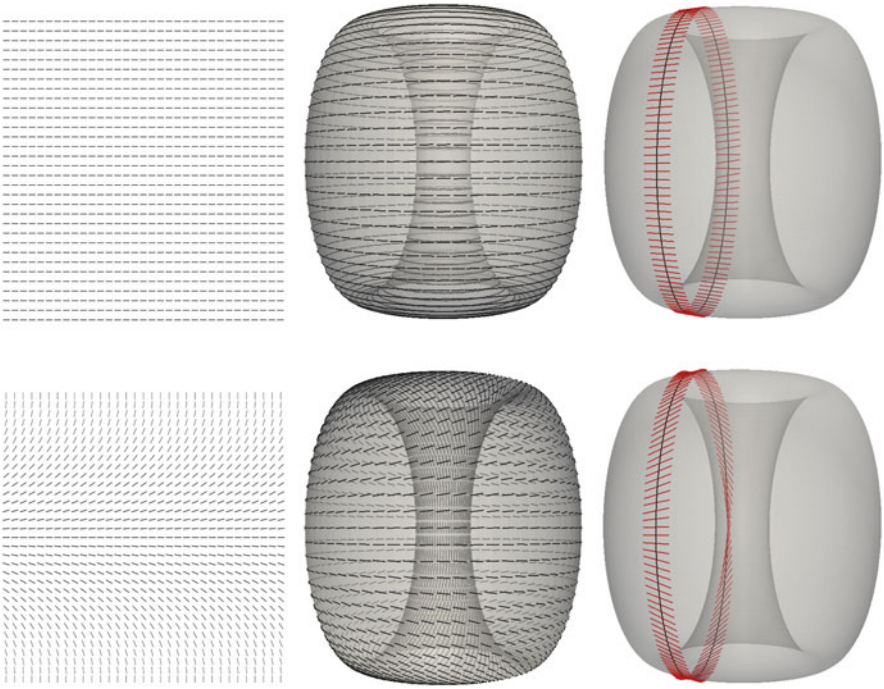


Fig. 8 Initial data for the torus experiments: On the fundamental domain $(0, 2\pi)^2$ of $\mathbb{T}^2$ we plot the line fields $Q^0_{h,1}$ and $Q^0_{h,2}$. Since $Q^0_{h,2}$ is a Moebius strip in one direction the resulting line field on the torus is non-orientable

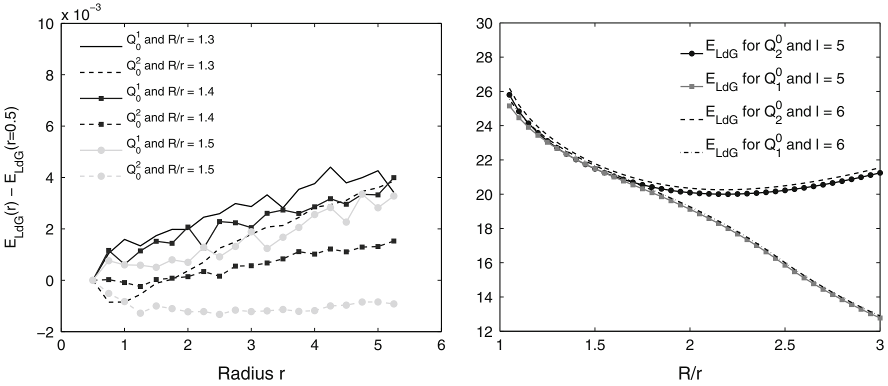


Fig. 9 Numerically confirmed invariance of the energy E_{LdG} under a rescaling $x \mapsto \lambda x$ of the surface (*left*) and final energies for ratios $R/r \in (1.05, 3)$ and two meshes (*right*). On the *left* we plot $E_{LdG}(r) - E_{LdG}(r = 0.5)$ for the ratios $R/r = 1.3, 1.4$ and 1.5 and initial data $Q^0_{h,1}$ and $Q^0_{h,2}$. On the *right* we plot the final energies for $r = 1$ and initial data $Q^0_{h,1}$ and $Q^0_{h,2}$. Note that the energies of the final states in the class of orientable line fields is much smaller than the energies of the non-orientable final states for $R/r > 2$ and they are of comparable size for $R/r \sim 1.5$

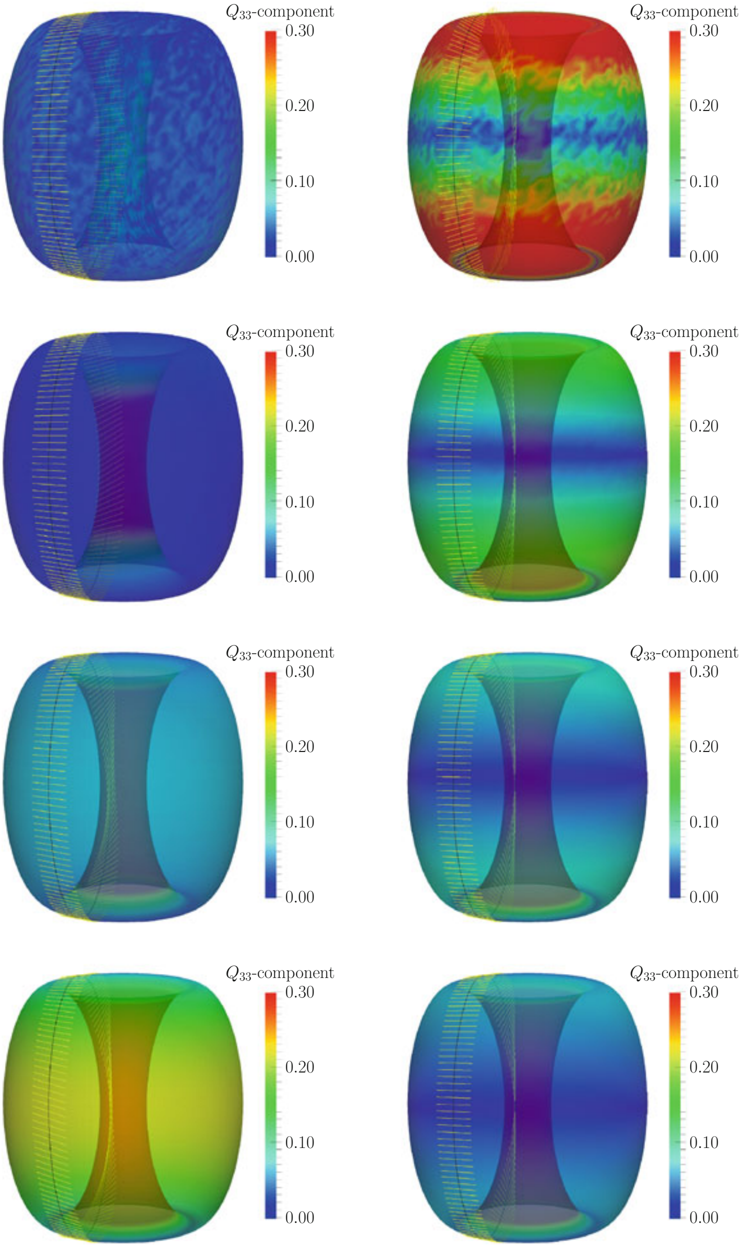


Fig. 10 Stationary points of E_{LdG} in the class of orientable (*left column*) and non-orientable (*right column*) tangential line fields on the torus: Snapshots of the evolution under the gradient flow with initial data $Q^0_{h,1}$ (*left*) and $Q^0_{h,2}$ (*right*). On the outer part of the torus we see different tilts of the liquid crystal when the evolution becomes stationary. We color the surface by $Q_{h,33} = |n^Q_{h,3}|^2$, where the color blue corresponds to $Q_{h,33} = 0$, that is, the molecules of the liquid crystal lie in a plane orthogonal to $e_3 = [0, 0, 1]^T$

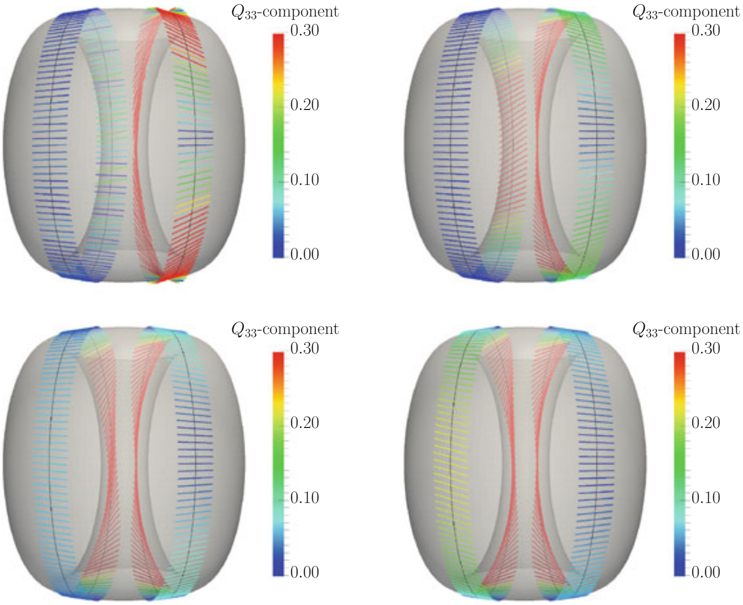


Fig. 11 Stationary points of E_{LdG} on the torus (colored by the Q_{33} component): We plot snapshots of a cut through the non-orientable (*left*) and orientable (*right*) line fields during the evolution

6.3.1 Different Ratios R/r

Since E_{LdG} and the corresponding Euler-Lagrange equations are invariant under a rescaling $x \mapsto \lambda x$ for $\lambda > 0$ on two-dimensional surfaces stationary points computed by a gradient flow algorithm only depend on the initial data Q^0 and the ratio R/r of the two radii that define the torus. We start the investigation of tangential line fields by computing final energies for the starting values $Q^0_{h,1}$, $Q^0_{h,2}$ and different ratios R/r. As can be seen in Fig. 9 on the right for $R/r > 2$ and our choice of the initial data the energies of the final states in the class of orientable line fields is much smaller than the energies of the non-orientable configurations. We will have a closer look at the ratio $R/r = 1.5$ where the energies are of comparable size and discuss properties of the stable orientable and non-orientable line fields.

6.3.2 Analyzing the Tilt for $R/r = 1.5$

The interest for physicists and engineers involved in the construction of bistable and multistable devices is the difference in the tilt of liquid crystal molecules in stable configurations. The tilt of liquid crystal molecules has an impact on polarized light crossing the device possibly leading to a new polarization. A polarizer at the end of the device measures the deviation of the outgoing from the ingoing polarization. Since the polarizer passes light of a specific polarisation, say the ingoing one, regions of different polarisations due to tilted molecules appear as darker spots.

