Dissertation

Adaptive Wavelet Methods for Inverse Problems: Acceleration Strategies, Adaptive Rothe Method and Generalized Tensor Wavelets

Ulrich Friedrich

2014

Adaptive Wavelet Methods for Inverse Problems: Acceleration Strategies, Adaptive Rothe Method and Generalized Tensor Wavelets

Dissertation
zur
Erlangung des Doktorgrades
der Naturwissenschaften
(Dr. rer. nat.)

dem

Fachbereich Mathematik und Informatik
der Philipps–Universität Marburg

vorgelegt von

Ulrich Friedrich

aus Marburg/Lahn

Marburg/Lahn Dezember 2014

Vom Fachbereich Mathematik und Informatik
der Philipps-Universität Marburg (Hochschulkennziffer: 1180)
als Dissertation angenommen am: 18. Dezember 2014

Erstgutachter: Prof. Dr. Stephan Dahlke, Philipps-Universität Marburg

Zweitgutachter: Prof. Dr. Rob Stevenson, University of Amsterdam

Tag der mündlichen Prüfung: 23. April 2015

Acknowledgments

In my time as a PhD student I had the privilege to contribute a tiny part to the beautiful and wonderful science of mathematics. My sincere gratitude goes to the people that allowed this to happen and that accompanied me along the way.

First of all I want to thank my supervisor Professor Stephan Dahlke for his guidance, patience and the many fruitful discussions. He gave me the opportunity to be part in his workgroup, which provided an inspiring, productive and very friendly atmosphere. Also, I was allowed to present some of my results at several conferences, in which I was able to both learn a lot and to make valuable experiences.

My thanks also go to Professor Rob P. Stevenson, who was the second referee of my thesis. The mutual visits were very productive and I will always remember the positive atmosphere.

I am grateful for all the support by Professor Thorsten Raasch. During his time as a PhD student he supported me in the preparation of my diploma thesis. Later, we had many discussions about subtle mathematical details and the source code that was the starting point for my own software development.

Further gratitude goes to Professor Massimo Fornasier. His style is both vivid and productive and I was able to learn a lot during our discussions in Munich and Marburg.

I also want to thank all my remaining coauthors: Professor Peter Maaß, Professor Klaus Ritter, Professor René L. Schilling, Professor Felix Lindner, Nabi G. Chegini, Petru A. Cioica-Licht, Rudolph A. Ressel, Nicolas Döhring, and Stefan Kinzel. The various discussions, emails and phone calls contributed a lot to this thesis. It was inspiring to experience the individual approaches to mathematics.

I want to thank all the friends and colleagues in Professor Dahlkes workgroup. Together we answered countless little questions. The friendly atmosphere was a steady source of motivation for my work.

Special thanks go to my family which always supported me, especially in difficult times.

My work was supported by the Deutsche Forschungsgesellschaft, grants DA 360/12-1 and DA 360/12-2. I want to thank the referees and all involved people that believed in the feasibility of this project.

Zusammenfassung

Allgemein kann man inverse Probleme als die Aufgabe beschreiben, aus gegebenen Daten $y^\dagger$ Rückschlüsse auf deren Ursache $u^\dagger$ zu ziehen. Mathematisch kann dies als die Inversion einer Operatorgleichung

$$\mathcal{K}(u^\dagger) = y^\dagger$$

beschrieben werden. Von besonderem Interesse ist hierbei der Fall, dass es sich um ein schlecht konditioniertes oder schlecht gestelltes Problem handelt. Um zu einem sinnvollen Lösungsbegriff zu kommen, können Regularisierungstechniken angewendet werden. Eines der wichtigsten Regularisierungsverfahren ist die sogenannten Tikhonov Regularisierung. Hier werden zusätzliche Annahmen an die Lösung gemacht, was mittels eines Strafterms $\mathscr{F}_\alpha : X \to \mathbb{R}$ modelliert wird. Zur Lösung des inversen Problems betrachtet man zunächst Näherungslösungen

$$u_\alpha^\delta = \arg\min_{u \in X} \frac{1}{2}\|\mathcal{K}(u) - y^\delta\|_Y^2 + \alpha\mathscr{F}(u), \tag{0.1}$$

die auf ungenauen Daten $\|y^\delta - y^\dagger\|_Y \leq \delta$ beruhen.

Die Konvergenz der Folge der Minimierer u_α^δ gegen $u^\dagger$ für eine gegebene Parameterwahl $\alpha = \alpha(\delta)$ wurde in der Vergangenheit intensiv studiert. Diese Regularitätstheorie für inverse Probleme ist in weiten Teilen nahezu vollständig, insbesondere für nichtlineare Operatoren $\mathcal{K}$.

Im Gegensatz dazu ist die Entwicklung von effizienten Lösungsverfahren zur Berechnung der Minimierer u_α^δ ein Feld aktueller Forschung. Ein häufig gemachter erster Ansatz ist, Strafterme $\mathscr{F}$ zu betrachten, die eine einfache Struktur der Lösung erzwingen. Für den Fall, dass X ein Hilbertraum ist, der mit Hilfe einer Riesz Basis diskretisiert wird, sind gewichtete ℓ_p-Normen der Entwicklungskoeffizeiten bezüglich der Basis von Interesse. Für $1 \leq p \leq 2$ ist für solche Strafterme bekannt, dass die Entwicklungskoeffizienten der Minimierer u_α^δ in $\ell_{2(p-1)}$ liegen, was insbesondere für $p = 1$ eine endliche Darstellung bedeutet. Zur Berechnung der Minimierer kommen typischerweise verallgemeinerte Gradientenabstiegsverfahren zum Einsatz. Für gewichtete ℓ_p-Norm Strafterme führen diese auf iterative Verfahren der Gestalt

$$u^{(n)} = \mathbb{S}_\alpha(u^{(n)} - (\mathcal{K}'(u^{(n)}))^*(\mathcal{K}(u^{(n)}) - y^\delta)), \tag{0.2}$$

wobei $\mathbb{S}_\alpha$ ein sogenannter Soft-Shrinkage-Operator ist, der einzelne Entwicklungskoeffizienten dämpft und $(\cdot)^*$ den adjungierten Operator bezeichnet.

Viele praktische Anwendungen führen auf allgemeine nichtlinearer Vorwärtsoperatoren $\mathcal{K}$. Die für diesen Fall bestehenen Verfahren der Bauart (0.2) teilen gemeinsame Schwächen. Im Allgemeinen kann nur gezeigt werden, dass die berechnete Folge $(u^{(n)})_{n\in\mathbb{N}}$ konvergente Teilfolgen besitzt, und das diese gegen stationäre Punkte der rechten Seite von (0.1) konvergieren. Diese Punkte besitzen allerdings keine Regularisierungseigenschaften. Erfahrungsgemäß konvergiert das unmodifizierte Verfahren (0.2) nur sehr langsam. Dabei umfasst jeder Iterationsschritt die Anwendung von $\mathcal{K}$ und der Adjungierten seiner Ableitung. Dies kann für sich genommen bereits numerisch sehr anspruchsvoll sein.

Die näherungsweise Lösung allgemeiner nichtlinearer inverser Probleme stellt eine sehr große numerische Herausforderung dar. Um diese anzugehen, werden in dieser Arbeit zwei Ansätze verfolgt:

$T1$. Entwicklung einer Strategie zur schnellen Minimierung des Problems (0.1) mit beweisbaren Konvergenzeigenschaften.

$T2$. Entwicklung, Analyse und Verallgemeinerung von effizienten numerischen Verfahren für die Teiloperatoren, die in jedem einzelnen Iterationsschritt auftreten.

Ein erstes Ergebnis dieser Arbeit ist eine Beschleunigungsstrategie für die Iteration (0.2), die auf einer streng monoton abfallenden Wahl für $\alpha^{(n)} \searrow \alpha$ beruht. Wenn eine Lipschitz Stetigkeitsannahme für den nichtlinearen Operator $\mathcal{K}$ und seine Ableitung erfüllt ist, und weiterhin eine lokale Kontraktionsbedingung erfüllt ist, konvergiert die Iterationsfolge $(u^{(n)})_{n\in\mathbb{N}}$ mit linearer Rate gegen das gesuchte globale Minimum u_α^δ. Die gemachten Annahmen sind für verschiedene Klassen von nichtlinearen Operatoren erfüllt. Für nichtlineare Probleme, die sich als nichtlineare Störung eines linearen Problems darstellen lassen, kann die erwünschte lokale Kontraktionsbedingung durch eine spezielle Vorkonditionierungsstrategie erfüllt werden.

Eine sehr wichtige Klasse von inversen Problemen stellen Parameteridentifikationsprobleme für partielle Differentialgleichungen dar. Bei diesen Problemen sollen aus gegebenen Daten die Parameter in der zugrundeliegenden partiellen Differentialgleichung rekonstruiert werden. Als Prototyp für diese Klasse von inversen Problemen und zur Motivation der weiteren Vorgehensweise beim Verfolgen von Ansatz $T2$ wird zunächst ein spezielles Parameteridentifikationsproblem für eine parabolische partielle Differentialgleichung betrachtet. Der Vorwärtsoperator wird eingehend analysiert, um die Anwendbarkeit von (0.2) zur Minimierung des Tikhonov Funktionals in (0.1) zu gewährleisten. Dabei muss beachtet werden, dass die zulässigen Parameter sowohl den Einschränkungen aus der Lösbarkeitstheorie parabolischer partieller Differentialgleichungen unterliegen als auch durch die Praxis gegebenen L_∞ Schranken. In diesem Beispiel ist es möglich, die Wirkung des Operators $(\mathcal{K}'(u^{(n)}))^*$ wiederum als die Lösung einer parabolischen Differentialgleichung zu charakterisieren.

Ein Weg parabolische Differentialgleichungen numerisch zu behandeln, ist die sogenannte horizontale Linienmethode, die oft auch als Rothe's Methode bezeichnet

wird. Hierbei wird das parabolische Problem als abstraktes Cauchy Problem aufgefasst und zunächst, üblicherweise implizit, in der Zeit diskretisiert. Dies wird dann mit einer Diskretisierung der resultierenden S-stufigen Systeme von räumlichen Problemen kombiniert. Die Effizienz von Rothe's Methode hängt von den numerischen Verfahren ab, die für die zeitliche und die räumliche Diskretisierung eingesetzt werden. Einen viel versprechenden Ansatz bilden adaptive Diskretisierungsverfahren. Generell handelt es sich hierbei um selbststeuernde Verfahren, die a posteriori Fehlerschätzer einsetzen, um die Diskretisierung an die aktuelle Näherungslösung anzupassen bis eine vorgegebene Fehlertoleranz erreicht ist. Hierdurch werden hochgradig nicht-uniforme Diskretisierungen realisiert. Im Vergleich zu klassischen Diskretisierungen wird dadurch die Anzahl der benötigten Freiheitsgrade, um eine vorgegebene Fehlertoleranz zu erreichen, drastisch reduziert.

Adaptive Verfahren können prinzipiell sowohl für die zeitliche als auch für die räumliche Diskretisierung in Rothe's Methode eingesetzt werden. In dieser Arbeit wird als erster Schritt in diese Richtung die Kombination aus zeitlich uniformer und räumlich adaptiver Diskretisierung betrachtet. Um die Konvergenz dieses Ansatzes sicher zu stellen, wird die S-stufige inexakte Rothe Methode, die sich durch die inexakte Lösung der S Stufengleichungen bis auf gewisse Toleranzen ergibt, im Detail analysiert. Es werden Schranken für diese Toleranzen angegeben, so dass die zeitliche Konvergenzrate der exakten Rothe Methode erhalten bleibt. Linear implizite Zeitdiskretisierungen führen auf S-stufige Systeme von elliptischen Differentialgleichungen im Raum. Besonders effiziente Verfahren zur Lösung der elliptischen Teilprobleme sind asymptotisch optimale adaptive Verfahren, die auf Diskretisierungen mit Wavelet Basen beruhen. Diese Verfahren konvergieren asymptotisch mit der selben Rate wie die beste m-Term-Approximation. Damit wird eine obere Schranke für die Anzahl der benötigten Freiheitsgrade hergeleitet, die nötig sind, um das gesamte parabolische Problem bis auf eine vorgegebene Fehlertoleranz zu lösen. Dieses Ergebnis hängt von den Approximationseigenschaften der Wavelet Diskretisierung und der Regularität der Lösungen der elliptischen Stufengleichungen ab. Für die Stufengleichungen, die sich bei der Diskretisierung der Wäremeleitungsgleichung ergeben, wird ein neues Regularitätsresultat bewiesen und eingesetzt, um die Gesamtkomplexität des Verfahrens in diesem Beispiel abzuschätzen.

Damit ist es nun völlig gerechtfertigt, die räumlich adaptive Rothe Methode zur Lösung des zuvor analysierten Parameteridentifikationsproblems anzuwenden. Numerische Tests werden für ein vereinfachtes Parameterrekonstruktionsproblem durchgeführt. Hierzu wird gezeigt, dass eine biorthogonale Tensor-Wavelet Diskretisierung in dem asymptotisch optimalen adaptiven Verfahren zur Lösung der gegebenen elliptischen Teilprobleme eingesetzt werden kann. Tensor-Wavelet Basen sind spezielle Wavelet Basen, die sich durch die Tensorierung univariater Wavelet Basen ergeben. Ihr Hauptvorteil gegenüber klassischen Wavelet Basen ist, dass die beste m-Term-Approximation mit Tensor-Wavelets mit einer dimensionsunabhängigen Rate konvergiert.

Die klassische Tensor-Wavelet Konstruktion ist auf Gebiete mit einer einfachen Pro-

duktgeometrie beschränkt, was die Anwendbarkeit dieser Wavelets stark einschränkt. Aus diesem Grund wird in dieser Arbeit eine Konstruktion für eine verallgemeinerte Tensor-Wavelet Basis für Sobolev Räume über einem Gebiet Ω mit relativ allgemeiner Geometrie entwickelt. Die Konstruktion basiert auf der Anwendung von Fortsetzungsoperatoren auf passende Basen auf Teilgebieten Ω_i, die eine nichtüberlappende Gebietszerlegung von Ω bilden. Zunächst werden Bedingungen identifiziert an die Randbedingungen, die von den lokalen Basen erfüllt werden müssen, damit geeignete Fortsetzungsoperatoren als beschränkte Abbildungen überhaupt existieren. Den Ausgangspunkt für eine rekursive Beschreibung der globalen Basiskonstruktion bildet eine Zerlegung von Ω in nichtüberlappende Hyperwürfel, die an einem Cartesischen Gitter ausgerichtet sind und mit Tensor-Wavelet Basen versehen werden. Die Basen auf den Teilgebieten werden sukzessive miteinander verschmolzen, indem rekursiv univariate Fortsetzungsoperatoren auf sie angewendet werden. Die beste m-Term-Approximation bezüglich der neuen Basis reproduziert die von klassischen Tensor-Wavelets bekannte dimensionsunabhängige Konvergenzrate. Es wird gezeigt, dass die hierfür notwendige Regularität für die Lösung elliptischer Gleichungen der Ordnung $2m = 2$ über einem polygonalen oder polyhedralen Gebiet für genügend glatte rechte Seiten gewährleistet ist. Numerische Tests mit asymptotisch optimalen adaptiven Verfahren werden durchgeführt, die belegen das die theoretisch optimale Rate in der Praxis auch realisiert wird.

Abstract

In general, inverse problems can be described as the task of inferring conclusions about the cause $u^\dagger$ from given observations $y^\dagger$ of its effect. This can be described mathematically as the inversion of an operator equation

$$\mathcal{K}(u^\dagger) = y^\dagger.$$

Particularly interesting is the case that this Problem is ill-posed or ill-conditioned. To arrive at a meaningful solution in this setting, regularization schemes need to be applied. One of the most important regularization methods is the so called Tikhonov regularization. In this approach a penalty term $\mathscr{F}_\alpha : X \to \mathbb{R}$ is introduced to enforce additional properties on the solution.

As an approximation to the unknown truth $u^\dagger$ it is possible to consider the minimizer

$$u_\alpha^\delta = \arg\min_{u \in X} \frac{1}{2}\|\mathcal{K}(u) - y^\delta\|_Y^2 + \alpha\mathscr{F}(u), \tag{0.3}$$

where y^δ is subject to $\|y^\delta - y^\dagger\|_Y \leq \delta$. The convergence of the sequence u_α^δ to $u^\dagger$ for appropriate parameter choice rules $a = a(\delta)$ and $\delta \to 0$ was studied intensely in the past. Indeed, the analysis of such regularization properties can almost be regarded as complete for many settings, including the important case of general nonlinear operators $\mathcal{K}$ over a Hilbert space X that are penalized with weighted ℓ_p-norm penalty terms on the coefficients with respect to a Riesz Basis.

In contrast the development of efficient minimization schemes for the computation of u_α^δ is a field of ongoing research. A popular approach is to consider penalty terms that enforce a simple structure on the minimizers. For Hilbert space settings, one such choice are the aforementioned weighted ℓ_p penalty terms. For $1 \leq p \leq 2$ they enforce that the coefficients of the minimizers u_α^δ with respect to the Riesz basis belong to $\ell_{2(p-1)}$. In particular, $p = 1$ implies the finiteness of the coefficient vector of the minimizer. Most computation schemes for u_α^δ are based on some generalized gradient descent approach. For problems with weighted ℓ_p-penalty terms this typically leads to iterations of the form

$$u^{(n)} = \mathbb{S}_\alpha(u^{(n)} - (\mathcal{K}'(u^{(n)}))^*(\mathcal{K}(u^{(n)}) - y^\delta)), \tag{0.4}$$

where $\mathbb{S}_\alpha$ is the well known soft shrinkage operator applied to each coefficient and $(\cdot)^*$ denotes the adjoint operator.

The schemes that are available for the numerical treatment of inverse problems related to general nonlinear operators $\mathcal{K}$ share some practical downsides. Convergence

of the sequence $(u^{(n)})_{n\in\mathbb{N}}$ is usually only guaranteed for subsequences and only to stationary points of the right-hand side of (0.3). In general, these points do not have any regularization properties. Also, the scheme (0.4) in its basic form is known to converge very poorly in practice. This is critical as each iteration step includes the application of $\mathcal{K}$ and the adjoint of its derivative. This in itself may already be numerically demanding.

The approximate solution of general nonlinear inverse problems poses a highly challenging numerical task. To approach this issue two strategies are investigated in this thesis:

$T1$. Development of a strategy for the fast minimization of (0.3) with provable convergence properties.

$T2$. Development, analysis and generalization of efficient numerical methods for the operators that are applied in each iteration step.

As the first result of this thesis an acceleration strategy for the iteration (0.4) which is based on a decreasing strategy $\alpha^{(n)} \searrow \alpha$ for the thresholding parameter is proposed. If the nonlinear Operator $\mathcal{K}$ and its derivative $\mathcal{K}'$ are Lipschitz continuous, and further a certain local contraction assumption holds, then the resulting algorithm is linearly convergent to a global minimizer and the iteration is monotone with respect to the Tikhonov functional. The assumptions are satisfied for important classes of operator equations. For operators $\mathcal{K}$ that consist of the sum of a linear part and a nonlinear perturbation a certain preconditioning strategy is introduced to promote the convergence assumptions.

A very important class of inverse problems are parameter identification problems for partial differential equations. Here the goal is to reconstruct parameters of the differential equations from given realizations. Both as a prototype for this class of inverse problems and further to motivate a certain approach to $T2$ a parameter identification problem for a parabolic differential equation is investigated. The forward operator $\mathcal{K}$ is analyzed in order to justify that the scheme (0.2) is applied to compute the minimizer of the Tikhonov functional in (0.1). The set of admissible parameters in this analysis is subject to L_p conditions arising from the solution theory of partial differential equations as well as L_∞ bounds arising from the underlying practical motivation. It is possible to describe the action of the adjoint of $\mathcal{K}$ as the solution of a parabolic differential equation.

One approach for the numerical treatment of parabolic differential equations is the so called horizontal method of lines, also known as Rothe's method. The parabolic problem is interpreted as an abstract Cauchy problem. It is discretized in time by means of an implicit scheme. This is combined with a discretization of the resulting system of spatial problems. The efficiency of the Rothe method depends on the numerical schemes that are applied for the temporal and spatial discretization. A promising approach is the application of adaptive discretization schemes. Such schemes are nonlinear approximation methods that utilize a posteriori error estimation to adapt the

discretization to the current approximation until a prescribed error tolerance is satisfied. They realize highly nonuniform discretizations. Therefore, such schemes tend to require much less degrees of freedom to than classical discretization schemes.

Adaptive methods may be employed in Rothe's method for both the temporal as well as the spatial discretization. As a first step in this thesis temporal uniform and spatially adaptive Rothe methods are considered. To ensure the convergence of such a scheme a rigorous convergence proof is given for the general setting that the temporal discretization leads to a system of S stage equations in space that are solved up to given tolerances. It is investigated how the tolerances in each time step have to be tuned in order to preserve the asymptotic temporal convergence order of the time stepping. In particular, the case of linearly-implicit time integrators and asymptotically optimal adaptive wavelet discretizations in space is discussed. Such spatial discretization schemes asymptotically converge with the same rate as the best-m-term approximation. Using concepts from regularity theory for partial differential equations and from nonlinear approximation theory, we determine an upper bound for the degrees of freedom for the overall scheme that are needed to adaptively approximate the solution up to a prescribed tolerance. As an important case study, the complexity of the approximate solution of the heat equation is investigated for a temporal discretization by means of a linearly implicit Euler scheme. To this end a regularity result for the resulting stage equations that are of Helmholtz type is proven.

The spatially adaptive Rothe method is applied to the previously investigated parabolic parameter identification problem. Numerical experiments are performed for a simplified parameter reconstruction problem. The the spatially adaptive scheme that is applied is based on a biorthogonal tensor wavelet basis. It is proven that such wavelet may be applied in an asymptotically optimal numerical scheme. Standard tensor wavelet bases are wavelet bases that consist of tensors of univariate wavelet bases. Their main advantage compared to classical constructions is that the best-m-term approximation with tensor wavelets exhibits dimension independent convergence rates.

The classical tensor wavelet construction is limit to domains with simple product geometry, seriously limiting the applicability of such wavelets. A construction for a generalized tensor wavelet basis for a range of Sobolev spaces over a domain Ω with a fairly general geometry is proposed. The construction is based on the application of extension operators on appropriate local bases on subdomains Ω_i that form a non-overlapping domain decomposition. Conditions on the boundary conditions on each subdomain are derived that need to be satisfied to guarantee that extension operators exist as bounded mappings. As subdomains, we take hypercubes, or smooth parametric images of those, and equip them with tensor product wavelet bases. The hypercubes are assumed to be aligned to a Cartesian grid. The bases on the subdomains are recursively merged by applying univariate extension operators. The approximation rates from the resulting piecewise tensor product basis are proven to be independent of the spatial dimension of Ω. For two- and three-dimensional polytopes it is shown that the solution of Poisson type problems satisfies the required regular-

ity condition. The dimension independent rates will be realized numerically in linear complexity by the application of the asymptotically optimal adaptive scheme.

Contents

1 Introduction and overview

The mathematical research area of *inverse problems* is concerned with the extraction of information about a model system from given data. An incomplete list of applications includes for example the fields of computer vision, geophysics, medical imaging, nondestructive testing.

The classical setting of an inverse problem involves a bounded linear operator

$$\mathcal{K} : X \to Y$$

acting between two Hilbert spaces X and Y with respective norms $\|\cdot\|_X$ and $\|\cdot\|_Y$. The task is to compute an approximation to the unknown truth $u^\dagger$ from noisy data y^δ such that the fidelity term satisfies

$$\|\mathcal{K}u^\dagger - y^\delta\|_Y \leq \delta.$$

This problem is usually ill-posed and the set of admissible solutions $\{u \in X : \|\mathcal{K}u - y^\delta\|_Y \leq \delta\}$ is unbounded. Regularization schemes need to be employed in order to arrive at a meaningful solution. A regularization scheme is a family of operators $\{T_\alpha\}_{\alpha \geq 0}$ such that the related sequence of minimizers $u_\alpha^\delta = T_\alpha y^\delta$ satisfies $u_{\alpha(\delta)}^\delta \to u^\dagger$ for $\delta \to 0$ and an appropriate parameter choice rule $\alpha(\delta)$. Most regularization schemes can either be characterized as iterative schemes, are based on a discretization approach or follow Tikhonov's approach.

For any given regularization scheme two important issues need to be addressed:

1. What are the regularization properties of the scheme? How should α be chosen and what can be said about convergence and convergence rates of $u_{\alpha(\delta)}^\delta$?

2. How can we efficiently compute approximations to the minimizers u_α^δ?

For an introduction to the regularization theory for classical linear inverse problems in Hilbert spaces we refer to [54, 90, 81, 106]. In recent years, many regularization results were generalized to the case of nonlinear operators $\mathcal{K} : D(\mathcal{K}) \subset X \to Y$. We will focus on the most widely used approach, that is Tikhonov regularization. One reason is that convergence and convergence rates of the minimizers $\{u_{\alpha(\delta)}^\delta\}_{\delta \geq 0}$ can be established under relatively mild assumptions. Moreover, there exist robust schemes to approximate those minimizers.

In Tikhonov regularization one is interested in the minimizers u_α^δ of

$$\Gamma(u) = \frac{1}{2}\|\mathcal{K}(u) - y^\delta\|_Y^2 + \alpha \mathscr{F}(u), \tag{1.0.1}$$

where $\mathscr{F} : X \to \mathbb{R}^+ \cup \{0\}$ is a proper and convex penalty term. In classical Tikhonov regularization the penalty is chosen as $\mathscr{F} = \|\cdot\|_X^2$. However, in practice this leads to overly smooth results. Therefore, in recent years, penalty terms were intensely studied that promote additional features of the minimizers, such as *sparsity*. We focus on the case that sparsity of a function $u \in X$ is understood as sparsity of its coefficient vector with respect to a given discretization of X. We call a coefficient vector sparse, if it is finitely supported or has rapidly decaying entries. This can be exploited by using weighted ℓ_p-penalty terms. In case of a discretization of X by an orthonormal basis $\{\eta_\mu\}_{\mu\in\mathcal{J}}$ we consider terms of the form

$$\mathscr{F}(u) = \sum_{\mu\in\mathcal{J}} w_\mu |\langle u, \eta_\mu\rangle_X|^p, \tag{1.0.2}$$

where $w_\mu \geq w > 0, \mu \in \mathcal{J}$ and $1 \leq p \leq 2$. For such penalty terms it has been shown in [87] that $u_\alpha^\delta \in \ell_{2(p-1)}(\mathcal{J})$ holds. In particular, this implies that for $p = 1$ the minimizers are finitely supported. This justifies the diction *sparsity promoting penalty term* for penalty terms as in (1.0.2).

So far we did not assume that the unknown truth $u^\dagger$ is sparse. Therefore, in general the minimizers u_α^δ only provide a sparse approximation of a non sparse function. This approach might nonetheless be numerically appealing as the computational effort for the computation of the minimizers is closely related to their support sizes. However, under the additional assumption that $u^\dagger$ is sparse, it is possible to derive improved convergence results. For instance, it has been shown in [63] that the choice $\alpha \sim \delta$ leads to the convergence result

$$\|u_\alpha^\delta - u^\dagger\|_X \sim \delta,$$

provided that $\mathcal{K}$ is Gâteux differentiable in $u^\dagger$, the restriction of its derivative to finite-dimensional subspaces of X is injective and the ℓ_p-penalty term is related to $\mathcal{K}$ by means of an apropriate source condition.

Sparsity promoting penalty terms were first introduced to Tikhonov regularization in the fundamental paper [46]. It is concerned with linear problems in a Hilbert space setting that are discretized by means of an orthonormal basis. Regularization properties of the Tikhonov approach are derived and convergence of an iterative scheme to compute the minimizers based on soft-shrinkage is proven. Since then, the linear theory has been successfully generalized in several ways. It is possible to consider discretizations by means of Riesz bases or frames, more general penalty terms and operators acting on Banach spaces, cf. the survey [78]. The regularization properties of Tikhonov regularization for nonlinear inverse problems in a Hilbert space setting with ℓ_p-sparsity constraints have been intensely studied in the last years and their analysis can almost be regarded as complete [76, 13, 15, 102]. For the treatment of nonlinear inverse problems in general abstract Banach space settings we refer to [108, 109].

For practical applications the efficient numerical approximation of the minimizers u_α^δ is essential. The most well-known iterative scheme for inverse problems with ℓ_p-penalty terms is the *iterative soft thresholding algorithm*, that has been proposed for linear inverse problems in [46]. By now, several algorithms for the treatment of nonlinear settings have been proposed. However, most of them can be interpreted as soft shrinkage schemes with differently chosen parameters, confer [76], and are special cases of the *generalized conditional gradient method*, which was considered in [13, 15].

The generalized conditional gradient scheme deals with the approximation of the minimizers of the functional

$$\Gamma(u) = \mathscr{E}(u) + \mathscr{F}(u). \tag{1.0.3}$$

Here it is assumed that $\mathscr{E} : X \to \mathbb{R}$ is continuous, has Lipschitz continuous Fréchet derivative $\mathscr{E}' : X \to \mathcal{L}(X, \mathbb{R})$, $\mathscr{E} + \mathscr{F}$ is proper, non-negative and coercive, and $\mathscr{F} : X \to \mathbb{R}$ is convex and lower weakly semi-continuous. However, it is neither assumed that $\mathscr{E}$ is convex nor that $\mathscr{F}$ is differentiable.

The key ingredient of the scheme is a first order necessary condition for the minimizers of (1.0.3), that is given by

$$\mathscr{E}'(u)(u) + \mathscr{F}(u) = \min_{v \in X} \mathscr{E}'(u)(v) + \mathscr{F}(v), \tag{1.0.4}$$

which has been proven for instance in [86]. An algorithm for computing the minimizers of (1.0.3) is then derived by minimizing the right-hand side of (1.0.4) in each iteration step:

1: Choose $u^{(0)} \in X$, such that $\mathscr{F}(u^{(0)}) < \infty$;

2: Determine $v^{(n)} \in X$ by

$$v^{(n)} = \arg\min_{v \in X} \mathscr{E}'(u^{(n)})(v) + \mathscr{F}(v); \tag{1.0.5}$$

3: Determine step size $s^{(n)} \in [0, 1]$ via

$$s^{(n)} = \arg\min_{s \in [0,1]} \mathscr{E}(u^{(n)} + s(v^{(n)} - u^{(n)})) + \mathscr{F}(u^{(n)} + s(v^{(n)} - u^{(n)})); \tag{1.0.6}$$

4: Put $u^{(n+1)} = u^{(n)} + s^{(n)}(v^{(n)} - u^{(n)})$. Return to step 2.

The connection to iterative soft thresholding schemes is given by the choice

$$\begin{aligned} \mathscr{E}(u) &= \frac{1}{2}\|\mathcal{K}(u) - y^\delta\|_X^2 - \frac{\lambda}{2}\|u\|_X^2, \\ \mathscr{F}(u) &= \frac{\lambda}{2}\|u\|_X^2 + \alpha \sum_{\mu \in \mathcal{J}} w_\mu |\langle u, \eta_\mu \rangle_X|^p, \end{aligned}$$

where $\lambda > 0$ will turn out to be an additional step size parameter for the descent direction. In order to derive a reformulation of (1.0.5) we state that the derivative $\mathscr{E}'(u) \in \mathcal{L}(X, Y)$ reads as

$$\mathscr{E}'(u)(\cdot) = \langle (\mathcal{K}'(u))^*(\mathcal{K}(u) - y^\delta) - \lambda u, \cdot \rangle_X,$$

where $(\mathcal{K}'(u))^* \in \mathcal{L}(Y,X)$ denotes the adjoint of $\mathcal{K}'(u)$. By inserting this into (1.0.5) and adding a quadratic term that does not influence the minimizer we conclude

$$\begin{aligned} v^{(n)} &= \arg\min_{v\in X} \langle (\mathcal{K}'(u^{(n)}))^*(\mathcal{K}(u^{(n)}) - y^\delta) - \lambda u^{(n)}, v\rangle_X + \frac{\lambda}{2}\|v\|_X^2 + \alpha \sum_{\mu\in\mathcal{J}} w_\mu |\langle v, \eta_\mu\rangle_X|^p \\ &= \arg\min_{v\in X} \|\frac{1}{\lambda}(\mathcal{K}'(u^{(n)}))^*(\mathcal{K}(u^{(n)}) - y^\delta) - u^{(n)} + v\|_X^2 + \frac{2\alpha}{\lambda} \sum_{\mu\in\mathcal{J}} w_\mu |\langle v, \eta_\mu\rangle_X|^p \\ &= \arg\min_{v\in X} \sum_{\mu\in\mathcal{J}} \langle \frac{1}{\lambda}(\mathcal{K}'(u^{(n)}))^*(\mathcal{K}(u^{(n)}) - y^\delta) - u^{(n)} + v, \eta_\mu\rangle_X + \frac{2\alpha}{\lambda} w_\mu |\langle v, \eta_\mu\rangle_X|^p. \end{aligned}$$

The minimizer of such a functional combining an ℓ_2-norm fidelity term and a weighted ℓ_p-norm penalty can be directly computed using a soft thresholding operation, see [18, 46]. It holds that

$$v^{(n)} = \mathbb{S}_{\frac{\alpha\mathbf{w}}{\lambda},p}(u^{(n)} - \frac{1}{\lambda}(\mathcal{K}'(u^{(n)}))^*(\mathcal{K}(u^{(n)}) - y^\delta)), \tag{1.0.7}$$

where $\mathbb{S}_{\frac{\alpha\mathbf{w}}{\lambda},p}$ is a shrinkage operator defined by

$$\mathbb{S}_{\frac{\alpha\mathbf{w}}{\lambda},p}(u) = \sum_{\mu\in\mathcal{J}} S_{\frac{\alpha w_\mu}{\lambda},p}(\langle u, \eta_\mu\rangle_X)\eta_\mu,$$

and the shrinkage maps $S_{\frac{\alpha w_\mu}{\lambda},p}$ are given by

$$S_{\alpha,p}(x) = \begin{cases} \operatorname{sgn}(x)[|x| - \alpha]_+, & p = 1, \\ G_{\alpha,p}^{-1}(x), & p > 1, \end{cases}$$

where $G_{\alpha,p}(x) = x + \alpha p \operatorname{sgn}(x)|x|^{p-1}$. We conclude that one iteration step of the basic iterated soft shrinkage algorithm is given by

$$u^{(n+1)} = u^{(n)} + s^{(n)}(\mathbb{S}_{\frac{\alpha\mathbf{w}}{\lambda},p}(u^{(n)} - \frac{1}{\lambda}(\mathcal{K}'(u^{(n)}))^*(\mathcal{K}(u^{(n)}) - y^\delta)) - u^{(n)}), \tag{1.0.8}$$

where $s^{(n)}$ is given by (1.0.6) and $\mathbf{w} = (w_\mu)_{\mu\in\mathcal{J}}$.

The line search (1.0.6) in step 3 of the generalized conditional gradient algorithm guarantees that $\Gamma(u^{(n)})$ decreases in each iteration step if $u^{(n)}$ is not already a stationary point. However, in many cases the optimal value for the step size $s^{(n)}$ presented in (1.0.6) is purely theoretical and a priori or heuristic choice rules need to be applied. Convergence of the scheme for such suboptimal choices of $s^{(n)}$ can nonetheless be ensured in many settings. Indeed, it has been shown in [13, Lemma 2.4] that if λ is chosen big enough, it is possible to choose $s^{(n)} = 1$ and to omit the line search completely.

A typical convergence result for the iterated soft thresholding algorithm holds under the assumption that $\mathcal{K}$ is continuous with Lipschitz continuous derivative, and

furthermore that for any sequence $(v^{(n)})_{n\in\mathbb{N}} \subset X$ its convergence $v^{(n)} \to v$ also implies the convergence $(\mathcal{K}'(v^{(n)}))^*(\mathcal{K}(v^{(n)}) - y^\delta) \to (\mathcal{K}'(v))^*(\mathcal{K}(v) - y^\delta)$. Then, it has been shown in [76, Theorem 4.3] that the sequence $(u^{(n)})_{n\in\mathbb{N}}$ has a subsequence that converges to a stationary point of Γ. This type of convergence result has to be expected since we are dealing with a gradient based iterative scheme and the nonlinearity of $\mathcal{K}$ implies that the target functional Γ may have several local minima. However, only the global minimizers of Γ have reliable regularizing properties.

Several different approaches to solve the central minimization problem related to (1.0.1) lead to iterated soft thresholding algorithms. Such algorithms have in common that they are relatively easy to implement and numerically robust. Furthermore, the related regularization theory is well established and convergence can be shown under relatively mild assumptions. However, iterated soft shrinkage algorithms share some serious practical downsides. Each shrinkage step (1.0.7) typically involves the application of the nonlinear operators $\mathcal{K}$ and $(\mathcal{K}'(\cdot))^*$. This in itself already poses a highly challenging numerical task. Further, the method in its basic form is known for its poor convergence speed and speed up strategies are pivotal for its numerical applicability.

Therefore, we have identified two tasks that need to be addressed in order to improve the convergence speed of the iterated soft thresholding algorithm:

$T1$. top level speed up strategies,

$T2$. fast solvers for the forward problem, that is, the for the application of $\mathcal{K}$ and the adjoint of $\mathcal{K}'$.

By now there exist several approaches for speed up strategies for linear inverse problems with sparsity constraints, see [89] for a comparison. First steps have been made to generalize some approaches to nonlinear settings, however, this is still a field of ongoing research.

One interesting strategy that has been recently considered for nonlinear inverse problems is based on a quadratic approximation of Γ in $u^{(n-1)}$. The resulting approximation $\Gamma(\cdot, u^{(n-1)})$ is then used to determine the descent direction in each iteration step. This is combined with an parameter choice rule for $\lambda^{(n)}$. In each iteration step the step size $\lambda^{(n)}$ is chosen according to a heuristic rule inside a prescribed interval $\underline{\lambda}, \overline{\lambda}$, where $\lambda^{(n)} = \overline{\lambda}$ would guarantee a decrease of Γ. If the current choice of the step size does not reduce Γ, it is increased a maximal finite number of times. This strategy was first applied to nonlinear finite-dimensional settings in [95]. In [88] it was applied to a general nonlinear Hilbert space setting and some modifications were discussed. The convergence analysis relies mainly on a Lipschitz continuity assumption on the derivative of $\mathcal{K}$. However, convergence is only proven in the sense that $|\Gamma(u^{(n)}) - \Gamma(\alpha^\delta)|$ behaves as $\mathcal{O}(n^{-1})$ for the basic algorithm. For convex fidelity terms this rate improves to $\mathcal{O}(n^{-2})$.

First steps are undertaken to apply semi-smooth Newton methods to sparsity constraint inverse problems [73, 64, 72]. The results seem promising, however, the con-

vergence analysis of these methods is still ongoing.

For $p = 1$ the application of the shrinkage operator $\mathbb{S}_{\alpha,p}$ ensures the finite support of the iterates $u^{(n)}$ in (1.0.8). Another way to enforce the finite support of the iterates is to replace the soft shrinkage by compressed hard thresholding. That is, discretization coefficients below a threshold are neglected and the remaining ones are not changed. The resulting algorithm has been studied for instance in [103], and performed quite well in computations. However, the iteration is not directly related to the minimization of a functional and therefore it is difficult to proof regularization properties.

Our main interest lies in the global minimizer of Γ. Convergence results for the strategies scetched above yield at best convergence of subsequences towards critical points of Γ. However, nothing is known about the regularizing properties of the stationary points of Γ. In Chapter 2, we propose a to consider a parameter choice strategy for the thresholding parameter α. We investigate conditions on $\mathcal{K}$ such that a decreasing strategy for the α leads to *linear convergence* of the iterates towards the *global minimizer* of Γ, thus ensuring regularization properties. Our work is completely covered by the the framework of Tikhonov regularization that we presented so far. However, to ease notation we will consider the discretized inverse problem for $K = \mathcal{K} \circ \mathcal{S} : \ell_2(\mathcal{J}) \to Y$, where

$$\mathcal{S} : \ell_2(\mathcal{J}) \to X, \quad \mathbf{v} \mapsto \sum_{\mu \in \mathcal{J}} v_\mu \eta_\mu$$

is the synthesis operator related to the discretization $\{\eta_\mu\}_{\mu \in \mathcal{J}}$ of X.

We propose to minimize the functional

$$\Gamma_{\boldsymbol{\alpha}^{(n)}}(\mathbf{u}) = \|K(\mathbf{u}) - y^\delta\|_Y^2 + \sum_{\mu \in \mathcal{J}} \alpha_\mu^{(n)} |u_\mu|,$$

in each iteration step, where the parameters $\boldsymbol{\alpha}^{(n)} \in \mathbb{R}_+^{\mathcal{J}}$ are chosen as entrywise decreasing sequences with $\lim_{n\to\infty} \boldsymbol{\alpha}^{(n)} = \boldsymbol{\alpha}$ that are bounded away from 0, that is, $\alpha_\mu^{(n)}, \alpha_\mu \geq \alpha \in \mathbb{R}_+, \mu \in \mathcal{J}$. This is achieved by applying a soft shrinkage operator

$$\mathbf{u}^{(n)} = \mathbb{S}_{\frac{1}{\lambda}\boldsymbol{\alpha}^{(n)}}\Big(\mathbf{u}^{(n)} - \frac{1}{\lambda}\big(K'(\mathbf{u}^{(n)})\big)^*\big(K(\mathbf{u}^{(n)}) - y^\delta\big)\Big). \tag{1.0.9}$$

We assume that λ is chosen big enough such that we may choose $s^{(n)} = 1$.

The approach is based on the investigations in [36], where a decreasing thresholding approach has been applied to a linear inverse problem with a ℓ_1-penalty term. Linear convergence of the scheme has been proven under the condition that K satisfies a restricted isometry property, that is, for some fixed $k \in \mathbb{N}$ and all $\Lambda \subset \mathcal{J}$ with $\#\Lambda \leq k$ it holds that

$$\|(\mathrm{Id} - K^*K)_{|\Lambda\times\Lambda}\|_{\mathcal{L}(\ell_2(\Lambda),\ell_2(\Lambda))} \leq \gamma_k < 1, \tag{1.0.10}$$

where $(\cdot)_{|\Lambda}$ denotes the restriction to the index set Λ. In this setting the limit of the thresholded iteration is indeed the global minimizer of (1.0.1), that is, it holds

that $\lim_{n\to\infty} u^{(n)} = u^{\delta}_{\boldsymbol{\alpha}^{(n)}}$. Furthermore, it was investigated in [36] how the restricted isometry property can be obtained by means of a multilevel preconditioning strategy.

The generalization of the decreasing thresholding approach to the full nonlinear setting relies on two fundamental assumptions. We need that the operator K and its derivative are Lipschitz continuous on closed and bounded sets. Moreover, we have to assume that the operator

$$T : \ell_2(\mathcal{J}) \to \ell_2(\mathcal{J}), \quad \mathbf{v} \mapsto T(\mathbf{v}) = \mathbf{v} - \frac{1}{\lambda}\big(K'(\mathbf{v})\big)^*\big(K(\mathbf{v}) - y^{\delta}\big), \tag{1.0.11}$$

restricted to finite index sets $\lambda \subset \mathcal{J}$, is a contraction on a sufficiently small ball around a critical point u^* of the functional (1.0.1). This condition on T is used as a nonlinear analog of the restricted isometry property assumption (1.0.10).

Then the main convergence result, specified in Theorem (2.2.4) holds: The iteration is linearly convergent, that is,

$$\|\mathbf{u}^* - \mathbf{u}^{(n)}\|_{\ell_2(\mathcal{J})} \leq \gamma^n \|\mathbf{u}^*\|_{\ell_2(\mathcal{J})}, \quad \text{for some } \gamma \leq 1,$$

whenever the $\boldsymbol{\alpha}^{(n)} \geq \boldsymbol{\alpha}$ are chosen according to

$$\max_{\mu\in\mathcal{J}} |\alpha_\mu^{(n)} - \alpha_\mu| \leq C\varepsilon^{(n)},$$

with C as in (2.2.35). Moreover, the iteration is monotone in the sense that

$$\Gamma_{\boldsymbol{\alpha}^{(n+1)}}(\mathbf{u}^{(n+1)}) < \Gamma_{\boldsymbol{\alpha}^{(n)}}(\mathbf{u}^{(n)}),$$

provided that $\mathbf{u}^{(n)}$ is not a critical point of $\Gamma_{\boldsymbol{\alpha}^{(n)}}$.

The local contraction condition on T may be hard to verify. In the Sections 2.2.2 and 2.2.3, we discuss in detail two classes of operators where it is satisfied. The first class consists of operators with bounded second derivatives and with first derivatives that satisfy the contraction property. The second class is given by nonlinear perturbations of linear operators that satisfy the restricted isometry property (1.0.10). It was shown in [36] that for large classes of linear operators where the restricted isometry property is not satisfied, it can actually be established by preconditioning. We investigate the application of such preconditioning strategies for linear operators with nonlinear perturbations.

The analysis is performed in an infinite-dimensional setting. In general the preconditioning strategy $D : \ell_2(\mathcal{J}) \to \mathrm{Ran}(D)$ is allowed to be unbounded in the topology of $\ell_2(\mathcal{J})$. Further the operators K and K' cannot be evaluated exactly in an infinite dimensional setting. Fortunately, we are able to address both issues by employing implementable numerical approximations schemes for the application of K and K'. We prove that the resulting inexact version of (1.0.9) again converges with linear rate. Moreover, the support of all iterates is contained in an a priori unknown finite index set $\Lambda_0 \subset \mathcal{J}$, which is constructed on the fly by the method. The problematic topology introduced by the preconditioning is circumvented by proving the equivalence of the

inexact scheme to a finite-dimensional scheme on $\mathbb{R}^{\lambda_0}$. Our analysis relies on adaptive numerical methods for the approximation of K and K', which we will present in more detail below.

The speed up strategy based on a decreasing thresholding parameter $\boldsymbol{\alpha}^{(n)}$ that we derive in Chapter 2 may to some degree be compatible with the strategy based on quadratic approximation proposed in [88]. The local contractivity assumption on the operator T (1.0.11) for some fixed λ implies the same property for a certain range $[\underline{\lambda}, \overline{\lambda}]$ of step sizes. The prospect is that such a range of admissible step sizes may be exploited as in [88] and should lead to a quantitative improvement of the convergence results therein. For settings where the local contractivity assumption on T is not satisfied it may be fruitful to investigate the decreasing thresholding strategy in the setting of [88]. This will probably not provide a qualitative improvement to the convergence rate. We can only expect the convergence of a subsequence with the error being bounded by $|\Gamma(u^{(n)}) - \Gamma(u_\alpha^\delta)|$. None the less, the combination of both strategies holds some potential for a further speed up of the iterated soft thresholding algorithm.

The remainder of this exposition is dedicated to the central task $T2$ of deriving fast numerical solvers for the forward operator $\mathcal{K}$ and the adjoint of its derivative $\mathcal{K}'$. The outline is as follows. We begin by presenting in detail an inverse problem that is given by the parameter reconstruction problem related to a parabolic partial differential equation. This problem can be considered to be prototypical for this central class of inverse problems. A class of highly efficient numerical schemes are adaptive numerical methods. The underlying strategy is described with a focus on adaptive discretizations with wavelets. Then we discuss how adaptive methods may be applied for the treatment of parabolic operator equations. One of the most widely used numerical schemes for the treatment of parabolic equations is the horizontal method of lines, often called Rothe's method. Here the problem is discretized first in time and then in space. A rigorous error and complexity analysis is performed for the special case that adaptive methods are applied in the spatial discretization. A method of this type is applied to a simplified problem derived from the initial prototypical inverse problem. One of the key building blocks of adaptive methods is the underlying discretization. Wavelet bases of tensor type are especially interesting as they provide dimension independent approximation rates under relatively mild smoothness assumptions. We generalize the classical tensor wavelet construction that is limited to simple product domains to fairly general bounded domains. The dimension independent approximation rates are again realized. Thus, the overall applicability of the tensor wavelet approach is greatly increased.

In genetic research advances in experimental techniques have lead to the availability of large-scale gene expression data sets. However, experiments for deriving information on the interaction of genes remain very challenging. A far reaching approach for the field of functional genomics opens up if one considers sophisticated mathematical models for the gene expression. If biologically interesting quantities such as the interaction of genes are the parameters of the model they are accessible as the solution

of the related inverse problem. The results can then be used to generate interesting hypotheses to direct further experiments. One of the most important biological model organisms is the fruit fly drosophila melanogaster. The inference of biological information by reconstructing model parameters related to the interaction of genes of the fruit fly is a field of current research [61, 62, 83]. We will focus on a model for the early development of the animal, the so called *embryogenesis*. Sophisticated reaction-diffusion models for the gene expression that take into account the interaction of different genes have been developed in [104, 49, 129, 22]. We will focus on the fundamental approach in [104]. Therein, the evolution of gene expression levels is modeled as a deterministic system of parabolic differential equations. The model contains several a priori unknown parameters that depend on time and space. Transport of gene products within the admissible domain, that is, the embryo, is modeled by a diffusion term D. The limited life span of gene products as well as consumption is addressed by a linear decay term λ. Finally, the synthesis is modeled as the product of a maximal synthesis rate R and a response function Φ. The signal response is modeled as a nonlinear function $\Phi : \mathbb{R} \to [0, 1]$, which takes as its argument the feedback Wg, where g is the vector of gene concentrations and W is an interaction matrix. The matrix W is the biologically most interesting parameter. Positive entries describe an amplifying influence of one gene on another, whilst negative entries correspond to an inhibiting effect.

The complete model is then given as follows. Let $\Omega \subset \mathbb{R}^n, n = 2, 3$ denote some bounded Lipschitz domain. The gene concentrations are modeled as real valued functions g_i, $i = 1, \dots, N$ on the spatial-temporal domain $\Omega \times [0, T]$. Then the evolution of the gene expression levels is modeled by the reaction-diffusion equation

$$\begin{aligned} &\frac{\partial g_i}{\partial t} - \operatorname{div}(D_i \nabla g_i) + \lambda_i \cdot g_i = R_i \Phi_i((Wg)_i) \quad \text{in } \Omega \times (0, T], \\ &\frac{\partial g_i}{\partial \nu} = 0 \quad \text{on } \partial\Omega \times (0, T], \quad g(\cdot, 0) = g_0 \quad \text{on } \Omega, \end{aligned} \tag{1.0.12}$$

where $\Phi_i : \mathbb{R} \to \mathbb{R}$, $\Phi_i(x) = \frac{1}{2}((x^2 + 1)^{-\frac{1}{2}} x + 1$, and $i = 1, \dots, N$.

The data y^δ, that is available in practice, consists of measurements of gene concentrations at certain points in time. This is modeled by introducing a sampling operator $\mathcal{M}$ which maps the solution space $\mathcal{W}$ of the partial differential equation (1.0.12) to an observation space $\mathcal{O}$ of finite temporal granularity. The forward operator related to (1.0.12) is denoted as

$$\mathscr{D} : \mathcal{P} \to \mathcal{W}, \quad \pi = (D, \lambda, R, W) \mapsto \text{ the solution } g \text{ of (1.0.12)}.$$

Then the inverse problem reads as

$$\mathcal{M} \circ \mathscr{D}(\pi) = y^\delta. \tag{1.0.13}$$

That this is indeed an ill-posed problem becomes apparent by checking that choosing the parameters as $\pi = (D, \lambda, \frac{1}{2}, W_0)$ and $(D, \lambda, \Phi(W\mathscr{D}(\pi)), 0)$ yields the same right-hand side in (1.0.12).

The analysis of the inverse problem (1.0.13) is performed in Chapter 3. Tikhonov regularization is applied and the solution by means of iterated soft shrinkage is discussed. Finally an adaptive solution scheme based on the concepts outlined in this exposition is proposed and applied to a simplified version of the parameter reconstruction problem.

In our analysis of the mapping properties of $\mathscr{D}$ we aim at a fairly general setting and incorporate quite recently established results on maximal L_p-regularity of the solution of parabolic equations [67, 4]. This allows us to choose a weak topology for the admissible set of parameters. It will turn out that we are able to consider parameters

$$\begin{aligned} &D \in L_\infty([0,T]\times U, \mathbb{R}^N), \quad \lambda \in L_{p_\lambda}([0,T]\times U, \mathbb{R}^N), \\ &R \in L_{p_R}([0,T]\times U, \mathbb{R}^N), \quad W \in L_{p_W}([0,T]\times U, \mathbb{R}^{N\times N}), \end{aligned}$$

that are additionally subject to the L_∞ bounds

$$0 < C_{\mathcal{P},1} \leq D, \lambda \leq C_{\mathcal{P},2}, \quad 0 \leq R \leq C_{\mathcal{P},2}, \quad \|W\|_{L_\infty} \leq C_{\mathcal{P},2}.$$

The L_∞ restrictions on the parameters have to be expected for real world parameters.

We denote the parameter space for D with $\mathcal{P}_D = \{D \in L_\infty : 0 < C_{\mathcal{P},1} \leq D \leq C_{\mathcal{P},2}\}$ and $\mathcal{P}_\lambda$, $\mathcal{P}_R$, and $\mathcal{P}_W$ analogously. Then the global parameter space is defined as

$$\mathcal{P} = \mathcal{P}_D \times \mathcal{P}_\lambda \times \mathcal{P}_R \times \mathcal{P}_W$$

equipped with the product norm of the individual L_p spaces.

The generality of the parameter space comes at a price. If at least one of the indices p_λ, p_R, p_W differs from ∞ then $\mathcal{P}$ is not a metric space. It is a subset of a vector space, however, relative open sets are not open in the global L_p topology. Therefor we have to clarify the meaning of differentiation with respect to $\pi \in \mathcal{P}$. Our definition of differentiation on non-open sets of vector spaces follows [75] and is given in Definition 3.2.1. Careful analysis shows that the regularization theory for inverse problems carries over to this general setting.

The analysis of the forward operator $\mathscr{D}$ begins with an existence and uniqueness result for the solution to (1.0.12). Further, we proof that $\mathscr{D}$ is continuously differentiable with a Lipschitz continuous derivative. Finally, in Remark 3.3.7 the action of the adjoint of $(\mathscr{D}(\pi))'$ is explicitly expressed as the solution of a parabolic differential equation similar to (1.0.12). Therefore, in order to address the issue $T2$ it is sufficient to investigate efficient schemes for parabolic differential equations.

In order to derive an efficient numerical scheme for the iterated soft shrinkage algorithm (1.0.8) we will focus on *adaptive methods*. Adaptive discretization schemes are nonlinear approximation methods that utilize a posteriori error estimation to adapt the discretization to the unknown solution until a prescribed error tolerance is satisfied. They realize highly nonuniform discretizations compared to classical discretization schemes. Likewise, they tend to require less degrees of freedom than classical schemes, leading to highly increased numerical performance.

A central quality to compare adaptive methods is the concept of *optimality*. We say that a method is asymptotically optimal if it converges with the same rate as the *best-m-term* approximation. For a given discretization we call an approximation to a signal a best-m-term approximation to that signal if it uses at most m degress of freedom and realizes the best possible approximation among all such approximations.

Adaptive schemes based on finite element discretizations have a long and successful history in applications. Despite their good practical performance, their convergence properties are still a field of current research for many settings. For instance, convergence in the classical setting of second order elliptic equations was only recently shown in [94]. In particular, results on the optimality of such schemes were proven only recently [12, 120].

We focus on adaptive schemes based on a discretization by *wavelets*. The classical wavelet basis is a hierarchical Riesz basis for $L_2(\mathbb{R}^n)$ that consists of translated, dilated and scaled versions of a single (or multiple) mother-wavelet. Wavelet bases excel because of their analytic properties:

- compact support of the individual wavelets,
- characterization of classical function spaces by means of weighted norms for the sequence space of the wavelet expansion coefficients,
- cancelation properties, that is, the inner product of the wavelets with a smooth function decays exponentially with increasing wavelet scales.

The characterization of function spaces by wavelets makes it possible to relate the convergence order of best m-term wavelet approximation to the smoothness of the function v, that one wants to approximate. We refer to the survey article [48] for a detailed discussion. One central result is the following. Let us denote the error of best m-term wavelet approximation in the Sobolev space $H^\nu(\Omega)$ by means of classical isotropic wavelets by $\sigma^{\mathtt{iso}}_{m,\nu}(v)$. Further, let $\nu \geq 0$ and v be contained in the Besov space

$$B^s_q(L_q(\Omega)), \quad \text{where} \quad \frac{1}{q} = \frac{s-\nu}{n} + \frac{1}{2}, \quad s > \nu.$$

Then, if the wavelets under consideration are of sufficiently high order, the error of best m-term wavelet approximation in $H^\nu(\Omega)$ can be estimated as follows:

$$\sigma^{\mathtt{iso}}_{m,\nu}(v) \leq C_{\mathtt{iso}} \, \|v\|_{B^s_q(L_q(\Omega))} \, m^{-\frac{s-\nu}{n}}, \tag{1.0.14}$$

with a constant $C_{\mathtt{iso}} > 0$, which does not depend on v or m.

Such a deterioration of convergence properties with respect to the space dimension is commonly referred to as the *curse of dimensionality*. One way to approach this issue is to consider discretizations by means of *tensor wavelets*. The classical tensor wavelet basis is derived as the tensor product of univariate wavelet bases and is limited to product domains □. Consequently tensor wavelets differ from classical

isotropic wavelets by the fact that wavelets on different levels are tensorized with each other, leading to rectangular and highly anisotropic supports. The main advantage of the tensor wavelet approach is that the rate of convergence of the best m-term approximation by means of tensor wavelets is independent of the space dimension. Indeed, it can be bounded by

$$\sigma_{m,\nu}^{\mathtt{ten}}(v) \leq C_{\mathtt{ten}} \, \|v\|_{\mathcal{H}^s_{m,\theta}(\square)} \, m^{-(s-\nu)}, \quad (u \in \mathcal{H}^s_{\nu,\theta}(\square) \cap H^\nu(\square)), \tag{1.0.15}$$

where $C_{\mathtt{ten}} > 0$ does not depend on v or m, and $\mathcal{H}^s_{\nu,\theta}(\square)$ is a weighted Sobolev space. We refer to Chapter 5 for details.

The tensor wavelet approach provides a mean to break the curse of dimensionality at least for moderate space dimensions. With increasing space dimension any asymptotically optimal numerical scheme based on tensor wavelets will give rise to an error bound reading as (1.0.15) with some constant $C_{\mathtt{asym}}$. However, in general the quotient $C_{\mathtt{asym}}(C_{\mathtt{ten}})^{-1}$ will grow exponentially in the space dimension. One reason is that the condition number of the tensor wavelet basis depends in general exponentially on the space dimension. Only if L_2-orthogonal univariate wavelets are used in the construction, the condition number of the basis may be bounded independently of n. The adaptive solution of a second order elliptic equation with constant coefficients with such wavelets was studied in [51]. In that setting $C(n)C_{\mathtt{ten}}^{-1}$ can indeed be bounded uniformly in n. Still, numerical experiments suggest that the constant $C_{\mathtt{ten}}$ itself grows with possible exponential rate in the space dimension. We refer to the discussion in [121]. Therefore, it is mandatory for the treatment of high-dimensional problems to utilize additional structural information. An interesting approach in this direction, that is closely related to classical tensor wavelet approximation, is to consider functions that admit a low rank tensor approximation. For first results in this direction we refer to [7].

For practical applications it is often mandatory to consider discretizations for general bounded domains. Most wavelet constructions for domains with complicated geometries are related to a nonoverlapping domain decomposition into subdomains with simple geometries. A basis on the whole domain may then be derived by applying extension operators to local wavelet bases on the subdomains, , as proposed in [44]. Another approach that was considered in [43, 17], is to glue wavelets from neighboring domains together along the interfaces. The former approach is very technical and indeed, the extension operators needed in the construction do not exist as bounded mappings for some combinations of geometries and boundary conditions. The latter approach yields bases with limited global smoothness. For a detailed discussion we refer to [24]. So far, tensor wavelets have only been considered on simple product domains. Below, we outline the construction of generalized tensor wavelets that we propose in Chapter 5.

An interesting alternative to wavelet bases is to consider a discretization based on wavelet frames. For a domain Ω with a complicated geometry, an elegant way to construct a frame is to consider an overlapping decomposition into subdomains

$\{\Omega_i\}$ with simple geometries. An aggregated wavelet frame is simply devised as the union of the extensions by 0 of local wavelet bases on the subdomains. However, the reduced complexity of the construction comes at the price of introducing redundancy. This complicates the design and the following analysis of numerical schemes based on a frame discretization. Moreover, in applications, redundancy of the discretization system may lead to an increased computational cost as there may be unneeded active coefficients in any given numerical solution. None the less this approach is feasible. For the treatment of elliptic operator equations by means of adaptive wavelet methods based on an aggregated frame discretization we refer to the Ph.D. thesis [130]. Because of the additionally technical difficulties when using frames and the availability of our generalized tensor wavelet, we focus on discretizations by means of wavelet bases in the following.

In the last years adaptive wavelet methods have become a well established tool for the treatment of operator equations, including problems on bounded domains and closed manifolds. We refer to the monographs [40, 25, 125] for an introduction to the treatment of operator equations with wavelets. One of the main reasons for the popularity of adaptive wavelet methods is the availability of provable asymptotically optimal schemes with *linear complexity*, that is, the number of operations needed to compute an approximation scales linear with the degrees of freedom involved. Such schemes were first developed in the fundamental papers [26, 27] for linear operator equations. For a comparison of these methods as well as an overview of adaptive wavelet methods for linear operator equations we refer to the survey [121]. Generalizations to the nonlinear case exist by now, see [28, 9, 45, 79]. However, the theory is only fully established for the classical isotropic wavelet constructions. For first results concerning the case of anisotropic tensor wavelets we refer to [112].

Our main focus lies on the *efficient numerical treatment of parabolic initial value problems*, such as (1.0.12). To this end, there exist three distinct approaches. The parabolic problem can be considered as an asymmetric problem over the full spatial-temporal domain $\Omega \times (0, T)$. A tensor wavelet basis over such a product domain can be derived as the tensor product of a wavelet basis for the temporal domain and a tensor wavelet basis over the spatial domain. For linear parabolic equations an adaptive method based on spatial-temporal tensor wavelets was investigated in [111, 19]. It was shown that the method converges with optimal rate and with linear complexity.

Different approaches are followed by the vertical method of lines and the horizontal method of lines. The former starts with a semidiscretization in space. Then, the remaining task is to solve a system of coupled ordinary differential equations in time. We refer to [68, 77, 124] for detailed information. The latter, which is also known as *Rothe's method*, starts with a semidiscretization in time, followed by a discretization in space. It has been studied in for example in [84, 91, 101]. In Rothe's method, the parabolic equation is interpreted as an abstract Cauchy problem, that is, an ordinary differential equation in time over a suitable function space over the spatial domain. This problem is usually stiff, therefore the temporal discretization must be based on

an implicit scheme. In particular, linearly-implicit schemes are of interest, because their realization leads to a system of linear elliptic stage equations, that has to be solved in each time step.

