Multilinear optimization in low-rank models

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

DISSERTATION

zur Erlangung des akademischen Grades

DOCTOR RERUM NATURALIUM
(Dr.rer.nat.)

im Fachgebiet

Mathematik

vorgelegt

von Master of Science Henrik Eisenmann
geboren am 08.06.1993 in Husum

Die Annahme der Dissertation wurde empfohlen von:

1. Prof Dr. André Uschmajew (Universität Augsburg)
2. Prof Dr. Bernd Sturmfels (Max-Planck-Institut MiS Leipzig)
3. Prof Dr. Reinhold Schneider (Technische Universität Berlin)

Die Verleihung des akademischen Grades erfolgt mit Bestehen der Verteidigung am 13.10.2022 mit dem Gesamtprädikat magna cum laude.

Abstract

Tensors play an important role in many applications and are fundamental objects of multilinear algebra and multivariate functions. When handling partial differential equations, the solutions often lie in the closure of a tensor space. In several fields of research, for example in psychometrics, chemometrics, recommendation systems, and signal processing, data appears in the form of multidimensional arrays. Multilinear maps and homogeneous polynomials are naturally identified as tensors and symmetric tensors, respectively.

The thesis deals with several topics on multilinear optimization and optimization using low-rank models. The contribution of this thesis starts by finding the maximum relative distance that a real rank-two tensor can have to the set of rank-one tensors in Chapter 1. Equivalently, one can ask for the smallest ratio of spectral and Frobenius norm. This question is easy to answer for matrices as the distance of a matrix to rank-one matrices is given by the singular value decomposition. A rank-r matrix of Frobenius norm one has a spectral norm of at least $1/\sqrt{r}$. For tensors, the topic *best rank-one approximation ratio* is under current research. This deals with the same question but typically for tensors restricted to a certain tensor space and no further algebraic restrictions. The answer is only known for certain formats of tensor spaces and underlying fields. Motivated by the simple answer for matrices, we are interested in the maximum distance of a rank-r tensor to rank-one tensors, independent of the underlying tensor space.

The ratio of spectral and Frobenius norm for special classes of tensors is of intrinsic interest. It gives norm bounds for two natural norms in tensor spaces. The spectral norm is natural when viewing a tensor as a multilinear form while the Frobenius norm is a natural norm for a tensor product of Hilbert spaces. The distance from rank-one tensors also appears in the context of quantum entanglement. Here, tensor rank is a discrete measure of entanglement, while the distance to rank-one tensors is a continuous measure of entanglement. This work can thus be seen as exploring the difference between these two measures of entanglement.

While the answer for general r is out of scope, we provide the maximum distance in the case $r = 2$ using techniques from non-smooth optimization and utilizing the small number of parameters necessary to describe a rank-two tensor.

The second chapter is concerned with low-rank approximations of solutions to parabolic partial differential equations. Oftentimes, the domain of a partial differential equation is separable and solutions can be well approximated by functions of low rank. It then makes sense to discretize on huge grits and impose low-rank constraints. The analysis of such an approach becomes difficult, as we are looking for solutions lying on a manifold and not on a linear space. The resulting equations are known in a broader context as the *Dirac-Frenkel variational principle*. We start an analysis of the underlying infinite dimensional problem. For this, we capture properties of a model problem, an anisotropic diffusion equation, and show existence and uniqueness of solutions in a more abstract setting. Furthermore, we provide stability estimates and a convergence proof of space-discrete solutions to the underlying space-continuous solution.

Finally, we treat multiparameter eigenvalue problems in the third chapter. They appear most notably when separation of variables is applied to boundary eigenvalue problems, but spectral parameters cannot be separated. Multiparameter eigenvalue problems generalize both linear systems of equations and generalized eigenvalue problems. We summarize classical notions of definiteness that guarantee that all solutions are real. In the case of linear systems of equations, these imply that the equations are of full rank. For the generalized eigenvalue problem $(A + \lambda B)u = 0$, they imply that $\operatorname{span}\{A, B\}$ consists of symmetric matrices and contains a positive definite matrix.

Multiparameter eigenvalue problems can be solved with the help of multilinear algebra techniques by solving associated linear eigenvalue problems. These can however be huge, even if the associated original problem is of moderate size. For the case of definite multiparameter eigenvalue problems, we propose Newton-type methods to find specific solutions with certain properties, which makes the resulting equations have unique solutions. We provide local and global convergence properties and demonstrate the performance of the resulting methods in numerical experiments.

Acknowledgements

First and foremost, I would like to thank my advisor André Uschmajew. I am grateful for all the fun and insightful mathematical discussions, and also for introducing me to many other researchers. This made the last three years a successful scientific experience for me.

I would also like to thank Bernd Sturmfels for a lot of inspiring advice and support. In general, the Nonlinear Algebra group is at the Max Planck Institute for Mathematics in the Sciences is a very hospital and stimulating environment to conduct research. I would like to thank all the past and present group members for the memorable time, be it a discussion in the library or one after work.

Furthermore, I would like to thank my other coauthors Arthur Bik, Markus Bachmayr, Emil Kieri, Felix Krahmer, Max Pfeffer, and Yuji Nakatsukasa for the fruitful colaborations. Hopefully, there are more projects to come. I would also like to thank Bart Vandereycken and Reinhold Schneider for agreeing to referee my thesis, and my friends Benjamin, Leon, Edina, Monika, and Max for their help proofreading my thesis.

I also want to express my gratitude for the support of the *International Max Planck Research School* ”Mathematics in the Sciences”, both financially and for providing a favorable working environment at the Max Planck Institute. I am thankful for all the work behind the scenes, be it from the library or computer group. A special thanks goes to Saskia Gutzschebauch for helping with all the administrative tasks.

Finally, I want to thank my friends and family for the continuous support and making the last years all the more enjoyable.

Contents

Introduction

Tensors play an important role in many applications and are fundamental objects of multilinear algebra and multivariate functions. When handling partial differential equations, the solutions often lie in the closure of a tensor space. In several fields of research, for example in psychometrics, chemometrics, recommendation systems, and signal processing, data appears in the form of multidimensional arrays. Multilinear maps and homogeneous polynomials are naturally identified as tensors and symmetric tensors, respectively.

In this thesis, we cover three different topics using low-rank tensor formats building on [EU21, BEKU21, EN22] each with a different aspect of optimization in tensor spaces. The first chapter builds on [EU21]. There, we find a more geometric result concerning the relative distance of the sets of rank-one and rank-two tensors. In the second chapter, we have a theoretical existence and uniqueness result for approximations of solutions to parabolic problems using low-rank tensor formats. These results appeared mostly already in [BEKU21]. The final chapter covers multiparameter eigenvalue problems, where rank-one tensors appear naturally as solutions. We use the structure of the problem to find efficient methods to compute solutions. This chapter is motivated by [EN22] and extends its results to a more general setting.

We now give a more general introduction to the topic of tensors and multilinear optimization. Before we formally define tensors and *tensor products*, let us review two examples appearing throughout this thesis. Note that the tensor product of spaces is only defined up to isomorphism. Therefore, the following examples only show one way to identify objects as tensors.

Example. The space of real $m \times n$ matrices $\mathbb{R}^{m\times n}$ is isomorphic to the tensor space $\mathbb{R}^m \otimes \mathbb{R}^n$. Every matrix is a finite sum of *rank-one matrices* uv^T. A bilinear map $a\colon \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ is uniquely represented by a matrix $A \in \mathbb{R}^{m\times n}$ via $a(x,y) = x^\mathsf{T} Ay$ and a quadratic map $a\colon \mathbb{R}^n \to \mathbb{R}$ is uniquely represented by a symmetric matrix $A \in \operatorname{Sym}_2 \mathbb{R}^n$ via $a(x) = x^\mathsf{T} Ax$.

Example. The space of square-integrable functions $L^2(\Omega)$ over a product domain $\Omega = \Omega_1 \times \Omega_2$ is the closure of $L^2(\Omega_1) \otimes L^2(\Omega_2)$ with respect to the norm $L^2(\Omega)$ norm $\|f\|^2_{L^2(\Omega)} = \iint_\Omega |f(x,y)|^2 \, dx\, dy$. The algebraic tensor space $L^2(\Omega_1) \otimes L^2(\Omega_2)$ consists of finite sums of functions $g(x,y) = g_1(x)g_2(y)$ where $g_i \in L^2(\Omega_i)$. The inner product $\langle \cdot,\cdot \rangle_{L^2(\Omega)}$ on $L^2(\Omega)$ is naturally induced by $\langle \cdot,\cdot \rangle_{L^2(\Omega_1)}$ and $\langle \cdot,\cdot \rangle_{L^2(\Omega_2)}$ as $\langle f_1 f_2, g_1 g_2 \rangle_{L^2(\Omega)} = \langle f_1, g_1 \rangle_{L^2(\Omega_1)} \langle f_2, g_2 \rangle_{L^2(\Omega_2)}$. Oftentimes, $L^2(\Omega_1) \otimes L^2(\Omega_2)$ already denotes the closure with respect to this norm [Hac19, Chapter 4].

An elegant way to define the tensor product is via the *universal property*; see e.g., [Gre67, Chapter I §2] or [Vak17, Chapter 1.3.5].

Definition (Tensor product via universal property)**.** Let $\mathcal{U}, \mathcal{V}$ and $\mathcal{T}$ be vector spaces and $\otimes\colon \mathcal{U} \times \mathcal{V} \to \mathcal{T}$, $(u, v) \mapsto u \otimes v$ be a bilinear map. The pair $(\mathcal{T}, \otimes)$ is called a tensor product for $\mathcal{U}$ and $\mathcal{V}$ if for every bilinear map $a : \mathcal{U} \times \mathcal{V} \to \mathcal{W}$ to a linear space $\mathcal{W}$ there is a unique linear map $L\colon \mathcal{T} \to \mathcal{W}$ such that $a(u, v) = L(u \otimes v)$.

The definition is summarized in the commutative diagram

$$\begin{tikzcd}
\mathcal{U} \times \mathcal{V} \arrow[r, "\otimes"] \arrow[dr, "a"'] & \mathcal{T} \arrow[d, dashed, "\exists! L"] \\
 & \mathcal{W}
\end{tikzcd} .$$

From this definition follows directly that any two tensor products for $\mathcal{U}$ and $\mathcal{V}$ are uniquely isomorphic. In absence of a representative, we write $\mathcal{U} \otimes \mathcal{V}$ for the tensor product of $\mathcal{U}$ and $\mathcal{V}$. In the definition, we may replace the existence of a unique linear map with just the existence of a linear map if in addition $\mathcal{T} = \operatorname{span}\{u \otimes v\colon u \in \mathcal{U},\ v \in \mathcal{V}\}$. The tensor product for two linear spaces $\mathcal{U}$ and $\mathcal{V}$ always exists and there is an explicit construction. This can be used as an alternative definition of the tensor product; see e.g., [Hac19, Chapter 3.2.1].

Example. Let $a\colon \mathbb{R}^m \times \mathbb{R}^n \to \mathcal{V}$ be a bilinear map. Define $L\colon \mathbb{R}^{m \times n}$ as the linear map satisfying $L(uv^\mathsf{T}) = a(u, v)$. This map is determined uniquely since $\{uv^\mathsf{T}\colon u \in \mathbb{R}^m,\ v \in \mathbb{R}^n\}$ is a generating system for $\mathbb{R}^{m \times n}$. Therefore, $(u, v) \mapsto uv^\mathsf{T}$ defines a tensor product.

The tensor products $\mathcal{U} \otimes (\mathcal{V} \otimes \mathcal{W})$ and $(\mathcal{U} \otimes \mathcal{V}) \otimes \mathcal{W}$ are uniquely isomorphic such that the diagram

$$\begin{tikzcd}
\mathcal{U} \times \mathcal{V} \times \mathcal{W} \arrow[r] \arrow[dr] & \mathcal{U} \times (\mathcal{V} \otimes \mathcal{W}) \arrow[r] & \mathcal{U} \otimes (\mathcal{V} \otimes \mathcal{W}) \arrow[d, leftrightarrow] \\
 & (\mathcal{U} \otimes \mathcal{V}) \times \mathcal{W} \arrow[r] & (\mathcal{U} \otimes \mathcal{V}) \otimes \mathcal{W}
\end{tikzcd}$$

commutes. By successively applying the universal property there is a unique linear map from $\mathcal{U} \otimes (\mathcal{V} \otimes \mathcal{W})$ to $(\mathcal{U} \otimes \mathcal{V}) \otimes \mathcal{W}$ and vice versa. Note that the maps from $\mathcal{U} \times \mathcal{V} \times \mathcal{W}$ to $\mathcal{U} \otimes (\mathcal{V} \otimes \mathcal{W})$ and $(\mathcal{U} \otimes \mathcal{V}) \otimes \mathcal{W}$ are trilinear. First, we can lift the trilinear map from $\mathcal{U} \times \mathcal{V} \times \mathcal{W}$ to $(\mathcal{U} \otimes \mathcal{V}) \otimes \mathcal{W}$ uniquely to a bilinear map from $\mathcal{U} \times (\mathcal{V} \otimes \mathcal{W})$ to $(\mathcal{U} \otimes \mathcal{V}) \otimes \mathcal{W}$ and finally a unique linear map from $\mathcal{U} \otimes (\mathcal{V} \otimes \mathcal{W})$ to $(\mathcal{U} \otimes \mathcal{V}) \otimes \mathcal{W}$. It is therefore reasonable to speak of the tensor product of d spaces for any $d \geq 2$. An element of a tensor product of d spaces is called a tensor of order d. It is one important aspect of the definition of tensors via the universal property that we get a direct way to make multilinear maps linear by going to the tensor product of the respective spaces.

Example. The space of real multidimensional arrays or hypermatrices $\mathbb{R}^{n_1 \times \ldots \times n_d}$ is isomorphic to the tensor space $\bigotimes_{i=1}^d \mathbb{R}^{n_i}$. The coordinate vectors $e_{i_1, \ldots, i_d}$ are then

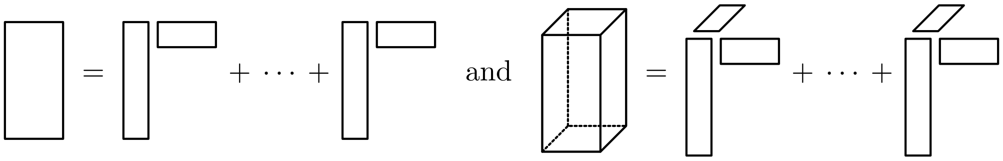


Figure 1: Decomposition of matrices and tensors into a sum of outer products.

associated with $e_{i_1} \otimes e_{i_2} \otimes \ldots \otimes e_{i_d}$. Every multidimensional array $a_{i_1,\ldots,i_d}$ is a sum of products of one-dimensional arrays $u_{1,i_1} u_{2,i_2} \ldots u_{d,i_d}$, for example as the weighted sum of the coordinate vectors $e_{i_1,\ldots,i_d}$.

The decomposition of a three-dimensional array into a sum of products of one-dimensional arrays can be visualized similarly to matrices. For matrices, we have a sum of outer products uv^{T} and for three-dimensional arrays, we multiply a third vector; see Figure 1. In data science, tensors and multidimensional arrays are taken as synonyms, and we also often refer to multidimensional arrays as tensors.

The tensor product can be defined using the universal property or by an explicit construction. A third way to define the tensor product is via multilinear maps; see e.g., [Lan12, Chapter 2.3]. Let $\mathcal{U}$ and $\mathcal{V}$ be vector spaces over a field k and $\mathcal{U}^*$ and $\mathcal{V}^*$ be their respective dual spaces. Then the set of linear maps $\mathcal{U}^* \to \mathcal{V}$ and the set of bilinear maps $\mathcal{U}^* \times \mathcal{V}^* \to k$ is also denoted by $\mathcal{U} \otimes \mathcal{V}$. If the spaces $\mathcal{U}$ and $\mathcal{V}$ are finite-dimensional, this defines the same object as in the definition via universal property up to isomorphism. This definition directly identifies tensors as multilinear maps. However, in infinite dimensions, this definition does not coincide with the one via the universal property.

Example. Let $\mathcal{U} = \ell^2 \cong \mathcal{U}^*$ be the space of square-summable real sequences and let $\langle \cdot, \cdot \rangle$ denote the inner product on ℓ^2. This defines a bilinear map from $\mathcal{U} \times \mathcal{U} \cong \mathcal{U}^* \times \mathcal{U}^* \to \mathbb{R}$, but is itself no element of $\mathcal{U} \otimes \mathcal{U}$ as it cannot be generated by finitely many bilinear maps of the form $(u, v) \mapsto \langle u, u_i \rangle \langle v, v_i \rangle$. Similarly, the identity map $\mathcal{U} \to \mathcal{U}$ is no element of $\mathcal{U} \otimes \mathcal{U}$ as $\mathcal{U}$ has no finite basis.

For infinite-dimensional spaces, the tensor product is usually not complete with respect to natural norms defined on it. It often makes sense to directly define a topological tensor product of vector spaces. Given a suitable norm, this is the closure of the algebraic definition of tensor spaces. For further details, we refer to [Hac19, Chapter 4].

Matrix and tensor rank and decompositions

Let us recall the rank of matrices and linear operators from linear algebra.

Definition (Rank of matrices and linear operators)**.** Let $A \in \mathbb{R}^{m \times n}$ be a matrix. Then the rank of A is

$$\begin{aligned}\operatorname{rank} A &= \dim \operatorname{span}\{a\colon a \text{ is a column of } A\} = \dim\{Ax\colon x \in \mathbb{R}^n\}\\ &= \dim \operatorname{span}\{a\colon a \text{ is a row of } A\} = \dim\{A^\mathsf{T}x\colon x \in \mathbb{R}^m\}\\ &= \min\{r\colon \text{there exist } u_i \text{ and } v_i \text{ such that } A = \sum_{i=1}^r u_i v_i^\mathsf{T}\}.\end{aligned}$$

Similarly let $L\colon \mathcal{U} \to \mathcal{V}$ be a linear operator and $L^*\colon \mathcal{V}^* \to \mathcal{U}^*$ be its dual operator. Then

$$\begin{aligned}&\operatorname{rank} L = \dim \operatorname{range} L = \dim \operatorname{range} L^*\\ &\qquad = \min\{r\colon \text{there exist } u_i \in \mathcal{U}^* \text{ and } v_i \in \mathcal{V} \text{ such that } L(x) = \sum_{i=1}^r u_i(x) v_i\}.\end{aligned}$$

It is a basic fact from linear algebra that all these definitions of rank coincide. One of the most important objects in this work will be the matrices of rank one uv^T and their pendant to tensors. A *rank-one tensor* is the tensor product $u_1 \otimes u_2 \otimes \ldots \otimes u_d$ of nonzero vectors $u_1, \ldots, u_d$. They constitute the nonzero image of the multilinear map $\bigtimes_{i=1}^d \mathcal{U}_i \to \bigotimes_{i=1}^d \mathcal{U}_i$ which is a cone. Its projective version is called the *Segre variety.*

A decomposition into a sum of rank-one matrices $u_i v_i^\mathsf{T}$ for a matrix A can also be expressed as a *matrix factorization* $A = UV^\mathsf{T}$ where the columns of U and V consist of the vectors u_i and v_i, respectively. Finding such a factorization plays an important role in data science. One famous example is the Netflix price [BL07]. This is a recommendation system problem, often solved using matrix factorization techniques [KBV09]. Here incomplete user and product data is stored in a matrix, where the entries measure how much a user likes a given item. A prediction for a user can be computed via matrix completion under the assumption that the data has a low-rank structure, see e.g., [CR09, Van13].

However, the use of matrices is sometimes limited. To examine this, let us review fluorescence spectroscopy, a method in chemical science, described in [SBG04, Chapter 10.2]. We have a mixture of k fluorescent substances, each having an excitation spectrum a_i and emission spectrum b_i. We assume linear behaviour, i.e., when exciting substance i with the wavelengths x, the measured emission is $b_i(a_i^\mathsf{T} x)$. Each substance has a concentration c_i. For the mixture, we get the matrix $M = \sum_{i=1}^k c_i b_i a_i^\mathsf{T}$ which we can measure. We would like to find the different substances and their excitation and emission spectrum, i.e., the factorization BCA^T, where A and B contain a_i and b_i

as its columns and C is a diagonal matrix containing the concentrations c_i. However, the matrices cannot be recovered uniquely, even when neglecting the trivial scaling ambiguities. Indeed, a matrix factorization is never unique. Let $A = UV^\mathsf{T}$ with $U \in \mathbb{R}^{m\times k}$ and $V \in \mathbb{R}^{n\times k}$. Then $A = (US)(S^{-1}V^\mathsf{T})$ holds true for any invertible matrix $S \in \mathbb{R}^{k\times k}$. One solution to this problem is to repeat the experiment ℓ times with different concentrations. We then measure the data in $\mathbf{M} \in \mathbb{R}^{\ell\times m\times n}$ with ℓ different mixtures, m different measurements of the emission spectrum, and n different excitation spectra. We now want to find matrices A and B such that for each different mixture j we find a diagonal matrix C_j containing the concentrations such that $M_j = BC_jA^\mathsf{T}$. Finding this parallel factorization is equivalent to finding the decomposition into a sum of rank-one tensors

$$\mathbf{M} = \sum_{i=1}^{k} c_i \otimes b_i \otimes a_i,$$

where c_i is now a vector storing the concentrations of substance i in the mixture j. A decomposition into a sum of rank-one tensors of order at least 3 is often unique when disregarding the order. Therefore, the rank-one tensors in the decomposition can have actual meaning and are of interest when analyzing data. This leads to the following generalization of rank to tensors.

Definition (Rank of a tensor)**.** Let $\mathbf{A} \in \bigotimes_{i=1}^{d} \mathcal{U}_i$ be a tensor of order d. The rank of $\mathbf{A}$ is the minimal number r such that

$$\mathbf{A} = \sum_{i=1}^{r} u_{1,i} \otimes \ldots \otimes u_{d,i}$$

denoted by $\operatorname{rank} \mathbf{A}$.

Note that tensors in the algebraic tensor product always have finite rank, but when at least two spaces are infinite-dimensional there is no upper bound for the rank. For a tensor $\mathbf{A}$ in the closure of the algebraic tensors product, it is possible that $\operatorname{rank} \mathbf{A} = \infty$. The decomposition of a tensor into a sum of rank-one tensors is known under different names in different fields. The name *canonical decomposition* (CANDECOMP) was introduced in [CC70], and the procedure as *parallel factor analysis* (PARAFAC) in [Har70]. In mathematics, it is now often known as the *canonical polyadic decomposition* (CPD).

The problem of uniqueness of a decomposition is handled in algebraic geometry under *identifiability* of tensors. Many results on uniqueness rely on *Kruskal's criterion* [Kru77]. There are also criteria from algebraic geometry, see e.g., [Lan12, Chapter 12.3], and the problem of identifiability is still being researched.

The tensor rank defined above only generalizes the last equality in the definition of the rank of matrices. The notion of column and row rank in the first two equalities is better generalized as the *multilinear rank* of a tensor.

Definition (Multilinear rank of a tensor)**.** Let $\mathbf{A} \in \bigotimes_{i=1}^{d} \mathcal{U}_i$ and view it as the linear maps $A_j \colon \mathcal{U}_j^* \to \bigotimes_{i \neq j} \mathcal{U}_i$ for $j = 1, \ldots, d$. Then its multilinear rank is the d-tuple $\operatorname{rank}_{\mathsf{multilin}} \mathbf{A} = (\operatorname{rank} A_1, \ldots, \operatorname{rank} A_d)$.

The multilinear rank is also called *tensor subspace rank* or *Tucker rank*. Matrices are the special case, since for a matrix A the multilinear rank is just $(\operatorname{rank} A, \operatorname{rank} A)$. For general tensors we only have the inequalities $r_j \leq \operatorname{rank} \mathbf{A} \leq \prod_{i \neq j} r_i$ where $(r_1, \ldots, r_d) = \operatorname{rank}_{\mathsf{multilin}} \mathbf{A}$ and in general no r_j has to be equal to $\operatorname{rank} \mathbf{A}$.

This is not the only way that tensor and matrix rank behave differently. For instance, let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix. Then A has the *symmetric decomposition* $A = \sum_{i=1}^{r} \lambda_i u_i u_i^{\mathsf{T}}$ for some $u_i \in \mathbb{R}^n$ and $\lambda_i \in \mathbb{R}$. Such a decomposition can be generated by an eigenvalue decomposition of $A = U \Lambda U^{\mathsf{T}}$ with an orthogonal matrix U. We can define the *symmetric rank* of a matrix as the smallest number r such that a symmetric decomposition exists. It is again a basic fact from linear algebra that the symmetric rank coincides with the usual rank. For real symmetric matrices, this is simply because the multiplicity of the eigenvalue zero coincides with the dimension of the kernel of the symmetric matrix, which determines the dimension of its image and therefore its rank. We can also define the symmetric rank of symmetric tensors.

Definition (Symmetric tensors and symmetric rank)**.** Let $\mathfrak{S}_d$ be the set of permutations of d elements. We associate every permutation $\sigma \in \mathfrak{S}_d$ with a linear map $\bigotimes_{i=1}^{d} \mathcal{U} \to \bigotimes_{i=1}^{d} \mathcal{U}$ generated by

$$u_1 \otimes \ldots \otimes u_d \mapsto u_{\pi(1)} \otimes \ldots \otimes u_{\pi(d)}$$

and denote it again by σ. A tensor $\mathbf{A}$ in $\bigotimes_{i=1}^{d} \mathcal{U}$ is symmetric if $\sigma(\mathbf{A}) = \mathbf{A}$ for all $\sigma \in \mathfrak{S}_d$. We denote the set of symmetric tensors in $\bigotimes_{i=1}^{d} \mathcal{U}$ as $\operatorname{Sym}_d \mathcal{U}$ and the symmetric rank-one tensor $u \otimes u \otimes \ldots \otimes u$ as u^d. The symmetric rank of a symmetric tensor $\mathbf{A}$ is the smallest number r such that $\mathbf{A} = \sum_{i=1}^{r} \lambda_i u_i^d$ denoted by $\operatorname{rank}_{\mathsf{S}} \mathbf{A} = r$.

In many special cases, it can be shown that $\operatorname{rank}_{\mathsf{S}} \mathbf{A} = \operatorname{rank} \mathbf{A}$ for symmetric tensors and therefore it was conjectured that symmetric rank and rank coincide in general [CGLM08]. However, Shitov provided a counterexample in [Shi18] and therefore both notions of rank for symmetric tensors differ in general.

Also unlike matrix rank, the rank of a tensor depends on the field. For example, tensors in $\mathbb{R}^{2 \times 2 \times 2}$ can also be considered as tensors in $\mathbb{C}^{2 \times 2 \times 2}$ but the ranks may differ. Consider the tensor $\mathbf{A} = e_{111} - e_{112} - e_{121} - e_{211}$ where $e_{ijk} = e_i \otimes e_j \otimes e_k$. Using *Cayley's hyperdeterminant* one can show that $\operatorname{rank} \mathbf{A} = 3$ as a real tensor [dSL08, Proposition 5.10] but we can be decompose $\mathbf{A} = \frac{1}{2}\left((e_1 + ie_2)^3 + (e_1 - ie_2)^3\right)$ using complex rank-one tensors.

Low-rank approximation

In many applications, data is not given directly as matrices or tensors of low rank. It can however make sense to find a good low-rank approximation for further analysis. Let for example N data points be given in a d-dimensional space. These can be stored in a matrix $A \in \mathbb{R}^{d\times N}$. Here, a *principal component analysis* can be helpful. For this, let $\mu \in \mathbb{R}^d$ be the mean of the data and $\mathbf{1}_N \in \mathbb{R}^N$ be the vector containing only ones. Assuming $d \le N$, we can then express A in the form

$$A = \mu\mathbf{1}_N^\mathsf{T} + \sum_{i=1}^{d} \sigma_i u_i v_i^\mathsf{T} = \mu\mathbf{1}_N^\mathsf{T} + U\Sigma V^\mathsf{T}$$

with orthonormal sets of vectors $u_i \in \mathbb{R}^d$, $v_i \in \mathbb{R}^N$ and descending positive real values σ_i. The vectors u_i are the eigenvectors of the covariance matrix $(A - \mu\mathbf{1}_N^\mathsf{T})(A - \mu\mathbf{1}_N^\mathsf{T})^\mathsf{T}$. It can be helpful to project the data onto the subspace spanned by only the first few eigenvectors u_i, for example, to lower the dimensions or to find clusters.

The principal component analysis is an instance of the *singular value decomposition* (SVD). Any $m \times n$ matrix can be decomposed as $A = U\Sigma V^\mathsf{T} = \sum_{i=1}^{\min\{m,n\}} \sigma_i u_i v_i^\mathsf{T}$ where the matrices U and V have orthonormal columns. The respective columns are the left and right singular vectors u_i and v_i, and the nonnegative descending entries $\sigma_1 \ge \sigma_2 \ge \ldots \ge 0$ of the diagonal matrix $\Sigma \in \mathbb{R}^{m\times n}$ are the singular values. The singular values σ_i contain important information of the matrix. For instance, its Frobenius norm is

$$\|A\|_\mathsf{F} := \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |A_{ij}|^2} = \sqrt{\sum_{i=1}^{\min\{n,m\}} \sigma_i^2}$$

and its spectral norm is

$$\|A\|_\sigma := \max_{\|u\|_2 = 1 = \|v\|_2} u^\mathsf{T} A v = \sigma_1.$$

The singular value decomposition is also the tool to find the best rank k approximation of a matrix in Frobenius or spectral norm.

Theorem (Schmidt, Eckhart-Young-Mirsky)**.** *Let $A = U\Sigma V^\mathsf{T}$ be a singular value decomposition for a matrix $A \in \mathbb{R}^{m\times n}$. Then*

$$\min_{\operatorname{rank} B \le k} \|A - B\|_\mathsf{F} = \sqrt{\sum_{i=k+1}^{\min\{n,m\}} \sigma_i^2} \quad \text{and} \quad \min_{\operatorname{rank} B \le k} \|A - B\|_\sigma = \sigma_{k+1}$$

and the minimum is attained for $\sum_{i=1}^{k} \sigma_i u_i v_i^\mathsf{T} = U_k \Sigma_k V_k^\mathsf{T}$, where $U_k \in \mathbb{R}^{m\times k}$ and $V_k \in \mathbb{R}^{n\times k}$ consist of the first k left and respectively right singular vectors and $\Sigma_k \in \mathbb{R}^{k\times k}$ is the diagonal matrix with the first k singular values as its entries.

This approximation was first discovered by Schmidt in 1907 for the case of functions in two variables and the L^2 norm in [Sch07, § 18]. Later Eckhart and Young formulated this theorem for matrices and the Frobenius norm in [EY36], and Mirsky found a generalization for unitarily invariant norms in [Mir60].

It is not obvious why data should be close to low rank. However, an approach to explain that this still works in many cases is evaluated in [UT19]. The authors show that matrices generated by certain *latent variable models* can be well approximated by low-rank matrices. It is also worth mentioning that even if N points of data are not close to a low-dimensional subspace, there is a linear map L onto a space of dimension $k \approx \frac{\log N}{\epsilon^2}$, which depends only on the number of points N and not their dimension, such that

$$(1-\epsilon)\|x-y\|_2 \leq \|L(x-y)\|_2 \leq (1+\epsilon)\|x-y\|_2$$

for any two of the data points x and y, i.e., distances are preserved approximately. This was first proven by Johnson and Lindenstrauss in [JL84] and is known as the *Johnson-Lindenstrauss lemma*. It is therefore possible to recover many properties using a low-rank model approximately even when the data is not inherently close to low rank.

One other application is to approximate functions on a product domain $\Omega_1 \times \Omega_2$, where $\Omega_i \subset \mathbb{R}^{n_i}$. We can discretize such a function f on an $N_1 \times N_2$ grid. Here, N_1 and N_2 can already be quite huge since they itself have to cover a multidimensional domain. Storing such an approximation directly can get infeasible quickly. However, if $f(x,y) \approx \sum_{i=1}^{k} g_i(x)h_i(y)$ with a small k, we can also store an approximation as a product of $N_1 \times k$ and $N_2 \times k$ matrices and make storing feasible. The naturally arising question is when functions in two variables can be approximated well by a small number of products of functions in one variable. If functions are sufficiently smooth, then a decay of singular values in the analog of the singular value decomposition for functions in $L^2(\Omega_1 \times \Omega_2) = \overline{L^2(\Omega_1) \otimes L^2(\Omega_2)}^{\|\cdot\|_{L^2(\Omega)}}$ can be observed. Here, the key is *mixed regularity*, i.e., the function is also bounded in the norm induced by the inner product $\langle \cdot, \cdot \rangle_{H^{s,s}_{\text{mix}}(\Omega_1 \times \Omega_2)} = \langle \cdot, \cdot \rangle_{H^s(\Omega_1)} \langle \cdot, \cdot \rangle_{H^s(\Omega_2)}$ [SU14, GH14], where H^s is the Sobolev space containing functions with square integrable weak derivatives up to order s. However, it is suggested in [GH14] that low-rank approximation is not optimal in dimension reduction and instead the use of *sparse grids* is proposed.

In general, it makes sense to reduce the complexity to store solutions of high-dimensional problems. For the *Sylvester equation*

$$AX - XB = C,$$

it was shown in [Gra04b] that, if the matrix C is of low rank, the solution X can be approximated with error ϵ with matrices of rank $\mathcal{O}(-\log(\epsilon))$, i.e., the error decays exponentially with respect to rank. This is generalized to finite-dimensional tensor equations of a similar form in [Gra04a]. For the infinite-dimensional *Lyapunov equation* with unbounded operators, it was shown that the error decays almost exponentially in [GK14] and a similar result for certain elliptic partial differential equations was shown

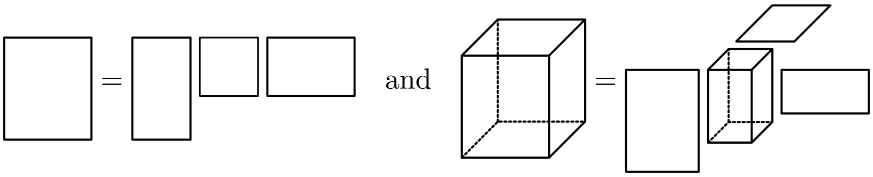


Figure 2: Subspace representation of tensors with order 2 and 3.

in [DDGS16]. In Chapter 2, motivated by the fact that solutions to partial differential equations are often well approximated by functions in a low-rank model, we provide existence and uniqueness results for a certain way to obtain low-rank approximations to solutions of parabolic problems.

We have collected arguments why data and functions are often close to low-rank matrices and tensors, and the best approximation of matrices in Frobenius norm can be found via the singular value decomposition. For tensors of higher order, we do not have an analogous tool. One problem is that tensors of bounded rank do not necessarily form a closed set. A prime example is the tensor

$$u \otimes u \otimes v + u \otimes v \otimes u + v \otimes u \otimes u = \lim_{t \to 0} \frac{1}{t} \left((u + tv)^3 - u^3 \right),$$

a rank-three tensor which is a limit of rank-two tensors. Only tensors of rank at most one form a closed set in every tensor space. Therefore, some tensors do not have a best rank-k approximation. In [dSL08] it is shown that over the real numbers there are instances where these tensors do not form isolated sets but have positive volume. Hence, the approximation problem is not always well-posed. Many other tensor-related problems, like determining rank, approximating analogs of singular and eigenvectors, and approximating its spectral norm, are NP-hard [HL13]. There is an analogy to the Eckhart-Young Theorem that was found in [DOT18]. For a sufficiently general complex tensor, its best rank-k approximation lies in the linear hull of its critical rank-one approximations. Note, however, that there are too many critical rank-one approximations. A general symmetric tensor $\mathbf{A} \in \operatorname{Sym}_d \mathbb{C}^n$ has $\frac{(d-1)^n - 1}{d-2} = \sum_{i=0}^{n-1} (d-1)^i$ different eigenvalues [CS13]. Chapter 1 is related to this topic. We explore properties of critical rank-one approximations to rank-two tensors.

For a stable approximation of tensors, subspace representations are used. Instead of representing a tensor $\mathbf{A}$ in the full space $\bigotimes_{i=1}^d \mathcal{U}_i$, an approximation $\tilde{\mathbf{A}} \in \bigotimes_{i=1}^d \mathcal{V}_i$ with $\mathcal{V}_i \subset \mathcal{U}_i$ is used. A quasi-optimal approximation can be found using the *multilinear singular value decomposition* [DLDMV00a] and, when needed, better approximations can be obtained using iterative methods [DLDMV00b]. Notably, the smallest possible subspaces $\mathcal{V}_i$ for an exact representation reveal $\operatorname{rank}_{\text{multilin}} \mathbf{A}$ just as a matrix factorization with smallest possible dimensions are rank revealing. For matrices, the singular value decomposition is a subspace representation, where the subspaces get stored in the singular vectors u_i and v_i; see Figure 2 for an illustration. For tensors of large order d, hierarchical tensor formats were introduced in [HK09, OT09] and have been widely used in numerical tensor calculus.

Models with exact low-rank solutions

There are many interesting optimization problems where the exact rank is already known. One class of problems appears when an original bilinear or quadratic problem gets lifted to a linear problem on the tensor space. One example is *blind deconvolution*. A convolution $u * v = w$ is bilinear in u and v and therefore by the universal property there is a linear map L such that $L(u \otimes v) = u * v = w$. Blurring of pictures and reverberation in acoustics are instances where a convolution of a signal u, say the picture or the sound, and a point spread function v, say a bad lens or reflections off walls, appears. When v is known, the process of deconvoluting u from w is solving a linear system. When v is unknown, the deconvolution becomes blind and both u and v have to be computed. Often, the point spread vector is also of interest, for example in seismology. To allow for unique recovery, additional assumptions on u and v have to be made, for example, they are assumed to lie in low-dimensional subspaces $\mathcal{U}$ and $\mathcal{V}$. In [ARR14] this problem is solved by recovering both u and v as the rank-one matrix $uv^\mathsf{T} = A$ from the linear measurements $L(A) = w$. Since the dimension of $\mathcal{U} \otimes \mathcal{V}$ is much higher than w, there are many possible solutions A. Among these solutions, the one with the lowest rank is desired. Minimizing rank is however a non-convex and NP-hard problem. Instead, the convex problem of minimizing the nuclear norm $\|A\|_* = \sum_{i=1}^n \sigma_i$ is utilized. The unit ball of matrices in nuclear norm is the convex envelope of rank-one matrices with Frobenius norm one. Minimizing the nuclear norm of an affine linear set can be solved with *semidefinite programming* [RFP10].

A similar problem is phase retrieval. Here, we have the measurements $|\langle u, v_i \rangle|^2$ of a complex vector $u \in \mathbb{C}^n$. We lost information on the phase in these measurements. In [CSV13] this problem is handled by lifting the real quadratic measurements to a linear map on Hermitian matrices and again minimizing nuclear norm to find uu^H.

Also, many eigenvalue problems have low-rank solutions. Consider for example the matrix-valued eigenvalue problem

$$AX + XB = \lambda X.$$

Its eigenvalues are given by $\mu_i + \nu_j$ with corresponding eigenvectors $u_i v_j^\mathsf{T}$, where (μ_i, u_i) and (ν_j, v_j) are the eigenvalues and eigenvectors of A and B^T, respectively. Similarly, the eigenfunctions of the Laplacian on a rectangular domain

$$\Delta f(x,y) := \frac{\partial^2}{\partial x^2} f(x,y) + \frac{\partial^2}{\partial y^2} f(x,y) = \lambda f(x,y) \quad \text{for } (x,y) \in (a,b) \times (c,d)$$

factorize as $f(x,y) = g(x)h(y)$ where $g''(x) = \mu g(x)$ and $h''(y) = \nu h(y)$. In both cases, the operator has the structure $L_1 \otimes \mathrm{id} + \mathrm{id} \otimes L_2$. Separating coordinates can often lead to factorized solutions. One other example is the Laplacian in polar coordinates. This is given by

$$\Delta f(r,\varphi) = \frac{\partial^2}{\partial r^2} f(r,\varphi) + \frac{1}{r}\frac{\partial}{\partial r} f(r,\varphi) + \frac{1}{r^2}\frac{\partial^2}{\partial \phi^2} f(r,\varphi)$$

and its eigenfunctions on a disc are given by $g(r)h(\varphi)$ where $h(\phi) = a\sin(n\varphi) + b\cos(n\varphi)$ and g is a fitting Bessel function. There are many possibilities to separate coordinates, but not always spectral parameters can be decoupled. Then a product ansatz can lead to multiparameter eigenvalue problems. These are the prime focus of Chapter 3.

Contribution of this thesis

The thesis deals with several topics on multilinear optimization and optimization using low-rank models. The contribution of this thesis starts by finding the maximum relative distance that a real rank-two tensor can have to the set of rank-one tensors in Chapter 1. Equivalently, one can ask for the smallest ratio of spectral and Frobenius norm. This question is easy to answer for matrices as the distance of a matrix to rank-one matrices is given by the singular value decomposition. A rank-r matrix of Frobenius norm one has a spectral norm of at least $1/\sqrt{r}$. For tensors, the topic *best rank-one approximation ratio* is under current research. This deals with the same question but typically for tensors restricted to a certain tensor space and no further algebraic restrictions. The answer is only known for certain formats of tensor spaces and underlying fields. Motivated by the simple answer for matrices, we are interested in the maximum distance of a rank-r tensor to rank-one tensors, independent of the underlying tensor space.