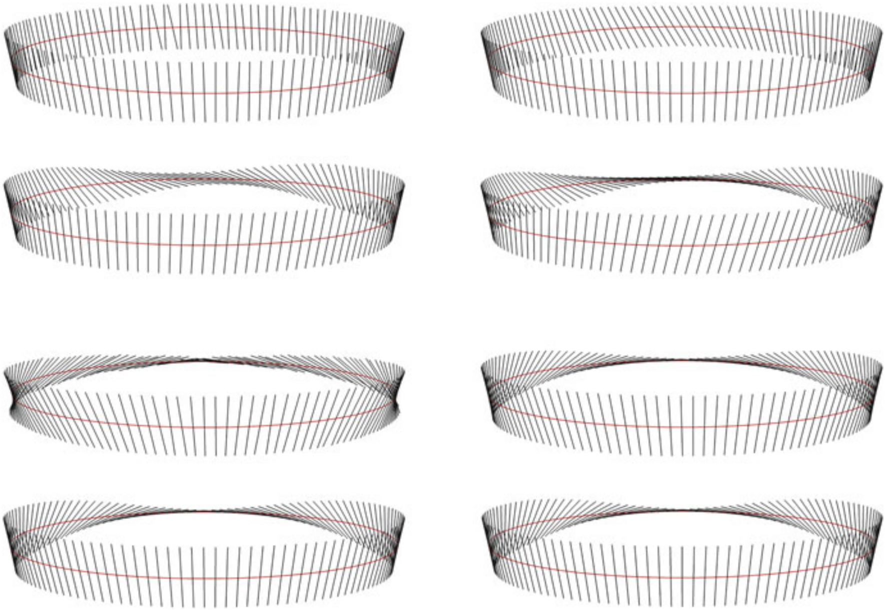


Fig. 12 Stationary points of E_{LdG} in the class of orientable (*upper two rows*) and non-orientable (*lower two rows*) tangential line fields on the torus: Some more snapshots of a cut through the line field. All observed stationary points in the torus experiments are rotational symmetric with respect to the x_3-axis

For a ratio $R/r = 1.5$ and a refinement step $\ell = 6$ we compute stationary points of E_{LdG} using orientable and non-orientable initial data $Q^0_{h,1}$ and $Q^0_{h,2}$, respectively. We measure the tilt of the molecules in terms of the x_3-component $Q_{h,33}$ of the line field. While the tilt of the crystal is almost the same on the inner part of the torus for orientable and non-orientable stable configurations we observe a difference on the outer part, see Figs. 10 and 11. We color all surfaces and line fields by $Q_{h,33} = |n^Q_{h,3}|^2$. In contrast to common bistable devices where different tilts of the crystal are obtained with defect and non-defect states [11, 17] we discuss, here, different stable configurations under the notion of orientability and non-orientability (Fig. 12).

Acknowledgements The authors acknowledge support by the DFG through the Collaborative Research Center (SFB) 611 *Singular Phenomena and Scaling in Mathematical Models*.

References

1. Alouges, F.: Liquid crystal configurations: the numerical point of view. In: Chipot, M., Saint Jean Paulin, J., Shafrir, I. (eds.) Progress in Partial Differential Equations: The Metz Surveys, 3. Volume 314 of Pitman Research Notes in Mathematics Series, pp. 3–17. Longman Scientific & Technical, Harlow (1994)

2. Alouges, F.: A new algorithm for computing liquid crystal stable configurations: the harmonic mapping case. SIAM J. Numer. Anal. **34**(5), 1708–1726 (1997)
3. Alouges, F., Ghidaglia, J.M.: Minimizing Oseen-Frank energy for nematic liquid crystals: algorithms and numerical results. Ann. Inst. H. Poincaré Phys. Théor. **66**(4), 411–447 (1997)
4. Ball, J.M., Zarnescu, A.: Orientability and energy minimization for liquid crystal models. Mol. Cryst. Liq. Cryst. **495**, 573–585 (2008)
5. Bartels, S.: Finite element approximation of harmonic maps between Surfaces. Habilitation thesis, Humboldt Universität zu Berlin, Berlin (2009)
6. Bartels, S.: Numerical analysis of a finite element scheme for the approximation of harmonic maps into surfaces. Math. Comput. **79**(271), 1263–1301 (2010)
7. Bartels, S.: Finite element approximation of large bending isometries. Numerische Mathematik **124**(3), 415–440 (2013)
8. Bartels, S., Dolzmann, G., Nochetto, R., Raisch, A.: Finite element methods for director fields on flexible surfaces. Interfaces Free Bound. **14**(2), 231–272 (2012)
9. Ciarlet, P.G.: The Finite Element Method for Elliptic Problems. Volume 40 of Classics in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (2002). Reprint of the 1978 original [North-Holland, Amsterdam; MR0520174 (58 25001)]
10. de Gennes, P.-G., Prost, J.: The Physics of Liquid Crystals. International Series of Monographs on Physics, 2nd edn. Oxford University Press, New York (1995)
11. Jones, J.C.: 40.1: Invited paper: the zenithal bistable device: from concept to consumer. SID Symp. Dig. Tech. Pap. **38**(1), 1347–1350 (2007)
12. Leslie, F.M.: Theory of flow phenomena in liquid crystals. Adv. Liq. Cryst. **4**, 1–81 (1979)
13. Lin, F., Liu, C.: Static and dynamic theories of liquid crystals. J. Partial Differ. Equ. **14**(4), 289–330 (2001)
14. Majumdar, A., Zarnescu, A.: Landau-De Gennes theory of nematic liquid crystals: the Oseen-Frank limit and beyond. Arch. Ration. Mech. Anal. **196**(1), 227–280 (2010)
15. Persson, P.-O., Strang, G.: A simple mesh generator in Matlab. SIAM Rev. **46**(2), 329–345 (2004). (Electronic)
16. Raisch, A.: Finite element methods for geometric problems. Phd thesis, Rheinische Friedrich-Wilhelms Universität Bonn, Bonn (2012)
17. Uche, C., Elston, S.J., Parry-Jones, L.A.: Microscopic observation of zenithal bistable switching in nematic devices with different surface relief structures. J. Phys. D Appl. Phys. **38**(13), 2283 (2005)
18. Virga, E.G.: Variational Theories for Liquid Crystals. Volume 8 of Applied Mathematics and Mathematical Computation. Chapman & Hall, London (1994)

A Fast and Accurate Numerical Method for the Computation of Unstable Micromagnetic Configurations

Sören Bartels, Mario Bebendorf, and Michael Bratsch

Abstract We present a fast and accurate numerical method to compute unstable micromagnetic configurations. The proposed scheme, which combines various state of the art methods, is able to treat the pointwise unit-length constraint of the magnetization field and to efficiently compute the stray field energy. Furthermore, numerical results are presented which are in agreement with the expected results in simple situations and allow predictions beyond theory.

1 Introduction

The computation of minimal switching energies between two given stable states and the detection of a corresponding unstable critical configuration is an important task in the mathematical modeling of many physical phenomena [2, 20, 29]. In this paper we address this problem and thereby aim at contributing to the understanding of the energy landscape for a mathematically challenging and well established model energy functional in micromagnetics; cf. [9, 22]. Its particular features are that it involves a unit-length constraint for the magnetization field and requires the computation of a stray field.

The finite element treatment of minimization problems and partial differential equations with pointwise constraints has been investigated intensively in recent years [1, 2]. Reliable and efficient methods are now available that typically impose the constraint at the nodes of a triangulation [4,5]. In iterative schemes the constraint

S. Bartels (✉)
Universität Freiburg, Hermann-Herder-Str. 10, D-79104 Freiburg, Germany
e-mail: bartels@mathematik.uni-freiburg.de

M. Bebendorf · M. Bratsch
Institut für Numerische Simulation, Rheinische Friedrich-Wilhelms-Universität Bonn,
Wegelerstr. 6, D-53115 Bonn, Germany
e-mail: bebendorf@ins.uni-bonn.de; bratsch@ins.uni-bonn.de

M. Griebel (ed.), *Singular Phenomena and Scaling in Mathematical Models*,
DOI 10.1007/978-3-319-00786-1_18,

is linearized and afterwards updates are made which consist of corrections of tangent spaces and a subsequent nodewise nearest-neighbor projection onto the given target manifold which lead to linear systems of equations.

Computing the solution of an exterior domain problem is often formulated as a boundary equation with a non-local operator and then requires the solution of linear systems of equations with large, fully populated matrices. This is can be done efficiently with the technique of so-called $\mathscr{H}$-matrices which can approximate the arising matrices with almost linear complexity, cf. [7, 16, 24].

Various methods are available to compute unstable critical points, i.e., saddle points, of energy functionals known as mountain pass algorithms [8]. They typically assume that the functional under consideration is defined on a linear space and therefore cannot be employed if this is not the case, i.e., if the linear interpolation between admissible configurations does not belong to the domain of the functional. A method that is capable to cope with related difficulties is the recently developed string method, see [28, 30], which evolves an entire path that connects two given states in the set of admissible configurations.

For the Landau-Lifschitz energy in micromagnetics certain stable critical configurations such as the so-called flower and vortex state are known [19, 26]. To efficiently switch between such states, e.g., by an applied field, it is important to determine the minimal energy required to achieve this change. We use the string method in combination with finite element methods to deal with the pointwise unit-length constraint and the $\mathscr{H}$-matrix technology to efficiently compute this energy and to identify a corresponding magnetization. The resulting numerical method is verified for a standard problem [23] and a so-called minimum energy path connecting the flower and vortex state is presented.

2 The Landau-Lifschitz Model

Let $\Omega \subset \mathbb{R}^3$ be a domain and let the magnetization $\mathbf{m} : \mathbb{R}^3 \to \mathbb{R}^3$ satisfying $\|\mathbf{m}(x)\|_2 = 1$ in Ω and $\mathbf{m}(x) = 0$ for all $x \notin \Omega$ be given. The energy associated with $\mathbf{m}$ is

$$E(\mathbf{m}) := \frac{1}{2}\int_\Omega \|D\mathbf{m}\|_F^2 \,\mathrm{d}x + \int_\Omega \varphi(\mathbf{m}) - \mathbf{f}\cdot\mathbf{m}\,\mathrm{d}x + E_s(\mathbf{m}), \qquad (1)$$

where $\varphi(\mathbf{m}) := 1 - (\mathbf{e}\cdot\mathbf{m})^2$ with given $\mathbf{e}, \mathbf{f} \in \mathbb{R}^3$ satisfying $\|\mathbf{e}\|_2 = 1$ and

$$E_s(\mathbf{m}) := \frac{\mu_0}{2}\int_{\mathbb{R}^3} \|H\|_2^2 \,\mathrm{d}x, \quad \mu_0 := 4\pi\cdot 10^{-7},$$

denotes the energy of the stray field H corresponding to $\mathbf{m}$. H can be computed from the Maxwell equations in the absence of electrical currents and charges

$$\operatorname{div} B = 0, \tag{2a}$$

$$\operatorname{curl}\ H = 0. \tag{2b}$$

H and the magnetic induction B are coupled by the equation $B = \mu_0(H + \mathbf{m})$.