Adaptive numerical methods may be implemented in Rothe's method in two possible ways. The temporal discretization scheme may utilize adaptive step size control based on an a posteriori error estimator. The local temporal error estimates may for instance be based on an embedded scheme of lower order or an extrapolation scheme. Clearly, it makes sense to adapt the temporal stepsizes to the temporal smoothness of the solution. However, results on the convergence properties of temporal adaptive schemes remain an field of ongoing research. For a discussion of temporal adaptive discretizations as well as numerical tests we refer to [101]. For linearly-implicit temporal discretizations, on which we focus, another way to incorporate adaptivity is to use adaptive numerical methods for the solution of the elliptic stage equations.

The combination of adaptive techniques for the temporal and spatial discretization seems natural. However, not much is known about this setting. In particular a comparison with the fully adaptive discretization on $\Omega \times (0,T)$, as proposed in [111, 19], would be fitting.

As a first step into this direction we investigate the combination of a uniform discretization in time with adaptive solution schemes for the stage equations. Concerning the convergence analysis of such inexact Rothe methods, the most far reaching results that we are aware of have been obtained in [84] for parabolic equations and finite element discretization in space. Therefore, we perform a thorough convergence analysis of Rothe's method, with uniform discretization in time and adaptive discretization in space in Chapter 4. Therein, we begin by considering an abstract temporal discretization that amounts to the solution of S stage equations in each time step and that is assumed to exhibit some overall temporal convergence rate. Then we investigate the inexact scheme, where the stage equations are only solved up to known tolerances. Under a Lipschitz continuity assumption for the operators, that describe the stage equations, we derive bounds for the tolerances of the solvers, such that the inexact scheme converges with the same rate as the exact scheme.

A large class of temporal discretization that fits our abstract assumptions are linearly implicit S-stage methods. Prominent examples of such schemes are methods of Rosenbrock type and the larger class of so called W-methods. For the convergence analysis of exact S-stage W-methods we refer to [91].

We apply our abstract analysis to the special case that adaptive wavelet methods are used to solve the stage equations. We focus on asymptotically optimal schemes with linear complexity. For such schemes approximation results similar to (1.0.14) and (1.0.15) hold, depending on the discretization. By the linear complexity of the method, our previous results on the tolerances needed in each stage equation translate into complexity estimates for the overall Rothe method under the assumption that all solutions to the stage equations belong to the appropriate smoothness spaces.

As an important case study, we investigate the discretization of the heat equation by means of a linearly implicit Euler scheme. In this example, there is only one stage

equation and the corresponding operator is of the form $(I-\tau\Delta_\Omega^D)^{-1}$, where Δ_Ω^D is the Dirichlet-Laplacian and τ is the temporal step size. For this setting we derive a new Besov regularity result that justifies the smoothness assumption on the solution of the stage equation and state the resulting complete complexity result for the solution of the heat equation.

On the basis of these theoretical considerations, we apply Rothe's method for the solution of the inverse problem (1.0.13). In Section 3.5 we consider a linearly implicit Rothe method with uniform discretization in time which utilizes an adaptive tensor wavelet solver for the spatial subproblems. The solver is asymptotically optimal with linear complexity and is based on biorthogonal univariate wavelets. For the closely related setting of a tensor wavelet basis based on L_2-orthogonal univariate wavelets it was shown in [51], that such schemes indeed exhibit optimal convergence rates in practice. We apply the new biorthogonal tensor wavelet method to a parameter reconstruction problem in a simplified setting and present numerical results in Section 3.6.

Classical tensor wavelet constructions are limited to product domains, significantly limiting their applicability. In Chapter 5 we consider a *generalized tensor wavelet basis* construction for fairly general domains. The new basis reproduces the dimension independent convergence rate of classical tensor wavelet bases. Our approach follows the ideas outlined in [23, 44]. The construction is based on a nonoverlapping domain decomposition of the global domain Ω into subdomains Ω_i. It is possible to consider parametric images of the subdomains. A global basis is constructed by applying extension operators to local bases on the subdomains. As a first step the abstract setting of a decomposition into two subdomains is considered. In this setting necessary conditions on the boundary conditions imposed on the local bases are stated such that an extension operator to the global domain exists as a bounded mapping. This approach can be applied recursively for the case of multiple subdomains. The abstract considerations are applied to general domains Ω that consist of nonoverlapping cubic subdomains that are aligned to a cartesian grid. In this setting it is possible to consider tensor wavelet bases for the subcubes and then to recursively apply *univariate extension operators*. This yields a global basis, where each individual basis function is again a tensor of extended univariate wavelets. To preserve the locality of the new basis, scale dependent univariate extension operators are considered that only extend wavelets with supports close to the boundary. The new generalized tensor wavelet basis reproduces the dimension independent approximation result of classical tensor wavelets under relatively mild assumptions. Indeed, the function that is approximated only needs to satisfy a piecewise weighted Sobolev smoothness assumption, that is, its restriction to the subcubes is assumed to satisfy the smoothness assumption required for classical tensor wavelet approximation. A regularity result for elliptic boundary value problems of order 2 on polygonal and polyhedral domains is derived that ensures the piecewise smoothness of the solution of the problem for smooth right-hand sides. Finally, numerical experiments confirm that the theoretical approximation rate of the basis is obtained in practice.

2 Multilevel preconditioning for sparse optimization of functionals with nonconvex fidelity terms

Authors: S. Dahlke, M. Fornasier, U. Friedrich, T. Raasch.
Journal: Journal of Inverse and Ill-Posed Problems, online December 2014.
Abstract: This paper is concerned with the development of numerical schemes for the minimization of functionals involving sparsity constraints and nonconvex fidelity terms. These functionals appear in a natural way in the context of Tikhonov regularization of nonlinear inverse problems with ℓ_1 penalty terms. Our method of minimization is based on a generalized conditional gradient scheme. It is well–known that these algorithms might converge quite slowly in practice. Therefore, we propose an acceleration which is based on a decreasing thresholding strategy. Its efficiency relies on certain spectral properties of the problem at hand. We show that under certain boundedness and contraction conditions the resulting algorithm is linearly convergent to a *global* minimizer and that the iteration is monotone with respect to the Tikhonov functional. We study important classes of operator equations to which our analysis can be applied. Moreover, we introduce a certain multilevel preconditioning strategy which in practice promotes the aforementioned spectral properties for problems where the nonlinearity is a perturbation of a linear operator.
MSC 2010: 65K10, 65J15, 41A25, 65N12, 65T60, 47J06, 47J25.
Key Words: Conditional gradient method, non-convex optimization, sparse minimization, (nonlinear) operator equations, iterative thresholding, multilevel preconditioning, wavelets.

2.1 Introduction

The aim of this paper is to derive an efficient numerical algorithm for the global minimization of functionals of the form

$$\Gamma_{\boldsymbol{\alpha}}(\mathbf{u}) := \big\|K(\mathbf{u}) - y\big\|_Y^2 + 2\|\mathbf{u}\|_{\ell_{1,\boldsymbol{\alpha}}(\mathcal{J})}, \quad \mathbf{u} \in \ell_2(\mathcal{J}), \tag{2.1.1}$$

where $K : \ell_2(\mathcal{J}) \to Y$ is a nonlinear, continuously Fréchet differentiable operator acting between the sequence space $\ell_2(\mathcal{J})$ over the countable index set $\mathcal{J}$ and a separable Hilbert space Y. Here $y \in Y$ is a given datum, and

$$\|\mathbf{u}\|_{\ell_{1,\boldsymbol{\alpha}}(\mathcal{J})} := \sum_{\mu\in\mathcal{J}} \alpha_\mu |u_\mu|$$

denotes the weighted ℓ_1-norm of $\mathbf{u}$ with respect to a positive weight sequence $\boldsymbol{\alpha} \in \mathbb{R}_+^{\mathcal{J}}$. We shall assume that there exists an $\alpha > 0$ such that $\alpha_\mu \geq \alpha$ for all $\mu \in \mathcal{J}$. Whenever the index set $\mathcal{J}$ is fixed and clear from the context, we will drop it in the notation and simply write ℓ_2 and $\ell_{1,\boldsymbol{\alpha}}$, respectively.

Typical examples where minimization problems of the form (2.1.1) arise are Tikhonov regularizations of nonlinear operator equations

$$\mathcal{K}(u) = y \tag{2.1.2}$$

when the forward operator $\mathcal{K} : X \to Y$ maps a separable Hilbert space X into Y. We refer, e.g., to [54, 107, 109] for a detailed discussion of Tikhonov regularization schemes. If the unknown solution is guaranteed to have a sparse expansion with respect to some suitable countable Riesz basis $\boldsymbol{\Psi} := \{\psi_\mu\}_{\mu\in\mathcal{J}}$ for X, it makes sense to utilize the ℓ_1-norm to promote sparse solutions. Denoting the linear synthesis operator associated to $\boldsymbol{\Psi}$ with

$$u = \sum_{\mu\in\mathcal{J}} u_\mu \psi_\mu =: \mathcal{F}(\mathbf{u}), \quad \mathbf{u} \in \ell_2(\mathcal{J}),$$

and setting $K := \mathcal{K}\circ\mathcal{F}$, the minimization of (2.1.1) will produce a sparsely populated coefficient array $\mathbf{u}$ with $K(\mathbf{u}) \approx y$. The modeling motivation is the search of the "simplest" (in this case modeled by the "sparsest") explanation to the given datum y, resulting from the nonlinear process $\mathcal{K}$, in the spirit of the Occam's razor. Moreover, it is known that, under certain smoothness conditions, the *global minimizers* of (2.1.1) are regularizers for the problem.

By now there is a vast literature concerning sparse regularization of nonlinear inverse problems, see for example [13, 15, 102, 123]. For most of the results in the literature related to minimizing algorithms for functionals of the type (2.1.1) usually only convergence to critical points is shown. Unfortunately, differently from global minimizers, nothing is really known concerning the regularization properties of critical points, significantly questioning the relevance of such convergence results.

The starting point of our present discussion is a generalized conditional gradient method which is known to guarantee the computation of subsequences converging to critical points of (2.1.1). The scope of this paper is to show under which *sufficient* conditions on $\mathcal{K}$ one may expect to have linear convergence of a suitable modification of this algorithm towards a *global minimizer*, hence guaranteeing regularization properties.

Several authors have independently proposed such an algorithm, see [53, 60, 116, 117] for the case of linear operators K and [13, 15] for the generalization to the nonlinear case. The general setting can be described as follows. One introduces an auxiliary parameter $\lambda \in \mathbb{R}_+$, and considers the splitting

$$\Gamma_{\boldsymbol{\alpha}}(\mathbf{u}) = \underbrace{\|K(\mathbf{u}) - y\|_Y^2 - \lambda\|\mathbf{u}\|_{\ell_2}^2}_{=:\Gamma_\lambda^{(1)}(\mathbf{u})} + \underbrace{\lambda\|\mathbf{u}\|_{\ell_2}^2 + 2\|\mathbf{u}\|_{\ell_{1,\boldsymbol{\alpha}}}}_{=:\Gamma_{\lambda,\boldsymbol{\alpha}}^{(2)}(\mathbf{u})}. \tag{2.1.3}$$

Then $\Gamma_\lambda^{(1)}$ is continuously Fréchet differentiable and $\Gamma_{\lambda,\boldsymbol{\alpha}}^{(2)}$ is convex, lower semicontinuous, and coercive with respect to $\|\cdot\|_{\ell_2}$, so that all the necessary properties to set up a generalized conditional gradient method are satisfied. The algorithm is given by

Algorithm 2.1.1 ISTA

1: Choose $\mathbf{u}^{(0)} \in \ell_{1,\boldsymbol{\alpha}}$; $n := 0$;

2: Determine descent direction $\mathbf{v}^{(n)}$

$$\mathbf{v}^{(n)} \in \arg\min_{\mathbf{v}\in\ell_2} \Big(2\langle \big(K'(\mathbf{u}^{(n)})\big)^* \big(K(\mathbf{u}^{(n)}) - y\big) - \lambda \mathbf{u}^{(n)}, \mathbf{v}\rangle_{\ell_2} + \lambda\|\mathbf{v}\|_{\ell_2}^2 + 2\|\mathbf{v}\|_{\ell_{1,\boldsymbol{\alpha}}}\Big); \tag{2.1.4}$$

3: Determine step size $s^{(n)}$

$$s^{(n)} \in \arg\min_{s\in[0,1]} \Gamma_{\boldsymbol{\alpha}}(\mathbf{u}^{(n)} + s(\mathbf{v}^{(n)} - \mathbf{u}^{(n)})); \tag{2.1.5}$$

4: Set $\mathbf{u}^{(n+1)} := \mathbf{u}^{(n)} + s^{(n)}(\mathbf{v}^{(n)} - \mathbf{u}^{(n)})$; $n := n+1$; return to step 2.

Here $\big(K'(\mathbf{u}^{(n)})\big)^* \in \mathcal{L}(Y, \ell_2)$ denotes the adjoint mapping of $K'(\mathbf{u}^{(n)}) \in \mathcal{L}(\ell_2, Y)$. We refer to [13] for a detailed discussion and convergence analysis of Algorithm 2.1.1. If the parameter λ is chosen large enough, it is possible to choose $s^{(n)} = 1$ and to omit the third step of the algorithm, see [13, Lemma 2.4]. Throughout this paper we always make this assumption, hence we focus on the minimization problem (2.1.4) in the following. Observe that by expanding the quadratic term below, (2.1.4) is equivalent to

$$\arg\min_{\mathbf{v}\in\ell_2} \|\mathbf{v} - \big(\mathbf{u}^{(n)} - \frac{1}{\lambda}\big(K'(\mathbf{u}^{(n)})\big)^*\big(K(\mathbf{u}^{(n)}) - y\big)\big)\|_{\ell_2}^2 + 2\|\mathbf{v}\|_{\ell_{1,\frac{\boldsymbol{\alpha}}{\lambda}}}. \tag{2.1.6}$$

The minimizer of such a functional combining an ℓ_2-norm fidelity term and weighted ℓ_1-norm penalization can be directly computed using a soft thresholding operation, see [18, 46]. By defining

$$S_\alpha(x) := \begin{cases} x - \alpha, & x > \alpha, \\ 0, & |x| \le \alpha, \\ x + \alpha, & x < -\alpha, \end{cases}$$

and $\mathbb{S}_{\boldsymbol{\alpha}}(\mathbf{u})_\mu := S_{\alpha_\mu}(u_\mu)$ it holds that

$$\mathbb{S}_{\boldsymbol{\alpha}}(\mathbf{a}) = \arg\min_{\mathbf{v}\in\ell_2} \|\mathbf{v} - \mathbf{a}\|_{\ell_2}^2 + 2\|\mathbf{v}\|_{\ell_{1,\boldsymbol{\alpha}}}. \tag{2.1.7}$$

Consequently, through (2.1.6), we obtain that (2.1.4) is uniquely solved by

$$\mathbf{v}^{(n)} = \mathbb{S}_{\frac{\boldsymbol{\alpha}}{\lambda}}\Big(\mathbf{u}^{(n)} + \frac{1}{\lambda}\big(K'(\mathbf{u}^{(n)})\big)^*\big(y - K(\mathbf{u}^{(n)})\big)\Big). \tag{2.1.8}$$

This explains why Algorithm 2.1.1 is also known as the *iterated soft thresholding algorithm (ISTA)* or the *thresholded Landweber iteration*.

The convergence of Algorithm 2.1.1 for nonlinear operators K was studied in [15]. There it was shown that the sequence $\big(\mathbf{u}^{(n)}\big)_{n\in\mathbb{N}}$ has subsequences which are guaranteed to converge to a *stationary point* $\mathbf{u}^*$ of (2.1.1), i.e.,

$$\mathbf{u}^* \in \arg\min_{\mathbf{v}\in\ell_2} 2\langle \big(K'(\mathbf{u}^*)\big)^*\big(K(\mathbf{u}^*)-y\big) - \lambda\mathbf{u}^*, \mathbf{v}\rangle_{\ell_2} + \lambda\|\mathbf{v}\|_{\ell_2}^2 + 2\|\mathbf{v}\|_{\ell_{1,\boldsymbol{\alpha}}}.$$

However, it is known from the linear case, that the algorithm in its most basic form converges rather slowly. Strategies to accelerate the convergence of the method are necessary for its applicability. In [36] three of us considered the case of *linear* operators K and proposed to choose a decreasing thresholding strategy for the parameters $\boldsymbol{\alpha}^{(n)}$. In the setting of [36], $\mathbf{u}^* = \mathbf{u}^*_{\boldsymbol{\alpha}}$ is a global minimizer of (2.1.1). Moreover it has been possible to show that the resulting scheme is guaranteed to converge linearly, under spectral conditions of K, the so-called restricted isometry property, see (2.2.51) below. Furthermore this property is obtainable for certain classes of operators by means of multilevel preconditioners, we refer to [36] for details. This paper is concerned with the generalization of this strategy to nonlinear operators K. That is, we are interested in the convergence analysis of the iteration

$$\mathbf{u}^{(n+1)} := \mathbb{S}_{\frac{1}{\lambda}\boldsymbol{\alpha}^{(n)}}\Big(\mathbf{u}^{(n)} + \frac{1}{\lambda}\big(K'(\mathbf{u}^{(n)})\big)^*\big(y - K(\mathbf{u}^{(n)})\big)\Big), \tag{2.1.9}$$

where $\boldsymbol{\alpha}^{(n)} \in \mathbb{R}_+^{\mathcal{J}}$ is an entrywise decreasing sequence with $\lim_{n\to\infty}\boldsymbol{\alpha}^{(n)} = \boldsymbol{\alpha}$ and $\alpha_\mu^{(n)}, \alpha_\mu \geq \alpha \in \mathbb{R}_+, \mu \in \mathcal{J}$.

The basic convergence analysis is outlined in Section 2.2. Our analysis relies on two fundamental assumptions. We need that the operator K satisfies certain boundedness and Lipschitz continuity conditions, see (2.2.12). Moreover, we have to assume that the operator

$$T : \ell_2 \to \ell_2, \quad \mathbf{v} \mapsto T(\mathbf{v}) := \mathbf{v} + \frac{1}{\lambda}\big(K'(\mathbf{v})\big)^*\big(y - K(\mathbf{v})\big)$$

is a contraction on a sufficiently small ball around a critical point $\mathbf{u}^*$ of the functional $\Gamma_{\boldsymbol{\alpha}}$, which will turn out to be the unique global minimizer there. Then the iteration is linearly convergent, i.e.,

$$\|\mathbf{u}^* - \mathbf{u}^{(n)}\|_{\ell_2} \leq \gamma^n\|\mathbf{u}^*\|_{\ell_2}, \quad \text{for some } \gamma \leq 1.$$

Moreover, the iteration is monotone in the sense that

$$\Gamma_{\boldsymbol{\alpha}^{(n+1)}}(\mathbf{u}^{(n+1)}) < \Gamma_{\boldsymbol{\alpha}^{(n)}}(\mathbf{u}^{(n)}), \tag{2.1.10}$$

provided that $\mathbf{u}^{(n)}$ is not a critical point of $\Gamma_{\boldsymbol{\alpha}^{(n)}}$. These properties are specified in Theorem 2.2.4 which is the main result of this paper.

The local contraction condition (2.2.32) on T may be hard to verify. In the Sections 2.2.2 and 2.2.3, we discuss in detail two classes of operators where it is satisfied. The first class consists of operators with bounded second derivatives and first derivative that satisfies the contraction property. The second class is given by nonlinear perturbations of linear operators satisfying the restricted isometry property (2.2.51). As already shown in [36] for large classes of linear operators $\mathcal{K}$ where (2.2.51) fails, it can actually be restored by preconditioning. Details will be outlined for the case of a nonlinear $\mathcal{K}$ which is a mild perturbation of a linear operator in Section 2.3.

The analysis in this paper is performed in an infinite-dimensional setting. In this general setting, clearly the operator K and K' cannot be evaluated exactly. Therefore, in Section 2.4, we discuss strategies to solve the infinite-dimensional problem by turning it into a finite-dimensional one and using the expected sparsity of the minimizer. If implementable approximations of the actions of K and K' up to prescribed tolerances are applied, then the resulting inexact, but implementable, version of the algorithm will again be linearly convergent. If the underlying Riesz basis is of wavelet type, then the desired approximations are known in the literature for certain classes of nonlinearities [45, 28, 79].

2.2 Convergence analysis

In this section we analyze the convergence properties of the iteration (2.1.9). As a first step we show that under relatively mild assumptions $\Gamma_{\boldsymbol{\alpha}^{(n)}}(\mathbf{u}^{(n)})$ decreases monotonically. It is known that in the case of constant thresholding parameters $\boldsymbol{\alpha}^{(n)} = \boldsymbol{\alpha}$, $n \in \mathbb{N}$, the sequence $\left(\mathbf{u}^{(n)}\right)_{n\in\mathbb{N}}$ has a convergent subsequence and every convergent subsequence converges to a stationary point of (2.1.1). However, we are particularly interested in the global minimizer of (2.1.1). Therefore, we prove that under more restrictive assumptions and for decreasing thresholing parameters $\boldsymbol{\alpha}^{(n)}$ the iterates converge linearly to the global minimizer of (2.1.1). In the remainder of this section we present examples of settings where our analysis can be applied. In Section 2.2.2, we describe how our assumptions can be fulfilled under smoothness conditions on the nonlinear operator K and its derivative. In Section 2.2.3 we present the important special case where K can be expressed as the sum of a linear operator satisfying the restricted isometry property, and a small nonlinear perturbation.

2.2.1 A general convergence result

We are particularly interested in computing approximations with the smallest possible number of nonzero entries to solutions of (2.1.2). As a benchmark, we recall that the most economical approximations of a given vector $\mathbf{v} \in \ell_2$ are provided by the *best N-term approximations* $\mathbf{v}_N$, defined by discarding in $\mathbf{v}$ all but the $N \in \mathbb{N}_0$ largest coefficients in absolute value. The error of best N-term approximation is defined as

$$\sigma_N(\mathbf{v}) := \|\mathbf{v} - \mathbf{v}_N\|_{\ell_2}. \tag{2.2.1}$$

The subspace of all ℓ_2 vectors with best N-term approximation rate $s > 0$, i.e., $\sigma_N(\mathbf{v}) \lesssim N^{-s}$ for some decay rate $s > 0$, is commonly referred to as the *weak ℓ_τ space* $\ell_\tau^w(\mathcal{J})$, for $\tau = (s + \frac{1}{2})^{-1}$, which, endowed with

$$|\mathbf{v}|_{\ell_\tau^w(\mathcal{J})} := \sup_{N \in \mathbb{N}_0} (N+1)^s \sigma_N(\mathbf{v}), \tag{2.2.2}$$

becomes the quasi-Banach space $(\ell_\tau^w(\mathcal{J}), |\cdot|_{\ell_\tau^w(\mathcal{J})})$. Moreover, for any $0 < \varepsilon \leq 2-\tau$, we have the continuous embeddings $\ell_\tau(\mathcal{J}) \hookrightarrow \ell_\tau^w(\mathcal{J}) \hookrightarrow \ell_{\tau+\varepsilon}(\mathcal{J})$, justifying why $\ell_\tau^w(\mathcal{J})$ is called weak $\ell_\tau(\mathcal{J})$. As before we omit the dependency on the index set $\mathcal{J}$ whenever it is clear from the context.

When it comes to the concrete computations of good approximations with a small number of active coefficients, one frequently utilizes certain thresholding procedures. Here small entries of a given vector are simply discarded, whereas the large entries may be slightly modified. In this paper, we will make use of *soft thresholding* that we already introduced in (2.1.7). It is well-known that $\mathbb{S}_{\boldsymbol{\alpha}}$ is non-expansive for any $\boldsymbol{\alpha} \in \mathbb{R}_+^{\mathcal{J}}$,

Moreover, for any fixed $x \in \mathbb{R}$, the mapping $\beta \mapsto S_\beta(x)$ is Lipschitz continuous with

$$\big|S_\beta(x) - S_{\beta'}(x)\big| \leq |\beta - \beta'|, \quad \text{for all } \beta, \beta' \geq 0. \tag{2.2.3}$$

We readily infer the following technical estimate (for the proof we refer the reader to [36]).

Lemma 2.2.1. *Assume $\mathbf{v} \in \ell_2$, $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}_+^{\mathcal{J}}$ such that $0 < \alpha = \min\big(\inf_\mu \alpha_\mu, \inf_\mu \beta_\mu\big)$, and define*

$$\Lambda_\alpha(\mathbf{v}) := \big\{\mu \in \mathcal{J} : |v_\mu| > \alpha\big\}.$$

Then

$$\big\|\mathbb{S}_{\boldsymbol{\alpha}}(\mathbf{v}) - \mathbb{S}_{\boldsymbol{\beta}}(\mathbf{v})\big\|_{\ell_2} \leq \Big(\#\Lambda_\alpha(\mathbf{v})\Big)^{1/2} \max_{\mu \in \Lambda_\alpha(\mathbf{v})} |\alpha_\mu - \beta_\mu|. \tag{2.2.4}$$

In the sequel, we shall also use the following support size estimate, the proof of which follows the lines of Lemma 5.1 in [26], more details are provided in [36].

Lemma 2.2.2. *Let $\mathbf{v} \in \ell_\tau^w$ and $\mathbf{w} \in \ell_2$ with $\|\mathbf{v} - \mathbf{w}\|_{\ell_2} \leq \varepsilon$. Assume $\boldsymbol{\alpha} = (\alpha_\mu)_{\mu \in \mathcal{J}} \in \mathbb{R}_+^{\mathcal{J}}$ and $\inf_\mu \alpha_\mu = \alpha > 0$. Then it holds that*

$$\#\operatorname{supp} \mathbb{S}_{\boldsymbol{\alpha}}(\mathbf{w}) \leq \#\Lambda_\alpha(\mathbf{w}) \leq \frac{4\varepsilon^2}{\alpha^2} + 4C|\mathbf{v}|_{\ell_\tau^w}^\tau \alpha^{-\tau}, \tag{2.2.5}$$

where $C = C(\tau) > 0$. In particular if $\mathbf{v} \in \ell_0$ then the estimate is refined

$$\#\operatorname{supp} \mathbb{S}_{\boldsymbol{\alpha}}(\mathbf{w}) \leq \#\Lambda_\alpha(\mathbf{w}) \leq \frac{4\varepsilon^2}{\alpha^2} + \|\mathbf{v}\|_{\ell_0}. \tag{2.2.6}$$

For the analysis of the iteration (2.1.9), we will always assume that the datum $y \in Y$ is fixed and contained in a bounded set, i.e.,

$$\|y\|_Y \leq C_Y < \infty. \tag{2.2.7}$$

In this setting, we define the operator

$$T : \ell_2 \to \ell_2, \quad \mathbf{v} \mapsto T(\mathbf{v}) := \mathbf{v} + \frac{1}{\lambda}\big(K'(\mathbf{v})\big)^*\big(y - K(\mathbf{v})\big). \tag{2.2.8}$$

In the following we want to show the convergence of the iteration (2.1.9) to stationary points of $\Gamma_{\boldsymbol{\alpha}}$ and to estimate the rate of convergence. In order to do that we shall in particular show that, under certain *local* contraction properties of the operator T, the stationary point is actually unique in a predetermined ball around $\mathbf{0}$ and coincides with the global minimizer of $\Gamma_{\boldsymbol{\alpha}}$. First of all, we need to characterize the ball where the interesting stationary points should be searched.

To this end, we recall that for each $\boldsymbol{\alpha} \in \mathbb{R}^{\mathcal{J}}_+, \alpha_\mu \geq \alpha > 0, \mu \in \mathcal{J}$, the related energy functional $\Gamma_{\boldsymbol{\alpha}}$ from (2.1.1) is coercive, i.e., $\Gamma_{\boldsymbol{\alpha}}(\mathbf{v}) \to \infty$ as $\|\mathbf{v}\|_{\ell_2} \to \infty$. In particular this implies that

$$R := \sup\big\{\|\mathbf{v}\|_{\ell_2} : \Gamma_{\boldsymbol{\alpha}}(\mathbf{v}) \leq \Gamma_{\boldsymbol{\alpha}^{(0)}}(\mathbf{0})\big\} \tag{2.2.9}$$

is finite, and we define

$$B(R) := \big\{\mathbf{v} \in \ell_2 : \|\mathbf{v}\|_{\ell_2} \leq R\big\}. \tag{2.2.10}$$

Notice that for $\mathbf{v} \in \ell_2$ such that $\Gamma_{\boldsymbol{\alpha}}(\mathbf{v}) \leq \Gamma_{\boldsymbol{\alpha}^{(0)}}(\mathbf{0})$, we have

$$2\alpha\|\mathbf{v}\|_{\ell_2} \leq 2\|\mathbf{v}\|_{\ell_{1,\boldsymbol{\alpha}}} \leq \Gamma_{\boldsymbol{\alpha}}(\mathbf{v}) \leq \Gamma_{\boldsymbol{\alpha}^{(0)}}(\mathbf{0}),$$

hence,

$$R \leq \frac{\Gamma_{\boldsymbol{\alpha}^{(0)}}(\mathbf{0})}{2\alpha}. \tag{2.2.11}$$

For the remainder of this section we will make the following additional assumption. The operators K and K' are Lipschitz continuous on closed bounded sets, i.e., for all closed and bounded $\mathcal{O} \subset \ell_2$ we assume

$$\begin{aligned} \|K(\mathbf{u}) - K(\mathbf{v})\|_Y &\leq C_K^{Lip}(\mathcal{O})\|\mathbf{u} - \mathbf{v}\|_{\ell_2}, \quad \mathbf{u}, \mathbf{v} \in \mathcal{O}, \\ \|K'(\mathbf{u}) - K'(\mathbf{v})\|_{\mathcal{L}(\ell_2, Y)} &\leq C_{K'}^{Lip}(\mathcal{O})\|\mathbf{u} - \mathbf{v}\|_{\ell_2}, \quad \mathbf{u}, \mathbf{v} \in \mathcal{O}. \end{aligned} \tag{2.2.12}$$

With a slight abuse of notation we denote the Lipschitz constants of K and K' on $B(R)$ by $C_K^{Lip}(R)$ and $C_{K'}^{Lip}(R)$, respectively. In particular (2.2.12) implies that K and K' are bounded on closed and bounded sets. Indeed, let $\mathcal{O} \subset \ell_2$ and $\mathbf{v}_0 \in \mathcal{O}$. Then we may bound K by estimating

$$\begin{aligned} \sup_{\mathbf{v}\in\mathcal{O}} \|K(\mathbf{v})\|_Y &\leq \sup_{\mathbf{v}\in\mathcal{O}} \|K(\mathbf{v}) - K(\mathbf{v}_0)\|_Y + \|K(\mathbf{v}_0)\|_Y \\ &\leq C_K^{Lip}(\mathcal{O}) \sup_{\mathbf{v}\in\mathcal{O}} \|\mathbf{v} - \mathbf{v}_0\|_{\ell_2} + \|K(\mathbf{v}_0)\|_Y < \infty, \end{aligned} \tag{2.2.13}$$

and K' by a similar estimate. We introduce the abbreviations

$$C_K^{bnd}(R) := \sup_{\mathbf{v}\in B(R)} \|K(\mathbf{v})\|_Y, \quad C_{K'}^{bnd}(R) := \sup_{\mathbf{v}\in B(R)} \|K'(\mathbf{v})\|_{\mathcal{L}(\ell_2,Y)}. \tag{2.2.14}$$

With these preliminaries, we can formulate the following proposition, which generalizes [13, Lemma 2.4].

Proposition 2.2.3. *Suppose that* (2.2.7) *and* (2.2.12) *hold. For some* $\lambda_0 > 0$ *and* R *as in* (2.2.9) *we define*

$$R' := R + \frac{1}{\lambda_0} C_{K'}^{bnd}(R)(C_K^{bnd}(R) + C_Y). \tag{2.2.15}$$

Then $\|K(\cdot) - y\|_Y^2$ *is locally Lipschitz. We choose in* (2.1.3)

$$\lambda > \lambda_{\min} := \max\left(\lambda_0, C_{K'}^{Lip}(R')(C_K^{bnd}(R') + C_Y) + C_{K'}^{bnd}(R')C_K^{Lip}(R')\right) \tag{2.2.16}$$

and denote by $\left(\mathbf{u}^{(n)}\right)_{n\in\mathbb{N}}$ *the iterates of the decreasing thresholding iteration* (2.1.9) *starting from* $\mathbf{u}^{(\mathbf{0})} = \mathbf{0}$. *Then it holds*

$$\Gamma_{\boldsymbol{\alpha}^{(n+1)}}(\mathbf{u}^{(n+1)}) < \Gamma_{\boldsymbol{\alpha}^{(n)}}(\mathbf{u}^{(n)}), \tag{2.2.17}$$

as long as $\mathbf{u}^{(n)}$ *is not yet a critical point of* $\Gamma_{\boldsymbol{\alpha}^{(n)}}$. *Furthermore the iterates fulfill the bound*

$$\|\mathbf{u}^{(n)}\|_{\ell_2} \le R, \quad n \in \mathbb{N}. \tag{2.2.18}$$

Proof. We shall prove by induction over n that

$$\|\mathbf{u}^{(n)}\|_{\ell_2} \le R \text{ and } \Gamma_{\boldsymbol{\alpha}^{(n)}}(\mathbf{u}^{(n)}) \le \Gamma_{\boldsymbol{\alpha}^{(0)}}(\mathbf{0}). \tag{2.2.19}$$

We will show that if λ is chosen according to (2.2.16) and $\mathbf{u}^{(n)} \neq \mathbf{u}^{(n+1)}$, which is the case if $\mathbf{u}^{(n)}$ is no critical point of $\Gamma_{\boldsymbol{\alpha}^{(n)}}$, then this implies

$$\Gamma_{\boldsymbol{\alpha}^{(n)}}(\mathbf{u}^{(n+1)}) < \Gamma_{\boldsymbol{\alpha}^{(n)}}(\mathbf{u}^{(n)}). \tag{2.2.20}$$

From the fact that $\boldsymbol{\alpha}^{(n)}$ decreases componentwise to $\boldsymbol{\alpha}$, together with (2.2.20) and (2.2.19) we obtain

$$\Gamma_{\boldsymbol{\alpha}}(\mathbf{u}^{(n+1)}) \le \Gamma_{\boldsymbol{\alpha}^{(n+1)}}(\mathbf{u}^{(n+1)}) \le \Gamma_{\boldsymbol{\alpha}^{(n)}}(\mathbf{u}^{(n+1)}) < \Gamma_{\boldsymbol{\alpha}^{(n)}}(\mathbf{u}^{(n)}) \le \Gamma_{\boldsymbol{\alpha}^{(0)}}(\mathbf{0}).$$

By (2.2.9) this also implies the validity of (2.2.19) for $n \to n+1$, and simultaneously of (2.2.17) and (2.2.18) for all $n \in \mathbb{N}$.

Notice that (2.2.19) in particular holds for $n = 0$. We begin by proving $\mathbf{u}^{(n+1)} \in B(R')$, where R' is defined in (2.2.15). We use the fact that soft shrinkage is nonexpansive, together with (2.2.14) and (2.2.7), to estimate

$$\begin{aligned} \|\mathbf{u}^{(n+1)}\|_{\ell_2} &\le \|\mathbf{u}^{(n)} + \frac{1}{\lambda}\big(K'(\mathbf{u}^{(n)})\big)^*\big(y - K(\mathbf{u}^{(n)})\big)\|_{\ell_2} \\ &\le \|\mathbf{u}^{(n)}\|_{\ell_2} + \frac{1}{\lambda} C_{K'}^{bnd}(R)(C_K^{bnd}(R) + C_Y) \le R'. \end{aligned} \tag{2.2.21}$$

Hence it follows that $\mathbf{u}^{(n)}, \mathbf{u}^{(n+1)} \in B(R')$. To prove (2.2.20) we shall use (2.1.4), reformulated for $\mathbf{u}^{(n)}$ and $\Gamma_{\boldsymbol{\alpha}}(\mathbf{u}^{(n)})$. In (2.1.3) we introduced the splitting

$$\Gamma_{\boldsymbol{\alpha}}(\mathbf{u}) = \Gamma_{\lambda}^{(1)}(\mathbf{u}) + \Gamma_{\lambda,\boldsymbol{\alpha}}^{(2)}(\mathbf{u}),$$

where $\Gamma_{\lambda}^{(1)}$ is continuously Fréchet differentiable. The derivative of $\Gamma_{\lambda}^{(1)}$ was already implicitly stated in (2.1.4) and may be reformulated by means of T as follows:

$$\big(\Gamma_{\lambda}^{(1)}\big)'(\mathbf{u})\mathbf{v} = 2\langle \big(K'(\mathbf{u})\big)^* \big(K(\mathbf{u}) - y\big) - \lambda \mathbf{u}, \mathbf{v}\rangle_{\ell_2} = -2\lambda \langle T(\mathbf{u}), \mathbf{v}\rangle_{\ell_2}. \tag{2.2.22}$$

Recall that by means of (2.1.7), (2.1.6), and (2.2.22), the definition of $\mathbf{u}^{(n+1)}$ in (2.1.9) can be reformulated as

$$\begin{aligned}
\mathbf{u}^{(n+1)} &= \arg\min_{\mathbf{v}\in\ell_2} \|\mathbf{v} - T(\mathbf{u}^{(n)})\|_{\ell_2}^2 + 2\|\mathbf{v}\|_{\ell_{1,\frac{1}{\lambda}\boldsymbol{\alpha}^{(n)}}} \\
&= \arg\min_{\mathbf{v}\in\ell_2} -2\lambda\langle T(\mathbf{u}^{(n)}), \mathbf{v}\rangle_{\ell_2} + \lambda\|\mathbf{v}\|_{\ell_2}^2 + 2\|\mathbf{v}\|_{\ell_{1,\boldsymbol{\alpha}^{(n)}}} \\
&= \arg\min_{\mathbf{v}\in\ell_2} \big(\Gamma_{\lambda}^{(1)}\big)'(\mathbf{u}^{(n)})\mathbf{v} + \Gamma_{\lambda,\boldsymbol{\alpha}^{(n)}}^{(2)}(\mathbf{v}).
\end{aligned}$$

In particular it follows that

$$\big(\Gamma_{\lambda}^{(1)}\big)'(\mathbf{u}^{(n)})\mathbf{u}^{(n+1)} + \Gamma_{\lambda,\boldsymbol{\alpha}^{(n)}}^{(2)}(\mathbf{u}^{(n+1)}) \le \big(\Gamma_{\lambda}^{(1)}\big)'(\mathbf{u}^{(n)})\mathbf{u}^{(n)} + \Gamma_{\lambda,\boldsymbol{\alpha}^{(n)}}^{(2)}(\mathbf{u}^{(n)})$$

holds, which is equivalent to

$$\big(\Gamma_{\lambda}^{(1)}\big)'(\mathbf{u}^{(n)})(\mathbf{u}^{(n+1)} - \mathbf{u}^{(n)}) \le \Gamma_{\lambda,\boldsymbol{\alpha}^{(n)}}^{(2)}(\mathbf{u}^{(n)}) - \Gamma_{\lambda,\boldsymbol{\alpha}^{(n)}}^{(2)}(\mathbf{u}^{(n+1)}). \tag{2.2.23}$$

Next, we apply the fundamental theorem of calculus to $\Gamma_{\lambda}^{(1)}$ and write

$$\begin{aligned}
&\Gamma_{\lambda}^{(1)}(\mathbf{u}^{(n+1)}) - \Gamma_{\lambda}^{(1)}(\mathbf{u}^{(n)}) \\
&\quad = \int_0^1 \big(\Gamma_{\lambda}^{(1)}\big)'(\mathbf{u}^{(n)} + \tau(\mathbf{u}^{(n+1)} - \mathbf{u}^{(n)}))(\mathbf{u}^{(n+1)} - \mathbf{u}^{(n)})\,\mathrm{d}\tau \\
&\quad = \int_0^1 \Big(\big(\Gamma_{\lambda}^{(1)}\big)'(\mathbf{u}^{(n)} + \tau(\mathbf{u}^{(n+1)} - \mathbf{u}^{(n)})) - \big(\Gamma_{\lambda}^{(1)}\big)'(\mathbf{u}^{(n)})\Big)(\mathbf{u}^{(n+1)} - \mathbf{u}^{(n)})\,\mathrm{d}\tau \\
&\qquad + \big(\Gamma_{\lambda}^{(1)}\big)'(\mathbf{u}^{(n)})(\mathbf{u}^{(n+1)} - \mathbf{u}^{(n)}).
\end{aligned}$$

This, together with (2.2.23) and (2.2.22) yields

$$\begin{aligned}
&\Gamma_{\boldsymbol{\alpha}^{(n)}}(\mathbf{u}^{(n+1)}) - \Gamma_{\boldsymbol{\alpha}^{(n)}}(\mathbf{u}^{(n)}) \\
&\quad = \Big(\Gamma_{\lambda}^{(1)}(\mathbf{u}^{(n+1)}) + \Gamma_{\lambda,\boldsymbol{\alpha}^{(n)}}^{(2)}(\mathbf{u}^{(n+1)})\Big) - \Big(\Gamma_{\lambda}^{(1)}(\mathbf{u}^{(n)}) + \Gamma_{\lambda,\boldsymbol{\alpha}^{(n)}}^{(2)}(\mathbf{u}^{(n)})\Big) \\
&\quad \le \int_0^1 \Big(\big(\Gamma_{\lambda}^{(1)}\big)'(\mathbf{u}^{(n)} + \tau(\mathbf{u}^{(n+1)} - \mathbf{u}^{(n)})) - \big(\Gamma_{\lambda}^{(1)}\big)'(\mathbf{u}^{(n)})\Big)(\mathbf{u}^{(n+1)} - \mathbf{u}^{(n)})\,\mathrm{d}\tau \\
&\quad = \int_0^1 2\langle \big(K'(\mathbf{u}^{(n)} + \tau(\mathbf{u}^{(n+1)} - \mathbf{u}^{(n)}))\big)^* \big(K(\mathbf{u}^{(n)} + \tau(\mathbf{u}^{(n+1)} - \mathbf{u}^{(n)})) - y\big) \\
&\qquad - \big(K'(\mathbf{u}^{(n)})\big)^* \big(K(\mathbf{u}^{(n)}) - y\big),\, (\mathbf{u}^{(n+1)} - \mathbf{u}^{(n)})\rangle_{\ell_2}\,\mathrm{d}\tau - \lambda\|\mathbf{u}^{(n+1)} - \mathbf{u}^{(n)}\|_{\ell_2}^2.
\end{aligned} \tag{2.2.24}$$

Moreover, by the assumptions (2.2.12) on K and K', we can estimate for $\mathbf{u}, \mathbf{v} \in B(R')$

$$\begin{aligned}
&\|\big(K'(\mathbf{u})\big)^*(K(\mathbf{u}) - y) - \big(K'(\mathbf{v})\big)^*(K(\mathbf{v}) - y)\|_{\ell_2} \\
&\quad = \|\Big(\big(K'(\mathbf{u})\big)^* - \big(K'(\mathbf{v})\big)^*\Big)(K(\mathbf{u}) - y) + \big(K'(\mathbf{v})\big)^*\big(K(\mathbf{u}) - K(\mathbf{v})\big)\|_{\ell_2} \\
&\quad \leq \lambda_{\min}\|\mathbf{u} - \mathbf{v}\|_{\ell_2}.
\end{aligned}$$

We apply this inequality for the special case $\mathbf{u} = \mathbf{u}^{(n)} + \tau(\mathbf{u}^{(n+1)} - \mathbf{u}^{(n)}), \mathbf{v} = \mathbf{u}^{(n)}$, to further estimate (2.2.24) as follows:

$$\begin{aligned}
&\Gamma_{\boldsymbol{\alpha}^{(n)}}(\mathbf{u}^{(n+1)}) - \Gamma_{\boldsymbol{\alpha}^{(n)}}(\mathbf{u}^{(n)}) \\
&\quad \leq \int_0^1 2\tau\lambda_{\min}\|\mathbf{u}^{(n+1)} - \mathbf{u}^{(n)}\|_{\ell_2}^2 \,\mathrm{d}\tau - \lambda\|\mathbf{u}^{(n+1)} - \mathbf{u}^{(n)}\|_{\ell_2}^2 \qquad (2.2.25) \\
&\quad = (\lambda_{\min} - \lambda)\|\mathbf{u}^{(n+1)} - \mathbf{u}^{(n)}\|_{\ell_2}^2.
\end{aligned}$$

Furthermore the right-hand side is negative if λ is chosen according to (2.2.16) and $\mathbf{u}^{(n)} \neq \mathbf{u}^{(n+1)}$, which is the case if $\mathbf{u}^{(n)}$ is no critical point of $\Gamma_{\boldsymbol{\alpha}^{(n)}}$, and this shows (2.2.20) and concludes the proof. □

Notice that we decided to start our iteration from $\mathbf{u}^{(\mathbf{0})} = \mathbf{0}$. On the one hand, this choice is motivated by the fact that a priori we do not dispose of any information on potentially interesting stationary points and an arbitrary choice of the initial iteration has to be made. On the other hand, as we will show below, under certain assumptions, we will be able to identify in this way the unique global minimizer of the functional $\Gamma_{\boldsymbol{\alpha}}$. As we are seeking for stationary points which are limits of the sequence $\big(\mathbf{u}^{(n)}\big)_{n\in\mathbb{N}}$ of the iterates of the decreasing thresholding iteration (2.1.9) starting from $\mathbf{u}^{(\mathbf{0})} = \mathbf{0}$, in view of Proposition 2.2.3 we can assume without loss of generality that interesting stationary points $\mathbf{u}^*$ belong to the ball $B(R)$. This assumption is not void, because a global minimizer $\mathbf{u}^\circ$ of $\Gamma_{\boldsymbol{\alpha}}$ necessarily has to lie in the ball $B(R)$, because $\Gamma_{\boldsymbol{\alpha}}(\mathbf{u}^\circ) < \Gamma_{\boldsymbol{\alpha}}(\mathbf{0})$. We shall also show below that all the iterates $\big(\mathbf{u}^{(n)}\big)_{n\in\mathbb{N}}$ are actually additionally located within the ball

$$\mathcal{B} := \{\mathbf{v} \in \ell_2 : \|\mathbf{u}^* - \mathbf{v}\|_{\ell_2} \leq \|\mathbf{u}^*\|_{\ell_2}\}, \qquad (2.2.26)$$

where $\mathbf{u}^*$ is an arbitrary stationary point of $\Gamma_{\boldsymbol{\alpha}}$ within $B(R)$. Hence, under the regularity assumption so far made for the operators K and K', the reference domain of the iterations of the algorithm is $\mathcal{B} \cap B(R)$. Within this setting we make the following assumption: T satisfies the Lipschitz condition

$$\|T(\mathbf{u}^*) - T(\mathbf{v})\|_{\ell_2} \leq C_T^{Lip}\|\mathbf{u}^* - \mathbf{v}\|_{\ell_2}, \quad \mathbf{v} \in \mathcal{B} \cap B(R),\, y \in Y,\, \|y\|_Y \leq C_Y, \quad (2.2.27)$$

for any fixed stationary point $\mathbf{u}^*$. (As we shall see below, such a condition is not so strong as we shall apply it to only *one* stationary point.) Furthermore we define for some fixed $\lambda_0 > 0$ the analogue of (2.2.15) on $\mathcal{B} \cap B(R)$, that is

$$R'' := R + \frac{1}{\lambda_0} C_{K'}^{bnd}(\mathcal{B} \cap B(R))(C_K^{bnd}(\mathcal{B} \cap B(R)) + C_Y). \qquad (2.2.28)$$

Then, the following convergence theorem holds.

Theorem 2.2.4. *Let $\mathbf{u}^*$ be a stationary point of (2.1.1) that satisfies $T(\mathbf{u}^*) \in \ell_\tau^w$ for some $0 < \tau < 2$. For some $\lambda_0 > 0$ and R'' as in (2.2.28) we choose*

$$\begin{aligned}\lambda > \max\Big(\lambda_0, \big(C_{K'}^{Lip}(\mathcal{B} \cap B(R'')(C_K^{bnd}(\mathcal{B} \cap B(R'')) + C_Y) \\ + C_{K'}^{bnd}(\mathcal{B} \cap B(R''))C_K^{Lip}(\mathcal{B} \cap B(R''))\big)\Big).\end{aligned} \tag{2.2.29}$$

Furthermore let $\boldsymbol{\alpha}^{(n)}, \boldsymbol{\alpha} \in \mathbb{R}_+^{\mathcal{J}}$ with $\alpha_\mu^{(n)} \geq \alpha_\mu \geq \alpha \in \mathbb{R}_+, \mu \in \mathcal{J}$. We set

$$L := \frac{4(C_T^{Lip})^2 \|\mathbf{u}^*\|_{\ell_2}^2 \lambda^2}{\alpha^2} + 4C|T(\mathbf{u}^*)|_{\ell_\tau^w}^\tau \left(\frac{\alpha}{\lambda}\right)^{-\tau}, \tag{2.2.30}$$

with C as in Lemma 2.2.2 and C_T^{Lip} as in (2.2.27). Moreover we define the set

$$\mathcal{B}_L := \{\mathbf{v} \in \ell_2 : \|\mathbf{u}^* - \mathbf{v}\|_{\ell_2} \leq \|\mathbf{u}^*\|_{\ell_2}, \#\operatorname{supp} \mathbf{v} \leq L\}. \tag{2.2.31}$$

Let us assume that there exists some $0 < \gamma_0 < 1$, such that for all $\mathbf{v} \in \mathcal{B}_L$ and $\operatorname{supp}(\mathbf{v}) \subset \Lambda \subset \mathcal{J}$ with $\#\Lambda \leq 2L$

$$\|\big(T(\mathbf{u}^*) - T(\mathbf{v})\big)_{|S^* \cup \Lambda}\|_{\ell_2(S^* \cup \Lambda)} \leq \gamma_0 \|\mathbf{u}^* - \mathbf{v}\|_{\ell_2}, \tag{2.2.32}$$

where $S^* := \operatorname{supp} \mathbf{u}^*$. Then, for any $\gamma_0 < \gamma < 1$, the sequence $\big(\mathbf{u}^{(n)}\big)_{n \in \mathbb{N}}$ obtained by (2.1.9) fulfills

$$\big(\mathbf{u}^{(n)}\big)_{n \in \mathbb{N}} \subset \mathcal{B}_L \cap B(R) \tag{2.2.33}$$

and converges to $\mathbf{u}^*$ at a linear rate

$$\|\mathbf{u}^* - \mathbf{u}^{(n)}\|_{\ell_2} \leq \varepsilon^{(n)} := \gamma^n \|\mathbf{u}^*\|_{\ell_2}, \tag{2.2.34}$$

whenever the $\boldsymbol{\alpha}^{(n)} \geq \boldsymbol{\alpha}$ are chosen according to

$$\max_{\mu \in \mathcal{J}} |\alpha_\mu^{(n)} - \alpha_\mu| \leq \lambda L^{-\frac{1}{2}} (\gamma - \gamma_0) \varepsilon^{(n)}. \tag{2.2.35}$$

Moreover, the iteration is monotone in the sense that

$$\Gamma_{\boldsymbol{\alpha}^{(n+1)}}(\mathbf{u}^{(n+1)}) < \Gamma_{\boldsymbol{\alpha}^{(n)}}(\mathbf{u}^{(n)}),$$

provided that $\mathbf{u}^{(n)}$ is not yet a critical point of $\Gamma_{\boldsymbol{\alpha}^{(n)}}$.

Remark 2.2.5. 1. Before proving Theorem 2.2.4, let us comment the following fundamental implication: by the Lipschitz condition (2.2.27) and local contraction property (2.2.32), the iterations $\big(\mathbf{u}^{(n)}\big)_{n \in \mathbb{N}}$ of the algorithm starting from $\mathbf{u}^{(0)} = \mathbf{0}$ must converge to any stationary point $\mathbf{u}^* \in B(R)$, hence implying automatically its uniqueness. In fact, if there was another stationary point, it would also coincide with the limit of this sequence. In particular, the global minimizer $\mathbf{u}^\circ$ of $\Gamma_{\boldsymbol{\alpha}}$ necessarily lies in the ball $B(R)$ and is a stationary point of $\Gamma_{\boldsymbol{\alpha}}$, and we have linear convergence of the iterates to $\mathbf{u}^\circ$.

2. The contractivity condition (2.2.32) is partially related to the convexity of $\Gamma_{\boldsymbol{\alpha}}$ in the vicinity of a stationary point $\mathbf{u}^*$. However, in general the condition (2.2.32) does not imply strict convexity of $\Gamma_{\boldsymbol{\alpha}}$ in a neighborhood of $\mathbf{u}^*$. We will discuss this relationship for simplicity in a finite-dimensional setting.

   On the one hand, if $\#\mathcal{J} = L < \infty$, (2.2.32) indeed implies the convexity of $\Gamma_{\boldsymbol{\alpha}}$ near $\mathbf{u}^*$. In this situation, (2.2.32) simplifies to

   $$\big\|T(\mathbf{u}^*) - T(\mathbf{v})\big\|_{\ell_2} \leq \gamma_0 \|\mathbf{u}^* - \mathbf{v}\|_{\ell_2}, \quad \text{for all } \mathbf{v} \in \ell_2,$$

   which entails $\|T'(\mathbf{u}^*)\|_{\mathcal{L}(\ell_2)} \leq \gamma_0 < 1$. If, additionally, $\Gamma_{\boldsymbol{\alpha}}$ is smooth in the vicinity of $\mathbf{u}^*$, the local convexity of $\Gamma_{\boldsymbol{\alpha}}$ follows from the monotonicity of its gradient

   $$\nabla\Gamma_{\boldsymbol{\alpha}}(\mathbf{u}) = 2\lambda(\mathbf{u} - T(\mathbf{u})) + 2\,\mathrm{sign}(\mathbf{u}^*)\boldsymbol{\alpha},$$

   because we have

   $$\big\langle \nabla\Gamma_{\boldsymbol{\alpha}}(\mathbf{u}) - \nabla\Gamma_{\boldsymbol{\alpha}}(\mathbf{v}), \mathbf{u} - \mathbf{v}\big\rangle_{\ell_2} \geq 0$$

   for all $\mathbf{u}, \mathbf{v}$ from a sufficiently small neighborhood of $\mathbf{u}^*$.

   On the other hand, in the generic case that L is small with respect to the cardinality of $\mathcal{J}$, the Tikhonov functional $\Gamma_{\boldsymbol{\alpha}}$ might be nonconvex in *each* neighborhood of $\mathbf{u}^*$ despite the validity of (2.2.32). As an example, let $\alpha_1, \alpha_2 > 0$ and $K : \mathbb{R}^2 \to \mathbb{R}^2$ be the smooth mapping

   $$K(\mathbf{u}) := \begin{pmatrix} u_1 - \alpha_1 - 1 \\ \big(\frac{1}{2} + \frac{1}{\pi}\arctan(2\alpha_2\pi u_2)\big)^{1/2} \end{pmatrix}, \quad \text{for all } \mathbf{u} \in \mathbb{R}^2.$$

   Further let $\mathbf{y} = \mathbf{0}$. By definition, $\Gamma_{\boldsymbol{\alpha}}$ separates into a sum of one-dimensional functionals,

   $$\Gamma_{\boldsymbol{\alpha}}(\mathbf{u}) = J_1(u_1) + J_2(u_2),$$

   with

   $$\begin{aligned} J_1(u_1) &:= (u_1 - \alpha_1 - 1)^2 + 2\alpha_1|u_1|, \\ J_2(u_2) &:= \tfrac{1}{2} + \tfrac{1}{\pi}\arctan(2\alpha_2\pi u_2) + 2\alpha_2|u_2|. \end{aligned}$$

   The functional J_1 is strictly convex, with unique minimizer $u_1^* = S_{\alpha_1}(1+\alpha_1) = 1$, whereas J_2 is nonconvex in each open neighborhood of its unique minimizer $u_2^* = 0$, see also Figure 2.1. Therefore, $\mathbf{u}^* = \binom{1}{0}$ is the unique minimizer of $\Gamma_{\boldsymbol{\alpha}}$, but $\Gamma_{\boldsymbol{\alpha}}$ is nonconvex in each open neighborhood of $\mathbf{u}^*$. However, if $\lambda \geq 1$, (2.2.32) holds for all $L < \frac{1}{2}$ and $\gamma_0 := 1 - \frac{1}{\lambda} \in [0, 1)$, because

   $$K'(\mathbf{u}) = \begin{pmatrix} 1 & 0 \\ 0 & \frac{\alpha_2}{(4\alpha_2^2\pi^2u_2^2+1)\sqrt{\frac{1}{2}+\frac{1}{\pi}\arctan(2\alpha_2\pi u_2)}} \end{pmatrix}$$

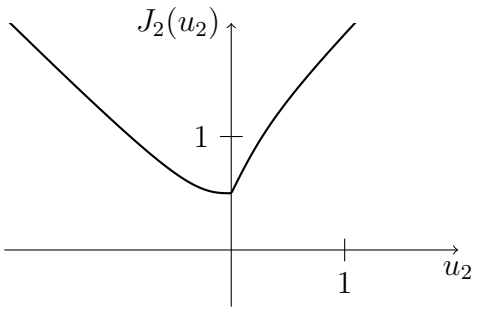


Figure 2.1: The functional J_2 from Remark 2.2.5, plotted for $\alpha_2 = 0.5$.

is diagonal. Therefore, the affinity of $K(\cdot)_1$ yields

$$\big(T(\mathbf{u}) - T(\mathbf{v})\big)_1 = \big(1 - \tfrac{1}{\lambda}\big)(u_1 - v_1), \quad \text{for all } \mathbf{u}, \mathbf{v} \in \mathbb{R}^2,$$

and hence

$$\big\|\big(T(\mathbf{u}) - T(\mathbf{v})\big)_{|\{1\}}\|_{\ell_2(\{1\})} \leq \big(1 - \tfrac{1}{\lambda}\big)\|\mathbf{u} - \mathbf{v}\|_{\ell_2}, \quad \text{for all } \mathbf{u}, \mathbf{v} \in \mathbb{R}^2. \tag{2.2.36}$$

Choosing $\mathbf{u} = \mathbf{u}^*$ and $\mathbf{v} = \mathbf{0}$ in (2.2.36) yields (2.2.32).

Let us now address the proof of Theorem 2.2.4.

Proof. The proof is performed by induction over n. There is nothing to show for $n = 0$. The first step is to prove that $\mathbf{u}^{(n+1)}$ is indeed contained in $\mathcal{B}_L$. Let $\mathbf{u}^{(n)} \in \mathcal{B}_L \cap B(R)$, then since $\boldsymbol{\alpha}^{(n)}$ is decreasing to $\boldsymbol{\alpha}$ it holds that

$$\operatorname{supp} \mathbf{u}^{(n+1)} = \operatorname{supp} \mathbb{S}_{\frac{1}{\lambda}\boldsymbol{\alpha}^{(n)}}\big(T(\mathbf{u}^{(n)})\big) \subset \operatorname{supp} \mathbb{S}_{\frac{1}{\lambda}\boldsymbol{\alpha}}\big(T(\mathbf{u}^{(n)})\big). \tag{2.2.37}$$

By using (2.2.27) for $\mathbf{v} = \mathbf{u}^{(n)}$, Lemma 2.2.2 tells us that

$$\begin{aligned} &\#\operatorname{supp} \mathbb{S}_{\frac{1}{\lambda}\boldsymbol{\alpha}}\big(T(\mathbf{u}^{(n)})\big) \leq \Lambda_{\frac{\alpha}{\lambda}}\big(T(\mathbf{u}^{(n)})\big) \\ &\quad \leq \frac{4(C_T^{Lip})^2\|\mathbf{u}^* - \mathbf{u}^{(n)}\|_{\ell_2}^2\lambda^2}{\alpha^2} + 4C|T(\mathbf{u}^*)|_{\ell_\tau^w}^\tau\Big(\frac{\alpha}{\lambda}\Big)^{-\tau} \leq L. \end{aligned} \tag{2.2.38}$$

We conclude that $\#\operatorname{supp} \mathbf{u}^{(n+1)} \leq L$. Let us denote $S^{(n)} = \operatorname{supp} \mathbf{u}^{(n)}$, $S^* = \operatorname{supp} \mathbf{u}^*$, and $\Lambda^{(n)} = S^* \cup S^{(n)} \cup S^{(n+1)}$. Notice that $\#S^{(n)} \cup S^{(n+1)} \leq 2L$. By the thresholding properties it is clear that after restriction to $\Lambda^{(n)}$

$$\mathbf{u}^*_{|\Lambda^{(n)}} = \mathbb{S}_{\frac{1}{\lambda}\boldsymbol{\alpha}}(T(\mathbf{u}^*)_{|\Lambda^{(n)}}), \tag{2.2.39}$$

and

$$\mathbf{u}^{(n+1)}_{|\Lambda^{(n)}} = \mathbb{S}_{\frac{1}{\lambda}\boldsymbol{\alpha}^{(n)}}(T(\mathbf{u}^{(n)})_{|\Lambda^{(n)}}) \tag{2.2.40}$$

hold. This, together with the nonexpansiveness of soft thresholding, Lemma 2.2.1 in the second inequality, and (2.2.32) together with $\#\operatorname{supp}\mathbf{u}^{(n+1)} \leq L$ in the third inequality yields

$$\begin{aligned}
&\|\mathbf{u}^* - \mathbf{u}^{(n+1)}\|_{\ell_2} \\
&\quad = \|\mathbf{u}^*_{|\Lambda^{(n)}} - \mathbf{u}^{(n+1)}_{|\Lambda^{(n)}}\|_{\ell_2(\Lambda^{(n)})} \\
&\quad = \|\mathbb{S}_{\frac{1}{\lambda}\boldsymbol{\alpha}}(T(\mathbf{u}^*)_{|\Lambda^{(n)}}) - \mathbb{S}_{\frac{1}{\lambda}\boldsymbol{\alpha}^{(n)}}(T(\mathbf{u}^{(n)})_{|\Lambda^{(n)}})\|_{\ell_2(\Lambda^{(n)})} \\
&\quad \leq \|\mathbb{S}_{\frac{1}{\lambda}\boldsymbol{\alpha}}(T(\mathbf{u}^*)_{|\Lambda^{(n)}}) - \mathbb{S}_{\frac{1}{\lambda}\boldsymbol{\alpha}}(T(\mathbf{u}^{(n)})_{|\Lambda^{(n)}})\|_{\ell_2(\Lambda^{(n)})} \\
&\qquad + \|\mathbb{S}_{\frac{1}{\lambda}\boldsymbol{\alpha}}(T(\mathbf{u}^{(n)})_{|\Lambda^{(n)}}) - \mathbb{S}_{\frac{1}{\lambda}\boldsymbol{\alpha}^{(n)}}(T(\mathbf{u}^{(n)})_{|\Lambda^{(n)}})\|_{\ell_2(\Lambda^{(n)})} \\
&\quad \leq \|T(\mathbf{u}^*)_{|\Lambda^{(n)}} - T(\mathbf{u}^{(n)})_{|\Lambda^{(n)}}\|_{\ell_2(\Lambda^{(n)})} \\
&\qquad + \frac{\big(\#\Lambda_{\frac{\alpha}{\lambda}}(T(\mathbf{u}^{(n)}))\big)^{1/2}}{\lambda}\big(\max_{\mu\in\Lambda_{\frac{\alpha}{\lambda}}(T(\mathbf{u}^{(n)}))}|\alpha_\mu - \alpha_\mu^{(n)}|\big) \\
&\quad \leq \gamma_0\|\mathbf{u}^* - \mathbf{u}^{(n)}\|_{\ell_2} + \frac{L^{1/2}}{\lambda}\big(\max_{\mu\in\Lambda_\alpha(T(\mathbf{u}^{(n)}))}|\alpha_\mu - \alpha_\mu^{(n)}|\big) \\
&\quad \leq \gamma_0\varepsilon^{(n)} + (\gamma - \gamma_0)\varepsilon^{(n)} = \gamma\varepsilon^{(n)} = \varepsilon^{(n+1)}.
\end{aligned}$$

The last inequality is a consequence of induction hypothesis and (2.2.35). This proves $\mathbf{u}^{(n+1)} \in \mathcal{B}_L$. Obviously $\mathbf{u}^{(n+1)} \in B(R)$ because of the monotonicity of the iterations:

$$\Gamma_{\boldsymbol{\alpha}}(\mathbf{u}^{(n+1)}) \leq \Gamma_{\boldsymbol{\alpha}^{(n+1)}}(\mathbf{u}^{(n+1)}) \leq \Gamma_{\boldsymbol{\alpha}^{(n)}}(\mathbf{u}^{(n+1)}) < \Gamma_{\boldsymbol{\alpha}^{(n)}}(\mathbf{u}^{(n)}) \leq \Gamma_{\boldsymbol{\alpha}^{(0)}}(\mathbf{0}). \qquad \square$$

2.2.2 Nonlinear operators with bounded second derivatives

In this section we state smoothness conditions on the nonlinear operator K which imply that the operator T defined in (2.2.8) fulfills (2.2.27) and (2.2.32). In the following we assume that S^* is the support of a global minimizer $\mathbf{u}^*$ of $\Gamma_{\boldsymbol{\alpha}}$ in $B(R)$. As discussed above, once we prove that T fulfills (2.2.27) and (2.2.32), then by Theorem 2.2.4 we automatically have that $\mathbf{u}^*$ is actually the unique stationary point of $\Gamma_{\boldsymbol{\alpha}}$ in $B(R)$.