The ratio of spectral and Frobenius norm for special classes of tensors is of intrinsic interest. It gives norm bounds for two natural norms in tensor spaces. The spectral norm is natural when viewing a tensor as a multilinear form while the Frobenius norm is a natural norm for a tensor product of Hilbert spaces. The distance from rank-one tensors also appears in the context of quantum entanglement. Here, tensor rank is a discrete measure of entanglement, while the distance to rank-one tensors is a continuous measure of entanglement. This work can thus be seen as exploring the difference between these two measures of entanglement.

While the answer for general r is out of scope, we provide the maximum distance in the case $r = 2$ using techniques from non-smooth optimization and utilizing the small number of parameters necessary to describe a rank-two tensor. Here, Section 1.2, which deals with symmetric rank-two tensors, appeared in slightly altered form in the preprint [EU21]. The content of Section 1.1 is a novel contribution.

The second chapter is concerned with low-rank approximations of solutions to parabolic partial differential equations. Oftentimes, the domain of a partial differential equation is separable and solutions can be well approximated by functions of low rank. It then makes sense to discretize on huge grits and impose low-rank constraints. The analysis of such an approach becomes difficult, as we are looking for solutions lying on a manifold and not on a linear space. The resulting equations are known in a broader context as the *Dirac-Frenkel variational principle.* We start the analysis of the underlying infinite-dimensional problem. For this, we capture properties of a model problem, an anisotropic diffusion equation, and show existence and uniqueness of solutions in a more abstract setting. Furthermore, we provide stability estimates and a convergence proof of space-discrete solutions to the underlying space-continuous solution. The contents of Chapter 2 appeared mostly in altered form in [BEKU21]. The results in Theorem 2.6 and Section 2.5 are new contributions.

Finally, we treat multiparameter eigenvalue problems in the third chapter. They appear most notably when separation of variables is applied to boundary eigenvalue problems, but spectral parameters cannot be separated. Multiparameter eigenvalue problems generalize both linear systems of equations and generalized eigenvalue problems. We summarize classical notions of definiteness that guarantee that all solutions are real. In the case of linear systems of equations, these imply that the equations are of full rank. For the generalized eigenvalue problem $(A + \lambda B)u = 0$, they imply that $\operatorname{span}\{A, B\}$ consists of symmetric matrices and contains a positive definite matrix.

Multiparameter eigenvalue problems can be solved with the help of multilinear algebra techniques by solving associated linear eigenvalue problems. These can however be huge, even if the associated original problem is of moderate size. For the case of definite multiparameter eigenvalue problems, we propose Newton-type methods to find specific solutions with certain properties, which makes the resulting equations have unique solutions. We provide local and global convergence properties and demonstrate the performance of the resulting methods in numerical experiments. Chapter 3 builds on the results of [EN22], where two-parameter eigenvalue problems are treated. The convergence results in [EN22] can be seen as special cases of the results in Section 3.3 and Section 3.4.

Chapter 1

Maximum relative distance of real rank-two to rank-one tensors

Let $\mathbf{A} \in \bigotimes_{i=1}^{d} \mathcal{H}_i$ be a tensor and $\mathcal{H}_i$ be Hilbert spaces. There are two very natural norms for the tensor $\mathbf{A}$. One norm is inherited from the induced inner product

$$\langle u_1 \otimes u_2 \otimes \ldots \otimes u_d, v_1 \otimes v_2 \otimes \ldots \otimes v_d \rangle_{\mathsf{F}} = \langle u_1, v_1 \rangle \langle u_2, v_2 \rangle \ldots \langle u_d, v_d \rangle.$$

The induced inner product and the respective norm are called Frobenius inner product and Frobenius norm, also known as the *Hilbert-Schmidt norm* for linear operators and the *Schur norm* for matrices. For multidimensional arrays $\mathbf{A} = [a_{i_1,\ldots,i_d}]_{i_1=1,\ldots,i_d=1}^{n_1,\ldots,n_d}$, this is just the Euclidean norm

$$\|\mathbf{A}\|_{\mathsf{F}} = \sqrt{\sum_{i_1=1}^{n_1} \cdots \sum_{i_d=1}^{n_d} |a_{i_1,\ldots,i_d}|^2}.$$

The other norm is inherited from the associated multilinear operator and is called the spectral norm. It is defined as

$$\|\mathbf{A}\|_\sigma = \max_{\|w_1\|=1,\ldots,\|w_d\|=1} |\langle \mathbf{A}, w_1 \otimes w_2 \otimes \ldots \otimes w_d \rangle_{\mathsf{F}}|. \tag{1.1}$$

The Frobenius and spectral norm determine the distance in Frobenius norm of a tensor to the closest rank-one tensor since

$$\min_{\operatorname{rank} \mathbf{B}=1} \|\mathbf{A} - \mathbf{B}\|_{\mathsf{F}}^2 = \|\mathbf{A}\|_{\mathsf{F}}^2 - \|\mathbf{A}\|_\sigma^2. \tag{1.2}$$

The relative distance $\min_{\operatorname{rank} \mathbf{B}=1} \|\mathbf{A} - \mathbf{B}\|_{\mathsf{F}} / \|\mathbf{A}\|_{\mathsf{F}}$ is determined by the ratio of both norms. Obviously, $\|\mathbf{A}\|_\sigma \leq \|\mathbf{A}\|_{\mathsf{F}}$. If the tensor space is finite dimensional, then there is a constant c such that $\|\mathbf{A}\|_\sigma \geq c \|\mathbf{A}\|_{\mathsf{F}}$ for all $\mathbf{A} \in \bigotimes_{i=1}^{d} \mathcal{H}_i$. The smallest such constant determines the maximum relative distance of a tensor to rank-one tensors

$$\max_{\mathbf{A} \in \bigotimes_{i=1}^{d} \mathcal{H}_i} \min_{\operatorname{rank} \mathbf{B}=1} \frac{\|\mathbf{A} - \mathbf{B}\|_{\mathsf{F}}}{\|\mathbf{A}\|_{\mathsf{F}}} = \max_{\mathbf{A} \in \bigotimes_{i=1}^{d} \mathcal{H}_i} \sqrt{1 - \frac{\|\mathbf{A}\|_\sigma^2}{\|\mathbf{A}\|_{\mathsf{F}}^2}}.$$

The minimal possible ratio $\|\mathbf{A}\|_\sigma / \|\mathbf{A}\|_{\mathsf{F}}$ that can be achieved is also called the *best rank-one approximation ratio* of the given tensor space [Qi11]. The ratio is of theoretical and practical relevance, for example in problems of low-rank approximation and also in quantum entanglement. In quantum physics, rank-one tensors describe separable, and therefore unentangled, states and the distance to such separable states is a geometric measure of entanglement [Shi95, WG03].

The Frobenius and spectral norm of a matrix, and therefore its distance to rank-one matrices, can be determined from the singular value decomposition. It follows directly that

$$\frac{\|A\|_\sigma}{\|A\|_{\mathsf{F}}} \geq \frac{1}{\sqrt{\min\{m,n\}}} \quad \text{and} \quad \min_{\operatorname{rank} B=1} \frac{\|A-B\|_{\mathsf{F}}}{\|A\|_{\mathsf{F}}} \leq \sqrt{1-\frac{1}{\min\{m,n\}}}$$

for $A \in \mathbb{R}^{m \times n}$ and equality is attained for matrices with identical singular values.

For tensors, the situation is less clear and is under current research; see e.g., [Qi11, KM15, LNSU18, LZ20, AKU20]. A trivial bound of the approximation ratio for a tensor $\mathbf{A} \in \bigotimes_{i=1}^{d} \mathbb{R}^{n_i}$ with $n_1 \leq n_2 \leq \ldots \leq n_d$ is

$$\frac{\|\mathbf{A}\|_\sigma}{\|\mathbf{A}\|_{\mathsf{F}}} \geq \frac{1}{\sqrt{\prod_{i=1}^{d-1} n_i}},$$

which is attained by approximating $\mathbf{A}$ with a rank-one tensor with the largest fiber $[a_{i_1,\ldots,i_d}]_{i_d=1}^{n_d}$ of $\mathbf{A}$ and zero otherwise.

The trivial bound is not always attained. One example are tensors in $\mathbb{R}^{3\times3\times3}$. Here the approximation ratio is [AKU20]

$$\min_{\mathbf{A}\in\mathbb{R}^{3\times3\times3}} \frac{\|\mathbf{A}\|_\sigma}{\|\mathbf{A}\|_{\mathsf{F}}} = \frac{1}{\sqrt{7}}.$$

For $n = 2, 4, 8$, the approximation ratio in $\mathcal{T} = \bigotimes_{i=1}^{d} \mathbb{R}^n$ is the trivial bound

$$\min_{\mathbf{A}\in\mathcal{T}} \frac{\|\mathbf{A}\|_\sigma}{\|\mathbf{A}\|_{\mathsf{F}}} = \frac{1}{\sqrt{n^{d-1}}}$$

and is attained only for *orthogonal tensors* up to scaling [LNSU18]. In the cases $n = 4$ and $n = 8$, orthogonal tensors of order at least three are never symmetric, which leads to the interesting conclusion, that for $\mathcal{S} = \operatorname{Sym}_d \mathbb{R}^n$

$$\min_{\mathbf{A}\in\mathcal{S}} \frac{\|\mathbf{A}\|_\sigma}{\|\mathbf{A}\|_{\mathsf{F}}} > \frac{1}{\sqrt{n^{d-1}}}.$$

Therefore, unlike for matrices, the approximation ratios for symmetric and general tensors do not always coincide.

Also unlike matrices, the approximation ratio depends on the field. In [CKP00] it was shown that

$$\min_{\mathbf{A}\in\mathbb{C}^{2\times2\times2}} \frac{\|\mathbf{A}\|_\sigma}{\|\mathbf{A}\|_{\mathsf{F}}} = \frac{2}{3},$$

which is larger than the approximation ratio for $\mathbb{R}^{2\times 2\times 2}$.

In this chapter, we want to start discussing the ratio not for a given tensor space, but for tensors of a given rank. In quantum physics, the tensor rank in the form of *Schmidt measure* is a discrete measure of entanglement, while the distance to rank-one tensors is a continuous one. From the perspective of quantum physics, the following can be interpreted as analyzing how far both measures can differ. For matrices, this ratio is

$$\min_{\operatorname{rank} A=r} \frac{\|A\|_\sigma}{\|A\|_{\mathsf{F}}} = \frac{1}{\sqrt{r}}$$

and is attained for matrices with r identical nonzero singular values. For tensors, the situation is again less clear. We have the following result for rank-two tensors.

Theorem 1.1. *Let $\mathcal{H}_i$ be real Hilbert spaces and $\mathbf{A} \in \bigotimes_{i=1}^d \mathcal{H}_i$ be a tensor of rank two. Then for $d \geq 3$*

$$\|\mathbf{A}\|_\sigma > \left(1 - \frac{1}{d}\right)^{\frac{d-1}{2}} \|\mathbf{A}\|_{\mathsf{F}}. \tag{1.3}$$

In particular, let $\mathbf{W} = \lim_{t\to 0} \frac{1}{t}\left(\bigotimes_{i=1}^d (u_i + tv_i) - \bigotimes_{i=1}^d u_i\right)$ for orthonormal vectors $u_i, v_i \in \mathcal{H}_i$. Then

$$\|\mathbf{W}\|_\sigma = \left(1 - \frac{1}{d}\right)^{\frac{d-1}{2}} \|\mathbf{W}\|_{\mathsf{F}},$$

i.e., the inequality (1.3) *is sharp.*

Proof. The inequality follows directly from the first observation in Section 1.1, Proposition 1.2, and Theorem 1.6 further below. The equality for $\mathbf{W}$ follows from Theorem 1.6 with a change of orthogonal bases. □

Together with (1.2) this implies

$$\min_{\operatorname{rank} \mathbf{B}=1} \|\mathbf{A} - \mathbf{B}\|_{\mathsf{F}} < \sqrt{1 - \left(1 - \frac{1}{d}\right)^{d-1}} \|\mathbf{A}\|_{\mathsf{F}}$$

for a rank-two tensor $\mathbf{A}$ of order d. Note that the bound is not attained for any rank-two tensor. This is because, unlike rank-two matrices, the set of tensors of rank two is not closed. We will show that equality is indeed only attained for tensors of *border rank two*, i.e., tensors that are limits of rank-two tensors, that are not of rank two themselves. Theorem 1.1 also implies the uniform bounds

$$\|\mathbf{A}\|_\sigma > \frac{1}{\sqrt{e}}\|\mathbf{A}\|_{\mathsf{F}} \quad \text{and} \quad \min_{\operatorname{rank} \mathbf{B}=1} \|\mathbf{A} - \mathbf{B}\|_{\mathsf{F}} < \sqrt{1 - \frac{1}{e}}\|\mathbf{A}\|_{\mathsf{F}}$$

for rank-two tensors $\mathbf{A}$ of any order since $(1 - 1/d)^{(d-1)/2} \searrow 1/\sqrt{e}$ as $d \to \infty$.

1.1 Reduction to binary forms

We start this section with the observation that we can restrict to the case of rank-two tensors in $\bigotimes_{i=1}^{d}\mathbb{R}^2$. Indeed, the tensor $\mathbf{A}=\bigotimes_{i=1}^{d}u_i+\bigotimes_{i=1}^{d}v_i$ lies in the tensor space $\bigotimes_{i=1}^{d}\operatorname{span}\{u_i,v_i\}$ and $\operatorname{span}\{u_i,v_i\}$ is isomorphic to $\mathbb{R}^2$ since u_i,v_i come from a real Hilbert space.

Our next proposition lets us reduce our study further to symmetric tensor spaces. Together with the previous observation, we can then reduce to $\operatorname{Sym}_d\mathbb{R}^2$ which can be identified as the space of real homogeneous polynomials of degree d in two variables, i.e., binary forms.

Proposition 1.2. *Let $\mathbf{A}\in\bigotimes_{i=1}^{d}\mathcal{H}_i$ be a real rank-two tensor. Then there is symmetric rank-two tensor $\mathbf{A}_{\mathsf{S}}\in\operatorname{Sym}_d\mathbb{R}^2$ with $\|\mathbf{A}\|_{\mathsf{F}}=\|\mathbf{A}_{\mathsf{S}}\|_{\mathsf{F}}$ and $\|\mathbf{A}\|_{\sigma}\geq\|\mathbf{A}_{\mathsf{S}}\|_{\sigma}$.*

For the proof, we require the following two lemmas on the behavior of successively taking geometric means of positive real numbers, and the relation of Frobenius and spectral norm of two certain 2×2 matrices.

Lemma 1.3. *Let $x,y\geq 0$ and define the sequence*

$$x_0=x,\quad x_1=\left(x^{d-1}y\right)^{\frac{1}{d}},\quad x_{k+2}=\left(x_{k+1}^{d-1}x_k\right)^{\frac{1}{d}}.$$

Then $\lim_{k\to\infty}x_k=\left(x^d y\right)^{\frac{1}{d+1}}$.

Proof. We may assume $x,y>0$, otherwise the result follows immediately. We show via induction that

$$x_k=\left(x^{d^{k+1}+(-1)^k}y^{d^k+(-1)^{k-1}}\right)^{\frac{1}{d^k(d+1)}}. \tag{1.4}$$

The cases $k=0$ and $k=1$ follow directly. Now let (1.4) be true for $1,\dots,k+1$. Then

$$\begin{aligned}x_{k+2}=\left(x_{k+1}^{d-1}x_k\right)^{\frac{1}{d}}&=x^{\left(\frac{(d-1)\left(d^{k+2}+(-1)^{k+1}\right)}{d^{k+2}(d+1)}+\frac{d^{k+1}+(-1)^k}{d^{k+1}(d+1)}\right)}y^{\left(\frac{(d-1)\left(d^{k+1}+(-1)^k\right)}{d^{k+1}(d+1)}+\frac{d^k+(-1)^{k-1}}{d^k(d+1)}\right)}\\&=\left(x^{d^{k+3}+(-1)^{k+2}}y^{d^{k+2}+(-1)^{k+1}}\right)^{\frac{1}{d^{k+2}(d+1)}},\end{aligned}$$

proving (1.4). Taking the limit $k\to\infty$ gives the result. □

Lemma 1.4. *Let $0\leq x,y\leq 1$ and define the matrices*

$$A=\begin{pmatrix}a+bxy & bx\sqrt{1-y^2}\\ by\sqrt{1-x^2} & b\sqrt{(1-x^2)(1-y^2)}\end{pmatrix}\quad\textit{and}\quad B=\begin{pmatrix}a+bxy & b\sqrt{xy-x^2y^2}\\ b\sqrt{xy-x^2y^2} & b(1-xy)\end{pmatrix}.$$

Then $\|A\|_{\mathsf{F}}=\|B\|_{\mathsf{F}}$ and $\|A\|_{\sigma}\geq\|B\|_{\sigma}$.

Proof. A direct calculation shows that $\|A\|_{\mathsf{F}} = \|B\|_{\mathsf{F}}$. The singular values of 2×2 matrices are given by $\sigma_{1,2}^2 = F^2/2 \pm \sqrt{F^4/4 - |D|^2}$, where F is the Frobenius norm and D is the determinant of the matrix. We have

$$|\det A|^2 = a^2b^2(1 - x^2 - y^2 + x^2y^2) \quad \text{and} \quad |\det B|^2 = a^2b^2(1 - 2xy + x^2y^2).$$

Since $2xy \leq x^2 + y^2$ implies $|\det A|^2 \leq |\det B|^2$, the largest singular value of A, which equals its spectral norm, is larger or equal to the largest singular value of B. □

Proof of Proposition 1.2. Let $\mathbf{A} = \alpha \bigotimes_{i=1}^d u_i + \beta \bigotimes_{i=1}^d v_i$ with $\|u_i\| = 1 = \|v_i\|$. Let

$$\mathbf{U} = \bigotimes_{i=1}^d u_i \quad \text{and} \quad \mathbf{V} = \bigotimes_{i=1}^d v_i.$$

Then $\|\mathbf{A}\|_{\mathsf{F}}^2 = \alpha^2\|\mathbf{U}\|_{\mathsf{F}}^2 + 2\alpha\beta\langle\mathbf{U},\mathbf{V}\rangle_{\mathsf{F}} + \beta^2\|\mathbf{V}\|_{\mathsf{F}}^2$. We may assume that $u_i, v_i \in \mathbb{R}^2$ and after an orthogonal change of bases and possibly changing sign of β, we may also assume that $u_i = e_1$ and $v_i = x_ie_1 + \sqrt{1-x_i^2}e_2$ with $0 \leq x_i \leq 1$. We will show that replacing k factors of $\mathbf{V}$, i.e., the vectors $v_{i_1}, \dots, v_{i_k}$, with $v = xe_1 + \sqrt{1-x^2}e_2$, where $x = \prod_{j=1}^k x_{i_j}^{1/k}$ is the geometric mean, leads to a tensor with the same Frobenius but smaller spectral norm. We may assume that we replace the first k vectors $v_1, \dots, v_k$. This tensor reads $\mathbf{A}_k = \alpha\mathbf{U} + \beta\mathbf{V}_k$ with $\mathbf{V}_k = \bigotimes_{i=1}^k v \otimes \bigotimes_{i=k+1}^d v_i$ and since

$$\langle\mathbf{U},\mathbf{V}_k\rangle_{\mathsf{F}} = \prod_{i=1}^k\langle u_i, v\rangle \prod_{i=k+1}^d \langle u_i, v_i\rangle = x^k \prod_{i=k+1}^d x_i = \prod_{i=1}^d x_i = \prod_{i=1}^d\langle u_i, v_i\rangle = \langle\mathbf{U},\mathbf{V}\rangle_{\mathsf{F}},$$

the Frobenius norm of $\mathbf{A}$ and $\mathbf{A}_k$ coincide. We show inductively that the spectral norm does not increase with k, i.e., $\|\mathbf{A}_{k+1}\|_\sigma \leq \|\mathbf{A}_k\|_\sigma \leq \|\mathbf{A}\|_\sigma$. For $k = d$, this gives a symmetric tensor with the desired properties. We start with $k = 2$. Let $w_1, \dots, w_d$ be the maximizers in

$$\max_{\|w_1\|=\ldots=\|w_d\|=1} \langle\mathbf{A}_2, \otimes_{i=1}^d w_i\rangle_{\mathsf{F}} = \|\mathbf{A}_2\|_\sigma.$$

Let $a = \alpha\prod_{i=3}^d\langle u_i, w_i\rangle$, $b = \beta\prod_{i=3}^d\langle v_i, w_i\rangle$, and define the matrices

$$A = ae_1e_1^{\mathsf{T}} + bv_1v_2^{\mathsf{T}} \quad \text{and} \quad B = ae_1e_1^{\mathsf{T}} + bvv^{\mathsf{T}}.$$

These matrices represent the bilinear forms in $\tilde{w}_1$ and $\tilde{w}_2$

$$\tilde{w}_1^{\mathsf{T}}A\tilde{w}_2 = \langle\mathbf{A}, \tilde{w}_1 \otimes \tilde{w}_2 \otimes \bigotimes_{i=3}^d w_i\rangle_{\mathsf{F}} \quad \text{and} \quad \tilde{w}_1^{\mathsf{T}}B\tilde{w}_2 = \langle\mathbf{A}_2, \tilde{w}_1 \otimes \tilde{w}_2 \otimes \bigotimes_{i=3}^d w_i\rangle_{\mathsf{F}}.$$

Therefore,

$$\|A\|_\sigma = \max_{\|\tilde{w}_1\|=\|\tilde{w}_2\|=1} \tilde{w}_1^{\mathsf{T}}A\tilde{w}_2 = \max_{\|\tilde{w}_1\|=\|\tilde{w}_2\|=1} \langle\mathbf{A}, \tilde{w}_1 \otimes \tilde{w}_2 \otimes \bigotimes_{i=3}^d w_i\rangle_{\mathsf{F}} \leq \|\mathbf{A}\|_\sigma$$

and

$$\|B\|_\sigma = \max_{\|\tilde{w}_1\|=\|\tilde{w}_2\|=1} \tilde{w}_1^\mathsf{T} B \tilde{w}_2 = \max_{\|\tilde{w}_1\|=\|\tilde{w}_2\|=1} \langle \mathbf{A}_2, \tilde{w}_1 \otimes \tilde{w}_2 \otimes \bigotimes_{i=3}^{d} w_i \rangle_\mathsf{F} = \|\mathbf{A}_2\|_\sigma.$$

Lemma 1.4 implies $\|B\|_\sigma \leq \|A\|_\sigma$ and therefore $\|\mathbf{A}_2\|_\sigma \leq \|\mathbf{A}\|_\sigma$.

Now assume that replacing k factors of $\mathbf{V}$ in this manner results in a tensor $\mathbf{A}_k$ with a smaller or equal spectral norm. We now replace the first k factors of $\mathbf{V}$ and the second to $k+1$-st factor of $\mathbf{V}$ successively, i.e., we first replace $\mathbf{V}$ with the rank-one tensor $\widetilde{\mathbf{V}}_0 = \bigotimes_{i=1}^k \tilde{v}_0 \otimes \bigotimes_{i=k+1}^d v_i$ and then with the rank-one tensor $\widetilde{\mathbf{V}}_1 = \tilde{v}_0 \otimes \bigotimes_{i=2}^{k+1} \tilde{v}_1 \otimes \bigotimes_{i=k+2}^d v_i$ and repeating. This leads to a sequence $\mathbf{B}_\ell = \alpha \mathbf{U} + \beta \widetilde{\mathbf{V}}_\ell$ with nonincreasing spectral norm. The vectors read $\tilde{v}_\ell = y_\ell e_1 + \sqrt{1 - y_\ell^2} e_2$ with

$$y_0 = \prod_{i=1}^k x_i^{1/k}, \quad y_1 = \left(x^{k-1} x_{k+1}\right)^{\frac{1}{k}}, \quad y_{\ell+2} = \left(y_{\ell+1}^{k-1} y_\ell\right)^{\frac{1}{k}}.$$

By Lemma 1.3, this sequence converges to

$$\left(\left(\prod_{i=1}^k x_i^{1/k}\right)^k x_{k+1}\right)^{1/(k+1)} = \prod_{i=1}^{k+1} x_i^{1/(k+1)},$$

i.e., the sequence of tensors $\mathbf{B}_\ell$ converges to $\mathbf{A}_{k+1}$. Since the spectral norm of the sequence $\mathbf{B}_\ell$ is nonincreasing, also $\|\mathbf{A}_{k+1}\|_\sigma \leq \|\mathbf{A}_k\|_\sigma \leq \|\mathbf{A}\|_\sigma$, concluding the proof. □

Proposition 1.2 allows us to restrict to symmetric tensors. Note that while for symmetric tensors the notions of rank and symmetric rank are not the same in general [Shi18], they coincide for rank-two tensors, see, e.g., [ZHQ16]. We can further restrict our search for rank-two tensors with the minimal ratio of spectral and Frobenius norm with the following observation.

Lemma 1.5. *Let $\mathbf{A} = a\mathbf{U} + b\mathbf{V}$ with rank-one tensor $\mathbf{U}$ and $\mathbf{V}$. Assume further that $\|\mathbf{U}\|_\mathsf{F} = \|\mathbf{V}\|_\mathsf{F} = 1$, $\langle \mathbf{U}, \mathbf{V} \rangle_\mathsf{F} \geq 0$ and $a, b > 0$. Then*

$$\|\mathbf{A}\|_\sigma \geq \frac{1}{\sqrt{2}} \|\mathbf{A}\|_\mathsf{F}.$$

Proof. Without loss of generality, assume $|a| \geq |b|$. Then

$$\|\mathbf{A}\|_\sigma^2 \geq \langle \mathbf{A}, \mathbf{U} \rangle_\mathsf{F}^2 \geq a^2 + 2ab\langle \mathbf{V}, \mathbf{U} \rangle_\mathsf{F}$$

and

$$\|\mathbf{A}\|_\mathsf{F}^2 = a^2 + 2ab\langle \mathbf{V}, \mathbf{U} \rangle_\mathsf{F} + b^2 \leq 2a^2 + 4ab\langle \mathbf{V}, \mathbf{U} \rangle_\mathsf{F}$$

since $0 \leq \langle \mathbf{V}, \mathbf{U} \rangle_\mathsf{F} \leq 1$ and $a \geq b > 0$. The claim follows. □

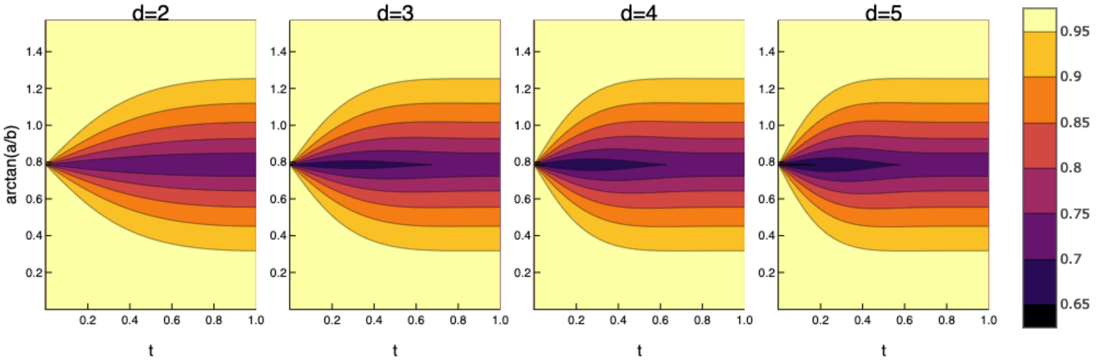


Figure 1.1: The ratio $\|p\|_\infty/\|p\|_{\mathsf{B}}$ norm of $p(x,y) = a(x+ty)^d - b(x-ty)^d$.

We can therefore restrict our search to symmetric tensors $\mathbf{A} = au^d - bv^d$ in $\operatorname{Sym}_d \mathbb{R}^2$ with $a, b > 0$, $\|u\| = \|v\| = 1$, and $\langle u, v\rangle \geq 0$. Due to a result by Banach [Ban38], the definition of the spectral norm (1.1) of a symmetric tensor simplifies to

$$\|\mathbf{A}\|_\sigma = \max_{\|w\|=1} \left|\langle \mathbf{A}, w^d\rangle_{\mathsf{F}}\right|.$$

It will be convenient to identify symmetric tensors with the associated polynomial $p_{\mathbf{A}}(w) = \langle \mathbf{A}, w^d\rangle_{\mathsf{F}}$. The spectral norm of the tensor $\mathbf{A}$ corresponds to the *uniform norm* of polynomials on the sphere

$$\|p\|_\infty \coloneqq \max_{\|w\|=1} |p(w)|.$$

If w is a maximizer of $|p_{\mathbf{A}}(w)|$ on the sphere, then $p_{\mathbf{A}}(w)w^d$ is a best rank-one approximations of $\mathbf{A}$ in Frobenius norm. Similarly, if $\pm w^d$ is a best rank-one approximation, then w maximizes $1/\|w\|^d |p_{\mathbf{A}}(w)|$.

The Frobenius norm corresponds to the *Bombieri norm*

$$\|p\|_{\mathsf{B}} \coloneqq \sqrt{\sum_{|\alpha|=1} \binom{d}{\alpha}^{-1} |c_\alpha|^2},$$

where $p(w) = \sum_{|\alpha|=1} c_\alpha w^\alpha$ with the multi index notation $\alpha = (\alpha_1, \dots, \alpha_n) \in \{0, \dots, d\}^n$, $w^\alpha = w_1^{\alpha_1} \dots w_d^{\alpha_d}$, $|\alpha| = \alpha_1 + \dots + \alpha_n$, and $\binom{d}{\alpha} = \frac{d!}{\alpha_1! \dots \alpha_d!}$.

Since all norms are invariant to orthogonal change of basis, the ratio of spectral and Frobenius norm of $au^d - bv^d$ depends only on the angle between the vectors u and v, and the ratio of a and b. We can restrict to the case $a, b > 0$, $\|u\| = \|v\| = 1$, and $\langle u, v\rangle \geq 0$ and the ratio is determined by the quantities $\arctan(a/b) \in (0, \pi/2)$ and $\langle u, v\rangle \in [0, 1)$. These quantities are attained for $a = \cos\varphi$, $b = \sin\varphi$, $u = (e_1 + te_2)/\sqrt{1+t^2}$, and

$v = (e_1 - te_2)/\sqrt{1+t^2}$ with $t \in (0,1]$ and $\varphi \in (0, \pi/2)$. Since scaling has no effect on the ratio, we can also use $u = e_1 + te_2$ $v = e_1 - te_2$. The corresponding polynomial is $p(x,y) = a(x+ty)^d - b(x-ty)^d$. We computed the ratio of the uniform and Bombieri norm in Figure 1.1 numerically for $d = 2, \dots, 5$. Figure 1.1 indicates that the smallest ratio is attained for $a = b$ and $t \to 0$.

1.2 Symmetric case

In this section, we prove a version of Theorem 1.1 for symmetric tensors. It is convenient to introduce notation for the symmetric part of rank-one tensors. We define

$$u_1 u_2 \dots u_d \coloneqq \frac{1}{d!} \sum_{\sigma \in \mathfrak{S}_d} u_{\sigma 1} \otimes u_{\sigma 2} \otimes \dots \otimes u_{\sigma_d},$$

where $\mathfrak{S}_d$ is the permutation group of d elements. This is consistent with the notation $u^d = \otimes_{i=1}^d u$ for symmetric rank-one tensors and it resembles the product of polynomials.

Theorem 1.6. *Let $\mathcal{H}$ be a real Hilbert space and $\mathbf{A} \in \operatorname{Sym}_d \mathcal{H}$ be a tensor of rank two. Then for $d \geq 3$*

$$\|\mathbf{A}\|_\sigma > \left(1 - \frac{1}{d}\right)^{\frac{d-1}{2}} \|\mathbf{A}\|_{\mathsf{F}}. \tag{1.5}$$

In particular, let $\mathbf{W} = du^{d-1}v = \lim_{t\to 0} \frac{1}{t}\left((u+tv)^d - u^d\right)$ for orthonormal vectors $u, v \in \mathcal{H}$. Then

$$\|\mathbf{W}\|_\sigma = \left(1 - \frac{1}{d}\right)^{\frac{d-1}{2}} \|\mathbf{W}\|_{\mathsf{F}},$$

i.e., the inequality (1.5) is sharp.

Proof. Let $\mathbf{A} = \mathbf{U} + \mathbf{V}$ where $\mathbf{U}$ and $\mathbf{V}$ are symmetric rank-one tensors. If $\langle \mathbf{U}, \mathbf{V}\rangle_{\mathsf{F}} < 0$ and $\|\mathbf{U}\|_{\mathsf{F}} = \|\mathbf{V}\|_{\mathsf{F}}$ or $\langle \mathbf{U}, \mathbf{V}\rangle_{\mathsf{F}} \geq 0$, then Proposition 1.12 further below and Lemma 1.5, respectively, imply (1.5). In the following, Proposition 1.8 and Proposition 1.9 imply that the minimal ratio cannot be attained for $\langle \mathbf{U}, \mathbf{V}\rangle_{\mathsf{F}} < 0$ and $\|\mathbf{U}\|_{\mathsf{F}} \neq \|\mathbf{V}\|_{\mathsf{F}}$. The infimum ratio of norms therefore has to be attained on the boundary of rank-two tensors. Finally, Proposition 1.7 and Proposition 1.14 show that the inequality is correct and sharp. □

1.2.1 Maximal distance

We first prove that inequality (1.5) is sharp. For this, it is helpful to switch to the perspective of polynomials.

Proposition 1.7. *The polynomial corresponding to* $\mathbf{W} = \lim_{t\to 0} \frac{1}{t}\left((e_1 + te_2)^d - e_1^d\right)$ *is* $p(x, y) = dx^{d-1}y$. *Furthermore,*

$$\|\mathbf{W}\|_{\mathsf{F}} = \|p\|_{\mathsf{B}} = \sqrt{d} \quad \textit{and} \quad \|\mathbf{W}\|_\sigma = \|p\|_\infty = (d-1)^{\frac{d-1}{2}} d^{\frac{d}{2}},$$

i.e., $\|\mathbf{W}\|_\sigma = (1 - 1/d)^{(d-1)/2}\|\mathbf{W}\|_{\mathsf{F}}$.

Proof. Note that $\mathbf{W} = \frac{d}{dt}(e_1 + te_2)^d|_{t=0}$. The polynomial corresponding to $(e_1 + te_2)^d$ is $p_t(x, y) = (x + ty)^d$ and $\frac{\partial}{\partial t} p_t(x, y)|_{t=0} = dx^{d-1}y = p(x, y)$. The stated Bombieri norm of p follows from the definition. The uniform norm is

$$\max dx^{d-1}y \quad \text{such that } x^2 + y^2 = 1.$$

The KKT conditions lead to the necessary optimality condition $(d-1)x^{d-2}y^2 = x^d$. We find that $x = \sqrt{1 - 1/d}$ and $y = 1/\sqrt{d}$ is a maximizer, and the stated uniform norm of p follows. □

1.2.2 Optimality conditions

For proving Theorem 1.6 we will determine the critical points of the optimization problem

$$\inf_{\substack{a,b\in\mathbb{R} \\ \|u\|=\|v\|=1}} F(a, b, u, v) = \frac{\|au^d - bv^d\|_\sigma^2}{\|au^d - bv^d\|_{\mathsf{F}}^2}. \tag{1.6}$$

The target function in (1.6) can be written as a composition

$$F(a, b, u, v) = G(\varphi(a, b, u, v))$$

where

$$G\colon \operatorname{Sym}_d \mathbb{R}^2 \to \mathbb{R}, \quad G(\mathbf{A}) = \frac{\|\mathbf{A}\|_\sigma^2}{\|\mathbf{A}\|_{\mathsf{F}}^2},$$

and

$$\varphi\colon \mathbb{R} \times \mathbb{R} \times \mathbb{R}^2 \times \mathbb{R}^2 \to \operatorname{Sym}_d \mathbb{R}^2, \quad \varphi(a, b, u, v) = au^d - bv^d.$$

While φ is smooth, the map G is not differentiable at all points. However, G is the quotient of the smooth function $\mathbf{A} \mapsto \|\mathbf{A}\|_{\mathsf{F}}^2$ and the convex function $\mathbf{A} \mapsto \|\mathbf{A}\|_\sigma^2$. Therefore, the rules for generalized gradients of regular functions are applicable; see [Cla90, Section 2.3]. It follows that the subdifferential of G in a point $\mathbf{A}$ can be computed using a quotient rule, which yields

$$\partial G(\mathbf{A}) = \frac{2\|\mathbf{A}\|_\sigma}{\|\mathbf{A}\|_{\mathsf{F}}^4}\left(\partial(\|\mathbf{A}\|_\sigma)\|\mathbf{A}\|_{\mathsf{F}}^2 - \mathbf{A}\|\mathbf{A}\|_\sigma\right).$$

Here $\partial(\|A\|_\sigma)$ denotes the subdifferential of the spectral norm in A. The derivative of φ equals

$$\varphi'(a, b, u, v)[\delta a, \delta b, \delta u, \delta v] = u^{d-1}(ad \cdot \delta u + \delta a \cdot u) - v^{d-1}(db \cdot \delta v + \delta b \cdot v), \tag{1.7}$$

which leads to

$$\begin{aligned}&\partial F(a,b,u,v)[\delta a,\delta b,\delta u,\delta v]\\ &\quad=\frac{2\|\mathbf{A}\|_\sigma}{\|\mathbf{A}\|_{\mathsf{F}}^4}\langle\partial(\|\mathbf{A}\|_\sigma)\|\mathbf{A}\|_{\mathsf{F}}^2-\mathbf{A}\|\mathbf{A}\|_\sigma, u^{d-1}(ad\cdot\delta u+\delta a\cdot u)-v^{d-1}(db\cdot\delta v+\delta b\cdot v)\rangle_{\mathsf{F}},\end{aligned} \tag{1.8}$$

where $\mathbf{A}=\varphi(a,b,u,v)$ for brevity. The subdifferential of the spectral norm can be characterized as the convex hull of normalized best rank-one approximations, i.e.,

$$\partial(\|\mathbf{A}\|_\sigma)=\operatorname{conv}\underset{\substack{\mathbf{B}\in\operatorname{Sym}_d\mathbb{R}^2\\ \operatorname{rank}\mathbf{B}=1\\ \|\mathbf{B}\|_{\mathsf{F}}=1}}{\arg\max}\langle\mathbf{A},\mathbf{B}\rangle_{\mathsf{F}}, \tag{1.9}$$

see [Cla75, Theorem 2.1] in general, and [AKU20, Section 2.3] in particular. In words, $\partial(\|\mathbf{A}\|_\sigma)$ equals the convex hull of the normalized symmetric best rank-one approximations of $\mathbf{A}$.

The first-order optimality condition $0\in\partial F(a,b,u,v)$ (see, e.g., [Cla90, Proposition 2.3.2]) for problem (1.6) together with (1.8) and (1.9) imply that

$$\lambda(au^d-bv^d)\in P_{u,v}\operatorname{conv}\underset{\substack{\mathbf{B}\in\operatorname{Sym}_d\mathbb{R}^2\\ \operatorname{rank}\mathbf{B}=1\\ \|\mathbf{B}\|_{\mathsf{F}}=1}}{\arg\max}\langle au^d-bv^d,\mathbf{B}\rangle_{\mathsf{F}}, \tag{1.10}$$

where $P_{u,v}$ is the orthogonal projection onto the image of the linear map $\phi'(a,b,u,v)$ in (1.7), i.e., the linear subspace $\{u^{d-1}\delta u+v^{d-1}\delta v\colon \delta u,\delta v\in\mathbb{R}^2\}$ of $\operatorname{Sym}_d\mathbb{R}^2$, and $\lambda\in\mathbb{R}$ is a Lagrange multiplier.

We now show that the optimality condition (1.10) cannot hold for a tensor au^d-bv^d admitting a unique best symmetric rank-one approximation. This is an interesting analogy to the fact that matrices achieving a minimal ratio of spectral and Frobenius norm have equal singular values, and therefore have no unique best rank-one approximations.

Proposition 1.8. *Let $\mathbf{A}=au^d-bv^d$ have rank two. If $\mathbf{A}$ has a unique best symmetric rank-one approximation, then $\mathbf{A}$ is not a critical point of the optimization problem* (1.6).

Proof. Let $\pm w^d$ be the best rank-one approximation of $\mathbf{A}$. Note that $\langle\mathbf{A},w^d\rangle\neq 0$. Since $\mathbf{A}=\varphi'(a,b,u,v)[1,1,0,0]$ in (1.7), this implies $P_{u,v}w^d\neq 0$. A direct computation shows $P_{u,v}w^d=\alpha u^{d-1}w+\beta v^{d-1}w$ for some $\alpha,\beta\in\mathbb{R}$. However, since u and v are linearly independent, we have the decomposition

$$\{u^{d-1}\delta u+v^{d-1}\delta v\colon \delta u,\delta v\in\mathbb{R}^n\}=\{u^{d-1}\delta u\colon \delta u\in\mathbb{R}^n\}\oplus\{v^{d-1}\delta v\colon \delta v\in\mathbb{R}^n\}$$

into two complementary subspaces. Therefore, (1.10) is only possible if w is both a multiple of u and v, which contradicts the linear independence of u and v. □

1.2.3 Unique symmetric best rank-one approximations

We now present a class of symmetric rank-two tensors admitting unique best symmetric rank-one approximations. By the result of Proposition 1.8, these can then be excluded from the further discussion on the minimal norm ratio.

Proposition 1.9. *Let* $\mathbf{A} = au^d - bv^d$ *with* $u \neq v$*,* $\|u\| = \|v\| = 1$*,* $\langle u, v\rangle > 0$ *and* $a > b > 0$*. Then* $\mathbf{A}$ *has exactly one best symmetric rank-one approximation.*

For the proof, we require auxiliary results. One is the following fact about polynomials.