From Eq. (2b) it follows that there is the so-called magnetostatic potential $u_{\mathbf{m}} : \mathbb{R}^3 \to \mathbb{R}$ satisfying $H = -\nabla u_{\mathbf{m}}$. Then (2a) becomes

$$\Delta u_{\mathbf{m}} = \operatorname{div} \mathbf{m},$$

which is equivalent with the weak formulation

$$\int_{\mathbb{R}^3} \nabla u_{\mathbf{m}} \cdot \nabla w \,\mathrm{d}x = \int_{\Omega} \mathbf{m} \cdot \nabla w \,\mathrm{d}x \quad \text{for all } w \in H^1(\mathbb{R}^3). \tag{3}$$

Notice that this defines a linear mapping $\mathbf{m} \mapsto u_{\mathbf{m}}$ with

$$\|\nabla u_{\mathbf{m}}\|_{L^2(\mathbb{R}^3)} \leq \|\mathbf{m}\|_{L^2(\Omega)}. \tag{4}$$

Using $w = u_{\mathbf{m}}$ in (3), it follows that

$$E_s(\mathbf{m}) = \frac{\mu_0}{2} \int_{\Omega} \mathbf{m} \cdot \nabla u_{\mathbf{m}} \,\mathrm{d}x. \tag{5}$$

In addition to the energy $E(\mathbf{m})$ also its derivative $E'(\mathbf{m})[\mathbf{v}]$ will be important. Let the direction $\mathbf{v}$ be given such that $\mathbf{v}(x) \cdot \mathbf{m}(x) = 0$ for almost every $x \in \Omega$. Then we obtain that

$$E'(\mathbf{m})[\mathbf{v}] = \int_{\Omega} \operatorname{trace}(D\mathbf{m})^T (D\mathbf{v}) \,\mathrm{d}x - 2\int_{\Omega} (\mathbf{e}\cdot\mathbf{m})(\mathbf{e}\cdot\mathbf{v}) \,\mathrm{d}x - \int_{\Omega} f \cdot \mathbf{v} \,\mathrm{d}x + E_s'(\mathbf{m})[\mathbf{v}],$$

where

$$E_s'(\mathbf{m})[\mathbf{v}] = \mu_0 \int_{\mathbb{R}^3} H(\mathbf{v}) H(\mathbf{m}) \,\mathrm{d}x = \mu_0 \int_{\Omega} \mathbf{v} \cdot \nabla u_{\mathbf{m}} \,\mathrm{d}x$$

due to (3).

3 Efficient Computation of the Stray Field Energy

The stray field energy represents the non-local effects of the magnetization. Hence, it is the numerically most challenging part of the computation of the Landau-Lifschitz model; see (1). In the following, a reformulation will be shown so that $\mathscr{H}$-matrices can be applied to approximate the local and non-local parts of the stray field energy.

3.1 Different Formulations

Our aim is to find an explicit expression for the magnetostatic potential $u_{\mathbf{m}}$. Equation (3) is equivalent to the following boundary value problem

$$\Delta u_{\mathbf{m}} = \begin{cases} \operatorname{div} \mathbf{m}, & \text{in } \Omega, \\ 0, & \text{in } \Omega^c := \mathbb{R}^3 \setminus \overline{\Omega}, \end{cases} \tag{6a}$$

$$[u_{\mathbf{m}}] = 0 \text{ on } \partial\Omega, \tag{6b}$$

$$[\partial_\nu u_{\mathbf{m}}] = -\mathbf{m} \cdot \nu \text{ on } \partial\Omega, \tag{6c}$$

which has the solution

$$u_{\mathbf{m}}(x) = \int_\Omega \nabla S(x-y) \cdot \mathbf{m}(y)\,\mathrm{d}y,$$

where $S(x) := -\frac{1}{4\pi}\|x\|_2^{-1}$ denotes the singularity function of the Laplacian. Note that $[\cdot]$ in (6) denotes the jump across the boundary $\partial\Omega$.

The computation of the stray field energy using the latter representation of $u_{\mathbf{m}}$ in combination with hierarchical matrices was already done in [24]. We favor the following representation (see [11]), because it leads to the interaction of Ω with its boundary $\partial\Omega$. Let u_1 and u_2 satisfy the following boundary value problems

$$\begin{aligned} \Delta u_1 &= \operatorname{div} \mathbf{m} \text{ in } \Omega, \\ u_1 &= 0 \text{ on } \partial\Omega \end{aligned}$$

and

$$\Delta u_2 = 0 \text{ in } \Omega \cup \Omega^c, \tag{7a}$$

$$[u_2] = 0 \text{ on } \partial\Omega, \tag{7b}$$

$$[\partial_\nu u_2] = g \text{ on } \partial\Omega, \tag{7c}$$

where $g := (\nabla u_1 - \mathbf{m}) \cdot \nu$. Then $u_{\mathbf{m}} = u_1 + u_2$ and the solution of the homogeneous problem (7) is

$$u_2(x) = \int_{\partial\Omega} S(x-y) g(y)\,\mathrm{d}s_y.$$

Hence, from (5)

$$E_s(\mathbf{m}) = \frac{\mu_0}{2}\int_\Omega \mathbf{m} \cdot \nabla u_1\,\mathrm{d}x + \frac{\mu_0}{2}\int_\Omega \int_{\partial\Omega} \mathbf{m}(x) \cdot \nabla S(x-y) g(y)\,\mathrm{d}s_y\,\mathrm{d}x.$$

From

$$\int_\Omega \mathbf{m}(x)\cdot\nabla S(x-y)\,\mathrm{d}x = -\int_\Omega S(x-y)(\operatorname{div}\mathbf{m})(x)\,\mathrm{d}x + \int_{\partial\Omega} S(x-y)\mathbf{m}(x)\cdot\nu_x\,\mathrm{d}s_x$$

we obtain that

$$\begin{aligned} E_s(\mathbf{m}) = \frac{\mu_0}{2}\int_\Omega \mathbf{m}\cdot\nabla u_1\,\mathrm{d}x - \frac{\mu_0}{2}\int_\Omega\int_{\partial\Omega} S(x-y)(\operatorname{div}\mathbf{m})g(y)\,\mathrm{d}s_y\,\mathrm{d}x \\ + \frac{\mu_0}{2}\int_{\partial\Omega}\int_{\partial\Omega} S(x-y)\mathbf{m}(x)\cdot\nu_x\,g(y)\,\mathrm{d}s_y\,\mathrm{d}s_x. \end{aligned}$$

3.2 Discretization

We assume that the computational domain Ω is decomposed into a set of tetrahedra $\mathscr{T}_h$ such that $\Omega = \cup_{\tau\in\mathscr{T}_h}\tau$. The finite element space consisting of linear ansatz functions $\Phi = (\varphi_i)_{i\in I}$ is denoted by $\mathscr{S}^1(\mathscr{T}_h)$, the corresponding set of nodes will be referred to as $\mathscr{N}_h$. We discretize the magnetization $\mathbf{m}_h \in \mathscr{S}^1(\mathscr{T}_h)^3$ such that

$$\mathbf{m}_h = \sum_{i\in I}\boldsymbol{\alpha}_i\varphi_i, \quad \boldsymbol{\alpha}_i \in \mathbb{R}^3.$$

Then $D\mathbf{m}_h = \sum_{i\in I}\boldsymbol{\alpha}_i(\nabla\varphi_i)^T$, and the first term in (1) can be computed from

$$\begin{aligned} \int_\Omega \|D\mathbf{m}_h\|_F^2\,\mathrm{d}x &= \int_\Omega \operatorname{trace}(D\mathbf{m}_h)^T(D\mathbf{m}_h)\,\mathrm{d}x = \sum_{i,j\in I}\int_\Omega \operatorname{trace}\nabla\varphi_i\boldsymbol{\alpha}_i^T\boldsymbol{\alpha}_j(\nabla\varphi_j)^T\,\mathrm{d}x \\ &= \sum_{i,j\in I}\boldsymbol{\alpha}_i\cdot\boldsymbol{\alpha}_j\int_\Omega \operatorname{trace}\nabla\varphi_i(\nabla\varphi_j)^T\,\mathrm{d}x = \sum_{i,j\in I}\boldsymbol{\alpha}_i\cdot\boldsymbol{\alpha}_j\int_\Omega \nabla\varphi_i\cdot\nabla\varphi_j\,\mathrm{d}x. \end{aligned}$$

The second and the third term in (1) have the values

$$\begin{aligned} &\int_\Omega \varphi(\mathbf{m}_h)\,\mathrm{d}x = \\ &\qquad \int_\Omega 1-(\mathbf{e}\cdot\mathbf{m}_h)^2\,\mathrm{d}x = \operatorname{vol}(\Omega) - \sum_{i,j\in I}(\mathbf{e}\cdot\boldsymbol{\alpha}_i)(\mathbf{e}\cdot\boldsymbol{\alpha}_j)\int_\Omega \varphi_i\varphi_j\,\mathrm{d}x \end{aligned}$$

and

$$\int_\Omega \mathbf{f}\cdot\mathbf{m}_h\,\mathrm{d}x = \sum_{i\in I}\mathbf{f}\cdot\boldsymbol{\alpha}_i\int_\Omega \varphi_i\,\mathrm{d}x.$$

Since u_1 vanishes on $\partial\Omega$, for the discretization of u_1 only the inner degrees of freedom are used, i.e.,

$$u_1^h = \sum_{j \in I_{in}} \beta_j \varphi_j$$

with $I_{in} := I \backslash I_{bd}$, where I_{bd} are the boundary indices. Let $I_{lay} \subset I_{in}$ be the vertices which have a neighbor in the set of boundary indices I_{bd}. Then the restriction of u_1^h to the boundary $\partial\Omega$ reads $(\nabla u_1^h)|_{\partial\Omega} = \sum_{i \in I_{lay}} \beta_i \nabla\varphi_i$ and

$$g_h = \sum_{j \in I_{lay}} \beta_j \nu \cdot \nabla\varphi_j - \sum_{j \in I_{bd}} \nu \cdot \boldsymbol{\alpha}_j \varphi_j .$$

Hence,

$$\begin{aligned} E_s(\mathbf{m}_h) &= \frac{\mu_0}{2} \sum_{i \in I} \left(\sum_{j \in I_{in}} \beta_j \boldsymbol{\alpha}_i \cdot \int_\Omega \varphi_i \nabla\varphi_j \,\mathrm{d}x - \boldsymbol{\alpha}_i \cdot \int_\Omega \int_{\partial\Omega} \nabla\varphi_i(x) S(x-y) g(y) \,\mathrm{d}s_y \,\mathrm{d}x \right) \\ &\quad + \frac{\mu_0}{2} \sum_{i \in I_{bd}} \boldsymbol{\alpha}_i \cdot \int_{\partial\Omega} \int_{\partial\Omega} \nu_x \varphi_i(x) S(x-y) g(y) \,\mathrm{d}s_y \,\mathrm{d}x \\ &= \frac{\mu_0}{2} \sum_{i \in I} \left(\sum_{j \in I_{in}} \beta_j \boldsymbol{\alpha}_i \cdot \int_\Omega \varphi_i \nabla\varphi_j \,\mathrm{d}x - \sum_{j \in I_{lay}} \beta_j \boldsymbol{\alpha}_i \cdot \mathbf{a}_{ij} + \sum_{j \in I_{bd}} \boldsymbol{\alpha}_i \cdot (B_{ij} \boldsymbol{\alpha}_j) \right) \\ &\quad + \frac{\mu_0}{2} \sum_{i \in I_{bd}} \left(\sum_{j \in I_{lay}} \beta_j \boldsymbol{\alpha}_i \cdot \mathbf{c}_{ij} - \sum_{j \in I_{bd}} \boldsymbol{\alpha}_i \cdot (D_{ij} \boldsymbol{\alpha}_j) \right), \end{aligned}$$

where

$$\mathbf{a}_{ij} = \int_\Omega \int_{\partial\Omega} \nabla\varphi_i(x) S(x-y) \nu \cdot \nabla\varphi_j(y) \,\mathrm{d}s_y \,\mathrm{d}x, \tag{8}$$

$$B_{ij} = \int_\Omega \int_{\partial\Omega} \nabla\varphi_i(x) S(x-y) \varphi_j(y) \nu_y^T \,\mathrm{d}s_y \,\mathrm{d}x, \tag{9}$$

and

$$\mathbf{c}_{ij} = \int_{\partial\Omega} \int_{\partial\Omega} \nu_x \varphi_i(x) S(x-y) \nu_y \cdot \nabla\varphi_j(y) \,\mathrm{d}s_y \,\mathrm{d}s_x, \tag{10}$$

$$D_{ij} = \int_{\partial\Omega} \int_{\partial\Omega} \nu_x \varphi_i(x) S(x-y) \varphi_j(y) \nu_y^T \,\mathrm{d}s_y \,\mathrm{d}s_x. \tag{11}$$

The efficient numerical evaluation of the singular integrals (8)–(11) using the Duffy transformation is described in the appendix.