Theorem 2.2.6. *Let the data fulfill assumption* (2.2.7)*. Assume that K is twice continuously differentiable on an open set that contains $\mathcal{B}$ and, together with its derivative K', is bounded on $\mathcal{B}$. Furthermore, assume that there exist $0 < \gamma_2 \leq \gamma_1 < 1$ such that for all $\Lambda \subset \mathcal{J}, \#\Lambda \leq 2L$ and $\boldsymbol{\zeta} \in \mathcal{B}, \operatorname{supp}\boldsymbol{\zeta} \subset S^* \cup \Lambda$, the following local contraction property*

$$\|\Big(\operatorname{Id} - \frac{1}{\lambda}\big(K'(\boldsymbol{\zeta})\big)^* K'(\boldsymbol{\zeta})\Big)_{|S^*\cup\Lambda\times S^*\cup\Lambda}\|_{\mathcal{L}(\ell_2(S^*\cup\Lambda),\ell_2(S^*\cup\Lambda))} \leq \gamma_2 \tag{2.2.41}$$

holds. Moreover, let us assume that the uniform spectral gap condition

$$\|\Big(\frac{1}{\lambda}\big(K''(\boldsymbol{\zeta})(\cdot)\big)^*\big(y-K(\boldsymbol{\zeta})\big)\Big)_{|S^*\cup\Lambda\times S^*\cup\Lambda}\|_{\mathcal{L}(\ell_2(S^*\cup\Lambda),\ell_2(S^*\cup\Lambda))}\leq\gamma_1-\gamma_2 \tag{2.2.42}$$

holds. Then T defined in (2.2.8) *fulfills* (2.2.27) *and* (2.2.32).

Proof. The proof is an application of the mean value theorem. In order to compute the derivative of T we introduce the auxiliary operator

$$G:(\mathbf{u},\mathbf{v})\mapsto\mathbf{u}+\frac{1}{\lambda}\big(K'(\mathbf{u})\big)^*\big(y-K(\mathbf{v})\big),\quad(\mathbf{u},\mathbf{v})\in\ell_2\times\ell_2$$

and observe

$$T=G\circ(\mathrm{Id},\mathrm{Id})^\intercal.$$

We compute

$$\begin{aligned}T'(\boldsymbol{\zeta})\,\mathbf{z}&=\Big(\frac{\partial G}{\partial\mathbf{u}},\frac{\partial G}{\partial\mathbf{v}}\Big)\big((\boldsymbol{\zeta},\boldsymbol{\zeta})\big)\circ\big(\,\mathrm{Id},\mathrm{Id}\,\big)^\intercal\,\mathbf{z}\\&=\Big(\,\mathrm{Id}+\frac{1}{\lambda}\big(K''(\boldsymbol{\zeta})(\cdot)\big)^*\big(y-K(\boldsymbol{\zeta})\big),-\frac{1}{\lambda}\big(K'(\boldsymbol{\zeta})\big)^*K'(\boldsymbol{\zeta})(\cdot)\Big)\circ\big(\,\mathrm{Id},\mathrm{Id}\,\big)^\intercal\,\mathbf{z}\\&=\mathbf{z}+\frac{1}{\lambda}\big(K''(\boldsymbol{\zeta})\,\mathbf{z}\big)^*\big(y-K(\boldsymbol{\zeta})\big)-\frac{1}{\lambda}\big(K'(\boldsymbol{\zeta})\big)^*K'(\boldsymbol{\zeta})\,\mathbf{z}.\end{aligned} \tag{2.2.43}$$

Observe that $K:\ell_2\to Y$, $K':\ell_2\to\mathcal{L}(\ell_2,Y)$, and $K'':\ell_2\to\mathcal{L}(\ell_2,\mathcal{L}(\ell_2,Y))$. Therefore $K''(\boldsymbol{\zeta})\,\mathbf{z}\in\mathcal{L}(\ell_2,Y)$ holds. Consequently $\big(K''(\boldsymbol{\zeta})\,\mathbf{z}\big)^*\in\mathcal{L}(Y,\ell_2)$, so that the composition in (2.2.43) is well defined. By our assumptions K,K', and K'' are bounded on the bounded set $\mathcal{B}$. This, together with (2.2.7) implies $\sup_{\xi\in\mathcal{B}}\|T'(\xi)\|_{\mathcal{L}(\ell_2,\ell_2)}<\infty$. Since $\mathcal{B}$ is convex we can use the mean value theorem to conclude the Lipschitz property (2.2.27). In order to prove (2.2.32), let $\mathbf{v}\in\mathcal{B}_L$ and $\mathrm{supp}(\mathbf{v})\subset\Lambda\subset\mathcal{J}$ with $\#\Lambda\leq 2L$. Then, by (2.2.43), (2.2.41), and (2.2.42) the estimate

$$\begin{aligned}&\|\big(T'(\boldsymbol{\zeta})\big)_{|S^*\cup\Lambda\times S^*\cup\Lambda}\|_{\mathcal{L}(\ell_2(S^*\cup\Lambda),\ell_2(S^*\cup\Lambda))}\\&\quad\leq\|\Big(\,\mathrm{Id}-\frac{1}{\lambda}\big(K'(\boldsymbol{\zeta})\big)^*K'(\boldsymbol{\zeta})\Big)_{|S^*\cup\Lambda\times S^*\cup\Lambda}\|_{\mathcal{L}(\ell_2(S^*\cup\Lambda),\ell_2(S^*\cup\Lambda))}\\&\quad+\|\Big(\frac{1}{\lambda}\big(K''(\boldsymbol{\zeta})(\cdot)\big)^*\big(y-K(\boldsymbol{\zeta})\big)\Big)_{|S^*\cup\Lambda\times S^*\cup\Lambda}\|_{\mathcal{L}(\ell_2(S^*\cup\Lambda),\ell_2(S^*\cup\Lambda))}\\&\quad\leq\gamma_1\end{aligned}$$

holds. The restriction $\mathcal{B}_{|S^*\cup\Lambda}:=\{\mathbf{u}_{|S^*\cup\Lambda},\mathbf{u}\in\mathcal{B}\}$ of $\mathcal{B}$ onto the index set $S^*\cup\Lambda$ is a convex set in $\ell_2(S^*\cup\Lambda)$. Hence, we can apply the mean value theorem again to finalize the proof as follows:

$$\begin{aligned}&\|\big(T(\mathbf{u}^*)-T(\mathbf{v})\big)_{|S^*\cup\Lambda}\|_{\ell_2(S^*\cup\Lambda)}\\&\leq\sup_{\boldsymbol{\zeta}\in\mathcal{B}_{|S^*\cup\Lambda}}\|\big(T'(\boldsymbol{\zeta})\big)_{|S^*\cup\Lambda\times S^*\cup\Lambda}\|_{\mathcal{L}(\ell_2(S^*\cup\Lambda),\ell_2(S^*\cup\Lambda))}\|\mathbf{u}^*-\mathbf{v}\|_{\ell_2}\\&\leq\gamma_1\|\mathbf{u}^*-\mathbf{v}\|_{\ell_2}.\end{aligned}$$

□

2.2.3 Nonlinear perturbation of linear operators

In this section we discuss the validity of (2.2.32) for the special case that K is given by the sum of a linear operator and a nonlinear perturbation. To be specific, we consider

$$K_\sigma = A + \sigma N, \tag{2.2.44}$$

where $\sigma \in \mathbb{R}_+$, $A \in \mathcal{L}(\ell_2, Y)$ and $N : \ell_2 \to Y$ is a nonlinear perturbation with Fréchet derivative $N' : \ell_2 \to \mathcal{L}(\ell_2, Y)$ and the property that

$$N,\ N' \text{ are Lipschitz continuous on closed bounded sets.} \tag{2.2.45}$$

Similarly to (2.2.12) and (2.2.14) we denote the respective Lipschitz constants and suprema on $B(R)$ with $C_N^{Lip}(R)$, $C_{N'}^{Lip}(R)$, $C_N^{bnd}(R)$, and $C_{N'}^{bnd}(R)$.

We begin by deriving uniform bounds for those constants. We denote

$$R(\sigma) = \sup\{\|\mathbf{v}\|_{\ell_2},\ \Gamma_{\boldsymbol{\alpha},\sigma}(\mathbf{v}) \le \Gamma_{\boldsymbol{\alpha}^{(0)},\sigma}(\mathbf{0})\},$$

where $\Gamma_{\boldsymbol{\alpha},\sigma}$ is the functional $\Gamma_{\boldsymbol{\alpha}}$ for $K = K_\sigma$ depending on σ. Accordingly we denote

$$R'(\sigma) := R(\sigma) + \frac{1}{\lambda_0} C_{K'}^{bnd}(R(\sigma))(C_K^{bnd}(R(\sigma)) + C_Y).$$

We denote with $\mathcal{C}(\sigma)$ the set of critical points of $\Gamma_{\boldsymbol{\alpha},\sigma}$ in $B(R(\sigma))$. For any $\mathbf{u}^*(\sigma) \in \mathcal{C}(\sigma)$ we denote

$$L(\mathbf{u}^*(\sigma)) := \frac{4(C_T^{Lip})^2 \|\mathbf{u}^*(\sigma)\|_{\ell_2}^2 \lambda^2}{\alpha^2} + 4C|T(\mathbf{u}^*(\sigma))|_{\ell_\tau^w}^\tau \left(\frac{\alpha}{\lambda}\right)^{-\tau}. \tag{2.2.46}$$

Lemma 2.2.7. *Let the data fulfill assumption* (2.2.7)*. Further let* $\sigma_0 \in \mathbb{R}_+$ *and* K_σ*,* $\sigma \in [0, \sigma_0]$*, be of the form* (2.2.44)*. Suppose that the assumptions* (2.2.45) *hold. Then it holds that*

$$\begin{aligned} R_0 &:= \sup_{\sigma\in[0,\sigma_0]} R(\sigma) < \infty, \\ R_0' &:= \sup_{\sigma\in[0,\sigma_0]} R'(\sigma) < \infty, \\ &\sup_{\sigma\in[0,\sigma_0]} C_{K_\sigma}^{Lip}(R'(\sigma)) < \infty, \\ &\sup_{\sigma\in[0,\sigma_0]} C_{K_\sigma'}^{Lip}(R'(\sigma)) < \infty. \end{aligned} \tag{2.2.47}$$

Under the additional assumption

$$\sup_{\sigma\in[0,\sigma_0]} \sup_{\mathbf{u}^*(\sigma)\in\mathcal{C}(\sigma)} |T_\sigma(\mathbf{u}^*(\sigma))|_{\ell_\tau^w} < \infty \tag{2.2.48}$$

it further holds that

$$L_0 := \sup_{\sigma\in[0,\sigma_0]} \sup_{\mathbf{u}^*(\sigma)\in\mathcal{C}(\sigma)} L(\mathbf{u}^*(\sigma)) < \infty. \tag{2.2.49}$$

Remark 2.2.8. Before proving this lemma, let us comment on the condition (2.2.48), requiring to consider a supremum over the set $\mathcal{C}(\sigma)$, which a priori can be very large. As we will show later, the boundedness of the quantities of this lemma and an additional spectral property on the operator A, the so-called *restricted isometry property*, see formula (2.2.51) below, will imply the operator T to fulfill (2.2.32). As already stated above, under this condition, the set $\mathcal{C}(\sigma)$ consists only of one point, i.e., the global minimizer of $\Gamma_{\boldsymbol{\alpha},\sigma}$. Hence the condition (2.2.48) will turn out to be much less restrictive as it seems at a first glance.

Let us now prove Lemma 2.2.7.

Proof. It is immediate to see that (2.2.11) implies

$$R(\sigma) \leq \frac{\sup_{\sigma' \in [0,\sigma_0]} \left(\Gamma_{\boldsymbol{\alpha}^{(0)},\sigma'}(\mathbf{0})\right)}{2\alpha}.$$

Furthermore the term $\sup_{\sigma' \in [0,\sigma_0]} \left(\Gamma_{\boldsymbol{\alpha}^{(0)},\sigma'}(\mathbf{0})\right)$ is finite, as

$$\sigma' \mapsto \Gamma_{\boldsymbol{\alpha}^{(0)},\sigma'}(\mathbf{0}) = \|\sigma' N(\mathbf{0}) - y\|_Y^2$$

is continuous and bounded on $[0, \sigma_0]$, because of the assumptions (2.2.7) and (2.2.45). Hence, we conclude the boundedness of R_0 in (2.2.47). By the assumption (2.2.45) we may bound the Lipschitz constant of K_σ on $B(R_\sigma)$ as follows:

$$C^{Lip}_{K_\sigma}(R(\sigma)) \leq \|A\|_{\mathcal{L}(\ell_2,Y)} + \sigma_0 C^{Lip}_N(R_0). \tag{2.2.50}$$

The constant $C^{Lip}_{K'_\sigma}(R(\sigma))$ may be bounded analogously. By the same reasoning as in (2.2.13) it follows that the two constants $C^{bnd}_{K_\sigma}(R(\sigma))$ and $C^{bnd}_{K'_\sigma}(R(\sigma))$ may be bounded independently of $\sigma \in [0, \sigma_0]$. This proves the existence of uniform bounds for the constants $C^{bnd}_{K_\sigma}(R(\sigma))$ and $C^{bnd}_{K'_\sigma}(R(\sigma))$, and consequently of R'_0 in (2.2.47). Using assumption (2.2.45) on $B(R'_0)$ allows us to estimate similarly to (2.2.50) uniform bounds for the constants $C^{Lip}_{K_\sigma}(R'(\sigma))$ and $C^{Lip}_{K'_\sigma}(R'(\sigma))$. It remains to prove the finiteness of L_0 in (2.2.49). To this end observe that the Lipschitz property of K_σ and K'_σ on $B(R_0)$ imply that T_σ, defined in (2.2.8), is Lipschitz on $B(R_0)$ and that the corresponding Lipschitz constants may be uniformly bounded in σ. The remaining terms in (2.2.46) are bounded uniformly in σ by assumption (2.2.48) and the estimate $\|\mathbf{u}^*(\sigma)\|_{\ell_2} \leq R(\sigma) \leq R_0$. □

We are now able to state conditions under which the fundamental contraction property (2.2.32) of the operators T_σ, defined by (2.2.8) can be ensured uniformly in σ for σ_0 sufficiently small.

Lemma 2.2.9. *Let the assumptions of Lemma 2.2.7 hold. Fix $\sigma_0 \in \mathbb{R}^+$. For all $\sigma \in [0, \sigma_0]$, we fix $\mathbf{u}^*(\sigma) \in \mathcal{C}(\sigma)$ and denote $S^*_\sigma := \operatorname{supp} \mathbf{u}^*(\sigma)$. We make the assumption that the linear part A of K_σ fulfills the restricted isometry property*

$$\|(\mathrm{Id} - \lambda^{-1} A^* A)_{|\Lambda^\circ \times \Lambda^\circ}\|_{\mathcal{L}(\ell_2(\Lambda^\circ), \ell_2(\Lambda^\circ))} \leq \gamma_1 < 1, \tag{2.2.51}$$

for all $\Lambda^\circ \subset \mathcal{J}$ with $\#\Lambda^\circ \leq 3L_0$. By using the notations as in (2.2.47)*, the constant*

$$\begin{aligned} C :=& \|A\|_{\mathcal{L}(\ell_2,Y)} C_N^{Lip}(R_0) \\ &+ C_{N'}^{Lip}(R_0)\big(\|A\|_{\mathcal{L}(\ell_2,Y)} R_0 + \sigma_0 C_N^{bnd}(R_0) + C_Y\big) \\ &+ C_{N'}^{bnd}(R_0)\big(\|A\|_{\mathcal{L}(\ell_2,Y)} + \sigma_0 C_N^{Lip}(R_0)\big) \end{aligned} \tag{2.2.52}$$

is bounded and for all σ with $0 \leq \sigma < \min\big(\sigma_0, (1-\gamma_1)\lambda C^{-1}\big)$ the following holds: For all $\mathbf{v} \in B(R_0)$ with $\#\operatorname{supp}\mathbf{v} \leq L_0$, and $\operatorname{supp}(\mathbf{v}) \subset \Lambda \subset \mathcal{J}$ with $\#\Lambda \leq 2L_0$, the contraction property

$$\|\big(T_\sigma(\mathbf{u}^*(\sigma)) - T_\sigma(\mathbf{v})\big)_{|S_\sigma^* \cup \Lambda}\|_{\ell_2(S_\sigma^* \cup \Lambda)} \leq \gamma_0 \|\mathbf{u}^*(\sigma) - \mathbf{v}\|_{\ell_2}, \tag{2.2.53}$$

holds with $\gamma_0 := \gamma_1 + \sigma\lambda^{-1}C < 1$.

Proof. We begin by proving that the constant C in (2.2.52) is bounded. To this end we apply Lemma 2.2.7 and observe that the Lipschitz property of N and N' on $B(R_0)$ implies similarly to (2.2.13) that N and N' are also bounded on $B(R_0)$.

Fix $\sigma \in [0, \sigma_0]$ and let $\mathbf{v} \in B(R_0)$, $\#\operatorname{supp}\mathbf{v} \leq L_0$ and $\operatorname{supp}\mathbf{v} \subset \Lambda \subset \mathcal{J}$ with $\#\Lambda \leq 2L_0$ and denote $\Lambda^\circ := S_\sigma^* \cup \Lambda$.

We use the splitting

$$\begin{aligned} T_\sigma(\mathbf{v}) - T_\sigma(\mathbf{u}^*(\sigma)) =& \mathbf{v} - \mathbf{u}^*(\sigma) - \lambda^{-1}A^*A(\mathbf{v} - \mathbf{u}^*(\sigma)) - \sigma\lambda^{-1}A^*\big(N(\mathbf{v}) - N(\mathbf{u}^*(\sigma))\big) \\ &- \sigma\lambda^{-1}\big(N'(\mathbf{v}) - N'(\mathbf{u}^*(\sigma))\big)^*\big((A+\sigma N)(\mathbf{v}) - y\big) \\ &- \sigma\lambda^{-1}\big(N'(\mathbf{u}^*(\sigma))\big)^*\big((A+\sigma N)(\mathbf{v}) - (A+\sigma N)(\mathbf{u}^*(\sigma))\big) \end{aligned}$$

together with the assumption (2.2.51) to estimate

$$\begin{aligned} &\|\big(T_\sigma(\mathbf{v}) - T_\sigma(\mathbf{u}^*(\sigma))\big)_{|\Lambda^\circ}\|_{\ell_2(\Lambda^\circ)} \\ &\leq \|\big((\mathrm{Id} - \lambda^{-1}A^*A)(\mathbf{v} - \mathbf{u}^*(\sigma))\big)_{|\Lambda^\circ}\|_{\ell_2(\Lambda^\circ)} + \sigma\lambda^{-1}\Big(\|\big(A^*\big(N(\mathbf{v}) - N(\mathbf{u}^*(\sigma))\big)\big)_{|\Lambda^\circ}\|_{\ell_2(\Lambda^\circ)} \\ &\quad + \|\big(\big(N'(\mathbf{v}) - N'(\mathbf{u}^*(\sigma))\big)^*\big((A+\sigma N)(\mathbf{v}) - y\big)\big)_{|\Lambda^\circ}\|_{\ell_2(\Lambda^\circ)} \\ &\quad + \|\big(\big(N'(\mathbf{u}^*(\sigma))\big)^*\big((A+\sigma N)(\mathbf{v}) - (A+\sigma N)(\mathbf{u}^*(\sigma))\big)\big)_{|\Lambda^\circ}\|_{\ell_2(\Lambda^\circ)}\Big) \\ &\leq \Big(\gamma_1 + \sigma\lambda^{-1}\big(\|A\|_{\mathcal{L}(\ell_2,Y)} C_N^{Lip}(R_0) \\ &\quad + C_{N'}^{Lip}(R_0)\big(\|A\|_{\mathcal{L}(\ell_2,Y)} R_0 + \sigma_0 C_N^{bnd}(R_0) + C_Y\big) \\ &\quad + C_{N'}^{bnd}(R_0)\big(\|A\|_{\mathcal{L}(\ell_2,Y)} + \sigma_0 C_N^{Lip}(R_0)\big)\big)\Big)\|\mathbf{v} - \mathbf{u}^*(\sigma)\|_{\ell_2} \\ &\quad = (\gamma_1 + \sigma\lambda^{-1}C)\|\mathbf{v} - \mathbf{u}^*(\sigma)\|_{\ell_2}, \end{aligned}$$

which implies that the contraction property (2.2.53) holds. □

The last lemma established the contraction property (2.2.53) uniformly in σ. Therefore, for the current choice of $K = K_\sigma$ as in (2.2.44), we are able to apply directly Theorem 2.2.4. Let us summarize the result as follows.

Theorem 2.2.10. *Let the assumptions of Lemma 2.2.9 hold for some λ with*

$$\lambda > \max\left(\lambda_0, C^{Lip}_{K'_\sigma}(R'_0)(C^{bnd}_{K_\sigma}(R'_0) + C_Y) + C^{bnd}_{K'_\sigma}(R'_0)C^{Lip}_{K_\sigma}(R'_0)\right).$$

Then, for all

$$0 \leq \sigma \leq \min\left(\sigma_0, (1-\gamma_1)\lambda C^{-1}\right)$$

and $\gamma_0 < \gamma < 1$, if we choose $\left(\boldsymbol{\alpha}^{(n)}\right)_{n\in\mathbb{N}}$ according to (2.2.35) the sequence $\left(\mathbf{u}^{(n)}(\sigma)\right)_{n\in\mathbb{N}}$ defined by (2.1.9) with initial guess $\mathbf{u}^{(0)} = \mathbf{0}$ satisfies

$$\left(\mathbf{u}^{(n)}(\sigma)\right)_{n\in\mathbb{N}} \subset B(R_0). \tag{2.2.54}$$

Furthermore, $\mathbf{u}^{(n)}(\sigma)$ converges to $\mathbf{u}^*(\sigma) \in \mathcal{C}(\sigma)$ at a linear rate, i.e.,

$$\|\mathbf{u}^*(\sigma) - \mathbf{u}^{(n)}(\sigma)\|_{\ell_2} \leq \gamma^n \|\mathbf{u}^*(\sigma) - \mathbf{u}^{(0)}(\sigma)\|_{\ell_2}, \tag{2.2.55}$$

and moreover

$$\Gamma_{\boldsymbol{\alpha}^{(n+1)},\sigma}(\mathbf{u}^{(n+1)}(\sigma)) < \Gamma_{\boldsymbol{\alpha}^{(n)},\sigma}(\mathbf{u}^{(n)}(\sigma)),$$

provided that $\mathbf{u}^{(n)}(\sigma)$ is not yet a critical point of $\Gamma_{\boldsymbol{\alpha}^{(n)},\sigma}$. In particular $\mathbf{u}^*(\sigma) \in \mathcal{C}(\sigma)$ has to be the only critical point of $\Gamma_{\boldsymbol{\alpha},\sigma}$ in $B(R_0)$ with $\#\operatorname{supp}\mathbf{u}^*(\sigma) \leq L_0$, actually it is its unique global minimizer in $B(R_0)$.

2.3 Preconditioning

The convergence analysis in Section 2.2 for the iteration (2.1.9) relies on the contraction property (2.2.32) of the operator T defined in (2.2.8). This property also ensures that, despite the fact that $\Gamma_{\boldsymbol{\alpha}}$ is a nonconvex functional, it has nevertheless a unique global minimizer in a prescribed ball centered at $\mathbf{0}$ and that the iteration (2.1.9) is guaranteed to converge to it with linear rate. Unfortunately, we can not expect this powerful property to hold in general, even in the case that the underlying operator K is linear and compact. Therefore, in this section, we present how preconditioning can be applied to promote property (2.2.32) in the case that K is a nonlinear perturbation of a linear operator. We have to imagine the action of this preconditioning as a sort of "stretching" of the functional $\Gamma_{\boldsymbol{\alpha}}$, so that no local minimizers or stationary points remain around $\mathbf{0}$ other than a unique global minimizer. Preconditioning also changes the topology of the minimization problem related to (2.1.1). Therefore, in Section 2.3.1, we begin by discussing the related topological issues. In Section 2.3.2 we present a preconditioning strategy and state conditions under which the restricted isometry property (2.2.51) will be satisfied. Finally in Section 2.3.3 we apply our findings to an interesting class of operators.

2.3.1 General setting

We shall consider the following modified functional

$$\big(\Gamma_{\boldsymbol{\alpha}} \circ D^{-1}\big)(\mathbf{z}) = \|(K \circ D^{-1})(\mathbf{z}) - y\|_Y^2 + 2\|D^{-1}\mathbf{z}\|_{\ell_{1,\boldsymbol{\alpha}}}, \quad \mathbf{z} \in \operatorname{Ran}(D), \tag{2.3.1}$$

where

$$D : \ell_2 \to \operatorname{Ran}(D)$$

is a suitable preconditioning matrix with well defined formal inverse $D^{-1} : \operatorname{Ran}(D) \to \ell_2$. Moreover we assume that D maps *finitely supported vectors* on *finitely supported vectors* and that

$$\|D^{-1}\mathbf{z}\|_{\ell_{1,\boldsymbol{\alpha}}} \sim \|\operatorname{diag}(D^{-1})\mathbf{z}\|_{\ell_{1,\boldsymbol{\alpha}}}, \tag{2.3.2}$$

which is the case, e.g., for block-diagonal matrices with invertible diagonal blocks.

Note, that preconditioning of the energy functional (2.1.1) changes the topology of the associated minimization problem. Moreover, the preconditioning operator D may be *unbounded* in the topology of ℓ_2. However, as we will see below, this is not an issue here. Indeed, Theorem 2.4.3, which will be proved later in Section 2.4, enables us to reduce the setting to a *finite-dimensional* one whenever needed, so that we can use the equivalence of norms on finite-dimensional vector spaces.

To this end we begin with the observation that any stationary point $\mathbf{u}^*$ of (2.1.1) can be characterized by having vanishing directional derivatives,

$$\mathbf{0} = \Gamma'_{\boldsymbol{\alpha}}(\mathbf{u}^*, \mathbf{d}) := \lim_{t \searrow 0} \frac{\Gamma_{\boldsymbol{\alpha}}(\mathbf{u}^* + t\,\mathbf{d}) - \Gamma_{\boldsymbol{\alpha}}(\mathbf{u}^*)}{t}, \quad \mathbf{d} \in \ell_2.$$

An analogous characterization holds for the stationary points $\mathbf{z}^*$ of (2.3.1). By the chain rule for directional derivatives, see [113, Proposition 3.6], we have

$$\mathbf{0} = \big(\Gamma_{\boldsymbol{\alpha}} \circ D^{-1}\big)'(\mathbf{z}^*, \mathbf{d}) = \Gamma'_{\boldsymbol{\alpha}}(D^{-1}\mathbf{z}^*, D^{-1}\mathbf{d}), \quad \mathbf{d} \in \ell_2.$$

In other words, there is a one-to-one relationship of the stationary points of (2.1.1) and (2.3.1). Moreover, by our assumptions on D, if $\mathbf{u}^*$ is a finitely supported stationary point of (2.1.1), the related stationary point $\mathbf{z}^* = D\mathbf{u}^*$ of (2.3.1) is also finitely supported.

We will use the assumption (2.3.2) to simplify the preconditioned energy functional $\Gamma_{\boldsymbol{\alpha}} \circ D^{-1}$. Indeed, motivated by the observation that $\|\operatorname{diag}(D^{-1})\mathbf{z}\|_{\ell_{1,\boldsymbol{\alpha}}} = \|\mathbf{z}\|_{\ell_{1,\operatorname{diag}(D^{-1})\boldsymbol{\alpha}}}$ and with a slight abuse of notation we will consider the modified energy functional

$$\Gamma^D_{\boldsymbol{\alpha}}(\mathbf{z}) := \|(K \circ D^{-1})(\mathbf{z}) - y\|_Y^2 + 2\|\mathbf{z}\|_{\ell_{1,\operatorname{diag}(D^{-1})\boldsymbol{\alpha}}}, \quad \mathbf{z} \in \operatorname{Ran}(D), \tag{2.3.3}$$

and the resulting minimization problem.

Remark 2.3.1. In the case of a nondiagonal weight matrix D, the modified energy functional $\Gamma^D_{\boldsymbol{\alpha}}$ will typically not have the same minimizer(s) as the original functional $\Gamma_{\boldsymbol{\alpha}} \circ D^{-1}$, and we do not assume that this is the case. However, this is not so much an issue in view of regularization problems. In the original topology of ℓ_2 the penalty term of $\Gamma^D_{\boldsymbol{\alpha}} \circ D$ is the weighted ℓ_1-norm

$$\|D\mathbf{u}\|_{\ell_{1,\operatorname{diag}(D^{-1})\boldsymbol{\alpha}}} = \|\operatorname{diag}(D^{-1})D\mathbf{u}\|_{\ell_{1,\boldsymbol{\alpha}}}.$$

By using (2.3.2) one can show that such penalty terms are proper, weakly lower semicontinuous and have bounded level sets. The regularization properties of such penalty terms have been studied in even more general settings [76].

We avoid to deal with the topology of $\operatorname{Ran}(D)$ in the following way. Let $\mathbf{z}^*$ be a fixed stationary point of the preconditioned energy functional (2.3.3) of finite support and $\Lambda_0 \subset \mathcal{J}$ an arbitrary finite set such that $\operatorname{supp} \mathbf{z}^* \subset \Lambda_0$. The restriction of (2.3.3) onto Λ_0 is then given by

$$\Gamma^D_{\boldsymbol{\alpha},\Lambda_0}(\mathbf{z}) := \|(K \circ D^{-1})_{|\Lambda_0}(\mathbf{z}) - y\|_Y^2 + 2\|\mathbf{z}\|_{\ell_{1,(\operatorname{diag}(D^{-1})\boldsymbol{\alpha})_{|\Lambda_0}}(\Lambda_0)}, \quad \mathbf{z} \in \mathbb{R}^{\Lambda_0}. \tag{2.3.4}$$

The minimization problem can now be considered in $\mathbb{R}^{\Lambda_0}$ endowed with the Euclidean norm. We denote the restriction of $\mathbf{z}^*$ onto Λ_0 by $\mathbf{z}^*_{|\Lambda_0}$ and by $E_{\Lambda_0} : \ell_2(\Lambda_0) \to \ell_2$ the trivial extension by 0. Then it follows by the chain rule for directional derivatives that

$$\begin{aligned}
\mathbf{0} &= \left(\Gamma^D_{\boldsymbol{\alpha}}\right)'(\mathbf{z}^*, E_{\Lambda_0}\mathbf{d}) \\
&= \left(\Gamma^D_{\boldsymbol{\alpha}}\right)'(E_{\Lambda_0}\mathbf{z}^*_{|\Lambda_0}, E_{\Lambda_0}\mathbf{d}) \\
&= \left(\Gamma^D_{\boldsymbol{\alpha},\Lambda_0}\right)'(\mathbf{z}^*_{|\Lambda_0}, \mathbf{d}), \quad \mathbf{d} \in \mathbb{R}^{\Lambda_0}.
\end{aligned}$$

Consequently $\mathbf{z}^*_{|\Lambda_0}$ is also a stationary point of the finite-dimensional energy functional (2.3.4). Unfortunately, the converse is not valid, because E_{Λ_0} does not have dense range and a stationary point for $\Gamma^D_{\boldsymbol{\alpha},\Lambda_0}$ does not necessarily correspond a priori to the restriction to a finite-dimensional set Λ_0 of a stationary point of $\Gamma^D_{\boldsymbol{\alpha}}$ in $\operatorname{Ran}(D)$.

Nevertheless, if one could assume that $\Gamma^D_{\boldsymbol{\alpha},\Lambda_0}$ has actually only one critical point in $\mathbb{R}^{\Lambda_0}$ for any choice of $\Lambda_0 \subset \mathcal{J}$ finite, then we can argue the uniqueness of the critical point of $\Gamma^D_{\boldsymbol{\alpha}}$ in $\operatorname{Ran}(D)$ as well. In fact, if there were two critical points $\mathbf{z}^*_1$ and $\mathbf{z}^*_2$ for $\Gamma^D_{\boldsymbol{\alpha}}$ in $\operatorname{Ran}(D)$, their support could be included in a finite set Λ_0' of indexes. Without loss of generality this set could be assumed to be a subset of Λ_0 for the latter large enough. Hence, the assumed uniqueness of the critical point in $\mathbb{R}^{\Lambda_0}$ for $\Gamma^D_{\boldsymbol{\alpha},\Lambda_0}$ would immediately imply that $(\mathbf{z}^*_1)_{|\Lambda_0} = (\mathbf{z}^*_2)_{|\Lambda_0}$ or, equivalently, that $\mathbf{z}^*_1 = \mathbf{z}^*_2$. In turn this means that, in the situation of a unique critical point in finite dimensions, the minimization of the finite-dimensional problem is actually equivalent to the minimization of the infinite-dimensional one.

In Section 2.4 we will present an implementable numerical scheme, which solves the finite-dimensional minimization problem related to (2.3.4). We shall also show that a priori knowledge of the set Λ_0 is not needed. In fact, it will be constructed on the fly by the presented adaptive scheme.

2.3.2 Multilevel preconditioning

In Section 2.2.3 we considered the case that K consists of a dominant linear part A and a nonlinear perturbation. In this setting, we were able to show that the contraction assumption (2.2.32) can be guaranteed if the linear part of the equation fulfills the restricted isometry property (2.2.51). In general this condition will fail to hold, even if A is a compact linear operator. Nevertheless, in this section, we show that this issue can be solved by a preconditioning strategy. To this end, we partly follow the lines of [36] and recall the corresponding results as far as they are needed for our purposes.

In the following we will assume that $\Omega \subset \mathbb{R}^d$ is a bounded Lipschitz domain and $\boldsymbol{\Psi} := \{\psi_\mu\}_{\mu\in\mathcal{J}}$ is a compactly supported wavelet basis or frame of wavelet type for $L_2(\Omega)$, see e.g. [25, Section 2.12]. Every $\mu \in \mathcal{J}$ is of the form $\mu = (j, k, e)$, where $j \in \mathbb{Z}$ is the *scale*, often denoted as $|\mu|$, $k \in \mathbb{Z}^d$ is the *spatial location* and e is the *type* of ψ_μ. We refer to [25, 26] for further details concerning this notation. We do not go into construction details concerning these bases or the alternative of wavelet frames. In fact, we simply assume the following properties for all $\mu \in \mathcal{J}$. Furthermore, for the ease of presentation, we formulate them for the case of an orthogonal wavelet basis on $\Omega = (0, 1)^d$:

- (W_1) The support $\Omega_\mu := \operatorname{supp}\psi_\mu$ fulfills $|\Omega_\mu| \sim 2^{-|\mu|d}$. Furthermore there exists a suitable cube Q, centered at the origin, such that, $\Omega_\mu \subset 2^{-|\mu|}k + 2^{-|\mu|}Q$, see [25, Section 2.12].
- (W_2) The basis has the cancellation property $\int_\Omega \xi^{\boldsymbol{\beta}}\psi_\mu(\xi)\,\mathrm{d}\xi = 0$, $|\boldsymbol{\beta}| = 0, \dots, d^* \in \mathbb{N}$.
- (W_3) $\|\psi_\mu\|_{L_\infty(\Omega)} \le C2^{d/2|\mu|}$.

Examples of wavelet bases satisfying these conditions can be found in [42]. In this setting the *synthesis map* related to $\boldsymbol{\Psi}$ reads as

$$\mathcal{F} : \ell_2 \to L_2(\Omega), \quad \mathcal{F}(\mathbf{u}) := \sum_{\mu\in\mathcal{J}} u_\mu\psi_\mu, \quad \mathbf{u} \in \ell_2. \tag{2.3.5}$$

Its adjoint is given by

$$\mathcal{F}^* : L_2(\Omega) \to \ell_2, \quad \mathcal{F}^*(u) := (\langle u, \psi_\mu\rangle_{L_2(\Omega)})_{\mu\in\mathcal{J}}. \tag{2.3.6}$$

Let $\mathcal{A} \in \mathcal{L}(L_2(\Omega), Y)$ be a linear operator and consider its discretization $A := \mathcal{A}\mathcal{F}$. In this section we aim at stating conditions under which (2.2.51) can be ensured by means of a preconditioning strategy. We will make technical assumptions on the matrix $G = \big(G_{\mu,\nu}\big)_{\mu,\nu\in\mathcal{J}}$ given by

$$G := A^*A = \big(\langle \mathcal{A}^*\mathcal{A}\psi_\nu, \psi_\mu\rangle_{L_2(\Omega)}\big)_{\mu,\nu\in\mathcal{J}}. \tag{2.3.7}$$

To be specific, we will assume that there exist constants $c_1, c_2, c_3, s, \eta, r \in \mathbb{R}_+, r > d$, such that the following conditions hold for all $\mu = (j, k, e), \nu = (j', k', e') \in \mathcal{J}$:

- The entries of G satisfy the decay estimate

$$|G_{\mu,\nu}| \leq c_1 \frac{2^{-s||\mu|-|\nu||}2^{-\eta \min(|\mu|,|\nu|)}}{\left(1 + 2^{\min(|\mu|,|\nu|)} \operatorname{dist}(\Omega_\mu, \Omega_\nu)\right)^r} \tag{2.3.8}$$

- On the diagonal, i.e., $\mu = \nu$, it holds that

$$|G_{\mu,\mu}| \geq c_2 2^{-\eta|\mu|}. \tag{2.3.9}$$

- For the same scale, i.e., $|\mu| = |\nu|$, the entries satisfy

$$|G_{\mu,\nu}| \leq c_3 \frac{2^{-2\eta|\mu|}}{(1 + |k - k'|)^r}. \tag{2.3.10}$$

Under these conditions the following holds.

Theorem 2.3.2 ([36, Theorem 4.6.]). *Suppose that G fulfills* (2.3.8), (2.3.9), *and* (2.3.10) *with $c_2 > c_3/(r-d)$. Let D be the block-diagonal matrix consisting of the square roots of the diagonal level blocks of G, i.e.,*

$$D_{\mu,\nu} := \begin{cases} G_{\mu,\nu}^{1/2} & |\mu| = |\nu|, \\ 0 & \text{otherwise.} \end{cases} \tag{2.3.11}$$

Then there exists a constant $C = C(c_1, c_2, c_3, r, d)$ such that for each finite set $\Lambda \subset \mathcal{J}$ with $|\Lambda| \leq 2^s C^{-1}$ the sub-matrix $(D^{-1}GD^{-1})_{|\Lambda\times\Lambda}$ satisfies

$$\|(\mathrm{Id} - D^{-1}GD^{-1})_{|\Lambda\times\Lambda}\| < C\, 2^{-(s-\frac{\eta}{2})}|\Lambda|$$

and

$$\kappa\big((D^{-1}GD^{-1})_{|\Lambda\times\Lambda}\big) \leq \frac{1 + C\, 2^{-(s-\frac{\eta}{2})}|\Lambda|}{1 - C\, 2^{-(s-\frac{\eta}{2})}|\Lambda|}.$$

2.3.3 Integral operators with Schwartz kernels on disjoint domains

In this section we study a class of operators which fits into the setting of Section 2.2.3. Let $\Omega, \hat{\Omega} \subset \mathbb{R}^d$ be two bounded Lipschitz domains with $\operatorname{dist}(\Omega, \hat{\Omega}) = \delta > 0$. For fixed $t \in \mathbb{R}_+$ we consider

$$\mathcal{K} = \mathcal{A} + \sigma\mathcal{N} : L_2(\Omega) \to H^t(\hat{\Omega}),$$

where $\sigma \in \mathbb{R}_+$, $\mathcal{A} \in \mathcal{L}(L_2(\Omega), H^t(\hat{\Omega}))$ is linear, and $\mathcal{N} : L_2(\Omega) \to H^t(\hat{\Omega})$ is a nonlinear operator. Furthermore, we assume that the linear part $\mathcal{A}$ is an integral operator with a Schwartz kernel. To be specific, we assume that $\mathcal{A}$ is given by

$$v \mapsto \mathcal{A}v := \int_\Omega \Phi(\cdot, \xi) v(\xi)\, \mathrm{d}\xi, \tag{2.3.12}$$

where $\Phi : \hat{\Omega} \times \Omega \to \mathbb{R}$ is a kernel of Schwartz type, i.e.,

$$|\partial_x^{\boldsymbol{\alpha}} \partial_\xi^{\boldsymbol{\beta}} \Phi(x,\xi)| \leq c_{\boldsymbol{\alpha},\boldsymbol{\beta}} |x-\xi|^{-(d+2t+|\boldsymbol{\alpha}|+|\boldsymbol{\beta}|)}, \quad \boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{N}^d, \tag{2.3.13}$$

holds. Concerning the nonlinear perturbation $\mathcal{N}$, we assume that it is given by

$$v \mapsto \mathcal{N}(v) := \int_\Omega \tilde{\Phi}(\cdot,\xi)|v(\xi)|^2 \,\mathrm{d}\xi,$$

where $\tilde{\Phi}$ also fulfills (2.3.13). This condition implies that $\mathcal{A}$ and $\mathcal{N}$ are well defined as operators mapping into $H^t(\hat{\Omega})$. Moreover, the nonlinear perturbation $\mathcal{N}$ is twice continuously differentiable and consequently indeed $\mathcal{N}$ and $\mathcal{N}'$ are Lipschitz continuous on bounded closed sets: To see this, we write $\mathcal{N} = \mathcal{N}_1 \circ \mathcal{N}_2$ with

$$\begin{aligned} \mathcal{N}_1 &: L_1(\Omega) \to H^t(\hat{\Omega}), \quad v \mapsto \int_\Omega \tilde{\Psi}(\cdot,\xi) v(\xi) \,\mathrm{d}\xi, \\ \mathcal{N}_2 &: L_2(\Omega) \to L_1(\Omega), \quad v \mapsto |v|^2. \end{aligned}$$

Here the operator $\mathcal{N}_1$ as well as the derivative of $\mathcal{N}_2$, i.e.,

$$\mathcal{N}_2' : L_2(\Omega) \to \mathcal{L}(L_2(\Omega), L_1(\Omega)), \quad v \mapsto 2v\,\cdot,$$

are linear. Recall that the synthesis map $\mathcal{F}$ associated to $\boldsymbol{\Psi}$ is given by (2.3.5). It is linear and hence Lipschitz. Together with the Lipschitz properties of $\mathcal{N}$, this implies that the discretized version of the nonlinear part, i.e., $N = \mathcal{N} \circ \mathcal{F}$, fulfills (2.2.45).

Let us now assume that the linear term $A = \mathcal{A}\mathcal{F}$ of $K = A + \sigma N$ does not fulfill already (2.2.51). We want to show that setting

$$K \circ D = A \circ D + \sigma N \circ D,$$

for a suitable preconditioning matrix D, will allow us now to fulfill it for $A \circ D$. Moreover the new nonlinear perturbation $N \circ D$ will again satisfy the Lipschitz continuity conditions (2.2.45) as soon as we will remember that, eventually, the problem will be turned into a finite-dimensional one. We shall construct the preconditioning matrix D by using the multilevel techniques presented in Section 2.3.2. To be specific, the remainder of this section is dedicated to the proof of property (2.3.8) of the matrix $G := \big(\mathcal{A}\mathcal{F}\big)^* \mathcal{A}\mathcal{F}$. (The other required properties (2.3.9) and (2.3.10) may be difficult to be shown, but they are often verified in practice.) To this end we follow the lines of [35]. To be explicit, with (2.3.7), the entries of G are given as

$$G_{\mu,\nu} = \langle \mathcal{A}^*\mathcal{A}\psi_\nu, \psi_\mu \rangle_{L_2(\Omega)} = \langle \mathcal{A}\psi_\nu, \mathcal{A}\psi_\mu \rangle_{H^t(\hat{\Omega})}. \tag{2.3.14}$$

We begin with the special case $t = 0$. and apply Taylor's formula to the kernel Φ around a point $\xi_0 \in \Omega_\mu$. For every $\xi \in \Omega_\mu$ there exists some $\theta \in [0,1]$ such that

$$\Phi(x,\xi) = \sum_{|\boldsymbol{\beta}| \leq d^*} \frac{\partial_\xi^{\boldsymbol{\beta}} \Phi(x,\xi_0)}{\boldsymbol{\beta}!} (\xi - \xi_0)^{\boldsymbol{\beta}} + \sum_{|\boldsymbol{\beta}| = d^*+1} \frac{\partial_\xi^{\boldsymbol{\beta}} \Phi(x, \xi_0 + \theta(\xi - \xi_0))}{\boldsymbol{\beta}!} (\xi - \xi_0)^{\boldsymbol{\beta}}. \tag{2.3.15}$$

With (2.3.13) we can estimate

$$\begin{aligned}
&|\sum_{|\boldsymbol{\beta}|=d^*+1} \frac{\partial_\xi^{\boldsymbol{\beta}}\Phi(x,\xi_0+\theta(\xi-\xi_0))}{\boldsymbol{\beta}!}(\xi-\xi_0)^{\boldsymbol{\beta}}| \\
&\quad\le \sum_{|\boldsymbol{\beta}|=d^*+1} \frac{1}{\boldsymbol{\beta}!}|(\xi-\xi_0)^{\boldsymbol{\beta}}| \sup_{\xi'\in\Omega_\mu} |\partial_\xi^{\boldsymbol{\beta}}\Phi(x,\xi')| \\
&\quad\le \sum_{|\boldsymbol{\beta}|=d^*+1} \frac{1}{\boldsymbol{\beta}!}|(\xi-\xi_0)^{\boldsymbol{\beta}}| c_{\mathbf{0},\boldsymbol{\beta}} \operatorname{dist}(x,\Omega_\mu)^{-(d+2t+d^*+1)}.
\end{aligned} \tag{2.3.16}$$

The cancellation property (W_2) of $\psi_\mu \in \boldsymbol{\Psi}$, together with (2.3.15) and (2.3.16) yields

$$\begin{aligned}
|\mathcal{A}\psi_\mu(x)| &= |\int_{\Omega_\mu} \Phi(x,\xi)\psi_\mu(\xi)\,\mathrm{d}\xi| \\
&\le \sum_{|\boldsymbol{\beta}|=d^*+1} \frac{1}{\boldsymbol{\beta}!} c_{\mathbf{0},\boldsymbol{\beta}} \operatorname{dist}(x,\Omega_\mu)^{-(d+2t+d^*+1)} \int_{\Omega_\mu} |(\xi-\xi_0)^{\boldsymbol{\beta}}||\psi_\mu(\xi)|\,\mathrm{d}\xi.
\end{aligned} \tag{2.3.17}$$

By our assumptions (W_1) and (W_3) on the wavelets, i.e., $\Omega_\mu \subset 2^{-|\mu|}k + 2^{-|\mu|}Q$ and $\|\psi_\mu\|_{L_\infty(\Omega)} \le C2^{d/2|\mu|}$, together with $\xi_0 \in \Omega_\mu$, it holds that

$$\begin{aligned}
\int_{\Omega_\mu} |(\xi-\xi_0)^{\boldsymbol{\beta}}||\psi_\mu(\xi)|\,\mathrm{d}\xi &\le C2^{\frac{d}{2}|\mu|} \int_{\Omega_\mu} |(\xi-\xi_0)^{\boldsymbol{\beta}}|\,\mathrm{d}\xi \\
&\le C2^{\frac{d}{2}|\mu|} \int_Q |(2^{-|\mu|}(\xi'+k)-\xi_0)^{\boldsymbol{\beta}}| 2^{-d|\mu|}\,\mathrm{d}\xi' \\
&\le C'2^{-|\mu|(\frac{d}{2}+|\boldsymbol{\beta}|)}.
\end{aligned}$$

The combination of (2.3.17) and the last estimate implies

$$|\mathcal{A}\psi_\mu(x)| \le C_{d^*} \operatorname{dist}(x,\Omega_\mu)^{-(d+2t+d^*+1)} 2^{-|\mu|(\frac{d}{2}+d^*+1)}. \tag{2.3.18}$$

Since we assumed that Ω and $\hat{\Omega}$ are disjoint domains, it holds for $\xi, \xi' \in \Omega$ with $\xi \ne \xi'$ and $\|\xi - x\|_2 \ge \delta, \|\xi' - x\|_2 \ge \delta$ that

$$\frac{1}{|\xi-x||\xi'-x|} \le C_{x,\delta} \frac{1}{|\xi-\xi'|}.$$

Furthermore $C_{x,\delta}$ can be bounded by C_δ, independently of x. With (2.3.18) we prove immediately, for $\mu \ne \nu$ with $\operatorname{dist}(\Omega_\mu,\Omega_\nu) > 0$ the estimate

$$\begin{aligned}
&|\langle \mathcal{A}\psi_\mu, \mathcal{A}\psi_\nu\rangle_{L_2(\hat{\Omega})}| \\
&\quad\le (C_{d^*})^2 2^{-(|\mu|+|\nu|)(\frac{d}{2}+d^*+1)} \int_{\hat{\Omega}} \big(\operatorname{dist}(x,\Omega_\mu)\operatorname{dist}(x,\Omega_\nu)\big)^{-(d+2t+d^*+1)}\,\mathrm{d}x \\
&\quad\le (C_{d^*})^2 C_\delta |\hat{\Omega}| 2^{-(|\mu|+|\nu|)(\frac{d}{2}+d^*+1)} \operatorname{dist}(\Omega_\mu,\Omega_\nu)^{-(d+2t+d^*+1)} \\
&\quad= (C_{d^*})^2 C_\delta |\hat{\Omega}| \frac{2^{-||\mu|-|\nu||(\frac{d}{2}+d^*+1)} 2^{-\min(|\mu|,|\nu|)(d^*+1-2t)}}{\big(2^{\min(|\mu|,|\nu|)}\operatorname{dist}(\Omega_\mu,\Omega_\nu)\big)^{d+2t+d^*+1}}.
\end{aligned} \tag{2.3.19}$$

For the case $\operatorname{dist}(\Omega_\mu, \Omega_\nu) = 0$ we use again (2.3.18) and apply $\operatorname{dist}(x, \Omega_\mu) \geq \delta$ directly to derive the simpler estimate

$$|\langle \mathcal{A}\psi_\mu, \mathcal{A}\psi_\nu \rangle_{L_2(\hat{\Omega})}| \leq (C_{d^*})^2 |\hat{\Omega}| 2^{-(|\mu|+|\nu|)(\frac{d}{2}+d^*+1)} \delta^{-2(d+2t+d^*+1)}. \tag{2.3.20}$$

Together (2.3.19) and (2.3.20) imply that $\big(\langle \mathcal{A}\psi_\mu, \mathcal{A}\psi_\nu \rangle_{L_2(\hat{\Omega})}\big)_{\mu,\nu\in\mathcal{J}}$ fulfills the assumption (2.3.8).

For the general case $t > 0$, we consider $\big(\langle \partial_x^{\boldsymbol{\alpha}}(\mathcal{A}\psi_\mu), \partial_x^{\boldsymbol{\alpha}}(\mathcal{A}\psi_\nu) \rangle_{L_2(\hat{\Omega})}\big)_{\mu,\nu\in\mathcal{J}}$, $\boldsymbol{\alpha} \in \mathbb{N}_0^d$. In this setting

$$\partial_x^{\boldsymbol{\alpha}}(\mathcal{A}\psi_\mu) = \int_\Omega \partial_x^{\boldsymbol{\alpha}} \Phi(\cdot, \xi) \psi_\mu(\xi) \, \mathrm{d}\xi$$

is again an integral with a Schwartz kernel. Indeed, an analogous argumentation as in the case $t = 0$ yields condition (2.3.8) for the case $t \in \mathbb{N}$, and consequently for $t \in \mathbb{R}_+$.

2.4 Equivalence to an inexact finite-dimensional scheme

In practice, whenever we deal with infinite-dimensional problems, the two operators K and K' can not be evaluated exactly, and one has to replace their output by suitable numerical approximations. In this section we study the convergence behavior of the resulting inexact algorithm to solve the preconditioned minimization problem (2.3.3). Although the original problem is posed in general in infinite dimensions, adaptive approximations will allow us to show the confinement of the iteration within a well-determined finite-dimensional space. In particular, in Theorem 2.4.3 below, we show that the global support of all iterates is contained in a finite set Λ_0. From a practical point of view, there would be no difference between the iterates produced by the adaptive scheme over the whole index set $\mathcal{J}$ or if we would restrict the set of possible indices to the (a priori unknown) set Λ_0. Therefore, by arguing as in Section 2.3.1, the combination of preconditioning and adaptive solvers yields an iterative scheme for the minimization of the unpreconditioned functional $\Gamma_{\boldsymbol{\alpha}}$.

We focus on the error introduced by the inexact evaluation of the nonlinear functional K and the linear operator $(K'(\cdot))^*$. To this end let us assume that for given tolerances $\varrho, \delta > 0$, there exist approximation schemes which for every $\mathbf{v} \in \ell_2$ and pairs $(\mathbf{v}, w) \in \ell_2 \times Y$, respectively, compute finite-dimensional approximations $[K(\mathbf{v})]_\varrho$ and $[\big(K'(\mathbf{v})\big)^*(w)]_\delta$ such that

$$\begin{aligned} \|K(\mathbf{v}) - [K(\mathbf{v})]_\varrho\|_Y &\leq \varrho, \\ \|\big(K'(\mathbf{v})\big)^*(w) - [\big(K'(\mathbf{v})\big)^*(w)]_\delta\|_{\ell_2} &\leq \delta. \end{aligned} \tag{2.4.1}$$

This assumption is realistic, e.g., if the exact application of K and $(K'(\cdot))^*$ involves the solution of partial differential or integral equations and the numerical approximations can be computed by means of adaptive discretization schemes. Let us mention

two prominent examples from the context of adaptive wavelet schemes of linear and nonlinear operator equations.

Example 2.4.1. Let $\mathbf{\Psi}_X = \{\psi_{X,\mu}\}_{\mu\in\mathcal{J}_X}$ and $\mathbf{\Psi}_Y = \{\psi_{Y,\mu}\}_{\mu\in\mathcal{J}_Y}$ be wavelet Riesz bases for X and Y, respectively, such that the assumptions of Section 2.3.2 are satisfied. We denote the associated synthesis operators by $\mathcal{F}_X$ and $\mathcal{F}_Y$. Furthermore let $K = \mathcal{K}\circ\mathcal{F}_X : \ell_2 \to Y$ for some nonlinear operator $\mathcal{K} : X \to Y$.

1. For the efficient approximate application of the linear operator $(K'(\mathbf{v}))^*$ to a given point $w \in Y$, it is advantageous if the coefficient array $\mathbf{w} \in \ell_2(\mathcal{J}_Y)$ of $w = \mathcal{F}_Y(\mathbf{w})$, or at least good approximations of it, has a fast decay [26]. In that case, one may exploit the representation

$$(K'(\mathbf{v}))^*(w) = (\mathcal{F}_X^* \circ (\mathcal{K}'(\mathcal{F}_X(\mathbf{v}))^* \circ \mFF_Y)(\mathbf{w}) =: \mathbf{A}_\mathbf{v}\mathbf{w}$$

   and the compressibility of the stiffness matrix $\mathbf{A}_\mathbf{v} \in \mathcal{L}(\ell_2(\mathcal{J}_Y), \ell_2(\mathcal{J}_X))$. In fact, if $\mathbf{A}_\mathbf{v} \in \mathcal{L}(\ell_\tau^w(\mathcal{J}_Y), \ell_\tau^w(\mathcal{J}_X))$ for all $0 < \tau_0 < \tau < 2$, then the second inequality in (2.4.1) can be ensured by suitable matrix compression techniques. In the special case of wavelet Riesz bases $\mathbf{\Psi}_X$, $\mathbf{\Psi}_Y$ and $\mathcal{K}'(\mathcal{F}_X(\mathbf{v}))$ being a differential operator or an integral operator with Schwartz kernel, e.g., we can expect that the stiffness matrix $\mathbf{A}_\mathbf{v}$ is s^*-compressible, i.e., there exist biinfinite matrices $\mathbf{A}_{\mathbf{v},j}$ with at most a constant multiple of 2^j nontrivial entries per row and column, such that $\|\mathbf{A}_\mathbf{v} - \mathbf{A}_{\mathbf{v},j}\|_2 \le C_s 2^{-js}$, $0 < s < s^*$. This property implies that $\mathbf{A}_\mathbf{v}$ boundedly maps $\ell_\tau^w(\mathcal{J}_Y)$ into $\ell_\tau^w(\mathcal{J}_X)$. We refer to [26, 119] and related works on the compressibility of operators in wavelet coordinates and the concrete realization of associated adaptive matrix-vector multiplications.

2. The approximate evaluation of the nonlinearity K itself at a given input $\mathbf{v} \in \ell_2$ is enabled under additional assumptions on the type of the nonlinearity. In the context of nonlinear operators, tree approximation techniques play an important role. Here a tree structure is imposed on the coefficient array of the output argument. For example, in the special case that X is a closed subspace of $H^s(\Omega)$, $s \ge 0$, $\Omega \subset \mathbb{R}^d$ a bounded domain, $Y = X'$ and $\mathcal{K}$ decomposes into $\mathcal{K} = \mathcal{A} + \mathcal{N}$ with a linear, boundedly invertible operator $\mathcal{A} : X \to X'$ and a Nemytskii-type nonlinearity

$$\mathcal{N} : X \to X', \quad (\mathcal{N}(v))(x) = f\big(\partial^{\beta_1} v(x), \dots, \partial^{\beta_k} v(x)\big), \quad \beta_j \in \mathbb{N}_0^d,$$

   adaptive wavelet tree approximation techniques have been developed and implemented in [9, 28, 45, 79].

For simplicity, we will assume in the sequel that y is given exactly. For convenience, we define the analogue of (2.2.8) by

$$\tilde{T}_{\varrho,\delta}(\mathbf{v}) := \mathbf{v} - \frac{1}{\lambda}[\big(K'(\mathbf{v})\big)^*([K(\mathbf{v})]_\varrho - y)]_\delta. \tag{2.4.2}$$

An implementable version of ISTA with decreasing threshold parameters $\boldsymbol{\alpha}^{(n)}$, i.e., (2.1.9), is then given by

$$\tilde{\mathbf{u}}^{(n+1)} = \mathbb{S}_{\frac{1}{\lambda}\boldsymbol{\alpha}^{(n)}}\big(\tilde{T}_{\varrho^{(n)},\delta^{(n)}}(\tilde{\mathbf{u}}^{(n)})\big). \tag{2.4.3}$$

The following theorem shows that if the parameters $\varrho^{(n)}, \delta^{(n)}$ are suitably chosen, the overall algorithm is still linearly convergent.