Lemma 1.10. *Let* $\alpha, \gamma > 0$*,* $\beta \geq 0$*, and* $d \geq 2$*. The equation* $x = \gamma(x-\alpha)(x+\beta)^{d-1}$ *has two real solutions for* x *if* d *is even, and three real solutions if* d *is odd.*

Proof. Let $p(x) = \gamma(x-\alpha)(x+\beta)^{d-1} - x$. Then by the intermediate value theorem, p must have at least two real zeros, namely one in the interval $[-\beta, 0]$ and another one in the interval (α, ∞). On the other hand,

$$p'(x) = \gamma d(x+\beta)^{d-2}\left(x - \frac{(d-1)\alpha - \beta}{d}\right) - 1,$$

has at most two sign changes, one at a value larger than $((d-1)\alpha - \beta)/d$ and another at one at a value smaller than $-\beta$ if d is odd. Therefore, p has at most three real zeros. The statement follows from the fact that the number of real zeros of a polynomial with real coefficients has the same parity as its degree. □

The second lemma narrows the possible locations of maximizers of the homogeneous form $|p_{\mathbf{A}}|$.

Lemma 1.11. *Under the assumptions of Proposition 1.9, let* w *be a maximizer of* $|p_{\mathbf{A}}(w)| = \left|\langle au^d - bv^d, w^d\rangle_{\mathsf{F}}\right|$ *subject to* $\|w\| = c > 0$*. Then* $|\langle u, w\rangle| \geq |\langle v, w\rangle|$*.*

Proof. Assume to the opposite that $|\langle u, w\rangle| < |\langle v, w\rangle|$ and without loss of generality $\langle v, w\rangle > 0$. Let Q be the symmetric orthogonal matrix mapping u to v and v to u (i.e. $Q = I - zz^{\mathsf{T}}$ with $z = (u+v)/\|u+v\|$), and let $\bar{w} = Qw$. Then $\langle u, w\rangle = \langle v, \bar{w}\rangle$ and $\langle v, w\rangle = \langle u, \bar{w}\rangle$. By assumption, we then have $\left|\langle au^d - bv^d, \bar{w}^d\rangle_{\mathsf{F}}\right| = \langle au^d - bv^d, \bar{w}^d\rangle_{\mathsf{F}}$. If $\left|\langle au^d - bv^d, w^d\rangle_{\mathsf{F}}\right| = \langle au^d - bv^d, w^d\rangle_{\mathsf{F}}$, this yields $\left|\langle au^d - bv^d, \bar{w}^d\rangle_{\mathsf{F}}\right| > \left|\langle au^d - bv^d, w^d\rangle_{\mathsf{F}}\right|$ (by using $(a+b)\langle v, w\rangle^d > (a+b)\langle u, w\rangle^d$) which contradicts the optimality of w. In the other case, $\left|\langle au^d - bv^d, w^d\rangle_{\mathsf{F}}\right| = -\langle au^d - bv^d, w^d\rangle_{\mathsf{F}}$, optimality implies

$$b(\langle u, w\rangle^d + \langle v, w\rangle^d) > a(\langle u, w\rangle^d + \langle v, w\rangle^d)$$

which contradicts $a > b$. □

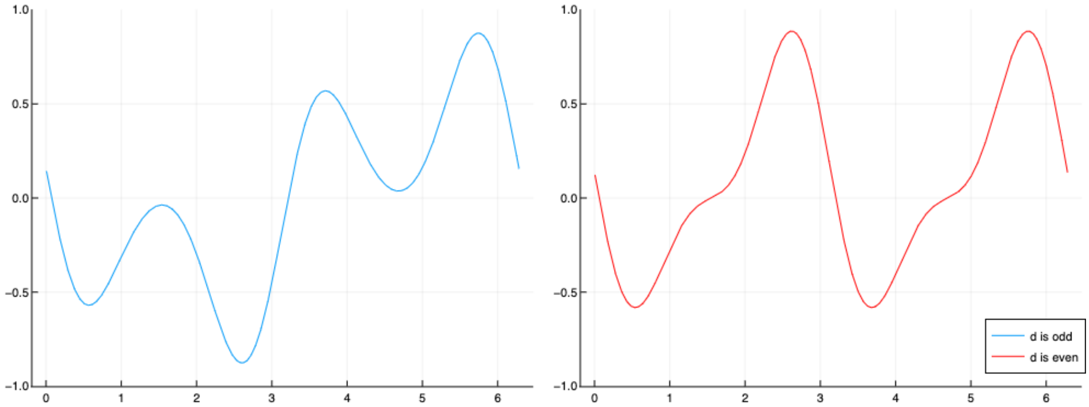


Figure 1.2: The values of the polynomial $p(x,y) = a(x+ty)^d - b(x-ty)^d$ with $a > b > 0$ on the circle $(x,y) = (\cos\varphi, \sin\varphi)$.

We are now in the position to prove Proposition 1.9. We use Lemma 1.10 to show that, depending on the parity of d, the associated polynomial $p_{\mathbf{A}}$ has either six or four critical points on the circle $x^2+y^2=1$. These critical points come in pairs (x,y) and $(-x,-y)$ and correspond to critical best rank-one approximations. Afterwards, we use Lemma 1.11 to show that only one of these pairs can correspond to the best rank-one approximation, which can be anticipated by Figure 1.2.

Proof of Proposition 1.9. We can assume that $\mathbf{A} \in \operatorname{Sym}_d \mathbb{R}^2$, so that $u, v \in \mathbb{R}^2$. Without loss of generality, since we can change coordinates, we can consider $a = 1$, $u = e_2$ and $\sqrt[d]{b}v = \alpha e_1 + \beta e_2$ with $\alpha, \beta > 0$ (since $\langle u, v\rangle > 0$), and $\alpha^2 + \beta^2 < 1$ (since $b < a = 1$). Let $\lambda^2(x^2+y^2) = 1$ and $\lambda > 0$, i.e., $\lambda(x,y)$ is a point on the unit circle. Then

$$p_{\mathbf{A}}(\lambda x, \lambda y) = \lambda^d[y^d - (\alpha x + \beta y)^d]. \tag{1.11}$$

Critical points on the circle are characterized by $x\frac{\partial}{\partial y}p_{\mathbf{A}} - y\frac{\partial}{\partial x}p_{\mathbf{A}} = 0$, which means

$$xy^{d-1} - (\beta x - \alpha y)(\alpha x + \beta y)^{d-1} = 0$$

independent of λ. Note that here $y = 0$ is not possible since both α and β are nonzero. Recall that a symmetric best rank-one approximation of $\mathbf{A}$ is given as $p_{\mathbf{A}}(w)w^d$, where w maximizes $|p_{\mathbf{A}}(w)|$ on the circle. Since $p_{\mathbf{A}}(-w) = (-1)^d p_{\mathbf{A}}(w)$, in order to prove the assertion it suffices to show that $|p_{\mathbf{A}}(\lambda x, \lambda y)|$ has exactly one maximizer $(\lambda x, \lambda y)$ with $y = 1$. The optimality condition at such a point reduces to

$$x = (\beta x - \alpha)(\alpha x + \beta)^{d-1}. \tag{1.12}$$

Hence, we only need to show that there is exactly one solution x of this equation corresponding to a global maximum of $|p_{\mathbf{A}}|$ on the unit circle.

If $y = 1$, then $p_{\mathbf{A}}$ in (1.11) has a zero at $x_0 = (1-\beta)/\alpha$. Then

$$x_0 = \frac{1-\beta}{\alpha} > \frac{\beta - \beta^2 - \alpha^2}{\alpha} = (\beta x_0 - \alpha)(\alpha x_0 + \beta)^{d-1}.$$

This shows that (1.12) has at least one solution $x^* > x_0$. We consider a solution x^* of this kind such that the corresponding unit vector $w = \lambda(x^* e_1 + e_2)$ is a local maximum of $|p_{\mathbf{A}}|$ on the unit circle. Taking the sign changes of $x - (\beta x - \alpha)(\alpha x + \beta)^{d-1}$ and the sign of $p_{\mathbf{A}}$ into account, x^* indeed corresponds to a local maximizer. We have

$$|\langle u, w\rangle| = \lambda < \frac{\lambda}{\sqrt[d]{b}} = \frac{\lambda}{\sqrt[d]{b}}(\alpha x_0 + \beta) < \lambda \frac{1}{\sqrt[d]{b}}(\alpha x^* + \beta) = |\langle v, w\rangle|.$$

By Lemma 1.11, w is not a global maximum of $|p_{\mathbf{A}}|$. If d is even, then by Lemma 1.10 equation (1.12) has exactly two solutions and therefore only one corresponds to a global maximum. If d is odd, then by the same lemma (1.12) has three solutions. Taking into account that $p_{\mathbf{A}}$ in (1.11) has only one zero for $y = 1$, one of these solutions corresponds to a local minimizer of $|p_{\mathbf{A}}|$. Hence, there is only one global maximizer. □

Proposition 1.8 and Proposition 1.9 show that the minimal ratio in (1.6) is not achieved for $a \neq b$.

1.2.4 Difference of two normalized rank-one tensors

In this section, we show by a direct calculation that $\|\mathbf{A}\|_\sigma / \|\mathbf{A}\|_{\mathsf{F}} > \|\mathbf{W}\|_\sigma / \|\mathbf{W}\|_{\mathsf{F}}$ when $\mathbf{A}$ is the difference of two symmetric rank-one tensors with the same norm and $\mathbf{W}$ is from Proposition 1.7. We will switch to the perspective of polynomials, and find two useful point evaluations to estimate the uniform norm.

Proposition 1.12. *Let $u \neq v$, $\|u\| = \|v\| \neq 0$, $\langle u, v\rangle \geq 0$ and $d \geq 3$. Then*

$$\frac{\|u^d - v^d\|_\sigma^2}{\|u^d - v^d\|_{\mathsf{F}}^2} > \left(1 - \frac{1}{d}\right)^{d-1}.$$

We require the following version of Jensen's inequality.

Lemma 1.13. *Let $f : [a, b] \to \mathbb{R}$ be convex and continuously differentiable. If $a + b = a' + b'$ and $a < a' < b' < b$, then*

$$\frac{1}{b-a}\int_a^b f(x)\,dx \geq \frac{1}{b'-a'}\int_{a'}^{b'} f(x)\,dx \geq f\left(\frac{a+b}{2}\right).$$

The inequalities are strict if f is strictly convex.

Proof. Without loss of generality let $a = -b$ and $a' = -b'$. A substitution results in

$$\begin{aligned}\frac{1}{b}\int_{-b}^{b} f(x)\,dx &= \frac{1}{b'}\int_{-b'}^{b'} f\left(\frac{b}{b'}x\right) - f(x) + f(x)\,dx \\ &= \frac{1}{b'}\int_{-b'}^{b'} f(x)\,dx + \frac{1}{b'}\int_{0}^{b'}\int_{x}^{\frac{bx}{b'}} f'(y) - f'(-y)\,dy\,dx \geq \frac{1}{b'}\int_{-b'}^{b'} f(x)\,dx,\end{aligned}$$

by monotonicity of the derivative of a convex function. This shows the first of the asserted inequalities. The second inequality is just Jensen's inequality, noting that $\frac{a+b}{2} = \frac{a'+b'}{2}$. If f is strictly convex, then f' is strictly monotone and the inequalities are strict. □

Proof of Proposition 1.12. We can assume that $u^d - v^d \in \operatorname{Sym}_d \mathbb{R}^2$ and identify $u^d - v^d$ with its polynomial. We only need to consider $p_t(x,y) = (x+ty)^d - (x-ty)^d$ with $t \in (0,1]$. The other cases follow after applying a rotation and scaling. Then

$$\|p_t\|_{\mathsf{B}}^2 = 2(1+t^2)^d - 2(1-t^2)^d =: g(t). \tag{1.13}$$

First, we apply the estimate

$$\|p_t\|_\infty \geq p_t\left(\frac{1}{\sqrt{1+t^2}}, \frac{t}{\sqrt{1+t^2}}\right) = \frac{(1+t^2)^d - (1-t^2)^d}{\sqrt{1+t^2}^d},$$

which yields

$$\frac{\|p_t\|_\infty^2}{\|p_t\|_{\mathsf{B}}^2} \geq \frac{(1+t^2)^d - (1-t^2)^d}{2(1+t^2)^d} = \frac{1}{2}\left(1 - \left(\frac{1-t^2}{1+t^2}\right)^d\right).$$

The right-hand side is monotonically increasing in the interval $(0,1]$. For $t = 1/\sqrt{d-1}$ it equals

$$\frac{1}{2}\left(1 - \left(\frac{d-2}{d}\right)^d\right) = \frac{d^d - (d-2)^d}{2d^d}.$$

This value is larger than $(1-1/d)^{d-1} = ((d-1)/d)^{d-1}$ since, using Lemma 1.13 with the function $f(t) = 2d\,t^{d-1}$, it holds that $\int_{d-2}^{d} f(t)\,dt = d^d - (d-2)^d > 2d(d-1)^{d-1}$ for $d \geq 3$. This shows

$$\frac{\|p_t\|_\infty^2}{\|p_t\|_{\mathsf{B}}^2} > \left(1 - \frac{1}{d}\right)^{d-1}$$

for all $t \in \left[1/\sqrt{d-1}, 1\right]$. It remains to verify this inequality for all $t \in \left(0, 1/\sqrt{d-1}\right)$, which is a little bit more involved. The starting point is another lower bound for the uniform norm, namely

$$\|p_t\|_\infty \geq p_t\left(\sqrt{1-\frac{1}{d}}, \frac{1}{\sqrt{d}}\right) = \frac{1}{\sqrt{d}^d}\left(\left(\sqrt{d-1}+t\right)^d - \left(\sqrt{d-1}-t\right)^d\right) =: h(t).$$

Note that $p_t(x,y)/t \to 2dx^{d-1}y = 2p_{\mathbf{W}}(x,y)$ as $t \to 0$ with $\mathbf{W}$ from Proposition 1.7 and therefore

$$\lim_{t\to 0} \frac{h(t)^2}{g(t)} = \frac{\|\mathbf{W}\|_\sigma^2}{\|\mathbf{W}\|_{\mathsf{F}}^2} = \left(1 - \frac{1}{d}\right)^{d-1}.$$

We now claim that

$$\frac{d}{dt}\frac{h(t)^2}{g(t)} > 0 \text{ for } t \in \left(0, \sqrt{\tfrac{1}{d-1}}\right)$$

which then proves the assertion. This claim is equivalent to the positivity of

$$\begin{aligned}
&\frac{\sqrt{d}^d}{4d}\big(2h'(t)g(t) - g'(t)h(t)\big) \\
&\quad= \left[\left(\sqrt{d-1}+t\right)^{d-1} + \left(\sqrt{d-1}-t\right)^{d-1}\right]\left[(1+t^2)^d - (1-t^2)^d\right] \\
&\qquad - t\left[\left(\sqrt{d-1}+t\right)^{d} - \left(\sqrt{d-1}-t\right)^{d}\right]\left[(1+t^2)^{d-1} + (1-t^2)^{d-1}\right].
\end{aligned}$$

Let

$$\begin{aligned}
a &:= \left(\sqrt{d-1}-t\right)(1-t^2) = \sqrt{d-1} - t - t^2\sqrt{d-1} + t^3,\\
b &:= \left(\sqrt{d-1}+t\right)(1+t^2) = \sqrt{d-1} + t + t^2\sqrt{d-1} + t^3,\\
a' &:= \left(\sqrt{d-1}-t\right)(1+t^2) = \sqrt{d-1} - t + t^2\sqrt{d-1} - t^3,\\
b' &:= \left(\sqrt{d-1}+t\right)(1-t^2) = \sqrt{d-1} + t - t^2\sqrt{d-1} - t^3.
\end{aligned}$$

Then elementary manipulations result in

$$\begin{aligned}
&\frac{\sqrt{d}^d}{4d}\big(2h'(t)g(t) - g'(t)h(t)\big) \\
&\quad= \left(b^{d-1} - a^{d-1}\right)\left(1 - t\sqrt{d-1}\right) - \left((b')^{d-1} - (a')^{d-1}\right)\left(1 + t\sqrt{d-1}\right). \qquad (1.14)
\end{aligned}$$

Note that for $t \in \left(0, 1/\sqrt{d-1}\right)$ we have $b > b' > a' > a$ and

$$b - a = 2t\left(1 + t\sqrt{d-1}\right), \quad b' - a' = 2t\left(1 - t\sqrt{d-1}\right).$$

Therefore, with $f(t) = (d-1)t^{d-2}$ we can rewrite (1.14) as

$$\frac{1}{4d}\sqrt{d}^d\big(2h'(t)g(t) - g'(t)h(t)\big) = \frac{1}{2t}\left[(b'-a')\int_a^b f(x)\,dx - (b-a)\int_{a'}^{b'} f(x)\,dx\right].$$

Moreover,

$$\frac{a+b}{2} = \sqrt{d-1} + 2t^3 > \sqrt{d-1} - 2t^3 = \frac{a'+b'}{2},$$

and therefore $a'' := \frac{a+b-(b'-a')}{2} > a' > a$ and $b > b'' := \frac{a+b+(b'-a')}{2} > b'$. Since $a'' + b'' = a + b$ and $a'' - b'' = a' - b'$, Lemma 1.13 yields

$$(b'-a') \int_a^b f(x)\,dx \geq (b-a) \int_{a''}^{b''} f(x)\,dx > (b-a) \int_{a'}^{b'} f(x)\,dx,$$

where the second inequality follows from the monotonicity of f. This shows that (1.14) is positive. □

1.2.5 Behaviour on the boundary

Finally, we consider tensors lying on the boundary of the set of symmetric rank-two tensors. We will again turn to the perspective of the respective polynomials and find two useful point evaluations to estimate the uniform norm.

Proposition 1.14. *Let* $\mathbf{A}$ *be a limit of symmetric rank-two tensors and* $\operatorname{rank} \mathbf{A} > 2$. *Then* $\|\mathbf{A}\|_\sigma^2 \geq \left(1 - \frac{1}{d}\right)^{d-1} \|\mathbf{A}\|_{\mathsf{F}}^2$ *and equality is attained if and only if* $\mathbf{A} = u^{d-1}v$ *for some orthogonal* u *and* v, *that is, for tensors arising from scaling and orthogonal transformations of the tensor* $\mathbf{W}$ *from Proposition 1.7.*

The boundary of rank-two tensors is well studied. We require the following well-known parametrization; see, e.g., [BL14]. We offer a self-contained proof for completeness.

Lemma 1.15. *Let* $\mathbf{A}$ *be a limit of symmetric rank-two tensors and* $\operatorname{rank} \mathbf{A} > 2$. *Then* $\mathbf{A}$ *is of the form*

$$\mathbf{A} = au^d + bdu^{d-1}v$$

with $\langle u, v\rangle = 0$ *and* $\|u\| = \|v\| = 1$.

Proof. Let $\mathbf{A}_n = u_n^d \pm v_n^d$ with $\lim_{n\to\infty} \mathbf{A}_n = \mathbf{A}$ or $\lim_{n\to\infty} \mathbf{A}_n = -\mathbf{A}$. It is not difficult to see that u_n and v_n must be unbounded since otherwise there is a subsequence of $\mathbf{A}_n$ converging to a tensor of rank at most two, contradicting $\operatorname{rank} \mathbf{A} > 2$. We write $v_n = s_n u_n + t_n w_n$ with $\|w_n\| = 1$ and $\langle u_n, w_n\rangle = 0$. Then

$$\mathbf{A}_n = (1 \pm s_n^d) u_n^d \pm \sum_{k=1}^{d} \binom{d}{k} s_n^{d-k} t_n^k u_n^{d-k} w_n^k,$$

and it can be checked that all terms are pairwise orthogonal. Hence, since $\mathbf{A}_n$ converges, all terms must be bounded and by passing to a subsequence we can assume that all of them converge. Due to $\|u_n\| \to \infty$ we have $1 \pm s_n^d \to 0$ for the first term, which implies that the sequence s_n is bounded. Therefore, considering the term $k = 1$, the sequence $t_n \|u_n\|^{d-1}$ is bounded which automatically implies $t_n^k \|u_n\|^{d-k} \to 0$ for all $k > 1$. We conclude that

$$\lim_{n\to\infty} \mathbf{A}_n = \lim_{n\to\infty} (1 \pm s_n^d) u_n^d + \lim_{n\to\infty} d s_n^{d-1} t_n u_n^{d-1} w_n = a u^d + b d u^{d-1} v$$

which proves the assertion. □

Proof of Proposition 1.14. Using Lemma 1.15, scaling and orthogonal transformations, we can assume $\mathbf{A} = a e_1^d + b d e_1^{d-1} e_2 \in \operatorname{Sym}_d \mathbb{R}^2$ with $a, b \geq 0$. We switch to the perspective of polynomials and study $p(x, y) = a x^d + b x^{d-1} y$. Then $\|p\|_{\mathsf{B}}^2 = a^2 + b^2 d$. We have the following two lower bounds for the uniform norm:

$$\|p\|_\infty \geq p\left(\sqrt{1 - \frac{1}{d}}, \frac{1}{\sqrt{d}}\right) = \frac{1}{\sqrt{d}^d}\left(a\sqrt{d-1}^d + bd\sqrt{d-1}^{d-1}\right) \tag{1.15}$$

and

$$\|p\|_\infty \geq p\,(1, 0) = a. \tag{1.16}$$

We can restrict to the case $\|p\|_{\mathsf{B}}^2 = a^2 + b^2 d = 1$ and need to show that

$$\|p\|_\infty > \left(1 - \frac{1}{d}\right)^{\frac{d-1}{2}}$$

whenever $a > 0$. The first lower bound (1.15) implies that this is true whenever $b > (\sqrt{d} - a\sqrt{d-1})/d$. Together with $1 = a^2 + b^2 d$ and $a, b \geq 0$ this verifies the claim for $0 < a < 2\sqrt{d(d-1)}/(2d-1)$. If $a \geq 2\sqrt{d(d-1)}/(2d-1)$, then the second lower bound (1.16) yields the desired estimate

$$\|p\|_\infty^2 \geq a^2 \geq \left(\frac{2\sqrt{d(d-1)}}{2d-1}\right)^2 > \frac{d-1}{d} > \left(1 - \frac{1}{d}\right)^{d-1}$$

for $d \geq 3$. □

1.3 Outlook

We have established the maximum relative distance of a real rank-two tensor to the set of rank-one tensors. However, it is not clear if similar techniques can be applied to gain results for tensors of higher rank. For rank-two tensors, we found that the maximum distance is attained for symmetric tensors. Again, it is not clear if this is the case for tensors of higher rank and results on tensors in full tensor spaces may suggest that this is not always the case. Another interesting question arises when we consider the influence of the field. We already discussed the influence of the field on the rank-one approximation ratio. When we consider the results of Theorem 1.1 for $d = 3$ and [CKP00], we observe that

$$\inf_{\substack{\mathbf{A}\in\mathbb{R}^{2\times2\times2}\\ \operatorname{rank}\mathbf{A}=2}} \frac{\|\mathbf{A}\|_\sigma}{\|\mathbf{A}\|_\mathsf{F}} = \frac{2}{3} = \min_{\mathbf{A}\in\mathbb{C}^{2\times2\times2}} \frac{\|\mathbf{A}\|_\sigma}{\|\mathbf{A}\|_\mathsf{F}} = \inf_{\substack{\mathbf{A}\in\mathbb{C}^{2\times2\times2}\\ \operatorname{rank}\mathbf{A}=2}} \frac{\|\mathbf{A}\|_\sigma}{\|\mathbf{A}\|_\mathsf{F}}$$

since tensors of complex rank two are dense in $\mathbb{C}^{2\times2\times2}$. That is, for tensors of order $d = 3$ and of rank two, the maximum relative distance coincides for real and complex tensors.

Chapter 2

Dynamical low-rank approximations to parabolic problems

Oftentimes, low-rank models need to be adapted when new data is observed. However, a new computation might be expensive or even unfeasible. Indeed, in many cases, the matrix $A(t_i)$ containing the underlying data at different times (t_i) is not stored and there is only access to the increment $A(t_{i+1}) + A(t_i)$ and the model $X(t_i)$. If the model at different times is a rank-r matrix $X(t_i)$, a natural update is

$$X(t_{i+1}) = \operatorname*{arg\,min}_{\operatorname{rank} X = r} \|X - X(t_i) - A(t_{i+1}) + A(t_i)\|_{\mathsf{F}}^2,$$

i.e., a low-rank model with a similar increment is chosen. The first-order optimality condition for this problem is

$$\langle X - X(t_i), Y\rangle_{\mathsf{F}} = \langle A(t_{i+1}) + A(t_i), Y\rangle_{\mathsf{F}} \quad \text{for all } Y \in T_X\mathcal{M}_r,$$

where $T_X\mathcal{M}_r$ is the *tangent space* to the manifold $\mathcal{M}_r$ of rank-r matrices at X. When passing to continuous times, the optimality condition becomes the evolution equation

$$\langle X'(t), Y\rangle_{\mathsf{F}} = \langle A'(t), Y\rangle_{\mathsf{F}} \quad \text{for all } Y \in T_{X(t)}\mathcal{M}_r$$

in weak form. In finite dimensions, since $X(t) \in \mathcal{M}_r$, this is equivalent to the explicit form

$$X'(t) = P_{T_{X(t)}\mathcal{M}_r} A'(t),$$

where $P_{T_{X(t)}\mathcal{M}_r}$ is the orthogonal projection onto the tangent space $T_{X(t)}\mathcal{M}_r$. Of course, nothing prevents us to exchange the increment A' by a general differential equation. We then have a problem of the form

$$\langle X'(t), Y\rangle_{\mathsf{F}} = \langle F(X(t), t), Y\rangle_{\mathsf{F}} \quad \text{for all } Y \in T_{X(t)}\mathcal{M}_r. \tag{2.1}$$

This approach was first considered in [KL07a] for low-rank matrices, but variational principles analogous to (2.1) were already used by Dirac in [Dir30] to compute approximations of wave functions as antisymmetrized rank-one tensors. In the literature, this approach is known as the *Dirac-Frenkel variational principle* and is used for many different non-linear reduced models; see e.g., [Lub08, Chapter II and Chapter IV].

In this chapter, we are concerned with the case where (2.1) comes from a parabolic partial differential equation with a separable spacial domain $\Omega = \Omega_1 \times \Omega_2$. When the solutions of partial differential equations can be well approximated by functions in a low-rank format, a typical strategy is to discretize on huge but finite grids, and afterwards impose low-rank constraints; see e.g., [Hac19, Chapter 16 and Chapter 17.3] and [BSU16]. Numerical methods for computing solutions on low-rank manifolds are under current research; see e.g., [LO14, CL22]. When analyzing such an approach, it is important to understand the underlying infinite-dimensional problems. A first step is the existence and uniqueness of solutions. For elliptic partial differential equations, it is not too difficult to provide a framework that ensures existence of solutions [BSU16, Section 4]. We will see, that the parabolic case is more delicate. First, we study a model problem in Section 2.1 and extract features that can be expected in a more general case. We formulate an abstract problem in Section 2.2 and provide a temporal discretization in Section 2.3. In Section 2.4, we show that the solutions of the time-discrete problem converge to a solution of the continuous problem and that this solution is essentially unique. Finally, we show that also solutions to a space-discrete problem converge to the unique solution in Section 2.5.

2.1 Model problem

As a model, we consider the anisotropic diffusion equation

$$\begin{aligned} \frac{\partial}{\partial t} u(x,t) - \nabla_x \cdot (B(t) \nabla_x u(x,t)) &= f(x,t) && \text{for } (x,t) \in \Omega \times (0,T), \\ u(x,t) &= 0 && \text{for } (x,t) \in \partial\Omega \times (0,T), \\ u(x,0) &= u_0(x) && \text{for } x \in \Omega \end{aligned} \tag{2.2}$$

on the product domain $\Omega = (0,1) \times (0,1)$. Here, $B(t)$ is a 2×2 matrix, and we assume it to be uniformly bounded and positive definite, as well as Lipschitz continuous with respect to t. When B is constant such an equation is for example attained when considering an isotropic diffusion equation after a linear change of variables.

Typically, one does not seek a classical solution for a problem of the form (2.2). Instead, the problem is formulated in the weak form on function spaces. The solution is a function $u \in L^2(0,T; H_0^1(\Omega))$ with values in the Hilbert space $H_0^1(\Omega)$ (the space of square-integrable functions vanishing on the boundary $\partial\Omega$ with square-integrable weak derivatives in space) and its derivative in time $u' \in L^2(0,T; H^{-1}(\Omega))$ has values in the dual space $H^{-1}(\Omega)$ of $H_0^1(\Omega)$. The problem in weak form reads: given $u_0 \in L^2(\Omega)$ and $f \in L^2(0,T; H^{-1}(\Omega))$, the solution is a function

$$u \in W(0,T; H_0^1(\Omega), H^{-1}(\Omega)) \coloneqq \{u \in L^2(0,T; H_0^1(\Omega)) \colon u' \in L^2(0,T; H^{-1}(\Omega))\}$$

such that for almost all $t \in (0,T)$

$$\begin{gathered}\langle u'(t), v\rangle + a(u(t), v; t) = \langle f(t), v\rangle \quad \text{for all } v \in H_0^1(\Omega)\ , \\ u(0) = u_0. \end{gathered} \tag{2.3}$$

Here, by $\langle \cdot, \cdot \rangle$ we denote the dual paring $H^{-1}(\Omega) \times H_0^1(\Omega)$ and $a\colon H_0^1(\Omega) \times H_0^1(\Omega) \times [0,T]$ is a bounded, symmetric and coercive bilinear form

$$a(u, v; t) := \int_\Omega (B(t)\nabla u(x)) \cdot \nabla v(x) dx$$

for every t. Classical theory provides a unique solution to (2.3); see e.g., [Zei90a, Theorem 23.A] and [Zei90b, Theorem 30.A].

Since $\Omega = (0,1)\times(0,1)$, the Hilbert spaces $L^2(\Omega)$ and $H_0^1(\Omega)$ admit certain structures. For a product $u(x,y) = u_1(x)u_2(y)$ and $v(x,y) = v_1(x)v_2(y)$ the inner products are

$$\langle u, v\rangle_{L^2(\Omega)} = \langle u_1, v_1\rangle_{L^2(0,1)} \langle u_2, v_2\rangle_{L^2(0,1)} \tag{2.4}$$

and

$$\langle u, v\rangle_{H_0^1(\Omega)} = \langle u_1, v_1\rangle_{H_0^1(0,1)} \langle u_2, v_2\rangle_{L^2(0,1)} + \langle u_1, v_1\rangle_{L^2(0,1)} \langle u_2, v_2\rangle_{H_0^1(0,1)}. \tag{2.5}$$

Also importantly, the inner product in the Hilbert space $H^{1,1}_{\text{mix}}(\Omega)$, which contains the functions $u \in H_0^1(\Omega)$ with square integrable mixed derivatives $\frac{\partial^2}{\partial x \partial y} u$, is given by

$$\langle u, v\rangle_{H^{1,1}_{\text{mix}}} = \int_\Omega \frac{\partial^2}{\partial x \partial y} u(x,y) \frac{\partial^2}{\partial x \partial y} v(x,y) dx dy = \langle u_1, v_1\rangle_{H_0^1(0,1)} \langle u_2, v_2\rangle_{H_0^1(0,1)}. \tag{2.6}$$

A function $u \in L^2(\Omega)$ admits a singular value decomposition

$$u(x,y) = \sum_{i=1}^\infty \sigma_i\, u_{1,i}(x) u_{2,i}(y) \quad \text{for almost every } x, y \in (0,1), \tag{2.7}$$

with $L^2(0,1)$ orthonormal $u_{1,i}$ and $u_{2,i}$ and a nonnegative, square-summable, and nonincreasing sequence (σ_i); see e.g., [Hac19, Chapter 4.4.3]. By $\operatorname{rank} u$ we denote the smallest number of nonzero terms needed. Note that $\operatorname{rank} u = \infty$ is possible.

As low-rank representations are convenient for several reasons, one can ask whether the parabolic equation (2.2) admits approximate solutions of low-rank. In dynamical low-rank approximation, one assumes this to be the case and attempts to directly evolve the solution on the set

$$\mathcal{M}_r = \{u \in L^2(\Omega)\colon \operatorname{rank} u = r\} \tag{2.8}$$

for a certain value r. The set $\mathcal{M}_r$ can be shown to be a differentiable manifold in various ways. In [FHN19] it is shown to be a Banach manifold, and in the appendix of [BEKU21] it is shown to be an embedded submanifold using submersions. There are many subtleties for manifolds in infinite dimensions, however we can circumvent many of these in this chapter. We only require the existence of a tangent space $T_u\mathcal{M}_r$ that contains the derivatives of curves through the point u in the manifold $\mathcal{M}_r$ and certain curvature estimates.

We study a modified version of (2.3). Given the initial point $u_0 \in T_{u(t)}\mathcal{M}_r \cap H_0^1(\Omega)$ and $f \in L^2(0,T;L^2(\Omega))$, we seek

$$u \in W(0,T;H_0^1(\Omega),L^2(\Omega)) \coloneqq \{u \in L^2(0,T;H_0^1(\Omega))\colon u' \in L^2(0,T;L^2(\Omega))\}$$

such that $u(t) \in \mathcal{M}_r$ for all $t \in (0,T)$ and

$$\begin{aligned} \langle u'(t), v\rangle + a(u(t),v;t) = \langle f(t), v\rangle \quad &\text{for all } v \in T_{u(t)}\mathcal{M}_r \cap H_0^1(\Omega), \\ u(0) = u_0 \quad & \end{aligned} \tag{2.9}$$

for almost every $t \in (0,T)$. This can be seen as a nonlinear version of the Galerkin method. In contrast to (2.3) we only require the equation to hold only for tangent vectors to the desired solution $u(t)$.

The situation is easier if $B(t)$ is diagonal and $f = 0$. Then the solutions of (2.3) and (2.9) coincide and can be computed by separation of variables. From an abstract perspective, this happens because the unbounded linear operator on $L^2(\Omega)$ induced by the bilinear form a maps into the tangent space at u. In [KL07b] a similar problem is studied, where the operator on $L^2(\Omega)$ is split into an unbounded part mapping to the tangent space and an arbitrary bounded part. If $B(t)$ is not diagonal this is not the case and these results are not applicable.

2.1.1 Properties of the manifold

First, we note that $\mathcal{M}_r$ is not closed. Its closure is indeed the set $\mathcal{M}_{\le r}$ of functions with rank at most r. The set $\mathcal{M}_{\le r}$ is even weakly sequentially closed; see e.g., [Hac19, Lemma 8.6]. In other words

$$\mathcal{M}_{\le r} = \mathcal{M}_{\le r-1} \cup \mathcal{M}_r = \overline{\mathcal{M}_r} = \overline{\mathcal{M}_r}^{\mathsf{w}}.$$

Also importantly, the set $\mathcal{M}_r$ is a cone. That is, $\alpha u \in \mathcal{M}_r$ for every $u \in \mathcal{M}_r$ and $\alpha > 0$.

2.1.2 Mixed regularity

For convenience, let us denote $u = u_1 \otimes u_2 \in L^2(\Omega)$ for the product of functions, that is, $u(x,y) = u_1(x)u_2(y)$ for almost every $(x,y) \in \Omega$. Every $u \in \mathcal{M}_r$ admits infinitely many decompositions $\sum_{i=1}^r u_{1,i} \otimes u_{2,i}$, one of them being the singular value decomposition (2.7).

A key observation, is that $u \in \mathcal{M}_r \cap H_0^1(\Omega)$ is also in $H_{\text{mix}}^{1,1}(\Omega)$. To see this, let $u = \sum_{i=1}^r \sigma_i\, u_{1,i} \otimes u_{2,i}$ be a singular value decomposition and $u \in H_0^1(\Omega)$. Since the singular vectors $u_{1,i}$ and $u_{2,i}$ are respectively orthonormal in $L^2(0,1)$, equation (2.5) results in

$$\|u\|^2_{H^1_0(\Omega)} = \sum \sigma_i^2 \left(\|u_{1,i}\|^2_{H^1_0(0,1)} + \|u_{2,i}\|^2_{H^1_0(0,1)} \right). \tag{2.10}$$

This already implies that the singular vectors belong to $H^1_0(0,1)$. Equation (2.6), the triangle inequality, and Young's inequality imply

$$\|u\|_{H^{1,1}_{\mathrm{mix}}(\Omega)} \leq \sum_{i=1}^{r} \sigma_i \|u_{1,i}\|_{H^1_0(0,1)} \|u_{2,i}\|_{H^1_0(0,1)} \leq \sum_{i=1}^{r} \frac{\sigma_i}{2} \left(\|u_{1,i}\|^2_{H^1_0(0,1)} + \|u_{2,i}\|^2_{H^1_0(0,1)} \right),$$

which gives

$$\|u\|_{H^{1,1}_{\mathrm{mix}}(\Omega)} \leq \frac{1}{2\sigma_r} \|u\|^2_{H^1_0(\Omega)}, \tag{2.11}$$

i.e., the $H^{1,1}_{\mathrm{mix}}(\Omega)$-norm of u is bounded by the inverse of its smallest singular value and its $H^1_0(\Omega)$ norm.

2.1.3 Tangent spaces and curvature estimates

Given a decomposition $u = \sum_{i=1}^r u_{1,i} \otimes u_{2,i}$, the tangent space at $u \in \mathcal{M}_r$ is given by

$$T_u\mathcal{M}_r = \left\{ \sum_{i=1}^{r} (u_{1,i} \otimes v_{2,i} + v_{1,i} \otimes u_{2,i}) : v_{1,i}, v_{2,i} \in L^2(0,1) \right\}. \tag{2.12}$$

It is quite apparent, that $T_u\mathcal{M}_r$ contains only tangent vectors. Indeed, the curve $\varphi(t) = \sum_{i=1}^r (u_{1,i} + tv_{1,i}) \otimes (u_{2,i} + tv_{2,i})$ lies in $\mathcal{M}_r$ for small t, $\varphi(0) = u$ and

$$\varphi'(0) = \sum_{i=1}^{r} (u_{1,i} \otimes v_{2,i} + v_{1,i} \otimes u_{2,i})$$

is of the form given in (2.12). For a proof, that $T_u\mathcal{M}_r$ contains all tangent vectors, we refer to the appendix of [BEKU21]. We denote by $\mathcal{U}_1 = \operatorname{span}\{u_{1,i} \colon i = 1, \dots, r\}$ and $\mathcal{U}_2 = \operatorname{span}\{u_{2,i} \colon i = 1, \dots, r\}$ the linear spaces containing the left and right singular vectors, respectively. Then (2.12) takes the form

$$\begin{aligned} T_u\mathcal{M}_r &= \mathcal{U}_1 \otimes L^2(0,1) + L^2(0,1) \otimes \mathcal{U}_2 \\ &= (\mathcal{U}_1 \otimes \mathcal{U}_2) \oplus \left(\mathcal{U}_1 \otimes \mathcal{U}_2^{\perp} \right) \oplus \left(\mathcal{U}_1^{\perp} \otimes \mathcal{U}_2 \right) = \left(\mathcal{U}_1^{\perp} \otimes \mathcal{U}_2^{\perp} \right)^{\perp}, \end{aligned} \tag{2.13}$$

where the superscript $\perp$ denotes the $L^2(0,1)$ and $L^2(\Omega)$ orthogonal complement and $\oplus$ denotes the direct sum of vector spaces, i.e., the intersection of the summands contains the zero element only. Hence, $T_u\mathcal{M}_r$ is closed and the $L^2(\Omega)$-orthogonal projection P_u onto $T_u\mathcal{M}_r$ is given by

$$\begin{aligned} P_u &= P_1 \otimes \mathrm{id}_{L^2(0,1)} + \mathrm{id}_{L^2(0,1)} \otimes P_2 - P_1 \otimes P_2 \\ &= P_1 \otimes P_2 + P_1 \otimes \left(\mathrm{id}_{L^2(0,1)} - P_2 \right) + \left(\mathrm{id}_{L^2(0,1)} - P_1 \right) \otimes P_2 \\ &= \mathrm{id}_{L^2(\Omega)} - \left(\mathrm{id}_{L^2(0,1)} - P_1 \right) \otimes \left(\mathrm{id}_{L^2(0,1)} - P_2 \right), \end{aligned} \tag{2.14}$$

where P_1 and P_2 are the $L^2(0,1)$-orthogonal projections onto $\mathcal{U}_1$ and $\mathcal{U}_2$, respectively.

The curvature of the manifold $\mathcal{M}_r$ at u and the $L^2(\Omega)$ distance of u to the relative boundary $\mathcal{M}_{\leq r-1}$ of $\mathcal{M}_r$ is given by the singular values of u. In particular, let σ_r be the smallest singular value of u. Then

$$\min_{w \in \mathcal{M}_{\leq r-1}} \|u - w\|_{L^2(\Omega)} = \sigma_r \tag{2.15}$$

and for $v \in \mathcal{M}_r$, we have the curvature estimates

$$\|P_u - P_v\|_{L^2(\Omega) \to L^2(\Omega)} \leq \frac{2}{\sigma_r} \|u - v\|_{L^2(\Omega)} \tag{2.16}$$

and

$$\| \left(\mathrm{id}_{L^2(\Omega)} - P_v \right) (u - v) \|_{L^2(\Omega)} \leq \frac{1}{\sigma_r} \|u - v\|^2_{L^2(\Omega)}; \tag{2.17}$$

see e.g., the appendix of [BEKU21] or [AJ14, CL10, LRSV13, WCCL16] for similar results.