4 Hierarchical Matrices

For the computation of the different energies in (1) it is necessary to efficiently treat fully populated matrices arising from the discretization of non-local operators, e.g. finite or boundary element discretizations of integral operators and inverses or the factors of the LU decomposition of FE discretizations of elliptic partial differential operators. For this purpose, Tyrtyshnikov [27] and Hackbusch et al. [15–17] introduced the structure of *mosaic skeleton matrices* or *hierarchical ($\mathscr{H}$-) matrices*. A similar approach, which is designed towards only the fast multiplication of a matrix by a vector, are the earlier developed fast summation methods *tree code* [3], *fast multipole methods* [13, 14], and *panel clustering* [18].

Many existing fast methods are based on multi-level structures. In contrast to multigrid methods, the efficiency of $\mathscr{H}$-matrices is based on a suitable hierarchy of partitions of the matrix indices. Let $I = \{1, \dots, M\}$ and $J = \{1, \dots, N\}$ be sets of indices corresponding to the rows and columns of a matrix $A \in \mathbb{R}^{M \times N}$. The efficiency of hierarchical matrices is based on the low-rank representation of sub-blocks from an appropriate partition $\mathscr{P}$ of the set of matrix indices $I \times J$; see Fig. 1.

The construction of $\mathscr{P}$ is usually done in the following way. First, cluster trees T_I and T_J are constructed by recursive subdivision of I and J. Each subdivision step is done such that indices that in some sense are close to each other are grouped into two clusters. In a second step the block cluster tree $T_{I \times J}$ is built by recursive subdivision of $I \times J$. Each block $t \times s$ is subdivided into the sons $t' \times s'$, where t' and s' are taken from the list of sons of t and s in T_I and T_J, respectively. The recursion stops at blocks $t \times s$ which either are small enough or satisfy a so-called *admissibility condition*. This condition guarantees that the restriction A_{ts} of A to $t \times s$ can be approximated by a matrix of low rank. It usually takes into account the geometry that is associated with the rows t and the columns s. The partition $\mathscr{P}$ is found as the leaves of $T_{I \times J}$. The set of hierarchical matrices on the partition $\mathscr{P}$ and blockwise rank k is then defined as

$$\mathscr{H}(\mathscr{P}, k) = \{A \in \mathbb{R}^{I \times J} : \text{rank}\ A_b \leq k \text{ for all } b \in \mathscr{P}\}.$$

The elements of this set can be stored with logarithmic-linear complexity and hence provide data-sparse representations of fully populated matrices. Additionally, exploiting the hierarchical structure of the partition, an approximate algebra can be defined [12] which is based on divide-and-conquer versions of the usual arithmetic operations. At least for discretizations of elliptic operators, these approximate operations have logarithmic-linear complexity (see [7]) and can be used to define substitutes for higher level matrix operations such as inversion, LU factorization, and QR factorization.

Using hierarchical matrices, we are able to efficiently compute the local and non-local parts of the energy functional (1) with almost optimal complexity. Especially approximations to the fully populated matrices (8)–(11) can be constructed with logarithmic-linear complexity via adaptive cross approximation (ACA) [6].

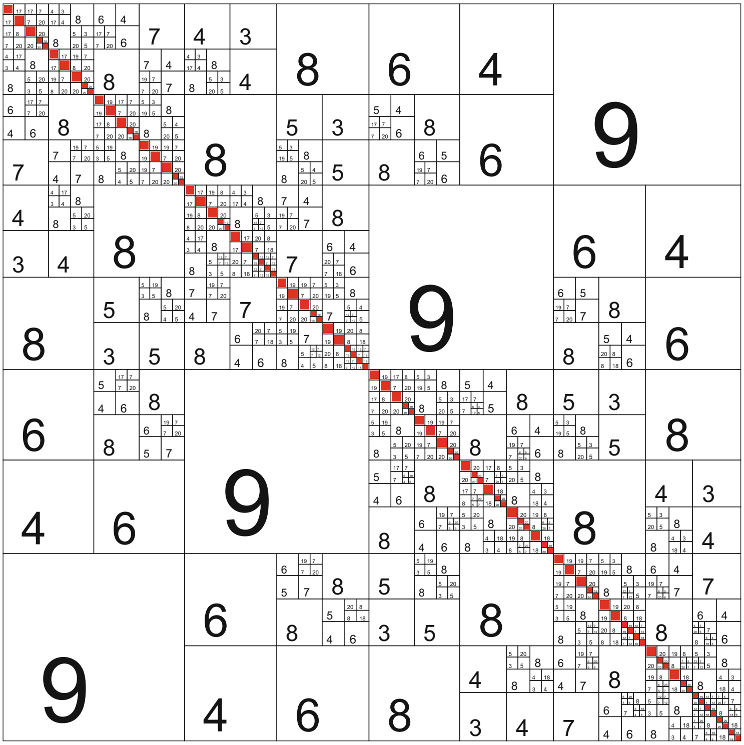


Fig. 1 A hierarchical matrix with its rank distribution

5 Iterative Minimization of the Energy

The numerical computation of local energy minima of the functional (1) is a challenging task due to the pointwise restriction $\|\mathbf{m}(x)\|_2 = 1$ for almost every $x \in \Omega$. In this section we will see how this constraint can be treated efficiently.

5.1 *The Minimization Algorithm*

We describe and analyze in this section our method to iteratively minimize the energy functional (1). To simplify the presentation, we consider the Dirichlet energy and a lower order term, i.e., the model functional

$$E(\mathbf{m}) = \Theta(\mathbf{m}) + \frac{1}{2}\int_{\Omega} \|D\mathbf{m}\|_F^2 \,\mathrm{d}x \tag{12}$$

with a smooth functional $\Theta : H^1(\Omega; \mathbb{R}^3) \to \mathbb{R}$. The following algorithm realizes a discrete H^1 flow for E with time-step size $\alpha > 0$ and employs ideas from [1, 2, 4].

Algorithm (minimization algorithm). Given $\mathbf{m}_h^0$ such that $\|\mathbf{m}_h^0(z)\|_2 = 1$ for all $z \in \mathscr{N}_h$ iterate for $n = 0, 1, 2, \ldots$ the following steps:

(1) Compute $\mathbf{w}_h^n \in \mathscr{S}^1(\mathscr{T}_h)^3$ with $\int_\Omega \mathbf{w}_h^n \,\mathrm{d}x = 0$ such that $\mathbf{w}_h^n(z) \cdot \mathbf{m}_h^n(z) = 0$ for all $z \in \mathscr{N}_h$ and

$$\int_\Omega \operatorname{trace}\,[(D\mathbf{w}_h^n)^T(D\mathbf{v}_h)]\,\mathrm{d}x = -\Theta'(\mathbf{m}_h^n)[\mathbf{v}_h] - \int_\Omega \operatorname{trace}\,[(D\mathbf{m}_h^n)^T(D\mathbf{v}_h)]\,\mathrm{d}x$$

for all $\mathbf{v}_h \in \mathscr{S}^1(\mathscr{T}_h)^3$ with $\int_\Omega \mathbf{v}_h \,\mathrm{d}x = 0$.

(2) Define $\mathbf{m}_h \in \mathscr{S}^1(\mathscr{T}_h)^3$ through

$$\mathbf{m}_h^{n+1}(z) = \frac{\mathbf{m}_h^n(z) + \alpha \mathbf{w}_h^n(z)}{\|\mathbf{m}_h^n(z) + \alpha \mathbf{w}_h^n(z)\|_2}$$

for all $z \in \mathscr{N}_h$ with a suitable $\alpha > 0$.

To ensure that the iteration is energy decreasing, we assume that the underlying triangulation $\mathscr{T}_h$ is weakly acute, i.e., that the off-diagonal entries of the corresponding $P1$ stiffness matrix are non-positive, cf. [4] for details.

Lemma 1 ([4]). *Assume that $\mathscr{T}_h$ is weakly acute and suppose that $\mathbf{m}_h, \mathbf{w}_h \in \mathscr{S}^1(\mathscr{T}_h)^3$ are such that $\|\mathbf{m}_h(z)\|_2 = 1$ and $\mathbf{m}_h(z) \cdot \mathbf{w}_h(z) = 0$ for all $z \in \mathscr{N}_h$. Then*

$$\left\| D\left(\frac{\mathbf{m}_h + \alpha \mathbf{w}_h}{\|\mathbf{m}_h + \alpha \mathbf{w}_h\|_2}\right)\right\|_{L^2(\Omega)} \le \|D(\mathbf{m}_h + \alpha \mathbf{w}_h)\|_{L^2(\Omega)}$$

for every $\alpha \in \mathbb{R}$.