Theorem 2.4.2. *Let $\mathbf{u}^*$ be a stationary point of (2.1.1) that satisfies $T(\mathbf{u}^*) \in \ell^w_\tau(\mathcal{J})$ for some $0 < \tau < 2$. Furthermore let $\boldsymbol{\alpha}^{(n)}, \boldsymbol{\alpha} \in \mathbb{R}^{\mathcal{J}}_+$ with $\alpha^{(n)}_\mu \geq \alpha_\mu \geq \alpha \in \mathbb{R}_+, \mu \in \mathcal{J}$. We set $\tilde{\mathbf{u}}^{(0)} = \mathbf{0}$ and assume that T fulfills condition (2.2.27). With C^{Lip}_T be as therein and C is as in Lemma 2.2.2 we set*

$$\tilde{L} := \frac{4\big((C^{Lip}_T + \tilde{\gamma} - \gamma)\|\mathbf{u}^*\|_{\ell_2}\big)^2\lambda^2}{\alpha^2} + 4C|T(\mathbf{u}^*)|^\tau_{\ell^w_\tau}\Big(\frac{\alpha}{\lambda}\Big)^{-\tau},$$

and define $\mathcal{B}_{\tilde{L}}$ analogously to (2.2.31). Let us assume that there exists some $0 < \gamma_0 < 1$, such that for all $\mathbf{v} \in \mathcal{B}_{\tilde{L}}$ and $\operatorname{supp}(\mathbf{v}) \subset \Lambda \subset \mathcal{J}$ with $\#\Lambda \leq 2\tilde{L}$

$$\|\big(T(\mathbf{u}^*) - T(\mathbf{v})\big)_{|S^*\cup\Lambda}\|_{\ell_2(S^*\cup\Lambda)} \leq \gamma_0\|\mathbf{u}^* - \mathbf{v}\|_{\ell_2}. \tag{2.4.4}$$

For the operator K' we assume

$$C^{bnd}_{K'}(\mathcal{B}_{\tilde{L}}) := \sup_{\mathbf{v}\in\mathcal{B}_{\tilde{L}}} \|K'(\mathbf{v})\|_{\mathcal{L}(\ell_2,Y)} < \infty. \tag{2.4.5}$$

Then, for any $\gamma_0 < \gamma < \tilde{\gamma} < 1$ the inexact thresholded iteration (2.4.3) fulfills

$$\big(\tilde{\mathbf{u}}^{(n)}\big)_{n\in\mathbb{N}} \subset \mathcal{B}_{\tilde{L}}$$

and converges to $\mathbf{u}^*$ at a linear rate

$$\|\mathbf{u}^* - \tilde{\mathbf{u}}^{(n)}\|_{\ell_2} \leq \tilde{\varepsilon}^{(n)} := \tilde{\gamma}^n\|\mathbf{u}^*\|_{\ell_2}, \tag{2.4.6}$$

whenever the parameters and tolerances are chosen according to

$$\frac{1}{\lambda}\Big(C^{bnd}_{K'}(\mathcal{B}_{\tilde{L}})\varrho^{(n)} + \delta^{(n)}\Big) \leq (\tilde{\gamma} - \gamma)\tilde{\varepsilon}^{(n)}, \tag{2.4.7}$$

$$\max_{\mu\in\mathcal{J}} |\alpha^{(n)}_\mu - \alpha_\mu| \leq \lambda\tilde{L}^{-\frac{1}{2}}(\gamma - \gamma_0)\tilde{\varepsilon}^{(n)}. \tag{2.4.8}$$

Proof. The proof is an induction over n. The case $n = 0$ is covered by the assumptions. Now let $\tilde{\mathbf{u}}^{(n)} \in \mathcal{B}_{\tilde{L}}$ and (2.4.6) hold for $n \in \mathbb{N}$. We begin by proving $\#\operatorname{supp}\tilde{\mathbf{u}}^{(n+1)} \leq \tilde{L}$.

To this end we use the standing assumption (2.4.1) on the inexact operator evaluations, together with the assumption (2.4.7) to estimate for $\mathbf{v} \in \mathcal{B}_{\tilde{L}}$

$$\begin{aligned}
&\|T(\mathbf{v}) - \tilde{T}_{\varrho^{(n)},\delta^{(n)}}(\mathbf{v})\|_{\ell_2} \\
&\quad = \frac{1}{\lambda}\|\big(K'(\mathbf{v})\big)^*(K(\mathbf{v}) - y) - [\big(K'(\mathbf{v})\big)^*([K(\mathbf{v})]_{\varrho^{(n)}} - y)]_{\delta^{(n)}}\|_{\ell_2} \\
&\quad \leq \frac{1}{\lambda}\Big(\|\big(K'(\mathbf{v})\big)^*(K(\mathbf{v}) - y) - \big(K'(\mathbf{v})\big)^*([K(\mathbf{v})]_{\varrho^{(n)}} - y)\|_{\ell_2} \\
&\quad + \|\big(K'(\mathbf{v})\big)^*([K(\mathbf{v})]_{\varrho^{(n)}} - y) - [\big(K'(\mathbf{v})\big)^*([K(\mathbf{v})]_{\varrho^{(n)}} - y)]_{\delta^{(n)}}\|_{\ell_2}\Big) \\
&\quad \leq \frac{1}{\lambda}\Big(\|\big(K'(\mathbf{v})\big)^*\|_{\mathcal{L}(Y,\ell_2)}\varrho^{(n)} + \delta^{(n)}\Big) \\
&\quad \leq (\tilde{\gamma} - \gamma)\tilde{\varepsilon}^{(n)}.
\end{aligned} \tag{2.4.9}$$

This inequality, applied for $\boldsymbol{v} = \tilde{\mathbf{u}}^{(n)} \in \mathcal{B}_{\tilde{L}}$, implies together with the Lipschitz continuity assumption (2.2.27) that

$$\begin{aligned}
&\|T(\mathbf{u}^*) - \tilde{T}_{\varrho^{(n)},\delta^{(n)}}(\tilde{\mathbf{u}}^{(n)})\|_{\ell_2} \\
&\quad = \|T(\mathbf{u}^*) - T(\tilde{\mathbf{u}}^{(n)}) + T(\tilde{\mathbf{u}}^{(n)}) - \tilde{T}_{\varrho^{(n)},\delta^{(n)}}(\tilde{\mathbf{u}}^{(n)})\|_{\ell_2} \\
&\quad \leq (C_T^{Lip} + \tilde{\gamma} - \gamma)\tilde{\varepsilon}^{(n)},
\end{aligned}$$

By invoking Lemma 2.2.2 we can conclude

$$\#\operatorname{supp}(\tilde{\mathbf{u}}^{(n+1)}) = \#\operatorname{supp}\mathbb{S}_{\frac{1}{\lambda}\boldsymbol{\alpha}^{(n)}}\big(\tilde{T}_{\varrho^{(n)},\delta^{(n)}}(\tilde{\mathbf{u}}^{(n)})\big) \leq \tilde{L}. \tag{2.4.10}$$

For the second part of the proof we set

$$\tilde{\Lambda}^{(n)} := S^* \cup \operatorname{supp}\tilde{\mathbf{u}}^{(n)} \cup \operatorname{supp}\tilde{\mathbf{u}}^{(n+1)}.$$

Notice that $\#\operatorname{supp}\tilde{\mathbf{u}}^{(n)} \cup \operatorname{supp}\tilde{\mathbf{u}}^{(n+1)} \leq 2\tilde{L}$. Because shrinkage is nonexpansive and by the assumption (2.4.4) we may estimate

$$\begin{aligned}
&\|\mathbf{u}^*_{|\tilde{\Lambda}^{(n)}} - \mathbb{S}_{\frac{1}{\lambda}\boldsymbol{\alpha}}\big(T(\tilde{\mathbf{u}}^{(n)})_{|\tilde{\Lambda}^{(n)}}\big)\|_{\ell_2(\tilde{\Lambda}^{(n)})} \\
&\quad = \|\mathbb{S}_{\frac{1}{\lambda}\boldsymbol{\alpha}}\big(T(\mathbf{u}^*)_{|\tilde{\Lambda}^{(n)}}\big) - \mathbb{S}_{\frac{1}{\lambda}\boldsymbol{\alpha}}\big(T(\tilde{\mathbf{u}}^{(n)})_{|\tilde{\Lambda}^{(n)}}\big)\|_{\ell_2(\tilde{\Lambda}^{(n)})} \leq \gamma_0\tilde{\varepsilon}^{(n)}.
\end{aligned} \tag{2.4.11}$$

Moreover, we may use the Lipschitz assumption (2.2.27) and invoke Lemma 2.2.2 directly to conclude

$$\begin{aligned}
\#\operatorname{supp}\mathbb{S}_{\frac{1}{\lambda}\boldsymbol{\alpha}}\big(T(\tilde{\mathbf{u}}^{(n)})\big) &\leq \frac{4\|T(\mathbf{u}^*) - T(\tilde{\mathbf{u}}^{(n)})\|_{\ell_2}^2\lambda^2}{\alpha^2} + 4C|T(\mathbf{u}^*)|_{\ell_\tau^w}^\tau\Big(\frac{\alpha}{\lambda}\Big)^{-\tau} \\
&\leq \frac{4(C_T^{Lip}\tilde{\varepsilon}^{(n)})^2\lambda^2}{\alpha^2} + 4C|T(\mathbf{u}^*)|_{\ell_\tau^w}^\tau\Big(\frac{\alpha}{\lambda}\Big)^{-\tau} \\
&\leq \tilde{L}.
\end{aligned} \tag{2.4.12}$$

Since $\boldsymbol{\alpha}^{(n)}$ is decreasing to $\boldsymbol{\alpha}$ it holds that

$$\operatorname{supp} \mathbb{S}_{\frac{1}{\lambda}\boldsymbol{\alpha}^{(n)}}\big(T(\tilde{\mathbf{u}}^{(n)})\big) \subset \operatorname{supp} \mathbb{S}_{\frac{1}{\lambda}\boldsymbol{\alpha}}\big(T(\tilde{\mathbf{u}}^{(n)})\big).$$

This, together with (2.4.8) gives

$$\begin{aligned} &\|\mathbb{S}_{\frac{1}{\lambda}\boldsymbol{\alpha}}\big(T(\tilde{\mathbf{u}}^{(n)})_{|\tilde{\Lambda}^{(n)}}\big) - \mathbb{S}_{\frac{1}{\lambda}\boldsymbol{\alpha}^{(n)}}\big(T(\tilde{\mathbf{u}}^{(n)})_{|\tilde{\Lambda}^{(n)}}\big)\|_{\ell_2(\tilde{\Lambda}^{(n)})} \\ &\quad \leq \frac{\big(\#\operatorname{supp} \mathbb{S}_{\frac{1}{\lambda}\boldsymbol{\alpha}}(T(\tilde{\mathbf{u}}^{(n)}))\big)^{\frac{1}{2}}}{\lambda} \max_{\mu\in\mathcal{J}} |\alpha_\mu^{(n)} - \alpha_\mu| \\ &\quad \leq (\gamma - \gamma_0)\tilde{\varepsilon}^{(n)}. \end{aligned} \tag{2.4.13}$$

Finally, we use that shrinkage in nonexpansive, together with (2.4.9) for $\tilde{\mathbf{u}}^{(n)}$ for the estimate

$$\|\mathbb{S}_{\frac{1}{\lambda}\boldsymbol{\alpha}^{(n)}}\big(T(\tilde{\mathbf{u}}^{(n)})_{|\Lambda^{(n)}}\big) - \mathbb{S}_{\frac{1}{\lambda}\boldsymbol{\alpha}^{(n)}}\big(\tilde{T}_{\varrho^{(n)},\delta^{(n)}}(\tilde{\mathbf{u}}^{(n)})_{|\tilde{\Lambda}^{(n)}}\big)\|_{\ell_2(\tilde{\Lambda}^{(n)})} \leq (\tilde{\gamma} - \gamma)\tilde{\varepsilon}^{(n)}. \tag{2.4.14}$$

The combination of (2.4.11), (2.4.13), and (2.4.14) finalizes the proof

$$\begin{aligned} \|\mathbf{u}^* - \tilde{\mathbf{u}}^{(n+1)}\|_{\ell_2} &\leq \|\mathbf{u}^*_{|\tilde{\Lambda}^{(n)}} - \mathbb{S}_{\frac{1}{\lambda}\boldsymbol{\alpha}}\big(T(\tilde{\mathbf{u}}^{(n)})_{|\tilde{\Lambda}^{(n)}}\big)\|_{\ell_2(\tilde{\Lambda}^{(n)})} \\ &\quad + \|\mathbb{S}_{\frac{1}{\lambda}\boldsymbol{\alpha}}\big(T(\tilde{\mathbf{u}}^{(n)})_{|\tilde{\Lambda}^{(n)}}\big) - \mathbb{S}_{\frac{1}{\lambda}\boldsymbol{\alpha}^{(n)}}\big(T(\tilde{\mathbf{u}}^{(n)})_{|\tilde{\Lambda}^{(n)}}\big)\|_{\ell_2(\tilde{\Lambda}^{(n)})} \\ &\quad + \|\mathbb{S}_{\frac{1}{\lambda}\boldsymbol{\alpha}^{(n)}}\big(T(\tilde{\mathbf{u}}^{(n)})_{|\tilde{\Lambda}^{(n)}}\big) - \tilde{\mathbf{u}}^{(n+1)}_{|\tilde{\Lambda}^{(n)}}\|_{\ell_2(\tilde{\Lambda}^{(n)})} \\ &\leq \gamma_0\tilde{\varepsilon}^{(n)} + (\gamma - \gamma_0)\tilde{\varepsilon}^{(n)} + (\tilde{\gamma} - \gamma)\tilde{\varepsilon}^{(n)} = \tilde{\varepsilon}^{(n+1)}. \end{aligned}$$ □

We have shown that the support size of each iterate $\tilde{\mathbf{u}}^{(n)}$ can be bounded by a uniform constant. As it turns out there also exists a bounded set $\Lambda_0 \subset \mathcal{J}$ that contains all those supports.

Theorem 2.4.3. *Let the assumptions of Theorem 2.4.2 hold. Let $N \in \mathbb{N}$ be large enough such that there exists some $\delta > 0$ with*

$$\tilde{\varepsilon}^{(N+1)} + \delta \leq \frac{1}{\lambda} \inf_{\mu\in\mathcal{J}} \alpha_\mu. \tag{2.4.15}$$

Then it holds that

$$\operatorname{supp}(\tilde{\mathbf{u}}^{(n)}) \subset \Lambda_\delta(T(\mathbf{u}^*)), \quad n \geq N,$$

and consequently

$$\operatorname{supp}(\tilde{\mathbf{u}}^{(n)}) \subset \Big(\bigcup_{j=0}^{N} \operatorname{supp}(\tilde{\mathbf{u}}^{(j)})\Big) \cup \Lambda_\delta(T(\mathbf{u}^*)) =: \Lambda_0, \quad n \in \mathbb{N}.$$

Proof. We prove that for any fixed $n \geq N$ and for all $\mu \in \mathcal{J} \setminus \Lambda_\delta(T(\mathbf{u}^*))$ it holds that $(\tilde{\mathbf{u}}^{(n+1)})_\mu = 0$. To this end let $0 < \gamma_0 < \gamma < \tilde{\gamma} < 1$ be as in Theorem 2.4.2 and denote $\tilde{\Lambda}^{(n)} := S^* \cup \operatorname{supp} \tilde{\mathbf{u}}^{(n)} \cup \tilde{\mathbf{u}}^{(n+1)}$. Recall that by estimating as in equation (2.4.9) it holds that

$$\|\big(T(\tilde{\mathbf{u}}^{(n)}) - \tilde{T}_{\varrho^{(n)},\delta^{(n)}}(\tilde{\mathbf{u}}^{(n)})\big)_{|\tilde{\Lambda}^{(n)}}\|_{\ell_2(\tilde{\Lambda}^{(n)})} \leq (\tilde{\gamma} - \gamma)\tilde{\varepsilon}^{(n)}, \tag{2.4.16}$$

and further that, since $\#\operatorname{supp} \tilde{\mathbf{u}}^{(n)} \leq \tilde{L}$, we can use (2.4.4) to estimate

$$\|\big(T(\mathbf{u}^*) - T(\tilde{\mathbf{u}}^{(n)})\big)_{|\tilde{\Lambda}^{(n)}}\|_{\ell_2(\tilde{\Lambda}^{(n)})} \leq \gamma_0 \tilde{\varepsilon}^{(n)}. \tag{2.4.17}$$

By definition $\mu \in \tilde{\Lambda}^{(n)} \setminus \Lambda_\delta(T(\mathbf{u}^*))$ implies that $|\big(T(\mathbf{u}^*)\big)_\mu| \leq \delta$. Therefore, for such μ we may use (2.4.16), (2.4.17), (2.4.15), and the fact that $\boldsymbol{\alpha}^{(n)}$ is decreasing to $\boldsymbol{\alpha}$ to estimate

$$\begin{aligned} |\big(\tilde{T}_{\varrho^{(n)},\delta^{(n)}}(\tilde{\mathbf{u}}^{(n)})\big)_\mu| &\leq \|\big(T(\mathbf{u}^*) - \tilde{T}_{\varrho^{(n)},\delta^{(n)}}(\tilde{\mathbf{u}}^{(n)})\big)_{|\tilde{\Lambda}^{(n)}}\|_{\ell_2(\tilde{\Lambda}^{(n)})} + |\big(T(\mathbf{u}^*)\big)_\mu| \\ &\leq \tilde{\varepsilon}^{(N+1)} + \delta \leq \frac{1}{\lambda} \inf_{\nu \in \mathcal{J}} \alpha_\nu \leq \frac{1}{\lambda} \inf_{\nu \in \mathcal{J}} \alpha_\nu^{(n)}. \end{aligned}$$

Finally, by the definition of $\mathbb{S}_{\frac{1}{\lambda}\boldsymbol{\alpha}^{(n)}}$ it follows that $(\tilde{\mathbf{u}}^{(n+1)})_\mu = 0$. □

3 An adaptive wavelet solver for a nonlinear parameter identification problem for a parabolic differential equation with sparsity constraints

Authors: S. Dahlke, U. Friedrich, P. Maass, R. A. Ressel, T. Raasch.
Journal: Journal of Inverse and Ill-Posed Problems **20** (2012), no. 2, 213–251.
Abstract: In this paper, we combine concepts from two different mathematical research topics: Adaptive wavelet techniques for well-posed problems and regularization theory for nonlinear inverse problems with sparsity constraints. We are concerned with identifying certain parameters in a parabolic reaction-diffusion equation from measured data. Analytical properties of the related parameter-to-state operator are summarized, which justify the application of an iterated soft shrinkage algorithm for minimizing a Tikhonov functional with sparsity constraints. The forward problem is treated by means of a new adaptive wavelet algorithm which is based on tensor wavelets.

In its general form, the underlying PDE describes gene concentrations in embryos at an early state of development. We implemented an algorithm for the related nonlinear parameter identification problem and numerical results are presented for a simplified test equation.
MSC 2010: 46N10, 47A52, 49M99, 65F20, 65F50, 65M32, 65M60, 65N12, 65T60.
Key Words: Regularization of ill-posed problems, sparsity, adaptive numerical schemes, tensor wavelets, parabolic partial differential equations, iterated soft shrinkage, embryogenesis.

3.1 Introduction

For about 30 years the advances of experimental techniques in genetic research have produced an abundance of data on gene expression. With full justification one may say that genetic research has matured enough for the application of mathematical methods permitting the extraction of structural information from this compiled data. A particularly popular object of genetic research is the Drosophila fly in which by genetic manipulation one may investigate the effect and the mutual interaction of certain genes on the development of the animal. However, conducting these experiments in-vitro is a challenging process. Therefore it is desirable to determine certain

critical parameters π in the animal's evolution from the measured expression of genes at certain times of its life cycle. One approach in studying gene regulation is to consider gene product concentrations as the state variables of a model and to assume that mutual gene interactions correspond to the synthesis rate of mentioned gene products [104].

Mathematically this amounts to solving an operator equation of the kind

$$\mathscr{D}(\pi) = y,$$

where $\mathscr{D}$ is the so-called *control-to-state operator* mapping the model parameters π to the solution of a system of parabolic partial differential equations, i.e., the data y. In the case of embryogenesis models, the set of parameters π includes reaction and diffusion coefficients, and the corresponding data y denotes the concentrations of different genes as a function in time and space. In practice the data y_{data} is usually only available at certain time instances, i.e., we have to deal with $y_{\text{data}} = \mathcal{M}y$, where $\mathcal{M}$ is a restriction operator to a finite set of time samples. Further the data is assumed to be contaminated with noise, i.e., it is given as y^δ, $\|y_{data} - y^\delta\| \leq \delta$. The operator $\mathscr{D}$ is nonlinear and ill-posed, so regularization techniques have to be employed. The reader unfamiliar with regularization may consult standard references such as the monographs [54, 90].

We use Tikhonov regularization to reformulate the inverse problem as finding the minimizer of the functional

$$\|\mathcal{M}\mathscr{D}(\pi) - y^\delta\|^2 + \alpha J(\pi). \tag{3.1.1}$$

The choice of the penalty term J gives some leeway to enforce certain characteristics of the minimizer, such as sparsity with respect to a chosen discretization. A typical choice that promotes sparsity are weighted sequence norms $J(\cdot) = \|\cdot\|_{w,q}, 1 \leq q \leq 2$, of the solution coefficients with respect to a Riesz basis or a frame. The biology of the underlying problem justifies this approach: the gene interactions are localized at very specific parts of the embryo and the mutual influence of all genes on the synthesis of one particular gene is limited, i.e., only few genes interfere with one particular gene.

This type of Tikhonov functionals has been investigated in the pioneering paper [46] for linear operators $\mathscr{D}$ in a Hilbert space setting and a penalty term with respect to an orthonormal basis. Subsequently, several approaches for generalizations to nonlinear operator equations have been proposed, see e.g. [15, 13, 102]. We will follow the approach of [15, 13] and use an iterated soft shrinkage method to compute the minimizer of (3.1.1). This method requires the solution of a forward problem and of some adjoint equation as well as a thresholding procedure in each iteration step.

Adaptive wavelet methods are ideally suited for solving the forward problem: the structure of the wavelet expansions matches well the biological qualities of the data, e.g., spatial localization of features. Further, the expansion coefficients provided by the wavelet solver almost immediately provide the needed coefficients for the thresholding

procedure. Furthermore, adaptive wavelet methods converge with the same order as the best N-term approximation which is particularly fast for sparse signals.

In the context of finite element schemes, adaptive algorithms for well-posed operator equations have a long, successful history, see [26] for an overview. Moreover, quite recently the design of adaptive algorithms based on wavelets has lead to a fundamental breakthrough. Indeed, in [26, 27] optimal adaptive algorithms in the above sense that are guaranteed to converge for a large class of problems, including operators of negative order, have been designed.

So far, the whole theory of adaptive wavelet solvers is well-developed for boundedly invertible operators. Some effort has been spent to generalize these ideas also to inverse problems, we refer for example to [36, 103], but this field is still in its infancy. However, since we utilize an iterative approach, we can take advantage of the mentioned adaptive algorithms at least for the forward problem, and embed this into a regularized iteration procedure for solving the inverse problem. The classical approach would be to use *isotropic* wavelets that span a complement space between consecutive spaces of a multi-resolution analysis. However, in this case the order of convergence of the wavelet algorithm deteriorates dramatically with the space dimension. Therefore in this paper, we use an algorithm based on recently developed *anisotropic* wavelets. Such a tensor basis contains the so-called sparse grid or hyperbolic cross spaces [16, 131]. It is known that a function with L_2 bounded mixed derivatives of a sufficiently large order can be approximated from sparse grid spaces at a rate that does not deteriorate as a function of the space dimension. In this sense the so-called *curse of dimensionality* is avoided. As demonstrated in [51, 110] also in the tensor product setting, adaptive wavelet methods realize the rate of best N-term approximation in linear complexity. Let us briefly mention that quite recently the construction of anisotropic tensor wavelets has been generalized to quadrangulable domains [20].

In summary, we are faced with the following tasks: First of all one needs to analyze the analytical properties of $\mathscr{D}$, in order to verify the assumptions of Tikhonov regularization with sparsity constraints. We do so in a fairly general setting and incorporate quite recently established results on maximal L_p-regularity of the solution of parabolic equations [67, 4]. Secondly, we need to compute the minimizer of (3.1.1). To this end we apply a generalized conditional gradient method, which is reformulated as an iterated soft shrinkage method. We give explicit formulas for the solutions to the forward problem and the adjoint problem. Moreover, as an important building block, an optimal solver for these problems utilizing tensor wavelets has to be designed.

Therefore the outline of the paper is as follows. As already mentioned, this paper aims at combining recent results on the analytic properties of nonlinear parameter identification problems for parabolic problems [105] with adaptive wavelet solvers for the underlying PDEs, see [101]. In order to make this paper self-contained, we review the major building blocks in the first sections. Nevertheless, these survey sections contain some new results, e.g., the specification of the iterated soft shrinkage method for the embryogenesis problem, the results in the Theorem 3.3.6 about local Lip-

schitz continuity of the derivative of the control-to-state operator and Theorem 3.5.8 regarding the fast decay of entries of stiffness matrices arising from tensor wavelet discretization of elliptic PDEs. We start in Section 3.2 with a formulation of the biological model problem as a nonlinear parabolic equation (3.2.1). Then the functional analytic setting is given. Properties of the underlying operators, such as Lipschitz continuity and Fréchet differentiability are derived and the existence and uniqueness result related to (3.2.1) is summarized. In Section 3.3 we analyze the mapping properties of the control-to-state map $\mathscr{D}$. We prove differentiability and local Lipschitz continuity and give an explicit formula for the adjoint of $\mathscr{D}'$. In Section 3.4 a regularization procedure is derived. We state a generalized conditional gradient method and its numerical implementation as a soft shrinkage procedure. Section 3.5 is dedicated to basic ideas of adaptive wavelet algorithms for elliptic equations with a special emphasis on the tensor wavelet setting. We describe how adaptive strategies can be used to treat parabolic equations by means of Rothe's method, i.e., the parabolic equation is first discretized in time and then in space. For stability reasons, one has to use an implicit scheme, so that an elliptic subproblem has to be solved in each time step. This is achieved by means of the proposed adaptive tensor wavelet solver. Finally in Section 3.6 the adaptive wavelet methods from Section 3.5 are combined with the algorithm for the inverse problem developed in Section 3.4. Numerical experiments for a simplified parabolic model problem in one and two space dimensions are presented.

3.2 Analysis of the forward problem

In this section we present and analyze the forward problem (3.2.1). We begin by presenting the biological model and the admissible set of parameters. Then the associated function spaces and operators are introduced. Finally, the solution theory is presented as far as it is needed and the existence and uniqueness of a solution of (3.2.1) is proved.

3.2.1 The biological model

The state variables, i.e., the concentrations of gene products, undergo permanent change over time. One of the assumed reasons for this change is direct regulation of the synthesis of one gene by the concentrations of other genes; further causes are diffusive processes of gene products through the admissible domain and decay, i.e., consumption, of the respective gene products. The synthesis requires some regulating function in a manner that reflects saturation in the signal response.

A mathematical formulation is given as follows. Let $U \subset \mathbb{R}^n, n = 2, 3$ denote some bounded Lipschitz domain and $U_T := (0, T] \times U$. The following model describes the interaction of N genes, the concentration of the i-th gene on U_T is denoted by g_i.

Then the gene expression evolution is modeled by the reaction-diffusion equation

$$\begin{aligned} &\frac{\partial g_i}{\partial t} - \operatorname{div}(D_i \nabla g_i) + \lambda_i \cdot g_i = R_i \Phi_i((Wg)_i) \text{ in } U_T \\ &\frac{\partial g_i}{\partial \nu} = 0 \quad \text{on } [0,T] \times \partial U, \quad g(0) = g_0 \text{ on } \{0\} \times U \end{aligned} \tag{3.2.1}$$

where $i = 1, \ldots, N$, $\frac{\partial}{\partial \nu}$ denotes the normal derivative. In the following we use a vector notation, e.g., $R := (R_1, \ldots, R_N)$, $\Phi := (\Phi_1, \ldots, \Phi_N)$. For the initial value we assume $g_0 \in W_2^1(U, \mathbb{R}^N)$. In our setting, a natural choice of the solution spaces to (3.2.1) are subspaces of Bochner integrable functions, i.e., generalized Sobolev spaces, see Section 3.2.2.

The model includes diffusion and decay of gene products via the parameters D and λ, both varying in time and space. The synthesis term $R\Phi(W\cdot)$ consists of a maximal synthesis rate R and the sigmoidal signal response function

$$\Phi_i : \mathbb{R} \to \mathbb{R}; \quad \Phi_i(y) = \frac{1}{2}\left(\frac{y}{\sqrt{y^2+1}} + 1\right). \tag{3.2.2}$$

Our particular choice of a sigmoidal signal response function is motivated by the investigations of [93]. However, other response functions are possible, see Section 3.2.2. The parameter that is most relevant from a biological point of view is the coupling matrix W. Here positive entries correspond to amplifying effects of gene products on others and negative ones describe an inhibiting influence.

The biological background of the model justifies certain assumptions on the parameters. First of all, all parameters are bounded, i.e., the admissible sets are subsets of L_∞ spaces. Further D and λ may only assume positive values. However, we want to apply generalized gradient methods which involve the dual space of the parameter space. The L_∞ topology would then require dealing with the very inconvenient dual of some L_∞ product space. Whenever theory permits, we will try to avoid this.

We therefor consider all parameters

$$\begin{aligned} &D \in L_\infty([0,T] \times U, \mathbb{R}^N), \quad \lambda \in L_{p_\lambda}([0,T] \times U, \mathbb{R}^N), \\ &R \in L_{p_R}([0,T] \times U, \mathbb{R}^N), \quad W \in L_{p_W}([0,T] \times U, \mathbb{R}^{N \times N}) \end{aligned}$$

that are additionally elements of the respective L_∞ spaces fulfilling the bounds

$$0 < C_{\mathcal{P},1} \leq D, \lambda \leq C_{\mathcal{P},2}, \quad 0 \leq R \leq C_{\mathcal{P},2}, \quad \|W\|_{L_\infty} \leq C_{\mathcal{P},2}.$$

The particular choice of $2 \leq p_\lambda, p_R, p_W < \infty$ will be specified later on. It will guarantee the existence and uniqueness of solutions for our PDE in some appropriate solution space. Moreover, our choice will ensure that the PDE solutions depend differentiable on the parameters. We denote the parameter space for D with $\mathcal{P}_D = \{D \in L_\infty : 0 < C_{\mathcal{P},1} \leq D \leq C_{\mathcal{P},2}\}$ and $\mathcal{P}_\lambda$, $\mathcal{P}_R$, and $\mathcal{P}_W$ analogously. The global parameter space is defined as

$$\mathcal{P} := \mathcal{P}_D \times \mathcal{P}_\lambda \times \mathcal{P}_R \times \mathcal{P}_W \tag{3.2.3}$$

equipped with the product norm of the individual L_p spaces. Observe that by the finiteness of U, the boundedness conditions of the individual parameters imply the boundedness of $\mathcal{P}$.

The analytic results presented in the following section are dealing with the full $(n+1)-$dimensional problem for N genes. However, the numerical results presented in Section 3.6 will deal with simplified models in 1 and 2 space dimensions. The simplified model aims at determining the biologically most relevant parameter W, which describes gene interaction as well as creation and absorption. Hence, for $N = 1$ and $\lambda = 0, D = 1, \phi(x) = x + \frac{1}{2}$ we obtain the test equation

$$u' - \Delta u - Wu = \frac{1}{2}.$$

This still poses a nonlinear inverse problem, which we will treat by Tikhonov regularization with sparsity constraints.

3.2.2 Function spaces and operators

Solution space and time derivatives

We will develop the solution theory for the well-known spaces of Bochner integrable functions. The general definitions and basic theory about these spaces can be found in [114, Ch. III.1–2].

Let us now fix the notation for our setting. We introduce the spaces

$$V_q = W_q^1(U, \mathbb{R}^N),$$
$$\mathcal{V}_{s,q} = L_s(0, T; V_q)$$

with $q > n$ and $s \in (1, \infty)$. The conjugate exponent q' is given by $1/q + 1/q' = 1$.

We call $u \in \mathcal{V}_{s,q}$ *differentiable* in time, if there exists $u' \in L_s(0, T; (V_{q'})')$, such that

$$\langle u', v\rangle = -\int_0^T \int_U uv' \,\mathrm{d}x \,\mathrm{d}t,$$

for all $v \in C_0^\infty([0, T], C^1(U, \mathbb{R}^N))$. u' is then the (*temporal*) *derivative* of u.

With these conventions we define the *generalized Sobolev space*

$$\mathcal{W}_s = \{u \in \mathcal{V}_{s,q} : u' \in L_s(0, T; (V_{q'})')\},$$
$$\|u\|_{\mathcal{W}_s} = \|u\|_{L_s(0,T;V_q)} + \|u'\|_{L_s(0,T;(V_{q'})')}.$$

The solution theory of (3.2.1) follows [67]. We fix $\mathfrak{q} \in (n, n + \varepsilon)$, with $\varepsilon = \varepsilon(U, C_{\mathcal{P},1}, C_{\mathcal{P},2})$ as in [67, Thm. 5.14] to ensure solvability of (3.2.1), see Section 3.2.3 for details. We set $s = \mathfrak{q}$ and use the simplified notation $\mathcal{V}_q := \mathcal{V}_{q,q}$. However, all results of this paper can be generalized to arbitrary $s \in (1, \infty)$.

The solution space we will use for (3.2.1) is then given by

$$\mathcal{W}_\mathfrak{q} = \{u \in \mathcal{V}_\mathfrak{q} : u' \in (\mathcal{V}_{\mathfrak{q}'})'\} \quad \text{where} \quad \|u\|_{\mathcal{W}_\mathfrak{q}} = \|u\|_{\mathcal{V}_\mathfrak{q}} + \|u'\|_{(\mathcal{V}_{\mathfrak{q}'})'}.$$

Bilinearform

Recall that the space of admissible parameters $\mathcal{P}$, defined in (3.2.3), is a non-open subset of a Cartesian product of L_p spaces that is bounded in L_∞. For any $\pi \in \mathcal{P}$ no L_p neighborhoods are contained in $\mathcal{P}$. Therefore, we need to clarify the meaning of differentiation with respect to π. We define the differentiation on non-open sets of vector spaces, following [75].

Definition 3.2.1. Let Z be a Banach space and Y a subset of a normed vector space X. A function $f : Y \to Z$ is called *strongly differentiable* at some $x \in Y$, if there is some $A(x) \in \mathcal{L}(X, Z)$, such that we have

$$\lim_{t\to 0} \sup_{\|e\|=t,\, e\in\mathcal{E}(x)} \left\| \frac{f(x+e) - f(x) - A(x)e}{\|e\|} \right\| = 0,$$

where the set of admissible displacement vectors at $x \in Y$ is

$$\mathcal{E}(x) = \{e : x + e \in Y\} \subset X.$$

In case the limit exists, $A(x)$ is called the *derivative of f at x*.

Note that this definition coincides with the usual Fréchet derivative if Y is an open subset of a normed vector space. We refer to [105] for a more detailed discussion of this setting.

We fix an arbitrary $\mathfrak{r} > \mathfrak{q}$ and consider p_λ according to

$$\frac{1}{p_\lambda} + \frac{1}{\mathfrak{r}} \leq \frac{1}{\mathfrak{q}}.$$

Then the elliptic part of our model in (3.2.1) defines the bilinear operator

$$\begin{aligned} \mathscr{A} &: \mathcal{P} \times \mathcal{W}_\mathfrak{q} \to (\mathcal{V}_{\mathfrak{q}'})', \\ \mathscr{A}(\pi, u)(v) &= \int_0^T \int_U \langle D\,\nabla u, \nabla v\rangle + \lambda\, u\, v \,\mathrm{d}x\,\mathrm{d}t. \end{aligned} \tag{3.2.4}$$

We introduce the notation $\mathscr{A}_\pi(\cdot)(\cdot) := \mathscr{A}(\pi, \cdot)(\cdot)$ and omit the dependency on π whenever it is clear from the context.

Since as usual a bounded linear operator coincides with its derivative, we obtain continuity and therefore continuous differentiability with respect to both input arguments u and π. Note, that by definition the differential operator $\frac{d}{dt} : \mathcal{W}_\mathfrak{q} \to (\mathcal{V}_{\mathfrak{q}'})'$, $u \mapsto u'$ is well-defined and bounded. Furthermore, these properties also hold if it is trivially extended to $\mathcal{P} \times \mathcal{W}_\mathfrak{q}$. Together this gives the following theorem.

Theorem 3.2.2. *The differential operator $\frac{d}{dt} + \mathscr{A} : \mathcal{P} \times \mathcal{W}_\mathfrak{q} \to (\mathcal{V}_{\mathfrak{q}'})'$ is well-defined and continuous.*

Nonlinear right-hand side

We want to prove the differentiability of the nonlinear right-hand side of (3.2.1). To do so we need to be able to embed $\mathcal{W}_\mathfrak{q}$ into spaces with higher integrability.

Theorem 3.2.3 ([1, Thm. III.4.10.2])**.** *For $r > \mathfrak{q}$ there exists a continuous embedding*

$$\mathcal{W}_\mathfrak{q} \hookrightarrow C([0,T], L_r(U, \mathbb{R}^N)).$$

Our analysis of the nonlinear right-hand side follows [3] and utilizes so-called superposition operators. With $\mathfrak{r} > \mathfrak{q}$ as before, our proofs rely on the specific choice

$$\frac{1}{p_R} + \frac{1}{p_W} + \frac{1}{\mathfrak{r}} \leq \frac{1}{\mathfrak{q}}.$$

Further the signal response functions Φ_i in (3.2.2) need to be smooth, globally Lipschitz continuous and globally bounded. In this sense the particular choice of Φ_i in (3.2.2) can be generalized to a larger class of right-hand sides.

Lemma 3.2.4 ([3, Thm. 3.8])**.** *Let the response functions Φ_i given by* (3.2.2)*. Then the map*

$$\begin{aligned} F : \mathcal{P} \times L_\mathfrak{r}(0,T; L_\mathfrak{r}(U,\mathbb{R}^N)) &\to L_\mathfrak{q}(0,T; L_\mathfrak{q}(U,\mathbb{R}^N)) \hookrightarrow (\mathcal{V}_{\mathfrak{q}'})', \\ (\pi, u) &\mapsto (F_i(\pi,u))_{i=1}^N := (R_i \Phi_i((Wu)_i))_{i=1}^N \end{aligned} \tag{3.2.5}$$

is continuous.

If it is clear from the context, we will omit the dependency of F in (3.2.5) on one parameter, denoting $F(\pi)$ or $F(u)$, respectively.

Theorem 3.2.5 ([3, Thm. 3.13, Rem. p. 105])**.** *The partial derivatives of F defined in* (3.2.5) *are Lipschitz continuous. Denoting $\pi_0 = (D_0, \lambda_0, R_0, W_0)$ they are given by*

$$\begin{aligned} \frac{\partial F}{\partial \pi}(\pi_0, u_0) : \mathcal{P} &\to L_\mathfrak{q}(0,T; L_\mathfrak{q}(U,\mathbb{R}^N)), \\ \pi &\mapsto (R_i \Phi_i((W_0 u_0)_i) + (R_0)_i \Phi_i'((W_0 u_0)_i)(Wu_0)_i)_{i=1}^N \end{aligned}$$

and

$$\begin{aligned} \frac{\partial F}{\partial u}(\pi_0, u_0) : L_\mathfrak{r}(0,T; L_\mathfrak{r}(U,\mathbb{R}^N)) &\to L_\mathfrak{q}(0,T; L_\mathfrak{q}(U,\mathbb{R}^N)), \\ u &\mapsto ((R_0)_i \Phi_i'((W_0 u_0)_i)(W_0)_{ij})_{i,j=1}^N u\,. \end{aligned}$$

3.2.3 Solvability

In order to define a suitable function space for the initial value of the PDE, we need the following embedding.

Theorem 3.2.6 ([4])**.** *Let $\sigma = 1 - 2/\mathfrak{q}$. Then there exists a continuous embedding*

$$\mathcal{W}_{\mathfrak{q}} \hookrightarrow C([0,T], G),$$

with G denoting the Besov space $B^{\sigma}_{\mathfrak{q},\mathfrak{q}}(U, \mathbb{R}^N)$.

With this, the weak formulation of our model PDE can be stated as follows.

Definition 3.2.7. A function $u \in \mathcal{V}_q$ is a *weak solution* of the PDE (3.2.1), iff

$$u \in \mathcal{W}_{\mathfrak{q}}: \quad u' + \mathscr{A}u = F(u) \quad \text{in} \quad (\mathcal{V}_{\mathfrak{q}'})', \quad u(0,\cdot) = u_0 \in G, \tag{3.2.6}$$

where $\mathscr{A}$ and F are defined in (3.2.4) and (3.2.5), respectively.

The main result of this section then reads as follows.

Theorem 3.2.8. *The Cauchy problem as stated in* (3.2.6) *has a unique weak solution.*

Proof. Consider the linearized problem

$$u \in \mathcal{W}_{\mathfrak{q}}: \quad u' + \mathscr{A}u = f \quad \text{in } (\mathcal{V}_{\mathfrak{q}'})', \quad u(0,\cdot) = u_0 \in G. \tag{3.2.7}$$

Then the results of [105] building on the main statements in [67, 4] imply that (3.2.7) has a unique solution that depends continuously on $f \in (\mathcal{V}_{\mathfrak{q}'})'$.

Next, we use the embedding from Theorem 3.2.6 and consider the map

$$\mathcal{B} : C([0,T], G) \to C([0,T], G)$$

which assigns to w the unique solution of (3.2.7) with right-hand side $f = F(w)$. By splitting $[0,T]$ into to sufficiently small subintervals we can obtain a contraction and apply Banach's fixed point theorem, see [59, pp. 500] for this classical technique. □

3.3 The control-to-state map

Knowledge of the properties of the control-to-state map $\mathscr{D} : \mathcal{P} \to \mathcal{W}_{\mathfrak{q}},\ \pi \mapsto u$, assigning to each tuple of parameters π the unique solution of (3.2.6) is the key for our analysis. In this section we summarize the needed results for the regularization scheme and steepest descent method we will employ later to solve the inverse problem. For detailed proofs we refer to the Ph.D. thesis [105].

3.3.1 Continuity and differentiability

In this section we present continuity and differentiability results and give an explicit formula for the derivative of the control-to-state map.

Our analysis relies on the auxiliary operator

$$\begin{aligned}\mathscr{C}: \quad & \mathcal{P} \times \mathcal{W}_{\mathfrak{q}} \to G \times (\mathcal{V}_{\mathfrak{q}'})', \\ & (\pi, u) \mapsto (u(0) - u_0, u' + \mathscr{A}_\pi u - F(\pi, u)).\end{aligned}$$

We equip the product spaces with the usual product norm. Then, by Theorem 3.2.6 and the assumptions on $\mathscr{A}$ and F, it follows that $\mathscr{C}$ is well defined.

For the next lemma we fix the first argument and show continuous differentiability with respect to the second argument.

Lemma 3.3.1. *Let $\pi_0 \in \mathcal{P}$ be fixed. The map $\mathscr{C}(\pi_0, \cdot) : \mathcal{W}_{\mathfrak{q}} \to G \times (\mathcal{V}_{\mathfrak{q}'})'$ is continuously differentiable and the derivative at any u is an isomorphism from $\mathcal{W}_{\mathfrak{q}}$ onto $G \times (\mathcal{V}_{\mathfrak{q}'})'$.*

Proof. Differentiability of the nonlinear part F was proved in Theorem 3.2.5. For the linear part it is a consequence of Theorem 3.2.2, as continuous linearity implies continuous differentiability. The continuous invertibility of the derivative is equivalent to the uniqueness, existence and stability that is guaranteed by solvability theory. □

By using similar arguments, one can also establish differentiability with respect to the first argument.

Lemma 3.3.2. *Let $u_0 \in \mathcal{W}_{\mathfrak{q}}$ be fixed. The map $\mathscr{C}(\cdot, u_0) : \mathcal{P} \to G \times (\mathcal{V}_{\mathfrak{q}'})'$ is continuously differentiable.*

In our setting, i.e., for subsets of non-complete normed vector spaces, a version of the implicit function theorem exists, we refer to [85]. Together with the last two lemmata this leads to the following statement.

Theorem 3.3.3. *The control-to-state map $\mathscr{D}$ is continuously differentiable. With a slight abuse of notation let $\pi_k = (D_k, \lambda_k, R_k, W_k)$, $k = 1, 2$ and $u_0 := \mathscr{D}(\pi_0)$. Then the derivative*

$$\frac{\partial \mathscr{D}}{\partial \pi}(\pi_0)(\pi_1) = -\left(\frac{\partial \mathscr{C}}{\partial u}(\pi_0, u_0)\right)^{-1} \circ \frac{\partial \mathscr{C}}{\partial \pi}(\pi_0, u_0)(\pi_1) =: v \tag{3.3.1}$$

of the control-to-state map $\mathscr{D}$ coincides with the solution v to the Cauchy problem

$$\begin{aligned} v' + \mathscr{A}_{\pi_0} v - R_0 \Phi'(W_0 u_0) W_0 v &= -\mathscr{A}_{\pi_1} u_0 + R_1 \Phi(W_0 u_0) + R_0 \Phi'(W_0 u_0) W_1 u_0, \\ v(0) &= 0. \end{aligned}$$

3.3.2 Adjoint of the derivative

The explicit formula for the derivative of the control-to-state map at some π_0 stated in Theorem 3.3.3, enables us to investigate further useful properties of $\mathscr{D}'$. We show local Lipschitz continuity of $\mathscr{D}'$ and give an explicit formula for its adjoint.

Lemma 3.3.4. *The map*

$$\left(\frac{\partial \mathscr{C}}{\partial u}(\cdot,\cdot)\right)^{-1} : \mathcal{P} \times \mathcal{W}_{\mathfrak{q}} \to \mathcal{L}(G \times (\mathcal{V}_{\mathfrak{q}'})', \mathcal{W}_{\mathfrak{q}})$$
$$(\pi_0, u_0) \mapsto \big((v_0, f) \mapsto \text{ solution of } v' + \mathscr{A}_{\pi_0} v - R_0 \Phi'(W_0 u_0) W_0 v = f, \; v(0) = v_0\big)$$

is uniformly continuous and locally Lipschitz continuous with uniform Lipschitz constant, i.e., each argument (π, u) is contained in a ball of uniform radius, such that the map is Lipschitz continuous with (globally) uniform Lipschitz constant on this ball.

Proof. The proof rests essentially on the Lipschitz continuity that was proved for the nonlinear right-hand side F in Theorem 3.2.5 and a theorem concerning the differentiability of operator inversion [10, Cor. 50.3]. For details we refer to [105]. □

Lemma 3.3.5. *The operator*

$$\frac{\partial \mathscr{C}}{\partial \pi} : \mathcal{P} \times \mathcal{W}_{\mathfrak{q}} \to \mathcal{L}(\mathcal{P}, G \times (\mathcal{V}_{\mathfrak{q}'})')$$
$$(\pi_0, u) \mapsto \big(\pi_1 \mapsto (0, \lambda_1 u - \operatorname{div}(D_1 \nabla u) - R_1 \Phi(W_0 u) - R_0 \Phi'(W_0 u) W_1 u)\big)$$

is Lipschitz continuous.

Proof. Similar to the proof the preceding lemma concerning concerning $\frac{\partial \mathscr{C}}{\partial u}$, the proof can be played back to the Lipschitz continuity of the nonlinearity F. For details we refer to [105]. □

The following theorem is an application of Lemmata 3.3.4 and 3.3.5.

Theorem 3.3.6. *The map $\mathscr{D}'$ is locally Lipschitz continuous in the sense of Lemma 3.3.4 and in particular uniformly continuous.*

Below, we will use the operator $(\mathscr{D}'(\pi_0))^*$ as part of an iterative method to compute the minimizer of the Tikhonov functional (3.1.1). To this end we need to derive an explicit expression for

$$(\mathscr{D}'(\pi_0))^* = -\left(\frac{\partial \mathscr{C}}{\partial \pi}(\pi_0, u)\right)^* \circ \left(\left(\frac{\partial \mathscr{C}}{\partial u}(\pi_0, u)\right)^{-1}\right)^*. \tag{3.3.2}$$

Remark 3.3.7. Concerning the inner part $((\frac{\partial \mathscr{C}}{\partial u}(\pi_0, u))^{-1})^*$ let $\mathscr{K}$ be defined by

$$\mathscr{K} : \mathcal{P} \times \mathcal{W}_{\mathfrak{q}} \to (\mathcal{V}_{\mathfrak{q}'})',$$
$$(\pi, u) \mapsto \big(v \mapsto \mathscr{A}_{\pi} v - R \Phi'(W u) W v\big).$$

Then a straightforward computation shows that $((\frac{\partial \mathscr{C}}{\partial u}(\pi_0, u))^{-1})^*$ maps $w \in (\mathcal{V}_{\mathfrak{q}'})'$ to the solution of the PDE problem

$$-v' + \mathscr{K}(\pi_0, u) v = w \text{ in } (\mathcal{V}_{\mathfrak{q}'})', \quad v(T) = 0.$$

Similarly, the adjoint of the operator

$$\frac{\partial \mathscr{C}}{\partial \pi}(\pi_0, u) : \mathcal{P} \to G \times (\mathcal{V}_{\mathfrak{q}'})'$$

is given by

$$(h_0, h) \mapsto (\langle \nabla h_i, \nabla u_i \rangle, h_i u_i, -h_i \cdot \Phi((W_0 u)_i), -h_i \cdot (R_0)_i \Phi'((W_0 u)_i) u^T)_{i=1}^N.$$

3.4 Regularization

3.4.1 Tikhonov regularization

In the applications we have in mind, we are given observed data y in some observation space $\mathcal{O}$. It is related to $\mathscr{D}(\pi)$ via a measurement operator $\mathcal{M} : \mathcal{W}_{\mathfrak{q}} \to \mathcal{O}$. A typical function space for data given by finitely many measurements at discrete times $0 = t_0 \leq \ldots \leq t_K = 1$, is given by

$$\mathcal{O} = L_2(\{t_i\}_{i=0}^K, L_2(U, \mathbb{R}^N)).$$

This model also allows to include deterministic noise in the data, i.e., if only noisy data y^δ with $\|y - y^\delta\| \leq \delta$ is available.

The existence of minimizers of Tikhonov functionals (3.1.1) as well as the related stability analysis have been studied in fairly general settings, cf. [76]. In our setting they are given in the following theorem. Its proof is a direct consequence of the analytic properties of $\mathscr{D}$ that we derived in the previous section. We refer to [105] for more details.

Theorem 3.4.1. *Let the Tikhonov functional be defined as*

$$\|\mathcal{M}\mathscr{D}(\pi) - y^\delta\|_{\mathcal{O}}^2 + \alpha J(\pi).$$

Assume there exists at least one exact solution to the noise-free equation

$$\mathcal{M}\mathscr{D}(\pi) = y$$

and that the penalty term $J : \mathcal{P} \to \mathbb{R}_+$ is lower semicontinuous with respect to an auxiliary L_∞-weak topology and has L_∞-weak* precompact level sets.*

Then for any such y^δ some minimizer of the Tikhonov functional can be found. Furthermore, for decreasing noise levels, i.e., $\delta_n \to 0$ and the parameter choice rule $\alpha = \alpha(\delta, y^\delta)$, such that

$$\alpha_n \to 0, \quad \frac{\delta_n^2}{\alpha_n} \to 0, \quad \textit{for} \quad n \to \infty,$$

the corresponding sequence of minimizers $(\pi_{\alpha_n})_{n \in \mathbb{N}}$ has a subsequence that converges $\mathcal{P}$-weakly to an exact solution of the noise-free equation.

Note that the auxiliary L_∞-weak* topology is solely used to verify the feasibility of our regularization procedure. The conditions of the theorem are fulfilled in all settings presented in this paper.

3.4.2 The generalized conditional gradient method

The major practical problem that remains is to compute a solution $\pi^\dagger = \pi(y^\delta)$ for a fixed regularization parameter α as in

$$\pi^\dagger = \arg\min_{\pi\in\mathcal{P}} \|\mathcal{M}\mathscr{D}(\pi) - y^\delta\|_{\mathcal{O}}^2 + \alpha J(\pi). \tag{3.4.1}$$

To this end we will employ a generalized conditional gradient method. This methods is a well-established tool for computing the Tikhonov minimizer of (3.4.1). For linear forward operators $\mathscr{D}$ it was analyzed in the paper by [46]. Its generalization to the nonlinear case has been discussed, e.g., in [15, 13, 102], at least in a Hilbert space setting. However, we are dealing with Banach space topology rather than Hilbert space topology. In the following we adapt the statements of [15] to our setting.

We consider a Banach space X and two functionals $\mathscr{E}, \mathscr{F} : X \to \mathbb{R}_0^+ \cup \infty$. In addition to the usual norm topology on X we introduce some topology τ on X, for which norm bounded sets are τ precompact.

The abstract goal is then to solve

$$\arg\min_{v\in X} \mathscr{E}(v) + \mathscr{F}(v). \tag{3.4.2}$$

In this setting we make the following assumptions.

Assumption 3.4.2. For $\mathscr{E}$ we assume continuous differentiability. $\mathscr{F}$ does not need to be differentiable, but needs to satisfy

C_1. $\mathscr{F}(x) < \infty$ for some $x \in X$,

C_2. $\mathscr{F}$ is convex,

C_3. $\mathscr{F}$ is sequentially τ lower semicontinuous, i.e.,

$$\mathscr{F}(x) \le \liminf_{n\to\infty} \mathscr{F}(x_n), \quad \text{whenever } x = \lim_{n\to\infty} x_n,$$

C_4. $\mathscr{F}$ is coercive, i.e., $\mathscr{F}(x_n) \to \infty$, whenever $\|x_n\| \to \infty$,

C_5. the problem $\arg\min_{v\in X} \mathscr{E}'(x)(v) + \mathscr{F}(v)$ has some solution,

C_6. $\mathscr{F}$ has weakly compact sublevel sets.

Remark 3.4.3. Two details of this approach demand closer attention. The first is the compactness requirement (C_6) on the sub-level sets of the penalty term $\mathscr{F}$. This assumption is satisfied for a weighted ℓ_q penalty term, $1 \leq q < 2$, if the weights are bounded away from 0, see e.g. [63]. The second aspect is the uniform continuity of the derivative of $\mathscr{E}$, which is covered by the analysis of the control-to-state operator in Section 3.3.

The generalized conditional gradient method (**GCGM**) that we will utilize to compute an approximation of the minimizer of (3.4.2) is given as follows.

Algorithm 3.4.4 GCGM

1: Choose $x_0 \in X$, such that $\mathscr{F}(x_0) < \infty$;
2: Determine $v_n \in X$ by

$$v_n = \arg\min_{v \in X} \mathscr{E}'(x_n)(v) + \mathscr{F}(v); \tag{3.4.3}$$

3: Determine step size $s_n \in [0,1]$ via

$$s_n = \arg\min_{s \in [0,1]} \mathscr{E}(x_n + s(v_n - x_n)) + \mathscr{F}(x_n + s(v_n - x_n)); \tag{3.4.4}$$

4: Put $x_{n+1} = x_n + s_n(v_n - x_n)$. Return to step 2.

3.4.3 Iterated soft shrinkage

In this section we present a specific choice for $\mathscr{E}$ and $\mathscr{F}$ that allows the reformulation of the minimization in (3.4.3) by the application of a soft shrinkage operator. In particular, this allows an efficient numerical treatment of the problem. Further, this explains why the reformulated algorithm is often called iterated soft shrinkage algorithm (**ISTA**).

We begin by considering $\mathcal{P}$ as a subset of

$$\mathscr{G} = \left(L_2([0,T] \times U, \mathbb{R}^N)\right)^3 \times L_2([0,T] \times U, \mathbb{R}^{N \times N}),$$

and fix a biorthogonal wavelet Riesz basis ($\Psi = \{\psi_\nu : \nu \in \mathcal{J}\}, \tilde{\Psi} = \{\tilde{\psi}_\nu : \nu \in \mathcal{J}\}$) for $\mathscr{G}$. A detailed description can be found in Section 3.5.1.

The connection between the minimization problem (3.4.1) induced by Tikhonov regularization and (3.4.3) is then given by the choice (compare [15, pp. 185])

$$\begin{aligned} \mathscr{E}(\pi) &:= \frac{1}{2}\|\mathcal{M}\mathscr{D}(\pi) - y^\delta\|_{\mathcal{O}}^2 - \frac{\sigma}{2}\|\pi\|_{\mathscr{G}}^2, \\ \mathscr{F}(\pi) &:= \frac{\sigma}{2}\|\pi\|_{\mathscr{G}}^2 + \alpha \sum_{\nu \in \mathcal{J}} w_\nu |\langle \pi, \tilde{\psi}_\nu \rangle|^q. \end{aligned}$$

where $w_\nu \geq w_0 > 0$ and $\sigma > 0$. This choice satisfies Assumption 3.4.2, cf. Remark 3.4.3. The minimization problem (3.4.1) then reads as

$$\pi^\dagger = \arg\min_{\pi\in\mathcal{P}} \frac{1}{2}\|\mathcal{M}\mathscr{D}(\pi) - y^\delta\|_{\mathcal{O}}^2 + \alpha \sum_{\nu\in\mathcal{J}} w_\nu |\langle \pi, \tilde{\psi}_\nu\rangle|^q. \tag{3.4.5}$$

The minimization problem in the second step of the **GCGM** algorithm is given by

$$\begin{aligned} v_n = \arg\min_{\pi\in\mathcal{P}} \sum_{\nu\in\mathcal{J}} &|\langle (\mathscr{D}'(\pi_n))^*\mathcal{M}^*(\mathcal{M}\mathscr{D}(\pi_n) - y^\delta) - \sigma\pi_n + \frac{\sigma}{2}\pi, \tilde{\psi}_\nu\rangle|^2 \\ &+ \alpha w_\nu |\langle \pi, \tilde{\psi}_\nu\rangle|^q, \end{aligned}$$

under the assumption that $(\mathscr{D}'(\pi_n))^*\mathcal{M}^*(\mathcal{M}\mathscr{D}(\pi_n) - y^\delta)$ is an element of $\mathscr{G}$.

The minimizer of such a functional combining an ℓ_2-norm and a weighted ℓ_q-norm can be directly computed using a soft thresholding operation, see [18, 46]. It holds that

$$(\langle v_n, \tilde{\psi}_\nu\rangle)_{\nu\in\mathcal{J}} = \mathbf{S}_{\frac{\alpha\mathbf{w}}{\sigma},q}((\langle \pi_n - \frac{1}{\sigma}(\mathscr{D}'(\pi_n))^*\mathcal{M}^*(\mathcal{M}\mathscr{D}(\pi_n) - y^\delta), \tilde{\psi}_\nu\rangle)_{\nu\in\mathcal{J}}), \tag{3.4.6}$$

where $\mathbf{S}_{\frac{\alpha\mathbf{w}}{\sigma},q}$ is a shrinkage operator that applies to each coefficient the shrinkage maps $S_{\frac{\alpha w_\nu}{\sigma},q}$. These maps are defined by

$$S_{\frac{\alpha w_\nu}{\sigma},q}(x) = \begin{cases} \operatorname{sgn}(x)[|x| - \frac{\alpha w_\nu}{\sigma}]_+, & q = 1, \\ G^{-1}_{\frac{\alpha w_\nu}{\sigma},q}(x), & q > 1, \end{cases} \tag{3.4.7}$$

where $G_{\frac{\alpha w_\nu}{\sigma},q}(x) = x + \frac{\alpha w_\nu}{\sigma} q \operatorname{sgn}(x)|x|^{q-1}$.

Remark 3.4.5.

1. In order to utilize an expansion with respect to Ψ, we need to ensure that $(\mathscr{D}'(\pi_n))^*\mathcal{M}^*(\mathcal{M}\mathscr{D}(\pi_n) - y^\delta)$ lies in $\mathscr{G}$ in each step of the iteration. Since the adjoint operator $\mathscr{D}'(\pi)^*$ maps to the dual space of $\mathcal{P}$, this is property automatically holds for all parameters but D by the L_p-maximal regularity theory we used.

2. If we choose σ large enough, it is possible to omit the line search in the third step of Algorithm 3.4.4 and to choose $s_n = 1$, $n \geq 0$, in (3.4.4). We refer to [13, Lem. 2.4] for details.

3.5 Discretization of the model PDE

In this section, we briefly explain how to apply adaptive wavelet methods for the numerical solution of the model PDE (3.2.6). First of all, in Section 3.5.1, we recall the wavelet setting. Then, in Section 3.5.2, we discuss adaptive wavelet schemes for elliptic problems. Finally, in Section 3.5.3, we are concerned with generalizations to parabolic equations.

3.5.1 Wavelets

Let us briefly recall the wavelet setting as far as it is needed for our purposes. We will not go into construction details and confine the discussion to the basic facts. For the anisotropic tensor wavelet construction in arbitrary dimensions we follow the paper [47].

We assume a univariate wavelet collection $\Psi = \{\psi_\nu : \nu \in \mathcal{J}\}$ on the unit interval $\mathcal{I} := (0,1)$ is available. The dual basis is denoted by $\tilde{\Psi} = \{\tilde{\psi}_\nu : \nu \in \mathcal{J}\}$. The indices $\nu \in \mathcal{J}$ encode several types of information, namely the *level*, denoted with $|\nu|$, and the spatial location.

For some fixed $t > 0$ we make the following assumptions on the univariate wavelets.

Assumption 3.5.1. Wavelet assumptions

P_1. $\{\psi_\nu : \nu \in \mathcal{J}\}$ is a Riesz basis for $L_2(\mathcal{I})$;

P_2. $\{2^{-|\nu|t}\psi_\nu : \nu \in \mathcal{J}\}$ is a Riesz basis for $W_2^t(\mathcal{I})$.

Furthermore we assume that for some $\mathbb{N} \ni d > t$

P_3. $|\langle \tilde{\psi}_\nu, u\rangle_{L_2(\mathcal{I})}| \lesssim 2^{-|\nu|d}|u|_{W_2^d(\operatorname{supp}\tilde{\psi}_\nu)}, \quad u \in W_2^d(\mathcal{I})$;

P_4. $\varrho = \sup\limits_{\nu\in\mathcal{J}} 2^{|\nu|}\max(\operatorname{diam}\operatorname{supp}\psi_\nu, \operatorname{diam}\operatorname{supp}\tilde{\psi}_\nu)$
$\eqsim \inf\limits_{\nu\in\mathcal{J}} 2^{|\nu|}\max(\operatorname{diam}\operatorname{supp}\psi_\nu, \operatorname{diam}\operatorname{supp}\tilde{\psi}_\nu)$,

P_5. $\sup\limits_{j,k\in\mathbb{N}_0} \#\big\{\nu \in \mathcal{J} : |\nu| = j$ and
$[k2^{-j}, (k+1)2^{-j}] \cap (\operatorname{supp}\tilde{\psi}_\nu \cup \operatorname{supp}\psi_\nu) \neq \emptyset\big\} < \infty$.

The properties (P_4) and (P_5) will be referred to by saying that both primal and dual wavelets are *localized* or *locally finite*, respectively. Denoting the unit cube for $n \in \mathbb{N}$ with $\square := \mathcal{I}^n$, the equalities

$$L_2(\square) = \bigotimes_{i=1}^n L_2(\mathcal{I})$$

and

$$W_2^t(\square) = W_2^t(\mathcal{I}) \otimes L_2(\mathcal{I}) \otimes \cdots \otimes L_2(\mathcal{I}) \cap \cdots \cap L_2(\mathcal{I}) \otimes \cdots \otimes L_2(\mathcal{I}) \otimes W_2^t(\mathcal{I})$$

hold.