We further note, that a weakly compact subset $\mathcal{M}' \subset \mathcal{M}_r$ has positive $L^2(\Omega)$-distance σ^* from $\mathcal{M}_{\leq r-1}$ and is attained for some $u^* \in \mathcal{M}'$. To see this, note first that for Banach spaces, by the Eberlein-Šmulian theorem, weak compactness is equivalent to weak sequential compactness; see e.g., [Die84, Chapter III]. Consider sequences $(u_n) \subset \mathcal{M}'$ and $(v_n) \subset \mathcal{M}_{\leq r-1}$ such that

$$\|u_n - v_n\|_{L^2(\Omega)} \leq \sigma_* + 1/n.$$

Both sequences are bounded, and hence (u_n, v_n) admits a weakly converging subsequence with limit (u_*, v_*). Then $u_* \in \mathcal{M}'$ and $v_* \in \mathcal{M}_{\leq r-1}$ since both sets are weakly sequentially closed. Since the norm is weakly sequentially lower semicontinuous, we obtain $\sigma_* \leq \|u_* - v_*\|_{L_2(\Omega)} \leq \sigma_*$, and thus equality. This shows

$$\sigma_* = \min_{u \in \mathcal{M}',\, v \in \mathcal{M}_{\leq r-1}} \|u - v\|_{L^2(\Omega)} > 0. \tag{2.18}$$

2.1.4 Compatibility of tangent spaces

We also use the intersection of the manifold and tangent spaces with the Sovolev space $H^1_0(\Omega)$. First, we require that for $u \in \mathcal{M}_r \cap H^1_0(\Omega)$ and $v \in T_u\mathcal{M}_r \cap H^1_0(\Omega)$ a curve $\varphi(t) \in \mathcal{M}_r \cap H^1_0(\Omega)$ can be chosen, such that $\varphi(0) = u$ and $\varphi'(0) = v$. To see this, let $u = \sum_{i=1}^r \sigma_i \, u_{1,i} \otimes u_{2,i}$ be a singular value decomposition. Then by the second line of (2.13), we can decompose $v = v_1 + v_2 + v_3$ with $v_1 \in \mathcal{U}_1 \otimes \mathcal{U}_2^\perp$, $v_2 \in \mathcal{U}_1^\perp \otimes \mathcal{U}_2$ and $v_3 \in \mathcal{U}_1 \otimes \mathcal{U}_2$. More explicitly,

$$v_1 = \sum_{i=1}^r u_{1,i} \otimes v_{2,i}, \quad v_2 = \sum_{i=1}^r v_{1,i} \otimes u_{2,i}, \quad \text{and} \quad v_3 = \sum_{i=1}^r \sum_{j=1}^r m_{ij} u_{1,i} \otimes u_{2,j},$$

with $v_{1,i} \in \mathcal{U}_1^\perp$, $v_{2,i} \in \mathcal{U}_2^\perp$, and $\sum_{i=1}^r \sum_{j=1}^r |m_{ij}|^2 < \infty$. Note that by (2.11)

$$v_3 \in \mathcal{U}_1 \otimes \mathcal{U}_2 \subset H_0^1(0,1) \otimes H_0^1(0,1) \subset H_0^1(\Omega),$$

and hence $v - v_3 \in \mathcal{T}_u\mathcal{M}_r \cap H_0^1(0,1)$. Again using $L^2(0,1)$-orthogonality and (2.5), we can bound the $H_0^1(0,1)$-norm of the factors $v_{1,i}$ and $v_{2,i}$ via

$$\begin{aligned}&\|v - v_3\|_{H_0^1(\Omega)}^2 \\ &= \sum_{i=1}^r \|v_{1,i}\|_{H_0^1(0,1)}^2 + \|v_{2,i}\|_{H_0^1(0,1)}^2 + \|u_{1,i}\|_{H_0^1(0,1)}^2 \|v_{1,i}\|_{L^2(0,1)}^2 + \|u_{2,i}\|_{H_0^1(0,1)}^2 \|v_{2,i}\|_{L^2(0,1)}^2.\end{aligned}$$

Therefore, $\varphi(t) = \sum_{i=1}^r (u_{1,i} + tv_{1,i}) \otimes \left(u_{2,i} + tv_{2,i} + t\sum_{j=1}^r m_{ij} u_{2,j}\right)$ is a smooth curve in $\mathcal{M}_r \cap H_0^1(\Omega)$ for small $|t|$ and has the desired properties $\varphi(0) = u$ and $\varphi'(0) = v$.

Based on the regularity of the singular vectors one can also show that if $u \in \mathcal{M}_r \cap H_0^1(\Omega)$, the tangent space projection P_u given in (2.14) can be bounded in $H_0^1(\Omega)$-norm as a map from $H_0^1(\Omega)$ to $T_u\mathcal{M}_r \cap H_0^1(\Omega)$ as follows:

$$\|P_u v\|_{H_0^1(\Omega)} \leq \left(1 + \frac{r}{\sigma_r(u)^2} \|u\|_{H_0^1(\Omega)}^2\right)^{1/2} \|v\|_{H_0^1(\Omega)}, \tag{2.19}$$

i.e., P_u is a continuous linear operator from $H_0^1(\Omega)$ into $H_0^1(\Omega)$ if $u \in \mathcal{M}_r \cap H_0^1(\Omega)$; see e.g., [BEKU21, Proposition A.4].

2.1.5 Operator splitting

We apply the results of Section 2.1.2 and Section 2.1.4 to split the bilinear form $a(\cdot,\cdot;t)$ into two parts $a_1(\cdot,\cdot;t)$ and $a_2(\cdot,\cdot;t)$. We can deal with these using different techniques to obtain existence and uniqueness results. The first part is coming from the diagonal entries $b_{11}(t)$ and $b_{22}(t)$ of $B(t)$ and the second is coming from the off-diagonal entry $b_{12}(t)$, i.e.,

$$a_1(u,v;t) = \iint_\Omega b_{11}(t)\frac{\partial}{\partial x}u(x,y)\frac{\partial}{\partial x}v(x,y) + b_{22}(t)\frac{\partial}{\partial y}u(x,y)\frac{\partial}{\partial y}v(x,y)\,dx\,dy$$

and

$$a_2(u,v;t) = \iint_\Omega b_{12}(t)\frac{\partial}{\partial x}u(x,y)\frac{\partial}{\partial y}v(x,y) + b_{12}(t)\frac{\partial}{\partial y}u(x,y)\frac{\partial}{\partial x}v(x,y)\,dx\,dy.$$

Note that the set of compactly supported and smooth functions $C_c^\infty(\Omega)$ is a dense subset of $H_0^1(\Omega)$ and it is not difficult to see that also $\mathcal{M}_r \cap C_c^\infty(\Omega)$ is a dense subset of $\mathcal{M}_r \cap H_0^1(\Omega)$. Therefore, there exists a sequence $(u_n) \subset \mathcal{M}_r \cap C_c^\infty(\Omega)$ converging to $u \in \mathcal{M}_r \cap H_0^1(\Omega)$ with respect to the $H_0^1(\Omega)$ norm. For a compactly supported and smooth function u, the bilinear forms take the form

$$a_1(u,v;t) = -\iint_\Omega \left(b_{11}(t)\frac{\partial^2}{\partial x^2}u(x,y) + b_{22}(t)\frac{\partial^2}{\partial y^2}u(x,y)\right) v(x,y)\,dx\,dy$$

and

$$a_2(u,v;t) = -\iint_\Omega 2b_{12}(t)\frac{\partial^2}{\partial x \partial y}u(x,y)v(x,y)\,dx\,dy.$$

For any $u \in \mathcal{M}_r \cap C_c^\infty(\Omega)$ let $u = \sum_{i=1}^r \sigma_i\, u_{1,i} \otimes u_{2,i}$ be its singular value decomposition. Then, by orthogonality, $\sigma_i\, u_{1,i}(x) = \int_0^1 u(x,y)u_{2,i}\,dy$ and $\sigma_i\, u_{2,i}(y) = \int_0^1 u(x,y)u_{1,i}\,dx$. It follows, that the singular vectors are smooth and compactly supported, i.e., the singular vectors $u_{1,i}, u_{2,i}$ lie in $C_c^\infty(0,1)$. It follows that

$$\begin{aligned} &b_{11}(t)\frac{\partial^2}{\partial x^2}u + b_{22}(t)\frac{\partial^2}{\partial y^2}u \\ &\qquad = \sum_{i=1}^r \sigma_i \left(b_{11}(t)\left(\frac{\partial^2}{\partial x^2}u_{1,i}\right) \otimes u_{2,i} + b_{22}(t)u_{1,i} \otimes \left(\frac{\partial^2}{\partial y^2}u_{2,i}\right)\right) \in T_u\mathcal{M}_r. \end{aligned}$$

Hence,

$$\begin{aligned} a_1(u,v;t) &= -\langle b_{11}(t)\frac{\partial^2}{\partial x^2}u + b_{22}(t)\frac{\partial^2}{\partial y^2}u, v\rangle_{L^2(\Omega)} \\ &= -\langle b_{11}(t)\frac{\partial^2}{\partial x^2}u + b_{22}(t)\frac{\partial^2}{\partial y^2}u, P_u v\rangle_{L^2(\Omega)} = a_1(u, P_u v;t) \end{aligned} \tag{2.20}$$

is fulfilled for any $u \in \mathcal{M}_r \cap C_c^\infty(\Omega)$ and $v \in H_0^1(\Omega)$. Both the left-hand side and the right-hand side of (2.20) are well defined for all $u \in \mathcal{M}_r \cap H_0^1(\Omega)$ due to (2.19). Indeed, let $(u_n) \subset \mathcal{M}_r \cap C_c^\infty(\Omega)$ converge to $u \in \mathcal{M}_r \cap H_0^1(\Omega)$. Then $a_1(u_n, v;t)$ converges to $a_1(u,v;t)$ since a_1 is a bounded bilinear form $H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$. Since $\|u - u_n\|_{L^2(\Omega)} \lesssim \|u - u_n\|_{H_0^1(\Omega)}$, the sequence $(P_{u_n}v)$ converges to $P_u v$ in $L^2(\Omega)$ due to (2.16). The set $L^2(\Omega)$ is a dense subset of $H^{-1}(\Omega)$ and $\|P_{u_n}v\|_{H_0^1(\Omega)}$ is bounded due to (2.19). Hence, the sequence $(P_{u_n}v)$ converges weakly to $P_u v$ in $H_0^1(\Omega)$; see e.g., [Zei90a, Proposition 21.23(g)]. By continuity, $a_1(u_n, \cdot;t)$ converges to $a_1(u,\cdot;t)$ in $H^{-1}(\Omega)$ as $n \to \infty$. This implies convergence of $a_1(u_n, P_{u_n}v;t)$ to $a_1(u, P_u v;t)$ as $n \to \infty$; see e.g., [Zei90a, Proposition 21.23(j)]. Ofcourse the limits coming from the right-hand side and the left-hand side of (2.20) have to coincide. Hence, equation (2.20) holds for any $u \in \mathcal{M}_r\ H_0^1(\Omega)$ and $v \in H_0^1(\Omega)$.

To handle the bilinear form a_2, we use (2.11) to bound $a_2(u,v;t)$ in terms of the $H_0^1(\Omega)$-norm of u and the $L^2(\Omega)$-norm of v. For $u \in \mathcal{M}_r \cap H_0^1(\Omega)$ we get

$$a_2(u,v;t) = -2b_{12}(t)\left\langle \frac{\partial^2}{\partial x \partial y}u, v\right\rangle_{L^2(\Omega)} \leq \frac{|b_{12}(t)|}{\sigma_r}\|u\|^2_{H_0^1(\Omega)}\|v\|_{L^2(\Omega)}, \tag{2.21}$$

where σ_r is the smallest singular value of u. That is, the bilinear form $a_2(\cdot,\cdot;t)$ induces a linear operator $A_2(t)$ such that $\|A_2(t)u\|_{L^2(\Omega)}$ is bounded in terms of the smallest singular value of u and its $H_0^1(\Omega)$-norm for $u \in \mathcal{M}_r \cap H_0^1(\Omega)$.

2.2 Abstract formulation

Motivated by the properties of the model problem discussed in Section 2.1, we formulate a more general setting. We consider a *Gelfand triplet* $\mathcal{V} \hookrightarrow \mathcal{H} \cong \mathcal{H}^* \hookrightarrow \mathcal{V}^*$ of Hilbert spaces, where $\mathcal{V}$ is compactly embedded in $\mathcal{H}$. This implies that the embedding is also continuous, i.e.,

$$\|u\|_{\mathcal{H}} \lesssim \|u\|_{\mathcal{V}}. \tag{2.22}$$

In our model problem $\mathcal{H} = L^2(\Omega)$ and $\mathcal{V} = H_0^1(\Omega)$, the compact embedding is due to the *Rellich–Kondrachov theorem* and (2.22) is the *Poincaré inequality*; see e.g., [Zei90a, Proposition 18.9 and Proposition 19.25].

By $\langle\cdot,\cdot\rangle : \mathcal{V}^* \times \mathcal{V} \to \mathbb{R}$ we denote the dual pairing between $\mathcal{V}^*$ and $\mathcal{V}$. Note that for $u \in \mathcal{H} \subset \mathcal{V}^*$ and $v \in \mathcal{V} \subset \mathcal{H}$ the dual pairing and the inner product on $\mathcal{H}$ coincide, i.e., $\langle u, v\rangle_{\mathcal{H}} = \langle u, v\rangle$. We will frequently identify $u \in \mathcal{V}$ as an element of $\mathcal{H}$ and in turn also as an element in $\mathcal{V}^*$. For every $t \in [0,T]$, let $a(\cdot,\cdot;t) : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ be a bilinear form which is assumed to be symmetric,

$$a(u,v;t) = a(v,u;t) \quad \text{for all } u,v \in \mathcal{V} \text{ and } t \in [0,T],$$

uniformly bounded, i.e., there exists $\beta > 0$ such that

$$|a(u,v;t)| \le \beta\|u\|_{\mathcal{V}}\|v\|_{\mathcal{V}} \quad \text{for all } u,v \in \mathcal{V} \text{ and } t \in [0,T],$$

and uniformly coercive, i.e., there exists $\mu > 0$ such that

$$a(u,u;t) \ge \mu\|u\|_{\mathcal{V}}^2 \quad \text{for all } u \in \mathcal{V} \text{ and } t \in [0,T].$$

Under these assumptions, $a(\cdot,\cdot;t)$ is an inner product on $\mathcal{V}$ defining an equivalent norm. Furthermore, it defines a bounded operator

$$A(t) : \mathcal{V} \to \mathcal{V}^* \tag{2.23}$$

such that

$$a(u,v;t) = \langle A(t)u, v\rangle \quad \text{for all } u,v \in \mathcal{V}.$$

We also assume that $a(u,v;t)$ is Lipschitz continuous with respect to t. In other words, there exists an $L \ge 0$ such that

$$|a(u,v;t) - a(u,v;s)| \le L\beta\|u\|_{\mathcal{V}}\|v\|_{\mathcal{V}}|t-s| \tag{2.24}$$

for all $u, v \in \mathcal{V}$ and $s, t \in [0, T]$. In the model problem, this corresponds to the Lipschitz continuity of the function $t \mapsto B(t)$.

We deal with evolution equations on a submanifold $\mathcal{M} \subset \mathcal{H}$. We do not require an exact notion of a manifold here. For our purpose it is sufficient that for every point $u \in \mathcal{M}$ there exists a closed subspace $T_u\mathcal{M} \subset \mathcal{H}$ such that $T_u\mathcal{M}$ contains all tangent vectors to $\mathcal{M}$ at u. Here, a tangent vector is any $v \in \mathcal{H}$ for which there exists a (strongly) differentiable curve $\varphi\colon (-\epsilon, \epsilon) \to H$ (for some $\epsilon > 0$) such that $\varphi(t) \in \mathcal{M}$ for all t and

$$\varphi(0) = u, \quad \varphi'(0) = v.$$

By $P_u : \mathcal{H} \to T_u\mathcal{M}$ we denote the $\mathcal{H}$-orthogonal projection onto $T_u\mathcal{M}$. We will also assume $\mathcal{M} \cap \mathcal{V}$ to be nonempty as well as $T_u\mathcal{M} \cap \mathcal{V}$ to be nonempty for $u \in \mathcal{M} \cap \mathcal{V}$.

The abstract problem takes the following form.

Problem 2.1. Given $f \in L^2(0, T; \mathcal{H})$ and $u_0 \in \mathcal{M} \cap \mathcal{V}$, find

$$u \in W(0, T; \mathcal{V}, \mathcal{H}) \coloneqq \{u \in L_2(0, T; \mathcal{V})\colon u' \in L_2(0, T; \mathcal{H})\}$$

such that for almost all $t \in [0, T]$,

$$\begin{aligned} u(t) &\in \mathcal{M}, \\ \langle u'(t), v\rangle + a(u(t), v; t) &= \langle f(t), v\rangle \quad \text{for all } v \in T_{u(t)}\mathcal{M} \cap \mathcal{V}, \\ u(0) &= u_0. \end{aligned} \tag{2.25}$$

We emphasize again that the main challenge for the analysis of solutions to this weak formulation is that according to the Dirac-Frenkel principle, the test functions are from the tangent space only. For now, we require additional smoothness of the initial value u_0 and the right-hand side f compared to solutions on the entire space $\mathcal{V}$ and we do not know if this is necessary. To show that Problem 2.1 admits solutions we will require several assumptions. These assumptions are abstractions of corresponding properties of the model problem of a low-rank manifold as discussed in Section 2.1, and hence the main results of this chapter apply to this setting. The assumptions are the following.

A1 (Cone property) $\mathcal{M}$ is a cone, that is, $u \in \mathcal{M}$ implies $su \in \mathcal{M}$ for all $s > 0$.

A2 (Curvature bound) For every subset $\mathcal{M}'$ of $\mathcal{M}$ that is weakly compact in $\mathcal{H}$, there exists a constant $\kappa = \kappa(\mathcal{M}')$ such that

$$\|P_u - P_v\|_{\mathcal{H}\to\mathcal{H}} \leq \kappa\|u - v\|_{\mathcal{H}}$$

and

$$\|(I - P_u)(u - v)\|_{\mathcal{H}} \leq \kappa\|u - v\|_{\mathcal{H}}^2$$

for all $u, v \in \mathcal{M}'$.

A3 (Compatibility of tangent spaces)

(a) For $u \in \mathcal{M} \cap \mathcal{V}$ and $v \in T_u\mathcal{M} \cap \mathcal{V}$ an admissible curve with $\varphi(0) = u$, $\varphi'(0) = v$ can be chosen such that

$$\varphi(t) \in \mathcal{M} \cap \mathcal{V}$$

for all $|t|$ small enough.

(b) If $u \in \mathcal{M} \cap \mathcal{V}$ and $v \in \mathcal{V}$ then $P_u v \in T_u\mathcal{M} \cap \mathcal{V}$.

A4 (Operator splitting) The associated operator $A(t)$ in (2.23) admits a splitting

$$A(t) = A_1(t) + A_2(t)$$

into two uniformly bounded operators $\mathcal{V} \to \mathcal{V}^*$ such that for all $t \in [0, T]$, all $u \in \mathcal{M} \cap \mathcal{V}$ and all $v \in \mathcal{V}$, the following holds:

(a) $A_1(t)$ maps to the tangent space, i.e.,

$$\langle A_1(t)u, v\rangle = \langle A_1(t)u, P_u v\rangle.$$

(b) $A_2(t)$ is locally bounded from $\mathcal{M} \cap \mathcal{V}$ to $\mathcal{H}$, i.e., for every subset $\mathcal{M}'$ of $\mathcal{M}$ that is weakly compact in $\mathcal{H}$, there exists $\gamma = \gamma(\mathcal{M}') > 0$ such that

$$A_2(t)u \in \mathcal{H} \quad \text{and} \quad \|A_2(t)u\|_{\mathcal{H}} \leq \gamma \|u\|_{\mathcal{V}}^{\eta} \quad \text{for all } u \in \mathcal{M}'$$

with an $\eta > 0$ independent of $\mathcal{M}'$.

For the model-problem, we explained **A1** in Section 2.1.1, the estimates in **A2** are (2.16) and (2.17) in Section 2.1.3, the compatibility of tangent spaces **A3** is discussed in Section 2.1.4, especially **A3**(b) follows from (2.19), and the operator splitting **A4** is elaborated in the equations (2.20) and (2.21) of Section 2.1.5. For **A4**(b) we take into account, that every weakly compact subset of $\mathcal{M}_r$ has a positive distance from its relative boundary as stated in (2.18) and $\eta = 2$.

We may weaken the uniform coercivity assumption to a uniform Gårding inequality

$$\langle A(t)u, u\rangle \geq \mu\|u\|_{\mathcal{V}}^2 - \alpha\|u\|_{\mathcal{H}}^2$$

as seen below. In the model problem, we can therefore not only handle *Dirichlet boundary conditions* but for instance also *Neumann boundary conditions*. To see this, suppose v is a solution (in the sense of Problem 2.1) of

$$\langle v'(t) + (A(t) + \alpha\,\mathrm{id})v(t), w\rangle = \langle e^{-\alpha t} f(t), w\rangle \quad \text{for all } w \in T_{v(t)}\mathcal{M} \cap \mathcal{V},$$
$$v(0) = u_0,$$

which, given the Gårding inequality, has a uniformly coercive operator $A(t) + \alpha\,\mathrm{id}$, that also fulfills assumption **A4** given that $\mathcal{M}$ is a cone. Then $u(t) = e^{\alpha t} v(t)$ solves the equation

$$\begin{aligned} \langle u'(t) + A(t)u(t), w\rangle &= \langle f(t), w\rangle \quad \text{for all } w \in T_{v(t)}\mathcal{M} \cap \mathcal{V}, \\ u(0) &= u_0. \end{aligned}$$

But since $\mathcal{M}$ is a cone, we have $T_{u(t)}\mathcal{M} \cap \mathcal{V} = T_{v(t)}\mathcal{M} \cap \mathcal{V}$, that is, u is indeed a solution of Problem 2.1 for the initial operator $A(t)$. For convenience, we can therefore restrict ourselves to the coercive case.

2.3 Temporal discretization

We discretize in time to show the existence of solutions to Problem 2.1 given assumptions **A1-A4**. An backward Euler method for Problem 2.1 is of the form

$$\left\langle \frac{u_{i+1} - u_i}{t_{i+1} - t_i}, v \right\rangle + a(u_{i+1}, v; t_{i+1}) = \langle f_{i+1}, v\rangle \quad \text{for all } v \in T_{u_{i+1}}\mathcal{M} \cap \mathcal{V} \tag{2.26}$$

with $t_{i+1} > t_i$ and $u_i \in \mathcal{M} \cap \mathcal{V}$. Here, f_{i+1} is the mean value of f on the interval $[t_i, t_{i+1}]$, that is,

$$f_{i+1} = \frac{1}{t_{i+1} - t_i} \int_{t_i}^{t_{i+1}} f(t)\, dt. \tag{2.27}$$

For constant time intervals $t_{i+1} - t_i = \tau$, we can define the piecewise constant interpolation $f_\tau(t) = f_i$ for $t \in ((i-1)\tau, i\tau]$. This is called the zero-order Clément quasi-interpolant of f and converges strongly to f in $L^2(0, T; \mathcal{H})$; see e.g., the footnote to equation (8.58) in [Rou13].

As the test space depends on the solution, this equation appears quite difficult to solve. However, when $a(\cdot, \cdot; t_{i+1})$ is symmetric, (2.26) is the first order optimality condition of the optimization problem

$$u_{i+1} = \operatorname*{arg\,min}_{u \in \overline{\mathcal{M}}^{\mathrm{w}} \cap \mathcal{V}} F_i(u) \tag{2.28}$$

with $F_i(u) := \frac{1}{2(t_{i+1} - t_i)} \|u - u_i\|_{\mathcal{H}}^2 + \frac{1}{2} a(u, u; t_{i+1}) - \langle f_{i+1}, u\rangle$. The existence of a solution to (2.28) and the corresponding optimality condition (2.26) can be shown by basic functional analysis results.

Lemma 2.2. *The optimization problem* (2.28) *has at least one solution.*

Proof. Since F_i is convex and continuous on $\mathcal{V}$ it is also weakly sequentially lower semicontinuous on $\mathcal{V}$; see, e.g., [Zei95, Section 2.5, Lemma 5]. Note that F_i has bounded sublevel sets on $\mathcal{V}$ since the bilinear form $a(\cdot,\cdot;t_{i+1})$ is coercive by assumption. We may therefore restrict the analysis to a sublevel set. Since $\mathcal{V}$ is a Hilbert space, it is a reflexive Banach space. By the Banach-Alaoglu theorem, a weakly closed and bounded subset of a reflexive Banach space is weakly compact and therefore weakly sequentially compact by the Eberlein-Šmulian theorem; see e.g., [Die84, Chapter II and Chapter III]. It now follows that F_i attains a minimum on every weakly sequentially closed subset of $\mathcal{V}$ since the intersection with a weakly closed and bounded set is weakly sequentially compact; see, e.g. [Zei85, Proposition 38.12(d)]. It hence remains to verify that $\overline{\mathcal{M}}^{\mathsf{w}} \cap \mathcal{V}$ is weakly sequentially closed in $\mathcal{V}$. Consider a sequence $(u_n) \subset \overline{\mathcal{M}}^{\mathsf{w}} \cap \mathcal{V}$ converging weakly in $\mathcal{V}$ to $u \in \mathcal{V}$. Obviously, since $\mathcal{H}^* \subset \mathcal{V}^*$, weak convergence in $\mathcal{V}$ implies weak convergence in $\mathcal{H}$, and since $\overline{\mathcal{M}}^{\mathsf{w}}$ is weakly sequentially closed in H, we get $u \in \overline{\mathcal{M}}^{\mathsf{w}} \cap \mathcal{V}$. This shows that this set is weakly sequentially closed in $\mathcal{V}$. □

Lemma 2.3. *Any local minimizer u_{i+1} of F_i on $\mathcal{M} \cap \mathcal{V}$ fulfils (2.26). If u_{i+1} is a local minimizer of F_i on $\overline{\mathcal{M}}^{\mathsf{w}} \cap \mathcal{V}$, then*

$$\left\langle \frac{u_{i+1}-u_i}{t_{i+1}-t_i}, u_{i+1} \right\rangle + a(u_{i+1},u_{i+1};t_{i+1}) = \langle f_{i+1}, u_{i+1} \rangle.$$

Proof. For the first case, let $v \in T_{u_{i+1}}\mathcal{M} \cap \mathcal{V}$. By the assumption **A3**(a), we can find a differentiable curve $\varphi(t)$ defined for $|t|$ small enough such that $\varphi(0) = u_{i+1}$, $\varphi'(0) = v$ and $\varphi(t) \in \mathcal{M} \cap \mathcal{V}$. Then $t \mapsto F_i(\varphi(t))$ has a local minimum at $t = 0$ and the derivative $F_i'(\varphi(0)) \circ \varphi'(0)$ is zero since F_i is continuously differentiable. This yields (2.26). If u_{i+1} is not in $\mathcal{M}$, we still have $\varphi(t) = (1+t)u_{i+1} \in \overline{\mathcal{M}}^{\mathsf{w}} \cap \mathcal{V}$ since $\overline{\mathcal{M}}^{\mathsf{w}} \cap \mathcal{V}$ is a cone. The same argument provides the final equation. □

In the following section, we ensure that given $u_0 \in \mathcal{M} \cap \mathcal{V}$ and small enough time interval $[0, T^*]$ the iterates u_i are in the intersection $\mathcal{M} \cap \mathcal{V}$. We can therefore provide an interval $[0, T^*]$ and timesteps $0 = t_0 < t_1 < \ldots < t_N = T^*$ on which the sequence (u_i) fulfills (2.26).

In the following, we set $\tau = T^*/N$ and $t_i = i\tau$. With $u_0 \in \mathcal{M} \cap \mathcal{V}$ we generate a sequence $(u_i)_{i=0}^N \subset \overline{\mathcal{M}}^{\mathsf{w}} \cap \mathcal{V}$. We construct the piecewise affine linear interpolation $\hat{u}_\tau \colon [0, T^*] \to \mathcal{V}$ and piecewise constant interpolation $\hat{v}_\tau \colon [0, T^*] \to \mathcal{V}$, i.e.,

$$\hat{u}_\tau(t) = \frac{(i+1)\tau - t}{\tau} u_i + \frac{t - i\tau}{\tau} u_{i+1} \quad \text{and} \quad \hat{v}_\tau(t) = u_{i+1}$$

for $t \in (i\tau, (i+1)\tau]$.

2.3.1 Energy estimates

We first provide discrete a priori estimates for the time-discrete solution and its finite differences. These show that $\hat{u}_\tau$ and $\hat{v}_\tau$ are uniformly bounded in $L^\infty(0, T; \mathcal{V})$ and that the derivatives $\hat{v}_\tau'$ are uniformly bounded in $L^2(0, T^*; \mathcal{H})$. In turn, we get a weakly converging subsequence. Its limit is a candidate for a solution to Problem 2.1. The cone property **A1** is crucial.

Lemma 2.4. *The sequence $(u_i)_{i=0}^N \subset \overline{\mathcal{M}}^{\mathsf{w}} \cap \mathcal{V}$ generated by (2.28) with the time step $\tau = T^*/N$ satisfies the estimates*

$$\|u_N\|_{\mathcal{H}}^2 + \sum_{i=1}^N \|u_i - u_{i-1}\|_{\mathcal{H}}^2 + \mu\tau \sum_{i=1}^N \|u_i\|_{\mathcal{V}}^2 \le \|u_0\|_{\mathcal{H}}^2 + C_1 \|f\|_{L^2(0,T^*;\mathcal{H})}^2, \tag{2.29}$$

$$\tau \sum_{i=1}^N \left\| \frac{u_i - u_{i-1}}{\tau} \right\|_{\mathcal{H}}^2 \le C_2\Big(\|u_0\|_{\mathcal{V}}^2 + \|f\|_{L^2(0,T^*;\mathcal{H})}^2\Big), \tag{2.30}$$

$$\|u_i\|_{\mathcal{V}}^2 \le C_3\Big(\|u_0\|_{\mathcal{V}}^2 + \|f\|_{L^2(0,T^*;\mathcal{H})}^2\Big), \qquad i = 1, \dots, N, \tag{2.31}$$

where $C_1, C_2, C_3 > 0$ depend on β, μ, L, and on the constant for the continuity of the embedding $\mathcal{V} \hookrightarrow \mathcal{H}$ in (2.22). As a result, $\hat{u}_\tau$ and $\hat{v}_\tau$ are bounded in $L^\infty(0, T^*; \mathcal{V})$, and $\hat{v}_\tau'$ is bounded in $L^2(0, T^*; \mathcal{H})$, uniformly for $\tau \to 0$.

Proof. We use coercivity of a, the optimality condition in Lemma 2.3, and the identity

$$\langle u_{i+1} - u_i, u_{i+1}\rangle = \langle u_{i+1} - u_i, u_{i+1}\rangle_{\mathcal{H}} = \frac{1}{2}\big(\|u_{i+1}\|_{\mathcal{H}}^2 - \|u_i\|_{\mathcal{H}}^2 + \|u_{i+1} - u_i\|_{\mathcal{H}}^2\big),$$

to get

$$\|u_{i+1}\|_{\mathcal{H}}^2 - \|u_i\|_{\mathcal{H}}^2 + \|u_{i+1} - u_i\|_{\mathcal{H}}^2 + 2\tau\mu\|u_{i+1}\|_{\mathcal{V}}^2 \le 2\tau\langle f_{i+1}, u_{i+1}\rangle.$$

We use Young's inequality to attain $2\tau\langle f_{i+1}, u_{i+1}\rangle \le \tau/\mu\|f_{i+1}\|_{\mathcal{V}^*}^2 + \mu\tau\|u_{i+1}\|_{\mathcal{V}}^2$ and hence

$$\|u_{i+1}\|_{\mathcal{H}}^2 - \|u_i\|_{\mathcal{H}}^2 + \|u_{i+1} - u_i\|_{\mathcal{H}}^2 + \tau\mu\|u_{i+1}\|_{\mathcal{V}}^2 \le \frac{\tau}{\mu}\|f_{i+1}\|_{\mathcal{V}^*}^2.$$

The constant of continuity $\mathcal{H} \cong \mathcal{H}^* \hookrightarrow \mathcal{V}^*$ depends only on the constant of continuity $\mathcal{V} \hookrightarrow \mathcal{H}$ in (2.22). We therefore get

$$\|f_{i+1}\|_{\mathcal{V}^*}^2 \le C_1\|f_{i+1}\|_{\mathcal{H}}^2 = C_1 \left\| \frac{1}{\tau} \int_{t_i}^{t_{i+1}} 1 \cdot f(t)\, dt \right\|_{\mathcal{H}}^2 \le \frac{C_1}{\tau} \int_{t_i}^{t_{i+1}} \|f(t)\|_{\mathcal{H}}^2\, dt \tag{2.32}$$

by the Cauchy-Schwarz inequality and the definition of f_i in (2.27). Summation over the index i results in (2.29).

Next we prove (2.30). Since u_{i+1} minimizes F_i, its value is smaller than $F_i(u_i)$, i.e.,

$$\begin{aligned} 2F_i(u_{i+1}) &= \frac{1}{\tau}\|u_{i+1} - u_i\|_{\mathcal{H}}^2 + a(u_{i+1}, u_{i+1}; t_{i+1}) - 2\langle f_{i+1}, u_{i+1}\rangle \\ &\le a(u_i, u_i; t_{i+1}) - 2\langle f_{i+1}, u_i\rangle = 2F_i(u_i). \end{aligned}$$

Using Young's inequality, it follows that

$$\begin{aligned}\tau\left\|\frac{u_{i+1}-u_i}{\tau}\right\|_{\mathcal{H}}^2 &\le a(u_i,u_i;t_{i+1})-a(u_{i+1},u_{i+1};t_{i+1})+2\tau\left\langle f_{i+1},\frac{u_{i+1}-u_i}{\tau}\right\rangle\\ &\le a(u_i,u_i;t_{i+1})-a(u_{i+1},u_{i+1};t_{i+1})+2\tau\|f_{i+1}\|_{\mathcal{H}}^2+\frac{\tau}{2}\left\|\frac{u_{i+1}-u_i}{\tau}\right\|_{\mathcal{H}}^2.\end{aligned}$$

This yields

$$\tau\left\|\frac{u_{i+1}-u_i}{\tau}\right\|_{\mathcal{H}}^2 \le 2a(u_i,u_i;t_{i+1})-2a(u_{i+1},u_{i+1};t_{i+1})+4\tau\|f_{i+1}\|_{\mathcal{H}}^2. \tag{2.33}$$

We sum over i and get

$$\begin{aligned}\tau\sum_{i=1}^N\left\|\frac{u_i-u_{i-1}}{\tau}\right\|_{\mathcal{H}}^2 \le 2a(u_0,u_0;0)+2\sum_{i=1}^N\big(a(u_{i-1},u_{i-1};t_i)-a(u_{i-1},u_{i-1};t_{i-1})\big)&\\ -2a(u_N,u_N;T^*)+4\tau\sum_{i=1}^N\|f_i\|_{\mathcal{H}}^2.&\end{aligned}$$

Using the Lipschitz continuity (2.24) in t of the bilinear form then allows for the estimate

$$\begin{aligned}\tau\sum_{i=1}^N\left\|\frac{u_i-u_{i-1}}{\tau}\right\|_{\mathcal{H}}^2 &\le 2\beta\|u_0\|_{\mathcal{V}}^2+2\beta L\tau\sum_{i=1}^N\|u_{i-1}\|_{\mathcal{V}}^2+4\tau\sum_{i=1}^N\|f_i\|_{\mathcal{H}}^2\\ &\le 2\beta(1+L\tau)\|u_0\|_{\mathcal{V}}^2+2\beta L\tau\sum_{i=1}^N\|u_i\|_{\mathcal{V}}^2+4\tau\sum_{i=1}^N\|f_i\|_{\mathcal{H}}^2.\end{aligned}$$

We may use (2.29) to get

$$\tau\sum_{i=1}^N\left\|\frac{u_i-u_{i-1}}{\tau}\right\|_{\mathcal{H}}^2 \le 2\beta(1+L\tau)\|u_0\|_{\mathcal{V}}^2+4\tau\sum_{i=1}^N\|f_i\|_{\mathcal{H}}^2+\frac{2\beta L}{\mu}\Big(\|u_0\|_{\mathcal{H}}^2+\frac{\tau}{\mu}\sum_{i=1}^N\|f_i\|_{\mathcal{V}^*}^2\Big).$$

The desired estimate (2.30) follows with (2.32) and $\|u_0\|_{\mathcal{H}}^2 \lesssim \|u_0\|_{\mathcal{V}}^2$.

For the last estimate (2.31) we start with (2.33). We readily obtain

$$0\le a(u_{j-1},u_{j-1};t_j)-a(u_j,u_j;t_j)+2\tau\|f_j\|_{\mathcal{H}}^2.$$

We sum over $j=1,\dots,i$ and rearrange to obtain

$$a(u_i,u_i;t_i)\le a(u_0,u_0;0)+\sum_{j=1}^i\big(a(u_{j-1},u_{j-1};t_j)-a(u_{j-1},u_{j-1};t_{j-1})\big)+2\tau\sum_{j=1}^i\|f_j\|_{\mathcal{H}}^2.$$

This implies

$$\begin{aligned}\mu\|u_i\|_{\mathcal{V}}^2 &\leq \beta\|u_0\|_{\mathcal{V}}^2 + \beta L\tau\sum_{j=1}^{i}\|u_{j-1}\|_{\mathcal{V}}^2 + 2\tau\sum_{j=1}^{i}\|f_j\|_{\mathcal{H}}^2 \\ &\leq \beta(1+L\tau)\|u_0\|_{\mathcal{V}}^2 + \beta L\tau\sum_{j=1}^{N}\|u_j\|_{\mathcal{V}}^2 + 2\tau\sum_{j=1}^{N}\|f_j\|_{\mathcal{H}}^2\end{aligned}$$

for any $i = 1, \dots, N$. Using (2.29) and (2.32) yields (2.31). □

We can now assure that the iterates u_i lie in $\mathcal{M}$ for $i = 1, \dots, N$ if T^* is sufficiently small. Let $u_0 \in \mathcal{V}$ have positive $\mathcal{H}$-distance σ to the relative boundary $\overline{\mathcal{M}}^{\mathsf{w}} \setminus \mathcal{M}$ and let $c = C_2(\|u_0\|_{\mathcal{V}}^2 + \|f\|_{L^2(0,T;\mathcal{H})}^2)$ be the right-hand side of (2.30) from Lemma 2.4 with $T^* = T$. Then

$$\|u_i - u_0\|_{\mathcal{H}}^2 \leq \left(\sum_{j=1}^{i}\tau\left\|\frac{u_j - u_{j-1}}{\tau}\right\|_{\mathcal{H}}\right)^2 \leq i\tau^2\sum_{j=1}^{N}\left\|\frac{u_j - u_{i-j}}{\tau}\right\|_{\mathcal{H}}^2 \leq T^* c \qquad (2.34)$$

and hence $\|u_i - u_0\|_{\mathcal{H}}^2 \leq \sigma^2$ for all i whenever $T^* < \sigma^2/c$. Choosing a small enough T^* therefore ensures $u_i \in \mathcal{M}$ for all $i = 1, \dots, N$.

2.4 Existence and uniqueness of solutions

We are now in the position to show the existence of a solution to Problem 2.1. Recall that $\hat{u}_\tau$ and $\hat{v}_\tau$ are the piecewise affine linear and piecewise constant interpolations of the sequence (u_i) generated by the implicit Euler method (2.26). We show via compactness arguments that the interpolations $\hat{u}_\tau$ and $\hat{v}_\tau$ have a common weak limit $\hat{u}$ in $L^2(0, T^*, \mathcal{V})$. The main difficulty is confirming (2.25) for the limit $\hat{u}$ of the time-discrete solutions since the tangent spaces at $\hat{u}(t)$ differ from the ones at $\hat{v}_\tau(t)$ and we therefore have to use the Lipschitz continuity of the tangent space projection in Assumption **A2**.

Theorem 2.5. *Let $a(u, v; t)$ define a bounded and coercive bilinear form in u and v, uniformly with respect to t, and let a be Lipschitz continuous with respect to t. Furthermore, let the assumptions stated in Problem 2.1 and* **A1**-**A4** *hold true and let $\hat{u}_\tau$ and $\hat{v}_\tau$ be the piecewise affine linear and piecewise constant interpolations given by*

$$\hat{u}_\tau(t) = \frac{(i+1)\tau - t}{\tau}u_i + \frac{t - i\tau}{\tau}u_{i+1} \quad \textit{and} \quad \hat{v}_\tau(t) = u_{i+1}$$

for $t \in (i\tau, (i+1)\tau]$.