To show convergence of the iterative algorithm, we argue as in [1, 2, 4]. For $\mathbf{m}_h^n$ and $\mathbf{w}_h^n$ as in the algorithm we have, upon choosing $\mathbf{v}_h = \mathbf{w}_h^n$ that

$$\int_\Omega \operatorname{trace}[(D\mathbf{m}_h^n)^T(D\mathbf{w}_h^n)]\,\mathrm{d}x = -\Theta'(\mathbf{m}_h^n)[\mathbf{w}_h^n] - \int_\Omega \|D\mathbf{w}_h^n\|_F^2 \,\mathrm{d}x.$$

Lemma 1 implies that

$$\begin{aligned}
\frac{1}{2}\int_\Omega \|D\mathbf{m}_h^{n+1}\|_F^2 - \|D\mathbf{m}_h^n\|_F^2 \,\mathrm{d}x &\le \frac{1}{2}\int_\Omega \|D(\mathbf{m}_h^n + \alpha \mathbf{w}_h^n)\|_F^2 - \|D\mathbf{m}_h^n\|_F^2 \,\mathrm{d}x \\
&= \alpha \int_\Omega \operatorname{trace}[(D\mathbf{m}_h^n)^T(D\mathbf{w}_h^n)] + \frac{\alpha^2}{2}\|D\mathbf{w}_h^n\|_F^2 \,\mathrm{d}x \\
&= -\alpha\Theta'(\mathbf{m}_h^n)[\mathbf{w}_h^n] + (\alpha^2/2 - \alpha)\int_\Omega \|D\mathbf{w}_h^n\|_F^2 \,\mathrm{d}x.
\end{aligned}$$

Hence, it follows that

$$\begin{aligned} E(\mathbf{m}_h^{n+1}) - E(\mathbf{m}_h^n) &= \Theta(\mathbf{m}_h^{n+1}) - \Theta(\mathbf{m}_h^n) + \frac{1}{2}\int_\Omega \|D\mathbf{m}_h^{n+1}\|_F^2 - \|D\mathbf{m}_h^n\|_F^2 \,\mathrm{d}x \\ &= \Theta(\mathbf{m}_h^{n+1}) - \Theta(\mathbf{m}_h^n) - \alpha\Theta'(\mathbf{m}_h^n)[\mathbf{w}_h^n] + \\ &\quad + (\alpha^2/2 - \alpha)\int_\Omega \|D\mathbf{w}_h^n\|_F^2 \,\mathrm{d}x. \end{aligned}$$

We assume here and will show below for the specification of Θ that corresponds to the model problem that we have

$$\left|\Theta(\mathbf{m}_h^{n+1}) - \Theta(\mathbf{m}_h^n) - \alpha\Theta'(\mathbf{m}_h^n)[\mathbf{w}_h^n]\right| \le C_\Theta \alpha^2 \|\mathbf{w}_h^n\|_{H^1(\Omega)}^2. \tag{13}$$

This estimate may be regarded as a Taylor expansion of Θ but its proof also requires to bound the difference between $\mathbf{m}_h^{n+1}$ and $\mathbf{m}_h^n + \alpha\mathbf{w}_h^n$. With a Poincaré inequality in (13) we thus have

$$\begin{aligned} E(\mathbf{m}_h^{n+1}) - E(\mathbf{m}_h^n) &\le (-\alpha + \alpha^2/2 + C_P C_\Theta \alpha^2)\int_\Omega \|D\mathbf{w}_h^n\|_F^2 \,\mathrm{d}x \\ &\le -\alpha(1 - \alpha/2 - C_P C_\Theta \alpha)\int_\Omega \|D\mathbf{w}_h^n\|_F^2 \,\mathrm{d}x. \end{aligned}$$

If α is sufficiently small so that $(1 - \alpha/2 - C_P C_\Theta \alpha) \ge 1/2$ then it follows that

$$E(\mathbf{m}_h^{n+1}) - E(\mathbf{m}_h^n) \le -\frac{\alpha}{2}\int_\Omega \|D\mathbf{w}_h^n\|_F^2 \,\mathrm{d}x$$

and after summation over $n = 0, 1, \ldots, N$

$$E(\mathbf{m}_h^{N+1}) + \frac{\alpha}{2}\sum_{n=0}^{N}\int_\Omega \|D\mathbf{w}_h^n\|_F^2 \,\mathrm{d}x \le E(\mathbf{m}_h^0).$$

This proves the stability and convergence of our numerical method.

5.2 Application to the Model Problem

In our model (12), the functional $\Theta(\mathbf{m})$ is given by

$$\Theta(\mathbf{m}) = \int_\Omega 1 - (\mathbf{e}\cdot\mathbf{m})^2 - \mathbf{f}\cdot\mathbf{m} + \frac{\mu_0}{2}\mathbf{m}\cdot\nabla u_{\mathbf{m}} \,\mathrm{d}x.$$

We claim that for this functional the estimate (13) is satisfied. For this we first show that

$$|\Theta(\mathbf{m}) - \Theta(\tilde{\mathbf{m}})| \le C_1 \|\mathbf{m} - \tilde{\mathbf{m}}\|_{L^2(\Omega)} \tag{14}$$

provided that $\|\mathbf{m}\|_{L^2(\Omega)}, \|\tilde{\mathbf{m}}\|_{L^2(\Omega)} \le C$. With Hölder's inequality we verify that

$$\begin{aligned}\Theta(\mathbf{m}) - \Theta(\tilde{\mathbf{m}}) \le \int_\Omega (\mathbf{e}\cdot\mathbf{m})^2 - (\mathbf{e}\cdot\tilde{\mathbf{m}})^2 \,\mathrm{d}x + \|\mathbf{f}\|_{L^2(\Omega)} \|\mathbf{m} - \tilde{\mathbf{m}}\|_{L^2(\Omega)} \\ + \frac{\mu_0}{2} \int_\Omega \nabla u_{\mathbf{m}} \cdot \mathbf{m} - \nabla u_{\tilde{\mathbf{m}}} \cdot \tilde{\mathbf{m}} \,\mathrm{d}x.\end{aligned}$$

Using Cauchy-Schwarz inequality and $\|\mathbf{e}\|_2 = 1$, the first term on the right-hand side can be bounded by

$$\int_\Omega (\mathbf{e}\cdot\mathbf{m})^2 - (\mathbf{e}\cdot\tilde{\mathbf{m}})^2 \,\mathrm{d}x \le \|\mathbf{m} - \tilde{\mathbf{m}}\|_{L^2(\Omega)} \|\mathbf{m} + \tilde{\mathbf{m}}\|_{L^2(\Omega)}.$$

Similarly, we have with (4) that

$$\begin{aligned}\int_\Omega \nabla u_{\mathbf{m}} \cdot \mathbf{m} - \nabla u_{\tilde{\mathbf{m}}} \cdot \tilde{\mathbf{m}} \,\mathrm{d}x &= \int_{\mathbb{R}^3} \|\nabla u_{\mathbf{m}}\|_2^2 - \|\nabla u_{\tilde{\mathbf{m}}}\|_2^2 \,\mathrm{d}x \\ &\le \|\nabla(u_{\mathbf{m}} - u_{\tilde{\mathbf{m}}})\|_{L^2(\mathbb{R}^3)} \|\nabla(u_{\mathbf{m}} + u_{\tilde{\mathbf{m}}})\|_{L^2(\mathbb{R}^3)} \\ &\le \|\mathbf{m} - \tilde{\mathbf{m}}\|_{L^2(\Omega)} (\|\mathbf{m}\|_{L^2(\Omega)} + \|\tilde{\mathbf{m}}\|_{L^2(\Omega)}).\end{aligned}$$

Hence, the property (14) follows from the assumed bounds on $\mathbf{m}$ and $\tilde{\mathbf{m}}$. We next show that

$$\left|\Theta(\mathbf{m} + \alpha\mathbf{w}) - \Theta(\mathbf{m}) - \alpha\Theta'(\mathbf{m})[\mathbf{w}]\right| \le C_2 \alpha^2 \|\mathbf{w}\|_{L^2(\Omega)}^2. \tag{15}$$

Straightforward calculations lead to

$$\begin{aligned}&\Theta(\mathbf{m} + \alpha\mathbf{w}) - \Theta(\mathbf{m}) - \alpha\Theta'(\mathbf{m})[\mathbf{w}] \\ &= \int_\Omega (\mathbf{e}\cdot(\mathbf{m} + \alpha\mathbf{w}))^2 - (\mathbf{e}\cdot\mathbf{m})^2 - 2\alpha(\mathbf{e}\cdot\mathbf{m})(\mathbf{e}\cdot\mathbf{w}) \,\mathrm{d}x \\ &\quad + \frac{\mu_0}{2} \int_\Omega \nabla u_{\mathbf{m}+\alpha\mathbf{w}} \cdot (\mathbf{m} + \alpha\mathbf{w}) - \nabla u_{\mathbf{m}} \cdot \mathbf{m} - 2\alpha \nabla u_{\mathbf{m}} \cdot \mathbf{w} \,\mathrm{d}x \\ &= \alpha^2 \int_\Omega (\mathbf{e}\cdot\mathbf{w})^2 \,\mathrm{d}x + \frac{\mu_0}{2} \int_\Omega \nabla u_{\mathbf{m}+\alpha\mathbf{w}} \cdot (\mathbf{m} + \alpha\mathbf{w}) - \nabla u_{\mathbf{m}} \cdot \mathbf{m} - 2\alpha \nabla u_{\mathbf{m}} \cdot \mathbf{w} \,\mathrm{d}x \\ &\le \alpha^2 \|\mathbf{w}\|_{L^2(\Omega)}^2 + \frac{\mu_0}{2} \int_\Omega \nabla u_{\mathbf{m}+\alpha\mathbf{w}} \cdot (\mathbf{m} + \alpha\mathbf{w}) - \nabla u_{\mathbf{m}} \cdot \mathbf{m} - 2\alpha \nabla u_{\mathbf{m}} \cdot \mathbf{w} \,\mathrm{d}x.\end{aligned}$$

Using (3) and $\nabla u_{\mathbf{m}+\alpha\mathbf{w}} = \nabla u_{\mathbf{m}} + \alpha\nabla u_{\mathbf{w}}$, we find that the second integral on the right-hand side of the previous estimate satisfies

$$\begin{aligned}
&\int_{\mathbb{R}^3} \|\nabla u_{\mathbf{m}+\alpha\mathbf{w}}\|_2^2 - \|\nabla u_{\mathbf{m}}\|_2^2 \,\mathrm{d}x - 2\alpha \int_{\Omega} \nabla u_{\mathbf{m}} \cdot \mathbf{w} \,\mathrm{d}x \\
&= \int_{\mathbb{R}^3} \|\nabla u_{\mathbf{m}}\|_2^2 + 2\alpha\nabla u_{\mathbf{m}} \cdot \nabla u_{\mathbf{w}} + \|\nabla u_{\alpha\mathbf{w}}\|_2^2 - \|\nabla u_{\mathbf{m}}\|_2^2 \,\mathrm{d}x - 2\alpha \int_{\Omega} \nabla u_{\mathbf{m}} \cdot \mathbf{w} \,\mathrm{d}x \\
&= \int_{\mathbb{R}^3} \|\nabla u_{\alpha\mathbf{w}}\|_2^2 \,\mathrm{d}x \le \alpha^2 \int_{\Omega} \|\mathbf{w}\|_2^2 \,\mathrm{d}x.
\end{aligned}$$