The *anisotropic tensor product wavelet* collection

$$\boldsymbol{\Psi} := \Big\{\psi_{\boldsymbol{\nu}} := \psi_{\nu_1} \otimes \cdots \otimes \psi_{\nu_n} : \boldsymbol{\nu} \in \boldsymbol{\mathcal{J}} := \prod_{i=1}^n \mathcal{J}\Big\},$$

and its renormalized version $\{(\sum_{i=1}^n 4^{t|\nu_i|})^{-1/2}\psi_{\boldsymbol{\nu}} : \boldsymbol{\nu} \in \boldsymbol{\mathcal{J}}\}$ are Riesz bases for $L_2(\square)$ and $W_2^t(\square)$, respectively, i.e., (P_1) and (P_2) hold for the multivariate case. The collection that is dual to $\boldsymbol{\Psi}$ reads as

$$\tilde{\boldsymbol{\Psi}} := \Big\{\tilde{\psi}_{\boldsymbol{\nu}} := \tilde{\psi}_{\nu_1} \otimes \cdots \otimes \tilde{\psi}_{\nu_n} : \boldsymbol{\nu} \in \boldsymbol{\mathcal{J}}\Big\}.$$

For $\boldsymbol{\nu} \in \boldsymbol{\mathcal{J}}$, we set $|\boldsymbol{\nu}| = (|\nu_1|, \ldots, |\nu_n|)$.

It is one of the most important advantages of anisotropic tensor wavelets that they give rise to dimension independent approximation rates, provided that the object one wants to approximate has sufficient smoothness in the weighted Sobolev scale. For $\theta \geq 0$, the *weighted Sobolev space* $H^d_\theta(\mathcal{I})$ is defined as the space of all measurable functions u on $\mathcal{I}$ for which the norm

$$\|u\|^2_{H^d_\theta(\mathcal{I})} := \sum_{j=0}^{d} \int_{\mathcal{I}} |x^\theta (1-x)^\theta u^{(j)}(x)|^2 \,\mathrm{d}x$$

is finite. For $m \in \{0, \ldots, \lfloor t \rfloor\}$ we will consider the weighted Sobolev space

$$H^d_{m,\theta}(\square) := \bigcap_{p=1}^{n} \bigotimes_{i=1}^{n} H^d_{\theta - \delta_{pi}\min(m,\theta)}(\mathcal{I}),$$

equipped with the norm

$$\|u\|^2_{H^d_{m,\theta}(\square)} := \sum_{p=1}^{n} \|u\|^2_{\bigotimes_{i=1}^n H^d_{\theta-\delta_{pi}\min(m,\theta)}(\mathcal{I})}.$$

The mentioned dimension independent approximation result is then given by the following theorem.

Theorem 3.5.2 ([47, Thm. 4.3])**.** *For any $\theta \in [0, d)$, there exist a (nested) sequence $(\boldsymbol{\mathcal{J}}_M)_{M\in\mathbb{N}} \subset \boldsymbol{\mathcal{J}}$ with $\#\boldsymbol{\mathcal{J}}_M \eqsim M$, such that for all $u \in H^d_{m,\theta}(\square) \cap W_2^m(\square)$*

$$\inf_{v \in \operatorname{span}\{\psi_{\boldsymbol{\nu}} : \boldsymbol{\nu} \in \boldsymbol{\mathcal{J}}_M\}} \|u - v\|_{W_2^m(\square)} \lesssim M^{-(d-m)} \|u\|_{H^d_{m,\theta}(\square)},$$

holds. For $m = 0$, $M^{-(d-m)}$ should be read as $(\log \#M)^{(n-1)(\frac{1}{2}+d)} M^{-d}$.

Remark 3.5.3. (i) The theory in this section remains valid if essential boundary conditions are considered.

(ii) The anisotropic tensor wavelet construction differs from standard *isotropic* tensor wavelet constructions by the fact that wavelets on different levels are tensorized with each other, leading to rectangular and highly anisotropic supports.

(iii) Suitable constructions of isotopic wavelets on domains can be found, e.g., in [42, 43, 44, 17]. We also refer to [24] for a detailed discussion. A generalized construction of anisotropic tensor wavelets on complex domains is developed in [20]. The dimension independent approximation result from Theorem 3.5.2 remains valid.

3.5.2 Adaptive wavelet schemes for elliptic problems

In this section, we briefly recall how wavelets can be used to treat elliptic operator equations of the form

$$\mathcal{A}u = f, \tag{3.5.1}$$

where we will assume $\mathcal{A}$ to be a boundedly invertible operator from some Hilbert space $\mathcal{H}$ into its normed dual $\mathcal{H}'$, i.e.,

$$\|\mathcal{A}u\|_{\mathcal{H}'} \sim \|u\|_{\mathcal{H}}, \quad u \in \mathcal{H}.$$

We shall only discuss the basic ideas, for further information, the reader is referred to [33, 26, 27]. In applications $\mathcal{H}$ is typically a Sobolev space $W_2^t(\Omega)$ on some domain $\Omega \subset \mathbb{R}^n$. We shall mainly focus on the special case where

$$a(v, w) := \langle \mathcal{A}v, w \rangle$$

defines a *symmetric* bilinear form on $\mathcal{H}$ which is *elliptic* in the sense that

$$a(v, v) \sim \|v\|_{\mathcal{H}}^2. \tag{3.5.2}$$

Usually, operator equations of the form (3.5.1) are solved by a Galerkin scheme, i.e., one defines an increasing sequence of finite-dimensional approximation spaces

$$S_{\Lambda_l} := \operatorname{span}\{\eta_\mu : \mu \in \Lambda_l\},$$

where $S_{\Lambda_l} \subset S_{\Lambda_{l+1}}$, and projects the problem onto these spaces, i.e.,

$$\langle \mathcal{A}u_{\Lambda_l}, v \rangle = \langle f, v \rangle \quad \text{for all} \quad v \in S_{\Lambda_l}.$$

To compute the actual Galerkin approximation, one has to solve a linear system

$$\mathbf{G}_{\Lambda_l}\mathbf{c}_{\Lambda_l} = \mathbf{f}_{\Lambda_l}, \quad \mathbf{G}_{\Lambda_l} = (\langle \mathcal{A}\eta_{\mu'}, \eta_\mu \rangle)_{\mu,\mu' \in \Lambda_l}, \quad \mathbf{f}_{\Lambda_l} = (\langle f, \eta_\mu \rangle)_{\mu \in \Lambda_l}.$$

Then the question arises how to choose the approximation spaces in a suitable way, for doing that in a somewhat clumsy fashion would yield a very inefficient scheme. One natural idea would be to use an *adaptive* scheme, i.e., an updating strategy which essentially consists of the following three steps:

solve	–	estimate	–	refine
$\mathbf{G}_{\Lambda_l}\mathbf{c}_{\Lambda_l} = \mathbf{f}_{\Lambda_l}$		$\|u - u_{\Lambda_l}\| =?$ a posteriori error estimator		add functions if necessary.

Already the second step is highly nontrivial since the exact solution u is unknown, so that clever a posteriori error estimators are needed. Then another challenging task is to show that the refinement strategy leads to a convergent scheme and to estimate its

order of convergence, if possible. In recent years, it has been shown that both tasks can be solved if wavelets are used as basis functions for the Galerkin scheme as we shall now explain.

The first step is to transform (3.5.1) into a discrete problem. By using the properties (P_1) and (P_2) of the multivariate wavelet basis (3.5.1) is equivalent to

$$\mathbf{A}\mathbf{u} = \mathbf{f} \tag{3.5.3}$$

where

$$\mathbf{A} := \mathbf{D}^{-1}\langle \mathcal{A}\mathbf{\Psi}, \mathbf{\Psi}\rangle^T\mathbf{D}^{-1}, \quad \mathbf{u} := \mathbf{D}\mathbf{c}, \quad u = \mathbf{c}^T\mathbf{\Psi}, \quad \mathbf{f} := \mathbf{D}^{-1}\langle f, \mathbf{\Psi}\rangle^T,$$

and

$$\mathbf{D} := \left(\left(\sum_{i=1}^{n} 4^{t|\nu_i|}\right)^{-1/2} \delta_{\boldsymbol{\nu},\boldsymbol{\nu}'}\right)_{\boldsymbol{\nu},\boldsymbol{\nu}'\in\boldsymbol{\mathcal{J}}}$$

is a diagonal scaling matrix.

Now (3.5.2) implies that

$$\|\mathbf{A}\|_{\mathcal{L}(\ell_2(\boldsymbol{\mathcal{J}}))} < \infty, \quad \|\mathbf{A}^{-1}\|_{\mathcal{L}(\ell_2(\boldsymbol{\mathcal{J}}))} < \infty,$$

and the computation of the Galerkin approximation amounts to solving the system

$$\mathbf{A}_\Lambda \mathbf{u}_\Lambda = \mathbf{f}_\Lambda := \mathbf{f}|_\Lambda, \qquad \mathbf{A}_\Lambda := (\mathbf{D}^{-1}\langle \mathcal{A}\mathbf{\Psi}, \mathbf{\Psi}\rangle^T\mathbf{D}^{-1})|_\Lambda.$$

Now, ellipticity (3.5.2) and the Riesz property yield

$$\|\mathbf{u} - \mathbf{u}_\Lambda\|_{\ell_2(\boldsymbol{\mathcal{J}})} \sim \|\mathbf{A}(\mathbf{u} - \mathbf{u}_\Lambda)\|_{\ell_2(\boldsymbol{\mathcal{J}})} \sim \|\mathbf{f} - \mathbf{A}\mathbf{u}_\Lambda\|_{\ell_2(\boldsymbol{\mathcal{J}})} \sim \|\mathbf{r}_\Lambda\|_{\ell_2(\boldsymbol{\mathcal{J}})},$$

so that the ℓ_2-norm of the *residual* $\mathbf{r}_\Lambda$ serves as an a posteriori error estimator. Each individual coefficient $(\mathbf{r}_\Lambda)_{\boldsymbol{\nu}}$ can be viewed as a local error indicator. Therefore a natural adaptive strategy would consist in catching the bulk of the residual, i.e., to choose the new index set $\hat{\Lambda}$ such that

$$\|\mathbf{r}_\Lambda|_{\hat{\Lambda}}\|_{\ell_2(\boldsymbol{\mathcal{J}})} \geq \zeta\|\mathbf{r}_\Lambda\|_{\ell_2(\boldsymbol{\mathcal{J}})}, \quad \text{for some} \quad \zeta \in (0,1).$$

However, such a scheme would not be implementable since the residual involves infinitely many coefficients. To transform this idea into an implementable scheme, the following three subroutines can be utilized:

- $\mathbf{RHS}[\varepsilon, \mathbf{g}] \to \mathbf{g}_\varepsilon$: determines for $\mathbf{g} \in \ell_2(\boldsymbol{\mathcal{J}})$ a finitely supported $\mathbf{g}_\varepsilon \in \ell_2(\boldsymbol{\mathcal{J}})$ such that
$$\|\mathbf{g} - \mathbf{g}_\varepsilon\|_{\ell_2(\boldsymbol{\mathcal{J}})} \leq \varepsilon.$$
- $\mathbf{APPLY}[\varepsilon, \mathbf{A}, \mathbf{v}] \to \mathbf{w}_\varepsilon$: determines for a finitely supported $\mathbf{v} \in \ell_2(\boldsymbol{\mathcal{J}})$ a finitely supported $\mathbf{w}_\varepsilon$ such that
$$\|\mathbf{A}\mathbf{v} - \mathbf{w}_\varepsilon\|_{\ell_2(\boldsymbol{\mathcal{J}})} \leq \varepsilon.$$

- **COARSE**$[\varepsilon, \mathbf{v}] \to \mathbf{v}_\varepsilon$: determines for a finitely supported $\mathbf{v} \in \ell_2(\boldsymbol{\mathcal{J}})$ a finitely supported $\mathbf{v}_\varepsilon \in \ell_2(\boldsymbol{\mathcal{J}})$ with at most M significant coefficients, such that

$$\|\mathbf{v} - \mathbf{v}_\varepsilon\|_{\ell_2(\boldsymbol{\mathcal{J}})} \leq \varepsilon. \tag{3.5.4}$$

  Moreover, $M \lesssim M_{\min}$ holds, $M_{\min}$ being the minimal number of entries for which (3.5.4) is valid.

Then, employing the key idea outlined above, the resulting fundamental algorithm reads as follows:

Algorithm 3.5.4 SOLVE$[\varepsilon, \mathbf{A}, \mathbf{f}] \to \mathbf{u}_\varepsilon$

$\Lambda_0 := \emptyset$; $\mathbf{r}_{\Lambda_0} := \mathbf{f}$; $\varepsilon_0 := \|\mathbf{f}\|_{\ell_2(\boldsymbol{\mathcal{J}})}$; $j := 0$; $u_0 := 0$;
while $\varepsilon_j > \varepsilon$ **do**
  $\varepsilon_{j+1} := 2^{-(j+1)}\|\mathbf{f}\|_{\ell_2(\boldsymbol{\mathcal{J}})}$; $\Lambda_{j,0} := \Lambda_j$; $\mathbf{u}_{j,0} := \mathbf{u}_j$;
  for $l = 1, ..., L$ **do**
    Compute Galerkin approximation $\mathbf{u}_{\Lambda_{j,l-1}}$ for $\Lambda_{j,l-1}$;
    Compute
      $\tilde{\mathbf{r}}_{\Lambda_{j,l-1}} := \mathbf{RHS}[C_1^{\mathrm{tol}}\varepsilon_{j+1}, \mathbf{f}] - \mathbf{APPLY}[C_1^{\mathrm{tol}}\varepsilon_{j+1}, \mathbf{A}, \mathbf{u}_{\Lambda_{j,l-1}}]$;
    Compute smallest set $\Lambda_{j,l}$,
      such that, $\|\tilde{\mathbf{r}}_{\Lambda_{j,l-1}}|_{\Lambda_{j,l}}\|_{\ell_2(\boldsymbol{\mathcal{J}})} \geq \frac{1}{2}\|\tilde{\mathbf{r}}_{\Lambda_{j,l-1}}\|_{\ell_2(\boldsymbol{\mathcal{J}})}$;
  end for
  COARSE$[C_2^{\mathrm{tol}}\varepsilon_{j+1}, \mathbf{u}_{\Lambda_{j,L}}] \to (\Lambda_{j+1}, \mathbf{u}_{j+1})$;
  $j := j + 1$;
end while

Remark 3.5.5. (i) We shall not discuss in detail the concrete numerical realization of the three fundamental subroutines. The subroutine **COARSE** consists of a thresholding step, whereas **RHS** essentially requires the computation of a best N-term approximation. The most complicated building block is **APPLY**. The subroutine can be realized and optimality of the resulting algorithm can be proved up to order s^*, if the stiffness matrix $\mathbf{A}$ is s^*-*compressible*, i.e., there exists matrices $\mathbf{A}^J$, $J \in \mathbb{N}$, with

$$\|\mathbf{A} - \mathbf{A}^J\|_{\mathcal{L}(\ell_2(\boldsymbol{\mathcal{J}}))} \lesssim M_J^{-s}$$

for all $s < s^*$, where $\mathbf{A}^J$ has $\mathcal{O}(M_J)$ nontrivial entries per column.

For elliptic operators with Schwartz kernels, the cancellation property of wavelets can be used to establish compressibility. For isotropic wavelets, the reader is referred to [26, 27, 118]. The anisotropic case using L_2-orthogonal wavelets has been dealt with in [51].

(ii) In Algorithm 3.5.4, c_1, c_2 and c_3 denote some suitably chosen constants whose concrete values depend on the problem at hand. Also the parameter L has to be chosen in a suitable way. We refer again to [26] for details.

Concerning the realization of **APPLY** in the biorthogonal anisotropic setting of Section 3.5.1 we generalize the findings in [51]. There, a stiffness matrix induced by the bilinear form

$$a(u,v) := \int_{\square} c_0 uv + \sum_{k=1}^{n} c_k \partial_k u \partial_k v \,\mathrm{d}x = f(v), \tag{3.5.5}$$

with constant coefficients $c_0 \geq 0$ and $c_k > 0, k = 1, \ldots, n$, is considered. Additional mild assumptions have to be imposed on the one-dimensional wavelets.

Assumption 3.5.6. Wavelet assumptions for compressibility of $\mathbf{A}$

P_6. $\|\psi_\nu\|_{L_\infty(\mathcal{I})} \lesssim 2^{1/2|\nu|}, \quad \|\dot{\psi}_\nu\|_{L_\infty(\mathcal{I})} \lesssim 2^{3/2|\nu|}$ for all $\nu \in \mathcal{J}$,

P_7. for all $x \in \mathcal{I}$ the cardinality of

$$\{\nu \in \mathcal{J} : |\nu| = j \text{ and } x \in \operatorname{conv}\{\operatorname{supp}\psi_\nu\}\}$$

is bounded independently of $j \in \mathbb{N}$,

P_8. ψ_ν is piecewise polynomial of order d with singular support uniformly bounded in $|\nu|$,

P_9. ψ_ν has d vanishing moments if $\operatorname{supp}\psi_\nu \subset (0,1)$.

The proof of compressibility consists of an application of the Schur lemma, together with estimates for the absolute value of $a_{\boldsymbol{\nu},\boldsymbol{\mu}} := a(\psi_{\boldsymbol{\nu}}, \psi_{\boldsymbol{\mu}})$ and for the cardinality of

$$\Lambda^l_{\boldsymbol{\nu}} := \{\boldsymbol{\mu} \in \boldsymbol{\mathcal{J}} : a_{\boldsymbol{\nu},\boldsymbol{\mu}} \neq 0 \text{ and } \||\boldsymbol{\nu}| - |\boldsymbol{\mu}|\|_1 = l\}, \quad \boldsymbol{\nu} \in \boldsymbol{\mathcal{J}},\, l \in \mathbb{N}.$$

The first estimate is not influenced by the change of setting. We just cite the result:

$$|a_{\boldsymbol{\nu},\boldsymbol{\mu}}| \lesssim 2^{-1/2\||\boldsymbol{\nu}|-|\boldsymbol{\mu}|\|_1} \|\psi_{\boldsymbol{\nu}}\|_{W_2^1(\square)} \|\psi_{\boldsymbol{\mu}}\|_{W_2^1(\square)}. \tag{3.5.6}$$

In our biorthogonal setting the following statement is true.

Lemma 3.5.7. *Let $\boldsymbol{\nu} \in \boldsymbol{\mathcal{J}}$ be fixed. It holds that*

$$\#\Lambda^l_{\boldsymbol{\nu}} \lesssim l^{n-1}$$

and consequently

$$\#\bigcup_{l \leq J} \Lambda^l_{\boldsymbol{\nu}} \lesssim J^n.$$

All constants are uniformly bounded in $\||\boldsymbol{\nu}|\|_1$.

Proof. We begin by considering the one-dimensional case $n = 1$ and fix $\nu \in \mathcal{J}$. We show that $\#\{\mu \in \mathcal{J} : |\mu| = l \text{ and } \langle \psi_\nu, \psi_\mu \rangle_{W_2^1(\mathcal{I})} \neq 0\}$ is bounded uniformly in $|\nu|$ and l.

The number of boundary adapted wavelets, i.e., those with

$$\{0, 1\} \cap \operatorname{supp} \psi_\mu \neq \emptyset,$$

is bounded independently of $|\mu|$.

For the interior, i.e., $\mu \in \mathcal{J}$ with $\operatorname{supp} \psi_\nu \cup \operatorname{supp} \psi_\mu \subset \mathcal{I}$, we use (P_8) and (P_9) to conclude that $\langle \psi_\nu, \psi_\mu \rangle_{L_2(\mathcal{I})} = 0$ if the intersection of the singular support of ψ_ν and $\operatorname{supp} \psi_\mu$ is empty. Partial integration and (P_9) yield the same result for the inner product of the derivatives $\langle \dot{\psi}_\nu, \dot{\psi}_\mu \rangle_{L_2(\mathcal{I})}$. The number of wavelets $\psi_\mu, |\mu| = l$, with nontrivial intersection of the singular support of ψ_ν and $\operatorname{supp} \psi_\mu$ is bounded by (P_7), uniformly in $|\nu|$.

For higher dimensions note, that the cardinality of $\{\mathbf{k} \in \mathbb{N}^n : \|\mathbf{k}\|_1 = l\}$ is dominated by l^{n-1} and consequently the cardinality of $\{\mathbf{k} \in \mathbb{N}^n : \|\mathbf{k}\|_1 \leq J\}$ is dominated by J^n. □

Theorem 3.5.8. *Let* $\mathbf{A}$ *as in* (3.5.3) *be induced by* (3.5.5) *and*

$$(\mathbf{A}^J)_{\boldsymbol{\nu},\boldsymbol{\mu}} := \begin{cases} (\mathbf{A})_{\boldsymbol{\nu},\boldsymbol{\mu}}, & \| |\boldsymbol{\nu}| - |\boldsymbol{\mu}| \|_1 < J, \\ 0, & \textit{otherwise.} \end{cases}$$

Then $\mathbf{A}^J$ *has* $\mathcal{O}(J^n)$ *nontrivial entries per column and*

$$\|\mathbf{A} - \mathbf{A}^J\|_{\mathcal{L}(\ell_2(\boldsymbol{\mathcal{J}}))} \lesssim 2^{-\frac{J}{2}} J^n.$$

Thus $\mathbf{A}$ *is compressible with* $s^* = \infty$.

Proof. The proof is an application of the Schur lemma to the matrix

$$\mathbf{B}^J := \mathbf{A} - \mathbf{A}^J.$$

Let $\boldsymbol{\nu} \in \boldsymbol{\mathcal{J}}$ be fixed. We use (3.5.6) and Lemma 3.5.7 to estimate

$$\sum_{\boldsymbol{\mu} \in \boldsymbol{\mathcal{J}}} |(\mathbf{B})_{\boldsymbol{\nu},\boldsymbol{\mu}}| = \sum_{l=J}^{\infty} \sum_{\boldsymbol{\mu} \in \Lambda_{\boldsymbol{\nu}}^l} |(\mathbf{D}^{-1})_{\boldsymbol{\nu},\boldsymbol{\nu}} (\mathbf{D}^{-1})_{\boldsymbol{\mu},\boldsymbol{\mu}} a_{\boldsymbol{\nu},\boldsymbol{\mu}}| \lesssim \sum_{l=J+1}^{\infty} l^{n-1} 2^{-\frac{l}{2}},$$

where the last sum is convergent and dominated by $2^{-\frac{J}{2}} J^n$. □

It can be shown that Algorithm 3.5.4 has the following basic properties:

- Algorithm 3.5.4 is guaranteed to converge for a huge class of problems, i.e.,

$$\|\mathbf{u} - \mathbf{u}_\varepsilon\|_{\ell_2(\boldsymbol{\mathcal{J}})} \lesssim \varepsilon.$$

- The order of convergence of Algorithm 3.5.4 is *optimal* in the sense that it asymptotically realizes the convergence order of best N-term wavelet approximation, i.e., if the best N-term approximation satisfies $\mathcal{O}(N^{-s})$, then

$$\|\mathbf{u} - \mathbf{u}_\varepsilon\|_{\ell_2(\mathcal{J})} = \mathcal{O}((\#\mathrm{supp}\mathbf{u}_\varepsilon)^{-s}).$$

- The number of arithmetic operations stays proportional to the number of unknowns, that is, the number of floating point operations needed to compute $\mathbf{u}_\varepsilon$ satisfies $\mathcal{O}(\#\mathrm{supp}\mathbf{u}_\varepsilon)$.

Remark 3.5.9. The analysis in this chapter was treated for the linear case. Generalizations to the nonlinear case exist by now, see [28, 9, 45, 79]. However, the theory is only fully established for the isotropic case. For first results concerning the anisotropic case we refer to [112]. These specific results are based on interpolets.

3.5.3 Adaptive wavelet schemes for parabolic problems

In this section, we turn to the development of adaptive wavelet-based numerical schemes for linear parabolic problems of the form (3.2.6). We assume that we are given a Gelfand triple $V \hookrightarrow X \hookrightarrow V'$ of Hilbert spaces and that $\mathcal{A}(t) : V \to V'$ fits into the setting of Section 3.5.2. Moreover, we assume that

$$-\mathcal{A}(t) : D(\mathcal{A}) \subset X \to X$$

is *sectorial*, i.e., there are constants $z_0 \in \mathbb{R}$, $\omega_0 \in (\frac{\pi}{2}, \pi)$ and $L > 0$, such that the resolvent set $\varrho(-\mathcal{A}(t))$ contains the open sector

$$\Sigma_{z_0,\omega_0} := \big\{z \in \mathbb{C} \setminus \{z_0\} : |\arg(z - z_0)| < \omega_0\big\},$$

and the resolvent operator $R(z, -\mathcal{A}(t)) := (zI + \mathcal{A}(t))^{-1}$ of $-\mathcal{A}(t)$ is bounded in norm by

$$\big\|R(z, -\mathcal{A}(t))\big\|_{\mathcal{L}(X)} \leq \frac{L}{|z - z_0|}, \quad z \in \Sigma_{z_0,\omega_0}.$$

We may then consider (3.2.6) as an abstract initial value problem for a Hilbert space-valued variable $u : [0, T] \to V$. For its numerical treatment, we use the Rothe method which is also known as the horizontal method of lines. Doing so, the discretization is performed in two major steps. Firstly, we consider a semidiscretization in time, where we will employ an S-stage linearly implicit scheme. We shall end up with an orbit of approximations $u^{(n)} \in X$ at intermediate times t_n that are implicitly given via the S elliptic stage equations. In a finite element context, this very approach has already been propagated in [84]. For the realization of the increment $u^{(n)} \mapsto u^{(n+1)}$ and the spatial discretization of the stage equations, we will then employ the adaptive wavelet scheme introduced in Section 3.5.2 as a black box solver.

Let us start with the time discretization. In order to obtain a convenient notation, we will consider (3.2.6) in the generalized form

$$u'(t) = H\big(t, u(t)\big), \quad t \in (0,T], \quad u(0) = u_0,$$

where $H : [0,T] \times V \to V'$ is given as

$$H(t,v) = -\mathcal{A}(t)v + F(t,v), \quad t \in [0,T], \quad v \in V.$$

We consider an S-stage linearly implicit method for the semidiscretization in time. By this we mean an iteration of the form

$$u^{(n+1)} = u^{(n)} + h \sum_{i=1}^{S} b_i k_i \tag{3.5.7}$$

with the *stage equations*

$$(I - h\gamma_{i,i}J)k_i = H\Big(t_n + \alpha_i h, u^{(n)} + h\sum_{j=1}^{i-1} \alpha_{i,j}k_j\Big) + hJ\sum_{j=1}^{i-1}\gamma_{i,j}k_j + h\gamma_i g, \tag{3.5.8}$$

$i = 1, \ldots, S$, where we set

$$\alpha_i := \sum_{j=1}^{i-1} \alpha_{i,j}, \quad \gamma_i := \sum_{j=1}^{i} \gamma_{i,j}.$$

The operator $I - h\gamma_{i,i}J$ in (3.5.8) has to be understood as a boundedly invertible operator from V to V', with the equality (3.5.8) in the sense of V'. Such a scheme is also known as a method of *Rosenbrock* type, see [66, 122] for details. All the quantities h, J, k_i and g in (3.5.8) do of course depend on the time step number n, but we drop the index n here for readability. The coefficients b_i, $\alpha_{i,j}$ and $\gamma_{i,j}$ have to be suitably chosen according to the desired properties of the Rosenbrock method. As a special case of (3.5.8), a *Rosenbrock-Wanner method* or *ROW-method* results if one chooses the exact derivatives $J = \partial_v H(t_n, u^{(n)})$ and $g = \partial_t H(t_n, u^{(n)})$. In this paper, we will confine the setting to these ROW-type methods.

In practice, a Rosenbrock scheme will be implemented in a slightly different way than given by (3.5.8). Introducing the variable $u_i := h\sum_{j=1}^{i}\gamma_{i,j}k_j$, the additional application of the operator J in the right-hand side of (3.5.8) can be avoided by rewriting (3.5.8) as

$$\left(\frac{1}{h\gamma_{i,i}}I - J\right)u_i = H\Big(t_n + \alpha_i h, u^{(n)} + \sum_{j=1}^{i-1} a_{i,j}u_j\Big) + \sum_{j=1}^{i-1}\frac{c_{i,j}}{h}u_j + h\gamma_i g, \tag{3.5.9}$$

$i = 1, \ldots, S$, and

$$u^{(n+1)} = u^{(n)} + \sum_{i=1}^{S} m_i u_i \tag{3.5.10}$$

where we have used the coefficients

$$\begin{aligned}\Gamma &= (\gamma_{i,j})_{i,j=1}^{S},\\ (a_{i,j})_{i,j=1}^{S} &= (\alpha_{i,j})_{i,j=1}^{S}\Gamma^{-1},\\ (c_{i,j})_{i,j=1}^{S} &= \operatorname{diag}(\gamma_{1,1}^{-1},\dots,\gamma_{S,S}^{-1}) - \Gamma^{-1},\\ (m_1,\dots,m_S)^{\top} &= (b_1,\dots,b_S)^{\top}\Gamma^{-1}.\end{aligned}$$

It is well-known that for a strongly $A(\theta)$-stable Rosenbrock method the numerical approximations according to (3.5.7) indeed converge to the exact solution as $h \to 0$, see [91] for details. However, a constant temporal step size h might not be the most economic choice. At least for times t close to 0 and in situations where the driving term f is not smooth at t, it is advisable to choose small values of h in order to track the behavior of the exact solution correctly. In regions where f and u are temporally smooth, larger time step sizes may be used. As a consequence, we have to employ an a posteriori temporal error estimator to control the current value of h. The traditional approach resorts to estimators for the local truncation error at t_n

$$\delta_h(t_n) := \Phi^{t_n,t_n+h}(u(t_n)) - u(t_n+h),$$

where

$$\Phi^{t_n,t_n+h} : X \to X$$

is the increment mapping of the given Rosenbrock scheme at time t_n with step size h. For the global error at $t = t_{n+1} = t_n + h_n$, we have the decomposition

$$e_{n+1} = u^{(n+1)} - u(t_{n+1}) = \Phi^{t_n,t_n+h_n}(u^{(n)}) - \Phi^{t_n,t_n+h_n}(u(t_n)) + \delta_{h_n}(t_n),$$

i.e., e_{n+1} comprises the local error at time t_n and the difference between the current Rosenbrock step $\Phi^{t_n,t_n+h_n}(u^{(n)})$ and the virtual step $\Phi^{t_n,t_n+h_n}(u(t_n))$ with starting point $u(t_n)$. Estimators for the local discretization error $\delta_{h_n}(t_n)$ can be either based on an embedded lower order scheme or on extrapolation techniques, see [65, 66]. For applications to partial differential equations, embedding strategies yield sufficient results and thus are our method of choice.

Since the iteration (3.5.7) cannot be implemented numerically, we will now finally address the numerical approximation of all the ingredients by finite-dimensional counterparts. Precisely, we have to find approximate, computable iterands $\tilde{u}^{(n+1)}$, such that the additional error $\tilde{u}^{(n+1)} - u^{(n+1)}$ introduced by the spatial discretization stays below some given tolerance ε when measured in an appropriate norm. Hence this perturbation of the virtual orbit $(u^{(n)})_{n\in\mathbb{N}_0}$ can be interpreted as a controllable additional error of the temporal discretization. The accumulation of local perturbations in the course of the iteration is then an issue for the step size controller. In order not to spoil the convergence behavior of the unperturbed iterands $u^{(n)}$ we will demand that $\tilde{u}^{(n+1)} - u^{(n+1)}$ stays small in the topology of X, which results in the requirement

$$\|\tilde{u}^{(n+1)} - u^{(n+1)}\|_X \leq \varepsilon$$

for the numerical scheme, where $\varepsilon > 0$ is the desired target accuracy. To achieve this goal, we want to use the convergent adaptive wavelet schemes as outlined in Section 3.5.2. Observe that by (3.5.10), the exact increment $u^{(n+1)}$ differs from $u^{(n)}$ by a linear combination of the exact solutions u_i of the S stage equations (3.5.9).

In case that the ellipticity constants of $\partial_v H(t_n, u^{(n)})$ do not depend on t and v, and we choose $J = \partial_v H(t_n, u^{(n)})$ as above, the operators involved in (3.5.9) take the form

$$B_\alpha := \alpha I + \mathcal{A}(t), \quad \alpha \geq 0,$$

where $\alpha = (h\gamma_{i,i})^{-1}$ for the i-th stage equation. By the estimate

$$\langle B_0 v, v\rangle \leq \langle B_\alpha v, v\rangle = \alpha\langle v, v\rangle_V + \langle B_0 v, v\rangle \leq (C\alpha + 1)\langle B_0 v, v\rangle, \quad v \in V,$$

we see that the energy norms $\|v\|_{B_\alpha} := |\langle B_\alpha v, v\rangle|^{1/2}$ differ from $\|v\|_{B_0} \eqsim \|v\|_V$ only by an α-dependent constant:

$$\|v\|_{B_0} \leq \|v\|_{B_\alpha} \leq (C\alpha + 1)^{1/2}\|v\|_{B_0}, \quad v \in V.$$

Consequently, if we define

$$(\mathbf{D}_\alpha)_{\boldsymbol{\nu},\boldsymbol{\nu}} := \|\psi_{\boldsymbol{\nu}}\|_{B_\alpha}, \quad \boldsymbol{\nu} \in \boldsymbol{\mathcal{J}},$$

then the system $\mathbf{D}_\alpha^{-1}\boldsymbol{\Psi}$ is a Riesz basis in the energy space $(V, \|\cdot\|_{B_\alpha})$, with Riesz constants independent of $\alpha \geq 0$:

$$\|\mathbf{c}\|_{\ell_2(\boldsymbol{\mathcal{J}})} \sim \|\mathbf{c}^\top \mathbf{D}_\alpha^{-1}\boldsymbol{\Psi}\|_{B_\alpha}, \quad \mathbf{c} \in \ell_2(\boldsymbol{\mathcal{J}}).$$

Therefore, we can use the Riesz basis $\mathbf{D}_\alpha^{-1}\boldsymbol{\Psi}$, $\alpha = (h\gamma_{i,i})^{-1}$ as test functions in a variational formulation of (3.5.9). Abbreviating the exact right-hand side of (3.5.9) by

$$r_{i,h} := H\Big(t_n + \alpha_i h, u^{(n)} + \sum_{j=1}^{i-1} a_{i,j} u_j\Big) + \sum_{j=1}^{i-1} \frac{c_{i,j}}{h} u_j + h\gamma_i g,$$

we get the system of equations

$$\langle B_\alpha u_i, \mathbf{D}_\alpha^{-1}\boldsymbol{\Psi}\rangle^\top = \langle r_{i,h}, \mathbf{D}_\alpha^{-1}\boldsymbol{\Psi}\rangle^\top. \tag{3.5.11}$$

Inserting a wavelet representation of $u_i = (\mathbf{D}_\alpha \mathbf{u}_i)^\top \mathbf{D}_\alpha^{-1}\boldsymbol{\Psi}$ into the variational formulation (3.5.11), we end up with the biinfinite linear system in $\ell_2(\boldsymbol{\mathcal{J}})$

$$\mathbf{D}_\alpha^{-1}\langle B_\alpha \boldsymbol{\Psi}, \boldsymbol{\Psi}\rangle^\top \mathbf{D}_\alpha^{-1}\mathbf{D}_\alpha \mathbf{u}_i = \mathbf{D}_\alpha^{-1}\langle r_{i,h}, \boldsymbol{\Psi}\rangle^\top. \tag{3.5.12}$$

Now we observe that problem (3.5.12) exactly fits into the setting of Section 3.5.2.

A detailed analysis of the concepts outlined above can be found in the Ph.D. thesis [101].

3.6 Numerical experiments

In this section we apply the algorithms outlined in Sections 3.4 and 3.5 to solve the parameter identification problem associated to the fundamental problem (3.2.1). We want to highlight the potential of sparsity constrained Tikhonov regularization in connection with an adaptive wavelet solver for the forward problem in this nonlinear inverse problem. Parameter identification problems for parabolic differential equations are amongst the most demanding inverse problems, both in terms of analytic as well as numerical complexity, see e.g. [76]. In our case we computed a parameter that appears in a (linearized) reaction-diffusion equation from the noise-contaminated solution sampled at certain timesteps.

As a first step and as a proof of concept, we consider a linearized version of (3.2.1) and the reconstruction of one parameter for one gene from synthetic data. After presentation of the simplified problem and the resulting algorithm we give numerical results for spatial domains $U = (0,1)^n$, $n = 1, 2$. Just to emphasize the numerical complexity, we want to stress that each iteration step of the iterated soft shrinkage procedure requires to solve two parabolic PDEs. Our numerical tests were routinely done with 10^4 iterations. For the shrinkage parameter α and the noise levels δ we considered 20 and 12 different values for the one-dimensional example and 7 and 13 values for the two-dimensional case, respectively. That is, our test runs required the solution of about $6.62 \cdot 10^6$ parabolic PDEs which amounts to several days of computing time assuming that each cycle in the iteration (two calls to the forward solver, one multiplication) is done in $5 - 15$ seconds, depending on the parameters. The computations presented in the following sections were done on the Linux clusters of the Center for Industrial mathematics at the University of Bremen (36 compute nodes: 2 or 4 x Intel Opteron 2376 or 8378, 16 or 32 GByte RAM; 320 cores in total) and of the Philipps-University Marburg (MARC; 142 compute nodes: 2xDualCore Intel Opteron 270 or 2216, 8 or 16 GByte RAM; 568 cores in total).

3.6.1 An algorithm for a model problem

The previous sections were dealing with analytic properties of the parameter identification problem (3.2.1) and with algorithms for approximating its solution. We now derive a simplified test problem, which aims at reconstructing the biologically most important parameter W. The simplified test problem is chosen in order to allow exhaustive numerical tests. We therefore derive a model problem by linearization of (3.2.1), that is, by the choice $\phi(x) = x + \frac{1}{2}$. Additionally we set $N = 1$, i.e., we consider a single gene, choose $T = 1$ and fix the parameters $D = 1$, $\lambda = 0$, and $R = 1$. The remaining task is to identify the interaction matrix

$$W^{\text{true}} \in L_2(0,1; L_2(U, \mathbb{R}))$$

from given synthetic data $y_{\text{data}} := \mathscr{D}(W^{\text{true}})$. This still poses a nonlinear inverse problem.

In practical applications the synthetic data y_{data} will often be corrupted by noise. We consider additive ℓ_2 noise of magnitude δ and denote the data we used for reconstruction as y^δ.

For simplicity we discretize the parameter space by means of isotropic Haar wavelets up to a given maximal level and consequently choose the ℓ_1-norm of the Haar coefficients as the penalty term for the regularized minimization problem (3.4.5). This choice determines the sense in which sparsity of the reconstruction has to be understood.

We apply Algorithm 3.4.4 with the specifications made in Section 3.4.3 to compute a minimizer of (3.4.5). The second step of the algorithm is given by (3.4.6) and consequently two parabolic equations have to be solved in each iteration step. The first one corresponds to the action of the control-to-state operator $\mathscr{D}$ which maps the parameter W to the solution u of

$$\begin{aligned} &u' - \Delta u - Wu = \frac{1}{2} \quad \text{in } (0,1] \times U, \\ &\frac{\partial u}{\partial \nu} = 0 \text{ on } [0,T] \times \partial U, \quad u(0) = u_0 \text{ on } \{0\} \times U. \end{aligned} \tag{3.6.1}$$

The second task is to solve the adjoint problem given by $(\mathscr{D}'(W))^*(w)$, with $w = \mathscr{D}(W) - y^\delta$. Using the explicit formula derived in Remark 3.3.7, the solution is given by

$$\begin{aligned} &-h' - \Delta h - Wh = w \quad \text{in } (0,1] \times U, \\ &\frac{\partial h}{\partial \nu} = 0 \text{ on } [0,T] \times \partial U, \quad h(1) = 0 \text{ on } \{0\} \times U. \end{aligned} \tag{3.6.2}$$

Let us note, that the steady state solution of this problem coincides with the standard test problem for nonlinear parameter identification, see e.g. [54, Ex. 10.16].

Both parabolic problems are solved with the Rothe method using inexact linearly implicit increments, see Section 3.5.3. Consequently we do not need to consider a measurement operator $\mathcal{M}$ since we have full knowledge of y_{data} and the solutions to (3.6.1) and (3.6.2) at all time steps. For the experiments, we choose the second-order Rosenbrock scheme ROS2. We use equidistant time steps. The elliptic subproblems are discretized by means of anisotropic tensor wavelets for $W_2^1(U)$ as described in Section 3.5.1. We solve the subproblems adaptively by means of Algorithm 3.5.4. The biorthogonal one-dimensional wavelets used in the construction are taken from [99] and are chosen to be piecewise polynomial of order d and to have $\tilde{d}$ vanishing moments. For simplification we set $\sigma = 1$ in (3.4.6) and $s_n = 1, n \geq 0$ to omit the line search (3.4.4) in Algorithm 3.4.4, cf. Remark 3.4.5.

The algorithm for solving the parameter identification problem is given by Algorith 3.6.1. There, $\mathbf{u}^{(n)}$ and $\mathbf{h}^{(n)}$ denote the coefficients of the solutions $u^{(n)}$ and $h^{(n)}$ of the problems (3.6.1) and (3.6.2) with respect to the spatial discretization, respectively. As a slight abuse of notation we also denote the Haar wavelet coefficients of the product $u^{(n)} \cdot h^{(n)}$ with boldface letters. It can be efficiently computed by using the

Haar generator coefficients of $u^{(n)}$ and $h^{(n)}$. The application of S_α has to be understood for each time step and coefficient. For the sake of comparability of the results we use a fixed number of iterations as a stopping criterion in our simulations.

Algorithm 3.6.1 RECONSTRUCT

$n := 0$; $\mathbf{W}^{(0)} := 0$;
repeat
Compute the solution $\mathbf{u}^{(n)}$ of (3.6.1) with parameter $\mathbf{W}^{(n)}$;
Compute the solution $\mathbf{h}^{(n)}$ of (3.6.2) with the right-hand side given by $\mathbf{u}^{(n)} - \mathbf{y}_{\text{data}}$;
Compute Haar wavelet coefficients $\mathbf{u}^{(n)} \cdot \mathbf{h}^{(n)}$ of the product $u^{(n)} \cdot h^{(n)}$;
Apply shrinkage $\mathbf{W}^{(n+1)} = S_\alpha(\mathbf{W}^{(n)} - \mathbf{u}^{(n)} \cdot \mathbf{h}^{(n)})$ (with S_α as in (3.4.7));
$n := n + 1$;
until $||\mathbf{W}^{(n)} - \mathbf{W}^{(n-1)}||_{\ell_1} / \|\mathbf{W}^{(n)}\|_{\ell_1} \leq \varepsilon$

3.6.2 Numerical results

For the case $n = 1$ we chose the order d and the vanishing moments $\tilde{d}$ of the primal wavelets as $d = \tilde{d} = 3$ for representing the solution u of the forward problem. Temporal discretization is done with 11 points.

We generated the synthetic data with W^{true} given by the projection of

$$W(t,x) = \frac{20000}{64} \max\left(0, (\frac{4}{10} - t)(t - \frac{8}{10})\right) \max(0, (\frac{1}{4} - x)(x - \frac{3}{4}))$$

onto the set of admissible Haar wavelets, see Figure 3.1.

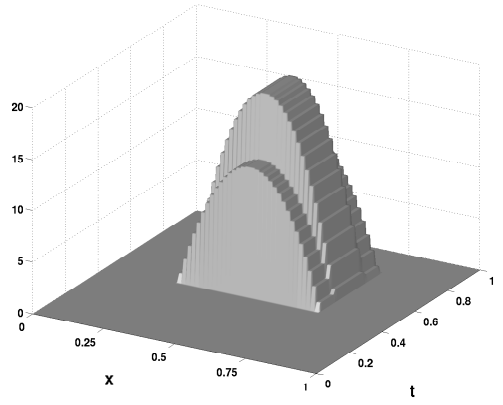


Figure 3.1: Exact unknown parameter W^{true}.

For given noise level δ and shrinkage parameter α we denote the computed reconstruction with W_α^δ. For choosing α we use an explorative approach and test all α

in

$$\{1,\ 20, 30, 50, 100, 120, 170, 230, 310, 390, 460, 550, 630, 750, 900,\\ 1300, 2000, 3500, 5000, 10000\} \cdot 10^{-7},$$

which are then compared with the L-curve criterion [70]. That is, α is determined by the point of maximal curvature in the $\|\mathbf{W}_\alpha^\delta\|_{\ell_1}$ over $\|y^\delta - \mathcal{D}(W_\alpha^\delta)\|_{L_2}$ plot. We denote this choice in the following with $\alpha(\delta, y^\delta)$.

First we present our results for the case without noise in Figure 3.2. Note that the best approximation is not obtained at the smallest tested value $\alpha = 10^{-7}$ but rather at $\alpha = 1.2 \cdot 10^{-5}$. This is due to the trade-off between discretization level, numerical artifacts and loss of detail. The reconstruction is given in Figure 3.5. The reconstruction obtained for zero noise level and with α determined by the L-curve criterion is denote by

$$W^{\mathrm{opt}} := W^0_{\alpha(0,y^0)}.$$

We regard this as the best achievable reconstruction within the restrictions of the chosen discretization scheme and compare our numerical reconstructions for noisy data with W^{opt}.

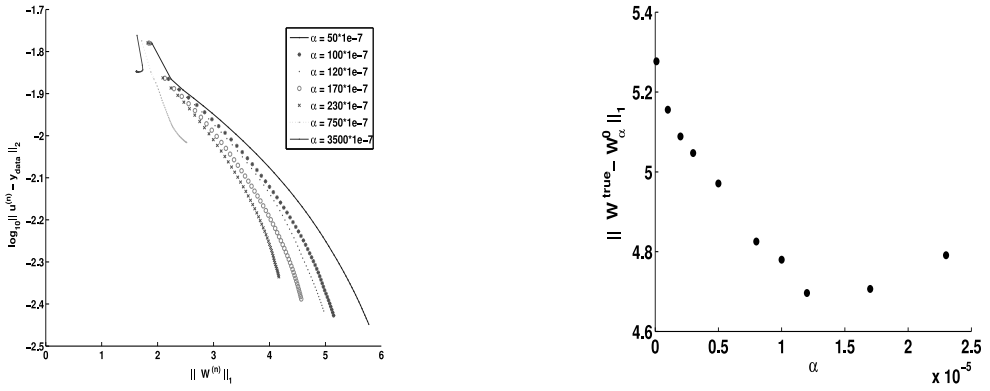


Figure 3.2: $\delta = 0$. Left: log-discrepancy over $\|\mathbf{W}^{(n)}\|_{\ell_1}$, right: unknown true reconstruction error.

Since the basic problem is ill-posed, understanding the dependence of the reconstruction method on the noise level is crucial. We explored ℓ_2 noise of magnitudes δ from $\{1, 3, 5, 7, 10, 12, 15, 25, 30\} \cdot 10^{-2}$, resulting in relative noise levels $\delta/\|\mathbf{y}_{\mathrm{data}}\|_{\ell_2}$ ranging from 0.313% to 9.36%.

As an example we present the special case of relative noise level 4.68%. The decline of the discrepancy over the course of iterations is rather fast during the first iterations and then levels out depending on α, see Figure 3.3. This seems to be a typical behavior. The reconstruction error is presented in Figure 3.4. To give a qualitative understanding of the reconstruction we present Figure 3.5.

As a practical proof of convergence we analyzed the reconstruction error

$$\|\mathbf{W}^{\mathrm{opt}} - \mathbf{W}^\delta_{\alpha(\delta,y^\delta)}\|_{\ell_1}$$

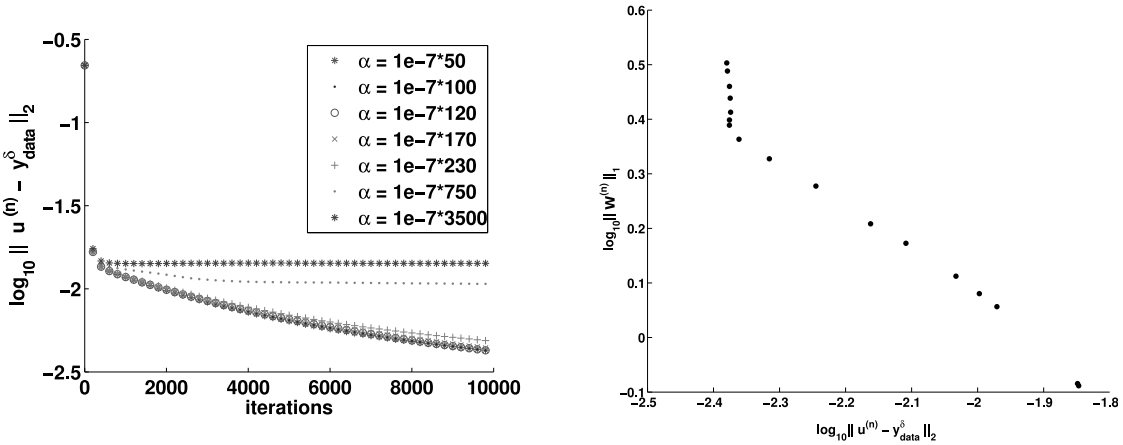


Figure 3.3: Noise 4.68%. Left: log-discrepancy over iterations, right: L-curve.

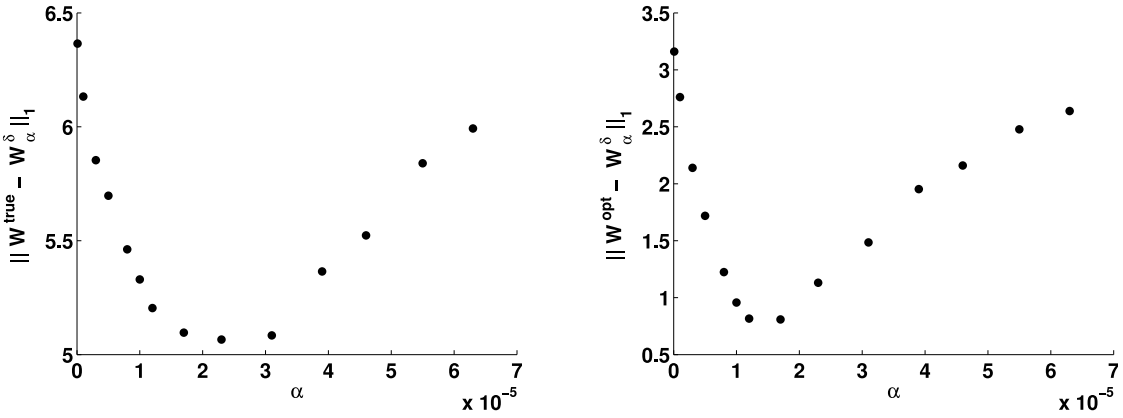


Figure 3.4: Noise 4.68%. Left: $\|\mathbf{W}^{\text{true}} - \mathbf{W}_\alpha^\delta\|_{\ell_1}$ over α, right: $\|\mathbf{W}^{\text{opt}} - \mathbf{W}_\alpha^\delta\|_{\ell_1}$ over α.

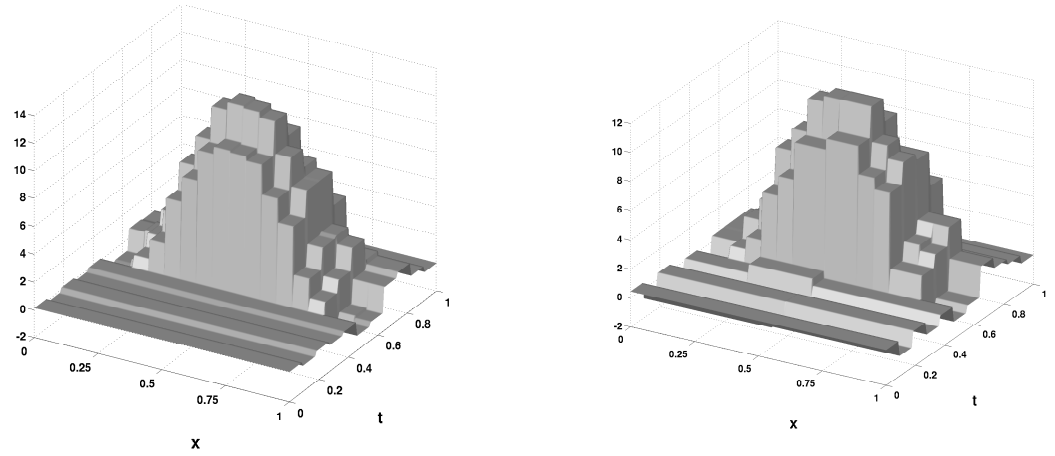


Figure 3.5: Reconstructions. Left: $\alpha = 1.2 \cdot 10^{-5}, \delta = 0$, right: $\alpha = 1.7 \cdot 10^{-5}$, noise 4.68%.

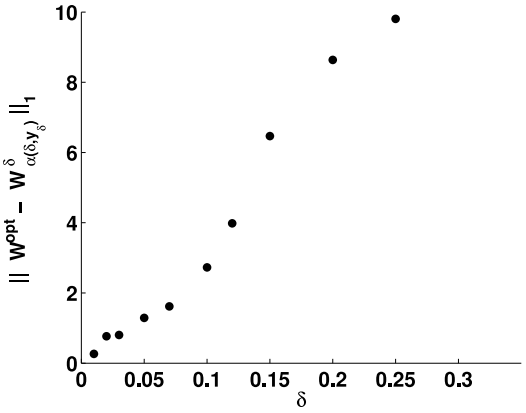


Figure 3.6: $\|\mathbf{W}^{\mathrm{opt}} - \mathbf{W}^{\delta}_{\alpha(\delta,y^{\delta})}\|_{\ell_1}$ over δ.

for decreasing noise levels δ. The convergence of the method is shown in Figure 3.6.

For the two-dimensional case $n = 2$, we used a similar approach. Order and vanishing moments of the primal univariate wavelets used in the anisotropic construction were chosen to be $d = \tilde{d} = 2$. We considered a temporal discretization with 9 points. The synthetic data were computed using W^{true} consisting of the projection of $W(t, \mathbf{x}) = h(t)g(\mathbf{x})$, given by

$$h(t) = \frac{64}{9} \max\Big(0, \Big(t - \frac{1}{8}\Big)\Big(\frac{7}{8} - t\Big)\Big),$$
$$g(\mathbf{x}) = \max\Big(0, \min\Big(4\Big(x_1 - \frac{1}{4}\Big), 1\Big)\Big) \max\Big(0, \min\Big(\frac{3}{2} - 4|x_2 - \frac{3}{8}|, 1\Big)\Big),$$

onto the set of admissible Haar wavelets. See Figure 3.7 for the value of W^{true} at $t = 0.5$.

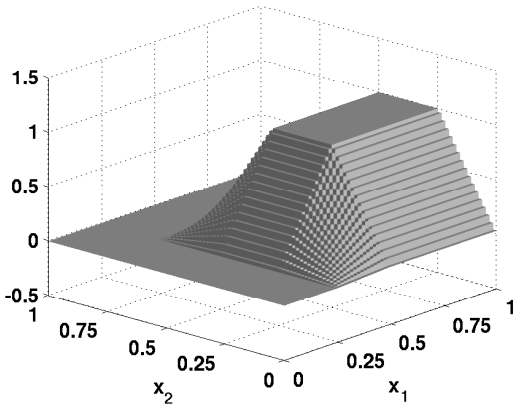


Figure 3.7: Exact parameter W^{true} at $t = 0.5$.

For our experiments we considered values for the shrinkage parameter α from $\{1, 5, 10, 15, 25, 50, 100\} \cdot 10^{-7}$ and 13 noise levels δ of magnitudes between 10^{-3} and

0.5. They correspond to relative noise levels between 0.023% and 11.41%. The convergence behavior of the method is quite similar to the one-dimensional case depicted in Figure 3.3. In this example we observed that the method is more sensitive to noise than in the one-dimensional case. To give a qualitative understanding of the dependency of the method on α and δ we present several reconstructions at $t = 0.5$ in Figure 3.8.

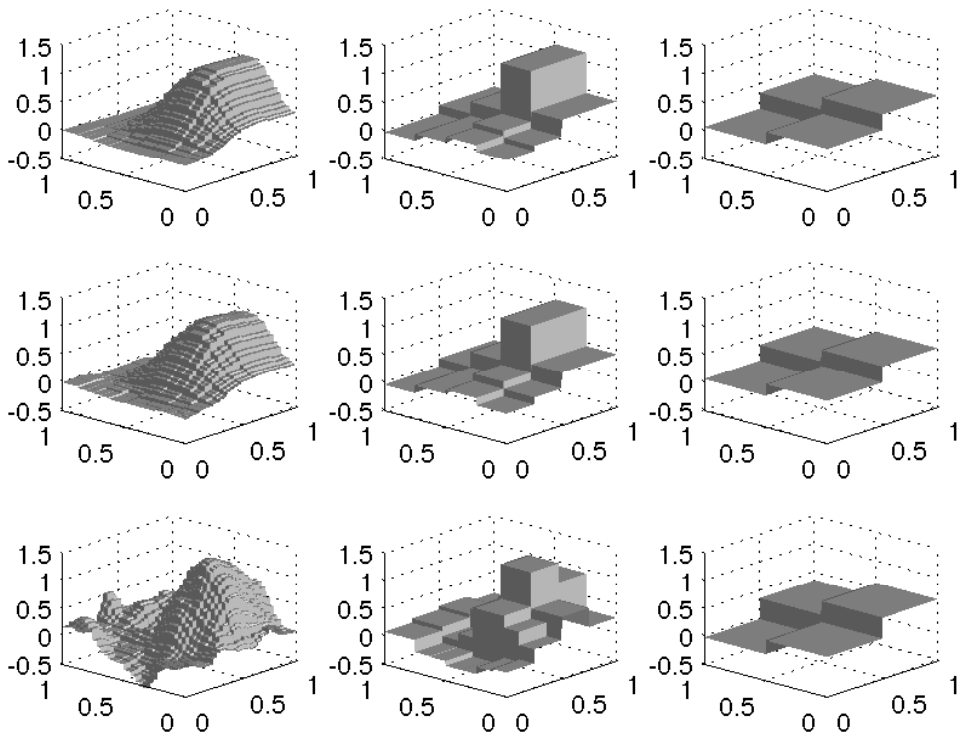


Figure 3.8: Reconstructions at $t = 0.5$. From left to right: $\alpha = 10^{-7}, 10^{-6}, 5{\cdot}10^{-6}$, from top to bottom: $\delta = 0, 6 \cdot 10^{-3}, 4 \cdot 10^{-2}$, i.e., relative noise 0%,0.14%,0.91%.

4 On the convergence analysis of spatially adaptive Rothe methods

Authors: P.A. Cioica, S. Dahlke, N. Döhring, U. Friedrich, S. Kinzel, F. Lindner, T. Raasch, K. Ritter, R.L. Schilling.
Journal: Foundations of Computational Mathematics **14** (2014), no. 5, 863–912.
Abstract: This paper is concerned with the convergence analysis of the horizontal method of lines for evolution equations of parabolic type. After a semidiscretization in time by S-stage one-step methods, the resulting elliptic stage equations per time step are solved with adaptive space discretization schemes. We investigate how the tolerances in each time step have to be tuned in order to preserve the asymptotic temporal convergence order of the time stepping also in the presence of spatial discretization errors. In particular, we discuss the case of linearly-implicit time integrators and adaptive wavelet discretizations in space. Using concepts from regularity theory for partial differential equations and from nonlinear approximation theory, we determine an upper bound for the degrees of freedom for the overall scheme that are needed to adaptively approximate the solution up to a prescribed tolerance.
MSC 2010: Primary: 35K90, 65J08, 65M20, 65M22, 65T60; secondary: 41A65, 46E35.
Key Words: Parabolic evolution equations, horizontal method of lines, S-stage linearly-implicit methods, adaptive wavelet methods, Besov spaces, nonlinear approximation.

4.1 Introduction

This paper is concerned with the numerical treatment of evolution equations of parabolic type, i.e.,

$$u'(t) = F(t, u(t)), \quad u(0) = u_0, \quad t \in [0, T]. \tag{4.1.1}$$

Typical examples are, for instance, semilinear equations of the form

$$u'(t) = Au(t) + f(t, u(t)), \quad u(0) = u_0, \quad t \in [0, T], \tag{4.1.2}$$

where, in practical applications, usually A is a differential operator and f a linear or nonlinear forcing term. Such equations describe diffusion processes and they are very often used for the mathematical modelling of biological, chemical and physical

processes. There are two principally different approaches to solve (4.1.1): the vertical method of lines and the horizontal method of lines. The former starts with an approximation first in space and then in time. We refer to [68], [77], [124] for detailed information. The latter starts with a discretization first in time and then in space; it is also known as *Rothe's method*. It has been studied in [84], [91]. These references are indicative and by no means complete.

In this paper, we concentrate on Rothe's method for the following reasons. In the horizontal method of lines, the parabolic equation can be interpreted as an abstract Cauchy problem, i.e., as an ordinary differential equation in some suitable function space. This immediately provides a way to employ adaptive strategies. Indeed, in time direction we might use one of the well known ODE solvers with step size control. This solver must be based on an implicit discretization scheme since the equation under consideration is usually stiff. Linearly-implicit one-step methods are of primary interest because their realization only requires to solve a system of linear elliptic stage equations per time step. To this end, as a second level of adaptivity, well-established adaptive numerical schemes based, e.g., on wavelets or finite elements, can be used. We refer to [26], [27], [37] for the wavelet case, and [5, 6, 8], [14], [55, 56, 57, 58], [69], [126], [127] for the finite element case. As before, these lists are not complete.

Although the combination of Rothe's method with adaptive strategies is natural, a rigorous convergence analysis seems to be still in its infancy. For parabolic equations and finite element discretization in space, the most far reaching results have been obtained in [84].

Not very much seems to be known about fully adaptive schemes. This paper can be seen as a first step in this direction. We still use uniform discretizations in time, but for the space discretization we use an arbitrary (non-uniform and adaptive) discretization scheme that allows to compute an approximation to the exact solution of an elliptic subproblem up to a prescribed accuracy. To treat the convergence problem, we start with the observation that at an abstract level, Rothe's method can be reformulated as the consecutive application of two types of operators, the inverse of a (linear) elliptic differential operator and certain (nonlinear) evaluation operators. Adaptivity enters via the inexact application of both types of operators in each time step, up to a given tolerance. Obviously, we need to know whether the whole scheme still converges with all these perturbations and how the tolerances in each time step have to be tuned to obtain convergence and corresponding convergence orders. These aspects are studied in Section 4.2.

Fortunately, it turns out that a huge class of concrete discretization schemes can be written as abstract Rothe schemes in the sense of Section 4.2. Indeed, in Section 4.3 we show that any linearly-implicit one-step scheme of W-type falls into this category. By combining our analysis with the convergence results for the unperturbed schemes, which, e.g., are outlined in [91], we are therefore able to provide rigorous convergence proofs for spatially adaptive versions of W-methods. The analysis is substantiated by several examples, where special emphasis is layed on the semilinear case (4.1.2).