(a) *The functions $\hat{u}_\tau$ and $\hat{v}_\tau$ converge, up to subsequences, weakly in $L^2(0,T^*;\mathcal{V})$ and strongly in $L^2(0,T^*;\mathcal{H})$, to the same function $\hat{u} \in L^\infty(0,T^*;\mathcal{V}) \cap W(0,T^*;\mathcal{V},\mathcal{H})$ with $\hat{u}(0) = u_0$, while the weak derivatives $\hat{u}_\tau'$ converge weakly to $\hat{u}'$ in $L^2(0,T^*;\mathcal{H})$, again up to subsequences. The functions $\hat{u}_\tau$ converge, up to subsequences, strongly in $C(0,T;\mathcal{H})$ to $\hat{u}$. Furthermore, the function values $\hat{u}(t)$ lie in $\overline{\mathcal{M}}^{\mathsf{w}} \cap \mathcal{V}$ for almost all $t \in [0,T^*]$.*

(b) *Let u_0 have positive $\mathcal{H}$-distance σ to the relative boundary $\overline{\mathcal{M}}^{\mathsf{w}} \setminus \mathcal{M}$. Then there exists a constant $c > 0$ independent of σ such that $\hat{u}$ solves Problem 2.1 on the time interval $[0,T^*]$ when $T^* < \sigma^2/c$.*

Proof of Theorem 2.5 (a). It follows from (2.29) and (2.30) in Lemma 2.4 that $\hat{u}_\tau$ and $\hat{v}_\tau$ are uniformly bounded in $L^2(0,T^*;\mathcal{V})$. Therefore, refinement in time generates sequences which converge weakly in $L^2(0,T^*;\mathcal{V})$, up to subsequence, i.e.,

$$\hat{u}_\tau \rightharpoonup \hat{u} \quad \text{and} \quad \hat{v}_\tau \rightharpoonup \hat{v} \quad \text{in } L^2(0,T^*;\mathcal{V}).$$

In particular, $\hat{u}_\tau - \hat{v}_\tau$ converges weakly in $L^2(0,T^*;\mathcal{H})$ to $\hat{u} - \hat{v}$. Comparing the two sequences in $L^2(0,T^*;\mathcal{H})$, we get

$$\begin{aligned}\int_0^{T^*} \|\hat{u}_\tau - \hat{v}_\tau\|_{\mathcal{H}}^2 \, dt &= \sum_{i=1}^N \int_{t_{i-1}}^{t_i} \|\hat{u}_\tau - \hat{v}_\tau\|_{\mathcal{H}}^2 \, dt \\ &= \tau \sum_{i=1}^N \int_0^1 \|(s-1)(u_i - u_{i-1})\|_{\mathcal{H}}^2 \, ds \\ &= \frac{\tau}{3} \sum_{i=1}^N \|u_i - u_{i-1}\|_{\mathcal{H}}^2,\end{aligned}$$

and together with (2.30) in Lemma 2.4 this results in

$$\int_0^{T^*} \|\hat{u}_\tau - \hat{v}_\tau\|_{\mathcal{H}}^2 \, dt \le \frac{C_2 \tau^2}{3} \Big(\|u_0\|_{\mathcal{V}}^2 + \|f\|_{L_2(0,T^*;\mathcal{H})}^2 \Big), \tag{2.35}$$

which tends to zero as $\tau \to 0$. We conclude $\hat{u} = \hat{v}$.

Likewise, $\hat{u}_\tau'$ is uniformly bounded in $L^2(0,T^*;\mathcal{H})$ and thus, up to subsequences, $\hat{u}_\tau' \rightharpoonup \hat{w}$ for some $\hat{w} \in L^2(0,T^*;\mathcal{H})$. We next show that $\hat{w}$ is the weak derivative of $\hat{u}$. For this, we need to verify that

$$\int_0^{T^*} \langle \hat{w}(t), v \rangle \, \phi(t) + \langle \hat{u}(t), v \rangle \, \phi'(t) \, dt = 0$$

for arbitrary $v \in \mathcal{V}$ and $\phi \in C_0^\infty(0,T^*)$. Adding and subtracting the weak derivative of

$\hat{u}_\tau$, we get

$$\int_{T^*} \langle \hat{w}(t), v\rangle\, \phi(t) + \langle \hat{u}(t), v\rangle\, \phi'(t)\, dt \\ = \int_{T^*} \langle \hat{w}(t) - \hat{u}_\tau'(t), v\rangle\, \phi(t) + \langle \hat{u}(t) - \hat{u}_\tau(t), v\rangle\, \phi'(t)\, dt.$$

Since $\hat{u}_\tau \rightharpoonup \hat{u}$ and $\hat{u}_\tau' \rightharpoonup \hat{w}$ in $L^2(0,T^*;\mathcal{V})$ and $L^2(0,T^*;\mathcal{H})$, respectively, and since $v\phi$, $v\phi' \in L^2(0,T^*;\mathcal{V})$, the right hand side converges to zero. Thus, $\hat{w} = \hat{u}'$.

The strong convergence of $\hat{u}_\tau$ in $L^2(0,T^*;\mathcal{H})$ follows from the theorem of Aubin and Lions; see e.g., [Sho97, Proposition III.1.3]. It states that when $\mathcal{V}$ is compactly embedded into $\mathcal{H}$, then the space $W(0,T^*;\mathcal{V},\mathcal{H})$ is compactly embedded into $L^2(0,T^*;\mathcal{H})$. Thereby, the weak convergence of $\hat{u}_\tau$ and $\hat{u}_\tau'$ that we just have proved implies the strong convergence of a subsequence $\hat{u}_\tau \to \hat{u}$ in $L^2(0,T^*;\mathcal{H})$. This together with (2.35) directly proves that also $\hat{u}_\tau \to \hat{v}$ in $L^2(0,T^*;\mathcal{H})$. By (2.31) the sequences $(\hat{u}_\tau)$ and $(\hat{v}_\tau)$ are also bounded in $L^\infty(0,T^*;\mathcal{V})$ and hence converge with respect to the weak* topology in $L^\infty(0,T^*;\mathcal{V})$, again up to subsequence. Since the $L^\infty(0,T^*;\mathcal{V})$-norm is weakly* sequentially semicontinuous, we even obtain $\hat{u} \in L^\infty(0,T^*;\mathcal{V})$.

Similarly, the strong convergence of $\hat{u}_\tau$ in $C(0,T^*;\mathcal{H})$ follows from another version of the Aubin-Lions theorem; see e.g., [Sim87, Section 8]. This version states when $\mathcal{V}$ is compactly embedded in $\mathcal{H}$, then a set that is bounded in $L^\infty(0,T^*;\mathcal{V})$ with derivatives bounded in $L^2(0,T^*;\mathcal{H})$ is relatively compact in $C(0,T^*;\mathcal{H})$. Hence, strong convergence follows from the estimates (2.30) and (2.31) in Lemma 2.4. This readily implies $\hat{u}(0) = \lim_{\tau\searrow 0} \hat{u}_\tau(0) = u_0$.

It remains to show that $\hat{u}(t) \in \overline{\mathcal{M}}^{\mathsf{w}}$ for almost all $t \in (0,T^*)$. We already proved $\underline{u(t)} = \lim_{\tau\searrow 0} \hat{u}_\tau(t)$. Therefore, it is sufficient to show that the $\mathcal{H}$-distance of $\hat{u}_\tau(t)$ from $\overline{\mathcal{M}}^{\mathsf{w}}$ converges to zero. Indeed, since $\hat{u}_\tau$ is an affine linear interpolation of the sequence $(u_i) \subset \overline{\mathcal{M}}^{\mathsf{w}}$, the estimate (2.30) implies

$$\|u_i - \hat{u}_\tau(t)\|_{\mathcal{H}}^2 \le \|u_i - u_{i-1}\|_{\mathcal{H}}^2 \le \tau C_2 \left(\|u_0\|_{\mathcal{V}}^2 + \|f\|_{L^2(0,T^*;\mathcal{H})}^2\right)$$

for $t \in ((i-1)\tau, i\tau]$. Hence, the distance of $\hat{u}_\tau(t)$ to $\overline{\mathcal{M}}^{\mathsf{w}}$ converges uniformly to zero. □

Proof of Theorem 2.5 (b). We provided a condition such that $(u_i) \subset \mathcal{M}' \subset \mathcal{M}$ lies in a weakly compact subset independent of the timestep τ in (2.34). Hence, there is a weakly compact subset $\mathcal{M}' \subset \mathcal{M}$ such that $\hat{u}(t), \hat{v}_\tau(t) \in \mathcal{M}'$ and the optimality condition (2.26) holds due to Lemma 2.3. In terms of $\hat{u}_\tau$ and $\hat{v}_\tau$ this may be written as

$$\langle \hat{u}_\tau'(t), w(t)\rangle + \langle A_\tau(t)\hat{v}_\tau(t), w(t)\rangle = \langle f_\tau(t), w(t)\rangle \quad \text{for } w(t) \in T_{\hat{v}_\tau(t)}\mathcal{M} \cap \mathcal{V},$$

where $A_\tau(t) = A(i\tau)$ and $f_\tau(t) = f_i$ for $t \in ((i-1)\tau, i\tau]$ are piecewise constant interpolations. It follows that

$$\int_0^{T^*} \langle \hat{u}_\tau'(t), w(t)\rangle + \langle A_\tau(t)\hat{v}_\tau(t), w(t)\rangle - \langle f_\tau(t), w(t)\rangle dt = 0 \tag{2.36}$$

for $w(t) \in T_{\hat{v}_\tau(t)}\mathcal{M} \cap \mathcal{V}$ almost everywhere. We will show

$$\int_0^{T^*} \langle \hat{u}'(t), w(t)\rangle + \langle A(t)\hat{u}(t), w(t)\rangle - \langle f(t), w(t)\rangle dt = 0 \tag{2.37}$$

for $w \in L^\infty(0, T^*; \mathcal{V})$ and $w(t) \in T_{\hat{u}(t)}\mathcal{M}$ almost everywhere. This implies (2.25) as desired, since in the opposite case there would be a subset $S \subseteq [0, T^*]$ of positive measure such that for all $t \in S$ we have

$$\langle \hat{u}'(t), w(t)\rangle + \langle A(t)\hat{u}(t), w(t)\rangle - \langle f(t), w(t)\rangle \neq 0$$

for some $w(t) \in T_{\hat{u}(t)}\mathcal{M} \cap \mathcal{V}$. By appropriately scaling these $w(t)$, we can then choose $w \in L^\infty(0, T^*; \mathcal{V})$ such that $w(t) \in T_{\hat{u}(t)}\mathcal{M} \cap \mathcal{V}$ almost everywhere and the left-hand side in (2.37) is positive.

Now let $w \in L^\infty(0, T^*; \mathcal{V})$ be given such that $w(t) \in T_{\hat{u}(t)}\mathcal{M}$ almost everywhere. Assumption **A3**(b) implies $P_{\hat{v}_\tau(t)}w(t) \in T_{\hat{v}_\tau(t)}\mathcal{M} \cap \mathcal{V}$ almost everywhere and hence

$$\int_0^{T^*} \langle \hat{u}_\tau'(t), P_{\hat{v}_\tau(t)}w(t)\rangle + \langle A_\tau(t)\hat{v}_\tau(t), P_{\hat{v}_\tau(t)}w(t)\rangle - \langle f_\tau(t), P_{\hat{v}_\tau(t)}w(t)\rangle dt = 0 \tag{2.38}$$

because of (2.36). We have

$$\int_0^{T^*} \|P_{\hat{v}_\tau(t)}w(t) - w(t)\|_{\mathcal{H}}^2 \, dt \le \int_0^{T^*} \|P_{\hat{v}_\tau(t)} - P_{\hat{u}(t)}\|_{\mathcal{H}\to\mathcal{H}}^2 \|w(t)\|_{\mathcal{H}}^2 \, dt$$

and hence Assumption **A2** implies that $P_{v_\tau}w$ converges strongly to w in $L^2(0, T^*; \mathcal{H})$. Since f_τ converges strongly to f and $\hat{u}_\tau'$ converges weakly to $\hat{u}'$ in $L^2(0, T^*; \mathcal{H})$, we have

$$\int_0^{T^*} \langle f_\tau(t), P_{\hat{v}_\tau(t)}w(t)\rangle \, dt \to \int_0^{T^*} \langle f(t), w(t)\rangle \, dt$$

and

$$\int_0^{T^*} \langle \hat{u}_\tau'(t), P_{\hat{v}_\tau(t)}w(t)\rangle \, dt \to \int_0^{T^*} \langle \hat{u}'(t), w(t)\rangle \, dt$$

see e.g., [Zei90a, Proposition 21.23(j)]. It remains to show

$$\int_0^{T^*} \langle A_\tau(t)\hat{v}_\tau(t), P_{\hat{v}_\tau(t)}w(t)\rangle \, dt \to \int_0^{T^*} \langle A(t)\hat{u}(t), w(t)\rangle \, dt. \tag{2.39}$$

We will proceed pointwise. We have

$$\begin{aligned} &\big|\langle A_\tau(t)\hat{v}_\tau(t), P_{\hat{v}_\tau(t)}w(t)\rangle - \langle A(t)\hat{u}(t), w(t)\rangle\big| \\ &\qquad \le \big|\langle A_\tau(t)\hat{v}_\tau(t), P_{\hat{v}_\tau(t)}w(t) - w(t)\rangle\big| + \big|\langle \big(A_\tau(t) - A(t)\big)\hat{v}_\tau(t), w(t)\rangle\big| \\ &\qquad\qquad + \big|\langle A(t)\big(\hat{v}_\tau(t) - \hat{u}(t)\big), w(t)\rangle\big|. \end{aligned}$$

For the second summand, the Lipschitz continuity of A implies

$$\left|\langle\big(A_\tau(t) - A(t)\big)\hat{v}_\tau(t), w(t)\rangle\right| \leq \tau L\beta \|v(t)\|_{\mathcal{V}}\|w(t)\|_{\mathcal{V}} \to 0 \quad \text{as} \quad \tau \to 0.$$

For the third summand, the weak convergence in $L^2(0,T^*;\mathcal{V})$ implies that

$$\int_0^{T^*} \left|\langle A(t)\big(\hat{v}_\tau(t) - \hat{u}(t)\big), w(t)\rangle\right| dt \to 0 \quad \text{as} \quad \tau \to 0$$

and therefore

$$\langle A(t)\big(\hat{v}_\tau(t) - \hat{u}(t)\big), w(t)\rangle \to 0 \quad \text{as} \quad \tau \to 0$$

for almost every t, possibly after passing to a subsequence; see e.g., [PW18, Corollary 2.3.20]. We have the integrable bound

$$\left|\langle A(t)\big(\hat{v}_\tau(t) - \hat{u}(t)\big), w(t)\rangle\right| \leq \beta\left(\|\hat{v}_\tau(t)\|_{\mathcal{V}} + \|\hat{u}(t)\|_{\mathcal{V}}\right)\|w(t)\|_{\mathcal{V}}.$$

For the first part, we denote $A_{1,\tau}(t) = A_1(i\tau)$ and $A_{2,\tau}(t) = A_2(i\tau)$ for $t \in ((i-1)\tau, i\tau]$ and use Assumption **A2** and Assumption **A4**. We get

$$\begin{aligned}\left|\langle A_\tau(t)\hat{v}_\tau(t), P_{\hat{v}_\tau(t)}w(t) - w(t)\rangle\right| &= \left|\langle A_{1,\tau}(t)\hat{v}_\tau(t) + A_{2,\tau}(t)\hat{v}_\tau(t), P_{\hat{v}_\tau(t)}w(t) - w(t)\rangle\right| \\ &\leq 0 + \gamma\|\hat{v}_\tau(t)\|_{\mathcal{V}}^{\eta}\|w(t)\|_{\mathcal{H}}\|\hat{u}(t) - \hat{v}_\tau(t)\|_{\mathcal{H}},\end{aligned}$$

where the part concerning $A_{1,\tau}$ vanishes due to Assumption **A4**(a) and $\gamma = \gamma(\mathcal{M}')$ and η are taken from Assumption **A4**(b). Due to the strong convergence of $\hat{v}_\tau$ to $\hat{u}$ in $L^2(0,T^*;\mathcal{H})$, this expression converges to zero for almost every t, again after passing to a subsequence. We also provided integrable bounds for every summand, hence (2.39) follows by dominated convergence. Therefore, the left-hand side in (2.38) converges to the left-hand side in (2.37) and the assertion follows as described. □

Next, we turn our attention to the stability of this problem. For this, let $v \in W(0,T^*;\mathcal{V},\mathcal{H})$ satisfy

$$\begin{aligned} v(t) &\in \mathcal{M}, \\ \langle v'(t), w\rangle + a(v(t), w; t) &= \langle g(t), w\rangle \quad \text{for all } w \in T_{v(t)}\mathcal{M} \cap \mathcal{V}, \\ v(0) &= v_0. \end{aligned} \tag{2.40}$$

for almost every t. We give bounds to pointwise distance $\|u(t) - v(t)\|_{\mathcal{H}}$ for a solution u of Problem 2.1 under the condition, that u and v are in $L^\eta(0,T;\mathcal{V})$. This is well defined since $W(0,T^*;\mathcal{V},\mathcal{H})$ is continuously embedded into $C(0,T^*;\mathcal{H})$; see e.g., [Zei90a, Proposition 23.23]. Note that solutions obtained from the time-stepping scheme are in $L^\infty(0,T;\mathcal{V}) \subset L^\eta(0,T;\mathcal{V})$. In the model problem of Section 2.1 and there is no further restriction since $\eta = 2$ is fulfilled. In the case when $\mathcal{M}$ is a linear space, then the difference $u - v$ satisfies a similar equation to (2.25) and (2.40) with the right-hand side replaced with $f-g$, and we can evaluate at $w = u(t) - v(t)$ to get classical stability estimates for linear parabolic problems. We follow a similar idea but since $u - v$ does not satisfy an appropriate equation and $u(t) - v(t)$ does not lie in the tangent space $T_{u(t)}\mathcal{M}$, we have to resort to Assumption **A2**. This results in a cruder estimate.

Theorem 2.6. *Let u be a solution of Problem 2.1 and v be a solution to* (2.40) *in the time interval $[0,T^*]$. Assume that the continuous representatives $u, v \in C(0,T^*;\mathcal{H})$ have values in a weakly compact subset $\mathcal{M}' \subset \mathcal{M}$. Moreover, assume that $u, v \in L^\eta(0,T^*;\mathcal{V})$ where η is from Assumption* **A4**(b). *Then*

$$\|u(t)-v(t)\|_{\mathcal{H}}^2 \leq \left(\|u_0 - v_0\|_{\mathcal{H}}^2 + \frac{1}{c}\int_0^t \|f(s)-g(s)\|_{\mathcal{H}}^2\,ds\right)\exp(\Lambda(t)+ct),$$

for any $c > 0$ and

$$\Lambda(t) := 2\kappa \int_0^t \|u'(s)\|_{\mathcal{H}} + \|v'(s)\|_{\mathcal{H}} + \gamma\left(\|u(s)\|_{\mathcal{V}}^\eta + \|v(s)\|_{\mathcal{V}}^\eta\right) + \|f(s)\|_{\mathcal{H}} + \|g(s)\|_{\mathcal{H}}\,ds,$$

where $\kappa = \kappa(\mathcal{M}')$ is from Assumption **A4**(b) *and*

$$\begin{aligned}0 \leq \Lambda(t) \leq 2\kappa\Big(&\|u'\|_{L^1(0,T^*;\mathcal{H})} + \|v'\|_{L^1(0,T^*;\mathcal{H})} + \gamma\left(\|u\|_{L^\eta(0,T^*;\mathcal{V})}^\eta + \|v\|_{L^\eta(0,T^*;\mathcal{V})}^\eta\right)\\ &+ \|f(s)\|_{L^1(0,T^*;\mathcal{H})} + \|g(s)\|_{L^1(0,T^*;\mathcal{H})}\Big) < \infty\end{aligned}$$

for all $t \in [0,T^*]$.

Proof. We use integration by parts in the sense of [Zei90a, Proposition 23.23(iv)]. This results in

$$\frac{1}{2}\frac{d}{dt}\|u(t)-v(t)\|_{\mathcal{H}}^2 \leq \langle u'(t)-v'(t)+A(t)(u(t)-v(t))-f(t)+g(t)+f(t)-g(t), u(t)-v(t)\rangle$$

for almost all $t \in [0,T^*]$ by coercivity of $A(t)$ and adding and subtracting $\langle f(t) - g(t), u(t)-v(t)\rangle$. We add and subtract (2.40) and (2.25) with $w = P_{v(t)}(u(t)-v(t))$ and $w = P_{u(t)}(u(t)-v(t))$, respectively. This results in

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\|u(t)-v(t)\|_{\mathcal{H}}^2\\ &\quad\leq \langle f(t)-g(t), u(t)-v(t)\rangle + \langle u'(t)+A(t)u(t)-f(t), (\mathrm{id}-P_{u(t)})(u(t)-v(t))\rangle\\ &\qquad - \langle v'(t)+A(t)v(t)-g(t), (\mathrm{id}-P_{v(t)})(u(t)-v(t))\rangle.\end{aligned}$$

We use Young's inequality to estimate

$$\langle f(t)-g(t), u(t)-v(t)\rangle \leq \frac{1}{2c}\|f(t)-g(t)\|_{\mathcal{H}}^2 + \frac{c}{2}\|u(t)-v(t)\|_{\mathcal{H}}^2$$

and Assumption **A4** to get

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}&\|u(t)-v(t)\|_{\mathcal{H}}^2\\ &\le \big(\|u'(t)\|_{\mathcal{H}}+\gamma\|u(t)\|_{\mathcal{V}}^{\eta}+\|f(t)\|_{\mathcal{H}}\big)\,\|(\operatorname{id}-P_{u(t)})(u(t)-v(t))\|_{\mathcal{H}}\\ &\quad+\big(\|v'(t)\|_{\mathcal{H}}+\gamma\|v(t)\|_{\mathcal{V}}^{\eta}+\|g(t)\|_{\mathcal{H}}\big)\,\|(\operatorname{id}-P_{v(t)})(u(t)-v(t))\|_{\mathcal{H}}\\ &\qquad+\frac{1}{2c}\|f(t)-g(t)\|_{\mathcal{H}}^2+\frac{c}{2}\|u(t)-v(t)\|_{\mathcal{H}}^2.\end{aligned}$$

Finally, Assumption **A2** implies

$$\begin{aligned}\frac{d}{dt}&\|u(t)-v(t)\|_{\mathcal{H}}^2\\ &\le \Big(2\kappa\left(\|u'(t)\|_{\mathcal{H}}+\|v'(t)\|_{\mathcal{H}}+\gamma\left(\|u(t)\|_{\mathcal{V}}^{\eta}+\|u(t)\|_{\mathcal{V}}^{\eta}\right)+\|f(t)\|_{\mathcal{H}}+\|g(t)\|_{\mathcal{H}}\right)+c\Big)\\ &\qquad\cdot\|u(t)-v(t)\|_{\mathcal{H}}^2+\frac{1}{c}\|f(t)-g(t)\|_{\mathcal{H}}^2\end{aligned}$$

and the results follows from Grönwall's lemma; see e.g. [Tes12, Lemma 2.7]. Here, we take into account that $L^2(0,T^*;\mathcal{H})\subset L^1(0,T^*;\mathcal{H})$. □

This stability estimate is meaningful since solutions coming from the time stepping scheme satisfy the integrability assumptions and lie in a compact subset $\mathcal{M}'\subset\mathcal{M}$ for small enough T^* due to (2.34).

Combining Theorem 2.5 with a continuation argument and invoking Theorem 2.6, we obtain a unique solution on a maximal time interval.

Theorem 2.7. *Let the assumptions stated in Section 2.2 hold and let u_0 have positive $\mathcal{H}$-distance from $\overline{\mathcal{M}}^{\mathsf{w}}\setminus\mathcal{M}$. There exists $T^*\in(0,T]$ and $u\in W(0,T^*;\mathcal{V},\mathcal{H})\cap L^\infty(0,T^*;\mathcal{V})$ such that u solves Problem 2.1 on the time interval $[0,T^*]$, and its continuous representative $u\in C(0,T^*;\mathcal{H})$ satisfies $u(t)\in\mathcal{M}$ for all $t\in[0,T^*)$. Here, T^* is maximal for the evolution on $\mathcal{M}$ in the sense that if $T^*<T$, then*

$$\liminf_{t\to T^*}\inf_{v\in\overline{\mathcal{M}}^{\mathsf{w}}\setminus\mathcal{M}}\|u(t)-v\|_{\mathcal{H}}=0.$$

In either case, u is the unique solution of Problem 2.1 in $W(0,T^*;\mathcal{V},\mathcal{H})\cap L^\eta(0,T^*;\mathcal{V})$. The solution satisfies

$$\|u\|_{L^2(0,T^*;\mathcal{V})}^2\le\|u_0\|_{\mathcal{H}}^2+C_1\|f\|_{L^2(0,T^*;\mathcal{H})}^2, \tag{2.41}$$

$$\|u'\|_{L^2(0,T^*;\mathcal{H})}^2\le C_2\left(\|u_0\|_{\mathcal{V}}^2+\|f\|_{L^2(0,T^*;\mathcal{H})}^2\right), \tag{2.42}$$

$$\|u\|_{L^\infty(0,T^*;\mathcal{V})}^2\le C_3\left(\|u_0\|_{\mathcal{V}}^2+\|f\|_{L^2(0,T^*;\mathcal{H})}^2\right), \tag{2.43}$$

where C_1, C_2, and C_3 are the constants from Lemma 2.4.

Proof. Let u, v be two solutions of Problem 2.1 in $W(0,T^*;\mathcal{V},\mathcal{H}) \cap L^\eta(0,T^*;\mathcal{V})$ and consider them as their continuous representatives in $C(0,T^*;\mathcal{H})$. Theorem 2.6 implies $u(t) = v(t)$ for every t in a compact subinterval of $[0,T^*)$. Since $[0,T^*)$ is the union of its compact sub intervals, we have $u(t) = v(t)$ for every $t \in [0,T^*)$. Hence, u is the unique solution of Problem 2.1 in $W(0,T^*;\mathcal{V},\mathcal{H}) \cap L^\eta(0,T^*;\mathcal{V})$.

Theorem 2.5 provides us with a solution u of Problem 2.1 on a time interval $[0,T_1]$ with $0 < T_1 \leq T$ such that $u \in L^\infty(0,T_1;\mathcal{V})$ and either $T_1 = T$ or $T_1 < \sigma_0^2/c$ where σ_0 is the $\mathcal{H}$-distance of u_0 from $\overline{\mathcal{M}}^{\mathsf{w}} \setminus \mathcal{M}$ and $c > 0$ is the constant from Theorem 2.5(b). In the latter case, we may assume without loss of generality that $u \in C(0,T_1;\mathcal{H})$ and $u(T_1) \in \mathcal{M} \cap \mathcal{V}$. Let σ_1 be the $\mathcal{H}$-distance of u_1 from $\overline{\mathcal{M}}^{\mathsf{w}} \setminus \mathcal{M}$. If $T_1 < T$, applying again Theorem 2.5 on $[T_1,T]$ with starting value $u_0 = u(T_1)$, we obtain a continuation of u to an interval $[0,T_2]$ with either $T_2 = T$ or $T_2 < T_1 + \sigma_1^2/c$. In the latter case, we can again assume $u \in C(0,T_2;\mathcal{H})$ and $u(T_2) \in \mathcal{V}$ with corresponding distance $\sigma_2 > 0$. We thus inductively obtain sequences (T_i) of final times and (σ_i) of positive distances which either terminate with $T_i = T$ for some i, in which case we are done. Otherwise, T_i is defined for all i and $T_i \to T^* \leq T$. Clearly, the constructed $u \in C(0,T^*;\mathcal{H})$ solves (2.25) on $[0,T^*)$. If $\inf_i \sigma_i > 0$, then $T_{i+1} - T_i$ is bounded from below, which contradicts $T_i \leq T^*$. Thus $\liminf_{i\to\infty} \sigma_i = 0$ holds, which implies the assertion.

The estimates (2.41), (2.42), and (2.43) follow from the weak convergence of $\hat{u}_\tau$ in $L^2(0,T^*;\mathcal{V})$ and $\hat{u}_\tau'$ in $L^2(0,T^*;\mathcal{H})$ and the weak* convergence of $\hat{u}_\tau$ in $L^\infty(0,T^*;\mathcal{V})$ and weakly and weakly* sequentially lower semicontinuity of the respective norms. □

With this result, the sequence $(\hat{u}_\tau)$ converges in the given time interval without passing to a subsequence. Every converging subsequence has the unique solution as its limit and since every subsequence has a converging subsequence, $(\hat{u}_\tau)$ converges to the unique solution $u \in W(0,T^*;\mathcal{V},\mathcal{H}) \cap L^\eta(0,T^*;\mathcal{V})$ by a subsequence of subsequence argument.

2.5 Convergence of spatial discretizations

Our result is only useful for numerical analysis if the unique solution in Theorem 2.7 is also the limit of space-discrete solutions. We provide a convergence result under compatibility assumptions on the discrete spaces $\mathcal{V}_h \subset \mathcal{V}$ with $\mathcal{M}$. The space-discrete problem is of the following form.

Problem 2.8. Given $f \in L^2(0,T;\mathcal{H})$ and $u_{0,h} \in \mathcal{M} \cap \mathcal{V}_h$, find $u_h \in W(0,T;\mathcal{V},\mathcal{H})$ such that for almost all $t \in [0,T]$,

$$\begin{aligned} u_h(t) &\in \mathcal{M} \cap \mathcal{V}_h, \\ \langle u_h'(t), v_h\rangle + a(u_h(t), v_h; t) &= \langle f(t), v_h\rangle \quad \text{for all } v_h \in T_{u_h(t)}\mathcal{M} \cap \mathcal{V}_h, \\ u_h(0) &= u_{h,0}. \end{aligned} \tag{2.44}$$

We require that the discrete subspaces $\mathcal{V}_h \subset \mathcal{V}$ have the following properties.

B1 (Approximation property) There is a projection $Q_h\colon \mathcal{V} \to \mathcal{V}_h$, such that

$$\|Q_h v\|_{\mathcal{V}} \leq c\|v\|_{\mathcal{V}}$$

for every $h > 0$ and $v \in \mathcal{V}$ and the sequence $(Q_h v)$ converges to v in $\mathcal{V}$ as $h \searrow 0$. For every $u \in \mathcal{M} \cap \mathcal{V}$ there exists a sequence (u_h) with $u_h \in \mathcal{M} \cap \mathcal{V}_h$ and u_h converges to u in $\mathcal{V}$ as $h \searrow 0$.

B2 (Compatibility of tangent spaces)

(a) For $u_h \in \mathcal{M} \cap \mathcal{V}_h$ and $v_h \in T_u\mathcal{M} \cap \mathcal{V}_h$ an admissible curve with $\varphi(0) = u_h$, $\varphi'(0) = v_h$ can be chosen such that

$$\varphi(t) \in \mathcal{M} \cap \mathcal{V}_h$$

for all $|t|$ small enough.

(b) If $u_h \in \mathcal{M} \cap \mathcal{V}_h$ and $v_h \in \mathcal{V}_h$ are fulfilled, it follows that $P_{u_h} v_h \in T_{u_h}\mathcal{M} \cap \mathcal{V}_h$.

The model problem described in Section 2.1 allows for such a space discretization. One instance is obtained from homogeneous *bilinear quadrilateral elements*. These can be interpreted as the tensor product $\mathcal{P}_h \otimes \mathcal{P}_h \subset H_0^1(\Omega)$ of the space of piecewise affine linear functions

$$\mathcal{P}_h := \{u \in C_0(0,1)\colon u \text{ is affine linear on the interval } ((i-1)h, ih) \text{ for } i = 1, \dots, N\},$$

where $h = 1/N$ and $N \in \mathbb{N}$. The set $\mathcal{M}_r \cap \mathcal{P}_h \otimes \mathcal{P}_h$ is nonempty when $N > r$ and Property **B2** is satisfied for $\mathcal{M} = \mathcal{M}_r$ and $\mathcal{V}_h = \mathcal{P}_h \otimes \mathcal{P}_h$. This can be seen by $u_h \in \mathcal{M}_r \cap \mathcal{P}_h \otimes \mathcal{P}_h$ representing as

$$u_h = \sum_{i=1}^{r} u_{1,i} \otimes u_{2,i}$$

with $u_{1,i}, u_{2,i} \in \mathcal{P}_h$. Then the interestion

$$T_{u_h}\mathcal{M}_r \cap \mathcal{P}_h \otimes \mathcal{P}_h = \left\{ \sum_{i=1}^{r} u_{1,i} \otimes v_{2,i} + v_{1,i} \otimes u_{2,i} \colon v_{1,i}, v_{2,i} \in \mathcal{P}_h \right\},$$

is the tangent space of the rank-r manifold in $\mathcal{P}_h \otimes \mathcal{P}_h$.

For every $u \in H_0^1(0,1)$ the sequence (u_h) with $u_h \in \mathcal{P}_h$ and $u_h(ih) = u(ih)$ for $i = 0, \dots, N$ converges strongly to u as $h \searrow 0$. It follows that for $u \in H^{1,1}_{\text{mix}}(\Omega)$ the interpolation $u_h \in P_h \otimes P_h$ with $u_h(ih, jh) = u(ih, jh)$ for $i, j = 0, \dots, N$ converges strongly to u in $H^{1,1}_{\text{mix}}(\Omega)$. Since $H^{1,1}_{\text{mix}}(\Omega)$ is a dense subset and continuously embedded in $H_0^1(\Omega)$, there is also a sequence (u_h) with $u_h \in P_h \otimes P_h$ converging to u in $H_0^1(\Omega)$. Hence, the $H_0^1(\Omega)$-orthogonal projection $Q_h\colon H_0^1(\Omega) \to \mathcal{P}_h \otimes \mathcal{P}_h$ satisfies the first part of Property **B1**. Since $u \in \mathcal{M}_r \cap H_0^1(\Omega)$ is also also in $H^{1,1}_{\text{mix}}(\Omega)$, we may also use the interpolation $u_h(ih, jh) = u(ih, jh)$ and $\operatorname{rank} u_h \leq r$. If $\operatorname{rank} u_h < r$, we may add a small perturbation $\epsilon_h \in P_h \otimes P_h$ such $\operatorname{rank}(u_h + \epsilon_h) = r$ and $\|\epsilon_h\|_{H_0^1(\Omega)} \to 0$ as $h \searrow 0$. Therefore, the model problem also satisfies Property **B1**.

Theorem 2.9. *Let $(u_{0,h}) \subset \mathcal{M} \cap \mathcal{V}_h$ be a sequence that converges to u_0 in $\mathcal{V}$ as $h \searrow 0$ and let u_0 have positive $\mathcal{H}$-distance σ to the relative boundary $\overline{\mathcal{M}}^{\mathsf{w}} \setminus \mathcal{M}$. Then there exists a constant $c > 0$ independent of σ and a constant $h_0 > 0$ such that there is a unique u_h in $W(0, T^*; \mathcal{V}, \mathcal{H}) \cap L^\eta(0, T^*; \mathcal{V})$ that solves Problem 2.8 on the time interval $[0, T^*]$ when $T^* < \sigma^2/c$ for all $h \leq h_0$. Furthermore, u_h converges to the unique solution u of Problem 2.1 in $W(0, T^*; \mathcal{V}, \mathcal{H}) \cap L^\eta(0, T^*; \mathcal{V})$ weakly in $L^2(0, T^*; \mathcal{V})$ and strongly in $C(0, T^*; \mathcal{H})$, while the weak derivatives u_h' converge weakly to u' in $L^2(0, T^*, \mathcal{H})$.*

Proof. Since $u_{0,h}$ converges to u_0 in $\mathcal{V}$, there exists an $h_0 > 0$ such that $\|u_{0,h} - u_0\|_{\mathcal{V}} \leq \sigma/2$ and $\|u_{0,h} - u_0\|_{\mathcal{H}} \leq \sigma/2$ for all $h \leq h_0$ due to (2.22). Therefore, the $\mathcal{H}$-distance of $u_{0,h}$ from $\overline{\mathcal{M}}^{\mathsf{w}} \setminus \mathcal{M}$ is at least $\sigma/2$. Hence, applying Theorem 2.5 with $\mathcal{V}_h$ in place of $\mathcal{V}$ provides solutions u_h to Problem 2.8 on a time interval $[0, T^*]$ with $T^* < \sigma^2/(4c)$ for every $h \leq h_0$. Furthermore, Theorem 2.7 provides the estimates

$$\begin{aligned}
\|u_h\|^2_{L^2(0,T^*;\mathcal{V})} &\leq \left(\|u_0\|_{\mathcal{H}} + \frac{\sigma}{2}\right)^2 + C_1 \|f\|^2_{L^2(0,T^*;\mathcal{H})}, \\
\|u_h'\|^2_{L^2(0,T^*;\mathcal{H})} &\leq C_2 \left(\left(\|u_0\|_{\mathcal{V}} + \frac{\sigma}{2}\right)^2 + \|f\|^2_{L^2(0,T^*;\mathcal{H})}\right), \\
\|u_h\|^2_{L^\infty(0,T^*;\mathcal{V})} &\leq C_3 \left(\left(\|u_0\|_{\mathcal{V}} + \frac{\sigma}{2}\right)^2 + \|f\|^2_{L^2(0,T^*;\mathcal{H})}\right)
\end{aligned}$$

and hence, there exists a subsequence of (u_h) converging weakly to $\tilde{u}$ in $L^2(0, T^*; \mathcal{V})$ and weakly* in $L^\infty(0, T^*; \mathcal{V})$ and the derivatives (u_h') converging weakly to $\tilde{w}$ in $L^2(0, T^*; \mathcal{H})$.

We next show that $\tilde{w}$ is the weak derivative of $\tilde{u}$. For this, we need to verify that

$$\int_0^{T^*} \langle \tilde{w}(t), v \rangle\rangle \, \phi(t) + \langle \tilde{u}(t), v \rangle \, \phi'(t) \, dt = 0$$

for arbitrary $v \in \mathcal{V}$ and $\phi \in C_0^\infty(0, T^*)$. For any $v_h \in \mathcal{V}_h$ we may add and subtract the weak derivative of u_h

$$\begin{aligned}
&\int_{T^*} \langle \tilde{w}(t), v_h \rangle \, \phi(t) + \langle \tilde{u}(t), v_h \rangle \, \phi'(t) \, dt \\
&\qquad = \int_{T^*} \langle \tilde{w}(t) - u_h'(t), v_h \rangle \, \phi(t) + \langle \tilde{u}(t) - u_h(t), v_h \rangle \, \phi'(t) \, dt.
\end{aligned}$$

Now let (v_h) be a sequence converging to v in $\mathcal{V}$. Then it follows that

$$\begin{aligned}\int_{T^*} \langle \tilde{w}(t), v\rangle\, \phi(t) + \langle \tilde{u}(t), v\rangle\, \phi'(t)\, dt &= \lim_{h\searrow 0} \int_{T^*} \langle \tilde{w}(t), v_h\rangle\, \phi(t) + \langle \tilde{u}(t), v_h\rangle\, \phi'(t)\, dt \\ &= \lim_{h\searrow 0} \int_{T^*} \langle \tilde{w}(t) - u_h'(t), v_h\rangle\, \phi(t)\; + \langle \tilde{u}(t) - u_h(t), v_h\rangle\, \phi'(t)\, dt = 0\end{aligned}$$

since $v_h\phi$ converges strongly to $v\phi$ in $L^2(0, T^*; \mathcal{V})$. Therefore, the sequence (u_h) converges weakly in $W(0, T^*; \mathcal{V}, \mathcal{H})$ and due to the Aubin-Lions theorem strongly in $C(0, T^*; \mathcal{H})$ to $\tilde{u}$, taking into account that the sequence (u_h) is bounded in $L^\infty(0, T^*; \mathcal{V})$. This also implies that the initial condition $\tilde{u}(0) = \lim_{h\searrow 0} u_h(0) = \lim_{h\searrow 0} u_{0,h} = u_0$.

It remains to show that $\tilde{u}$ satisfies (2.25) and is therefore the unique solution of Problem 2.1 in $W(0, T^*; \mathcal{V}, \mathcal{H}) \cap L^\eta(0, T^*; \mathcal{V})$. It then follows that the entire sequence converges to $\tilde{u}$. For $v \in T_{\tilde{u}(t)}\mathcal{M} \cap \mathcal{V}$, we consider $(v_h) = (Q_h v)$ which converges strongly to v in $\mathcal{V}$. Property **B1** implies that the sequence is uniformly bounded in $\mathcal{V}$ and due to (2.22) also in $\mathcal{H}$. By (2.44), we have

$$\langle u_h'(t), P_{u_h(t)} v_h\rangle + a(u_h(t), P_{u_h(t)} v_h; t) = \langle f(t), P_{u_h(t)} v_h\rangle$$

for almost every t since $P_{u_h(t)} v_h \in \mathcal{M} \cap \mathcal{V}_h$ by Property **B2**(b). We have chosen the time interval in a way such that $u_h(t) \in \mathcal{M}' \subset \mathcal{M}$ lie in a weakly compact subset for all $t \in [0, T^*]$. Hence, using Assumption **A2**,

$$\begin{aligned}\|v - P_{u_h(t)} v_h\|_{\mathcal{H}} &\le \|v - P_{u_h(t)} v\|_{\mathcal{H}} + \|P_{u_h(t)}(v - v_h)\|_{\mathcal{H}} \\ &\le \kappa(\mathcal{M}')\|\tilde{u}(t) - u_h(t)\|_{\mathcal{H}}\|v\|_{\mathcal{H}} + \|v - v_h\|_{\mathcal{H}},\end{aligned} \tag{2.45}$$

i.e., $P_{u_h(t)} v_h$ converges strongly to v in $\mathcal{H}$. Using a similar argument as in Theorem 2.5(b), it is sufficient to show

$$\int_0^{T^*} \langle \tilde{u}'(t), v(t)\rangle + a(\tilde{u}(t), v(t); t) - \langle f(t), v(t)\rangle\, dt = 0$$

for all $v \in L^\infty(0, T^*; \mathcal{V})$ with $v(t) \in T_{\tilde{u}(t)}\mathcal{M} \cap \mathcal{V}$ for almost every t.