This leads to the estimate (15). The combination of (14) and (15) implies that for iterates $\mathbf{m}_h^n$, $\mathbf{w}_h^n$, and $\mathbf{m}_h^{n+1}$ we have, noting that $\|\mathbf{m}_h^n\|_{L^2(\Omega)}, \|\mathbf{w}_h^n\|_{L^2(\Omega)} \le C$,

$$\begin{aligned}
&\big|\Theta(\mathbf{m}_h^{n+1}) - \Theta(\mathbf{m}_h^n) - \alpha\Theta'(\mathbf{m}_h^n)[\mathbf{w}_h]\big| \\
&\qquad \le \big|\Theta(\mathbf{m}_h^{n+1}) - \Theta(\mathbf{m}_h^n + \alpha\mathbf{w}_h^n)\big| + C_2\alpha^2\|\mathbf{w}_h^n\|_{L^2(\Omega)}^2 \\
&\qquad \le C_1\|\mathbf{m}_h^{n+1} - (\mathbf{m}_h^n + \alpha\mathbf{w}_h^n)\|_{L^2(\Omega)} + C_2\alpha^2\|\mathbf{w}_h^n\|_{L^2(\Omega)}^2.
\end{aligned}$$

For every node $z \in \mathcal{N}_h$ we have

$$\begin{aligned}
\|\mathbf{m}_h^{n+1}(z) - \mathbf{m}_h^n(z) - \alpha\mathbf{w}_h^n(z)\|_2 &= \left\| \frac{\mathbf{m}_h^n(z) + \alpha\mathbf{w}_h^n(z)}{\|\mathbf{m}_h^n(z) + \alpha\mathbf{w}_h^n(z)\|_2} - \mathbf{m}_h^n(z) - \alpha\mathbf{w}_h^n(z) \right\|_2 \\
&= \|\mathbf{m}_h^n(z) + \alpha\mathbf{w}_h^n(z)\|_2 - 1 \\
&= \big(1 + \alpha^2\|\mathbf{w}_h^n(z)\|_2^2\big)^{1/2} - 1 \le \frac{\alpha^2}{2}\|\mathbf{w}_h^n(z)\|_2^2
\end{aligned}$$

and the continuous embedding $H^1(\Omega) \hookrightarrow L^4(\Omega)$ yields

$$\|\mathbf{m}_h^{n+1} - (\mathbf{m}_h^n + \alpha\mathbf{w}_h^n)\|_{L^2(\Omega)} \le C\alpha^2\|\mathbf{w}_h^n\|_{L^4(\Omega)}^2 \le C'\alpha^2\|\mathbf{w}_h^n\|_{H^1(\Omega)}^2.$$

Hence, the model problem fulfills the desired property (13).

6 The String Method

We aim at computing an unstable critical configuration whose energy is minimal among all maxima of curves connecting two given states, i.e., we compute a saddle point. For this, we adopt a method proposed in [30] that does not require an energy that is defined on a linear space as it is needed for classical mountain pass algorithms. The difficulty for the energy functional under consideration is the appropriate incorporation of the pointwise unit length constraint.

To describe the problem in a continuous setting, we assume that we are given two local minima $\mathbf{m}_0 \in \mathscr{A}$ and $\mathbf{m}_1 \in \mathscr{A}$ of the energy functional $E : \mathscr{A} \to \mathbb{R}$, where the space of admissible magnetic configurations $\mathscr{A}$ is defined by

$$\mathscr{A} = \{\mathbf{m} \in H^1(\Omega; \mathbb{R}^3) : \|\mathbf{m}(x)\|_2 = 1 \text{ for almost every } x \in \Omega\}.$$

We then consider a family of curves connecting $\mathbf{m}_0$ and $\mathbf{m}_1$ parametrized by $t \geq 0$, i.e., a continuous mapping

$$\varphi(t, \cdot) : [0, 1] \to \mathscr{A}$$

such that $\varphi(t, 0) = \mathbf{m}_0$ and $\varphi(t, 1) = \mathbf{m}_1$.

Our aim is it to compute a curve connecting $\mathbf{m}_0, \mathbf{m}_1 \in \mathscr{A}$ such that the component of ∇E normal to φ vanishes, i.e.

$$(\nabla E)^{\perp}(\varphi) = 0, \tag{16}$$

where

$$(\nabla E)^{\perp}(\varphi) = \nabla E(\varphi) - (\nabla E(\varphi) \cdot \tau)\tau$$

and τ denotes the unit tangent vector of φ. A path φ satisfying (16) is called a minimum energy path (MEP).

To numerically evaluate such a path, we deploy the string method which was proposed in [28, 29] and modified in [30]. The modified string method stands out due to its simplicity. Another known method to compute MEPs is the nudged elastic band (NEB) method; cf. [21].

Algorithm (modified string method). Let two local minima $\mathbf{m}_0, \mathbf{m}_1 \in \mathscr{A}$ of the energy functional E be given. Define a path $\varphi^0 : \{0, \ldots, N\} \to \mathscr{A}$ as a collection of $N + 1$ points with $\varphi^0(0) = \mathbf{m}_0$ and $\varphi^0(N) = \mathbf{m}_1$. The points $\varphi^0(i)$, $i = 1, \ldots, N - 1$, are computed via interpolation. Iterate for $n = 0, 1, 2, \ldots$ the following steps:

(1) Let $\varphi^n_*(0) = \varphi^n(0)$ and $\varphi^n_*(N) = \varphi^n(N)$. Compute for each configuration $\varphi^n(i)$ with $i = 1, \ldots, N - 1$ a single iteration step of the minimization algorithm as proposed in Sect. 5.1 and assign the result to $\varphi^n_*(i)$.

(2) Compute via interpolation the new path $\big(\varphi^{n+1}(i)\big)_{i=0,\ldots,N}$ as a reparametrization of $\big(\varphi^n_*(i)\big)_{i=0,\ldots,N}$.

The interpolation used in the modified string method can be done in arbitrary ways. We choose to interpolate geodesically on the sphere to obtain a reparametrization of the string.

The advantage of the modified string method is that the point-wise constraint $\|\mathbf{m}(x)\|_2 = 1$, $x \in \Omega$, is inherited from the above minimization algorithm. In Fig. 2, an initial path and an MEP as the result of the modified string method are shown.

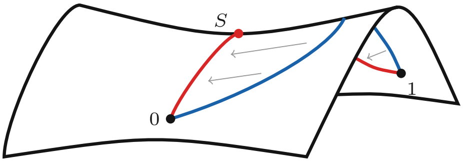


Fig. 2 A scheme for the modified string method showing the initial path (*blue line*) and the MEP (*red line*) connecting the two energy minima $\mathbf{m}_0$ and $\mathbf{m}_1$. The configuration $\mathbf{m}_S$ is the saddle point on the MEP

7 Numerical Examples

In this section, we first verify our implementation using a standard problem, afterwards we investigate the overall complexity of the presented numerical method and finally we compute an MEP for a cubic magnetic particle.

All tests were performed on a single core of an Intel Xeon X5482 processor with 3.2 GHz and 64 GB of RAM. The programming was done in `C++` and is based on the hierarchical matrix library A$\mathscr{H}$MED.[1] Furthermore, we set the minimal block cluster size of the created $\mathscr{H}$-matrices to 50 and the relative blockwise approximation accuracy to $1\mathrm{e}-3$.

7.1 Validation of Implementation

A problem proposed by A. Hubert, University of Erlangen-Nuremberg, to check the computation of the different energies in (1) is to calculate the single domain limit of a cubic magnetic particle.

This is the length of the cube for which the so-called flower and vortex state have equal energies. The test is also known as the μ-mag standard problem #3; see [23].

In our tests, the cube was discretized into 24,576 tetrahedra and 3,072 triangles. Figure 3 shows the reduced energy, as proposed in [23], relative to the size of the cube. Our numerical tests show a single domain limit of 8.23, which represents the theoretically expected value of approximately 8.

7.2 Time and Memory Consumption

To demonstrate the almost linear behavior of the presented numerical algorithms in terms of memory and time consumption, we chose different discretizations of the unit cube which were created using netgen.[2]

[1] http://bebendorf.ins.uni-bonn.de/AHMED.html

[2] http://www.hpfem.jku.at/netgen/

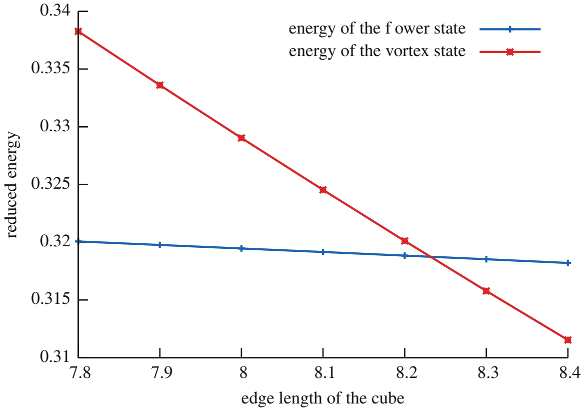


Fig. 3 The resulting energy minima of the μ-mag standard problem #3

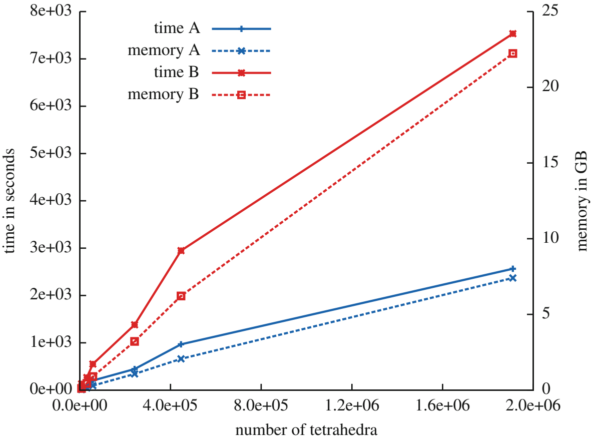


Fig. 4 The time and memory consumption of the matrices A and B; cf. (8) and (9)

As a first test, we look at the construction of the matrices (8)–(11). These matrices have to be set up only once for a certain geometry and approximation accuracy in an initialization step. Figure 4 shows that the time and memory needed to construct the matrices (8) and (9) is linear up to logarithmic terms with respect to the number of tetrahedra. Furthermore, from Fig. 5 it can be seen that the construction of (10) and (11) is quasi-optimal with respect to the number of triangles. For the string method, the minimization algorithm from Sect. 5.1 is the dominant part of the computational time. As can be seen from Fig. 6, these minimization steps of the algorithm presented in Sect. 5.1 are almost linear in terms of the number of tetrahedra. Hence, using $\mathscr{H}$-matrices to compute the different energies results in a numerical scheme which has logarithmic-linear complexity.