The analysis in the Section 4.2 and 4.3, respectively, holds for any spatially adaptive

numerical scheme that provides an approximation of the unknown solution up to any prescribed tolerance. In the finite element setting, such strategies have been derived in [12], [52], [120], however, for several reasons, we are in particular interested in spatially adaptive schemes based on *wavelets*. In recent studies, it has turned out that the strong analytical properties of wavelets can be used to design adaptive numerical schemes that are guaranteed to converge with optimal order, i.e., with the same order as best m-term wavelet approximation. We refer, e.g., to [26], [27] for details. These relationships pave a way to rigorous complexity estimates in the wavelet case. Indeed, it is well-known that the convergence order of best m-term approximation depends on the smoothness of the object one wants to approximate in specific scales of Besov spaces [48]. So, the overall complexity can be determined by combining our abstract analysis with estimates for the Besov smoothness of the solutions to the elliptic subproblems in each time step. In Section 4.4, we show that, although technically quite involved, this way is indeed passable. In particular, we study the case of the classical heat equation and the linearly-implicit Euler schemes.

4.2 Abstract description of Rothe's method

We begin with an example that motivates our perspective on the analysis of Rothe's method. The setting and notation will be given in Section 4.2.2, and in Section 4.2.3 we state and prove one of our main results, that is an abstract convergence proof.

4.2.1 Motivation

To introduce our abstract setting of Rothe's method, let us consider the heat equation

$$\left.\begin{aligned} u'(t) &= \Delta u(t) + f(t,u(t)) && \text{on } \Omega,\ t \in (0,T], \\ u(0) &= u_0 && \text{on } \Omega, \\ u &= 0 && \text{on } \partial\Omega,\ t \in (0,T], \end{aligned}\right\} \tag{4.2.1}$$

where $\Omega \subset \mathbb{R}^d$, $d \geq 1$, denotes a bounded Lipschitz domain. We discretize this equation by means of a linearly-implicit Euler scheme with uniform time steps. Let $K \in \mathbb{N}$ be the number of subdivisions of the time interval $[0,T]$, where the step size will be denoted by $\tau := T/K$, and the k-th point in time is denoted by $t_k := \tau k$, $k \in \{0,\dots,K\}$. The linearly-implicit Euler scheme, starting at u_0, is given by

$$\frac{u_{k+1}-u_k}{\tau} = \Delta u_{k+1} + f(t_k,u_k),$$

i.e.,

$$(I-\tau\Delta)u_{k+1} = u_k + \tau f(t_k,u_k), \tag{4.2.2}$$

for $k = 0,\dots,K-1$. If we assume that the elliptic problem

$$L_\tau v := (I-\tau\Delta)v = w \quad \text{on } \Omega, \quad v|_{\partial\Omega} = 0,$$

can be solved exactly, then one step of the scheme (4.2.2) can be written as

$$u_{k+1} = L_\tau^{-1} R_{\tau,k}(u_k), \tag{4.2.3}$$

where

$$R_{\tau,k}(v) := v + \tau f(t_k, v)$$

and L_τ is a boundedly invertible operator between suitable Hilbert spaces. That is, we can look at this equation in a Gelfand triple setting $(H_0^1(\Omega),\ L_2(\Omega),\ H^{-1}(\Omega))$ with L_τ as an operator from $H_0^1(\Omega)$ to $H^{-1}(\Omega)$, where $H_0^1(\Omega)$ denotes the H^1-Sobolev space with Dirichlet boundary conditions, $H^{-1}(\Omega)$ its normed dual, and $L_2(\Omega)$ the Lebesgue space of quadratic integrable functions. We may also consider (4.2.3) in $L_2(\Omega)$, since $H_0^1(\Omega)$ is embedded in $L_2(\Omega)$ and $L_2(\Omega)$ is embedded in $H^{-1}(\Omega)$, provided that $R_{\tau,k} : L_2(\Omega) \to L_2(\Omega)$ is well defined.

Having the above simple example in mind, we observe that the fundamental form of (4.2.3) essentially remains the same even if we introduce more sophisticated discretizations in time, e.g., as outlined below and in Section 4.3.

4.2.2 Setting and assumptions

In many applications not only one-stage approximation methods, such as the linearly-implicit Euler scheme, are used, but also more sophisticated S-stage schemes. The reason is, S-stage schemes can lead to higher temporal convergence orders, see Section 4.3 for further details. Therefore, in this subsection we state a scheme with the same form as in (4.2.3) that provides an abstract interpretation of linearly-implicit S-stage schemes, where $S \in \mathbb{N}$.

As above, we begin with a uniform discretization of the time interval $[0, T]$ with $K \in \mathbb{N}$ subdivisions, step size $\tau := T/K$, and $t_k := k\tau$ for $k \in \{0, \dots, K\}$. Taking an abstract point of view, we introduce separable real Hilbert spaces $\mathcal{H}$, $\mathcal{G}$, and consider a mapping $u : [0, T] \to \mathcal{H}$. Furthermore, let $L_{\tau,i}$ be a family of, possibly unbounded, linear operators which have bounded inverses $L_{\tau,i}^{-1} : \mathcal{G} \to \mathcal{H}$, and let

$$R_{\tau,k,i} : \underbrace{\mathcal{H} \times \cdots \times \mathcal{H}}_{i} \to \mathcal{G} \tag{4.2.4}$$

be a family of (nonlinear) evaluation operators for $k \in \{0, \dots, K-1\}$ and $i = 1, \dots, S$. As the norm on the Cartesian product in (4.2.4) we set

$$\big\|(v_1, \dots, v_i)\big\|_{\mathcal{H}\times\cdots\times\mathcal{H}} := \sum_{l=1}^{i} \|v_l\|_{\mathcal{H}}.$$

Remark 4.2.1. **(i)** The function $u : [0, T] \to \mathcal{H}$ is understood to be a solution of a parabolic partial differential equation of the form (4.1.2).

(ii) In most cases $L_{\tau,i}^{-1}$ is not given explicitly and, for this reason, we need an efficient numerical scheme for its evaluation. The situation is completely different with $R_{\tau,k,i}$,

which is usually given explicitly and does not require the solution of an operator equation for its evaluation. Concrete examples of these operators will be presented and studied in Section 4.3.

(iii) In a Gelfand triple setting (V, U, V^*) typical choices for the spaces $\mathcal{H}$ and $\mathcal{G}$ are $\mathcal{H} = V$, $\mathcal{G} = V^*$ or $\mathcal{H} = \mathcal{G} = U$. However, also a more general setting such as

$$V \subseteq \mathcal{H} \subseteq U \subseteq V^* \subseteq \mathcal{G}$$

is possible. Observe that our motivating example from Section 4.2.1 fits in this setting with $H_0^1(\Omega) = \mathcal{H} \subseteq L_2(\Omega)$ and $\mathcal{G} = H^{-1}(\Omega)$.

Starting from the given value $u_0 := u(0) \in \mathcal{H}$, we define the *abstract exact S-stage scheme* iteratively by

$$\left.\begin{aligned} u_{k+1} &:= \sum_{i=1}^{S} w_{k,i}, \\ w_{k,i} &:= L_{\tau,i}^{-1} R_{\tau,k,i}(u_k, w_{k,1}, \ldots, w_{k,i-1}), \quad i = 1, \ldots, S, \end{aligned}\right\} \tag{4.2.5}$$

for $k = 0, \ldots, K-1$. One step of this iteration can be described as an application of the operator

$$\left.\begin{aligned} E_{\tau,k,k+1} : \mathcal{H} &\to \mathcal{H}, \\ v &\mapsto \sum_{i=1}^{S} w_{k,i}(v), \\ w_{k,i}(v) &:= L_{\tau,i}^{-1} R_{\tau,k,i}\big(v, w_{k,1}(v), \ldots, w_{k,i-1}(v)\big), \quad i = 1, \ldots, S. \end{aligned}\right\} \tag{4.2.6}$$

If we define the family of operators

$$E_{\tau,j,k} := \begin{cases} E_{\tau,k-1,k} \circ \ldots \circ E_{\tau,j,j+1}, & j < k \\ I, & j = k, \end{cases} \tag{4.2.7}$$

then the output of the exact S-stage scheme (4.2.5) is given by the sequence

$$u_k = E_{\tau,0,k}(u_0), \quad k = 0, \ldots, K. \tag{4.2.8}$$

The convergence analysis which we present relies on a crucial technical assumption on the operators defined in (4.2.7).

Assumption 4.2.2. For all $0 \le j, k \le K$ the operators

$$E_{\tau,j,k} : \mathcal{H} \to \mathcal{H} \text{ are globally Lipschitz continuous}$$

with Lipschitz constants $C_{\tau,j,k}^{\mathrm{Lip}}$.

Remark 4.2.3. Assumption 4.2.2 is relatively mild, as it is usually fulfilled in the applications we have in mind. Concrete examples will be given at the end of this section, as well as in Section 4.3.

We call the sequence (4.2.8) the *output of the exact S-stage scheme*, since the operators involved in the definition of $E_{\tau,0,k}$ are evaluated exactly. In practical applications this is very often not possible; the operators $L_{\tau,i}^{-1}$ and $R_{\tau,k,i}$ can only be evaluated up to a prescribed accuracy. Therefore, as a start, we make the following

Assumption 4.2.4. For all τ, $k \in \{0, \dots, K-1\}$, and for any prescribed tolerance $\varepsilon_k > 0$ and arbitrary $v \in \mathcal{H}$, we have an approximation $\tilde{E}_{\tau,k,k+1}(v)$ of $E_{\tau,k,k+1}(v)$ at hand, such that

$$\big\|E_{\tau,k,k+1}(v) - \tilde{E}_{\tau,k,k+1}(v)\big\|_{\mathcal{H}} \leq \varepsilon_k$$

with a known upper bound $M_{\tau,k}(\varepsilon_k, v) < \infty$ for the degrees of freedom needed to achieve the prescribed tolerance ε_k.

Remark 4.2.5. In this abstract setting the term *degrees of freedom* might be a bit vague, since the precise meaning of this term depends on the concrete form of the applied approximation scheme. For instance, in the finite element and the wavelet setting, the degrees of freedom refer to the number of basis functions, which are needed for the approximant to achieve the tolerance.

For simplicity, we make the following

Assumption 4.2.6. The initial value is given exactly, i.e.,

$$\tilde{u}_0 := u(0).$$

Remark 4.2.7. The case where Assumption 4.2.6 does not hold, i.e., $\tilde{u}_0 \neq u(0)$, can be handled in a similar way. However, this only increases *notational* complexity.

Given an approximation scheme satisfying Assumption 4.2.4 and using Assumption 4.2.6, the *abstract inexact* variant of (4.2.5) is defined by

$$\left.\begin{aligned} \tilde{u}_0 &:= u(0), \\ \tilde{u}_{k+1} &:= \tilde{E}_{\tau,k,k+1}(\tilde{u}_k) \quad \text{for } k = 0, \dots, K-1. \end{aligned}\right\} \tag{4.2.9}$$

We will show in Theorem 4.2.18 how to tune the tolerances $(\varepsilon_k)_{k=0,\dots,K-1}$ in such a way that the scheme (4.2.9) has the same qualitative properties as the exact scheme (4.2.5). As in (4.2.7), we define

$$\tilde{E}_{\tau,j,k} := \begin{cases} \tilde{E}_{\tau,k-1,k} \circ \dots \circ \tilde{E}_{\tau,j,j+1}, & j < k \\ I, & j = k. \end{cases}$$

Consequently, the output of the inexact S-stage scheme (4.2.9) is given by

$$\tilde{u}_k = \tilde{E}_{\tau,0,k}\big(u(0)\big), \quad k = 0, \dots, K.$$

Now, we are faced with the following problems. In practice, the Lipschitz constants $C^{\mathrm{Lip}}_{\tau,j,k}$ of $E_{\tau,j,k}$, given in Assumption 4.2.2, might be hard to estimate directly and are only available in very specific situations. As we shall see in Section 4.4, the individual operators $L^{-1}_{\tau,i}R_{\tau,k,i}$, $i = 1,\dots,S$, are much easier to handle. Moreover, a direct approximation scheme for $E_{\tau,k,k+1}$, as required by Assumption 4.2.4, might also be hard to get directly. Nevertheless, very often, one has convergent numerical schemes for the individual operators $L^{-1}_{\tau,i}R_{\tau,k,i}$. Therefore, with these observations in mind, we are now going to state the corresponding assumptions for these individual operators.

Assumption 4.2.8. For $k = 0,\dots,K-1$ and $i = 1,\dots,S$ the operators

$$L^{-1}_{\tau,i}R_{\tau,k,i} : \underbrace{\mathcal{H}\times\cdots\times\mathcal{H}}_{i} \to \mathcal{H} \text{ are globally Lipschitz continuous}$$

with Lipschitz constants $C^{\mathrm{Lip}}_{\tau,k,(i)}$.

Remark 4.2.9. Note that, on the one hand, Assumption 4.2.2 is slightly more general than Assumption 4.2.8, since it is easy to see that a composition of non-Lipschitz continuous operators can be Lipschitz continuous. On the other hand, Assumption 4.2.8 implies Assumption 4.2.2. This is a consequence of the fact, that, if we introduce the constants

$$C'_{\tau,k,(i)} := \prod_{l=i+1}^{S}\left(1 + C^{\mathrm{Lip}}_{\tau,k,(l)}\right) \tag{4.2.10}$$

for $k = 0,\dots,K-1$ and $i = 0,\dots,S$, we can estimate the Lipschitz constants $C^{\mathrm{Lip}}_{\tau,j,k}$ of $E_{\tau,j,k}$ as follows:

$$C^{\mathrm{Lip}}_{\tau,j,k} \le \prod_{r=j}^{k-1}\left(C'_{\tau,r,(0)} - 1\right), \quad 0 \le j \le k \le K. \tag{4.2.11}$$

This will be worked out in detail in the proof of Theorem 4.2.21.

The analogue to Assumption 4.2.4 is

Assumption 4.2.10. For all τ, $k \in \{0,\dots,K-1\}$, $i \in \{1,\dots,S\}$, there exists a numerical scheme that, for any prescribed tolerance $\varepsilon_{k,i} > 0$ and arbitrary $v_0,\dots,v_{i-1} \in \mathcal{H}$, yields an approximation $[v]_{\varepsilon_{k,i}}$ of

$$v := L^{-1}_{\tau,i}R_{\tau,k,i}(v_0,\dots,v_{i-1}),$$

such that

$$\left\|v - [v]_{\varepsilon_{k,i}}\right\|_{\mathcal{H}} \le \varepsilon_{k,i}$$

with a known upper bound $M_{\tau,k,i}(\varepsilon_{k,i}, v) < \infty$ for the degrees of freedom needed to achieve the prescribed accuracy $\varepsilon_{k,i}$.

For any numerical scheme satisfying Assumption 4.2.10, and given tolerances $\varepsilon_{k,i} > 0$, $k = 0, \dots, K-1$, $i = 1, \dots, S$, the corresponding *inexact* variant of (4.2.5) is defined by

$$\left.\begin{aligned} \tilde{u}_0 &:= u(0), \\ \tilde{u}_{k+1} &:= \sum_{i=1}^{S} \tilde{w}_{k,i}, \\ \tilde{w}_{k,i} &:= \big[L_{\tau,i}^{-1} R_{\tau,k,i}(\tilde{u}_k, \tilde{w}_{k,1}, \dots, \tilde{w}_{k,i-1})\big]_{\varepsilon_{k,i}}, \quad i = 1, \dots, S, \end{aligned}\right\} \tag{4.2.12}$$

for $k = 0, \dots, K-1$. Note that (4.2.12) is consistent with (4.2.9), since it corresponds to the specific choice

$$\left.\begin{aligned} \tilde{E}_{\tau,k,k+1} : \mathcal{H} &\to \mathcal{H}, \\ v &\mapsto \sum_{i=1}^{S} \tilde{w}_{k,i}(v), \\ \tilde{w}_{k,i}(v) &:= \big[L_{\tau,i}^{-1} R_{\tau,k,i}(v, \tilde{w}_{k,1}(v), \dots, \tilde{w}_{k,i-1}(v))\big]_{\varepsilon_{k,i}}, \quad i = 1, \dots, S. \end{aligned}\right\} \tag{4.2.13}$$

In Theorem 4.2.23 we will show how to tune the tolerances in the scheme (4.2.12) in such a way that the approximation of u in $\mathcal{H}$ has the same qualitative properties as the exact scheme (4.2.5).

Remark 4.2.11. **(i)** For $\tilde{E}_{\tau,k,k+1}$ as in (4.2.13) and arbitrary $v \in \mathcal{H}$, the estimate

$$\big\| E_{\tau,k,k+1}(v) - \tilde{E}_{\tau,k,k+1}(v) \big\|_{\mathcal{H}} \le \sum_{i=1}^{S} C'_{\tau,k,(i)} \varepsilon_{k,i} \tag{4.2.14}$$

holds with $C'_{\tau,k,(i)}$ given by (4.2.10). Thus, for any prescribed tolerance ε_k, if Assumptions 4.2.8 and 4.2.10 are fulfilled, we can choose $\varepsilon_{k,i}$, $i = 1, \dots, S$, in such a way that the error we make by applying $\tilde{E}_{\tau,k,k+1}$ from (4.2.13) instead of $E_{\tau,k,k+1}$ is bounded by ε_k, uniformly in $\mathcal{H}$. In this sense Assumption 4.2.10 implies Assumption 4.2.4. Detailed arguments for the validity of estimate (4.2.14) will be given in the proof of Theorem 4.2.21.

(ii) We do not specify the numerical scheme $[\,\cdot\,]_\varepsilon$ at this point. It might be based on, e.g., a spectral method, an (adaptive) finite element scheme, or an adaptive wavelet solver. The latter case will be discussed in detail in Section 4.4. There, $M_{\tau,k,i}(\varepsilon, v)$ will be an upper bound for the number of elements of the spatial wavelet system that is needed to achieve the desired tolerance.

(iii) Later on, in Section 4.4, we will assume that the operators $R_{\tau,k,i}$ can be evaluated exactly which, of course, may not always be possible. We postpone the analysis of these additional difficulties to a forthcoming paper.

4.2.3 Controlling the error of the inexact schemes

We want to use the schemes described in Section 4.2.2 to compute approximations to a solution $u : [0,T] \to \mathcal{H}$ of a parabolic partial differential equation. The analysis presented in this section is based on the central assumption that the *exact* scheme (4.2.5) converges to the solution with a given approximation order δ, cf. Assumption 4.2.14. In Theorem 4.2.18 and Theorem 4.2.23, we state conditions how to tune the tolerances in the *inexact* schemes (4.2.9) and (4.2.12), respectively, so that they still converge to u and inherit the approximation order δ of the exact scheme. We start with a natural assumption.

Assumption 4.2.12. There exists a unique solution $u : [0,T] \to \mathcal{H}$ to the problem under consideration, i.e., to (4.1.1) or (4.1.2), respectively.

Remark 4.2.13. Of course, the type of such solutions depends on the form of the specific parabolic partial differential equation. We avoid, on purpose, a detailed discussion of this aspect in this section. Further information are given in Remark 4.3.7.

The analysis presented in this section is based on the following central

Assumption 4.2.14. The *exact* scheme (4.2.5) converges to $u(T)$ with order $\delta > 0$, i.e., for some constant $C_{\text{exact}} > 0$,

$$\left\| u(T) - E_{\tau,0,K}(u(0)) \right\|_{\mathcal{H}} \leq C_{\text{exact}}\, \tau^{\delta},$$

where the constant may depend on f, T, and u_0, but not on τ.

Remark 4.2.15. Error estimates as in Assumption 4.2.14 are quite natural and hold very often, see Section 4.3 and the references therein, in particular, [91, Theorem 6.2].

At first, we give an estimate for the error propagation of the scheme (4.2.9) measured in the norm of $\mathcal{H}$.

Theorem 4.2.16. *Suppose that Assumptions 4.2.2, 4.2.4, 4.2.6, and 4.2.12 hold. Let $(u_k)_{k=0}^K$, $K \in \mathbb{N}$, be the output of the exact scheme* (4.2.5)*, and let $(\tilde{u}_k)_{k=0}^K$ be the output of the scheme* (4.2.9) *with given tolerances ε_k, $k = 0, \dots, K-1$. Then, for all $0 \leq k \leq K$,*

$$\|u(t_k) - \tilde{u}_k\|_{\mathcal{H}} \leq \|u(t_k) - u_k\|_{\mathcal{H}} + \sum_{j=0}^{k-1} C^{\text{Lip}}_{\tau,j+1,k}\, \varepsilon_j.$$

Proof. The triangle inequality yields

$$\|u(t_k) - \tilde{u}_k\|_{\mathcal{H}} \leq \left\| u(t_k) - u_k \right\|_{\mathcal{H}} + \|u_k - \tilde{u}_k\|_{\mathcal{H}},$$

so it remains to estimate the second term. We write $E_{j,k} := E_{\tau,j,k}$ for simplicity. Using $u_0 = \tilde{u}_0$ and writing $u_k - \tilde{u}_k$ as a telescopic sum, we get

$$\begin{aligned} u_k - \tilde{u}_k &= \big(E_{0,k}(\tilde{u}_0) - E_{1,k}\tilde{E}_{0,1}(\tilde{u}_0)\big) \\ &\quad + \big(E_{1,k}\tilde{E}_{0,1}(\tilde{u}_0) - E_{2,k}\tilde{E}_{0,2}(\tilde{u}_0)\big) \\ &\quad \dots \\ &\quad + \big(E_{k-1,k}\tilde{E}_{0,k-1}(\tilde{u}_0) - \tilde{E}_{0,k}(\tilde{u}_0)\big) \\ &= \sum_{j=0}^{k-1} \big(E_{j,k}\tilde{E}_{0,j}(u_0) - E_{j+1,k}\tilde{E}_{0,j+1}(u_0)\big). \end{aligned}$$

Another application of the triangle inequality yields

$$\|u_k - \tilde{u}_k\|_{\mathcal{H}} \leq \sum_{j=0}^{k-1} \big\|E_{j,k}\tilde{E}_{0,j}(u_0) - E_{j+1,k}\tilde{E}_{0,j+1}(u_0)\big\|_{\mathcal{H}}.$$

Due to the Lipschitz continuity of $E_{\tau,j,k}$, cf. Assumption 4.2.2, each term in the sum can be estimated from above by

$$\begin{aligned} &\big\|E_{j,k}\tilde{E}_{0,j}(u_0) - E_{j+1,k}\tilde{E}_{0,j+1}(u_0)\big\|_{\mathcal{H}} \\ &\quad = \big\|E_{j+1,k}E_{j,j+1}\tilde{E}_{0,j}(u_0) - E_{j+1,k}\tilde{E}_{0,j+1}(u_0)\big\|_{\mathcal{H}} \\ &\quad \leq C^{\mathrm{Lip}}_{\tau,j+1,k}\big\|E_{j,j+1}\tilde{E}_{0,j}(u_0) - \tilde{E}_{0,j+1}(u_0)\big\|_{\mathcal{H}}. \end{aligned} \tag{4.2.15}$$

With $\tilde{E}_{0,j}(u_0) = \tilde{u}_j$ and using Assumption 4.2.4, we observe

$$\big\|E_{j,j+1}\tilde{E}_{0,j}(u_0) - \tilde{E}_{0,j+1}(u_0)\big\|_{\mathcal{H}} = \big\|E_{j,j+1}(\tilde{u}_j) - \tilde{E}_{j,j+1}(\tilde{u}_j)\big\|_{\mathcal{H}} \leq \varepsilon_j. \qquad \square$$

Remark 4.2.17. In the description of our abstract setting we have chosen the spaces $\mathcal{H}$ and $\mathcal{G}$ to be the same in all time steps. However, at the expense of a slightly more involved notation, the result of Theorem 4.2.16 stays true with $\mathcal{H}$ replaced by variable spaces $\mathcal{H}_k$, $k = 0, \dots, K-1$, as long as we can guarantee the Lipschitz continuity of the mappings $E_{\tau,j,k} : \mathcal{H}_j \to \mathcal{H}_k$ with corresponding Lipschitz constants $C^{\mathrm{Lip}}_{\tau,j,k}$, $1 \leq j \leq k$.

Based on Theorem 4.2.16 we are now able to state the conditions on the tolerances $(\varepsilon_k)_{k=0,\dots,K-1}$ such that for the scheme (4.2.9) our main goal is achieved.

Theorem 4.2.18. *Suppose that Assumptions 4.2.2, 4.2.4, 4.2.6, and 4.2.12 hold. Let Assumption 4.2.14 hold for some $\delta > 0$. If we consider the case of inexact operator evaluations as described in* (4.2.9) *and choose*

$$0 < \varepsilon_k \leq (C^{\mathrm{Lip}}_{\tau,k+1,K})^{-1}\, \tau^{1+\delta},$$

$k = 0, \dots, K-1$, then we get

$$\big\|u(T) - \tilde{E}_{\tau,0,K}(u(0))\big\|_{\mathcal{H}} \leq (C_{\mathrm{exact}} + T)\, \tau^{\delta}.$$

Proof. Applying Theorem 4.2.16, Assumption 4.2.14 and $K = T/\tau$, we obtain

$$\begin{aligned}\big\|u(t_K) - \tilde{u}_K\big\|_{\mathcal{H}} &\leq \big\|u(t_K) - u_K\big\|_{\mathcal{H}} + \sum_{k=0}^{K-1} C^{\mathrm{Lip}}_{\tau,k+1,K}\,\varepsilon_k \\ &\leq C_{\mathrm{exact}}\,\tau^{\delta} + \sum_{k=0}^{K-1} C^{\mathrm{Lip}}_{\tau,k+1,K}\,(C^{\mathrm{Lip}}_{\tau,k+1,K})^{-1}\,\tau^{1+\delta} \\ &= C_{\mathrm{exact}}\,\tau^{\delta} + K\tau^{1+\delta} = (C_{\mathrm{exact}} + T)\,\tau^{\delta}.\end{aligned}$$

□

One of the final goals of our analysis is to provide upper estimates for the overall complexity of the resulting scheme. As a first step, in this direction, we provide a quite abstract version, which is a direct consequence of Theorem 4.2.18.

Corollary 4.2.19. *Suppose that the assumptions of Theorem 4.2.18 are satisfied. Choose*

$$\varepsilon_k := (C^{\mathrm{Lip}}_{\tau,k+1,K})^{-1}\,\tau^{1+\delta},$$

for $k = 0, \dots, K-1$, then the realization of $\tilde{E}_{\tau,0,K}(u_0)$ requires at most

$$M_{\tau,T}(\delta, (\varepsilon_k)) := \sum_{k=0}^{K-1} M_{\tau,k}(\varepsilon_k, E_{\tau,k,k+1}(\tilde{u}_k))$$

degrees of freedom.

Remark 4.2.20. At this point, without specifying an approximation scheme and therefore without a concrete knowledge of $M_{\tau,k}(\varepsilon,\cdot)$, Corollary 4.2.19 might not look very deep. Nevertheless, it will be filled with content in Section 4.4. There, we will discuss the specific case of adaptive wavelet solvers for which concrete estimates for $M_{\tau,k}(\varepsilon,\cdot)$ are available.

The next step is to play the same game for the inexact scheme (4.2.12). We start again by controlling the error propagation.

Theorem 4.2.21. *Suppose that Assumptions 4.2.6, 4.2.8, 4.2.10, and 4.2.12 hold. Let $(u_k)_{k=0}^K$, $K \in \mathbb{N}$, be the output of the exact scheme (4.2.5), and let $(\tilde{u}_k)_{k=0}^K$ be the output of the inexact scheme (4.2.12) with prescribed tolerances $\varepsilon_{k,i}$, $k = 0, \dots, K-1$, $i = 1, \dots, S$. Then, for all $0 \leq k \leq K$,*

$$\big\|u(t_k) - \tilde{u}_k\big\|_{\mathcal{H}} \leq \big\|u(t_k) - u_k\big\|_{\mathcal{H}} + \sum_{j=0}^{k-1}\Big(\prod_{l=j+1}^{k-1}(C'_{\tau,l,(0)} - 1)\Big)\sum_{i=1}^{S} C'_{\tau,j,(i)}\varepsilon_{j,i}.$$

Proof. We just have to repeat the proof of Theorem 4.2.16 with the special choice (4.2.13) for $\tilde{E}_{\tau,k,k+1}$, and to include two modifications. First, instead of the exact Lipschitz constants $C_{\tau,j+1,k}$ in (4.2.15), we can use their estimates (4.2.11) presented in

Remark 4.2.9. Second, in the last step of the proof of Theorem 4.2.16, we may estimate the error we make when using $\tilde{E}_{\tau,j,j+1}$ instead of $E_{\tau,j,j+1}$ as in Remark 4.2.11(i). Thus, to finish the proof we have to show that the estimates (4.2.11) and (4.2.14) hold.

We start with (4.2.11). Note that it is enough to show that

$$C^{\mathrm{Lip}}_{\tau,k,k+1} \le C'_{\tau,k,(0)} - 1, \quad 0 \le k \le K-1, \tag{4.2.16}$$

since, obviously,

$$C^{\mathrm{Lip}}_{\tau,j,k} \le \prod_{r=j}^{k-1} C^{\mathrm{Lip}}_{\tau,r,r+1}, \quad 0 \le j \le k \le K.$$

Thus, let us prove that (4.2.16) is true, if Assumption 4.2.8 holds. To this end, we fix $k \in \{0,\dots,K-1\}$ as well as arbitrary $u, v \in \mathcal{H}$. Using (4.2.6) and the triangle inequality, we obtain

$$\big\|E_{\tau,k,k+1}(u) - E_{\tau,k,k+1}(v)\big\|_{\mathcal{H}} \le \sum_{i=1}^{S} \|w_{k,i}(u) - w_{k,i}(v)\|_{\mathcal{H}}. \tag{4.2.17}$$

Applying Assumption 4.2.8, we get for each $i \in \{1,\dots,S\}$:

$$\big\|w_{k,i}(u) - w_{k,i}(v)\big\|_{\mathcal{H}} \le C^{\mathrm{Lip}}_{\tau,k,(i)}\Big(\|u-v\|_{\mathcal{H}} + \sum_{l=1}^{i-1} \big\|w_{k,l}(u) - w_{k,l}(v)\big\|_{\mathcal{H}}\Big).$$

Hence, for $r = 0,\dots,S-1$, we have

$$\begin{aligned}\sum_{i=1}^{r+1} \big\|w_{k,i}(u) - w_{k,i}(v)\big\|_{\mathcal{H}} \le \big(1 + C^{\mathrm{Lip}}_{\tau,k,(r+1)}\big) \sum_{i=1}^{r} \big\|w_{k,i}(u) - w_{k,i}(v)\big\|_{\mathcal{H}} \\ + C^{\mathrm{Lip}}_{\tau,k,(r+1)}\|u-v\|_{\mathcal{H}}.\end{aligned} \tag{4.2.18}$$

By induction, it is easy to show that $e_{r+1} \le a_r e_r + b_r$ and $e_0 = 0$ imply

$$e_r \le \sum_{j=1}^{r} b_{j-1} \prod_{l=j+1}^{r} a_{l-1}. \tag{4.2.19}$$

In our situation, this fact leads to the estimate

$$\sum_{i=1}^{S} \big\|w_{k,i}(u) - w_{k,i}(v)\big\|_{\mathcal{H}} \le \sum_{i=1}^{S} C^{\mathrm{Lip}}_{\tau,k,(i)}\|u-v\|_{\mathcal{H}} \prod_{l=i+1}^{S} \big(1 + C^{\mathrm{Lip}}_{\tau,k,(l)}\big),$$

since (4.2.18) holds for $r = 0,\dots,S-1$. Furthermore, we can use the equality

$$\sum_{i=1}^{S} C^{\mathrm{Lip}}_{\tau,k,(i)} \prod_{l=i+1}^{S} (1 + C^{\mathrm{Lip}}_{\tau,k,(l)}) = \prod_{i=1}^{S}(1 + C^{\mathrm{Lip}}_{\tau,k,(i)}) - 1 = C'_{\tau,k,(0)} - 1$$

to obtain

$$\sum_{i=1}^{S} \big\|w_{k,i}(u) - w_{k,i}(v)\big\|_{\mathcal{H}} \le \big(C'_{\tau,k,(0)} - 1\big)\|u - v\|_{\mathcal{H}}.$$

Together with (4.2.17), this proves (4.2.16).

Finally, let us move to the estimate (4.2.14). Fix $k \in \{0, \dots, K-1\}$ and let $\tilde{E}_{\tau,k,k+1}$ be given by (4.2.13) with the prescribed tolerances $\varepsilon_{k,i}$, $i = 1, \dots, S$, from our assertion. Then, for arbitrary $v \in \mathcal{H}$, we have

$$\big\|E_{\tau,k,k+1}(v) - \tilde{E}_{\tau,k,k+1}(v)\big\|_{\mathcal{H}} \le \sum_{i=1}^{S} \big\|w_{k,i}(v) - \tilde{w}_{k,i}(v)\big\|_{\mathcal{H}}. \tag{4.2.20}$$

Using the triangle inequality, as well as Assumption 4.2.8, we obtain for every $i = 1, \dots, S$,

$$\begin{aligned}
&\big\|w_{k,i}(v) - \tilde{w}_{k,i}(v)\big\|_{\mathcal{H}} \\
&= \Big\|L_{\tau,i}^{-1} R_{\tau,k,i}(v, w_{k,1}(v), ..., w_{k,i-1}(v)) \\
&\qquad\qquad - \big[L_{\tau,i}^{-1} R_{\tau,k,i}(v, \tilde{w}_{k,1}(v), ..., \tilde{w}_{k,i-1}(v))\big]_{\varepsilon_{k,i}}\Big\|_{\mathcal{H}} \\
&\le \Big\|L_{\tau,i}^{-1} R_{\tau,k,i}(v, w_{k,1}(v), ..., w_{k,i-1}(v)) - L_{\tau,i}^{-1} R_{\tau,k,i}(v, \tilde{w}_{k,1}(v), ..., \tilde{w}_{k,i-1}(v))\Big\|_{\mathcal{H}} \\
&\quad + \Big\|L_{\tau,i}^{-1} R_{\tau,k,i}(v, \tilde{w}_{k,1}(v), ..., \tilde{w}_{k,i-1}(v)) \\
&\qquad\qquad - \big[L_{\tau,i}^{-1} R_{\tau,k,i}(v, \tilde{w}_{k,1}(v), ..., \tilde{w}_{k,i-1}(v))\big]_{\varepsilon_{k,i}}\Big\|_{\mathcal{H}} \\
&\le C^{\mathrm{Lip}}_{\tau,k,(i)} \sum_{l=1}^{i-1} \big\|w_{k,l}(v) - \tilde{w}_{k,l}(v)\big\|_{\mathcal{H}} + \varepsilon_{k,i}.
\end{aligned}$$

Thus, for $r = 0, \dots, S-1$,

$$\sum_{i=1}^{r+1} \big\|w_{k,i}(v) - \tilde{w}_{k,i}(v)\big\|_{\mathcal{H}} \le \big(1 + C^{\mathrm{Lip}}_{\tau,k,(r+1)}\big) \sum_{i=1}^{r} \big\|w_{k,i}(v) - \tilde{w}_{k,i}(v)\big\|_{\mathcal{H}} + \varepsilon_{k,r+1}.$$

Arguing as above, cf. (4.2.19), we get

$$\sum_{i=1}^{S} \big\|w_{k,i}(v) - \tilde{w}_{k,i}(v)\big\|_{\mathcal{H}} \le \sum_{i=1}^{S} \varepsilon_{k,i} \prod_{l=i+1}^{S} (1 + C^{\mathrm{Lip}}_{\tau,k,(l)}) = \sum_{i=1}^{S} C'_{\tau,k,(i)} \varepsilon_{k,i}.$$

Together with (4.2.20), this proves (4.2.14). □

Remark 4.2.22. By construction, Theorem 4.2.21 is slightly weaker than Theorem 4.2.16, but from the practical point of view Theorem 4.2.21 is more realistic. As already outlined in Section 4.2.2, in many cases, estimates for the Lipschitz constants according to Assumption 4.2.8 and convergent numerical schemes according to Assumption 4.2.10 are available.

Based on Theorem 4.2.21, we are able to state the conditions on the tolerances $\varepsilon_{k,i}$, $k = 0, \dots, K-1$, $i = 1, \dots, S$, such that the scheme (4.2.12) converges with the desired order. We put

$$C''_{\tau,k} := \prod_{l=k+1}^{K-1} \left(C'_{\tau,l,(0)} - 1 \right) \tag{4.2.21}$$

for $k = 0, \dots, K-1$, where $C'_{\tau,l,(0)}$ is given by (4.2.10).

Theorem 4.2.23. *Suppose that Assumptions 4.2.6, 4.2.8, 4.2.10, and 4.2.12 hold. Let Assumption 4.2.14 hold for some $\delta > 0$. If we consider the case of inexact operator evaluations as described in* (4.2.12) *and choose*

$$0 < \varepsilon_{k,i} \leq \frac{1}{S} \Big(C''_{\tau,k} C'_{\tau,k,(i)} \Big)^{-1} \tau^{1+\delta}, \tag{4.2.22}$$

then we get

$$\big\| u(T) - \tilde{u}_K \big\|_{\mathcal{H}} \leq (C_{\text{exact}} + T)\, \tau^{\delta}. \tag{4.2.23}$$

Proof. Applying Theorem 4.2.21, Assumption 4.2.14, and choosing $\varepsilon_{k,i}$ as in (4.2.22), we obtain

$$\begin{aligned} \big\| u(t_K) - \tilde{u}_K \big\|_{\mathcal{H}} &\leq \big\| u(t_K) - u_K \big\|_{\mathcal{H}} + \sum_{k=0}^{K-1} \sum_{i=1}^{S} C''_{\tau,k} C'_{\tau,k,(i)}\, \varepsilon_{k,i} \\ &= (C_{\text{exact}} + T)\, \tau^{\delta}. \end{aligned}$$

□

Remark 4.2.24. **(i)** Let us take a closer look at condition (4.2.22). The number of factors in $C''_{\tau,k}$ is proportional to $K-k$, so that the tolerances are allowed to grow with k (if all factors in $C''_{\tau,k}$ are greater than or equal to 1, which is usually the case). This means that the stage equations at earlier time steps have to be solved with higher accuracy compared to those towards the end of the iteration. Furthermore, the number of factors in $C'_{\tau,k,(i)}$ is proportional to $S - i$, but independent of k. Consequently, also the early stages have to be solved with higher accuracy compared to the later ones.

(ii) In Theorem 4.2.23, i.e., (4.2.22) a specific choice for the tolerances $\varepsilon_{k,i}$, $k = 0, \dots, K-1$, $i = 1, \dots, S$, has been used. Essentially, it is an equilibrium strategy. However, also alternative choices are possible. Indeed, an inspection of the proof of Theorem 4.2.23 shows that any choice of $\varepsilon_{k,i}$ satisfying

$$\sum_{i=1}^{S} C'_{\tau,k,(i)} \varepsilon_{k,i} \leq \left(C''_{\tau,k} \right)^{-1} \tau^{1+\delta}$$

would also be sufficient.

(iii) In practical applications, it would be natural to use the additional flexibility for the choice of $\varepsilon_{k,i}$ as outlined in (ii) to minimize the overall degrees of freedom of the method, given by

$$M_{\tau,T}(\delta) := M_{\tau,T}(\delta, (\varepsilon_{k,i})_{k,i}) := \sum_{k=0}^{K-1} \sum_{i=1}^{S} M_{\tau,k,i}(\varepsilon_{k,i}, \hat{w}_{k,i}), \tag{4.2.24}$$

where for $k = 0, \dots, K-1$, $i = 1, \dots, S$,

$$\hat{w}_{k,i} := L_{\tau,i}^{-1} R_{\tau,k,i}(\tilde{u}_k, \tilde{w}_{k,1}, \dots, \tilde{w}_{k,i-1}), \tag{4.2.25}$$

and $M_{\tau,k,i}(\varepsilon_{k,i}, \hat{w}_{k,i})$ as in Assumption 4.2.10. We will omit the dependency on $(\varepsilon_{k,i})_{k,i}$ whenever the tolerances are clear from the context. This leads to the abstract minimization problem

$$\min_{(\varepsilon_{k,i})_{k,i}} \sum_{k=0}^{K-1} \sum_{i=1}^{S} M_{\tau,k,i}(\varepsilon_{k,i}, \hat{w}_{k,i}) \quad \text{subject to} \quad \sum_{k=0}^{K-1} \sum_{i=1}^{S} C''_{\tau,k} C'_{\tau,k,(i)}\, \varepsilon_{k,i} \leq T\tau^{\delta}.$$

We conclude this section with first applications of Theorem 4.2.18.

Example 4.2.25. Let us continue the example from the very beginning of this section and consider Eq. (4.2.1) in the Gelfand triple $(H_0^1(\Omega), L_2(\Omega), H^{-1}(\Omega))$. We want to interpret the linearly-implicit Euler scheme as an abstract one-stage method with $\mathcal{H} = \mathcal{G} = L_2(\Omega)$. To this end, let

$$\Delta_\Omega^D : D(\Delta_\Omega^D) \subseteq L_2(\Omega) \to L_2(\Omega),$$

denote the *Dirichlet Laplacian* with domain

$$D(\Delta_\Omega^D) := \Big\{ u \in H_0^1(\Omega) : \Delta u := \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2} u \in L_2(\Omega) \Big\},$$

which is defined as in Appendix A.1, starting with the elliptic, symmetric and bounded bilinear form

$$\begin{aligned} a : H_0^1(\Omega) \times H_0^1(\Omega) &\to \mathbb{R} \\ (u, v) &\mapsto a(u, v) := \int_\Omega \langle \nabla u, \nabla v \rangle \, \mathrm{d}x. \end{aligned} \tag{4.2.26}$$

Here, we pick a smooth initial value $u_0 \in D(\Delta_\Omega^D)$, and consider a continuously differentiable function

$$f : [0, T] \times L_2(\Omega) \to L_2(\Omega),$$

which we assume to be Lipschitz continuous in the second variable, uniformly in $t \in [0, T]$. We denote the Lipschitz constant by $C^{\mathrm{Lip},f}$. Since Δ_Ω^D generates a strongly continuous contraction semigroup on $L_2(\Omega)$ (cf. Appendix A.1), Eq. (4.2.1) has a unique classical solution, see, e.g. [98, Theorems 6.1.5 and 6.1.7]. Thus, there exists a unique continuous function $u : [0, T] \to L_2(\Omega)$, continuously differentiable in $(0, T]$, taking only values in $D(\Delta_\Omega^D)$, and fulfilling

$$u(0) = u_0, \quad \text{as well as} \quad u'(t) = \Delta_\Omega^D u(t) + f(t, u(t)), \ \text{for } t \in (0, T).$$

In this setting, we can state the exact linearly-implicit Euler scheme (4.2.2) in the form of an abstract one-stage scheme as follows: With $\mathcal{H} = \mathcal{G} = L_2(\Omega)$ and $\tau = T/K$, we define the operators

$$\begin{aligned} L_{\tau,1}^{-1} : L_2(\Omega) &\to L_2(\Omega) \\ v &\mapsto L_{\tau,1}^{-1} v := (I - \tau \Delta_\Omega^D)^{-1} v, \end{aligned}$$

as well as

$$\begin{aligned} R_{\tau,k,1} : L_2(\Omega) &\to L_2(\Omega) \\ v &\mapsto R_{\tau,k,1}(v) := v + \tau f(t_k, v), \end{aligned}$$

for $k = 0, \dots, K-1$. Then the exact linearly-implicit Euler scheme fits perfectly into the abstract exact scheme (4.2.5) with $S = 1$.

Under our assumptions on the initial value u_0 and the forcing term f, this scheme converges to the exact solution of Eq. (4.2.1) with order $\delta = 1$, i.e., there exists a constant $C_{\text{exact}} > 0$, such that

$$\left\| u(T) - u_K \right\|_{L_2(\Omega)} \leq C_{\text{exact}} \tau^1,$$

see for instance [30]. Therefore, Assumption 4.2.14 is satisfied.

Assumption 4.2.2 can be verified by the following argument: It is well known that for any $\tau > 0$, the operator $L_{\tau,1}^{-1}$ defined above is bounded with norm less than or equal to one (cf. Appendix A.1). Because of the Lipschitz continuity of f, for each $k \in \{0, \dots, K-1\}$, the composition

$$E_{\tau,k,k+1} := L_{\tau,1}^{-1} R_{\tau,k,1} : L_2(\Omega) \to L_2(\Omega)$$

is Lipschitz continuous with Lipschitz constant

$$C_{\tau,k,k+1}^{\text{Lip}} \leq 1 + \tau C^{\text{Lip},f}.$$

Thus, if we define $E_{\tau,j,k} : L_2(\Omega) \to L_2(\Omega)$ for $0 \leq j \leq k \leq K$ as in (4.2.7), these operators are Lipschitz continuous with Lipschitz constants

$$C_{\tau,j,k}^{\text{Lip}} \leq (1 + \tau C^{\text{Lip},f})^{k-j},$$

i.e., Assumption 4.2.2 is fulfilled. Furthermore, these constants can be estimated uniformly for all j, k and τ, since

$$1 \leq C_{\tau,j,k}^{\text{Lip}} \leq (1 + \tau C^{\text{Lip},f})^K \leq \exp(T C^{\text{Lip},f}).$$

Now, let us assume that we have an approximation $\tilde{E}_{\tau,k,k+1}(v)$, $v \in L_2(\Omega)$, such that Assumption 4.2.4 is fulfiled. We want to use the abstract results from above and present a concrete way to choose the tolerances $(\varepsilon_k)_{k=0}^{K-1}$, so that the output $(\tilde{u}_k)_{k=0}^{K}$

of the inexact linearly-implicit Euler scheme (4.2.9) converges to the exact solution with the same order $\delta = 1$. Therefore, if we choose

$$\varepsilon_k \leq \frac{\tau^2}{\exp(TC^{\mathrm{Lip},f})} \quad \text{for } k = 0, \ldots, K-1,$$

we can conclude from Theorem 4.2.18 that the inexact linearly-implicit Euler-scheme (4.2.9) converges to the exact solution of Eq. (4.2.1) with order $\delta = 1$, i.e.,

$$\|u(T) - \tilde{u}_K\|_{L_2(\Omega)} \leq (C_{\mathrm{exact}} + T)\,\tau^1,$$

for all $K \in \mathbb{N}$.

Example 4.2.26. In the situation from Example 4.2.25, let us consider a specific form of $f : (0,T] \times L_2(\Omega) \to L_2(\Omega)$, namely

$$(t, v) \mapsto f(t, v) := \bar{f}(v),$$

where $\bar{f} : \mathbb{R} \to \mathbb{R}$ is continuously differentiable with bounded, strictly negative derivative, i.e., there exists a constant $\bar{B} > 0$, so that

$$-\bar{B} < \frac{\mathrm{d}}{\mathrm{d}x}\bar{f}(x) < 0 \text{ for all } x \in \mathbb{R}.$$

Then, for arbitrary $v_1, v_2 \in L_2(\Omega)$ we get for any $k = 0, \ldots, K-1$,

$$\begin{aligned}
&\left\|L_{\tau,1}^{-1} R_{\tau,k,1}(v_1) - L_{\tau,1}^{-1} R_{\tau,k,1}(v_2)\right\|_{L_2(\Omega)} \\
&\qquad \leq \left\|R_{\tau,k,1}(v_1) - R_{\tau,k,1}(v_2)\right\|_{L_2(\Omega)} \\
&\qquad = \left\|v_1 + \tau\bar{f}(v_1) - (v_2 + \tau\bar{f}(v_2))\right\|_{L_2(\Omega)} \\
&\qquad \leq \sup_{x\in\mathbb{R}} \left|1 + \tau\frac{\mathrm{d}}{\mathrm{d}x}\bar{f}(x)\right| \|v_1 - v_2\|_{L_2(\Omega)}.
\end{aligned}$$

Thus, if $\tau < 2/\bar{B}$, we have a contraction. For $K \in \mathbb{N}$ big enough, and $\varepsilon_k \leq \tau^2$, $k = 0, \ldots, K-1$, we can argue as in Example 4.2.25 to show that

$$\left\|u(T) - \tilde{u}_K\right\|_{L_2(\Omega)} \leq (C_{\mathrm{exact}} + T)\,\tau^1,$$

i.e., the inexact linearly-implicit Euler scheme (4.2.9) again converges to the exact solution of Eq. (4.2.1) with order $\delta = 1$, but for much larger values of ε_k, thus, with much less degrees of freedom.

Remark 4.2.27. In principle, the abstract description of Rothe's method might potentially also be applied to stochastic evolution equations of the form

$$\mathrm{d}u(t) = F(t, u(t))\,\mathrm{d}t + B(u(t))\,\mathrm{d}W(t), \quad u(0) = u_0,$$

where u is a stochastic process and W a Wiener process. By proceeding in this way one would end up with an abstract approximation scheme that provides a unified treatment of deterministic and stochastic equations. However, a detailed elaboration of the stochastic setting would be beyond this manuscript and will therefore be presented in a forthcoming paper.

4.3 Application to linearly-implicit one-step schemes

In this section we substantiate our abstract convergence analysis to the case when Rothe's method is induced by a linearly-implicit S-stage time integrator. We want to compute solutions $u : (0,T] \to V$ to initial value problems of the form (4.1.1) where $F : [0,T] \times V \to V^*$ is a nonlinear right-hand side and $u_0 \in V$ is some initial value. Consequently, we consider the Gelfand triple setting (V, U, V^*).

Essentially this section consists of two parts. First, we show that linearly-implicit S-stage schemes fit nicely into the abstract setting as outlined in Section 4.2 with $\mathcal{H} = V$ and $\mathcal{G} = V^*$, see Observation 4.3.2. In the second part, given in Observation 4.3.10, we analyse the case $\mathcal{H} = \mathcal{G} = U$, since error estimates for the discretization in time are often formulated in the norm of U and then a higher order of convergence might be achieved, cf. Theorem 4.3.6.

In their general form, linearly-implicit S-stage methods are given by

$$u_{k+1} = u_k + \tau \sum_{i=1}^{S} m_i y_{k,i}, \quad k = 0, 1, \dots, K-1, \tag{4.3.1}$$

with S linear *stage equations*

$$(I - \tau\gamma_{i,i} J) y_{k,i} = F\Big(t_k + a_i\tau, u_k + \tau \sum_{j=1}^{i-1} a_{i,j} y_{k,j}\Big) + \sum_{j=1}^{i-1} c_{i,j} y_{k,j} + \tau\gamma_i g, \tag{4.3.2}$$

and

$$a_i := \sum_{j=1}^{i-1} a_{i,j} \sum_{l=1}^{j} \frac{\gamma_{j,l}}{\gamma_{j,j}}, \quad \gamma_i := \sum_{l=1}^{i} \gamma_{i,l}, \tag{4.3.3}$$

for $i = 1, \dots, S$. By J and g we denote (approximations of) the partial derivatives $F_u(t_k, u_k)$ and $F_t(t_k, u_k)$, respectively. This particular choice for a_i ensures that J does not enter the right-hand side of (4.3.2). The parameters $a_{i,j}$, $c_{i,j}$, $\gamma_{i,j}$ and $m_i \neq 0$ have to be suitably chosen according to the desired properties of the scheme.

Remark 4.3.1. If $J = F_u(t_k, u_k)$ and $g = F_t(t_k, u_k)$ are the exact derivatives, the corresponding scheme is also known as a method of *Rosenbrock* type. However, this specific choice of J and g is not needed to derive a convergent time discretization scheme. In the larger class of *W–methods*, J and g are allowed to be approximations to the exact Jacobians. Often one chooses $g = 0$, which is done at the price of a significantly lower order of convergence and a substantially more complicated stability analysis.

First, we consider the case $\mathcal{H} = V$, $\mathcal{G} = V^*$. The scheme (4.3.1) immediately fits into the abstract setting of Section 4.2, as long as we interpret the term u_k as the

solution to an additional 0th stage equation given by the identity operator I on V. Now, if we define

$$\begin{aligned} L_{\tau,i} &: V \to V^*, \\ &\quad v \mapsto (I - \tau\gamma_{i,i}J)v \end{aligned} \tag{4.3.4}$$

and use the right-hand side of the stage equations (4.3.2) to define the operators

$$\begin{aligned} &R_{\tau,k,i} : V \times \cdots \times V \to V^*, \\ &(v_0, \dots, v_{i-1}) \mapsto \tau m_i \Big(F\big(t_k + a_i\tau, v_0 + \sum_{j=1}^{i-1} \frac{a_{i,j}}{m_j} v_j\big) + \sum_{j=1}^{i-1} \frac{c_{i,j}}{\tau m_j} v_j + \tau\gamma_i g \Big), \end{aligned} \tag{4.3.5}$$

for $k = 0, \dots, K-1$ and $i = 1, \dots, S$, then the scheme (4.3.1),(4.3.2),(4.3.3) is related to the abstract Rothe method (4.2.5) as follows.

Observation 4.3.2. *For $k = 0, \dots, K-1$ and $i = 1, \dots, S$ let $L_{\tau,i}$ and $R_{\tau,k,i}$ be defined by* (4.3.4) *and* (4.3.5)*, respectively, and set $L_{\tau,0}^{-1} R_{\tau,k,0} := I_{V \to V}$. Then the linearly-implicit S-stage scheme* (4.3.1),(4.3.2),(4.3.3) *is an abstract $(S+1)$-stage scheme in the sense of* (4.2.5) *with $\mathcal{H} = V$, $\mathcal{G} = V^*$. We have*

$$\begin{aligned} u_{k+1} &:= \sum_{i=0}^{S} w_{k,i}, \\ w_{k,i} &:= L_{\tau,i}^{-1} R_{\tau,k,i}(u_k, w_{k,1}, \dots, w_{k,i-1}), \quad i = 0, \dots, S, \end{aligned}$$

for $k = 0, \dots, K-1$.

Remark 4.3.3. Of course, since the operators $R_{\tau,k,i}$ are derived from the right-hand side F, it might happen that they contain, e.g., nontrivial partial differential operators. Nevertheless, even in this case these differential operators are only *applied* to the current iterate and do not require the numerical solution of an operator equation, and that is why the operators $R_{\tau,k,i}$ can still be interpreted as evaluation operators.

Let us now look at an example, where a simple one-stage scheme of the form (4.3.1),(4.3.2),(4.3.3) with $\mathcal{H} = V$ and $\mathcal{G} = V^*$ is translated into a scheme with $\mathcal{H} = \mathcal{G} = U$.

Example 4.3.4. Let $\Omega \subseteq \mathbb{R}^d$ be a bounded Lipschitz domain. Consider the heat equation (4.2.1) in the Gelfand triple $(H_0^1(\Omega), L_2(\Omega), H^{-1}(\Omega))$ with $\Delta_\Omega^D u + f(t,u) =: F(t,u)$ and $F : [0,T] \times H_0^1(\Omega) \to H^{-1}(\Omega)$. Assume that f fulfils the conditions from Example 4.2.25. The scheme (4.3.1),(4.3.2),(4.3.3) with $S = 1$, $\gamma_{1,1} = m_1 = 1$, $J = \Delta_\Omega^D : H_0^1(\Omega) \to H^{-1}(\Omega)$, and $g = 0$ leads to

$$u_{k+1} = u_k + \tau\big(I - \tau\Delta_\Omega^D\big)^{-1}\big(\Delta_\Omega^D u_k + f(t_k, u_k)\big), \quad k = 0, \dots, K-1,$$

which fits perfectly into the setting of Section 4.2. It can be rewritten as a 2-stage scheme of the form (4.2.5) with $\mathcal{H} = V$ and $\mathcal{G} = V^*$, cf. Observation 4.3.2. However,

since the Dirichlet-Laplacian is not bounded on $L_2(\Omega)$, it can not be understood directly as an S-stage scheme of the form (4.2.5) with $\mathcal{H} = \mathcal{G} = L_2(\Omega)$, but a short computation shows that it can be rewritten as

$$u_{k+1} = \big(I - \tau\Delta_\Omega^D\big)^{-1}\big(u_k + \tau f(t_k, u_k)\big), \quad k = 0, \dots, K-1.$$

Thus, if we start with $u_0 \in D(\Delta_\Omega^D)$, and consider the Dirichlet Laplacian as an unbounded operator on $L_2(\Omega)$, this scheme can be interpreted as an abstract one-stage scheme of the form (4.2.5) with $\mathcal{H} = \mathcal{G} = U$: It is just the linearly-implicit Euler scheme for the heat equation we have already discussed in Example 4.2.25 and, as we have seen, it converges with rate $\delta = 1$. It is worth noting that this result stays true for a wider class of operators A instead of Δ_Ω^D, see [30] for details.

The next step is to discuss the case $\mathcal{H} = \mathcal{G} = U$ in detail. In order to avoid technical difficulties, we restrict the discussion to the case of semi-linear problems (4.1.1) with a right-hand side of the form

$$F : [0,T] \times V \to V^*, \quad F(t,u) := A(t)u + f(t,u), \tag{4.3.6}$$

where $A(t)$ is given for all $t \in (0,T)$ in the sense of Appendix A.1. Furthermore, we will focus on W-methods with the specific choice

$$J(t_k) := A(t_k), \quad g := 0, \tag{4.3.7}$$

in (4.3.2). We restrict our analysis to these methods for the following reasons. First, the linearly-implicit Euler scheme, which is the most important example, is a W-method and not a Rosenbrock method. Second, the choice of J in (4.3.7) avoids the evaluation of the Jacobian in each time step, which might be numerically costly.

In our setting, the overall convergence rate that can be expected is limited by the convergence rate of the exact scheme, cf. Theorem 4.2.23 and Assumption 4.2.14. Therefore, to obtain a reasonable result, it is clearly necessary to discuss the approximation properties of the exact S-stage scheme. To the best of our knowledge, the most far reaching results concerning the convergence of S-stage W-methods for evolution problems have been derived by Lubich and Ostermann [91]. For the reader's convenience, we discuss their results as far as it is needed for our purposes. To do so, we need the following definitions and assumptions.

The method (4.3.1),(4.3.2),(4.3.3) is called $A(\theta)$-*stable* if the related stability function

$$R(z) := 1 + z\,\mathbf{m}^\top \Big(\mathbf{I} - (c_{i,j})_{i,j=1}^S - z\big(\operatorname{diag}(\gamma_{i,i}) + (a_{i,j})_{i,j=1}^S\big)\Big)^{-1}\mathbf{1},$$

where $\mathbf{1}^\top := (1, \dots, 1)^\top$ and $\mathbf{m}^\top := (m_1, \dots, m_S)^\top$, fulfils

$$|R(z)| \le 1 \quad \text{for all } z \in \mathbb{C} \text{ with } |\arg(z)| \ge \pi - \theta.$$

If, additionally, the limit $|R(\infty)| := \lim_{|z|\to\infty} |R(z)| < 1$, then the method is called *strongly* $A(\theta)$-*stable*.

We say that the scheme (4.3.1),(4.3.2),(4.3.3) is of *order* $p \in \mathbb{N}$, if the error of the method, when applied to ordinary differential equations defined on open subsets of $\mathbb{R}^d$ with sufficiently smooth right-hand sides, satisfies

$$\|u(t_k) - u_k\|_{\mathbb{R}^d} \leq C_{\mathrm{ord}}\, \tau^p,$$

uniformly on bounded time intervals.

Assumption 4.3.5. Let $C_{\mathrm{offset}} \geq 0$ and denote $\hat{J}(t) := A(t) + C_{\mathrm{offset}} I$.

(i) For both instances $G(t) := F_u(t, u(t))$ and $G(t) := \hat{J}(t)$ it holds that $G(t) : V \to V^*$, $t \in [0,T]$, is a uniformly bounded family of linear operators in $\mathcal{L}(V, V^*)$. Each $G(t)$ is boundedly invertible and the family $G(t)^{-1}$, $t \in [0,T]$, is uniformly bounded in $\mathcal{L}(V^*, V)$.

(ii) There exist constants $\phi < \pi/2$, $C_i^{\mathrm{sect}} > 0$, $i = 1, 2$ such that for all $t \in [0,T]$ and $z \in \mathbb{C}$ with $|\arg(z)| \leq \pi - \phi$ the operators $zI - F_u(t, u(t))$ and $zI - \hat{J}(t)$ are invertible, and their resolvents are bounded on V, i.e.,

$$\big\|(zI - F_u(t, u(t)))^{-1}\big\|_{\mathcal{L}(V,V)} \leq \frac{C_1^{\mathrm{sect}}}{|z|}, \quad \big\|(zI - \hat{J}(t))^{-1}\big\|_{\mathcal{L}(V,V)} \leq \frac{C_2^{\mathrm{sect}}}{|z|}.$$

(iii) The mapping $t \mapsto F_u(t, u(t)) \in \mathcal{L}(V, V^*)$ is sufficiently often differentiable on $[0,T]$ and fulfils the Lipschitz condition

$$\big\|F_u(t, u(t)) - F_u(t', u(t'))\big\|_{\mathcal{L}(V,V^*)} \leq C_u^F |t - t'| \quad \text{for } 0 \leq t \leq t' \leq T.$$

(iv) The following bounds hold uniformly for v varying in bounded subsets of V and $0 \leq t \leq T$:

$$\big\|F_{tu}(t, v)w\big\|_{V^*} \leq C_{tu}^F \|w\|_V, \quad \big\|F_{uu}(t, v)[w_1, w_2]\big\|_{V^*} \leq C_{uu}^F \|w_1\|_V \|w_2\|_V.$$

(v) There exists a splitting

$$f_u(t, u(t)) =: S_k^{(l)} + S_k^{(r)} \tag{4.3.8}$$

and constants $\mu < 1$, $\beta \geq \mu$ (positive), $C_k^{(l)}$ (sufficiently small) as well as $C_{k,\mu}^{(r)}$, $C_{k,\beta}$, and $C_{k,\beta}^{(r)}$, such that

$$\begin{aligned}
\|S_k^{(l)}\|_{\mathcal{L}(V,V^*)} &\leq C_k^{(l)}, \\
\big\|S_k^{(r)} \hat{J}^{-\mu}(t_k)\big\|_{\mathcal{L}(V^*,V^*)} &\leq C_{k,\mu}^{(r)}, \\
\big\|\hat{J}^{\beta}(t_k)(F_u(t_k, u(t_k)))^{-\beta}\big\|_{\mathcal{L}(V,V)} &\leq C_{k,\beta}, \\
\big\|\hat{J}^{\beta}(t_k) S_k^{(l)} \hat{J}^{-\beta}(t_k)\big\|_{\mathcal{L}(V,V^*)} &\leq C_k^{(l)}, \\
\big\|S_k^{(r)} \hat{J}^{-\beta}(t_k)\big\|_{\mathcal{L}(V,V)} &\leq C_{k,\beta}^{(r)}.
\end{aligned}$$

Now, given above terms, [91, Thm. 6.2] reads as follows.