Since $P_{u_h(t)} Q_h v(t)$ converges to $v(t)$ in $\mathcal{H}$ for almost all $t \in [0, T^*]$ and we have the square integrable bound (2.45), it follows that $P_{u_h(t)} Q_h v(t)$ converges strongly to v in $L^2(0, T^*; \mathcal{H})$. This together with weak convergence of (u_h') in $L^2(0, T^*; \mathcal{H})$ implies

$$\lim_{h\searrow 0} \int_0^{T^*} \langle u_h'(t), P_{u_h(t)} Q_h v(t)\rangle - \langle f(t), P_{u_h(t)} Q_h v(t)\rangle\, dt = \int_0^{T^*} \langle \tilde{u}'(t), v(t)\rangle - \langle f(t), v(t)\rangle\, dt.$$

Finally, we use Assumption **A4**. We have

$$\begin{aligned}&a(u_h(t), P_{u_h(t)} Q_h v(t); t) - a(\tilde{u}(t), v(t); t) \\ &\qquad = \langle A_1(t) u_h(t), P_{u_h(t)} Q_h v(t)\rangle - \langle A_1(t)\tilde{u}(t), v(t)\rangle \\ &\qquad\qquad + \langle A_2(t) u_h(t), P_{u_h(t)} Q_h v(t)\rangle - \langle A_2(t)\tilde{u}(t), v(t)\rangle\end{aligned}$$

and due to Assumption **A4**(a)

$$\langle A_1(t)u_h(t), P_{u_h(t)}Q_hv(t)\rangle = \langle A_1(t)u_h(t), Q_hv(t)\rangle.$$

This implies

$$\lim_{h\searrow 0}\int_0^{T^*} \langle A_1(t)u_h(t), P_{u_h(t)}Q_hv(t)\rangle - \langle A_1(t)u(t), v(t)\rangle \, dt = 0$$

as u_h converges weakly to $\tilde{u}$ and Q_hv converges strongly to v in $L^2(0,T^*;\mathcal{V})$. For the second summand, we have

$$\begin{aligned}&\langle A_2(t)u_h(t), P_{u_h(t)}Q_hv(t)\rangle - \langle A_2(t)\tilde{u}(t), v(t)\rangle\\ &\qquad\qquad = \langle A_2(t)u_h(t), P_{u_h(t)}Q_hv(t) - v(t)\rangle + \langle A_2(t)(\tilde{u}(t) - u_h(t)), v(t)\rangle,\end{aligned}$$

where

$$\left|\langle A_2(t)u_h(t), P_{u_h(t)}Q_hv(t) - v(t)\rangle\right| \le \gamma\|u_h(t)\|_{\mathcal{V}}^{\eta}\|P_{u_h(t)}Q_hv(t) - v(t)\|_{\mathcal{H}}$$

and $\int_0^{T^*} \|u_h(t)\|_{\mathcal{V}}^{\eta}\|P_{u_h(t)}Q_hv(t) - v(t)\|_{\mathcal{H}}\, dt \to 0$ by (2.45) and the uniform bound of u_h in $L^\infty(0,T^*;\mathcal{V})$. Moreover, since u_h converges weakly to $\tilde{u}$ in $L^2(0,T^*;\mathcal{V})$, we have $\int_0^{T^*} \langle A_2(t)(\tilde{u}(t) - u_h(t)), v(t)\rangle \, dt \to 0$ as $h \searrow 0$. Combining these results, we have

$$\int_0^{T^*} a(u_h(t), P_{u_h(t)}Q_hv(t);t) - a(\tilde{u}(t), v(t);t)\, dt \to 0 \quad \text{as} \quad h\searrow 0$$

and hence

$$\begin{aligned}&\int_0^{T^*} \langle \tilde{u}'(t), v(t)\rangle + a(\tilde{u}(t), v(t);t) - \langle f(t), v(t)\rangle \, dt\\ &= \lim_{h\searrow 0}\int_0^{T^*} \langle u_h'(t), P_{u_h(t)}Q_hv(t)\rangle + a(u_h(t), P_{u_h(t)}Q_hv(t);t) - \langle f(t), P_{u_h(t)}Q_hv(t)\rangle \, dt = 0\end{aligned}$$

for all $v \in L^\infty(0,T^*;\mathcal{V})$ with $v(t) \in T_{\tilde{u}(t)}\mathcal{M} \cap \mathcal{V}$ for almost every t. □

2.6 Outlook

We have established results for the existence and uniqueness of dynamical low-rank approximations to solutions of certain parabolic problems. Furthermore, we showed convergence of either time-discrete or space-discrete solutions to the underlying continuous approximation. We expect that our results are also applicable for higher-dimensional parabolic problems for suitable low-rank tensor formats [KL10, LOV15, LRSV13]. Naturally, it is of interest under which conditions dynamical low-rank approximations are good approximations to the underlying unrestrained problem and how to establish convergence rates for discrete low-rank approximations.

Chapter 3

Solving definite multiparameter eigenvalue problems

In this chapter, we are concerned with the *multiparameter eigenvalue problem* (MEP). For this, let $A_{k\ell} \in \mathbb{C}^{n_k \times n_k}$ be square matrices. We look for unit vectors $u_k \in \mathcal{U}_k \subset \mathbb{C}^{n_k}$ and $(\lambda_1, \ldots, \lambda_m) \in \mathbb{C}^m$ such that

$$\begin{array}{ccccccc} \big(A_{10} & + \lambda_1 A_{11} & + \ldots + & \lambda_m A_{1m}\big) & u_1 & = 0 \\ \big(A_{20} & + \lambda_1 A_{21} & + \ldots + & \lambda_m A_{2m}\big) & u_2 & = 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ \big(A_{m0} & + \lambda_1 A_{m1} & + \ldots + & \lambda_m A_{mm}\big) & u_m & = 0 \end{array} \tag{3.1}$$

or for $(\lambda_0, \ldots, \lambda_m) \neq 0$ in the homogeneous version

$$\begin{array}{ccccccc} \big(\lambda_0 A_{10} & + \lambda_1 A_{11} & + \ldots + & \lambda_m A_{1m}\big) & u_1 & = 0 \\ \big(\lambda_0 A_{20} & + \lambda_1 A_{21} & + \ldots + & \lambda_m A_{2m}\big) & u_2 & = 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ \big(\lambda_0 A_{m0} & + \lambda_1 A_{m1} & + \ldots + & \lambda_m A_{mm}\big) & u_m & = 0, \end{array} \tag{3.2}$$

We call $u_1 \otimes \ldots \otimes u_m$ an eigenvector corresponding to the eigenvalue $\lambda = (\lambda_1, \ldots, \lambda_m)$ in case of (3.1) and the eigenvalue $\lambda = (\lambda_0, \ldots, \lambda_m)$ in case of (3.2). The MEP generalizes linear systems of equations and generalized eigenvalue problems. In the case $m = 1$, the MEP is a generalized eigenvalue problem, and if $n_k = 1$ for $k = 1, \ldots, m$, then the MEP reduces to a linear system of equations.

The solutions to the MEP can be obtained using multilinear algebra techniques; see e.g., [Atk72, Theorem 6.8.1]. The idea is to apply a multilinear version of Cramer's rule for linear systems. For $u_k, v_k \in \mathbb{C}^{n_k}$ and $k = 1, \ldots, m$, define the matrix

$$\tilde{W}(v_1, \ldots, v_m, u_1, \ldots, u_m) = [\tilde{w}_{k\ell}(u_k, v_k)]_{k=1,\ell=0}^{m,m} = \left[v_k^{\mathsf{H}} A_{k\ell} u_k\right]_{k=1,\ell=0}^{m,m}.$$

If $u_1 \otimes \ldots \otimes u_m$ is an eigenvector corresponding to $\lambda = (\lambda_0, \ldots, \lambda_m)$, then

$$\tilde{W}(v_1, \ldots, v_m, u_1, \ldots, u_m)\lambda = 0$$

for all $v_k \in \mathbb{C}^{n_k}$. Now assume there is $\mu = (\mu_0, \dots, \mu_m)$ such that for all $u_k \in \mathcal{U}_k$

$$\begin{pmatrix} \mu_0 & \mu_1 & \dots & \mu_m \\ \tilde{w}_{10}(u_1, v_1) & \tilde{w}_{11}(u_1, v_1) & \dots & \tilde{w}_{1m}(u_1, v_1) \\ \vdots & \vdots & & \vdots \\ \tilde{w}_{m0}(u_m, v_m) & \tilde{w}_{m1}(u_m, v_m) & \dots & \tilde{w}_{mm}(u_m, v_m) \end{pmatrix} = \begin{pmatrix} \mu^{\mathsf{T}} \\ \tilde{W}(v_1, \dots, v_m, u_1, \dots, u_m) \end{pmatrix}$$

is invertible for some $v_k \in \mathbb{C}^{n_k}$. Let $u_1 \otimes \ldots \otimes u_m$ be an eigenvector corresponding to $\lambda = (\lambda_0, \dots, \lambda_m)$. It follows from Cramer's rule that

$$\begin{aligned} \lambda_m \det \begin{pmatrix} \mu \\ \tilde{W}(v_1, \dots, v_m, u_1, \dots, u_m) \end{pmatrix} \\ = \det \begin{pmatrix} \mu_0 & \dots & \mu_{m-1} & \mu^{\mathsf{T}} \lambda \\ \tilde{w}_{10}(u_1, v_1) & \dots & \tilde{w}_{1,m-1}(u_1, v_1) & 0 \\ \vdots & & \vdots & \vdots \\ \tilde{w}_{m0}(u_m, v_m) & \dots & \tilde{w}_{m,m-1}(u_m, v_m) & 0 \end{pmatrix}. \end{aligned}$$

Note that the determinants above are linear in each u_k and antilinear in each v_k. We can therefore define linear operators $\Delta, \Delta_m \colon \bigotimes_{k=1}^m \mathbb{C}^{n_k} \to \bigotimes_{k=1}^m \mathbb{C}^{n_k}$ that satisfy

$$(v_1 \otimes \ldots \otimes v_m)^{\mathsf{H}} \Delta (u_1 \otimes \ldots \otimes u_m) = \det \begin{pmatrix} \mu^{\mathsf{T}} \\ \tilde{W}(v_1, \dots, v_m, u_1, \dots, u_m) \end{pmatrix} \tag{3.3}$$

and

$$(v_1 \otimes \ldots \otimes v_m)^{\mathsf{H}} \Delta_m (u_1 \otimes \ldots \otimes u_m) = (-1)^m \det \begin{pmatrix} \tilde{w}_{10}(u_1, v_1) & \dots & \tilde{w}_{1,m-1}(u_1, v_1) \\ \vdots & & \vdots \\ \tilde{w}_{m0}(u_m, v_m) & \dots & \tilde{w}_{m,m-1}(u_m, v_m) \end{pmatrix}.$$

Hence, the eigenpair $(\lambda, u_1 \otimes \ldots \otimes u_m)$ satisfies

$$\mu^{\mathsf{T}} \lambda \, \Delta_m (u_1 \otimes \ldots \otimes u_m) = \lambda_m \, \Delta (u_1 \otimes \ldots \otimes u_m)$$

and we can define similar linear operators $\Delta_0, \dots, \Delta_{m-1}$ that satisfy

$$\mu^{\mathsf{T}} \lambda \, \Delta_\ell (u_1 \otimes \ldots \otimes u_m) = \lambda_\ell \, \Delta (u_1 \otimes \ldots \otimes u_m). \tag{3.4}$$

Hence, all eigenpairs of the MEP can be obtained as the solutions to the simultaneous eigenvalue problems (3.4). In [Atk72, Chapter 6] the converse is also shown. In addition, the linear operators $\Delta^{-1} \Delta_\ell$ commute, and their eigenvectors consist of decomposible tensors $u_1 \otimes \ldots \otimes u_m$.

Multiparameter eigenvalue problems have been extensively studied; see e.g., [Atk72, Vol88]. They naturally arise in mathematical physics when variables can be separated but resulting spectral parameters cannot. We will discuss this case more thoroughly in Section 3.2. There are many other applications that lead to (3.1) or (3.2) including delay differential equations [JH09] and optimization problems [SNTI16]. They can also appear when the domain of a boundary eigenvalue problem is decomposed [RJ21, Section 5.2]. We also want to remark that some nonlinear eigenvalue problems can be expressed by (3.1) as described in [RJ21]. For example, the polynomial eigenvalue problem

$$M(\lambda)u = \sum_{k=0}^{m} \lambda^k A_k u = 0$$

can be expressed as (3.1) by setting $\lambda_k = \lambda^k$ and adding equations

$$(A_{k0} + \lambda_1 A_{k1} + \lambda_{k-1} A_{k,k-1} + \lambda_k A_{k,k}) u_k = 0$$

such that $\det(A_{k0} + \lambda_1 A_{k1} + \lambda_{k-1} A_{k,k-1} + \lambda_k A_{k,k}) = \lambda_k - \lambda_1 \lambda_{k-1}$. This however leads to a singular MEP if $m > 2$, i.e., there is no μ such that Δ in (3.4) is nonsingular. The more simple polynomial eigenvalue problem

$$M(\lambda)u = (A + \lambda B + \lambda^k C)u$$

can be transformed into a nonsingular two-parameter eigenvalue problem whose linear eigenvalue problem (3.4) can take various forms, one being the linearization described in [MW02]. Similarly, the MEP contains many nonlinear eigenvalue problems with algebraic relations of the spectral parameters $\lambda_0, \ldots, \lambda_m$.

There are various approaches to solve (3.1), many of which use the linear eigenvalue problem (3.4). For the case that all matrices are Hermitian and Δ_0 is positive definite, this was first considered in [ST86], and for the more general case, that Δ_0 is invertible in [HKP05] by using the generalized Schur decomposition. These approaches work well when $\prod_{k=1}^{m} n_k$ is not too large, as they have a complexity of order $O(\prod_{k=1}^{m} n_k^3)$. This is only feasible for small n_k and m, as otherwise $\prod_{k=1}^{m} n_k$ is getting too large. For $m = 2$ and larger n_k, subspace methods are being used to find a selection of eigenvalues. In [HP02] and [HKP05] a Jacobi-Davidson type method for the two-parameter case was proposed and in [MP15] an Arnoldi type method was considered. For $m = 3$ various subspace methods were proposed in [HMMP19]. Another possibility is based on homotopy continuation, for example discussed in [Ple00] and [DYY16]. These aim to find all eigenvalues.

In this chapter, we extend results from [EN22]. In [EN22], we were considering the two-parameter eigenvalue problem under the assumption that all matrices are Hermitian and Δ_0 is positive definite. We used the notion of the signed index of an eigenvalue and employed an alternating algorithm that has a geometric interpretation closely related to Newton's method and also an optimization perspective. We will thoroughly discuss the signed index of an eigenvalue in Section 3.1. In this chapter, we use similar ideas for the MEP (3.1) and its homogeneous form (3.2) under certain definiteness assumptions.

These are satisfied for a class of boundary eigenvalue problems which we introduce in Section 3.2. We introduce Newton-type methods in Section 3.3 and discuss a perspective coming from optimization problems in Section 3.4. These methods also apply to MEPs with large n_k and m when only a few eigenvalues of certain multiindex are sought; the complexity per eigenvalue is in $O(\sum_{k=1}^{m} n_k^3)$. Finally, in Section 3.5 we explore the performance of these methods in numerical experiments.

Using Newton's method for MEPs is not a new idea and was proposed previously, e.g. in [Pod08] by looking for joint zeros of $f_k(\lambda) = \det(\sum_{\ell=0}^{m} \lambda_\ell A_{k\ell})$ and $k = 1, \ldots, m$. However, for general MEPs, this is sensitive to initialization and a good starting guess is required. We circumvent this problem by applying Newton's method to functions that have unique zeros. Our proposed methods are related to methods solving multiparameter Sturm-Liouville eigenvalue problems by looking for eigenfunctions with a certain amount of internal zeros; see e.g., [Lev94, Lev99]. The concept of internal zeros of eigenfunctions of Sturm-Liouville eigenvalue problems is a special case of the multiindex of an eigenvalue which will be explained in the upcoming two sections.

3.1 Definite multiparameter eigenvalue problems

In this section, we aim to discuss a generalization of the well-known fact from linear algebra that any Hermitian matrix $A = A^{\mathsf{H}}$ has only real eigenvalues. For this purpose, we collect some results from [Vol88, Chapter 1]. From now on, we will always assume that all matrices $A_{k\ell}$ in (3.1) and (3.2) are Hermitian. Now assume that $\lambda = (\lambda_0, \ldots, \lambda_m)$ is real. Then

$$\sum_{\ell=0}^{m} \lambda_\ell A_{k\ell} \quad \text{for } k = 1, \ldots, m$$

are Hermitian matrices and all their respective eigenvalues are real. This observation leads to the following definition.

Definition 3.1. Let $\lambda \in \mathbb{R}^{m+1}$ be an eigenvalue of (3.2). The multiindex of λ is the m-tuple $\mathbf{i} = (i_1, \ldots, i_m)$ such that 0 is the i_k-th largest eigenvalue of $\sum_{\ell=0}^{m} \lambda_\ell A_{k\ell}$.

In the case $m = 1$, a simple condition that eigenvalues can be chosen real, is that there is a positive definite matrix $\mu_0 A_{10} + \mu_1 A_{11}$ for some real $(\mu_0, \mu_1) \neq 0$. For MEPs, we consider the matrix

$$W(u_1, \ldots, u_m) = \begin{pmatrix} u_1^{\mathsf{H}} A_{10} u_1 & \ldots & u_1^{\mathsf{H}} A_{1m} u_1 \\ \vdots & & \vdots \\ u_m^{\mathsf{H}} A_{m0} u_m & \ldots & u_m^{\mathsf{H}} A_{mm} u_m \end{pmatrix}. \tag{3.5}$$

Definition 3.2. The MEP (3.2) is called locally definite if $\operatorname{rank} W(u_1, \ldots, u_m) = m$ for all $u_k \in \mathcal{U}_k$, $k = 1, \ldots, m$.

Now let $u_1 \otimes \ldots \otimes u_m$ be an eigenvector corresponding to the eigenvalue λ. Then λ is in the nullspace of $W(u_1, \ldots, u_m)$. If the MEP is locally definite, this nullspace is one-dimensional. Since $W(u_1, \ldots, u_m)$ is real, the eigenvalue can thus be chosen real as well. By scaling, we may restrict the search for eigenvalues to the sphere $\mathbb{S}^m = \{\lambda \in \mathbb{R}^{m+1} \colon \|\lambda\| = 1\}$. By the above consideration, we may further restrict to the set

$$\mathcal{P} = \{\lambda \in \mathbb{S}^m \colon W(u)\lambda = 0 \text{ for some } u \in \mathcal{U}_1 \times \ldots \times U_m\}. \tag{3.6}$$

This set obviously has the symmetry $\mathcal{P} = -\mathcal{P}$. In fact, it is the disjoint union of two connected sets $\mathcal{P}^+$ and $\mathcal{P}^- = -\mathcal{P}^+$; see e.g., [Vol88, Lemma 1.2.1]. Here, $\mathcal{P}^+$ is the set

$$\mathcal{P}^+ = \left\{\lambda \in \mathbb{S}^m \colon W(u)\lambda = 0 \text{ and } \det\begin{pmatrix} \lambda^{\mathsf{T}} \\ W(u) \end{pmatrix} > 0 \text{ for some } u \in \mathcal{U}_1 \times \ldots \times U_m\right\}.$$

With this machinery, we can characterize all eigenvalues of a locally definite MEP (3.2).

Theorem 3.3. [Vol88, Theorem 1.4.1] *Let the MEP* (3.2) *be locally definite. Then for every multiindex* $\mathbf{i} \in \{1, \ldots, n_1\} \times \ldots \times \{1, \ldots, n_m\}$ *and every sign* $\sigma \in \{+, -\}$ *there is a unique eigenvalue* $\lambda \in \mathcal{P}^\sigma$ *of multiindex* $\mathbf{i}$. *We say* λ *is of signed index* $(\mathbf{i}, \sigma)$.

It is also useful to consider stronger definiteness assumptions, which in practice are often satisfied.

Definition 3.4. The MEP (3.2) is called definite with respect to $\mu \in \mathbb{S}^m$ if

$$\det\begin{pmatrix} \mu^{\mathsf{T}} \\ W(u_1, \ldots, u_m) \end{pmatrix} > 0$$

for all $u_k \in \mathcal{U}_k$, $k = 1, \ldots, m$.

This definition is equivalent to the condition that $\mathcal{P}$ in (3.6) is the disjoint union of the connected sets

$$\mathcal{P}^+ = \{\lambda \in \mathcal{P} \colon \mu^{\mathsf{T}}\lambda > 0\}$$

and $\mathcal{P}^- = -\mathcal{P}^+$. It also follows that the linear operator Δ defined in (3.3) is positive definite; see e.g., [Vol88, Theorem 4.4.1]. For the inhomogeneous MEP (3.1) definiteness with respect to $\mu = (1, 0, \ldots, 0)$ is naturally a good condition. In this case, Theorem 3.3 implies, that every eigenvalue of signed index $(\mathbf{i}, +)$ can be scaled to $\lambda = (1, \lambda_1, \ldots, \lambda_m)$. If the MEP is definite with respect to $(1, 0, \ldots, 0)$, it is called *right definite*. When we consider a right definite MEP of the form (3.1), we denote

$$\mathcal{Q} = \left\{\lambda \in \mathbb{R}^m \colon W(u)\begin{pmatrix} 1 \\ \lambda \end{pmatrix} = 0 \text{ for some } u \in \mathcal{U}_1 \times \ldots \times U_m\right\}$$

as an analog of $\mathcal{P}$. The sets $\mathcal{P}^\sigma$ and $\mathcal{Q}$ can be seen as analogies to the numerical range of a linear operator.

Another commonly used assumption is *left definiteness.* The MEP is left definite with respect to $\mu = (0, \mu_1, \dots, \mu_m)$ if the matrices A_{k0} are negative definite and

$$\det \begin{pmatrix} u_1^{\mathsf{H}} A_{11} u_1 & \dots & u_m^{\mathsf{H}} A_{1m} u_m \\ \vdots & & \vdots \\ u_{k-1}^{\mathsf{H}} A_{k-1,1} u_{k-1} & \dots & u_{k-1}^{\mathsf{H}} A_{k-1,m} u_{k-1} \\ \mu_1 & \dots & \mu_m \\ u_{k+1}^{\mathsf{H}} A_{k+1,1} u_{k+1} & \dots & u_{k+1}^{\mathsf{H}} A_{k+1,m} u_{k+1} \\ \vdots & & \vdots \\ u_m^{\mathsf{H}} A_{m1} u_m & \dots & u_m^{\mathsf{H}} A_{mm} u_m \end{pmatrix} > 0 \tag{3.7}$$

for $k = 1, \dots, m$. Left definiteness implies definiteness with respect to μ.

Many problems coming from application are either right or left definite or both. To end this section, we want to state a useful equivalent condition for local definiteness.

Lemma 3.5. [Vol88, Theorem 1.4.3] *The MEP* (3.2) *is locally definite if and only if for every* $\sigma = (\sigma_1, \dots, \sigma_m) \in \{-1, 1\}^m$ *there is* $\alpha = (\alpha_0, \dots, \alpha_m) \in \mathbb{R}^{m+1}$ *such that* $\sum_{\ell=0}^{m} \sigma_k u_k^{\mathsf{H}} A_{k\ell} u_k \alpha_\ell > 0$ *for all* $u_k \in \mathcal{U}_k, k = 1, \dots, m$.

This condition states that local definiteness is equivalent to existence of parameters $\alpha = (\alpha_0, \dots, \alpha_m)$ for every sign $\sigma_k \in \{-1, 1\}$ such that $\sigma_k \sum_{\ell=0}^{m} \alpha_\ell A_{k\ell}$ is positive definite for $k = 1, \dots, m$. With this lemma, it is not difficult to see that local definiteness and definiteness are equivalent for $m = 2$. If the MEP is definite with respect to μ, it follows directly that it is local definite as well. For the converse direction, choose $\alpha = (\alpha_0, \alpha_1, \alpha_2)$ and $\beta = (\beta_0, \beta_1, \beta_2)$ such that $\sum_{\ell=0}^{2} \alpha_\ell A_{k\ell}$ and $(-1)^k \sum_{\ell=0}^{2} \beta_\ell A_{k\ell}$ are positive definite for $k = 1, 2$. Next, choose μ orthogonal to α and β such that $\det \begin{pmatrix} \mu & \alpha & \beta \end{pmatrix} > 0$. Then

$$\det \begin{pmatrix} \mu^{\mathsf{T}} \\ W(u_1, u_2) \end{pmatrix} \det \begin{pmatrix} \mu & \alpha & \beta \end{pmatrix} \\ = \det \begin{pmatrix} \mu^{\mathsf{T}} \mu & 0 & 0 \\ \sum_{\ell=0}^{2} \mu_\ell u_1^{\mathsf{H}} A_{1\ell} u_1 & \sum_{\ell=0}^{2} \alpha_\ell u_1^{\mathsf{H}} A_{1\ell} u_1 & \sum_{\ell=0}^{2} \beta_\ell u_1^{\mathsf{H}} A_{1\ell} u_1 \\ \sum_{\ell=0}^{2} \mu_\ell u_2^{\mathsf{H}} A_{2\ell} u_2 & \sum_{\ell=0}^{2} \alpha_\ell u_2^{\mathsf{H}} A_{2\ell} u_2 & \sum_{\ell=0}^{2} \beta_\ell u_2^{\mathsf{H}} A_{2\ell} u_2 \end{pmatrix} > 0,$$

i.e., the MEP is definite with respect to μ. For $m \geq 3$ local definiteness no longer implies definiteness. A counter example is given in [Vol88, Chapter 1.5] and in Section 3.5.4.

3.2 Multiparameter Sturm-Liouville problems

As a motivation, consider the Helmholtz equation with Dirichlet boundary conditions

$$\begin{aligned} \Delta u(x) + \lambda\, u(x) &= 0 \qquad \text{for } x \in \Omega \\ u(x) &= 0 \qquad \text{for } x \in \partial\Omega \end{aligned} \tag{3.8}$$

on a domain Ω. If the domain is the hyperrectangle $\Omega = (a_1, b_1) \times \ldots \times (a_m, b_m)$, then the solutions of (3.8) can be easily obtained by separating the variables and solving the one-dimensional boundary value problems

$$u_k''(x_k) + \lambda_k \, u_k(x_k) = 0 \quad \text{for } x_k \in (a_k, b_k) \quad \text{and} \quad u_k(a_k) = 0 = u_k(b_k),$$

which are readily solved. The solutions of (3.8) are given by $u(x_1, \ldots, x_m) = \prod_{k=1}^m u_k(x_k)$ and $\lambda = \sum_{k=1}^m \lambda_k$. If the domain is the disc $\Omega = \{(x, y) \in \mathbb{R}^2 \colon x^2 + y^2 < 1\}$, we can apply a more involved separation. Using polar coordinates, one obtains the solution via

$$\begin{array}{ccccccll} \Phi''(\varphi) & & & - & \nu\, \Phi(\varphi) & = & 0 & \text{for } \varphi \in (0, 2\pi), \\ (rR'(r))' & + & \lambda\, rR(r) & + & \nu\, \frac{1}{r} R(r) & = & 0 & \text{for } r \in (0, 1) \end{array}$$

with suitable boundary conditions. These boundary value problems can still be solved one after another and the solutions of (3.8) are again the products of solutions $R\Phi$ and with the eigenvalue λ.

For our purpose, the situation is more interesting if the domain is the ellipse $\Omega = \{(x, y) \in \mathbb{R}^2 \colon x^2/a^2 + y^2/b^2 < 1\}$ with $b > a > 0$. In the elliptical coordinates

$$x = c\cosh(\rho)\cos(\varphi), \quad y = c\sinh(\rho)\sin(\varphi),$$

the equation separates into

$$\begin{array}{ccccccll} P''(\rho) & + & \lambda\, \frac{c^2}{2}\cosh(2\rho)\, P(\rho) & - & \nu\, P(\rho) & = & 0 & \text{for } \rho \in (0, r), \\ \Phi''(\varphi) & - & \lambda\, \frac{c^2}{2}\cos(2\varphi)\, \Phi(\varphi) & + & \nu\, \Phi(\varphi) & = & 0 & \text{for } \varphi \in (0, 2\pi), \end{array}$$

again with suitable boundary conditions. Here, c describes the focal points of the ellipse and $c^2 \cosh^2 r = a^2$. The resulting equations are *Mathieu's modified differential equation* and *Mathieu's differential equation*, and the spectral parameters cannot be separated.

In higher dimensions, if $B = \operatorname{Diag}(1/(b_{m+1} - b_1), \ldots, 1/(b_{m+1} - b_m))$ is a diagonal matrix with m different positive eigenvalues $0 < b_{m+1} - b_m < \ldots < b_{m+1} - b_1$ and the domain is the ellipsoid $\Omega = \{x \in \mathbb{R}^m \colon x^\mathsf{T} B x < 1\}$, we can separate using the ellipsoidal coordinates $\xi_1, \ldots, \xi_m$ defined by

$$\sum_{\ell=1}^m \frac{x_\ell^2}{\xi_k - b_\ell} = 1 \quad \text{for } \xi_k \in (b_k, b_{k+1}).$$

We obtain the m coupled differential equations

$$(p(\xi_k)\, u_k'(\xi_k))' + \frac{(-1)^{m-k}}{4p(\xi_k)} \left(\sum_{\ell=1}^m \lambda_\ell \, \xi_k^{\ell-1} \right) u_k(\xi_k) = 0 \quad \text{for } \xi_k \in (b_k, b_{k+1}) \tag{3.9}$$

with $p(\xi) = \sqrt{\prod_{\ell=1}^m |b_\ell - \xi|}$ and appropriate boundary conditions; see e.g., [Vol88, Chapter 6.9], [SW79, Section 1.2], or [MS54, Chapter 1.13].

All of these problems are coupled Sturm-Liouville eigenvalue problems. They are of the general form

$$(p_k(x_k)\,u_k'(x_k))' + q_k(x_k)\,u_k(x_k) + \left(\sum_{\ell=1}^{m} \lambda_\ell\, a_{k\ell}(x_k)\right) u_k(x_k) = 0 \quad \text{for } x_k \in (a_k, b_k) \tag{3.10}$$

and boundary conditions. Usual assumptions are $p_k(x_k) > 0$ for $x_k \in [a_k, b_k]$ and p_k is continuously differentiable. This is not the case for (3.9) but oftentimes the positivity assumption can be weakened; see e.g., [Tes12, Chapter 5.3]. Similar to the finite dimensional problem (3.1), real eigenvalues $\lambda = (\lambda_1, \dots, \lambda_m)$ have an index since the occurring operators are self-adjoint with respect to the inner product on $L^2(a_k, b_k)$ and are bounded from above. There are according existence results for eigenvalues of signed index; see e.g., [Vol88, Theorem 2.5.3 and Theorem 2.7.1]. Notably, if the eigenvalue problem is right definite, then there is an eigenvalue with signed index $(\mathbf{i}, +)$. Right and left definiteness of (3.10) can be checked pointwise. It is right definite if

$$\det \begin{pmatrix} a_{11}(x_1) & \dots & a_{1m}(x_1) \\ \vdots & & \vdots \\ a_{m1}(x_m) & \dots & a_{mm}(x_m) \end{pmatrix} > 0$$

on a dense subset of $[a_1, b_1] \times \ldots \times [a_m, b_m]$ and the analog of (3.7) for left definiteness is satisfied if

$$\det \begin{pmatrix} a_{11}(x_1) & \dots & a_{1m}(x_1) \\ \vdots & & \vdots \\ a_{k-1,1}(x_{k-1}) & \dots & a_{k-1,m}(x_{k-1}) \\ \mu_1 & \dots & \mu_m \\ a_{k+1,1}(x_{k+1}) & \dots & a_{k+1,m}(x_{k+1}) \\ \vdots & & \vdots \\ a_{m1}(x_m) & \dots & a_{mm}(x_m) \end{pmatrix} > 0$$

on a dense subset of $[a_1, b_1] \times \ldots \times [a_m, b_m]$; see e.g., [Vol88, Theorem 3.6.2].

The multiindex of an eigenvalue also has an interpretation as the internal zeros of the corresponding eigenfunctions u_k of (3.10). An eigenfunction u of the Sturm-Liouville eigenvalue problem

$$(p(x)\,u'(x))' + q(x)\,u(x) = \lambda\, u(x) \quad \text{for } x \in (a, b)$$

has n internal zeros if λ is the $n+1$-th largest eigenvalue under the boundary conditions

$$\cos(\alpha)\,u(a) + \sin(\alpha)\,u'(a) = 0 \quad \text{and} \quad \cos(\beta)\,u(b) + \sin(\beta)\,u'(b) = 0,$$

or $2n$ internal zeros if λ is the $2n$-th or $2n+1$-th largest eigenvalue under periodic boundary conditions, and $2n-1$ internal zeros if λ is the $2n-1$-th or $2n$-th largest eigenvalue under antiperiodic boundary conditions; see e.g, [Tes12, Theorem 5.17 and Theorem 5.37] and [CL55, Chapter 8, Theorem 2.1 and Theorem 3.1]. The analog is true for the multiindex of an eigenvalue in the case of (3.10). If $\lambda = (\lambda_1, \dots, \lambda_m)$ has multiindex $\mathbf{i} = (i_1, \dots, i_m)$, then u_k has $i_k - 1$ or i_k internal zeros, depending on the boundary conditions; see e.g., [Vol88, Theorem 3.5.1].

The methods we construct in Section 3.3 compute eigenvalues based on its index. When the problem comes from a discretization of a multiparameter Sturm-Liouville eigenvalue problem (3.10), this has the additional interpretation of finding an eigenvalue whose eigenfunctions have the corresponding number of interior zeros.

3.3 Newton-type methods

For each multiindex $\mathbf{i} \in \{1, \dots, n_1\} \times \dots \times \{1, \dots, n_m\}$, we define the function

$$F_{\mathbf{i}} \colon \mathbb{R}^m \to \mathbb{R}^m, \quad \lambda = (\lambda_1, \dots, \lambda_m) \mapsto F_{\mathbf{i}}(\lambda) = (\varepsilon_{1,i_1}(\lambda), \dots, \varepsilon_{m,i_m}(\lambda)), \tag{3.11}$$

where $\varepsilon_{k,i_k}(\lambda)$ is the i_k-th largest eigenvalue of the matrix $\sum_{\ell=0}^{m} \lambda_\ell A_{k\ell}$ with $\lambda_0 = 1$. This is zero if and only if λ is an eigenvalue of the MEP (3.1) with the corresponding multiindex. Similarly, we define

$$\tilde{F}_{\mathbf{i}} \colon \mathbb{S}^m \to \mathbb{R}^m, \quad \lambda = (\lambda_0, \dots, \lambda_m) \mapsto \tilde{F}_{\mathbf{i}}(\lambda) = (\varepsilon_{1,i_1}(\lambda), \dots, \varepsilon_{m,i_m}(\lambda)) \tag{3.12}$$

for the homogeneous MEP (3.2). If the MEP is locally definite, it follows from Theorem 3.3 that $\tilde{F}_{\mathbf{i}}$ has exactly two zeros for every multiindex, one of positive sign and one of negative sign. If the MEP is right definite, then $F_{\mathbf{i}}$ has a unique zero.

Next, we employ Newton's method for the functions $F_{\mathbf{i}}$. For this, we require the derivative of $F_{\mathbf{i}}$. It follows from [Kat76, Chapter 2, Theorem 5.15, Theorem 5.16, and Theorem 6.1] that $\varepsilon_{k,i_k}(\lambda)$ is analytic in λ if the eigenvalue $\varepsilon_{k,i_k}(\lambda)$ of $\sum_{\ell=0}^{m} \lambda_\ell A_{k\ell}$ is simple. Otherwise, $\varepsilon_{k,i_k}(\lambda)$ still admits a generalized gradient in the sense of [Cla75, Definition 1.1]. If $\varepsilon_{k,i_k}(\lambda)$ is a simple eigenvector with corresponding eigenvalue u_k in the unit sphere $\mathcal{U}_k$, then a simple calculation shows

$$\varepsilon'_{k,i_k}(\lambda)\Delta\lambda = \sum_{\ell=1}^{k} u_k^{\mathsf{H}} \Delta\lambda_\ell A_{k\ell} u_k. \tag{3.13}$$

If $\varepsilon_{k,i_k}(\lambda)$ is no simple eigenvalue, the generalized gradient $\partial\varepsilon_{k,i_k}(\lambda)$ consists of the convex hull of the linear maps given by the one-sided directional derivatives

$$\{J_{u_k} : J_{u_k}\Delta\lambda = \sum_{\ell=1}^{k} u_k^{\mathsf{H}} \Delta\lambda_\ell A_{k\ell} u_k \text{ and } u_k \text{ is an eigenvector corresponding to } \varepsilon_{k,i_k}(\lambda)\}.$$

This can be seen by applying the characterization for generalized gradients of max-functions in [Cla75, Theorem 2.1] two times to

$$\varepsilon_{k,i_k}(\lambda) = \max_{\substack{U\in\mathbb{C}^{n_k\times i_k}\\ U^{\mathsf{H}}U=\mathrm{id}_{i_k}}} \min_{\substack{x\in\mathbb{C}^{i_k}\\ \|x\|=1}} \sum_{\ell=1}^{k} x^{\mathsf{H}}U^{\mathsf{H}}\lambda_\ell A_{k\ell}Ux.$$

This characterization of eigenvalues is known as the minimax principle for Hermitian matrices; see e.g., [Bha97, Chapter III]. The functions ε_{k,i_k} are even strongly semismooth [SS02], i.e.,

$$\sup_{J\in\partial\varepsilon_{k,i_k}(\lambda+\Delta\lambda)} \|\varepsilon_{k,i_k}(\lambda+\Delta\lambda)-\varepsilon_{k,i_k}(\lambda)+J\Delta\lambda\| \in O(\|\Delta\lambda\|^2). \tag{3.14}$$

If a function from $\mathbb{R}^n$ into $\mathbb{R}^n$ is semismooth, a semismooth Newton method can be applied. If the function is strongly semismooth, then local quadratic convergence is retained; see e.g., [IS14, Chapter 2] for an overview.

It follows that, whenever all eigenvalues $\varepsilon_{k,i_k}(\lambda)$ for $k=1,\dots,m$ are simple, $F_{\mathbf{i}}(\lambda)$ is analytic in λ and its derivative is given by

$$F_{\mathbf{i}}'(\lambda) = \begin{pmatrix} u_1^{\mathsf{H}}A_{11}u_1 & \dots & u_1^{\mathsf{H}}A_{1m}u_1 \\ \vdots & & \vdots \\ u_m^{\mathsf{H}}A_{m1}u_m & \dots & u_m^{\mathsf{H}}A_{mm}u_m \end{pmatrix}, \tag{3.15}$$

where $u_k \in \mathcal{U}_k$ are the eigenvectors of $\sum_{\ell=0}^{m}\lambda_\ell A_{k\ell}$ corresponding to the eigenvalue $\varepsilon_{k,i_k}(\lambda)$. If the MEP is right definite, then $F'(\lambda)$ is invertible. Similarly, Clarke's generalized Jacobian $\partial F_{\mathbf{i}}(\lambda)$ is given by the convex hull of

$$\left\{\begin{pmatrix} u_1^{\mathsf{H}}A_{11}u_1 & \dots & u_1^{\mathsf{H}}A_{1m}u_1 \\ \vdots & & \vdots \\ u_m^{\mathsf{H}}A_{m1}u_m & \dots & u_m^{\mathsf{H}}A_{mm}u_m \end{pmatrix} : u_k\in\mathcal{U}_k \text{ is an eigenvector corresponding to } \varepsilon_{k,i_k}(\lambda)\right\}.$$

Since the functions ε_{k,i_k} are strongly semismooth, $F_{\mathbf{i}}$ is also strongly semismooth; see e.g., [IS14, Proposition 1.73]. It follows that a semismooth Newton method applied to $F_{\mathbf{i}}$ converges locally quadratic; see e.g., [IS14, Theorem 2.42].

Given $\lambda^{(j)}$, the semismooth Newton method computes the next iterate by

$$J\in\partial F_{\mathbf{i}}\left(\lambda^{(j)}\right),\quad F_{\mathbf{i}}\left(\lambda^{(j)}\right)+J\Delta\lambda=0,\quad \lambda^{(j+1)}=\lambda^{(j)}+\Delta\lambda.$$

We can compute a particular choice of $J\in\partial F_{\mathbf{i}}\left(\lambda^{(j)}\right)$ with eigenvectors $u_k\in\mathcal{U}_k$ corresponding to $\varepsilon_{k,i_k}\left(\lambda^{(j)}\right)$ as in (3.15). Then $F_{\mathbf{i}}\left(\lambda^{(j)}\right)$ is given by

$$F_{\mathbf{i}}\left(\lambda^{(j)}\right) = W(u_1,\dots,u_m)\begin{pmatrix}1\\ \lambda^{(j)}\end{pmatrix} = \begin{pmatrix} u_1^{\mathsf{H}}A_{10}u_1 & \dots & u_1^{\mathsf{H}}A_{1m}u_1 \\ \vdots & & \vdots \\ u_m^{\mathsf{H}}A_{m0}u_m & \dots & u_m^{\mathsf{H}}A_{mm}u_m \end{pmatrix}\begin{pmatrix}1\\ \lambda_1^{(j)}\\ \vdots\\ \lambda_m^{(j)}\end{pmatrix}.$$

Algorithm 1 Semismooth Newton method for right definite MEPs

```
Require: A right definite MEP of the form (3.1) and a multiindex $\mathbf{i}$
Ensure: approximation of an eigenvalue $\lambda$ of index $\mathbf{i}$
  intial guess $\lambda^{(0)} \in \mathbb{R}^m$
  for $j = 1, 2, \dots$ do
      for $k = 1, \dots, m$ do
          compute $i_k$-th largest eigenvalue of $\sum_{\ell=0}^{m} \lambda_\ell^{(j-1)} A_{k\ell}$
          and a corresponding eigenvector $u_k \in \mathcal{U}_k$
      end for
      solve $W(u_1, \dots, u_m) \begin{pmatrix} 1 \\ \lambda^{(j)} \end{pmatrix} = 0$
  end for
  return $\lambda^{(j)}$
```

The next iterate $\lambda^{(j+1)}$ is given by

$$0 = F_{\mathbf{i}}\left(\lambda^{(j)}\right) + J\Delta\lambda = \begin{pmatrix} u_1^{\mathsf{H}} A_{10} u_1 & \dots & u_1^{\mathsf{H}} A_{1m} u_1 \\ \vdots & & \vdots \\ u_m^{\mathsf{H}} A_{m0} u_m & \dots & u_m^{\mathsf{H}} A_{mm} u_m \end{pmatrix} \begin{pmatrix} 1 \\ \lambda_1^{(j)} + \Delta\lambda_1 \\ \vdots \\ \lambda_m^{(j)} + \Delta\lambda_m \end{pmatrix},$$

i.e., as the solution of the linear equation

$$W(u_1, \dots, u_m) \begin{pmatrix} 1 \\ \lambda^{(j+1)} \end{pmatrix} = 0. \tag{3.16}$$

If the MEP (3.1) is right definite, this equation has a unique solution. We summarize this procedure in Algorithm 1.