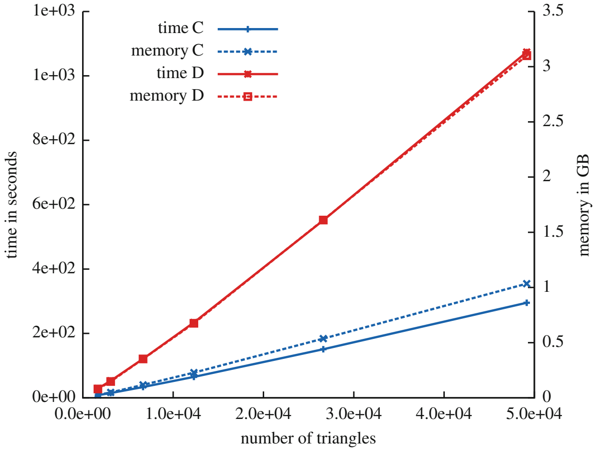


Fig. 5 The time and memory consumption of the matrices C and D; cf. (10) and (11)

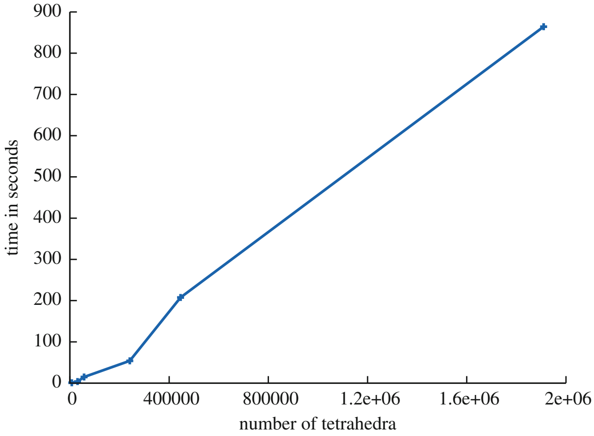


Fig. 6 The time needed for a single minimization step

7.3 Minimum Energy Paths

In the following test, we use the simplified string method from Sect. 6 to compute a MEP. The geometry and the parameter configuration are chosen as in the μ-mag standard problem #3 with a cube edge length of 8.2. Hence, the two energy minima are the flower and the vortex state. For the tests, the cube was discretized into 24,576 tetrahedra and 3,072 triangles. For each iteration step the discretized path consists of 41 single magnetization points.

In Fig. 7, the two energy minima (flower and vortex state) which are used to define our initial path are depicted. Different colors of the arrows representing the

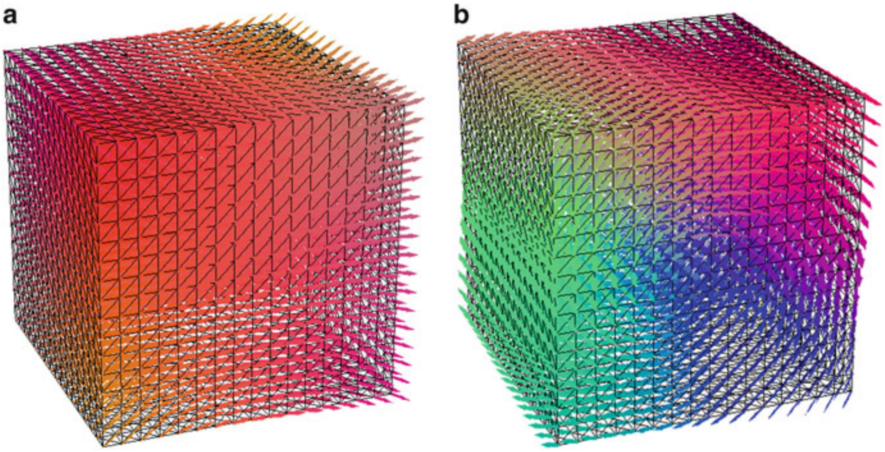


Fig. 7 A model of the two energy minima used to compute the MEP. (**a**) Flower state. (**b**) Vortex state

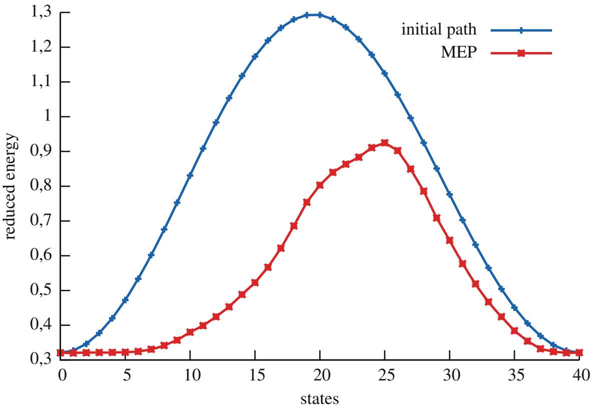


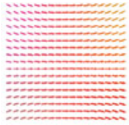
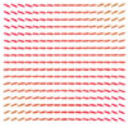
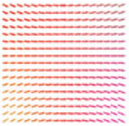
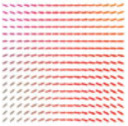
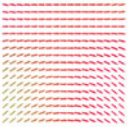
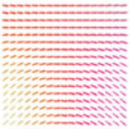
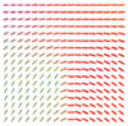
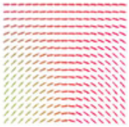
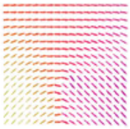
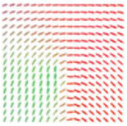
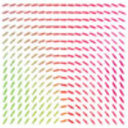
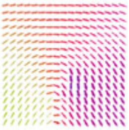
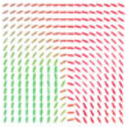
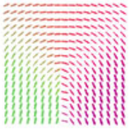
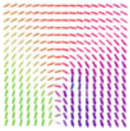
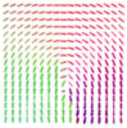
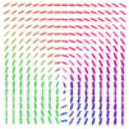
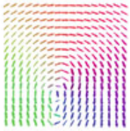
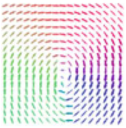
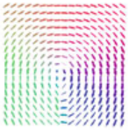
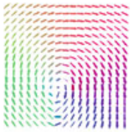
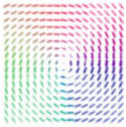
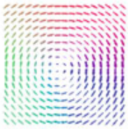
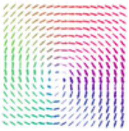
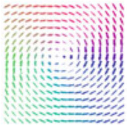
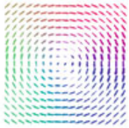
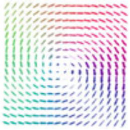
Fig. 8 The reduced energy of the magnetization points of the initial path and the MEP connecting the flower (left) and the vortex state (right)

magnetization indicate different directions in space. The initial magnetization points in between the energy minima were computed using geodesical interpolation.

A comparison of the reduced energies of the initialized path and the MEP is shown in Fig. 8.

Here, the magnetization point 25 maximizes the reduced energy of the MEP and is about 28.5 % lower than the energy barrier of the initial path. Furthermore, in Table 1 slices of selected magnetization states of the MEP are shown. It is remarkable that in the magnetization point 40 the vortex is opening on the front side and closing on the back.

Table 1 The different magnetization states of the MEP connecting the vortex and the flower state

point #	magnetization of the cube		
	front	mid	back
0			
5			
10			
15			
20			
25			
30			
35			
40			

The limiting factor of the numerical experiments is the computational time since all the single magnetization points along the discretized path need to be minimized consecutively. Even with the employed fast methods approximately 2 days were needed to perform the computations.

Appendix

The efficient numerical evaluation of the singular integrals from Sect. 3, see (8)–(11), is a challenging task. One way to overcome difficulties is to use spherical coordinates, which has the disadvantage of awkward integration bounds.

Another way is to use the Duffy transformation as proposed in [10]. The general principle is to transform a triangle onto a square to get rid of the singularity. The following example demonstrates this approach. Using the transformation $x = \xi$ and $y = \xi\eta$, one obtains that

$$\int_0^1 \int_0^x \frac{1}{x+y}\,\mathrm{d}y\,\mathrm{d}x = \int_0^1 \int_0^1 \frac{1}{1+\eta}\,\mathrm{d}\eta\,\mathrm{d}\xi$$

and integration can be done by using standard methods.

This principle has already been applied to the integration on pairs of triangles in [25] and can be used to evaluate the integrals in (10) and (11). Similar ideas can be applied to the combination of a triangle and a tetrahedron. For the kernel function $\kappa : \mathbb{R}^3 \times \mathbb{R}^2 \to \mathbb{R}$ we need to evaluate the following integration on the reference triangle and tetrahedron

$$I := \int\limits_0^1 \int\limits_0^{1-x_1} \int\limits_0^{1-x_1-x_2} \int\limits_0^1 \int\limits_0^{1-y_1} \kappa(\mathbf{x}, \mathbf{y})\,\mathrm{d}\mathbf{y}\,\mathrm{d}\mathbf{x}$$

with $\mathbf{x} = (x_1, x_2, x_3)$ and $\mathbf{y} = (y_1, y_2)$. By introducing relative coordinates, the integral I can be transformed to

$$I = \int\limits_0^1 \int\limits_0^{\tilde{x}_1} \int\limits_0^{\tilde{x}_1-\tilde{x}_2} \int\limits_{-\tilde{x}_1}^{1-\tilde{x}_1} \int\limits_{-\tilde{x}_2}^{\tilde{y}_1+\tilde{x}_1-\tilde{x}_2} \kappa\left(\begin{pmatrix} 1-\tilde{x}_1 \\ \tilde{x}_2 \\ \tilde{x}_3 \end{pmatrix}, \begin{pmatrix} 1-\tilde{y}_1-\tilde{x}_1 \\ \tilde{y}_2+\tilde{x}_2 \\ 0 \end{pmatrix}\right) \mathrm{d}\tilde{\mathbf{y}}\,\mathrm{d}\tilde{\mathbf{x}}.$$

The kernel is singular only for $\tilde{\mathbf{y}} = 0$, $\tilde{x}_3 = 0$, which we can be eliminated using a Duffy transformation. Similar to the approach in [25], we split the integral into six different domains.

$$\left\{\begin{matrix} -1 \le \tilde{y}_1 \le 0 \\ -1 \le \tilde{y}_2 \le \tilde{y}_1 \\ -\tilde{y}_2 \le \tilde{x}_1 \le 1 \\ -\tilde{y}_2 \le \tilde{x}_2 \le \tilde{x}_1 \\ 0 \le \tilde{x}_3 \le \tilde{x}_1 - \tilde{x}_2 \end{matrix}\right\} \cup \left\{\begin{matrix} -1 \le \tilde{y}_1 \le 0 \\ \tilde{y}_1 \le \tilde{y}_2 \le 0 \\ -\tilde{y}_1 \le \tilde{x}_1 \le 1 \\ -\tilde{y}_2 \le \tilde{x}_2 \le \tilde{x}_1 + \tilde{y}_1 - \tilde{y}_2 \\ 0 \le \tilde{x}_3 \le \tilde{x}_1 - \tilde{x}_2 \end{matrix}\right\} \cup \left\{\begin{matrix} -1 \le z_1 \le 0 \\ 0 \le \tilde{y}_2 \le 1 + \tilde{y}_1 \\ \tilde{y}_2 - \tilde{y}_1 \le \tilde{x}_1 \le 1 \\ 0 \le \tilde{x}_2 \le \tilde{x}_1 + \tilde{y}_1 - \tilde{y}_2 \\ 0 \le \tilde{x}_3 \le \tilde{x}_1 - \tilde{x}_2 \end{matrix}\right\}$$

$$\cup \left\{\begin{matrix} 0 \le \tilde{y}_1 \le 1 \\ -1 + \tilde{y}_1 \le \tilde{y}_2 \le 0 \\ -\tilde{y}_2 \le \tilde{x}_1 \le 1 - \tilde{y}_1 \\ -\tilde{y}_2 \le \tilde{x}_2 \le \tilde{x}_1 \\ 0 \le \tilde{x}_3 \le \tilde{x}_1 - \tilde{x}_2 \end{matrix}\right\} \cup \left\{\begin{matrix} 0 \le \tilde{y}_1 \le 1 \\ 0 \le \tilde{y}_2 \le \tilde{y}_1 \\ 0 \le \tilde{x}_1 \le 1 - \tilde{y}_1 \\ 0 \le \tilde{x}_2 \le \tilde{x}_1 \\ 0 \le \tilde{x}_3 \le \tilde{x}_1 - \tilde{x}_2 \end{matrix}\right\} \cup \left\{\begin{matrix} 0 \le \tilde{y}_1 \le 1 \\ \tilde{y}_1 \le \tilde{y}_2 \le 1 \\ \tilde{y}_2 - \tilde{y}_1 \le \tilde{x}_1 \le 1 - \tilde{y}_1 \\ 0 \le \tilde{x}_2 \le \tilde{y}_1 - \tilde{y}_2 + \tilde{x}_1 \\ 0 \le \tilde{x}_3 \le \tilde{x}_1 - \tilde{x}_2 \end{matrix}\right\}.$$

The resulting integrals $I := I_1 + \ldots + I_6$, can be transformed onto a five-dimensional unit cube. The single terms are given in the following way.