Theorem 4.3.6. *Suppose that the solution u of Eq. (4.1.1), together with (4.3.6), is unique and has sufficiently regular temporal derivatives. Let Assumption 4.3.5 hold. Suppose that the scheme (4.3.1),(4.3.2),(4.3.3) is a W-method of order $p \geq 2$ that is strongly $A(\theta)$-stable with $\theta > \phi$ and $\phi < \pi/2$, cf. 4.3.5(ii). Let $\beta \in [0,1]$ be as in 4.3.5(v) such that $D(A(t)^\beta)$ is independent of t (with uniformly equivalent norms), $A^\beta u' \in L_2(0,T;V)$. Then the error provided by the numerical solution u_k, $k = 0, \dots, K$ is bounded in $\tau \leq \tau_0$ by*

$$\begin{aligned}
&\Big(\tau \sum_{k=0}^{K} \big\|u_k - u(t_k)\big\|_V^2\Big)^{1/2} + \max_{0\leq k\leq K} \big\|u_k - u(t_k)\big\|_U \\
&\quad \leq C_1^{\mathrm{conv}} \tau^{1+\beta} \Big(C_2^{\mathrm{conv}} + C_1^{\mathrm{conv}} C_k^{(l)}\Big) C_k^{(l)} \Big(\int_0^T \big\|A^\beta u'(t)\big\|_V^2 \,\mathrm{d}t\Big)^{1/2} \\
&\quad + C_1^{\mathrm{conv}} \tau^2 \Big(\int_0^T \big\|A^\beta u'(t)\big\|_V^2 \,\mathrm{d}t + \int_0^T \big\|u''(t)\big\|_V^2 \,\mathrm{d}t + \int_0^T \big\|u'''(t)\big\|_{V^*}^2 \,\mathrm{d}t\Big)^{1/2}.
\end{aligned} \tag{4.3.9}$$

The constants C_1^{conv}, C_2^{conv}, and τ_0 depend on the concrete choice of the W-method, the constants in the assumptions, and on T. The maximal time step size τ_0 depends in addition on the size of the integral terms in (4.3.9).

Remark 4.3.7. As in Theorem 4.3.6, and throughout this section, we assume that a unique solution exists, i.e., Assumption 4.2.12 holds. This is the starting point for our convergence analysis of inexact S-stage schemes. Thus, we will not discuss the solvability and uniqueness theory for PDEs in detail. However, since in the forthcoming examples we will use the results from [91], let us briefly recall which solution concept is used in the following standard situation: Consider a linear operator $A : V \to V^*$ fulfilling the conditions from Appendix A.1, and assume that F in (4.1.1) has the form $F(t,u) := Au + f(t)$. Then, a *weak formulation* of Eq. (4.1.1) is: find

$$u \in C([0,T];U) \cap L_2(0,T;V),$$

such that

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle u(t), v\rangle_U = \langle Au(t), v\rangle_{V^*\times V} + \langle f(t), v\rangle_U \qquad \text{for all } v \in V, t \in (0,T].$$

Before we continue our analysis, let us present a well-known W-method which fulfils the assumptions of Theorem 4.3.6.

Example 4.3.8. For $S = 2$, we present the following scheme taken from [128], which is a strongly $A(\theta)$-stable ($\theta = \pi/2$) W-method of order $p = 2$. It is sometimes called *ROS2* in the literature and is given by

$$u_{k+1} = u_k + \frac{3}{2}\tau y_{k,1} + \frac{1}{2}\tau y_{k,2},$$

where

$$\begin{aligned} y_{k,1} &= \big(I - \tau\tfrac{1}{2+\sqrt{2}}A(t_k)\big)^{-1}\big(A(t_k)u_k + f(t_k,u_k)\big),\\ y_{k,2} &= \big(I - \tau\tfrac{1}{2+\sqrt{2}}A(t_k)\big)^{-1}\big(A(t_k+\tau)(u_k+\tau y_{k,1})\\ &\qquad + f(t_k+\tau, u_k+\tau y_{k,1}) - 2y_{k,1}\big). \end{aligned}$$

It fits into the setting of (4.3.1),(4.3.2),(4.3.3) with $m_1 = 3/2$, $m_2 = 1/2$, $\gamma_{1,1} = \gamma_{2,2} = (2+\sqrt{2})^{-1}$, $a_{2,1} = 1$ and $c_{2,1} = -2$.

To avoid technical difficulties we will restrict the setting of (4.3.6) for the remainder of this section to the special case

$$F : [0,T] \times V \to V^*, \quad F(t,u) := Au + f(t), \tag{4.3.10}$$

where $A : V \to V^*$ is given in the sense of Appendix A.1, and $f : [0,T] \to U$ is a continuously differentiable function. In this case, as already mentioned in Example 4.2.25, Eq. (4.1.1) has a unique classical solution, provided $u_0 \in D(A;U)$, see e.g., [98, Cor. 2.5]. It is worth noting that this unique solution is a also a weak solution in the sense of [91], as addressed in Remark 4.3.7.

Using the abstract results from Section 4.2, we will analyse the inexact S-stage method corresponding to the W-method with

$$J := A \quad \text{and} \quad g := 0. \tag{4.3.11}$$

In the sequel, we will restrict the discussion to the case $S = 2$. This is not a major restriction for the following reason: In Theorem 4.3.6, the maximal convergence order of W-methods is bounded by $\delta = 1 + \beta$, where $\beta \in [0,1]$. In Example 4.3.11 we will show that an F of the form (4.3.10) fulfils Assumption 4.3.5(v) with $\beta = 1$. If we additionally impose the asserted regularity assumptions with $\beta = 1$, see (4.3.17) below, then we can apply Theorem 4.3.6 with $\beta = 1$ to the *ROS2*-method from Example 4.3.8 above (which is a 2-stage method), and get the optimal order in this context.

The structure (4.3.10) of the right-hand side F in (4.1.1), allows the following reformulation of the W-method with (J,g) as in (4.3.11) (a proof can be found in Appendix A.2).

Lemma 4.3.9. *Consider the S-stage W-method given by* (4.3.1),(4.3.2),(4.3.3) *with $S = 2$ and F and (J,g) as in* (4.3.10) *and* (4.3.11)*, respectively. Then, if $\gamma_{i,i} \neq 0$, for $i = 1,2$, we have*

$$u_{k+1} = \Big(1 - \tfrac{m_1}{\gamma_{1,1}} - \tfrac{m_2}{\gamma_{2,2}}\big(1 - \tfrac{a_{2,1}}{\gamma_{1,1}}\big)\Big)u_k + \Big(\tau m_1 - \tau m_2 \tfrac{a_{2,1}}{\gamma_{2,2}}\Big)v_{k,1} + \tau m_2 v_{k,2},$$

where

$$\begin{aligned} v_{k,1} &= L_{\tau,1}^{-1}\big(\tfrac{1}{\tau\gamma_{1,1}}u_k + f(t_k)\big),\\ v_{k,2} &= L_{\tau,2}^{-1}\Big(\big(\tfrac{1}{\tau\gamma_{2,2}}(1 - \tfrac{a_{2,1}}{\gamma_{1,1}}) - \tfrac{c_{2,1}}{\tau\gamma_{1,1}}\big)u_k + \big(\tfrac{a_{2,1}}{\gamma_{2,2}} + c_{2,1}\big)v_{k,1} + f(t_k + a_2\tau)\Big). \end{aligned}$$

Observation 4.3.10. *Note that, if $\gamma_{i,i} \neq 0$, for $i = 1, 2$, and $m_1\gamma_{2,2} \neq m_2 a_{2,1}$, the scheme under consideration perfectly fits into the setting of Section 4.2 with $\mathcal{H} = \mathcal{G} = U$. It can be written in the form of the abstract Rothe method* (4.2.5)*. More precisely,*

$$\left.\begin{aligned} u_{k+1} &= \sum_{i=0}^{2} w_{k,i}, \\ w_{k,i} &:= L_{\tau,i}^{-1} R_{\tau,k,i}(u_k, w_{k,1}, \ldots, w_{k,i-1}), \quad i = 0, 1, 2, \end{aligned}\right\} \tag{4.3.12}$$

with

$$\begin{aligned} L_{\tau,i}^{-1} : U &\to U, \\ v &\mapsto \big(I - \tau\gamma_{i,i}A\big)^{-1} v, \quad \text{for } i = 1, 2, \end{aligned} \tag{4.3.13}$$

the evaluation operators

$$\begin{aligned} R_{\tau,k,1} : U &\to U, \\ v &\mapsto \big(\tfrac{m_1}{\gamma_{1,1}} - \tfrac{m_2 a_{2,1}}{\gamma_{2,2}\gamma_{1,1}}\big)v + \tau\big(m_1 - m_2\tfrac{a_{2,1}}{\gamma_{2,2}}\big) f(t_k), \end{aligned} \tag{4.3.14}$$

as well as

$$\begin{aligned} R_{\tau,k,2} : U \times U &\to U, \\ (v_0, v_1) &\mapsto \big(\tfrac{m_2}{\gamma_{2,2}}(1 - \tfrac{a_{2,1}}{\gamma_{1,1}}) - \tfrac{c_{2,1}m_2}{\gamma_{1,1}}\big)v_0 \\ &\quad + \tfrac{m_2 a_{2,1} + m_2\gamma_{2,2}c_{2,1}}{m_1\gamma_{2,2} - m_2 a_{2,1}} v_1 + \tau m_2 f(t_k + a_2\tau), \end{aligned} \tag{4.3.15}$$

and a 0th step given by

$$\begin{aligned} L_{\tau,0}^{-1} R_{\tau,k,0} : U &\to U, \\ v &\mapsto \big(1 - \tfrac{m_1}{\gamma_{1,1}} - \tfrac{m_2}{\gamma_{2,2}}(1 - \tfrac{a_{2,1}}{\gamma_{1,1}})\big)v. \end{aligned} \tag{4.3.16}$$

This is an immediate consequence of Lemma 4.3.9.

An easy computation, together with the fact that $L_{\tau,1}^{-1}$ and $L_{\tau,2}^{-1}$ are contractions on U (cf. Appendix A.1), yield the Lipschitz constant

$$C_{\tau,k,(1)}^{\mathrm{Lip}} = \big|\tfrac{m_1}{\gamma_{1,1}} - \tfrac{m_2 a_{2,1}}{\gamma_{2,2}\gamma_{1,1}}\big|$$

of $L_{\tau,1}^{-1} R_{\tau,k,1}$. Simultaneously, the Lipschitz constant of $L_{\tau,2}^{-1} R_{\tau,k,2}$ can be estimated as follows:

$$C_{\tau,k,(2)}^{\mathrm{Lip}} \leq \max\Big(\big|\tfrac{m_2}{\gamma_{2,2}}(1 - \tfrac{a_{2,1}}{\gamma_{1,1}}) - \tfrac{m_2 c_{2,1}}{\gamma_{1,1}}\big|, \big|\tfrac{m_2 a_{2,1} + m_2\gamma_{2,2}c_{2,1}}{m_1\gamma_{2,2} - m_2 a_{2,1}}\big|\Big).$$

Note that both constants are independent of k and τ.

Example 4.3.11. As a first step towards the case of inexact operator evaluations we need to check the applicability of Theorem 4.3.6 in the current setting (4.3.10), (4.3.11). Therefore, we now check Assumption 4.3.5. We begin by choosing $C_{\mathrm{offset}} = 0$.

As a consequence it holds that $\hat{J} = F_u(t, u(t)) = A$, independently of t. Assumption 4.3.5(i) holds by the assumptions on A, see Appendix A.1. This, together with the ellipticity assumption (A.1.1) already implies Assumption 4.3.5(ii), see [80]. Further, $A = F_u(t, v)$ is independent of (t, v), and as a consequence Assumptions 4.3.5(iii) and (iv) hold with $C_u^F = C_{tu}^F = C_{uu}^F = 0$. Finally, since J is the exact Jacobian, it is possible to choose $S_k^{(l)} = S_k^{(r)} = 0$ in (4.3.8), such that Assumption 4.3.5(v) holds with $C_k^{(l)} = C_{k,\mu}^{(r)} = C_{k,\beta}^{(r)} = 0$, $C_{k,\beta} = 1$ and $\beta = 1$. Concerning the W-method (4.3.1),(4.3.2),(4.3.3) we assume it to be of order $p \geq 2$ and strongly $A(\theta)$-stable with $\theta > \phi$, where ϕ is as in Assumption 4.3.5(ii). E.g., the scheme from Example 4.3.8 could be employed. If for the solution of Eq. (4.1.1) with F as in (4.3.10) the regularity assumptions

$$Au', u'' \in L_2(0, T; V), \ u''' \in L_2(0, T; V^*) \tag{4.3.17}$$

hold, then we can apply Theorem 4.3.6. Using $C_k^{(l)} = 0$ and $\beta = 1$, the convergence result (4.3.9) reads as

$$\begin{aligned}&\Big(\tau \sum_{k=0}^{K} \big\|u_k - u(t_k)\big\|_V^2\Big)^{1/2} + \max_{0\leq k\leq K} \big\|u_k - u(t_k)\big\|_U \\ &\quad \leq C_1^{\mathrm{conv}} \tau^2 \Big(\int_0^T \big\|Au'(t)\big\|_V^2 \,\mathrm{d}t + \int_0^T \big\|u''(t)\big\|_V^2 \,\mathrm{d}t + \int_0^T \big\|u'''(t)\big\|_{V^*}^2 \,\mathrm{d}t\Big)^{1/2}.\end{aligned}$$

That means, the error measured in the norm $\|\cdot\|_U$ is of order $\delta = 2$.

Example 4.3.12. We employ the method *ROS2* from Example 4.3.8 to our general convergence results for the case of inexact solution of the stage equations, cf. Theorem 4.2.23. First, we present the method in its reformulation on $\mathcal{H} = \mathcal{G} = U$, as given in Observation 4.3.10. Inserting the coefficients

$$m_1 = \tfrac{3}{2}, \ m_2 = \tfrac{1}{2}, \ \gamma_{1,1} = \gamma_{2,2} = (2 + \sqrt{2})^{-1}, \ a_{2,1} = 1, \ \text{and } c_{2,1} = -2$$

into (4.3.12), (4.3.13), (4.3.14), (4.3.15), and (4.3.16) yields

$$\begin{aligned} u_{k+1} &= \sum_{i=0}^{2} w_{k,i}, \\ w_{k,i} &:= L_{\tau,i}^{-1} R_{\tau,k,i}(u_k, w_{k,1}, \ldots, w_{k,i-1}), \quad i = 0, 1, 2, \end{aligned}$$

where the 0th stage vanishes, i.e., $L_{\tau,0}^{-1} R_{\tau,k,0} \equiv 0$,

$$\begin{aligned} L_{\tau,1}^{-1} = L_{\tau,2}^{-1} : U &\to U \\ v &\mapsto \big(I - \tau \tfrac{1}{2+\sqrt{2}} A\big)^{-1} v, \end{aligned}$$

and the evaluation operators are given by

$$\begin{aligned} R_{\tau,k,1} : U &\to U, \\ v &\mapsto -\tfrac{\sqrt{2}}{2} v + \tau \tfrac{1-\sqrt{2}}{2} f(t_k), \end{aligned}$$

and

$$R_{\tau,k,2} : U \times U \to U,$$
$$(v_0, v_1) \mapsto -\tfrac{\sqrt{2}}{2} v_0 + \tfrac{\sqrt{2}}{1-\sqrt{2}} v_1 + \tau \tfrac{1}{2} f(t_k + \tau).$$

This scheme fits perfectly into the abstract Rothe method (4.2.5) with $S = 2$. By Observation 4.3.10, we get the following estimates of the Lipschitz constants of $L_{\tau,i}^{-1} R_{\tau,k,i}$, $i = 1, 2$:

$$C^{\mathrm{Lip}}_{\tau,k,(1)} = \tfrac{\sqrt{2}}{2}, \quad \text{and} \quad C^{\mathrm{Lip}}_{\tau,k,(2)} \leq \max\left(\tfrac{\sqrt{2}}{2}, \tfrac{-\sqrt{2}}{1-\sqrt{2}}\right) \leq \tfrac{\sqrt{2}}{2}.$$

As in Example 4.3.11, we assume that the exact solution u satisfies (4.3.17). Furthermore, we assume we have a method at hand, such that Assumption 4.2.10 is satisfied. Then, by Theorem 4.3.6 and Theorem 4.2.23, if we choose the tolerances $\varepsilon_{k,i}$, for $k = 0, \ldots, K-1$ and $i = 1, 2$, so that they satisfy

$$0 < \varepsilon_{k,i} \leq \tfrac{1}{2} \tau^3 \left(\tfrac{1}{2} + \sqrt{2}\right)^{K-k-1} \prod_{l=i+1}^{2} \left(1 + \tfrac{\sqrt{2}}{2}\right),$$

the corresponding inexact 2-stage scheme (4.2.12) converges with order $\delta = 2$. The computational cost can be estimated by

$$\sum_{k=0}^{K-1} \left(M_{\tau,k,1}(\varepsilon_{k,1}, \hat{w}_{k,1}) + M_{\tau,k,2}(\varepsilon_{k,2}, \hat{w}_{k,2}) \right)$$

with $M_{\tau,k,i}(\cdot,\cdot)$ as in Assumption 4.2.10 and $\hat{w}_{k,i}$ as in Remark 4.2.24(iii).

Remark 4.3.13. For methods of Rosenbrock type, i.e., under the assumption that we use exact Jacobians J and g, a result similar to Theorem 4.3.6 holds. In [91, Theorem 5.2] it is shown that for methods of order $p \geq 3$ and under certain additional regularity assumptions on the exact solution u of Eq. (4.1.1) the error can be bounded similar to (4.3.9) with rate $\tau^{2+\beta}$, $\beta \in [0, 1]$.

4.4 Spatial approximation by wavelet methods

For the inexact Rothe method in Section 4.2 we assumed, cf. Assumption 4.2.10, that we have a numerical solver which enables us to compute the solution of the subproblem arising at the k-th time step and i-th stage up to a prescribed tolerance $\varepsilon_{k,i}$. In practice, this goal can be achieved by employing *adaptive* discretization strategies with a posteriori error control and guaranteed convergence properties, like adaptive discretizations based on finite element or wavelet methods.

Our analysis will focus on the application of adaptive wavelet schemes. As will be explained in Section 4.4.3, there exist adaptive strategies based on wavelets that are guaranteed to converge for a large range of problems. Moreover, they are *asymptotically optimal*, i.e., they (asymptotically) realize the same convergence order as best

m-term wavelet approximation and the computational effort is proportional to the degrees of freedom m.

We start in Section 4.4.1 by introducing the wavelet setting as far as it is needed for our purposes. In Section 4.4.2, we combine estimates for optimal wavelet solvers with the complexity results for the inexact Rothe method (4.2.12) of Section 4.2. For equations of the form (4.4.2), we derive estimates on the degrees of freedom, which are needed to guarantee that the inexact scheme converges with the same order as the exact scheme. In Section 4.4.3 we outline the construction of an optimal adaptive wavelet solver in practice.

As always, $\Omega \subseteq \mathbb{R}^d$, $d \geq 1$, will denote a bounded Lipschitz domain.

4.4.1 Wavelet setting

Let us briefly recall the wavelet setting. In general, a *wavelet* basis $\Psi = \{\psi_\mu : \mu \in \mathcal{J}\}$ is a Riesz basis for $L_2(\Omega)$, that is, there exists two positive constants, c_R and C_R, such that

$$c_R \Big(\sum_{\mu\in\mathcal{J}} |a_\mu|^2 \Big) \leq \Big\| \sum_{\mu\in\mathcal{J}} a_\mu \psi_\mu \Big\|^2 \leq C_R \Big(\sum_{\mu\in\mathcal{J}} |a_\mu|^2 \Big)$$

holds for all $(a_\mu)_{\mu\in\mathcal{J}} \in \ell_2(\mathcal{J})$ and $\operatorname{clos}(\operatorname{span} \psi_\mu) = L_2(\Omega)$. This property is numerically essential, since small errors in the coefficients have a controllable impact on the wavelet expansion. The indices $\mu \in \mathcal{J}$ typically encode several types of information, namely the *scale* (often denoted by $|\mu|$), the *spatial location* as well as the *type* of the wavelet. For instance, on the real line, μ can be identified with a pair of integers (j, k), where $j = |\mu|$ denotes the dyadic refinement level and $2^{-j}k$ the location of the wavelet.

We will ignore any explicit dependence on the type of the wavelet from now on, since this only produces additional constants. Hence, we frequently use $\mu = (j, k)$ and

$$\mathcal{J} = \big\{(j, k) \,:\, j \geq j_0, k \in \mathcal{J}_j\big\},$$

where $\mathcal{J}_j$ is some countable index set and $|(j, k)| = j$. Moreover,

$$\tilde{\Psi} = \{\tilde{\psi} \,:\, \mu \in \mathcal{J}\}$$

denotes the *dual wavelet basis*, which is biorthogonal to Ψ, i.e.,

$$\langle \psi_\mu, \tilde{\psi}_\nu \rangle = \delta_{\mu,\nu}, \quad \mu, \nu \in \mathcal{J}.$$

We will not discuss any technical description of the basis Ψ. Instead, we assume that the domain Ω enables us to construct a wavelet basis Ψ with the following properties:

(W1) The wavelets are *local* in the sense that there exist two constants $c_{\mathrm{loc}}, C_{\mathrm{loc}} > 0$, independent of $\mu \in \mathcal{J}$, such that

$$c_{\mathrm{loc}} 2^{-|\mu|} \leq \operatorname{diam}(\operatorname{supp} \psi_\mu) \leq C_{\mathrm{loc}} 2^{-|\mu|}, \quad \mu \in \mathcal{J}.$$

(W2) The wavelets satisfy the *cancellation property*, i.e., for some parameter $\tilde{m} \in \mathbb{N}$,

$$|\langle v, \psi_\mu \rangle_{L_2(\Omega)}| \leq C_{\mathrm{can}} 2^{-|\mu|(\frac{d}{2}+\tilde{m})} |v|_{W^{\tilde{m}}(L_\infty(\operatorname{supp}\psi_\mu))}$$

for $|\mu| > j_0$, with a constant $C_{\mathrm{can}} > 0$, which does not depend on v and μ.

(W3) The wavelet basis induces characterizations of Besov spaces $B^s_q(L_p(\Omega))$, i.e., there exist constants $c_{\mathrm{norm}}, C_{\mathrm{norm}} > 0$, independent of v, such that

$$\begin{aligned} c_{\mathrm{norm}} \|v\|^q_{B^s_q(L_p(\Omega))} &\leq \sum_{j=j_0}^{\infty} 2^{j(s+d(\frac{1}{2}-\frac{1}{p}))q} \Big(\sum_{\mu \in \mathcal{J}_j} \big|\langle v, \tilde{\psi}_\mu \rangle_{L_2(\Omega)}\big|^p \Big)^{q/p} \\ &\leq C_{\mathrm{norm}} \|v\|^q_{B^s_q(L_p(\Omega))} \end{aligned} \tag{4.4.1}$$

holds for $0 < p, q < \infty$ and all s with $s_1 > s > d(1/p-1)_+$ for some parameter s_1.

In (W3) the upper bound s_1 depends, in particular, on the smoothness and the approximation properties of the wavelet basis.

From now on we always include the following Assumption.

Assumption 4.4.1. There exists a biorthogonal wavelet basis Ψ for $L_2(\Omega)$ that satisfies the properties (W1), (W2), (W3) for a sufficiently large range of parameters s_1, s, p, q and $\tilde{m}$.

Remark 4.4.2. (i) The norm equivalence (4.4.1) and the fact that $B^s_2(L_2(\Omega)) = H^s(\Omega)$ imply that a simple rescaling immediately yields a Riesz basis for $H^s(\Omega)$ with $0 < s < s_1$. We will also assume that Dirichlet boundary conditions can be included, so that a characterization of the type (4.4.1) also holds for $H^s_0(\Omega)$.

(ii) Suitable constructions of wavelet systems on domains can be found, e.g., in [17], [42, 43, 44]. For a detailed discussion we refer to [24].

4.4.2 Complexity estimates for a wavelet-Rothe method

In this section, we study Rothe schemes based on wavelets. In 4.4.2 we combine the abstract analysis presented in Section 4.2 with complexity estimates for adaptive wavelet solvers. We derive estimates for the degrees of freedom, which are needed to guarantee that the inexact scheme (4.2.12) converges with the same order as the exact scheme (4.2.5) within the wavelet setting. As it turns out, among other things, regularity estimates for the exact solution in specific Besov spaces are essential. Then, in 4.4.2, we substantiate the analysis further and discuss regularity estimates for the heat equation. It turns out, that in this case concrete Besov regularity estimates and an explicit estimate of the overall complexity can be derived.

Complexity estimates using adaptive wavelet solvers

To keep the technical difficulties at a reasonable level, we restrict ourselves to parabolic evolution equations of the form

$$u'(t) = A(t)u(t) + f(t, u(t)), \quad t \in (0, T], \quad u(0) = u_0, \tag{4.4.2}$$

where $A : (0, T] \times V \to V^*$, $f : (0, T] \times U \to U$, and (V, U, V^*) is a Gelfand triple with $V = H_0^{\hat{s}}(\Omega)$, $U = L_2(\Omega)$, and $V^* = H^{-\hat{s}}(\Omega)$, $\hat{s} > 0$. So, we are in the setting of Section 4.2 with $\mathcal{H} = H^\nu(\Omega)$ (for some smoothness parameter $0 \le \nu \le \hat{s}$) and $\mathcal{G} \supseteq H^{-\hat{s}}(\Omega)$. Recall, that we assume 4.2.12 and that an exact scheme (4.2.5) is given which satisfies Assumption 4.2.8 and 4.2.14.

We split our analysis into two parts. In the first part, we concentrate on the (rather theoretical) case, where the solutions of the stage equations are approximated by using best m-term wavelet approximation; and the complexity estimate is given in Theorem 4.4.8. Unfortunately, best m-term approximation is not implementable in our case, since the solutions to the subproblems are not known explicitly, so the m largest wavelet coefficients cannot be extracted directly. Therefore, in the second part, we turn our attention to the case where the stage equations are solved numerically by using an implementable wavelet solver which is asymptotically optimal. In Theorem 4.4.10 we show that the complexity estimate, derived in Theorem 4.4.8, immediately extends to this case.

Now to the first part. We apply best m-term wavelet approximation as an approximation scheme in place of Assumption 4.2.10. The error of best m-term wavelet approximation in $H^\nu(\Omega)$ is defined as

$$\sigma_{m,\nu}(v) := \inf\Big\{ \big\| v - \sum_{\mu\in\Lambda} c_\mu \psi_\mu \big\|_{H^\nu(\Omega)} \,:\, c_\mu \in \mathbb{R}, \Lambda \subset \mathcal{J}, \#\Lambda = m \Big\}.$$

The following theorem can be derived from [48, Section 7.7], see also [32, Chapter 7].

Theorem 4.4.3. *Let $\nu \ge 0$ and $v \in B_q^s(L_q(\Omega))$, where*

$$\frac{1}{q} = \frac{s-\nu}{d} + \frac{1}{2}, \quad s > \nu. \tag{4.4.3}$$

Furthermore, let Assumption 4.4.1 hold with $s_1 > s$. Then the error of best m-term wavelet approximation in $H^\nu(\Omega)$ can be estimated as follows:

$$\sigma_{m,\nu}(v) \le C_{\mathrm{nlin}} \, \|v\|_{B_q^s(L_q(\Omega))} \, m^{-\frac{s-\nu}{d}}, \tag{4.4.4}$$

with a constant $C_{\mathrm{nlin}} > 0$, which does not depend on v or m.

Remark 4.4.4. (i) In the case that Ψ is an orthonormal wavelet basis, a best m-term approximation to a function v can be derived by selecting the m largest wavelet coefficients (in absolut value) in the wavelet expansion of v. In the biorthogonal

case, choosing the m largest coefficients yields a best m-term approximation up to a constant. In this sense best m-term approximation is an approximation scheme that fulfils Assumption 4.2.10.

(ii) The scale of Besov spaces $B^s_q(L_q(\Omega))$, where the parameters (s,q) are linked by (4.4.3) for a $\nu \geq 0$ is called the *nonlinear approximation line for approximation in* $H^\nu(\Omega)$. In Fig. 4.1(a) it is displayed in a DeVore-Triebel diagram for the case $\nu = 0$, i.e., $H^\nu(\Omega) = L_2(\Omega)$ and for the case $\nu = 1$.

(iii) We refer to the survey article [48] for a detailed discussion of best m-term wavelet approximation and other nonlinear approximation schemes.

Now, consider the inexact scheme (4.2.12) based on best m-term approximation in each stage. We can apply Theorem 4.2.23 and derive an estimate for the degrees of freedom needed to compute a solution up to a tolerance $(C_{\text{exact}} + T)\,\tau^\delta$.

Lemma 4.4.5. *Suppose that Assumptions 4.2.6, 4.2.8, 4.2.12, and 4.4.1 hold. Let Assumption 4.2.14 hold for some $\delta > 0$ and let the inexact scheme* (4.2.12) *be based on best m-term wavelet approximation with the tolerances given by*

$$\varepsilon_{k,i} := \left(S\, C''_{\tau,k} C'_{\tau,k,(i)}\right)^{-1} \tau^{1+\delta} \tag{4.4.5}$$

with $C'_{\tau,k,(i)}$ as in (4.2.10) *and $C''_{\tau,k}$ as in* (4.2.21). *Let the exact solutions $\hat{w}_{k,i}$ of the stage equations in* (4.2.12), *be given by* (4.2.25), *and assume that all $\hat{w}_{k,i}$ are contained in the same Besov space $B^s_q(L_q(\Omega))$ with* (4.4.3). *Then we have* (4.2.23), *i.e.,*

$$\left\|u(T) - \tilde{u}_K\right\|_{H^\nu(\Omega)} \leq (C_{\text{exact}} + T)\,\tau^\delta,$$

and the number of the degrees of freedom $M_{\tau,T}(\delta)$, given by (4.2.24), *that are needed for the computation of $(\tilde{u}_k)_{k=0}^K$ is bounded from above by*

$$M_{\tau,T}(\delta) \leq \sum_{k=0}^{K-1} \sum_{i=1}^{S} \left\lceil C_{\text{nlin}}^{\frac{d}{s-\nu}} \, \|\hat{w}_{k,i}\|^{\frac{d}{s-\nu}}_{B^s_q(L_q(\Omega))} \left(\left(S\, C''_{\tau,k} C'_{\tau,k,(i)}\right)^{-1} \tau^{1+\delta} \right)^{-\frac{d}{s-\nu}} \right\rceil,$$

with C_{nlin} as in (4.4.4), *and where $\lceil \cdot \rceil$ denotes the upper Gauss-bracket.*

Proof. We are in the setting of Theorem 4.2.23. By Theorem 4.4.3 we may, for each stage equation, choose $m \in \mathbb{N}_0$ as the smallest possible integer, such that

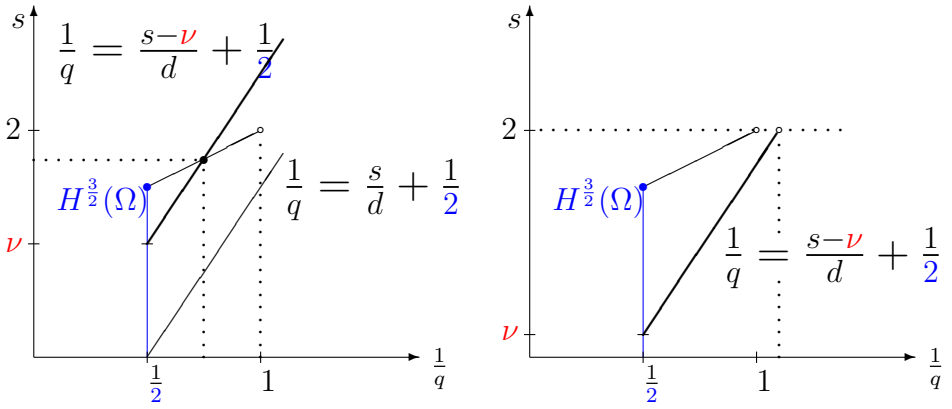
$$\sigma_{m,\nu}(\hat{w}_{k,i}) \leq C_{\text{nlin}} \|\hat{w}_{k,i}\|_{B^s_q(L_q(\Omega))} m^{-\frac{s-\nu}{d}} \leq \varepsilon_{k,i},$$

holds, that is

$$m = \left\lceil \left(C_{\text{nlin}} \|\hat{w}_{k,i}\|_{B^s_q(L_q(\Omega))}\right)^{\frac{d}{s-\nu}} \varepsilon_{k,i}^{-\frac{d}{s-\nu}} \right\rceil.$$

Using (4.4.5) and summing over k and i completes the proof. □

Lemma 4.4.5 shows that we need estimates for the Besov norms of the exact solutions $\hat{w}_{k,i}$ of the stage equations in (4.2.12). We can provide an estimate in the following setting.



(a) $\nu \geq 0$, $(2-\varepsilon) - \frac{d}{2} < \nu < \frac{3}{2}$ (b) $0 \leq \nu \leq (2-\varepsilon) - \frac{d}{2}$

Figure 4.1: DeVore-Triebel diagrams, $d = 3$

Lemma 4.4.6. *Suppose* $L_{\tau,i}^{-1} \in \mathcal{L}(L_2(\Omega), B_q^s(L_q(\Omega)))$ *with* (4.4.3), $i = 1, \dots, S$, *and assume that the operators* $R_{\tau,k,i} : L_2(\Omega) \times \cdots \times L_2(\Omega) \to L_2(\Omega)$ *are Lipschitz continuous with Lipschitz constants* $C_{\tau,k,(i)}^{\mathrm{Lip},R}$ *for all* $k = 0, \dots, K-1$, $i = 1, \dots S$. *With* $C'_{\tau,j,(i)}$ *as in* (4.2.10), *we define*

$$\begin{aligned} C_{k,i}^{\mathrm{Bes}} := \Big(& \prod_{l=1}^{i-1} \big(1+\max(C_{\tau,k,(l)}^{\mathrm{Lip}}, \|L_{\tau,l}^{-1} R_{\tau,k,l}(0,\dots,0)\|_{L_2(\Omega)})\big)\big(1+\|u_k\|_{L_2(\Omega)}\big) \\ & + \prod_{l=1}^{i-1} \big(1 + C_{\tau,k,(l)}^{\mathrm{Lip}}\big) \sum_{j=0}^{k-1} \Big(\prod_{n=j+1}^{k-1} \big(C'_{\tau,n,(0)} - 1\big)\Big) \sum_{r=1}^{S} C'_{\tau,j,(r)} \varepsilon_{j,r} \\ & + \sum_{j=1}^{i-1} \varepsilon_{k,j} \prod_{l=j+1}^{i-1} \big(1 + C_{\tau,k,(l)}^{\mathrm{Lip}}\big)\Big). \end{aligned} \tag{4.4.6}$$

Then all $\hat{w}_{k,i}$, *as defined in* (4.2.25), *are contained in the same Besov space* $B_q^s(L_q(\Omega))$ *with* (4.4.3), *and their norms can be estimated by*

$$\begin{aligned} \|\hat{w}_{k,i}\|_{B_q^s(L_q(\Omega))} \leq & \|L_{\tau,i}^{-1}\|_{\mathcal{L}(L_2(\Omega), B_q^s(L_q(\Omega)))} \\ & \times \max\big(C_{\tau,k,(i)}^{\mathrm{Lip},R}, \|R_{\tau,k,i}(0,\dots,0)\|_{L_2(\Omega)}\big)\, C_{k,i}^{\mathrm{Bes}}. \end{aligned} \tag{4.4.7}$$

Proof. The proof is similar to the proof of Theorem 4.2.21. It can be found in Appendix A.2. □

Remark 4.4.7. In Lemma 4.4.6, the assumption $L_{\tau,i}^{-1} \in \mathcal{L}(L_2(\Omega), B_q^s(L_q(\Omega)))$ with (4.4.3), and the Lipschitz continuity of $R_{\tau,k,i}$ imply Assumption 4.2.8 with $\mathcal{H} = H^\nu(\Omega)$. However, this Lipschitz constant may not be optimal.

The combination of Lemma 4.4.5 and 4.4.6 yields the main result of the first part, i.e., the complexity estimate for the case that best m-term approximations are used for the solution of the stage equations.

Theorem 4.4.8. *Let the assumptions of the Lemmas 4.4.5 and 4.4.6 be satisfied. With $C'_{\tau,k,(i)}$ as in* (4.2.10) *and $C''_{\tau,k}$ as in* (4.2.21)*, we have*

$$\begin{aligned} &M_{\tau,T}(\delta) \\ &\leq \sum_{k=0}^{K-1}\sum_{i=1}^{S}\Big\lceil C_{\mathrm{nlin}}^{\frac{d}{s-\nu}}\Big(\max\big(C^{\mathrm{Lip},R}_{\tau,k,(i)},\|R_{\tau,k,i}(0,\dots,0)\|_{L_2(\Omega)}\big)\,C^{\mathrm{Bes}}_{k,i}\Big)^{\frac{d}{s-\nu}} \\ &\quad\times\big(\|L^{-1}_{\tau,i}\|_{\mathcal{L}(L_2(\Omega),B^s_q(L_q(\Omega)))}\big)^{\frac{d}{s-\nu}}\Big(\big(S\,C''_{\tau,k}C'_{\tau,k,(i)}\big)^{-1}\tau^{1+\delta}\Big)^{-\frac{d}{s-\nu}}\Big\rceil. \end{aligned} \tag{4.4.8}$$

As outlined above, the next step is to discuss the complexity of Rothe's method in the case that *implementable* numerical wavelet schemes instead of the best m-term approximation are employed for the stage equations. We make the following assumptions, cf. Assumption 4.2.10 and Remark 4.2.11(iii).

Assumption 4.4.9. (i) There exists an implementable asymptotically optimal numerical wavelet scheme for the stage equations arising in (4.2.12). That is, if the best m-term approximation in $H^\nu(\Omega)$ converges with rate $m^{-(s-\nu)/d}$, for some $s>\nu>0$, then the scheme computes finite index sets $\Lambda_l\subset\mathcal{J}$ and coefficients $(c_\mu)_{\mu\in\Lambda_l}$ with

$$\Big\|L^{-1}_{\tau,i}v-\sum_{\mu\in\Lambda_l}c_\mu\psi_\mu\Big\|_{H^\nu(\Omega)}\leq C^{\mathrm{asym}}_{\tau,i,s,\nu}(L^{-1}_{\tau,i}v)\,(\#\Lambda_l)^{-\frac{s-\nu}{d}} \tag{4.4.9}$$

for some constant $C^{\mathrm{asym}}_{\tau,i,s,\nu}(L^{-1}_{\tau,i}v)$. Further, for all $\varepsilon>0$ there exists an $l(\varepsilon)$ such that

$$\Big\|L^{-1}_{\tau,i}v-\sum_{\mu\in\Lambda_l}c_\mu\psi_\mu\Big\|_{H^\nu(\Omega)}\leq\varepsilon,\quad l\geq l(\varepsilon),$$

and such that

$$\#\Lambda_{l(\varepsilon)}\leq C^{\mathrm{asym}}_{\tau,i,s,\nu}(L^{-1}_{\tau,i}v)\,\varepsilon^{-\frac{d}{s-\nu}}.$$

(ii) The operators $R_{\tau,k,i}$ can be evaluated exactly.

In Section 4.4.3 we discuss a prototype of an adaptive wavelet method, fulfilling Assumption 4.4.9(i), which has been derived in [26]. It satisfies an optimality estimate of the form (4.4.9) for the energy norm (4.4.22). However, since the energy norm is equivalent to some Sobolev norm $\|\cdot\|_{H^\nu}$, cf. (4.4.23), the estimate (4.4.9) also holds for this case. Moreover, it has been shown in [26] that the constant is of a specific form, which is similar to (4.4.4). Therefore, we specify Assumption 4.4.9(i) in the following way.

Assumption 4.4.9. (iii) The constant $C^{\mathrm{asym}}_{\tau,i,s,\nu}(L^{-1}_{\tau,i}v)$ in (4.4.9) is of the form

$$C^{\mathrm{asym}}_{\tau,i,s,\nu}(L^{-1}_{\tau,i}v)=\hat{C}^{\mathrm{asym}}_{\tau,i}\|L^{-1}_{\tau,i}v\|_{B^s_q(L_q(\Omega))},\quad \frac{1}{q}=\frac{s-\nu}{d}+\frac{1}{2},$$

with a constant $\hat{C}^{\mathrm{asym}}_{\tau,i}$ independent of $L^{-1}_{\tau,i}v$.

In this setting we are immediately able to state our main result.

Theorem 4.4.10. *Let the assumptions of the Lemmas 4.4.5 and 4.4.6 be satisfied. If an optimal numerical wavelet scheme, that satisfies Assumption 4.4.9, is used for the numerical solution of the stage equations, then the necessary degrees of freedom can be estimated as in Theorem 4.4.8 with $\hat{C}_{\tau,i}^{\mathrm{asym}}$ instead of C_{nlin}, i.e.,*

$$\begin{aligned} &M_{\tau,T}(\delta) \\ &\leq \sum_{k=0}^{K-1} \sum_{i=1}^{S} \Big\lceil \big(\hat{C}_{\tau,i}^{\mathrm{asym}}\big)^{\frac{d}{s-\nu}} \Big(\max\big(C_{\tau,k,(i)}^{\mathrm{Lip},R}, \|R_{\tau,k,i}(0,\dots,0)\|_{L_2(\Omega)}\big)\, C_{k,i}^{\mathrm{Bes}} \Big)^{\frac{d}{s-\nu}} \\ &\times \big(\|L_{\tau,i}^{-1}\|_{\mathcal{L}(L_2(\Omega),B_q^s(L_q(\Omega)))}\big)^{\frac{d}{s-\nu}} \Big(\big(S\, C''_{\tau,k} C'_{\tau,k,(i)}\big)^{-1} \tau^{1+\delta} \Big)^{-\frac{d}{s-\nu}} \Big\rceil. \end{aligned} \tag{4.4.10}$$

Remark 4.4.11. The constant $\hat{C}_{\tau,i}^{\mathrm{asym}}$ depends on the concrete design of the adaptive method at hand. As an example this constant may depend on the design of the routines **APPLY**, **RHS** and **COARSE**, as presented in Section 4.4.3. Moreover the value of $\hat{C}_{\tau,i}^{\mathrm{asym}}$ depends on the equivalence constants of the energy norm and the Sobolev norm in (4.4.23). Therefore this constant may grow as τ gets small. However, the reader should observe that this is an intrinsic problem and not caused by our approach.

Now the question arises if and how the Besov norms of the exact solutions of the stage equations $\hat{w}_{k,i}$, cf. (4.4.7) can be specified, and moreover how all the constants involved in (4.4.8) and (4.4.10) can be estimated. In the next subsection we will present a detailed study for the most important model problem, that is the linearly-implicit Euler scheme applied to the heat equation.

Complexity estimates for the heat equation

In this subsection, we conclude the discussion of Example 4.2.25. We derive a concrete estimate of the overall complexity of the linearly-implicit Euler scheme applied to the heat equation in the wavelet setting. Recall,

$$\begin{aligned} u'(t) &= \Delta u(t) + f(t,u(t)) && \text{on } \Omega,\ t \in (0,T], \\ u(0) &= u_0 && \text{on } \Omega, \\ u &= 0 && \text{on } \partial\Omega,\ t \in (0,T], \end{aligned}$$

on a bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$, and consider the case $\mathcal{H} = \mathcal{G} = U = L_2(\Omega)$. The operators $L_{\tau,1}^{-1}$ and $R_{\tau,k,1}$ are given by

$$L_{\tau,1}^{-1} = (I - \tau\Delta_\Omega^D)^{-1}, \quad R_{\tau,k,1} = I + \tau f(t_k,\cdot). \tag{4.4.11}$$

The first step is to estimate the Besov regularity of the solutions to the stage equations. To this end, the mapping properties of $L_{\tau,1}^{-1}$ with respect to the adaptivity

scale of Besov spaces (4.4.3) have to be analysed. Recall that for special cases bounds for the Lipschitz constant of $R_{\tau,k,1} : L_2(\Omega) \to L_2(\Omega)$ have already been proven in the Examples 4.2.25, 4.2.26, i.e.,

$$C^{\mathrm{Lip},R}_{\tau,k,(1)} \leq 1 + \tau C^{\mathrm{Lip},f} \quad \text{and} \quad C^{\mathrm{Lip},R}_{\tau,k,(1)} \leq \sup_{x\in\mathbb{R}} \left| 1 + \tau \frac{\mathrm{d}}{\mathrm{d}x} \bar{f}(x) \right|$$

are shown in Example 4.2.25 and Example 4.2.26, respectively. We put

$$C^{\mathrm{Lap}}_{\mathrm{Bes},\varepsilon} := \left\| (\Delta^D_\Omega)^{-1} \right\|_{\mathcal{L}(L_2(\Omega),B_1^{2-\varepsilon}(L_1(\Omega)))} \tag{4.4.12}$$

and

$$C^{\mathrm{Lap}}_{\mathrm{Sob}} := \left\| (\Delta^D_\Omega)^{-1} \right\|_{\mathcal{L}(L_2(\Omega),H^{3/2}(\Omega))},$$

where $(\Delta^D_\Omega)^{-1} \in \mathcal{L}(L_2(\Omega), B_1^{2-\varepsilon}(L_1(\Omega)))$ has been shown in [34], see also [39, Cor. 1] for details. The fundamental result $(\Delta^D_\Omega)^{-1} \in \mathcal{L}(L_2(\Omega),H^{3/2}(\Omega))$ has been shown in [74, Thm. B].

Lemma 4.4.12. *Let $\varepsilon > 0$. Then the operator $(I-\tau\Delta^D_\Omega)^{-1}$ is contained in the spaces $\mathcal{L}(L_2(\Omega), B_1^{2-\varepsilon}(L_1(\Omega)))$ and $\mathcal{L}(L_2(\Omega), H^{3/2}(\Omega))$. The respective operator norms can be estimated by*

$$\left\| (I - \tau\Delta^D_\Omega)^{-1} \right\|_{\mathcal{L}(L_2(\Omega),B_1^{2-\varepsilon}(L_1(\Omega)))} \leq \frac{1}{\tau} C^{\mathrm{Lap}}_{\mathrm{Bes},\varepsilon} \tag{4.4.13}$$

and

$$\left\| (I - \tau\Delta^D_\Omega)^{-1} \right\|_{\mathcal{L}(L_2(\Omega),H^{3/2}(\Omega))} \leq \frac{1}{\tau} C^{\mathrm{Lap}}_{\mathrm{Sob}}, \tag{4.4.14}$$

respectively.

Proof. We start by proving (4.4.13). The observation

$$(I - \tau\Delta^D_\Omega)^{-1} = (-\tau\Delta^D_\Omega)^{-1}\big(I - (I - \tau\Delta^D_\Omega)^{-1}\big)$$

leads to

$$\left\| (I - \tau\Delta^D_\Omega)^{-1} \right\|_{\mathcal{L}(L_2(\Omega),B_1^{2-\varepsilon}(L_1(\Omega)))} \leq \tau^{-1}\, C^{\mathrm{Lap}}_{\mathrm{Bes},\varepsilon} \left\| I - (I - \tau\Delta^D_\Omega)^{-1} \right\|_{\mathcal{L}(L_2(\Omega))};$$

and the last term can be bounded from above by

$$\begin{aligned}
\left\| I - (I - \tau\Delta^D_\Omega)^{-1} \right\|^2_{\mathcal{L}(L_2(\Omega))} &= \sup_{\|v\|_{L_2(\Omega)}=1} \sum_{k\in\mathbb{N}} \left| (1 - (1-\tau\lambda_k)^{-1}) \langle v, e_k \rangle_{L_2(\Omega)} \right|^2 \\
&= \sup_{\|v\|_{L_2(\Omega)}=1} \sum_{k\in\mathbb{N}} \left| \frac{-\tau\lambda_k}{1-\tau\lambda_k} \langle v, e_k \rangle_{L_2(\Omega)} \right|^2 \\
&\leq \sup_{\|v\|_{L_2(\Omega)}=1} \sum_{k\in\mathbb{N}} \left| \langle v, e_k \rangle_{L_2(\Omega)} \right|^2 \\
&= 1.
\end{aligned}$$

The estimate (4.4.14) follows in a similar fashion. □

With Lemma 4.4.12 at hand, we are now ready to prove the desired mapping properties for $L_{\tau,1}^{-1} : L_2(\Omega) \to B_q^s(L_q(\Omega))$, where (4.4.3) holds. We put

$$C_{\text{inter}}^{\text{Lap}}(\theta) := (C_{\text{Sob}}^{\text{Lap}})^{1-\theta} \, (C_{\text{Bes},\varepsilon}^{\text{Lap}})^{\theta}, \quad \theta \in (0,1). \tag{4.4.15}$$

Lemma 4.4.13. *Let $\varepsilon > 0$, $d \geq 2$, $\nu \geq 0$.* **(i)** *For $(2-\varepsilon) - \frac{d}{2} < \nu < \frac{3}{2}$, that is*

$$\theta := \frac{3-2\nu}{d-1+2\varepsilon} \in (0,1), \tag{4.4.16}$$

we have

$$(I - \tau\Delta_\Omega^D)^{-1} \in \mathcal{L}(L_2(\Omega), B_q^s(L_q(\Omega))) \quad \textit{with} \quad s = \frac{3d - 2\nu + 4\varepsilon\nu}{2d - 2 + 4\varepsilon}$$

and $1/q = (s-\nu)/d + 1/2$. Its norm can be bounded in the following way

$$\left\|(I - \tau\Delta_\Omega^D)^{-1}\right\|_{\mathcal{L}(L_2(\Omega), B_q^s(L_q(\Omega)))} \leq \frac{1}{\tau} C_{\text{inter}}^{\text{Lap}}(\theta). \tag{4.4.17}$$

(ii) *For $0 \leq \nu \leq (2-\varepsilon) - \frac{d}{2}$, we have*

$$(I - \tau\Delta_\Omega^D)^{-1} \in \mathcal{L}(L_2(\Omega), B_q^{2-\varepsilon}(L_q(\Omega))) \quad \textit{with} \quad q = \frac{2d}{4 - 2\varepsilon - 2\nu + d}.$$

and its norm can be bounded by $\tau^{-1} C_{\text{Bes},\varepsilon}^{\text{Lap}}$ as in (4.4.13).

Proof. **(i)** The proof is based on interpolation properties of Besov spaces, see [11] for details. For real interpolation it holds that

$$\left(B_{p_0}^{s_0}(L_{p_0}(\Omega)), B_{p_1}^{s_1}(L_{p_1}(\Omega))\right)_{\bar\theta,p} = B_p^s(L_p(\Omega))$$

in the sense of equivalent (quasi) norms, provided that the parameters satisfy

$$0 < \bar\theta < 1, \quad s = (1-\bar\theta)s_0 + \bar\theta s_1, \quad \frac{1}{p} = \frac{1-\bar\theta}{p_0} + \frac{\bar\theta}{p_1} \tag{4.4.18}$$

and $s_0, s_1 \in \mathbb{R}$, $0 < p_0, p_1 < \infty$. Furthermore, if (4.4.18) holds, a linear operator $\mathcal{T}$ that is contained in $\mathcal{L}(L_2(\Omega), B_{p_0}^{s_0}(L_{p_0}(\Omega)))$ and $\mathcal{L}(L_2(\Omega), B_{p_1}^{s_1}(L_{p_1}(\Omega)))$ is also an element of $\mathcal{L}(L_2(\Omega), B_p^s(L_p(\Omega)))$. Its norm can be estimated by

$$\|\mathcal{T}\|_{\mathcal{L}(L_2(\Omega), B_p^s(L_p(\Omega)))} \leq \|\mathcal{T}\|^{1-\bar\theta}_{\mathcal{L}(L_2(\Omega), B_{p_0}^{s_0}(L_{p_0}(\Omega)))} \|\mathcal{T}\|^{\bar\theta}_{\mathcal{L}(L_2(\Omega), B_{p_1}^{s_1}(L_{p_1}(\Omega)))}.$$

Observe that $H^{3/2}(\Omega) = B_2^{3/2}(L_2(\Omega))$ and that we can apply Lemma 4.4.12. We need to determine the value for $\bar\theta$, such that the resulting interpolation space lies on the nonlinear approximation line $1/p = (s-\nu)/d + 1/2$. This is the case for $\bar\theta = (3-2\nu)/(d-1+2\varepsilon)$, cf. Figure 4.1(a).

(ii) The proof is a combination of (4.4.13) in Lemma 4.4.12 and the continuous embedding of $B_1^{2-\varepsilon}(L_1(\Omega)) \hookrightarrow B_q^{2-\varepsilon}(L_q(\Omega))$. The value of q is determined by the intersection of the lines $s = (2-\varepsilon)$ and $1/p = (s-\nu)/d + 1/2$, cf. Figure 4.1(b). □

Remark 4.4.14. Our findings for the discretization of the heat equation by means of the linearly-implicit Euler scheme carry over to discretizations with $S > 1$ stages. For the case $S = 2$ the operators $L_{\tau,i}^{-1}$, $R_{\tau,k,i}$, $i = 1, 2$, are provided by Observation 4.3.10 and are similar to (4.4.11), e.g.,

$$L_{\tau,i}^{-1} = (I - \tau\gamma_{i,i}\Delta_\Omega^D)^{-1}, \quad i = 1, 2.$$

Lemma 4.4.13 can be reformulated with $\tau\gamma_{i,i}$ replacing τ, and the Lipschitz continuity of $R_{\tau,k,i}$ can be established directly as before.

We are now able to give specific bounds for the degrees of freedom needed to compute the solution of the heat equation by means of the linearly-implicit Euler scheme. Again, we split our analysis into two parts. First, we apply Theorem 4.2.23 to the case when best m-term approximation (with respect to the $H^\nu(\Omega)$ norm, $\nu \geq 0$) is used in each step of the inexact scheme (4.2.12).

Theorem 4.4.15. *Let the assumptions of the Lemmas 4.4.5, 4.4.6 and 4.4.13 hold. Let τ be small enough such that*

$$(1 + \tau C^{\mathrm{Lip},f})^{-1}\tau\big\|f(0)\big\|_{L_2(\Omega)} \leq 1.$$

We put $C_{\mathrm{sup,u}} := \sup_{t\in[0,T]} \|u(t)\|_{L_2(\Omega)}$ and

$$C_{\mathrm{short}}(\tau) := \begin{cases} C_{\mathrm{nlin}}\, C_{\mathrm{Bes},\varepsilon}^{\mathrm{Lap}}\,(1 + \tau C^{\mathrm{Lip},f}) & : \ 0 \leq \nu \leq (2-\varepsilon) - \frac{d}{2}, \\ C_{\mathrm{nlin}}\, C_{\mathrm{inter}}^{\mathrm{Lap}}(\theta)\,(1 + \tau C^{\mathrm{Lip},f}) & : \ (2-\varepsilon) - \frac{d}{2} < \nu < \frac{3}{2},\ \nu > 0, \end{cases}$$

where C_{nlin}, $C_{\mathrm{Bes},\varepsilon}^{\mathrm{Lap}}$, $C_{\mathrm{inter}}^{\mathrm{Lap}}$, and θ are given by (4.4.4), (4.4.12), (4.4.15), *and* (4.4.16), *respectively. Let C_{exact} be given as in Assumption 4.2.14. In the setting of Example 4.2.25, if best m-term wavelet approximation for the spatial approximation of the stage equations is applied, then the degrees of freedom $M_{\tau,T}$ needed to compute a solution up to a tolerance $(C_{\mathrm{exact}} + T)\,\tau$ can be estimated by*

$$M_{\tau,T} \leq T\tau^{-1} + \tfrac{1}{2}\big(2C_{\mathrm{short}}(\tau)\big)^{\frac{2}{\theta}}\big(T^{\frac{2}{\theta}+1}\tau^{-(\frac{2}{\theta}+1)} + C_{\mathrm{lim}}(\tau)\tau^{-(\frac{6}{\theta}+1)}\big),$$

with

$$C_{\mathrm{lim}}(\tau) := \big(1 + C_{\mathrm{sup,u}} + C_{\mathrm{exact}}\tau\big)^{\frac{2}{\theta}}\tau\frac{(1 + \tau C^{\mathrm{Lip},f})^{\frac{2}{\theta}T\tau^{-1}} - 1}{1 - (1 + \tau C^{\mathrm{Lip},f})^{-\frac{2}{\theta}}}.$$

Furthermore,

$$\lim_{\tau\to 0} C_{\mathrm{lim}}(\tau) = \tfrac{\theta}{2}\big(1 + C_{\mathrm{sup,u}}\big)^{\frac{2}{\theta}}(C^{\mathrm{Lip},f})^{-1}\big(\exp\big(C^{\mathrm{Lip},f}\tfrac{2}{\theta}T\big) - 1\big).$$

Proof. We apply Theorem 4.4.8 with $S = 1$ and $\delta = 1$. In the setting of Example 4.2.25 it holds that

$$C_{\tau,k,(1)}^{\mathrm{Lip},R} = 1 + \tau C^{\mathrm{Lip},f}, \quad C'_{\tau,k,(1)} = 1, \quad C'_{\tau,k,(0)} = 2 + \tau C^{\mathrm{Lip},f},$$

independently of k. Thus (4.2.21) reads as $C''_{\tau,k} = (1+\tau C^{\mathrm{Lip},f})^{K-k-1}$ and (4.4.6) can be simplified to

$$C^{\mathrm{Bes}}_{k,1} = 1 + \|u_k\|_{L_2(\Omega)} + k(C^{\mathrm{Lip},R}_{\tau,k,(1)})^{k-K}\tau^2.$$

The norm of $\|u_k\|_{L_2(\Omega)}$ can be bounded as follows. By Assumption 4.2.14 we have $\|u(t_k) - u_k\|_{L_2(\Omega)} \le C_{\mathrm{exact}}\tau$ and as a consequence

$$\|u_k\|_{L_2(\Omega)} \le \big\|u(t_k) - u_k\big\|_{L_2(\Omega)} + \|u(t_k)\|_{L_2(\Omega)} \le C_{\mathrm{exact}}\tau + C_{\mathrm{sup,u}},$$

where $C_{\mathrm{sup,u}}$ is finite since $[0,T]$ is compact and u is continuous. Using the bound (4.4.17) of Lemma 4.4.13(i) in the estimate (4.4.8) we obtain

$$\begin{aligned} M_{\tau,T} &\le \sum_{k=0}^{K-1} \left\lceil \Big(C_{\mathrm{short}}\big((C^{\mathrm{Lip},R}_{\tau,k,(1)})^{K-k}\tau^{-3}(1 + C_{\mathrm{sup,u}} + C_{\mathrm{exact}}\tau) + k\big)\Big)^{\frac{2}{\theta}} \right\rceil \\ &\le K + \sum_{k=0}^{K-1} \Big(C_{\mathrm{short}}\big((C^{\mathrm{Lip},R}_{\tau,k,(1)})^{K-k}\tau^{-3}(1+C_{\mathrm{sup,u}}+C_{\mathrm{exact}}\tau) + k\big)\Big)^{\frac{2}{\theta}}. \end{aligned}$$

An application of Jensen's inequality and the geometric series formula yield

$$\begin{aligned} M_{\tau,T} &\le K + C^{\frac{2}{\theta}}_{\mathrm{short}} 2^{\frac{2}{\theta}-1} \\ &\quad \times \sum_{k=0}^{K-1} \Big(\big((C^{\mathrm{Lip},R}_{\tau,k,(1)})^{K-k}\tau^{-3}(1 + C_{\mathrm{sup,u}} + C_{\mathrm{exact}}\tau)\big)^{\frac{2}{\theta}} + k^{\frac{2}{\theta}} \Big) \\ &\le K\big(1 + \tfrac{1}{2}(2C_{\mathrm{short}}K)^{\frac{2}{\theta}}\big) \\ &\quad + \tau^{-\frac{6}{\theta}}\tfrac{1}{2}\big(2C_{\mathrm{short}}(1 + C_{\mathrm{sup,u}} + C_{\mathrm{exact}}\tau)\big)^{\frac{2}{\theta}} \frac{(1+\tau C^{\mathrm{Lip},f})^{\frac{2}{\theta}K} - 1}{1 - (1+\tau C^{\mathrm{Lip},f})^{-\frac{2}{\theta}}}. \end{aligned}$$

The proof is finalized by the insertion of $K = T\tau^{-1}$ and the observations

$$\begin{aligned} &\lim_{\tau\to 0}(1+\tau C^{\mathrm{Lip},f})^{\frac{2}{\theta}T\tau^{-1}} - 1 = \exp\big(C^{\mathrm{Lip},f}\frac{2}{\theta}T\big) - 1, \\ &\lim_{\tau\to 0} \frac{\tau}{1 - (1+\tau C^{\mathrm{Lip},f})^{-\frac{2}{\theta}}} = \frac{1}{\frac{2}{\theta}C^{\mathrm{Lip},f}}. \end{aligned}$$

The case, where Lemma 4.4.13(ii) is applied to (4.4.8), is analogous. □

Now, we turn to the case when an optimal numerical wavelet scheme is used for the numerical solution of the stage equations in (4.2.12). The wavelet schemes we have in mind are optimal with respect to the energy norm (4.4.22), see Section 4.4.3. In our setting it is induced by L_τ and equivalent to the Sobolev norm $H^1(\Omega)$. For this reason, we now state the estimate for the degrees of freedom in the case of the Sobolev norm $H^1(\Omega)$, i.e., $\nu = 1$.