We can employ a similar algorithm for the homogeneous MEP (3.2). In that case, we require a signed index $(\mathbf{i}, \sigma)$ and solve the linear equation

$$W(u_1, \dots, u_m)\lambda^{(j+1)} = 0 \quad \text{with} \quad \lambda^{(j+1)} \in \mathcal{P}^\sigma. \tag{3.17}$$

The other steps are analogous.

Let us shortly consider Algorithm 1 in the case $m = 1$. In that case, the MEP (3.1) is just the generalized eigenvalue problem

$$Au + \lambda Bu = 0$$

with positive definite B. Algorithm 1 computes an eigenvalue of a given index by a Newton method that requires solving an eigenvalue problem in each step. Obviously, it is computationally less expensive to solve the generalized eigenvalue problem directly. The following consequence of Sylvester's law of inertia shows that the index of λ corresponds to the ordering of the eigenvalues of the generalized eigenvalue problem.

Algorithm 2 Semismooth splitted Newton method for right definite MEPs

```
Require: A right definite MEP of the form (3.1) and a multiindex $\mathbf{i}$
Ensure: approximation of an eigenvalue $\lambda$ of index $\mathbf{i}$
  intial guess $\lambda^{(0)} \in \mathbb{R}^m$
  for $j = 1, 2, \dots$ do
      for $k = 2, \dots, m$ do
          compute $i_k$-th largest eigenvalue of $\sum_{\ell=0}^m \lambda_\ell^{(j-1)} A_{k\ell}$
          and a corresponding eigenvector $u_k \in \mathcal{U}_k$
      end for
      solve $W^{(1)}(u_2, \dots, u_m) \begin{pmatrix} 1 \\ \lambda^{(\text{part})} + t\,\lambda^{(\text{hom})} \end{pmatrix} = 0$ for $\lambda^{(\text{part})}$ and $\lambda^{(\text{hom})}$
      find $t$ such that 0 is the $i_1$-th largest eigenvalue of
      $A_{10} + \sum_{\ell=1}^m \lambda_\ell^{(\text{part})} A_{1\ell} + t \sum_{\ell=1}^m \lambda_\ell^{(\text{hom})} A_{1\ell}$
      set $\lambda^{(j)} = \lambda^{(\text{part})} + t\,\lambda^{(\text{hom})}$
  end for
  return $\lambda^{(j)}$
```

Lemma 3.6. *Let A, B be symmetric matrices and B be positive definite. Then zero is the i-th largest eigenvalue of the matrix $A + \lambda B$ if and only if λ is the i-th smallest eigenvalue of the generalized eigenvalue problem $Au + \lambda Bu = 0$.*

Proof. Since B is positive definite, it admits a positive definite square root $B^{\frac{1}{2}}$. By Sylvester's law of inertia [HJ90, Theorem 4.5.8] the matrix $A+\lambda B$ has the same number of positive and negative eigenvalues as the matrix $B^{-\frac{1}{2}}AB^{-\frac{1}{2}}+\lambda\,\mathrm{id}$. Since the eigenvalues of the generalized eigenvalue problem $Au + \lambda Bu = 0$ are the same as the ones of the matrix $-B^{-\frac{1}{2}}AB^{-\frac{1}{2}}$, the claim follows. □

As a consequence, we can solve the generalized eigenvalue problem $A + \lambda B$ and directly recover the index of the eigenvalue. We can utilize this idea to construct a modified method, where one component of $F_{\mathbf{i}}$ is forced to be zero in each step. For this, first consider the case $n_2 = \dots = n_m = 1$. Then (3.1) is of the form

$$A_{10}u_1 + \lambda_1 A_{11} u_1 + \dots + \lambda_m A_{1m} u_1 = 0, \quad b + G\lambda = 0,$$

where $b \in \mathbb{R}^{m-1}$ and $G \in \mathbb{R}^{m-1\times m}$ contain the second to m-th equation of the MEP. If the MEP is right definite, then $\operatorname{rank} G = m-1$ and its kernel is given by $\operatorname{span}\{\lambda^{(\text{hom})}\}$ and the solution of $b + G\lambda = 0$ is given by $\lambda = \lambda^{(\text{part})} + t\lambda^{(\text{hom})}$. Hence, the MEP reduces to

$$A_{10}u_1 + \left(\sum_{\ell=1}^m \lambda_\ell^{(\text{part})} A_{1\ell}\right) u_1 + t\left(\sum_{\ell=1}^m \lambda_\ell^{(\text{hom})} A_{1\ell}\right) u_1 = 0,$$

a generalized eigenvalue problem. Here, $\sum_{\ell=1}^m \lambda_\ell^{(\text{hom})} A_{1\ell}$ is definite and we may choose $\lambda^{(\text{hom})}$ such that $\sum_{\ell=1}^m \lambda_\ell^{(\text{hom})} A_{1\ell}$ is postive definite. This can be seen as follows. Right

Algorithm 3 Semismooth cyclic splitted Newton method for right definite MEPs

```
Require: A right definite MEP of the form (3.1) and a multiindex $\mathbf{i}$
Ensure: approximation of an eigenvalue $\lambda$ of index $\mathbf{i}$
  intial guess $\lambda^{(0)} \in \mathbb{R}^m$
  for $j = 1, 2, \dots$ do
    set $\tilde{k} = j \mod m$
    for $k \neq j \mod m$ do
      compute $i_k$-th largest eigenvalue of $\sum_{\ell=0}^{m} \lambda_\ell^{(j-1)} A_{k\ell}$
      and a corresponding eigenvector $u_k \in \mathcal{U}_k$
    end for
    solve $W^{(\tilde{k})}(u_1, \dots, u_{\tilde{k}-1}, u_{\tilde{k}+1}, \dots, u_m) \begin{pmatrix} 1 \\ \lambda^{(\mathrm{part})} + t\,\lambda^{(\mathrm{hom})} \end{pmatrix} = 0$
    for $\lambda^{(\mathrm{part})}$ and $\lambda^{(\mathrm{hom})}$
    find $t$ such that 0 is the $i_{\tilde{k}}$-th largest eigenvalue of
    $A_{\tilde{k}0} + \sum_{\ell=1}^{m} \lambda_\ell^{(\mathrm{part})} A_{\tilde{k}\ell} + t \sum_{\ell=1}^{m} \lambda_\ell^{(\mathrm{hom})} A_{\tilde{k}\ell}$
    set $\lambda^{(j)} = \lambda^{(\mathrm{part})} + t\,\lambda^{(\mathrm{hom})}$
  end for
  return $\lambda^{(j)}$
```

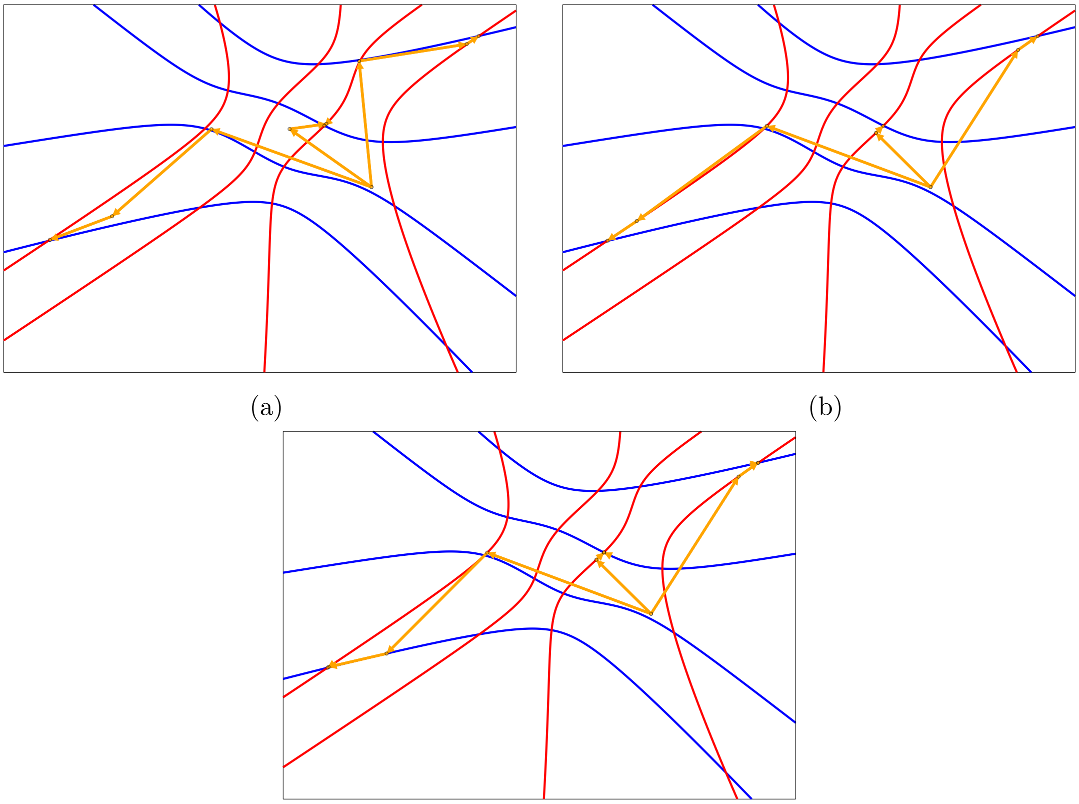
definiteness implies that not both $u_1^{\mathsf{H}} \left(\sum_{\ell=1}^{m} \lambda_\ell^{(\mathrm{hom})} A_{1\ell} \right) u_1 = 0$ and $G\lambda^{(\mathrm{hom})} = 0$ for every $u_1 \in \mathcal{U}_1$. Since $\lambda^{(\mathrm{hom})}$ is in the nullspace of G, this implies $u_1^{\mathsf{H}} \left(\sum_{\ell=1}^{m} \lambda_\ell^{(\mathrm{hom})} A_{1\ell} \right) u_1 \neq 0$ for every $u_1 \in \mathcal{U}_1$ and therefore $\sum_{\ell=1}^{m} \lambda_\ell^{(\mathrm{hom})} A_{1\ell}$ is either positive or negative definite. Therefore, we can use Lemma 3.6 to find λ such that $b + G\lambda = 0$ and $\varepsilon_{1,i_1}(\lambda) = 0$ for any $i_1 \in \{1, \dots, n_1\}$.

We can now apply this idea to a single step of the Newton method. Given $\lambda^{(j)}$, we compute $\varepsilon_{k,i_k}\left(\lambda^{(j)}\right)$ and corresponding eigenvectors $u_k \in \mathcal{U}_k$ for $k \neq \tilde{k}$. We denote by $W^{(\tilde{k})}(u_1, \dots, u_{\tilde{k}-1}, u_{\tilde{k}+1}, \dots, u_m)$ the matrix $W(u_1, \dots, u_m)$ without its $\tilde{k}$-th row, i.e.,

$$W^{(\tilde{k})}(u_1, \dots, u_{\tilde{k}-1}, u_{\tilde{k}+1}, \dots, u_m) = \begin{pmatrix} u_1^{\mathsf{H}} A_{10} u_1 & \dots & u_m^{\mathsf{H}} A_{1m} u_m \\ \vdots & & \vdots \\ u_{\tilde{k}-1}^{\mathsf{H}} A_{\tilde{k}-1,0} u_{\tilde{k}-1} & \dots & u_{\tilde{k}-1}^{\mathsf{H}} A_{\tilde{k}-1,m} u_{\tilde{k}-1} \\ u_{\tilde{k}+1}^{\mathsf{H}} A_{\tilde{k}+1,0} u_{\tilde{k}+1} & \dots & u_{\tilde{k}+1}^{\mathsf{H}} A_{\tilde{k}+1,m} u_{\tilde{k}+1} \\ \vdots & & \vdots \\ u_m^{\mathsf{H}} A_{m0} u_m & \dots & u_m^{\mathsf{H}} A_{mm} u_m \end{pmatrix}.$$

Next, we find $\lambda^{(\mathrm{hom})} \neq 0$ and $\lambda^{(\mathrm{part})}$ such that

$$W^{(\tilde{k})}(u_1, \dots, u_{\tilde{k}-1}, u_{\tilde{k}+1}, \dots, u_m) \begin{pmatrix} 1 \\ \lambda^{(\mathrm{part})} + t\,\lambda^{(\mathrm{hom})} \end{pmatrix} = 0$$



(a)

(b)

(c)

Figure 3.1: Visualization of Algorithm 1 in (a), Algorithm 2 in (b), and Algorithm 3 in (c) for $m = 2$ and three different multiindices.

for every t. Finally, we use Lemma 3.6 and definiteness of $\sum_{\ell=1}^{m} \lambda_{\ell}^{(\mathrm{hom})} A_{\tilde{k}\ell}$ to find the next iterate of the form $\lambda^{(j+1)} = \lambda^{(\mathrm{part})} + t\,\lambda^{(\mathrm{hom})}$ such that $\varepsilon_{\tilde{k},i_{\tilde{k}}}\left(\lambda^{(j)}\right) = 0$. We summarize this procedure in Algorithm 2 in the case that $\tilde{k} = 1$ for every step. Another possibility is use a different $\tilde{k}$ in every step, for example $\tilde{k} = j \mod m$. This is summarized in Algorithm 3. In the case $m = 2$, Algorithm 3 coincides with the method proposed in [EN22].

Changing $\tilde{k}$ in each step has the advantage, that one of the required eigenvalue problems was already solved in the previous step, and thus reducing the computational complexity of this method.

We visualized the three methods in Figure 3.1. In there, a two-parameter problem is shown where the red curves correspond to the eigencurves of the first equation and the blue curves correspond to the eigencurves of the second equation. It can be seen that Algorithm 2 forces the first equation to be satisfied as the iterates in orange stay on the red curves. The Algorithm 3 cycles between the red and blue curves. One can also observe that the next iterate of Algorithm 3 is given by the intersection of the tangent at the current curve and the other curve as described in [EN22, Section 2.3].

All of these methods admit local quadratic convergence. To see this, we have the following more general result.

Theorem 3.7. *Let $F_1\colon \mathbb{R}^m \to \mathbb{R}^{m_1}$ and $F_2\colon \mathbb{R}^m \to \mathbb{R}^{m_2}$ with $m_1 + m_2 = m$ be strongly semismooth, and $(F_1, F_2)(x^*) = 0$. Furthermore, assume there is $\gamma < \infty$ such that*

$$\left\| \begin{pmatrix} J_1 \\ \int_0^1 J_2(t)\, dt \end{pmatrix}^{-1} \right\| \leq \gamma$$

for all $J_1 \in \partial F_1(x)$, $J_2(t) \in \partial F_2(ty + (1-t)z)$ and $x, y, z \in \mathbb{R}^m$. Given $x^{(j)} \in \mathbb{R}^m$, let $J_1 \in \partial F_1\left(x^{(j)}\right)$ and let $x^{(j+1)} \in \mathbb{R}^m$ satisfy the equations

$$F_1\left(x^{(j)}\right) + J_1\left(x^{(j+1)} - x^{(j)}\right) = 0 \quad \text{and} \quad F_2\left(x^{(j+1)}\right) = 0.$$

Then $\left\|x^* - x^{(j+1)}\right\| \in O\left(\left\|x^* - x^{(j)}\right\|^2\right)$.

Proof. Since F_2 is semismooth, it is also locally Lipschitz continuous. Therefore, $f(t) = F_2\left(tx^* + (1-t)x^{(j+1)}\right)$ is locally Lipschitz continuous, almost everywhere differentiable, and the fundamental theorem of calculus applies, i.e.,

$$0 = F_2\left(x^*\right) - F_2\left(x^{(j+1)}\right) = \int_0^1 f'(t)\, dt.$$

Hence, by [IS14, Proposition 1.63], there is $J_2(t) \in \partial F_2\left(tx^* + (1-t)x^{(j+1)}\right)$ such that $J_2(t)\left(x^* - x^{(j+1)}\right) = f'(t)$, that is, there is an element in the generalized Jacobian that attains the directional derivative, and we get

$$0 = F_2(x^*) - F_2\left(x^{(j+1)}\right) = \int_0^1 J_2(t)\, dt\ \left(x^* - x^{(j+1)}\right).$$

Furthermore, since F_1 is strongly semismooth, the first equation for $x^{(j+1)}$ implies

$$\left\|J_1\left(x^* - x^{(j+1)}\right)\right\| = \left\|F_1\left(x^{(j)}\right) - F_1\left(x^*\right) + J_1\left(x^* - x^{(j)}\right)\right\| \in O\left(\left\|x^* - x^{(j)}\right\|^2\right).$$

Hence,

$$\begin{pmatrix} J_1 \\ \int_0^1 J_2(t)\, dt \end{pmatrix} \left(x^* - x^{(j+1)}\right) = \begin{pmatrix} F_1\left(x^{(j)}\right) - F_1\left(x^*\right) + J_1\left(x^* - x^{(j)}\right) \\ 0 \end{pmatrix}$$

and the result follows as the matrix on the left-hand side has a bounded inverse. □

This result is applicable for Algorithm 1, Algorithm 2, and Algorithm 3. We already noted that the functions ε_{k,i_k} are strongly semismooth. It remains to provide that the matrix consisting of elements of the generalized Jacobian and averaged elements of the generalized Jacobian has a uniformly bounded inverse. This is a consequence of the following statement for right definite MEPs.

Proposition 3.8. *Let the MEP* (3.1) *be right definite and let*

$$\mathcal{J} := \left\{ \begin{pmatrix} u_1^{\mathsf{H}} A_{11} u_1 & \dots & u_1^{\mathsf{H}} A_{1m} u_1 \\ \vdots & & \vdots \\ u_m^{\mathsf{H}} A_{m1} u_m & \dots & u_m^{\mathsf{H}} A_{mm} u_m \end{pmatrix} : u_k \in \mathcal{U}_k \text{ for } k = 1, \dots, m \right\}.$$

Then there is $\gamma < \infty$ such that $\left\| J^{-1} \right\| \leq \gamma$ for all $J \in \operatorname{conv} \mathcal{J}$.

Proof. First note that $\mathcal{J}$ is a compact set in $\mathbb{R}^{m \times m}$ as it is the image of the continuous map

$$J \colon \mathcal{U}_1 \times \ldots \times \mathcal{U}_m \to \mathbb{R}^{m \times m}, \quad u = (u_1, \dots, u_d) \mapsto J(u) = \begin{pmatrix} u_1^{\mathsf{H}} A_{11} u_1 & \dots & u_1^{\mathsf{H}} A_{1m} u_1 \\ \vdots & & \vdots \\ u_m^{\mathsf{H}} A_{m1} u_m & \dots & u_m^{\mathsf{H}} A_{mm} u_m \end{pmatrix}$$

of the compact set $\mathcal{U}_1 \times \ldots \times \mathcal{U}_m$. It follows from Carathéodory's theorem that $\operatorname{conv} \mathcal{J}$ is itself compact since $\mathbb{R}^{m \times m}$ is finite dimensional; see e.g., [Tuy16, Theorem 1.1 and Corollary 1.9]. Finally, we show that $\det J \neq 0$ for $J \in \operatorname{conv} \mathcal{J}$. Then $\operatorname{conv} \mathcal{J}$ is a compact subset of invertible matrices and therefore the set $\{J^{-1} \colon J \in \operatorname{conv} \mathcal{J}\}$ is compact as inverting a matrix is a continuous map on invertible matrices.

We proceed by induction. We denote by $\mathcal{J}_k$ the set consisting of the k-th row of matrices in $\mathcal{J}$, i.e.,

$$\mathcal{J}_k = \{ \begin{pmatrix} u_k^{\mathsf{H}} A_{k1} u_k & \dots & u_k^{\mathsf{H}} A_{km} u_k \end{pmatrix} : u_k \in \mathcal{U}_k \}.$$

Furhthermore, we denote by $\mathcal{V}_k$ the set of matrices where the rows $\ell = 1, \dots, k$ consist of the convex combinations $J_\ell \in \operatorname{conv} \mathcal{J}_j$ and the other rows consist of $J_\ell \in \mathcal{J}_\ell$ for $\ell = k+1, \dots, m$, i.e.,

$$\mathcal{V}_k = \left\{ \begin{pmatrix} J_1 \\ \vdots \\ J_m \end{pmatrix} : J_\ell \in \operatorname{conv} \mathcal{J}_\ell \text{ for } \ell = 1, \dots, k \text{ and } J_\ell \in \mathcal{J}_\ell \text{ for } \ell = k+1, \dots, m \right\}.$$

From right definiteness, it follows that $\det J > 0$ for $J \in \mathcal{J} = \mathcal{V}_0$. Now let $\det J > 0$ for all $J \in \mathcal{V}_{k-1}$. Then any $J \in \mathcal{V}_k$ is of the form $J = \sum_{j=1}^{M} \sigma_j J^{(j)}$ where $\sigma_j \geq 0$,,

$\sum_{j=1}^{M} \sigma_j = 1$, and

$$J^{(j)} = \begin{pmatrix} J_1 \\ \vdots \\ J_{k-1} \\ J_k^{(j)} \\ J_{k+1} \\ \vdots \\ J_m \end{pmatrix} \quad \text{for} \quad \begin{cases} J_\ell \in \operatorname{conv} \mathcal{J}_\ell \text{ for } \ell = 1, \dots, k-1, \\ J_k^{(j)} \in \mathcal{J}_k, \\ \text{and} \\ J_\ell \in \mathcal{J}_\ell \text{ for } \ell = k+1, \dots, m. \end{cases}$$

Then by linearity of the determinant in its rows, we have $\det J = \sum_{j=1}^{M} \sigma_j \det J^{(j)} > 0$ as $J^{(j)} \in \mathcal{V}_{k-1}$. It follows that $\det J > 0$ for all $J \in \mathcal{V}_k$ and $k = 0, \dots, m$.

Let $J \in \operatorname{conv} \mathcal{J}$. By definition, $J = \sum_{j=1}^{M} \sigma_j J^{(j)}$ for some $J^{(j)} \in \mathcal{J}$, $\sum_{j=1}^{M} \sigma_j = 1$, and $\sigma_j \geq 0$. Then

$$J = \begin{pmatrix} \sum_{j=1}^{M} \sigma_j J_1^{(j)} \\ \vdots \\ \sum_{j=1}^{M} \sigma_j J_m^{(j)} \end{pmatrix} \quad \text{for some } J_k^{(k)} \in \mathcal{J}_k,\ k = 1, \dots m.$$

It follows that $\operatorname{conv} \mathcal{J} \subset \mathcal{V}_m$ and the assertion is proven. □

The local quadratic convergence of Algorithm 1, Algorithm 2, and Algorithm 3 readily follows.

Theorem 3.9. *Let the MEP (3.1) be right definite and λ be the unique eigenvalue of multiindex $\mathbf{i}$. Then the sequences $\lambda^{(j)}$ generated by Algorithm 1, Algorithm 2, and Algorithm 3 satisfy*

$$\left\| \lambda - \lambda^{(j+1)} \right\| \in O\left(\left\| \lambda - \lambda^{(j)} \right\|^2 \right),$$

i.e., they admit local quadratic convergence.

Proof. We only consider Algorithm 2 as the proofs for the other ones are analogous. Let $F_2 = \varepsilon_{1,i_1}$ and

$$F_1 = \begin{pmatrix} \varepsilon_{2,i_2} \\ \vdots \\ \varepsilon_{m,i_m} \end{pmatrix}.$$

Both are strongly semicontinuous [SS02]. The iterates satisfy the equations of Theorem 3.7 and Proposition 3.8 implies that the inequality concerning generalized Jacobian is satisfied. The result follows. □

The convergence results are only of local nature. In principle globalization strategies involving a line search are applicable, however, quadratic convergence would be lost. Our numerical results however suggest that this is not necessary. In Section 3.4 we discuss global convergence results for some eigenvalues.

The results of Theorem 3.7 and Proposition 3.8 also imply local quadratic convergence when two or more lines of $F_{\mathbf{i}}$ are kept zero. This is in principle applicable when an effective method to compute an eigenvalue of multiindex for a two-parameter eigenvalue is available.

3.4 Extreme eigenvalues

The problem of finding an eigenvalue of extreme multiindix $\mathbf{i} \in \{1, n_1\} \times \ldots \times \{1, n_m\}$ has an optimization perspective. Note again, that eigenvalues of the MEP (3.1) lie in the set

$$\mathcal{Q} = \left\{ \lambda \in \mathbb{R}^m \colon W(u) \begin{pmatrix} 1 \\ \lambda \end{pmatrix} = 0 \text{ for some } u \in \mathcal{U}_1 \times \ldots \times U_m \right\}.$$

The set $\mathcal{Q}$ is not convex, as can be anticipated by Figure 3.1. The eigenvalues of extreme indices are however extreme points of its convex hull.

Proposition 3.10. *Let the MEP* (3.1) *be right definite. Then*

$$\mathcal{Q} \subset \operatorname{conv} \left\{ \lambda^{\mathbf{i}} \colon \mathbf{i} \in \{1, n_1\} \times \ldots \times \{1, n_m\} \right\}$$

where $\lambda^{\mathbf{i}}$ denotes the eigenvalue of multiindex $\mathbf{i}$.

Proof. Let $\lambda \in \mathcal{Q}$ and let $u \in \mathcal{U}_1 \times \ldots \times \mathcal{U}_m$ be a corresponding vector such that $W(u) \begin{pmatrix} 1 \\ \lambda \end{pmatrix} = 0$. We denote $\mathcal{E} = \{1, n_1\} \times \ldots \times \{1, n_m\}$ as the set of extreme multiindices. Next, notice that

$$W(u) \begin{pmatrix} 1 \\ \lambda^{\mathbf{i}} \end{pmatrix} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_m \end{pmatrix}$$

where $\epsilon_k \geq 0$ if $i_k = n_k$ and $\epsilon_k \leq 0$ if $i_k = 1$. Hence,

$$0 \in \operatorname{conv} \left\{ W(u) \begin{pmatrix} 1 \\ \lambda^{\mathbf{i}} \end{pmatrix} : \mathbf{i} \in \mathcal{E} \right\} = W(u) \left\{ \begin{pmatrix} 1 \\ \lambda \end{pmatrix} : \lambda \in \operatorname{conv} \left\{ \lambda^{\mathbf{i}} \colon \mathbf{i} \in \mathcal{E} \right\} \right\}.$$

Since $W(u)$ is of full rank, the solution of $W(u) \begin{pmatrix} 1 \\ \lambda \end{pmatrix} = 0$ is unique and the assertion follows. □

It follows directly that

$$\min_{\lambda \in \mathcal{Q}} \mu^{\mathsf{T}} \lambda$$

is attained at an eigenvalue of extreme multiindex. If the MEP (3.1) is left definite with respect to μ and right definite, then we know which extreme multiindex corresponds to a maximizing eigenvalue.

Proposition 3.11. *Let the MEP* (3.1) *be right definite and left definite with respect to μ. Then*

$$\min_{\lambda \in \mathcal{Q}} \mu^{\mathsf{T}} \lambda$$

is attained at the eigenvalue $\lambda^{\mathbf{1}}$ of multiindex $\mathbf{1} = (1, \dots, 1)$. Furthermore, there is a $\beta > 0$ such that $\mu^{\mathsf{T}} \left(\lambda - \lambda^{\mathbf{1}} \right) \geq \beta \left\| \lambda - \lambda^{\mathbf{1}} \right\|$ for all $\lambda \in \mathcal{Q}$.

Proof. Let $\lambda \in \mathcal{Q}$ and choose $u \in \mathcal{U}_1 \times \dots \times \mathcal{U}_m$ such that $W(u) \begin{pmatrix} 1 \\ \lambda \end{pmatrix} = 0$. Note that the entries of $W(u) \begin{pmatrix} 1 \\ \lambda^{\mathbf{1}} \end{pmatrix}$ are nonpositive since 0 is the largest eigenvalue of the matrices $\sum_{\ell=0}^{m} \lambda_\ell^{\mathbf{1}} A_{k\ell}$ for $k = 1, \dots, m$. Hence,

$$J(u) \left(\lambda^{\mathbf{1}} - \lambda \right) := \begin{pmatrix} u_1^{\mathsf{H}} A_{11} u_1 & \dots & u_1^{\mathsf{H}} A_{1m} u_1 \\ \vdots & & \vdots \\ u_m^{\mathsf{H}} A_{m1} u_m & \dots & u_m^{\mathsf{H}} A_{mm} u_m \end{pmatrix} \left(\lambda^{\mathbf{1}} - \lambda \right) = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_m \end{pmatrix} \tag{3.18}$$

with $\epsilon_k \leq 0$ for $k = 1, \dots, m$. If $\lambda \neq \lambda^{\mathbf{1}}$, then at least one inequality is strict and by Proposition 3.8 there is $\tilde{\gamma}$

$$\tilde{\gamma} \max_{k=1,\dots,m} |\epsilon_k| \geq \left\| \lambda^{\mathbf{1}} - \lambda \right\|.$$

By right definiteness the matrix $J(u)$ is invertible, and by Cramer's rule and the condition (3.7) for left definiteness the entries of $\mu^{\mathsf{T}} (J(u))^{-1}$ are positive. Furthermore, by compactness of the sets $\mathcal{U}_1, \dots, \mathcal{U}_m$ there is an $\alpha > 0$ such that the entries are at least α for every $u \in \mathcal{U}_1 \times \ldots \times \mathcal{U}_m$. Hence,

$$\mu^{\mathsf{T}} \left(\lambda^{\mathbf{1}} - \lambda \right) = \mu^{\mathsf{T}} (J(u))^{-1} J(u) \left(\lambda^{\mathbf{1}} - \lambda \right) \leq \frac{\alpha}{\tilde{\gamma}} \left\| \lambda^{\mathbf{1}} - \lambda \right\|,$$

which shows the assertion. □

A similar argument also shows that $\mu^{\mathsf{T}} \left(\lambda^{\mathbf{i}} - \lambda^{\mathbf{j}} \right) \leq 0$ whenever $i_k \leq j_k$ for $k = 1, \dots, m$. That is, the map $\mathbf{i} \mapsto \mu^{\mathsf{T}} \lambda^{\mathbf{i}}$ is order preserving with respect to the product order on multiindices.

Proposition 3.12. *Let the MEP* (3.1) *be right definite and left definite with respect to μ. Let $\lambda^{\mathbf{i}}$ and $\lambda^{\mathbf{j}}$ be eigenvalues of multiindices $\mathbf{i}$ and $\mathbf{j}$. Then*

$$\mu^{\mathsf{T}} \lambda^{\mathbf{i}} \leq \mu^{\mathsf{T}} \lambda^{\mathbf{j}}$$

if $i_k \leq j_k$ for $k = 1, \dots, m$.

Proof. We show that there is a $u \in \mathcal{U}_1 \times \cdots \times \mathcal{U}_m$ such that $J(u)\left(\lambda^{\mathbf{i}} - \lambda^{\mathbf{j}}\right)$ has nonpositive entries. Then the result follows analogously as in the proof of Proposition 3.11. Since 0 is the j_k-th largest eigenvalue of $\sum_{\ell=0}^{m} \lambda_\ell^{\mathbf{j}} A_{k\ell}$, the i_k-th largest eigenvalue of $\sum_{\ell=0}^{m} \lambda_\ell^{\mathbf{i}} A_{k\ell}$, and $i_k \leq j_k$, there is a $u_k \in \mathcal{U}_k$ such that $u_k^{\mathsf{H}} \sum_{\ell=0}^{m} \left(\lambda_\ell^{\mathbf{i}} - \lambda_\ell^{\mathbf{j}}\right) A_{k\ell} u_k \leq 0$. It follows that $J(u_1, \ldots, u_m)\left(\lambda_\ell^{\mathbf{i}} - \lambda_\ell^{\mathbf{j}}\right)$ has nonpositive entries and the result follows. □

Hence, if the MEP (3.1) is left and right definite, we can compute eigenvalues that have small $\mu^{\mathsf{T}}\lambda$ using Algorithm 1, Algorithm 2, or Algorithm 3 by requiring small multiindices. Finally, left definiteness also implies global convergence of the methods when the desired eigenvalue has multiindex $\mathbf{1}$.

Theorem 3.13. *Let the MEP 3.1 be left and right definite. The sequences $\lambda^{(j)}$ generated by Algorithm 1, Algorithm 2, and Algorithm 3 converge globally to $\lambda^{\mathbf{1}}$ when the required multiindex is $\mathbf{1}$.*

Proof. First notice that $\lambda^{(j)} \in \mathcal{Q}$ for $j \geq 1$. Thus, due to Proposition 3.11 it is enough to show that $\mu^{\mathsf{T}}\lambda^{(j)}$ converges to $\mu^{\mathsf{T}}\lambda^{\mathbf{1}}$. From Proposition 3.8 and the characterization of the directional derivatives of $F_{\mathbf{1}}$, it follows that

$$\gamma \|F_{\mathbf{1}}(\lambda)\| \geq \left\|\lambda - \lambda^{\mathbf{1}}\right\|,$$

and furthermore the entries of $F_{\mathbf{1}}$ are nonnegative by the definition of $\mathcal{Q}$ and $F_{\mathbf{1}}(\lambda)$. It follows, that the largest entry of $F_{\mathbf{1}}(\lambda)$ is at least $\delta\mu^{\mathsf{T}}(\lambda - \lambda^{\mathbf{1}})$ with some $\delta > 0$ independent of λ. The difference $\lambda^{(j)} - \lambda^{(j+1)}$ satisfies

$$J\left(\lambda^{(j)} - \lambda^{(j+1)}\right) = F_{\mathbf{1}}\left(\lambda^{(j)}\right)$$

for some $J \in \operatorname{conv}\mathcal{J}$ of Proposition 3.8. With a similar argument as in the proof of Proposition 3.8, we get that $\mu^{\mathsf{T}}J^{-1}$ has entries larger than some α for every $J \in \operatorname{conv}\mathcal{J}$. Hence,

$$\mu^{\mathsf{T}}\left(\lambda^{(j)} - \lambda^{(j+1)}\right) \geq \alpha\delta\mu^{\mathsf{T}}\left(\lambda^{(j)} - \lambda^{\mathbf{1}}\right)$$

and thus

$$\mu^{\mathsf{T}}\left(\lambda^{(j+1)} - \lambda^{\mathbf{1}}\right) \leq (1 - \alpha\delta)\,\mu^{\mathsf{T}}\left(\lambda^{(j)} - \lambda^{\mathbf{1}}\right),$$

that is global linear convergence of the sequence $\mu^{\mathsf{T}}\lambda^{(j)}$ to $\mu^{\mathsf{T}}\lambda^{\mathbf{1}}$. □

We want to note that the proofs of Proposition 3.11, Proposition 3.12, and Theorem 3.13 only require the condition (3.7) and not the negative definiteness of the matrices A_{k0}. However, assuming left definiteness is not a real restriction. Indeed, the negative definiteness can be achieved by a translation of eigenvalues; see e.g., [Vol88, Lemma 2.7.3]. In the case $m = 2$ right definiteness even implies condition (3.7).

Proposition 3.14. *Let the MEP 3.1 be right definite and $m = 2$. Then there is a $\mu = (0, \mu_1, \mu_2)$ such that condition* (3.7) *is satisfied.*

Proof. The condition (3.7) is equivalent to positive definiteness of the matrices

$$\mu_2 A_{11} - \mu_1 A_{12} \quad \text{and} \quad \mu_1 A_{22} - \mu_2 A_{21}.$$

We use Lemma 3.5 with $\sigma_1 = 1$ and $\sigma_2 = -1$ and get positive definite matrices

$$\alpha_1 A_{11} + \alpha_2 A_{12} \quad \text{and} \quad -\alpha_1 A_{21} - \alpha_2 A_{22}.$$

The result follows with $\mu_1 = -\alpha_2$ and $\mu_2 = \alpha_1$. □

It follows that Algorithm 1, Algorithm 2, and Algorithm 3 converge for extreme indices when (3.1) is a right definite two-parameter eigenvalue problem.

Corollary 3.15. *Let the MEP 3.1 be right definite and $m = 2$. The sequences $\lambda^{(j)}$ generated by Algorithm 1, Algorithm 2, and Algorithm 3 converge globally to $\lambda^{\mathbf{i}}$ when the required multiindex $\mathbf{i}$ is extreme, i.e., $\mathbf{i} \in \{1, n_1\} \times \{1, n_2\}$.*

Proof. If $\mathbf{i} = \mathbf{1}$, the result follows directly from Theorem 3.13 and Proposition 3.14. The case $\mathbf{i} = (1, n_2)$ follows by permuting equations and changing signs of the matrices A_{20}, A_{21}, and A_{22}. This is again a right definite two-parameter eigenvalue problem with the same eigenvalues as the original but the eigenvalue of multiindex $(1, n_2)$ of the original one now has multiindex $\mathbf{1}$. The other two cases follow analogously. □

3.5 Numerical experiments

In this section, we demonstrate the perfomance of Algorithm 1, Algorithm 2, and Algorithm 3 and compare it with the performance of methods in the MATLAB toolbox MultiParEig [Ple22]. All numerical experiments were run on one core Intel Xeon Gold 6144 at 3.5 GHz.

We want to note that we did not implement our methods to the highest efficiency. For example, all methods are parallelizable in multiple fashions. The computation of an eigenvalue of a certain multiindex is independent of the computation of an eigenvalue with a different multiindex. These can therefore be computed in parallel. When computing only one eigenvalue, the computation of the m different eigenvalues in the second loop of the algorithms can be done in parallel as well. The methods can also be made more efficient by some precalculations. For example, Algorithm 2 and Algorithm 3 are more efficient, when the matrices $A_{k\ell}$ for $k, \ell = 1, \ldots, m$ are diagonal, as the generalized eigenvalue problems can then easily be transformed into a symmetric eigenvalue problem. For $m = 2$, this can be achieved by congruence transformations beforehand, reducing computational complexity.

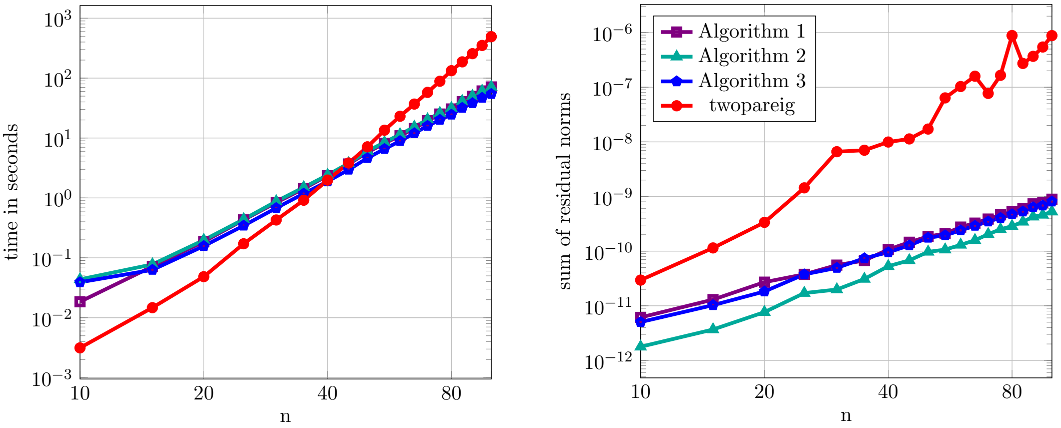


Figure 3.2: Comparison of the performances of Algorithm 1, Algorithm 2, Algorithm 3, and twopareig [Ple22] for $m = 2$ and different sizes of matrices. The matrices are generated randomly such that the MEP is right definite as explained in Section 3.5.1 with $n_1 = n_2 = n$. We solve for all n^2 eigenvalues.

Throughout the experiments, we track the precision of the found eigenvalues and eigenvectors by the sum of residual norms $\sum_{k=1}^{m} \|\sum_{\ell=0}^{m} \lambda_\ell A_{k\ell} u_k\|$. If we look for more than a single eigenvalue, we sum up these sums of residual norms.

3.5.1 Randomly generated MEPs

At first, we generated matrices for an MEP randomly such that the resulting MEP is right definite. For $m = 2$, we achieve this by setting

$$A_{11} = A_{22} = \mathrm{id}_n$$

and setting A_{12} and A_{21} as diagonal matrices with entries chosen uniform at random in the interval $[-1, 1]$. Now with probability one the MEP is right definite. For A_{10} and A_{20} we generate symmetric matrices at random with entries distributed normally.