- I_1:

$$I_1 = \int\limits_{(0,1)^5} p_1 \kappa \left(\begin{pmatrix} 1 - \eta_5 \\ \eta_5(1 - \eta_1 + \eta_1\eta_2) \\ \eta_1\eta_4\eta_5(1 - \eta_2) \end{pmatrix}, \begin{pmatrix} 1 - \eta_5 + \eta_1\eta_2\eta_3\eta_5 \\ \eta_5(1 - \eta_1) \end{pmatrix} \right) \mathrm{d}\boldsymbol{\eta},$$

$$p_1 := \eta_1^3\eta_2\eta_5^4(1 - \eta_2)$$

- I_2:

$$I_2 = \int\limits_{(0,1)^5} p_2 \kappa \left(\begin{pmatrix} 1 - \eta_5 \\ \eta_1\eta_5(1 - \eta_2 + \eta_2\eta_3) \\ \eta_4\eta_5(1 - \eta_1 + \eta_1\eta_2 - \eta_1\eta_2\eta_3) \end{pmatrix}, \begin{pmatrix} 1 - \eta_5 + \eta_1\eta_2\eta_5 \\ \eta_1\eta_5(1 - \eta_2) \end{pmatrix} \right) \mathrm{d}\boldsymbol{\eta},$$

$$p_2 := \eta_1^2\eta_2\eta_5^4(1 - \eta_1 + \eta_1\eta_2 - \eta_1\eta_2\eta_3)$$

- I_3:

$$I_3 = \int\limits_{(0,1)^5} p_3 \kappa \left(\begin{pmatrix} 1 - \eta_5 \\ \eta_1\eta_5(1 - \eta_2) \\ \eta_4\eta_5(1 - \eta_1 + \eta_1\eta_2) \end{pmatrix}, \begin{pmatrix} 1 - \eta_5 + \eta_1\eta_2\eta_3\eta_5 \\ \eta_1\eta_5(1 - \eta_2\eta_3) \end{pmatrix} \right) \mathrm{d}\boldsymbol{\eta},$$

$$p_3 := \eta_1^2\eta_2\eta_5^4(1 - \eta_1 + \eta_1\eta_2),$$

- I_4:

$$I_4 = \int\limits_{(0,1)^5} p_4 \kappa \left(\begin{pmatrix} 1 - \eta_5 + \eta_1\eta_2\eta_3\eta_5 \\ \eta_1\eta_5(1 - \eta_2\eta_3) \\ \eta_4\eta_5(1 - \eta_1) \end{pmatrix}, \begin{pmatrix} 1 - \eta_5 \\ \eta_1\eta_5(1 - \eta_2) \end{pmatrix} \right) \mathrm{d}\boldsymbol{\eta},$$

$$p_4 := \eta_1^2\eta_2\eta_5^4(1 - \eta_1)$$

- I_5:

$$I_5 = \int\limits_{(0,1)^5} p_5 \kappa \left(\begin{pmatrix} 1-\eta_5+\eta_1\eta_2\eta_5 \\ \eta_1\eta_5(1-\eta_2) \\ \eta_4\eta_5(1-\eta_1) \end{pmatrix}, \begin{pmatrix} 1-\eta_5 \\ \eta_1\eta_5(1-\eta_2+\eta_2\eta_3) \end{pmatrix} \right) \mathrm{d}\boldsymbol{\eta},$$

$$p_5 := \eta_1^2\eta_2\eta_5^4(1-\eta_1)$$

- I_6:

$$I_6 = \int\limits_{(0,1)^5} p_6 \kappa \left(\begin{pmatrix} 1-\eta_5+\eta_1\eta_2\eta_3\eta_5 \\ \eta_5(1-\eta_1) \\ \eta_1\eta_4\eta_5(1-\eta_2\eta_3) \end{pmatrix}, \begin{pmatrix} 1-\eta_5 \\ \eta_5(1-\eta_1+\eta_1\eta_2) \end{pmatrix} \right) \mathrm{d}\boldsymbol{\eta},$$

$$p_6 := \eta_1^3\eta_2\eta_5^4(1-\eta_2\eta_3)$$

The integrals $I_1, \ldots, I_6$, can be evaluated efficiently using standard quadrature formulas, e.g., Gaussian quadrature.

Acknowledgements The author wishes to thank Felix Otto and Matthias Kurzke for stimulating discussion on the topic and Dirk Praetorius for pointing out the Ref. [11] on the employed splitting method for the computation of the stray field.

References

1. Alouges, F.: A new algorithm for computing liquid crystal stable configurations: the harmonic mapping case. SIAM J. Numer. Anal. **34**, 1708–1726 (1997)
2. Alouges, F., Conti, S., DeSimone, A., Pokern, Y.: Energetics and switching of quasi-uniform states in small ferromagnetic particles. Math. Model. Numer. Anal. **38**, 235–248 (2004)
3. Barnes, J., Hut, P.: A hierarchical $\mathcal{O}(N \log N)$ force calculation algorithm. Nature **324**, 446–449 (1986)
4. Bartels, S.: Stability and convergence of finite-element approximation schemes for harmonic maps. SIAM J. Numer. Anal. **43**, 220–238 (2005)
5. Bartels, S., Prohl, A.: Convergence of an implicit finite element method for the Landau-Lifshitz-Gilbert equation. SIAM J. Numer. Anal. **44**(4), 1405–1419 (2006)
6. Bebendorf, M.: Approximation of boundary element matrices. Numer. Math. **86**(4), 565–589 (2000)
7. Bebendorf, M.: Hierarchical Matrices: A Means to Efficiently Solve Elliptic Boundary Value Problems. Volume 63 of Lecture Notes in Computational Science and Engineering (LNCSE). Springer, Berlin (2008). ISBN:978-3-540-77146-3
8. Choi, Y.S., McKenna, P.J.: A mountain pass method for the numerical solution of semilinear elliptic problems. Nonlinear Anal. **20**(4), 417–437 (1993)
9. DeSimone, A., Kohn, R.V., Müller, S., Otto, F.: Magnetic microstructures – a paradigm of multiscale problems. In: ICIAM 99, Edinburgh, pp. 175–190. Oxford University Press, Oxford (2000)
10. Duffy, M.: Quadrature over a pyramid or cube of integrands with a singularity at a vertex. SIAM J. Numer. Anal. **19**, 1260–1262 (1982)

11. García-Cervera, J.M.: Numerical micromagnetics: a review. Bol. Soc. Esp. Matem. Apl. **39**, 103–135 (2007)
12. Grasedyck, L., Hackbusch, W.: Construction and arithmetics of $\mathscr{H}$-matrices. Computing **70**, 295–334 (2003)
13. Greengard, L.F., Rokhlin, V.: A fast algorithm for particle simulations. J. Comput. Phys. **73**(2), 325–348 (1987)
14. Greengard, L.F., Rokhlin, V.: A new version of the fast multipole method for the Laplace equation in three dimensions. In: Iserles, A. (ed.) Acta Numerica, 1997. Volume 6 of Acta Numerica, pp. 229–269. Cambridge University Press, Cambridge (1997)
15. Hackbusch, W.: A sparse matrix arithmetic based on $\mathscr{H}$-matrices. Part I: introduction to $\mathscr{H}$-matrices. Computing **62**(2), 89–108 (1999)
16. Hackbusch, W.: Hierarchische Matrizen. Springer, Berlin/Heidelberg (2009)
17. Hackbusch, W., Khoromskij, B.N.: A sparse $\mathscr{H}$-matrix arithmetic. Part II: application to multi-dimensional problems. Computing **64**(1), 21–47 (2000)
18. Hackbusch, W., Nowak, Z.P.: On the fast matrix multiplication in the boundary element method by panel clustering. Numer. Math. **54**(4), 463–491 (1989)
19. Hertel, R., Kronmüller, H.: Finite element calculations on the single-domain limit of a ferromagnetic cube – a solution to μmag standard problem no. 3 J. Magn. Magn. Mater. **238**(2–3), 185–199 (2002)
20. Hubert, A., Schäfer, R.: Magnetic Domains – The Analysis of Magnetic Microstructures. Springer, Berlin (2009)
21. Jónsson, H., Mills, G., Jacobsen, K.W.: Nudged elastic band method for finding minimum energy paths of transitions. In: Berne, B.J. et al. (eds.) Classical and Quantum Dynamics in Condensed Phase Simulations. World Scientific, Singapore (1998)
22. Kruzík, M., Prohl, A.: Recent developments in the modeling, analysis, and numerics of ferromagnetism. SIAM Rev. **48**(3), 439–483 (2006)
23. McMichael, R.D.: Standard problem number 3 – problem specification and reported solutions. http://www.ctcms.nist.gov/~rdm/mumag.html (2008)
24. Popović, N., Praetorius, D.: Applications of $\mathscr{H}$-matrix techniques in micromagnetics. Computing **74**(3), 177–204 (2005)
25. Sauter, S., Schwab, C.: Boundary Element Methods. Springer, Heidelberg/New York (2011)
26. Schabes, M.E., Bertram, H.N.: Magnetization processes in ferromagnetic cubes. J. Appl. Phys. **64**(3), 1347–1357 (1988)
27. Tyrtyshnikov, E.E.: Mosaic-skeleton approximations. Calcolo **33**(1–2), 47–57 (1996/1998). Toeplitz matrices: structures, algorithms and applications (Cortona, 1996)
28. Weinan, E., Ren, W., Vanden-Eijnden, E.: String method for the study of rare events. Phys. Rev. B **66**, 052301 1–4 (2002)
29. Weinan, E., Ren, W., Vanden-Eijnden, E.: Energy landscape and thermally activated switching of submicron-sized ferromagnetic elements. J. Appl. Phys. **93**, 2275–2282 (2003)
30. Weinan, E., Ren, W., Vanden-Eijnden, E.: Simplified and improved string method for computing the minimum energy paths in barrier-crossing events. J. Chem. Phys. **126**, 164103 1–8 (2007)