We used Algorithm 1, Algorithm 2, Algorithm 3, and twopareig [Ple22], which uses the generalized Schur decomposition, to find every eigenvalue for $n \times n$ matrices $A_{\ell k}$ and $n = 10, 15, \ldots, 100$. The results are summarized in Figure 3.2. Our methods find eigenvalues with higher precision by orders of magnitude, where Algorithm 2 has a slight edge, and for $n > 40$ our methods need less time to find all eigenvalues. This was to be expected, as the complexity of finding one eigenvalue with our methods is in $O(mn^3)$ if we assume that the number of iterations does not depend on n and m. This leads to a computational complexity of $O(n^5)$ for all eigenvalues in the case $m = 2$ compared to a complexity of $O(n^6)$ using the generalized Schur decomposition. Figure 3.2 even suggests a complexity of $O(n^4)$.

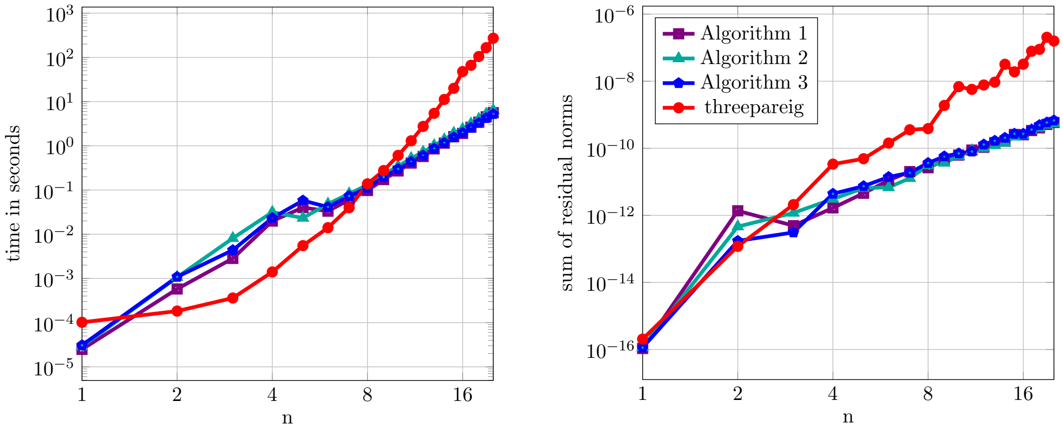


Figure 3.3: Comparison of the performances of Algorithm 1, Algorithm 2, Algorithm 3, and threepareig [Ple22] for $m = 3$ and different sizes of matrices. The matrices are generated randomly such that the MEP is right definite as explained in Section 3.5.1 with $n_1 = n_2 = n_3 = n$. We solve for all n^3 eigenvalues.

As a next example, we generated right definite three-parameter eigenvalue problems at random. Here, we set

$$A_{11} = A_{22} = A_{33} = \mathrm{id}_n$$

and set A_{12}, A_{13}, A_{21}, A_{23}, A_{31}, A_{32} as diagonal matrices with entries uniformly distributed in $[-\frac{1}{2}, \frac{1}{2}]$ and A_{k0} again as symmetric matrices with Gaussian distributed entries. This also leads to a right definite three-parameter eigenvalue problem with probability one.

For these three-parameter eigenvalue problems, we compared our methods to threepareig [Ple22], which also uses the generalized Schur decomposition. We used our methods and threepareig to find all n^3 eigenvalues with $n \times n$ matrices $A_{\ell k}$ with $n = 1, 2, \ldots, 20$. The results are summarized in Figure 3.3. Threepareig is only able to solve MEPs with small n, as it explicitly constructs the $n^3 \times n^3$ matrices Δ_k in (3.4) and runs into memory problems even for matrices of moderate size. For $n > 8$ our methods has a smaller time demand, and Algorithm 1, Algorithm 2, and Algorithm 3 have a similar performance. For $m = 3$ the complexity of our method is $O(n^6)$ for finding all eigenvalues, again assuming the number of iterations is independent of n. This is small compared to the computational complexity using the generalized Schur decomposition, which is of the order $O(n^9)$. The time plot in Figure 3.3 suggests this is the case. It is therefore more feasible to find all eigenvalues for larger three-parameter eigenvalue problems with our methods. The measured precision is again higher by orders of magnitude.

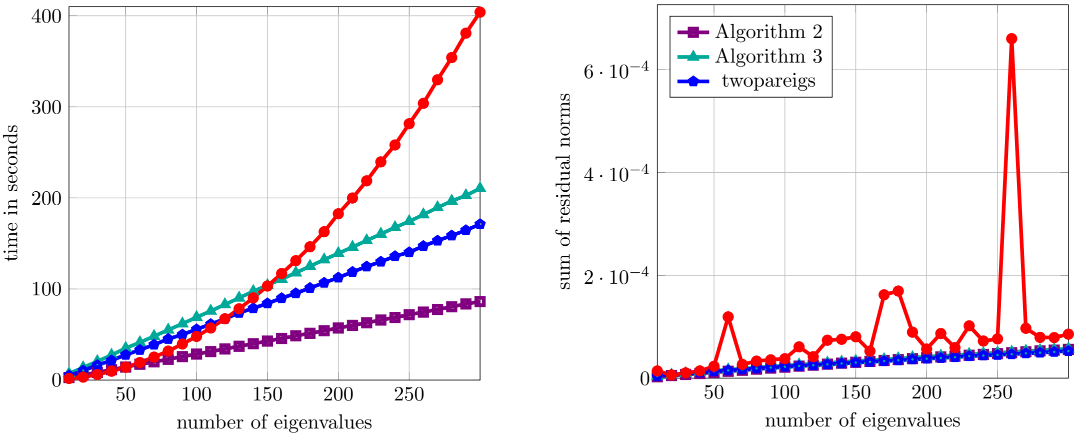


Figure 3.4: Comparison of the performances of Algorithm 1, Algorithm 2, Algorithm 3, and twopareigs [Ple22] for a discretization of (3.19) on polynomial bases of degree 300. We compute eigenvalues with small λ.

3.5.2 Coupled Mathieu's differential equations

As a first example coming from boundary eigenvalue problems, we considered the coupled version of Mathieu's modified and Mathieu's equation

$$\begin{aligned} P''(\rho) &+ \lambda \cosh(\rho)\, P(\rho) - \nu\, P(\rho) = 0 \quad \text{for } \rho \in (0, r), \quad P(0) = 0 = P(1), \\ \Phi''(\varphi) &- \lambda \cos(\varphi)\, \Phi(\varphi) + \nu\, \Phi(\varphi) = 0 \quad \text{for } \varphi \in (0, \pi), \quad \Phi(0) = 0 = \Phi(\pi). \end{aligned} \tag{3.19}$$

This is one of four different configurations arising when separation of variables is applied to the Helmholtz equation

$$\begin{aligned} \Delta u(x) + \lambda\, u(x) &= 0 \qquad \text{for } x \in \Omega \\ u(x) &= 0 \qquad \text{for } x \in \partial\Omega \end{aligned}$$

on the ellipse $\Omega = \{(x, y) \colon x^2/\cosh(1) + y^2/\sinh(1) < 1\}$, see e.g., the introduction of [Vol88]. We discretized these equations on a basis of polynomials of degree 300 using the MATLAB toolbox Chebfun [DHT14]. We computed $L^2(0,1)$- and $L^2(0,\pi)$-orthonormal bases B_1 and B_2 satisfying the boundary conditions as column vectors and computed the symmetric matrices

$$\begin{aligned} A_{10} &= -\int_0^1 B_1'(\rho) B_1'(\rho)^\mathsf{T}\, d\rho, & A_{20} &= -\int_0^\pi B_2'(\varphi) B_2'(\varphi)^\mathsf{T}\, d\varphi, \\ A_{11} &= \int_0^1 \cosh(\rho) B_1(\rho) B_1(\rho)^\mathsf{T}\, d\rho, & A_{21} &= -\int_0^\pi \cos(\varphi) B_2(\varphi) B_2(\varphi)^\mathsf{T}\, d\varphi, \\ A_{12} &= -\int_0^1 B_1(\rho) B_1(\rho)^\mathsf{T}\, d\rho, & A_{22} &= \int_0^\pi B_2(\varphi) B_2(\varphi)^\mathsf{T}\, d\varphi \end{aligned}$$

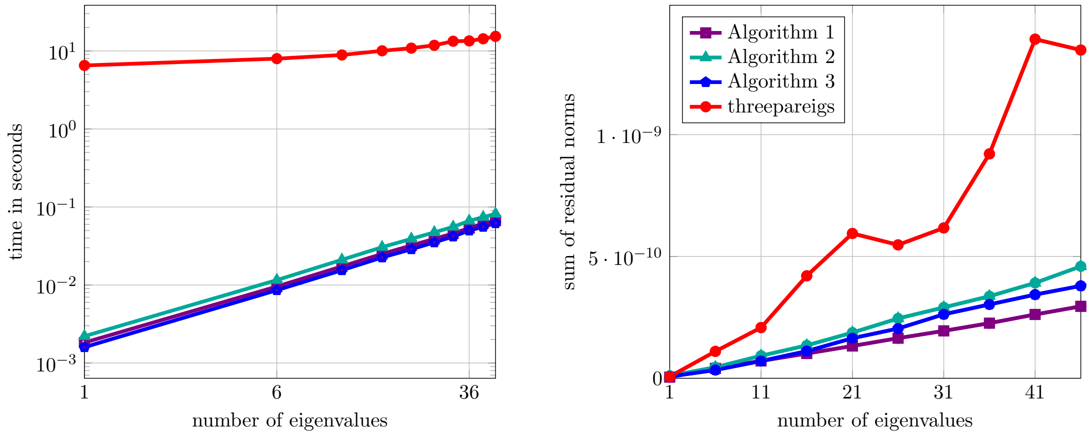


Figure 3.5: Comparison of the performances of Algorithm 1, Algorithm 2, Algorithm 3, and threepareigs [Ple22] for a discretization of (3.20) on polynomial bases of degree 16 with boundary conditions (3.21). We compute eigenvalues with small λ_3.

with Chebfun. The resulting MEP is left definite with respect to $\mu = (0, 1, 0)$ and right definite. We can thus apply Proposition 3.12 to find eigenvalues and eigenvectors with small λ by requiring small multiindices in Algorithm 1, Algorithm 2, and Algorithm 3. For this we impose the product order on multiindices, i.e., $\mathbf{i} \leq \mathbf{j}$ if $i_k \leq j_k$ for $k = 1, \ldots, m$. At first we compute the eigenvalue of multiindex $\mathbf{1}$. Next we choose a minimal multiindex $\mathbf{i}$ among the ones we have not computed an eigenvalue to and that minimizes $\min_{\mathbf{j} \in \mathcal{N}(\mathbf{i})} \lambda^{\mathbf{j}}$, where $\mathcal{N}(\mathbf{i}) = \{\mathbf{j} \in \{1, \ldots, n_1\} \times \ldots \times \{1, \ldots, n_m\} \colon i_{\tilde{k}} - j_{\tilde{k}} = 1$ for a single $\tilde{k}$ and $i_k - j_k = 0$ otherwise$\}$ is a set of neighbors of $\mathbf{i}$.

Using Algorithm 1, Algorithm 2, and Algorithm 3 and this strategy, we computed $5, 10, \ldots, 300$ eigenvalues with small λ and compared their performances to twopareigs [Ple22], which uses a Krylov-Schur method [MP15]. The results are summarized in Figure 3.4. For this problem, Algorithm 1 performed the best. Our methods need the same time for every eigenvalue, whereas twopareigs required more time per eigenvalue when computing many eigenvalues. When computing more than 50 eigenvalues Algorithm 1 requires less time than twopareigs. The precision of the computed eigenvalues and eigenvectors is slightly higher when computed with our methods.

3.5.3 Ellipsoidal wave equation

Now we consider a three-parameter eigenvalue problem, that arises from the Helmholtz equation on the ellipsoid $\Omega = \{(x, y, z) \in \mathbb{R}^3 \colon x^2/x_0^2 + y^2/y_0^2 + z^2/z_0^2\}$. A fitting choice of ellipsoidal coordinates and separation of variables leads to the equations

$$\begin{array}{rcccccccl} p(\xi_1)\left(p(\xi_1)u_1'(\xi_1)\right)' & + & \lambda_1 u_1(\xi_1) & + & \lambda_2\xi_1 u_1(\xi_1) & + & \xi_1^2\lambda_3 u_1(\xi_1) & = & 0 \quad \text{for } \xi_1 \in (c,d),\\ p(\xi_2)\left(p(\xi_2)u_2'(\xi_2)\right)' & - & \lambda_1 u_2(\xi_2) & - & \lambda_2\xi_2 u_2(\xi_2) & - & \xi_2^2\lambda_3 u_2(\xi_2) & = & 0 \quad \text{for } \xi_2 \in (1,c),\\ p(\xi_3)\left(p(\xi_3)u_3'(\xi_3)\right)' & + & \lambda_1 u_3(\xi_3) & + & \lambda_2\xi_3 u_3(\xi_3) & + & \xi_3^2\lambda_3 u_3(\xi_3) & = & 0 \quad \text{for } \xi_3 \in (0,1), \end{array} \tag{3.20}$$

with $p(\xi) = \sqrt{|\xi||\xi-1||\xi-c|}$ and appropriate values of c and d. Here, only λ_3 is connected to the original spectral parameter λ. We again discretize using polynomial bases B_1, B_2, and B_3 with $B_1(d) = 0$ due to Dirichlet boundary conditions. Again using the MATLAB toolbox Chebfun, we computed the symmetric matrices

$$\begin{aligned} A_{10} &= -\int_c^d p(\xi)B_1'(\xi)B_1'(\xi)^\mathsf{T}\,d\xi, & A_{20} &= -\int_1^c p(\xi)B_2'(\xi)B_2'(\xi)^\mathsf{T}\,d\xi,\\ A_{30} &= -\int_0^1 p(\xi)B_3'(\xi)B_3'(\xi)^\mathsf{T}\,d\xi, & A_{11} &= \int_c^d \frac{1}{p(\xi)}B_1(\xi)B_1(\xi)^\mathsf{T}\,d\xi,\\ A_{21} &= -\int_1^c \frac{1}{p(\xi)}B_2(\xi)B_2(\xi)^\mathsf{T}\,d\xi, & A_{31} &= \int_0^1 \frac{1}{p(\xi)}B_3(\xi)B_3(\xi)^\mathsf{T}\,d\xi,\\ A_{12} &= \int_c^d \frac{\xi}{p(\xi)}B_1(\xi)B_1(\xi)^\mathsf{T}\,d\xi, & A_{22} &= -\int_1^c \frac{\xi}{p(\xi)}B_2(\xi)B_2(\xi)^\mathsf{T}\,d\xi,\\ A_{32} &= \int_0^1 \frac{\xi}{p(\xi)}B_3(\xi)B_3(\xi)^\mathsf{T}\,d\xi, & A_{13} &= \int_c^d \frac{\xi^2}{p(\xi)}B_1(\xi)B_1(\xi)^\mathsf{T}\,d\xi,\\ A_{23} &= -\int_1^c \frac{\xi^2}{p(\xi)}B_2(\xi)B_2(\xi)^\mathsf{T}\,d\xi, & A_{33} &= \int_0^1 \frac{\xi^2}{p(\xi)}B_3(\xi)B_3(\xi)^\mathsf{T}\,d\xi. \end{aligned}$$

This discretization leads to the boundary conditions

$$\begin{array}{lcccccccl} cu_3'(0) & + & \lambda_1 u_3(0) & & & & & = & 0,\\ (c-1)u_3'(1) & + & \lambda_1 u_3(1) & + & \lambda_2 u_3(1) & + & \lambda_3 u_3(1) & = & 0,\\ (c-1)u_2'(1) & - & \lambda_1 u_2(1) & - & \lambda_2 u_2(1) & - & \lambda_3 u_2(1) & = & 0,\\ (c^2-c)u_2'(c) & - & \lambda_1 u_2(c) & - & \lambda_2 c u_2(c) & - & \lambda_3 c^2 u_2(c) & = & 0,\\ (c^2-c)u_1'(c) & + & \lambda_1 u_1(c) & + & \lambda_2 c u_1(c) & + & \lambda_3 c^2 u_1(c) & = & 0,\\ u_1(d) & & & & & & & = & 0. \end{array} \tag{3.21}$$

This is one of 8 possible configurations for handling singularities at the boundaries of the interval, see e.g., [HMMP19, Section 2.1]. We set $c = 12/7$ and $d = 16/7$ which corresponds to the ellipsoid with $x_0 = 1$, $y_0 = 1.5$, and $z_0 = 2$, which is also considered in [HMMP19, Section 5.1]. The resulting eigenvalue problem is right definite and left definite with respect to $\mu = (0,0,0,1)$. Thus, we can apply the strategy described in the previous example to find eigenvalues with small λ_3.

First, we discretized on polynomial bases of degree 16, which leads to $n_1 = n_2 = 17$ and $n_3 = 16$. We used our methods and threepareigs [Ple22] to compute $1, 6, \ldots, 46$ eigenvalues with small λ_3. The results are summarized in Figure 3.5. The method

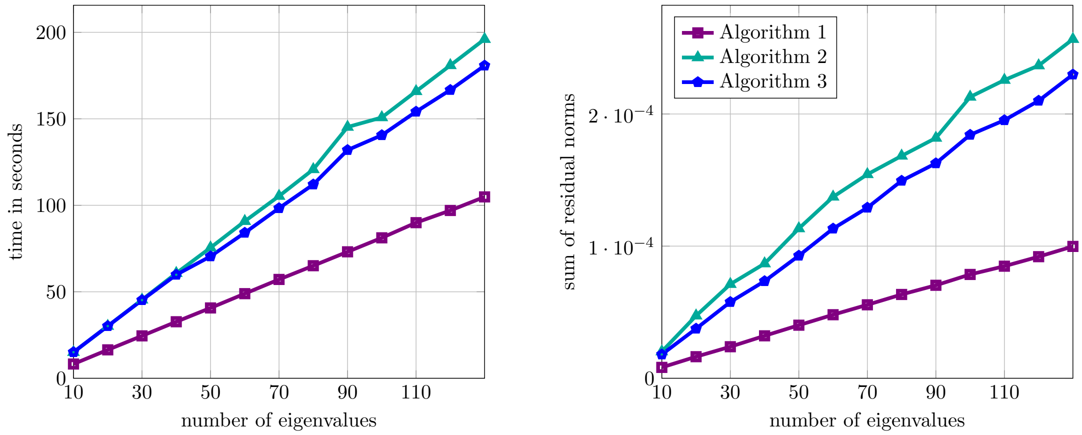


Figure 3.6: Comparison of the performances of Algorithm 1, Algorithm 2, and Algorithm 3 for a discretization of (3.20) on polynomial bases of degree 300 with boundary conditions (3.21). We compute eigenvalues with small λ_3.

threepareigs requires the matrices Δ_0 and Δ_3 in (3.4). Our methods do not require these matrices, which is one reason why the required time for computing some eigenvalues is smaller using our methods. The precision of our methods is again higher. For small matrices, Algorithm 1, Algorithm 2, and Algorithm 3 perform similarly. Algorithm 1 is slightly more precise in this experiment, and Algorithm 3 is slightly faster.

Next, we discretized on polynomial bases of degree 300 in the same way and computed $10, 20, \dots, 130$ eigenvalues with small λ_3. For this size of matrices, threepareigs is no longer applicable, as the matrices Δ_0 and Δ_3 are too large. The methods from in [HMMP19] are not directly applicable for these generated matrices and require a preconditioning step described in their article. Our methods are still applicable, as they only require solving eigenvalue problems with matrices of size ~ 300. The results are summarized in Figure 3.6. In this experiment, Algorithm 1 performed the best. Again, the experiment suggests, that the time required for computing an eigenvalue does not depend on its multiindex.

3.5.4 Locally definite problem

Finally, we apply our methods to a homogeneous MEP of the form (3.2) that is locally definite but not definite. The following example is taken from [Vol88, Chapter 1.5]. Let

$$A_{10} = A_{21} = A_{32} = \begin{pmatrix} 1 & & & \\ & 5 & & \\ & & 1 & \\ & & & 1 \end{pmatrix}, \qquad A_{11} = A_{20} = A_{33} = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & 5 & \\ & & & 1 \end{pmatrix},$$

$$A_{12} = A_{23} = A_{30} = \begin{pmatrix} 5 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{pmatrix}, \qquad A_{13} = A_{22} = A_{31} = -\begin{pmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 5 \end{pmatrix}.$$

The resulting MEP is locally definite which can be seen by applying Lemma 3.5. Indeed, for every sign $\sigma \in \{-1,1\}^3$ one can choose α as one of the vectors $\pm e_i \in \mathbb{R}^4$. It is however not definite. To see this, note that

$$\lambda_1 = \frac{1}{\sqrt{12}}(-1,-3,1,1), \quad \lambda_2 = \frac{1}{\sqrt{12}}(-1,1,1,-3),$$
$$\lambda_3 = \frac{1}{\sqrt{12}}(-1,1,-3,1), \quad \lambda_4 = \frac{1}{\sqrt{12}}(3,1,1,1)$$

are eigenvalues in $\mathcal{P}^+$ and $\lambda_1+\lambda_2+\lambda_3+\lambda_4 = 0$. Hence, there is no μ such that $\mu^\mathsf{T}\lambda > 0$ for all $\lambda \in \mathcal{P}^+$ and the MEP cannot be definite. The eigenvectors are given by the coordinate vectors $e_{i_1} \otimes e_{i_2} \otimes e_{i_3}$. We used the variant of Algorithm 1 that uses (3.17) instead of (3.16) to find the next iterate. We computed the eigenvectors and eigenvalues of each signed index to machine precision indicating that Algorithm 1 is also applicable to locally definite MEPs that are not definite.

3.6 Concluding remarks and outlook

We presented new methods for computing eigenvalues of definite multiparameter eigenvalue problems based on their signed index. Our approaches only require finding certain eigenpairs coming from the original problem. This makes it feasible to find eigenvalues of definite multiparameter eigenvalue problems for larger m using our methods.

Our methods heavily rely on the problem being at least locally definite, so that there is a one-to-one correspondence of signed indices and eigenvalues. Additionally, local definiteness forces all eigenvalues to be real. For the case $m = 1$, we are in the situation of generalized eigenvalue problems. When a generalized eigenvalue problem is almost definite, one can show, using inertia laws, that many eigenvalues are real [NN19]. It would be of interest to investigate the applicability of these results in a situation when the multiparameter eigenvalue problem is almost definite.

Bibliography

[AJ14] A. Arnold and T. Jahnke. On the approximation of high-dimensional differential equations in the hierarchical Tucker format. *BIT*, 54(2):305–341, 2014.

[AKU20] A. Agrachev, K. Kozhasov, and A. Uschmajew. Chebyshev polynomials and best rank-one approximation ratio. *SIAM J. Matrix Anal. Appl.*, 41(1):308–331, 2020.

[ARR14] A. Ahmed, B. Recht, and J. Romberg. Blind deconvolution using convex programming. *IEEE Trans. Inform. Theory*, 60(3):1711–1732, 2014.

[Atk72] F. V. Atkinson. *Multiparameter eigenvalue problems.* Mathematics in Science and Engineering, Vol. 82. Academic Press, New York-London, 1972. Volume I: Matrices and compact operators.

[Ban38] S. Banach. Über homogene Polynome in (L^2). *Studia Math.*, 7:36–44, 1938.

[BEKU21] M. Bachmayr, H. Eisenmann, E. Kieri, and A. Uschmajew. Existence of dynamical low-rank approximations to parabolic problems. *Math. Comp.*, 90(330):1799–1830, 2021.

[Bha97] R. Bhatia. *Matrix analysis*, volume 169 of *Graduate Texts in Mathematics.* Springer-Verlag, New York, 1997.

[BL07] J. Bennett and S. Lanning. The netflix prize. In *KDD Cup and Workshop in conjunction with KDD*, 2007.

[BL14] J. Buczyński and J. M. Landsberg. On the third secant variety. *J. Algebraic Combin.*, 40(2):475–502, 2014.

[BSU16] M. Bachmayr, R. Schneider, and A. Uschmajew. Tensor networks and hierarchical tensors for the solution of high-dimensional partial differential equations. *Found. Comput. Math.*, 16(6):1423–1472, 2016.

[CC70] J. D. Carroll and J.-J. Chang. Analysis of individual differences in multidimensional scaling via an n-way generalization of “Eckart-Young” decomposition. *Psychometrika*, 35(3):283–319, 1970.

[CGLM08] P. Comon, G. Golub, L.-H. Lim, and B. Mourrain. Symmetric tensors and symmetric tensor rank. *SIAM J. Matrix Anal. Appl.*, 30(3):1254–1279, 2008.

[CKP00] F. Cobos, T. Kühn, and J. Peetre. Extreme points of the complex binary trilinear ball. *Studia Math.*, 138(1):81–92, 2000.

[CL55] E. A. Coddington and N. Levinson. *Theory of ordinary differential equations.* McGraw-Hill, New York, 1955.

[CL10] D. Conte and C. Lubich. An error analysis of the multi-configuration time-dependent Hartree method of quantum dynamics. *M2AN Math. Model. Numer. Anal.*, 44(4):759–780, 2010.

[CL22] G. Ceruti and C. Lubich. An unconventional robust integrator for dynamical low-rank approximation. *BIT*, 62(1):23–44, 2022.

[Cla75] F. H. Clarke. Generalized gradients and applications. *Trans. Amer. Math. Soc.*, 205:247–262, 1975.

[Cla90] F. H. Clarke. *Optimization and nonsmooth analysis.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second edition, 1990.

[CR09] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009.

[CS13] D. Cartwright and B. Sturmfels. The number of eigenvalues of a tensor. *Linear Algebra Appl.*, 438(2):942–952, 2013.

[CSV13] E. J. Candès, T. Strohmer, and V. Voroninski. PhaseLift: exact and stable signal recovery from magnitude measurements via convex programming. *Comm. Pure Appl. Math.*, 66(8):1241–1274, 2013.

[DDGS16] W. Dahmen, R. DeVore, L. Grasedyck, and E. Süli. Tensor-sparsity of solutions to high-dimensional elliptic partial differential equations. *Found. Comput. Math.*, 16(4):813–874, 2016.

[DHT14] T. A Driscoll, N. Hale, and L. N. Trefethen. *Chebfun Guide.* Pafnuty Publications, 2014.

[Die84] J. Diestel. *Sequences and series in Banach spaces*, volume 92 of *Graduate Texts in Mathematics.* Springer-Verlag, New York, 1984.

[Dir30] P. A. M. Dirac. Note on exchange phenomena in the thomas atom. *Mathematical Proceedings of the Cambridge Philosophical Society*, 26(3):376–385, 1930.

[DLDMV00a] L. De Lathauwer, B. De Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.*, 21(4):1253–1278, 2000.

[DLDMV00b] L. De Lathauwer, B. De Moor, and J. Vandewalle. On the best rank-1 and rank-$(R_1, R_2, \cdots, R_N)$ approximation of higher-order tensors. *SIAM J. Matrix Anal. Appl.*, 21(4):1324–1342, 2000.

[DOT18] J. Draisma, G. Ottaviani, and A. Tocino. Best rank-k approximations for tensors: generalizing Eckart-Young. *Res. Math. Sci.*, 5(2):Paper No. 27, 13, 2018.

[dSL08] V. de Silva and L.-H. Lim. Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM J. Matrix Anal. Appl.*, 30(3):1084–1127, 2008.

[DYY16] B. Dong, B. Yu, and Y. Yu. A homotopy method for finding all solutions of a multiparameter eigenvalue problem. *SIAM J. Matrix Anal. Appl.*, 37(2):550–571, 2016.

[EN22] H. Eisenmann and Y. Nakatsukasa. Solving two-parameter eigenvalue problems using an alternating method. *Linear Algebra Appl.*, 643:137–160, 2022.

[EU21] H. Eisenmann and A. Uschmajew. Maximum relative distance between symmetric rank-two and rank-one tensors. arXiv:2111.12611, 2021.

[EY36] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[FHN19] A. Falcó, W. Hackbusch, and A. Nouy. On the Dirac-Frenkel variational principle on tensor Banach spaces. *Found. Comput. Math.*, 19(1):159–204, 2019.

[GH14] M. Griebel and H. Harbrecht. Approximation of bi-variate functions: singular value decomposition versus sparse grids. *IMA J. Numer. Anal.*, 34(1):28–54, 2014.

[GK14] L. Grubišić and D. Kressner. On the eigenvalue decay of solutions to operator Lyapunov equations. *Systems Control Lett.*, 73:42–47, 2014.

[Gra04a] L. Grasedyck. Existence and computation of low Kronecker-rank approximations for large linear systems of tensor product structure. *Computing*, 72(3-4):247–265, 2004.

[Gra04b] L. Grasedyck. Existence of a low rank or $\mathcal{H}$-matrix approximant to the solution of a Sylvester equation. *Numer. Linear Algebra Appl.*, 11(4):371–389, 2004.

[Gre67] W. H. Greub. *Multilinear algebra*. Die Grundlehren der mathematischen Wissenschaften, Band 136. Springer-Verlag New York, Inc., New York, 1967.

[Hac19] W. Hackbusch. *Tensor spaces and numerical tensor calculus*, volume 56 of *Springer Series in Computational Mathematics*. Springer, Cham, second edition, 2019.

[Har70] R. A. Harshman. Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-model factor analysis. In *UCLA Working Papers in Phonetics*, 1970.

[HJ90] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge, 1990. Corrected reprint of the 1985 original.

[HK09] W. Hackbusch and S. Kühn. A new scheme for the tensor representation. *J. Fourier Anal. Appl.*, 15(5):706–722, 2009.

[HKP05] M. E. Hochstenbach, T. Košir, and B. Plestenjak. A Jacobi-Davidson type method for the two-parameter eigenvalue problem. *SIAM J. Matrix Anal. Appl.*, 26(2):477–497, 2004/05.

[HL13] C. J. Hillar and L.-H. Lim. Most tensor problems are NP-hard. *J. ACM*, 60(6):Art. 45, 39, 2013.

[HMMP19] E. Hochstenbach, M, K. Meerbergen, E. Mengi, and B. Plestenjak. Subspace methods for three-parameter eigenvalue problems. *Numer. Linear Algebra Appl.*, 26(4):e2240, 22, 2019.

[HP02] M. E. Hochstenbach and B. Plestenjak. A Jacobi-Davidson type method for a right definite two-parameter eigenvalue problem. *SIAM J. Matrix Anal. Appl.*, 24(2):392–410, 2002.

[IS14] A. F. Izmailov and M. V. Solodov. *Newton-type methods for optimization and variational problems*. Springer Series in Operations Research and Financial Engineering. Springer, Cham, 2014.

[JH09] E. Jarlebring and M. E. Hochstenbach. Polynomial two-parameter eigenvalue problems and matrix pencil methods for stability of delay-differential equations. *Linear Algebra Appl.*, 431(3-4):369–380, 2009.

[JL84] W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in modern analysis and probability (New Haven, Conn., 1982)*, volume 26 of *Contemp. Math.*, pages 189–206. Amer. Math. Soc., Providence, RI, 1984.

[Kat76] T. Kato. *Perturbation theory for linear operators.* Grundlehren der Mathematischen Wissenschaften, Band 132. Springer-Verlag, Berlin-New York, second edition, 1976.

[KBV09] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[KL07a] O. Koch and C. Lubich. Dynamical low-rank approximation. *SIAM J. Matrix Anal. Appl.*, 29(2):434–454, 2007.

[KL07b] O. Koch and C. Lubich. Regularity of the multi-configuration time-dependent Hartree approximation in quantum molecular dynamics. *M2AN Math. Model. Numer. Anal.*, 41(2):315–331, 2007.

[KL10] O. Koch and C. Lubich. Dynamical tensor approximation. *SIAM J. Matrix Anal. Appl.*, 31(5):2360–2375, 2010.

[KM15] X. Kong and D. Meng. The bounds for the best rank-1 approximation ratio of a finite dimensional tensor space. *Pac. J. Optim.*, 11(2):323–337, 2015.

[Kru77] J. B. Kruskal. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra Appl.*, 18(2):95–138, 1977.

[Lan12] J. M. Landsberg. *Tensors: geometry and applications*, volume 128 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2012.

[Lev94] T. Levitina. On numerical solution of multiparameter Sturm-Liouville spectral problems. In *Numerical analysis and mathematical modelling*, volume 29 of *Banach Center Publ.*, pages 275–281. Polish Acad. Sci. Inst. Math., Warsaw, 1994.

[Lev99] T. V. Levitina. On a numerical solution of some three-parameter spectral problems. *Zh. Vychisl. Mat. Mat. Fiz.*, 39(11):1787–1801, 1999.

[LNSU18] Z. Li, Y. Nakatsukasa, T. Soma, and A. Uschmajew. On orthogonal tensors and best rank-one approximation ratio. *SIAM J. Matrix Anal. Appl.*, 39(1):400–425, 2018.

[LO14] C. Lubich and I. V. Oseledets. A projector-splitting integrator for dynamical low-rank approximation. *BIT*, 54(1):171–188, 2014.

[LOV15] C. Lubich, I. V. Oseledets, and B. Vandereycken. Time integration of tensor trains. *SIAM J. Numer. Anal.*, 53(2):917–941, 2015.

[LRSV13] C. Lubich, T. Rohwedder, R. Schneider, and B. Vandereycken. Dynamical approximation by hierarchical Tucker and tensor-train tensors. *SIAM J. Matrix Anal. Appl.*, 34(2):470–494, 2013.

[Lub08] C. Lubich. *From quantum to classical molecular dynamics: reduced models and numerical analysis*. Zurich Lectures in Advanced Mathematics. European Mathematical Society (EMS), Zürich, 2008.

[LZ20] Z. Li and Y.-B. Zhao. On norm compression inequalities for partitioned block tensors. *Calcolo*, 57(1):Paper No. 11, 27, 2020.

[Mir60] L. Mirsky. Symmetric gauge functions and unitarily invariant norms. *Quart. J. Math. Oxford Ser. (2)*, 11:50–59, 1960.

[MP15] K. Meerbergen and B. Plestenjak. A Sylvester-Arnoldi type method for the generalized eigenvalue problem with two-by-two operator determinants. *Numer. Linear Algebra Appl.*, 22(6):1131–1146, 2015.

[MS54] J. Meixner and F. W. Schäfke. *Mathieusche Funktionen und Sphäroidfunktionen mit Anwendungen auf physikalische und technische Probleme*. Die Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen mit besonderer Berücksichtigung der Anwendungsgebiete, Band LXXI. Springer-Verlag, Berlin-Göttingen-Heidelberg, 1954.

[MW02] V. Mehrmann and D. Watkins. Polynomial eigenvalue problems with Hamiltonian structure. *Electron. Trans. Numer. Anal.*, 13:106–118, 2002.

[NN19] Y. Nakatsukasa and V. Noferini. Inertia laws and localization of real eigenvalues for generalized indefinite eigenvalue problems. *Linear Algebra Appl.*, 578:272–296, 2019.

[OT09] I. V. Oseledets and E. E. Tyrtyshnikov. Breaking the curse of dimensionality, or how to use SVD in many dimensions. *SIAM J. Sci. Comput.*, 31(5):3744–3759, 2009.

[Ple00] B. Plestenjak. A continuation method for a right definite two-parameter eigenvalue problem. *SIAM J. Matrix Anal. Appl.*, 21(4):1163–1184, 2000.

[Ple22] B. Plestenjak. MultiParEig. `https://www.mathworks.com/matlabcentral/fileexchange/47844-multipareig`, MATLAB Central File Exchange. Retrieved April 26, 2022, 2022.

[Pod08] B. M. Podlevskiĭ. On the application of Newton's method to the determination of the eigenvalue of some two-parameter (multiparameter) spectral problems. *Zh. Vychisl. Mat. Mat. Fiz.*, 48(12):2107–2112, 2008.

[PW18] N. S. Papageorgiou and P. Winkert. *Applied nonlinear functional analysis*. De Gruyter Graduate. De Gruyter, Berlin, 2018. An introduction.

[Qi11] L. Qi. The best rank-one approximation ratio of a tensor space. *SIAM J. Matrix Anal. Appl.*, 32(2):430–442, 2011.

[RFP10] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, 52(3):471–501, 2010.

[RJ21] E. Ringh and E. Jarlebring. Nonlinearizing two-parameter eigenvalue problems. *SIAM J. Matrix Anal. Appl.*, 42(2):775–799, 2021.

[Rou13] T. Roubíček. *Nonlinear partial differential equations with applications*, volume 153 of *International Series of Numerical Mathematics*. Birkhäuser/Springer Basel AG, Basel, second edition, 2013.

[SBG04] A. Smilde, R. Bro, and P. Geladi. *Multi-way Analysis: Applications in the Chemical Sciences*. Wiley, 2004.

[Sch07] E. Schmidt. Zur Theorie der linearen und nichtlinearen Integralgleichungen. I. Teil: Entwicklung willkürlicher Funktionen nach Systemen vorgeschriebener. *Mathematische Annalen*, 63:433–476, 1907.

[Shi95] A. Shimony. Degree of entanglement. *Annals of the New York Academy of Sciences*, 755(1):675–679, 1995.

[Shi18] Y. Shitov. A counterexample to Comon's conjecture. *SIAM J. Appl. Algebra Geom.*, 2(3):428–443, 2018.

[Sho97] R. E. Showalter. *Monotone operators in Banach space and nonlinear partial differential equations*, volume 49 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 1997.

[Sim87] J. Simon. Compact sets in the space $L^p(0,T;B)$. *Ann. Mat. Pura Appl. (4)*, 146:65–96, 1987.

[SNTI16] S. Sakaue, Y. Nakatsukasa, A. Takeda, and S. Iwata. Solving generalized CDT problems via two-parameter eigenvalues. *SIAM J. Optim.*, 26(3):1669–1694, 2016.

[SS02] D. Sun and J. Sun. Strong semismoothness of eigenvalues of symmetric matrices and its application to inverse eigenvalue problems. *SIAM J. Numer. Anal.*, 40(6):2352–2367 (2003), 2002.

[ST86] T. Slivnik and G. Tomšič. A numerical method for the solution of two-parameter eigenvalue problems. *J. Comput. Appl. Math.*, 15(1):109–115, 1986.

[SU14] R. Schneider and A. Uschmajew. Approximation rates for the hierarchical tensor format in periodic Sobolev spaces. *J. Complexity*, 30(2):56–71, 2014.

[SW79] D. Schmidt and G. Wolf. A method of generating integral relations by the simultaneous separability of generalized schrödinger equations. *SIAM Journal on Mathematical Analysis*, 10(4):823–838, 1979.

[Tes12] G. Teschl. *Ordinary differential equations and dynamical systems*, volume 140 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2012.

[Tuy16] H. Tuy. *Convex analysis and global optimization*, volume 110 of *Springer Optimization and Its Applications*. Springer, [Cham], 2016. Second edition [of MR1615096].

[UT19] M. Udell and A. Townsend. Why are big data matrices approximately low rank? *SIAM J. Math. Data Sci.*, 1(1):144–160, 2019.

[Vak17] R. Vakil. THE RISING SEA: Foundations of Algebraic Geometry. http://math.stanford.edu/ vakil/216blog/FOAGnov1817public.pdf, 2017.

[Van13] B. Vandereycken. Low-rank matrix completion by Riemannian optimization. *SIAM J. Optim.*, 23(2):1214–1236, 2013.

[Vol88] H. Volkmer. *Multiparameter eigenvalue problems and expansion theorems*, volume 1356 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1988.

[WCCL16] K. Wei, J.-F. Cai, T. F. Chan, and S. Leung. Guarantees of Riemannian optimization for low rank matrix recovery. *SIAM J. Matrix Anal. Appl.*, 37(3):1198–1222, 2016.

[WG03] T.-C. Wei and P. M. Goldbart. Geometric measure of entanglement and applications to bipartite and multipartite quantum states. *Phys. Rev. A*, 68:042307, Oct 2003.

[Zei85] E. Zeidler. *Nonlinear functional analysis and its applications. III*. Springer-Verlag, New York, 1985. Variational methods and optimization, Translated from the German by Leo F. Boron.

[Zei90a] E. Zeidler. *Nonlinear functional analysis and its applications. II/A*. Springer-Verlag, New York, 1990. Linear monotone operators, Translated from the German by the author and Leo F. Boron.

[Zei90b] E. Zeidler. *Nonlinear functional analysis and its applications. II/B*. Springer-Verlag, New York, 1990. Nonlinear monotone operators, Translated from the German by the author and Leo F. Boron.

[Zei95] E. Zeidler. *Applied functional analysis*, volume 109 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1995. Main principles and their applications.

[ZHQ16] X. Zhang, Z.-H. Huang, and L. Qi. Comon's conjecture, rank decomposition, and symmetric rank decomposition of symmetric tensors. *SIAM J. Matrix Anal. Appl.*, 37(4):1719–1728, 2016.

Selbstständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialien oder erbrachten Dienstleistungen als solche gekennzeichnet.

Leipzig, den 7. Juni 2022

. . . .
(Henrik Eisenmann)

Daten zum Autor

Name: Henrik Eisenmann
Geburtsdatum: 08.06.1993 in Husum

10/2012 - 6/2019	Studium der Mathematik Technische Universität Berlin
seit 07/2019	Doktorand Max-Planck-Institut MiS Leipzig

Bibliographische Daten

Multilinear optimization in low-rank models
(Multilineare Optimimierung in niedrig Rang Modellen)
Eisenmann, Henrik
Universität Leipzig, Dissertation, 2022
95 Seiten, 10 Abbildungen, 104 Referenzen