Theorem 4.4.16. *Let the assumptions of Theorem 4.4.15 hold, whereas we now employ an implementable asymptotically optimal numerical scheme, such that Assumption 4.4.9 holds for $\nu = 1$. Using $\hat{C}_{\mathrm{short}}(\tau) := \hat{C}_{\tau,1}^{\mathrm{asym}} C_{\mathrm{inter}}^{\mathrm{Lap}}(\theta)\,(1 + \tau C^{\mathrm{Lip},f})$, the degrees of freedom needed to compute a solution up to a tolerance $(C_{\mathrm{exact}} + T)\,\tau$ can be estimated by*

$$M_{\tau,T} \leq T\tau^{-1} + \tfrac{1}{2}\big(2\hat{C}_{\mathrm{short}}(\tau)\big)^{\frac{2}{\theta}}\big(T^{\frac{2}{\theta}+1}\tau^{-(\frac{2}{\theta}+1)} + \hat{C}_{\mathrm{lim}}(\tau)\tau^{-(\frac{6}{\theta}+1)}\big), \tag{4.4.19}$$

with

$$\hat{C}_{\mathrm{lim}}(\tau) := \big((1 + C_{\mathrm{sup,u}} + C_{\mathrm{exact}}\tau)\big)^{\frac{2}{\theta}}\tau\frac{(1 + \tau C^{\mathrm{Lip},f})^{\frac{2}{\theta}T\tau^{-1}} - 1}{1 - (1 + \tau C^{\mathrm{Lip},f})^{-\frac{2}{\theta}}}$$

and

$$\theta := \frac{1}{d - 1 + 2\varepsilon}. \tag{4.4.20}$$

Furthermore,

$$\lim_{\tau\to 0}\hat{C}_{\mathrm{lim}}(\tau) = \tfrac{\theta}{2}\big(1 + C_{\mathrm{sup,u}}\big)^{\frac{2}{\theta}}(C^{\mathrm{Lip},f})^{-1}\big(\exp(C^{\mathrm{Lip},f}\tfrac{2}{\theta}T) - 1\big).$$

Remark 4.4.17. (i) The calculations above shows that, among other things, the overall complexity of the resulting scheme heavily depends on the Besov smoothness of the exact solutions to the stage equations. Due to the Lipschitz character of the domain Ω, and since we are working in the L_2-setting, this Besov regularity is limited by $s = 2$. However, for more specific domains, e.g., polygonal domains in $\mathbb{R}^2$ and smoother right-hand sides, much higher Besov smoothness can be achieved, see, e.g., [31], [38] for details. Therefore, for polygonal domains and smoother source terms f we expect that also in our case higher Besov smoothness for the solutions of the stage equations arises, yielding a much lower overall complexity. The details will be discussed in a forthcoming paper.

(ii) Let us further discuss the asymptotic behavior of $M_{\tau,T}$ as τ tends to zero. For simplicity, let us consider the case $d = 2$, then we can choose θ arbitrary close to 1. Asymptotically optimal schemes are usually described in the energy norm induced by the operator $L_{\tau,1}$, whith a constant analogous to (4.4.9) that is independent of $L_{\tau,1}$, see, e.g., [26]. With the notation as (4.4.23) the following consideration for the energy norm induced by $L_{\tau,1}$

$$\big\langle (I + \tau\Delta_\Omega^D)u, u\big\rangle_{L_2(\Omega)} \geq \langle u, u\rangle_{L_2(\Omega)} + \tau\, c_{\mathrm{energy}}^2(\Delta_\Omega^D)\,\|u\|_{H^1(\Omega)}^2,$$

implies $c_{\mathrm{energy}}(I + \tau\Delta_\Omega^D) \geq \tau^{\frac{1}{2}} c_{\mathrm{energy}}(\Delta_\Omega^D)$, so that we can conclude

$$\hat{C}_{\tau,1}^{\mathrm{asym}} = \hat{C}_1\,\tau^{-\frac{1}{2}}$$

with some constant $\hat{C}_1$ independent of τ. In this case (4.4.19) reads as

$$M_{\tau,T} \leq T\tau^{-1} + \frac{1}{2}\big(2\hat{C}_1 C_{\mathrm{inter}}^{\mathrm{Lap}}(1 + \tau C^{\mathrm{Lip},f})\big)^2\big(T^3\tau^{-4} + \hat{C}_{\mathrm{lim}}(\tau)\tau^{-8+\varepsilon'}\big),$$

i.e., for small τ the last term is dominating and therefore the number of degrees of freedom behaves as $\tau^{-8+\varepsilon'}$.

4.4.3 Adaptive wavelet schemes for elliptic problems

We show how wavelets can be used for the adaptive numerical treatment of elliptic operator equations. To be specific, we are interested in equations of the form

$$\mathcal{A}u = f, \tag{4.4.21}$$

where we will assume $\mathcal{A}$ to be a boundedly invertible operator from some Hilbert space V into its normed dual V^*, i.e.,

$$c_{\text{ell}}\|v\|_V \leq \|\mathcal{A}v\|_{V^*} \leq C_{\text{ell}}\|v\|_V, \quad v \in V.$$

Consequently, we are again in a Gelfand triple setting (V, U, V^*). We will only discuss some basic ideas. For further information, the reader is referred to [26], [27], [33]. In our setting, that is the setting of the Rothe method, the operator $\mathcal{A}$ will be one of the operators $L_{\tau,i}$ that arise in the treatment of the stage equations. Therefore, in the applications we have in mind V will always be one of the Sobolev space $H^\nu(\Omega)$ or $H_0^\nu(\Omega)$.

We will focus on the special case where

$$a(v, w) := \langle \mathcal{A}v, w\rangle_{V^*\times V}$$

defines a continuous, symmetric and elliptic bilinear form on V in the sense of (A.1.1). Then, of course, $\mathcal{A}$ corresponds to the operator $-A$ in (A.1.2). In this setting the bilinear form induces a norm on V, the *energy norm*, by setting

$$\|\cdot\| := a(\cdot,\cdot)^{\frac{1}{2}}. \tag{4.4.22}$$

It is equivalent to the Sobolev norm, i.e.,

$$c_{\text{energy}}\|\cdot\|_{H^\nu(\Omega)} \leq \|\cdot\| \leq C_{\text{energy}}\|\cdot\|_{H^\nu(\Omega)}. \tag{4.4.23}$$

Usually, operator equations of the form (4.4.21) are solved by a Galerkin scheme, i.e., one defines an increasing sequence of finite dimensional approximation spaces $S_{\Lambda_l} := \text{span}\{\eta_\mu : \mu \in \Lambda_l\}$, where $S_{\Lambda_l} \subset S_{\Lambda_{l+1}}$, and projects the problem onto these spaces, i.e.,

$$\langle \mathcal{A}u_{\Lambda_l}, v\rangle_{V^*\times V} = \langle f, v\rangle_{V^*\times V} \quad \text{for all} \quad v \in S_{\Lambda_l}.$$

To compute the current Galerkin approximation, one has to solve a linear system

$$\mathbf{G}_{\Lambda_l}\mathbf{c}_{\Lambda_l} = \mathbf{f}_{\Lambda_l},$$

with $\mathbf{G}_{\Lambda_l} := (\langle \mathcal{A}\eta_{\mu'}, \eta_\mu\rangle_{V^*\times V})_{\mu,\mu'\in\Lambda_l}$, $(\mathbf{f}_\Lambda)_\mu := \langle f, \eta_\mu\rangle_{V^*\times V}$, $\mu \in \Lambda_l$.

It is a somewhat delicate task to choose the approximation spaces in the right way. Doing it in an arbitrary way might result in a very inefficient scheme. A natural idea is to use an *adaptive* scheme, i.e., an updating strategy which essentially consists of the following steps

solve	–	estimate	–	refine
$\mathbf{G}_{\Lambda_l}\mathbf{c}_{\Lambda_l} = \mathbf{f}_{\Lambda_l}$		$\|u - u_{\Lambda_l}\| = \ ?$ a posteriori error estimator		add functions if necessary.

The second step is highly nontrivial since the exact solution u is unknown, so that clever *a posteriori* error estimators are needed. An equally challenging task is to show that the refinement strategy leads to a convergent scheme and to estimate its order of convergence, if possible. In recent years, it has been shown that both tasks can be solved if wavelets are used as basis functions for the Galerkin scheme as we will now explain.

The first step is to transform (4.4.21) into a discrete problem. From the norm equivalences (4.4.1) it is easy to see that (4.4.21) is equivalent to

$$\mathbf{A}\mathbf{u} = \mathbf{f},$$

where

$$\mathbf{A} := \mathbf{D}^{-1}\langle \mathcal{A}\Psi, \Psi\rangle_{V^*\times V}^{\top}\mathbf{D}^{-1}, \quad \mathbf{u} := \mathbf{D}\mathbf{c}, \quad \mathbf{f} := \mathbf{D}^{-1}\langle f, \Psi\rangle_{V^*\times V}^{\top},$$

and $\mathbf{D} := (2^{-s|\mu|}\,\delta_{\mu,\mu'})_{\mu,\mu'\in\mathcal{J}}$. Computing a Galerkin approximation amounts to solving the system

$$\mathbf{A}_\Lambda \mathbf{u}_\Lambda = \mathbf{f}_\Lambda := \mathbf{f}|_\Lambda, \qquad \mathbf{A}_\Lambda := (2^{-s(|\mu|+|\nu|)}\langle \psi_\mu, \mathcal{A}\psi_\nu\rangle_{V^*\times V})_{\mu,\nu\in\Lambda}.$$

Now, ellipticity and the norm equivalences (4.4.1) yield

$$\begin{aligned}\|\mathbf{u} - \mathbf{u}_\Lambda\|_{\ell_2(\mathcal{J})} &\le c_{\mathrm{dis}}\|\mathbf{A}(\mathbf{u} - \mathbf{u}_\Lambda)\|_{\ell_2(\mathcal{J})}\\ &\le C_{\mathrm{dis}}\|\mathbf{f} - \mathbf{A}(\mathbf{u}_\Lambda)\|_{\ell_2(\mathcal{J})}\\ &= C_{\mathrm{dis}}\|\mathbf{r}_\Lambda\|_{\ell_2(\mathcal{J})},\end{aligned}$$

so that the $\ell_2(\mathcal{J})$-norm of the *residual* $\mathbf{r}_\Lambda$ serves as an *a posteriori* error estimator. Each individual coefficient $(\mathbf{r}_\Lambda)_\mu$ can be viewed as a local error indicator. Therefore a natural adaptive strategy would consist in catching the bulk of the residual, i.e., to choose the new index set $\hat{\Lambda}$ such that

$$\|\mathbf{r}_\Lambda|_{\hat{\Lambda}}\|_{\ell_2(\mathcal{J})} \ge \zeta\|\mathbf{r}_\Lambda\|_{\ell_2(\mathcal{J})}, \quad \text{for some} \quad \zeta \in (0,1).$$

However, such a scheme cannot be implemented since the residual involves infinitely many coefficients. To transform this idea into an implementable scheme, the following three subroutines can be utilized

(S1) **RHS**$[\varepsilon, \mathbf{g}] \to \mathbf{g}_\varepsilon$ determines for $\mathbf{g} \in \ell_2(\mathcal{J})$ a finitely supported $\mathbf{g}_\varepsilon \in \ell_2(\mathcal{J})$ such that

$$\|\mathbf{g} - \mathbf{g}_\varepsilon\|_{\ell_2(\mathcal{J})} \leq \varepsilon.$$

(S2) **APPLY**$[\varepsilon, \mathbf{G}, \mathbf{v}] \to \mathbf{w}_\varepsilon$ determines for $\mathbf{G} \in \mathcal{L}(\ell_2(\mathcal{J}))$ and for a finitely supported $\mathbf{v} \in \ell_2(\mathcal{J})$ a finitely supported $\mathbf{w}_\varepsilon \in \ell_2(\mathcal{J})$ such that

$$\|\mathbf{G}\mathbf{v} - \mathbf{w}_\varepsilon\|_{\ell_2(\mathcal{J})} \leq \varepsilon.$$

(S3) **COARSE**$[\varepsilon, \mathbf{v}] \to \mathbf{v}_\varepsilon$ determines for a finitely supported $\mathbf{v} \in \ell_2(\mathcal{J})$ a finitely supported $\mathbf{v}_\varepsilon \in \ell_2(\mathcal{J})$ with at most m significant coefficients, such that

$$\|\mathbf{v} - \mathbf{v}_\varepsilon\|_{\ell_2(\mathcal{J})} \leq \varepsilon. \tag{4.4.24}$$

Moreover, $m \leq C m_{\min}$ holds, $m_{\min}$ being the minimal number of entries for which (4.4.24) is valid.

Then, by employing the key idea outlined above, we get the following fundamental algorithm:

Algorithm 4.4.18 SOLVE$[\varepsilon, \mathbf{A}, \mathbf{f}] \to \mathbf{u}_\varepsilon$

$\Lambda_0 := \emptyset$; $\mathbf{r}_{\Lambda_0} := \mathbf{f}$; $\varepsilon_0 := \|\mathbf{f}\|_{\ell_2(\mathcal{J})}$; $j := 0$; $u_0 := 0$;
while $\varepsilon_j > \varepsilon$ **do**
  $\varepsilon_{j+1} := 2^{-(j+1)}\|\mathbf{f}\|_{\ell_2(\mathcal{J})}$; $\Lambda_{j,0} := \Lambda_j$; $\mathbf{u}_{j,0} := \mathbf{u}_j$;
  for $l = 1, ..., L$ **do**
    Compute Galerkin approximation $\mathbf{u}_{\Lambda_{j,l-1}}$ for $\Lambda_{j,l-1}$;
    Compute
      $\tilde{\mathbf{r}}_{\Lambda_{j,l-1}} := \mathbf{RHS}[C_1^{\mathrm{tol}}\varepsilon_{j+1}, \mathbf{f}] - \mathbf{APPLY}[C_1^{\mathrm{tol}}\varepsilon_{j+1}, \mathbf{A}, \mathbf{u}_{\Lambda_{j,l-1}}]$;
    Compute smallest set $\Lambda_{j,l}$,
      such that, $\|\tilde{\mathbf{r}}_{\Lambda_{j,l-1}}|_{\Lambda_{j,l}}\|_{\ell_2(\mathcal{J})} \geq \frac{1}{2}\|\tilde{\mathbf{r}}_{\Lambda_{j,l-1}}\|_{\ell_2(\mathcal{J})}$;
  end for
  COARSE$[C_2^{\mathrm{tol}}\varepsilon_{j+1}, \mathbf{u}_{\Lambda_{j,L}}] \to (\Lambda_{j+1}, \mathbf{u}_{j+1})$;
  $j := j + 1$;
end while

Remark 4.4.19. In [26], it has been shown that Algorithm 4.4.18 exactly fits into the setting of Assumptions 4.4.9. Let us denote by $\Lambda_\varepsilon \subset \mathcal{J}$ the final index set when Algorithm 4.4.18 terminates (the method of updating ε_j ensures termination). Then Algorithm 4.4.18 has the following properties.

(P1) Algorithm 4.4.18 is guaranteed to converge for a huge class of problems, in particular for the differential operators $L_{\tau,i}$ that we have in mind. Denoting with $H^\nu(\Omega)$ the Sobolev space according to (4.4.23), we have

$$\|u - \sum_{\mu \in \Lambda_\varepsilon} c_\mu \psi_\mu\|_{H^\nu(\Omega)} \leq C(u)\varepsilon.$$

(P2) Algorithm 4.4.18 is asymptotically optimal in the sense of Assumption 4.4.9, i.e., with $1/q = (s-\nu)/d + 1/2$, we have

$$\Big\|u - \sum_{\mu\in\Lambda_\varepsilon} c_\mu \psi_\mu\Big\|_{H^\nu(\Omega)} \le \hat{C}^{\mathrm{asym}} \|u\|_{B^s_q(L_q(\Omega))} (\#\Lambda_\varepsilon)^{-\frac{(s-\nu)}{d}}.$$

Remark 4.4.20. (i) We will not discuss the concrete numerical realization of the three fundamental subroutines in detail. The subroutine **COARSE** consists of a thresholding step, whereas **RHS** essentially requires the computation of a best m-term approximation. The most complicated building block is **APPLY**. Let us just mention that its existence can be established for elliptic operators with Schwartz kernels by using the cancellation property of wavelets.

(ii) In Algorithm 4.4.18, C_1^{tol} and C_2^{tol} denote some suitably chosen constants whose concrete values depend on the problem under consideration. The parameter L has to be chosen in a suitable way. We refer again to [26] for details.

(iii) It has been shown in [26] that Algorithm 4.4.18 has the additional property that the number of arithmetic operations stays proportional to the number of unknowns, i.e., the number of floating point operations needed to compute $\mathbf{u}_\varepsilon$ is bounded by $C_{\mathrm{supp}} \# \mathrm{supp}\, \mathbf{u}_\varepsilon$.

A.1 Variational operators

In the preceding sections, we very often considered the same problem on different spaces, e.g., we switched from an operator equation defined on V to the same equation defined on U. In this section we want to clarify in more detail why this is justified.

Let $(V, \langle\cdot,\cdot\rangle_V)$ be a separable real Hilbert space. Furthermore, let

$$a(\cdot,\cdot) : V \times V \to \mathbb{R}$$

be a continuous, symmetric and elliptic bilinear form. This means that there exist two constants $c_{\mathrm{ell}}, C_{\mathrm{ell}} > 0$, such that for arbitrary $u, v \in V$, the bilinear form fulfills the following conditions:

$$c_{\mathrm{ell}} \|u\|_V^2 \le a(u,u), \qquad a(u,v) = a(v,u), \qquad |a(u,v)| \le C_{\mathrm{ell}} \|u\|_V \|v\|_V. \tag{A.1.1}$$

Then, by the Lax-Milgram theorem, the operator

$$\begin{aligned} A : V &\to V^* \\ v &\mapsto Av := -a(v,\cdot) \end{aligned} \tag{A.1.2}$$

is boundededly invertible. Let us now assume that V is densely embedded into a real Hilbert space $(U, \langle\cdot,\cdot\rangle_U)$ via a linear embedding j. We write

$$V \overset{j}{\hookrightarrow} U.$$

Furthermore, we identify the Hilbert space U with its topological dual space U^* via the Riesz isomorphism $U \ni u \mapsto \Phi u := \langle u, \cdot \rangle_U \in U^*$. The adjoint map $j^* : U^* \to V^*$ of j embeds U^* densely into the topological dual V^* of V. All in all we have a so called Gelfand triple (V, U, V^*),

$$V \overset{j}{\hookrightarrow} U \overset{\Phi}{\equiv} U^* \overset{j^*}{\hookrightarrow} V^*.$$

Using $\langle \cdot, \cdot \rangle_{V^* \times V}$ to denote the dual pairs of V and V^*, we have

$$\langle j(v_1), j(v_2) \rangle_U = \langle j^* \, \Phi \, j(v_1), v_2 \rangle_{V^* \times V} \qquad \text{for all } v_1, v_2 \in V. \tag{A.1.3}$$

In this setting, we can consider the operator $A : V \to V^*$ as an unbounded operator on the intermediate space U. More precisely, set

$$D(A) := D(A; U) := \{u \in V \, : \, Au \in j^* \Phi(U)\},$$

and define the operator

$$\begin{aligned} \tilde{A} : D(\tilde{A}) := j(D(A;U)) \subseteq U &\to U \\ u &\mapsto \tilde{A}u := \Phi^{-1} j^{*-1} A \, j^{-1} u. \end{aligned}$$

Such an (unbounded) linear operator is sometimes called *variational*. It is densely defined, since U^* is densely embedded in V^*. Furthermore, the symmetry of the bilinear form $a(\cdot, \cdot)$ implies that $\tilde{A}$ is self-adjoint. At the same time, it is strictly negative definite, because of the ellipticity of a. Moreover, since $A : V \to V^*$ is boundedly invertible, the operator $\tilde{A}^{-1} : U \to U$, defined by $\tilde{A}^{-1} := j A^{-1} j^* \Phi$ is the bounded inverse of $\tilde{A}$. It is compact if the embedding j of V in U is compact.

Let us now fix $\tau > 0$ and consider the bilinear form

$$\begin{aligned} a_\tau : V \times V &\to \mathbb{R} \\ (u, v) &\mapsto a_\tau(u, v) := \tau \langle j(u), j(v) \rangle_U + a(u, v), \end{aligned}$$

which is also continuous, symmetric and elliptic in the sense of (A.1.3). Obviously, for $u, v \in V$, we have the identity

$$\begin{aligned} a_\tau(u, v) &= \tau \langle j^* \Phi j(u), v \rangle_{V^* \times V} - \langle Au, v \rangle_{V^* \times V} \\ &= \langle (\tau j^* \Phi j - A) u, v \rangle_{V^* \times V}, \end{aligned}$$

so that applying again the Lax-Milgram theorem, we can conclude that $(\tau j^* \Phi j - A) : V \to V^*$ is boundedly invertible. Therefore, the operator

$$\begin{aligned} (\tau I - \tilde{A}) : D(\tilde{A}) \subseteq U &\to U \\ u &\mapsto (\tau I - \tilde{A}) u := \tau u - \tilde{A} u, \end{aligned}$$

which coincides with $\Phi^{-1}j^{*-1}(\tau j^*\Phi j - A)j^{-1}$ on $D(\tilde{A})$, possesses a bounded inverse $(\tau I - \tilde{A})^{-1} = j(\tau j^*\Phi j - A)^{-1}j^*\Phi : U \to U$. Thus, the resolvent set $\varrho(\tilde{A})$ of $\tilde{A}$ contains all $\tau \geq 0$. In particular, for any $\tau > 0$, the range of the operator $(\tau I - \tilde{A})$ is all of U. Since, furthermore, $\tilde{A}$ is dissipative, the Lumer-Phillips theorem implies that $\tilde{A}$ generates a strongly continuous semigroup $\{e^{t\tilde{A}}\}_{t\geq 0}$ of contractions on U, see, e.g. [98, Theorem 1.4.3]. Thus, an application of the Hille-Yosida theorem (see, e.g. [98, Theorem 1.3.1]) shows that the operator $L_\tau^{-1} := (I - \tau\tilde{A})^{-1} = \tau(\tau I - \tilde{A})^{-1} : U \to U$ is a contraction for each $\tau > 0$.

By an abuse of notation, we sometimes write A instead of $\tilde{A}$.

A.2 Proofs of Lemma 4.3.9 and Lemma 4.4.6

Proof of Lemma 4.3.9. By (4.3.10) and (4.3.11) the stage equations (4.3.2) read as

$$\begin{aligned}(I - \tau\gamma_{1,1}A)w_{k,1} &= Au_k + f(t_k),\\ (I - \tau\gamma_{2,2}A)w_{k,2} &= A(u_k + \tau a_{2,1}w_{k,1}) + f(t_k + a_2\tau) + c_{2,1}w_{k,1}.\end{aligned}$$

We begin with an application of the following basic observation, that

$$I = (I - CA)^{-1}(I - CA)$$

implies

$$(I - CA)^{-1}A = -\frac{1}{C}I + \frac{1}{C}(I - CA)^{-1}.$$

It follows that

$$\begin{aligned}w_{k,1} &= \big((-\tfrac{1}{\tau\gamma_{1,1}}I + \tfrac{1}{\tau\gamma_{1,1}}(I - \tau\gamma_{1,1}A)^{-1}\big)u_k + (I - \tau\gamma_{1,1}A)^{-1}f(t_k)\\ &= -\tfrac{1}{\tau\gamma_{1,1}}u_k + L_{\tau,1}^{-1}\big(\tfrac{1}{\tau\gamma_{1,1}}u_k + f(t_k)\big).\end{aligned}$$

We denote

$$v_{k,1} = L_{\tau,1}^{-1}\big(\tfrac{1}{\tau\gamma_{1,1}}u_k + f(t_k)\big).$$

A similar computation for the second stage equation yields

$$\begin{aligned}w_{k,2} &= \big(-\tfrac{1}{\tau\gamma_{2,2}}I + \tfrac{1}{\tau\gamma_{2,2}}(I - \tau\gamma_{2,2}A)^{-1}\big)(u_k + \tau a_{2,1}w_{k,1})\\ &\quad + (I - \tau\gamma_{2,2}A)^{-1}\big(f(t_k + a_2\tau) + c_{2,1}w_{k,1}\big)\\ &= -\tfrac{1}{\tau\gamma_{2,2}}\big((1 - \tfrac{a_{2,1}}{\gamma_{1,1}})u_k + \tau a_{2,1}v_{k,1}\big)\\ &\quad + L_{\tau,2}^{-1}\Big(\tfrac{1}{\tau\gamma_{2,2}}\big((1 - \tfrac{a_{2,1}}{\gamma_{1,1}})u_k + \tau a_{2,1}v_{k,1}\big)\\ &\qquad + f(t_k + a_2\tau) + c_{2,1}(-\tfrac{1}{\tau\gamma_{1,1}}u_k + v_{k,1})\Big)\\ &= -\tfrac{1}{\tau\gamma_{2,2}}(1 - \tfrac{a_{2,1}}{\gamma_{1,1}})u_k - \tfrac{a_{2,1}}{\gamma_{2,2}}v_{k,1}\\ &\quad + L_{\tau,2}^{-1}\Big(\big(\tfrac{1}{\tau\gamma_{2,2}}(1 - \tfrac{a_{2,1}}{\gamma_{1,1}}) - \tfrac{c_{2,1}}{\tau\gamma_{1,1}}\big)u_k\\ &\qquad + (\tfrac{a_{2,1}}{\gamma_{2,2}} + c_{2,1})v_{k,1} + f(t_k + a_2\tau)\Big).\end{aligned}$$

We denote

$$v_{k,2} = L_{\tau,2}^{-1}\Big(\big(\tfrac{1}{\tau\gamma_{2,2}}(1-\tfrac{a_{2,1}}{\gamma_{1,1}}) - \tfrac{c_{2,1}}{\tau\gamma_{1,1}}\big)u_k + (\tfrac{a_{2,1}}{\gamma_{2,2}} + c_{2,1})v_{k,1} + f(t_k + a_2\tau)\Big)$$

and arrive at

$$\begin{aligned} u_{k+1} &= u_k + \tau m_1(-\tfrac{1}{\tau\gamma_{1,1}}u_k + v_{k,1}) \\ &\qquad + \tau m_2\big(-\tfrac{1}{\tau\gamma_{2,2}}(1-\tfrac{a_{2,1}}{\gamma_{1,1}})u_k - \tfrac{a_{2,1}}{\gamma_{2,2}}v_{k,1} + v_{k,2}\big) \\ &= \big(1 - \tfrac{m_1}{\gamma_{1,1}} - \tfrac{m_2}{\gamma_{2,2}}(1-\tfrac{a_{2,1}}{\gamma_{1,1}})\big)u_k + (\tau m_1 - \tau m_2\tfrac{a_{2,1}}{\gamma_{2,2}})v_{k,1} + \tau m_2 v_{k,2}. \end{aligned}$$ □

Proof of Lemma 4.4.6. We start with the estimate

$$\begin{aligned} \|\hat{w}_{k,i}\|_{B_q^s(L_q(\Omega))} &= \big\|L_{\tau,i}^{-1}R_{\tau,k,i}(\tilde{u}_k, \tilde{w}_{k,1}, ..., \tilde{w}_{k,i-1})\big\|_{B_q^s(L_q(\Omega))} \\ &\le \|L_{\tau,i}^{-1}\|_{\mathcal{L}(L_2(\Omega),B_q^s(L_q(\Omega)))}\,\big\|R_{\tau,k,i}(\tilde{u}_k, \tilde{w}_{k,1}, ..., \tilde{w}_{k,i-1})\big\|_{L_2(\Omega)}. \end{aligned}$$

The Lipschitz continuity of $R_{\tau,k,i}$ implies the linear growth property

$$\begin{aligned} &\big\|R_{\tau,k,i}(\tilde{u}_k, \tilde{w}_{k,1}, ..., \tilde{w}_{k,i-1})\big\|_{L_2(\Omega)} \\ &\le C_{\tau,k,(i)}^{\mathrm{Lip,R}}\Big(\|\tilde{u}_k\|_{L_2(\Omega)} + \sum_{j=1}^{i-1}\|\tilde{w}_{k,j}\|_{L_2(\Omega)}\Big) + \big\|R_{\tau,k,i}(0, ..., 0)\big\|_{L_2(\Omega)} \\ &\le \max\Big(C_{\tau,k,(i)}^{\mathrm{Lip,R}}, \big\|R_{\tau,k,i}(0, ..., 0)\big\|_{L_2(\Omega)}\Big) \times \Big(1 + \|\tilde{u}_k\|_{L_2(\Omega)} + \sum_{j=1}^{i-1}\|\tilde{w}_{k,j}\|_{L_2(\Omega)}\Big) \\ &\le \max\Big(C_{\tau,k,(i)}^{\mathrm{Lip,R}}, \big\|R_{\tau,k,i}(0, ..., 0)\big\|_{L_2(\Omega)}\Big) \times \Big(1 + \|u_k\|_{L_2(\Omega)} + \sum_{j=1}^{i-1}\|w_{k,j}\|_{L_2(\Omega)} \\ &\qquad + \|u_k - \tilde{u}_k\|_{L_2(\Omega)} + \sum_{j=1}^{i-1}\|w_{k,j} - \tilde{w}_{k,j}\|_{L_2(\Omega)}\Big). \end{aligned}$$

As before, the Lipschitz continuity of $L_{\tau,i}^{-1}R_{\tau,k,i}$ implies

$$\begin{aligned} &\|w_{k,i}\|_{L_2(\Omega)} = \big\|L_{\tau,i}^{-1}R_{\tau,k,i}(u_k, w_{k,1}, \ldots, w_{k,i-1})\big\|_{L_2(\Omega)} \\ &\le \max\Big(C_{\tau,k,(i)}^{\mathrm{Lip}}, \big\|L_{\tau,i}^{-1}R_{\tau,k,i}(0, ..., 0)\big\|_{L_2(\Omega)}\Big) \times \Big(1 + \|u_k\|_{L_2(\Omega)} + \sum_{j=1}^{i-1}\|w_{k,j}\|_{L_2(\Omega)}\Big). \end{aligned}$$

By induction, we estimate

$$\begin{aligned} &1 + \|u_k\|_{L_2(\Omega)} + \sum_{j=1}^{i-1}\|w_{k,j}\|_{L_2(\Omega)} \\ &\le \prod_{l=1}^{i-1}\Big(1 + \max\Big(C_{\tau,k,(l)}^{\mathrm{Lip}}, \big\|L_{\tau,l}^{-1}R_{\tau,k,l}(0, \ldots, 0)\big\|_{L_2(\Omega)}\Big)\Big)\big(1 + \|u_k\|_{L_2(\Omega)}\big). \end{aligned}$$

Note that

$$\|\tilde{w}_{k,i} - \hat{w}_{k,i}\|_{L_2(\Omega)} \leq \|\tilde{w}_{k,i} - \hat{w}_{k,i}\|_{H^\nu(\Omega)} \leq \varepsilon_{k,i}.$$

This enables us to follow similar lines as in the proof of Theorem 4.2.21. We estimate

$$\begin{aligned}
&\|u_k - \tilde{u}_k\|_{L_2(\Omega)} + \sum_{j=1}^{i-1} \|w_{k,j} - \tilde{w}_{k,j}\|_{L_2(\Omega)} \\
&\quad \leq (1 + C^{\mathrm{Lip}}_{\tau,k,(i-1)}) \Big(\|u_k - \tilde{u}_k\|_{L_2(\Omega)} + \sum_{j=1}^{i-2} \|w_{k,j} - \tilde{w}_{k,j}\|_{L_2(\Omega)} \Big) \\
&\qquad + \Big\| L^{-1}_{\tau,i-1} R_{\tau,k,i-1}(\tilde{u}_k, \tilde{w}_{k,1}, \ldots, \tilde{w}_{k,i-2}) \\
&\qquad\qquad - \big[L^{-1}_{\tau,i-1} R_{\tau,k,i-1}(\tilde{u}_k, \tilde{w}_{k,1}, \ldots, \tilde{w}_{k,i-2}) \big]_{\varepsilon_{k,i-1}} \Big\|_{L_2(\Omega)} \\
&\quad \leq (1 + C^{\mathrm{Lip}}_{\tau,k,(i-1)}) \Big(\|u_k - \tilde{u}_k\|_{L_2(\Omega)} + \sum_{j=1}^{i-2} \|w_{k,j} - \tilde{w}_{k,j}\|_{L_2(\Omega)} \Big) + \varepsilon_{k,i-1}
\end{aligned}$$

and conclude by induction

$$\begin{aligned}
&\|u_k - \tilde{u}_k\|_{L_2(\Omega)} + \sum_{j=1}^{i-1} \|w_{k,j} - \tilde{w}_{k,j}\|_{L_2(\Omega)} \\
&\qquad \leq \big(\prod_{l=1}^{i-1} (1 + C^{\mathrm{Lip}}_{\tau,k,(l)}) \big) \|u_k - \tilde{u}_k\|_{L_2(\Omega)} + \sum_{j=1}^{i-1} \varepsilon_{k,j} \prod_{l=j+1}^{i-1} (1 + C^{\mathrm{Lip}}_{\tau,k,(l)}).
\end{aligned}$$

The proof is finished by

$$\|u_k - \tilde{u}_k\|_{L_2(\Omega)} \leq \sum_{j=0}^{k-1} \Big(\prod_{l=j+1}^{k-1} (C'_{\tau,l,(0)} - 1) \Big) \sum_{i=1}^{S} C'_{\tau,j,(i)} \varepsilon_{j,i},$$

which is shown as in Theorem 4.2.21. □

5 Piecewise tensor product wavelet bases by extensions and approximation rates

Authors: Nabi Chegini, Stephan Dahlke, Ulrich Friedrich, Rob Stevenson.
Journal: Mathematics of Computation **82** (2013), no. 284, 2157–2190.
Abstract: Following [*Studia Math.*, 76(2) (1983), pp. 1–58 and 95–136] by Z. Ciesielski and T. Figiel and [*SIAM J. Math. Anal.*, 31 (1999), pp. 184–230] by W. Dahmen and R. Schneider, by the application of extension operators we construct a basis for a range of Sobolev spaces on a domain Ω from corresponding bases on subdomains that form a non-overlapping decomposition. As subdomains, we take hypercubes, or smooth parametric images of those, and equip them with tensor product wavelet bases. We prove approximation rates from the resulting piecewise tensor product basis that are independent of the spatial dimension of Ω. For two- and three-dimensional polytopes we show that the solution of Poisson type problems satisfies the required regularity condition. The dimension independent rates will be realized numerically in linear complexity by the application of the adaptive wavelet-Galerkin scheme.
AMS 2000 subject classification: 15A69, 35B65, 41A25, 41A63, 42C40, 65N12, 65T60.
Key Words: Wavelets, tensor product approximation, domain decomposition, extension operators, weighted anisotropic Sobolev space, regularity, adaptive wavelet scheme, best approximation rates, Fichera corner.

5.1 Introduction

Let $\Omega = \cup_{k=0}^{N} \Omega_k \subset \mathbb{R}^n$ be a non-overlapping domain decomposition. By the use of extension operators, we will construct isomorphisms from the Cartesian product of Sobolev spaces on the subdomains, which incorporate suitable boundary conditions, to Sobolev spaces on Ω. By applying such an isomorphism to the union of Riesz bases for the Sobolev spaces on the subdomains, the result is a Riesz basis for the Sobolev space on Ω.

Since the approach can be applied recursively, to understand the construction of such an isomorphism, it is sufficient to consider the case of having two subdomains. For $i \in \{1, 2\}$, let R_i be the restriction of functions on Ω to Ω_i, let η_2 be the extension

by zero of functions on Ω_2 to functions on Ω, and let E_1 be some extension of functions on Ω_1 to functions on Ω which, for some $m \in \mathbb{N}_0$, is bounded from $H^m(\Omega_1)$ to the target space $H^m(\Omega)$. Then $\begin{bmatrix} R_1 \\ R_2(\mathrm{Id} - E_1 R_1) \end{bmatrix} : H^m(\Omega) \to H^m(\Omega_1) \times H^m_{0,\partial\Omega_1\cap\partial\Omega_2}(\Omega_2)$ is boundedly invertible with inverse $[E_1 \quad \eta_2]$, see Figure 5.1 ($H^m_{0,\partial\Omega_1\cap\partial\Omega_2}(\Omega_2)$ is the space of $H^m(\Omega_2)$ functions that vanish up to order $m-1$ at $\partial\Omega_1 \cap \partial\Omega_2$). Consequently, if

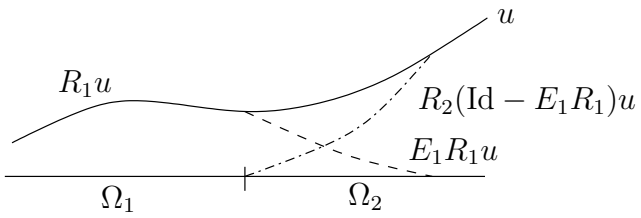


Figure 5.1: Splitting of u into a sum of functions on the subdomains.

Ψ_1 is a Riesz basis for $H^m(\Omega_1)$ and Ψ_2 is a Riesz basis for $H^m_{0,\partial\Omega_1\cap\partial\Omega_2}(\Omega_2)$, then $E_1\Psi_1 \cup \eta_2\Psi_2$ is a Riesz basis for $H^m(\Omega)$.

The principle to construct a basis for a function space on Ω by applying an isomorphism from this space onto the product of corresponding function spaces on non-overlapping subdomains was introduced in [23]. In [44] (see also [82]), this idea was revisited with the aim to practically construct such a basis for doing computations, rather than to show its existence.

In addition to the findings from [44], in the current work we derive precise conditions on the ordering of the subdomains so that the corresponding "true" extension operators (not the trivial zero extensions), being the building blocks of the isomorphism, actually do exist as bounded mappings. To explain this, as an example, consider the construction of a basis for $H^1(\Omega)$ where Ω is an L-shaped domain subdivided into 3 subdomains as illustrated in Figure 5.2. The arrows depict the direction and the or-

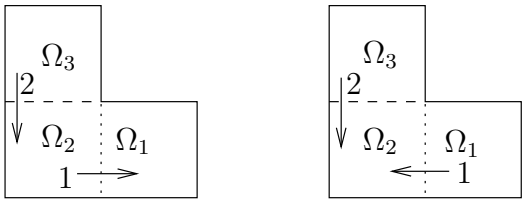


Figure 5.2: A feasible and a non-feasible configuration for $H^1(\Omega)$.

dering of the extensions. The construction requires homogeneous boundary conditions on incoming interfaces and no boundary conditions on outgoing interfaces. In the left case we begin by constructing a basis for $H^1_{0,\partial\Omega_2\cap\partial\Omega_3}(\Omega_1 \cup \Omega_2)$ as the union of a basis for

$H^1_{0,\partial\Omega_1\cap\partial\Omega_2}(\Omega_1)$ and the image of a basis for $H^1_{0,\partial\Omega_2\cap\partial\Omega_3}(\Omega_2)$ under the first extension, which has to be bounded as an operator from $H^1_{0,\partial\Omega_2\cap\partial\Omega_3}(\Omega_2)$ to $H^1_{0,\partial\Omega_2\cap\partial\Omega_3}(\Omega_1\cup\Omega_2)$. The full basis is constructed by adding the image of a basis for $H^1(\Omega_3)$ under the second extension, which needs to be a bounded operator from $H^1(\Omega_3)$ to $H^1(\Omega)$.

Choosing the action of the extension operators as illustrated in the right case yields an invalid configuration. This is due to the fact that in the first step we would need a bounded extension operator from $H^1(\Omega_1)$ to $H^1_{0,\partial\Omega_2\cap\partial\Omega_3}(\Omega_1\cup\Omega_2)$. In view of the boundary condition incorporated in the latter space, this is, however, impossible.

The conditions on the directions of the arrows depend on the boundary conditions imposed on $\partial\Omega$, e.g., they will be different when a basis for $H^1_0(\Omega)$ is sought.

Our main interest in the construction of a basis from bases from subdomains lies in the use of *piecewise tensor product approximation*. On the hypercube

$$\square := (0,1)^n,$$

one can construct a basis for the Sobolev space $H^m(\square)$ (or for a subspace incorporating Dirichlet boundary conditions) by taking an n-fold tensor product of a collection of univariate functions that forms a Riesz basis for $L_2(0,1)$ as well as, properly scaled, for $H^m(0,1)$. Thinking of a univariate wavelet basis of order $d > m$, the advantage of this approach is that the rate of nonlinear best M-term approximation of a sufficiently smooth function u is $d-m$, compared to $\frac{d-m}{n}$ for standard best M-term isotropic (wavelet) approximation of order d on $\square$. The multiplication of the one-dimensional rate $d-m$ by the factor $\frac{1}{n}$ is commonly referred to as the *curse of dimensionality*.

One may argue that for any fixed n, a rate $d-m$ can also be obtained by isotropic approximation by increasing the order from d to $nd-(n-1)m$. Concerning the required smoothness of u, however, in the latter case it is (essentially) necessary and sufficient that for $1 \le i \le n$, $0 \le k \le m$, it holds that $\partial^\alpha \partial_i^k u \in L_p(\square)$ for $p = (d-m+\frac{1}{2})^{-1}$ and $\|\alpha\|_1 \le n(d-m)$, where α denotes a multiindex, i.e., $\alpha \in \mathbb{N}_0^n$. With tensor product approximation the last condition reads as the much milder one $\|\alpha\|_\infty \le d-m$ (a precise formulation of the smoothness conditions in terms of (tensor products of) Besov spaces can be found in [97, 115]).

Actually, the above conditions guarantee only any rate $s < d-m$. Arguments from interpolation space theory that are used do not give a result for the "endpoint" $s = d-m$.

In any case, for dimensions $n \ge 3$, the solution of an elliptic boundary value problem of order $2m = 2$ generally does *not* satisfy the conditions such that isotropic approximation converges with the best, or any near best possible rate allowed by the polynomial order, i.e., $\frac{d-m}{n}$ for order d. In order to achieve this rate, generally anisotropic approximation is mandatory (cf. [2]).

In addition to avoiding the curse of dimensionality, the possibility of anisotropic approximation is automatically included in (adaptive) tensor product approximation. In [47], see also [96], it was shown that best approximations of u from a suitably chosen nested sequence of spaces spanned by tensor product wavelets realizes the

best possible rate $d-m$, so not only any near best possible rate, when for $1 \leq i \leq n$, $0 \leq k \leq m$ and $\|\alpha\|_\infty \leq d-m$, $\partial^\alpha \partial_i^k u$ is in a *weighted* $L_2(\square)$ space, with a weight being an n-fold product of univariate weights on $(0,1)$ that vanish at the endpoints. Clearly, the optimal rate $d-m$ for this linear approximation scheme implies this rate for the nonlinear best M-term approximation from the tensor product basis. What is more, in [47] it was shown that for a sufficiently smooth right-hand side, the solution of Poisson's problem on the n-dimensional unit cube $\square$ satisfies this regularity condition.

In view of these results on $\square$, we consider a domain Ω whose closure is the union of subdomains $\tau + \overline{\square}$ for some $\tau \in \mathbb{Z}^n$, or a domain Ω that is a parametric image of such a domain under a piecewise sufficiently smooth, globally C^{m-1} diffeomorphism κ, being a homeomorphism when $m = 1$. We equip $H^m(\Omega)$ (or a subspace incorporating Dirichlet boundary conditions) with a Riesz basis that is constructed using extension operators as discussed before from tensor product wavelet bases of order d on the subdomains, or from push-forwards of such bases. Our restriction to decompositions of Ω into subdomains from a topological Cartesian partition will allow us to rely on univariate extensions. We will show the best possible approximation rate $d-m$ for any u that restricted to any of these subdomains has a pull-back whose derivatives of sufficiently high order are in the aforementioned weighted $L_2(\square)$-spaces. The latter proof turns out to be technically hard. Indeed, in order to end up with locally supported wavelets, we will apply local, scale-dependent extension operators – i.e., only wavelets that are non-zero near an interface will be extended, – which do not preserve more smoothness than essentially membership of H^m.

Furthermore, using anisotropic regularity results recently shown in [29], we show that if, additionally, Ω is a two- or, more interesting, a three-dimensional polytope, then for a sufficiently smooth right-hand side, the solution of Poisson's problem satisfies this piecewise regularity condition. For that to hold in three dimensions, it will be needed that the parametrization map κ is piecewise trilinear, and it may require a refinement of the initial decomposition of Ω.

Since it defines a boundedly invertible mapping from a Hilbert space, being $H_0^1(\Omega)$, to its dual, the Poisson problem is an example of a well-posed operator equation. Equipping $H_0^1(\Omega)$ with a Riesz basis constructed using extension operators from tensor product wavelet bases of order d on the subdomains, the operator equation is equivalently formulated as a boundedly invertible bi-infinite matrix vector equation. Approximate solutions produced by the adaptive wavelet-Galerkin method ([26, 121]) were proven to converge with the best possible rate in linear complexity. We perform numerical tests in two and three dimensions with wavelets of order $d = 5$ that confirm that this rate is $d-m$.

This paper is organized as follows: In Sect. 5.2, we present the abstract idea behind the construction of isomorphisms from a Sobolev space on a domain onto the product of corresponding Sobolev spaces on subdomains that form a non-overlapping decomposition.

In Sect. 5.3, we recall results on tensor product approximation on a hypercube, and collect assumptions on the univariate wavelets, being the building blocks of the tensor product wavelets.

In Sect. 5.4, we consider a domain Ω that is the union of hypercubes from a Cartesian partition of $\mathbb{R}^n$ into hypercubes. We formulate precise conditions on the order in which univariate extensions over interfaces have to applied, and which boundary conditions have to be imposed, such that for a range of smoothness indices the composition of these extensions is an isomorphism from a Sobolev spaces on Ω onto the product of the corresponding Sobolev spaces on the collection of hypercubes. Equipping these hypercubes with tensor product wavelet bases, we end up with a piecewise tensor product wavelet basis on Ω.

In order to obtain locally supported primal and dual wavelets, in Sect. 5.5 the extension operators are replaced by scale-dependent modifications, in the sense that only wavelets with supports "near" the interfaces are extended. It is shown that approximation from the resulting piecewise tensor product basis gives rise to rates that are independent of the spatial dimension, assuming the function that is approximated satisfies some mild, piecewise weighted Sobolev smoothness conditions.

In Sect. 5.6, these regularity conditions are verified for the solution of Poisson's problem with sufficiently smooth right-hand side in two and three-dimensional polytopes.

The best possible rates from the piecewise tensor product basis can be realized in linear complexity by the application of the adaptive wavelet-Galerkin scheme. In Sect. 5.7, we present numerical results obtained with this scheme for the two-dimensional slit domain, and the three-dimensional thick L-shaped domain and the Fichera corner domain.

5.2 Construction of the isomorphisms

In an abstract setting, for a class of mappings from a Banach space to the Cartesian product of two other Banach spaces, we give conditions on such mappings to be isomorphisms. The results will be applied to construct isomorphisms from a Sobolev space on a domain onto the product of Sobolev spaces on subdomains.

Proposition 5.2.1. *For normed linear spaces V and V_i $(i = 1, 2)$, let $E_1 \in B(V_1, V)$, $\eta_2 \in B(V_2, V)$, $R_1 \in B(V, V_1)$, and $R_2 \in B(\Im\eta_2, V_2)$ be such that*

$$R_1 E_1 = \mathrm{Id}, \quad R_2\eta_2 = \mathrm{Id}, \quad R_1\eta_2 = 0, \quad \Im(\mathrm{Id} - E_1R_1) \subset \Im\eta_2.$$

Then

$$E = [E_1 \quad \eta_2] \in B(V_1 \times V_2, V) \text{ is boundedly invertible,}$$

with inverse

$$E^{-1} = \begin{bmatrix} R_1 \\ R_2(\mathrm{Id} - E_1R_1) \end{bmatrix}.$$

Proof. Using that $R_1E_1 = \mathrm{Id}$, $R_1\eta_2 = 0$, $R_2\eta_2 = \mathrm{Id}$, we have

$$\begin{bmatrix} R_1 \\ R_2(\mathrm{Id} - E_1R_1) \end{bmatrix} [E_1 \quad \eta_2] = \begin{bmatrix} \mathrm{Id} & 0 \\ 0 & \mathrm{Id} \end{bmatrix},$$

and using that $\Im(\mathrm{Id} - E_1R_1) \subset \Im\eta_2$ and $R_2\eta_2 = \mathrm{Id}$, we have

$$[E_1 \quad \eta_2] \begin{bmatrix} R_1 \\ R_2(\mathrm{Id} - E_1R_1) \end{bmatrix} = E_1R_1 + \eta_2R_2(\mathrm{Id} - E_1R_1) = \mathrm{Id}. \quad \square$$

In applications V (V_i) will be densely embedded in a Hilbert space H (H_i). Questions about boundedness of E or E^{-1} in dual spaces then reduce to properties of the Hilbert adjoint of E. Study of the Hilbert adjoint will also be relevant for the investigation of dual bases.

Proposition 5.2.2. *For Hilbert spaces H and H_i ($i = 1, 2$), let $R_i \in B(H, H_i)$, and isometries $\eta_i \in B(H_i, H)$ be such that*

$$R_i\eta_j = \delta_{ij} \quad (i, j \in \{1, 2\}), \quad H = \Im\eta_1 \oplus^{\perp} \Im\eta_2,$$

and let $E_1 \in B(H_1, H)$ be such that $R_1E_1 = \mathrm{Id}$.

Then $\eta_1R_1 + \eta_2R_2 = \mathrm{Id}$, $E \in B(H_1 \times H_2, H)$ is boundedly invertible, $\eta_i^* = R_i$, and

$$E^* = \begin{bmatrix} E_1^* \\ R_2 \end{bmatrix}, \quad E^{-*} = [\eta_1 \quad (\mathrm{Id} - \eta_1E_1^*)\eta_2].$$

Proof. The first statement statement follows from $\eta_1R_1 + \eta_2R_2 = \mathrm{Id}$ on $\Im\eta_i$. The second statement follows from Proposition 5.2.1 once we have verified that $\Im(\mathrm{Id} - E_1R_1) \subset \Im\eta_2$. Writing $(\mathrm{Id} - E_1R_1)x = \eta_1x_1 + \eta_2x_2$, and applying R_1 to both sides, we find $x_1 = 0$ as required. For any $u \in H_i$, $v \in H$,

$$\langle \eta_iu, v\rangle_H = \langle \eta_iu, \sum_j \eta_jR_jv\rangle_H = \langle \eta_iu, \eta_iR_iv\rangle_H = \langle u, R_iv\rangle_{H_i},$$

or $\eta_i^* = R_i$. Now the last statement follows from the formulas for E and E^{-1} given in Proposition 5.2.1. $\square$

Remark 5.2.3. The formulas for E and E^{-*}, and so those for E^{-1} and E^* are symmetric, with reversed roles of H_1 and H_2, in the sense that with $E_2 := (\mathrm{Id} - \eta_1E_1^*)\eta_2$, it holds that $(\mathrm{Id} - \eta_2E_2^*)\eta_1 = E_1$.

Let $\tilde{V}$ and $\tilde{V}_i$ ($i = 1, 2$) be reflexive Banach spaces with

$$\tilde{V} \hookrightarrow H, \quad \tilde{V}_i \hookrightarrow H_i \text{ with dense embeddings.}$$

In this setting, we have that boundedness, or bounded invertibility of

$$E : \tilde{V}_1' \times \tilde{V}_2' \to \tilde{V}'$$

is equivalent to boundedness, or to bounded invertibility of

$$E^* : \tilde{V} \to \tilde{V}_1 \times \tilde{V}_2.$$

Proposition 5.2.4. *Let the assumptions of Proposition 5.2.2 be valid. Let*

$$R_2 \in B(\tilde{V}, \tilde{V}_2), \quad \eta_1 \in B(\tilde{V}_1, \tilde{V}), \quad E_1^* \in B(\tilde{V}, \tilde{V}_1).$$

Then $E^* \in B(\tilde{V}, \tilde{V}_1 \times \tilde{V}_2)$, and so $E \in B(\tilde{V}_1' \times \tilde{V}_2', \tilde{V}')$, is boundedly invertible if and only if R_2 has a right-inverse $\hat{E}_2 \in B(\tilde{V}_2, \tilde{V})$.

Proof. The assumptions imply that $E^* \in B(\tilde{V}, \tilde{V}_1 \times \tilde{V}_2)$, and that for $E^{-*} \in B(\tilde{V}_1 \times \tilde{V}_2, \tilde{V})$ it suffices to show that $E_2 := (\mathrm{Id} - \eta_1 E_1^*)\eta_2 \in B(\tilde{V}_2, \tilde{V})$. If the latter is true, then, since $R_2 E_2 = \mathrm{Id}$, we can take $\hat{E}_2 = E_2$.

Conversely, let $\hat{E}_2 \in B(\tilde{V}_2, \tilde{V})$ be a right-inverse of R_2. We have that

$$\begin{aligned} R_1(\mathrm{Id} - E_2 R_2) &= R_1 - R_1\eta_2 R_2 + R_1\eta_1 E_1^*\eta_2 R_2 = R_1 + E_1^*\eta_2 R_2 \\ &= E_1^*(\eta_1 R_1 + \eta_2 R_2) = E_1^* \in B(\tilde{V}, \tilde{V}_1). \end{aligned}$$

So

$$\mathrm{Id} - E_2 R_2 = (\eta_1 R_1 + \eta_2 R_2)(\mathrm{Id} - E_2 R_2) = \eta_1 R_1(\mathrm{Id} - E_2 R_2) \in B(\tilde{V}, \tilde{V}),$$

or $E_2 R_2 \in B(\tilde{V}, \tilde{V})$. But then $E_2 = E_2 R_2 \hat{E}_2 \in B(\tilde{V}_2, \tilde{V})$. □

Finally in this section, we apply arguments from interpolation space theory to conclude boundedness of E in scales of Banach spaces.

Proposition 5.2.5. *(a). Let V, $\underline{V}$, and V_i, $\underline{V}_i$ $(i = 1, 2)$ be Banach spaces with*

$$\underline{V} \hookrightarrow V, \quad \underline{V}_i \hookrightarrow V_i \quad \textit{with dense embeddings.}$$

Let the mappings (R_1, R_2, E_1, η_2) satisfy the conditions from Proposition 5.2.1 for both *triples (V, V_1, V_2) and $(\underline{V}, \underline{V}_1, \underline{V}_2)$. Then for $s \in [0, 1]$, $q \in [1, \infty]$,*

$$E \in B([V_1, \underline{V}_1]_{s,q} \times [V_2, \underline{V}_2]_{s,q}, [V, \underline{V}]_{s,q}) \textit{ is boundedly invertible.}$$

(b). Let $\tilde{V}$, $\underline{\tilde{V}}$, and $\tilde{V}_i$, $\underline{\tilde{V}}_i$ be reflexive Banach spaces, and H and H_i be Hilbert spaces $(i = 1, 2)$ with

$$\underline{\tilde{V}} \hookrightarrow \tilde{V} \hookrightarrow H, \quad \underline{\tilde{V}}_i \hookrightarrow \tilde{V}_i \hookrightarrow H_i \quad \textit{with dense embeddings.}$$

Let the conditions of Proposition 5.2.2 be satisfied, as well as the conditions of Proposition 5.2.4 for both *triples $(\tilde{V}, \tilde{V}_1, \tilde{V}_2)$ and $(\underline{\tilde{V}}, \underline{\tilde{V}}_1, \underline{\tilde{V}}_2)$. Then for $s \in [0, 1]$, $q \in [1, \infty]$,*

$$E \in B([\tilde{V}_1, \underline{\tilde{V}}_1]'_{s,q} \times [\tilde{V}_2, \underline{\tilde{V}}_2]'_{s,q}, [\tilde{V}, \underline{\tilde{V}}]'_{s,q}) \textit{ is boundedly invertible.}$$

5.3 Approximation by tensor product wavelets on the hypercube

We will study non-overlapping domain decompositions, where the subdomains are either unit n-cubes or smooth images of those. Sobolev spaces on these cubes, that appear with the construction of a Riesz basis for a Sobolev space on the domain as a whole, will be equipped with tensor product wavelet bases. From [47], we recall the construction of those bases, as well as results on the rate of approximation from spans of suitably chosen subsets of these bases.

For $t \in [0,\infty) \setminus (\mathbb{N}_0 + \{\frac{1}{2}\})$ and $\vec{\sigma} = (\sigma_\ell, \sigma_r) \in \{0,\dots,\lfloor t+\frac{1}{2}\rfloor\}^2$, with $\mathcal{I} := (0,1)$, let

$$H^t_{\vec{\sigma}}(\mathcal{I}) := \{v \in H^t(\mathcal{I}) : v(0) = \cdots = v^{(\sigma_\ell - 1)}(0) = 0 = v(1) = \cdots = v^{(\sigma_r-1)}(1)\}.$$

Remark 5.3.1. Later, we will use this definition also with $\mathcal{I}$ reading as a general non-empty interval, with 0 and 1 reading as its left and right boundary.

For t and $\vec{\sigma}$ as above, and for $\tilde{t} \in [0,\infty) \setminus (\mathbb{N}_0 + \{\frac{1}{2}\})$ and $\vec{\tilde{\sigma}} = (\tilde{\sigma}_\ell, \tilde{\sigma}_r) \in \{0,\dots,\lfloor \tilde{t} + \frac{1}{2}\rfloor\}^2$, we assume *univariate wavelet* collections

$$\Psi_{\vec{\sigma},\vec{\tilde{\sigma}}} := \left\{\psi_\lambda^{(\vec{\sigma},\vec{\tilde{\sigma}})} : \lambda \in \nabla_{\vec{\sigma},\vec{\tilde{\sigma}}}\right\} \subset H^t_{\vec{\sigma}}(\mathcal{I})$$

such that

$\mathcal{W}_1$. $\Psi_{\vec{\sigma},\vec{\tilde{\sigma}}}$ is a Riesz basis for $L_2(\mathcal{I})$,

$\mathcal{W}_2$. $\{2^{-|\lambda|t}\psi_\lambda^{(\vec{\sigma},\vec{\tilde{\sigma}})} : \lambda \in \nabla_{\vec{\sigma},\vec{\tilde{\sigma}}}\}$ is a Riesz basis for $H^t_{\vec{\sigma}}(\mathcal{I})$,

where $|\lambda| \in \mathbb{N}_0$ denotes the *level* of λ. Denoting the dual basis of $\Psi_{\vec{\sigma},\vec{\tilde{\sigma}}}$ for $L_2(\mathcal{I})$ as $\tilde{\Psi}_{\vec{\sigma},\vec{\tilde{\sigma}}} := \{\tilde{\psi}_\lambda^{(\vec{\sigma},\vec{\tilde{\sigma}})} : \lambda \in \nabla_{\vec{\sigma},\vec{\tilde{\sigma}}}\}$, furthermore we assume that

$\mathcal{W}_3$. $\{2^{-|\lambda|\tilde{t}}\tilde{\psi}_\lambda^{(\vec{\sigma},\vec{\tilde{\sigma}})} : \lambda \in \nabla_{\vec{\sigma},\vec{\tilde{\sigma}}}\}$ is a Riesz basis for $H^{\tilde{t}}_{\vec{\tilde{\sigma}}}(\mathcal{I})$,

and that for some

$$\mathbb{N} \ni d > t,$$

$\mathcal{W}_4$. $|\langle \tilde{\psi}_\lambda^{(\vec{\sigma},\vec{\tilde{\sigma}})}, u\rangle_{L_2(\mathcal{I})}| \lesssim 2^{-|\lambda|d}\|u\|_{H^d(\operatorname{supp}\tilde{\psi}^{(\vec{\sigma},\vec{\tilde{\sigma}})})}$ $(u \in H^d(\mathcal{I}) \cap H^t_{\vec{\sigma}}(\mathcal{I}))$,

$\mathcal{W}_5$. $\varrho := \sup_{\lambda\in\nabla_{\vec{\sigma},\vec{\tilde{\sigma}}}} 2^{|\lambda|}\max(\operatorname{diam}\operatorname{supp}\tilde{\psi}_\lambda^{(\vec{\sigma},\vec{\tilde{\sigma}})}, \operatorname{diam}\operatorname{supp}\psi_\lambda^{(\vec{\sigma},\vec{\tilde{\sigma}})})$
$\eqsim \inf_{\lambda\in\nabla_{\vec{\sigma},\vec{\tilde{\sigma}}}} 2^{|\lambda|}\max(\operatorname{diam}\operatorname{supp}\tilde{\psi}_\lambda^{(\vec{\sigma},\vec{\tilde{\sigma}})}, \operatorname{diam}\operatorname{supp}\psi_\lambda^{(\vec{\sigma},\vec{\tilde{\sigma}})})$,

$\mathcal{W}_6$. $\sup_{j,k\in\mathbb{N}_0} \#\{|\lambda| = j : [k2^{-j},(k+1)2^{-j}] \cap (\operatorname{supp}\tilde{\psi}_\lambda^{(\vec{\sigma},\vec{\tilde{\sigma}})} \cup \operatorname{supp}\psi_\lambda^{(\vec{\sigma},\vec{\tilde{\sigma}})}) \neq \emptyset\} < \infty$.

The conditions $(\mathcal{W}_5)$ and $(\mathcal{W}_6)$ will be referred to by saying that *both primal and dual wavelets* are *local* or *locally finite*, respectively. For some arguments, it will be used that by increasing the coarsest scale, the constant ϱ can always be assumed to be sufficiently small.

With, for $n \in \mathbb{N}$,

$$\square := \mathcal{I}^n,$$

one has $L_2(\square) = \otimes_{i=1}^n L_2(\mathcal{I})$. For

$$\boldsymbol{\sigma} = (\vec{\sigma}_i = ((\sigma_i)_\ell, (\sigma_i)_r))_{1\le i\le n} \in (\{0, \dots, \lfloor t + \tfrac{1}{2}\rfloor\}^2)^n,$$

we define

$$H^t_{\boldsymbol{\sigma}}(\square) := H^t_{\vec{\sigma}_1}(\mathcal{I}) \otimes L_2(\mathcal{I}) \otimes \cdots \otimes L_2(\mathcal{I}) \cap \cdots \cap L_2(\mathcal{I}) \otimes \cdots \otimes L_2(\mathcal{I}) \otimes H^t_{\vec{\sigma}_n}(\mathcal{I}),$$

which is the space of $H^t(\square)$-functions whose normal derivatives of up to orders $(\sigma_i)_\ell$ and $(\sigma_i)_r$ vanish at the facets $\overline{\mathcal{I}}^{i-1} \times \{0\} \times \overline{\mathcal{I}}^{n-i}$ and $\overline{\mathcal{I}}^{i-1} \times \{1\} \times \overline{\mathcal{I}}^{n-i}$, respectively $(1 \le i \le n)$ (the proof of this fact given in [47] for $t \in \mathbb{N}_0$ can be generalized to $t \in [0,\infty) \setminus (\mathbb{N}_0 + \frac{1}{2})$).

The *tensor product wavelet* collection

$$\boldsymbol{\Psi}_{\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}}} := \otimes_{i=1}^n \Psi_{\vec{\sigma}_i,\vec{\tilde{\sigma}}_i} = \big\{\boldsymbol{\psi}^{(\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}})}_{\boldsymbol{\lambda}} := \otimes_{i=1}^n \psi^{(\vec{\sigma}_i,\vec{\tilde{\sigma}}_i)}_{\lambda_i} : \boldsymbol{\lambda} \in \boldsymbol{\nabla}_{\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}}} := \prod_{i=1}^n \nabla_{\vec{\sigma}_i,\vec{\tilde{\sigma}}_i}\big\},$$

and its renormalized version $\big\{\big(\sum_{i=1}^n 4^{t|\lambda_i|}\big)^{-1/2} \boldsymbol{\psi}^{(\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}})}_{\boldsymbol{\lambda}} : \boldsymbol{\lambda} \in \boldsymbol{\nabla}_{\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}}}\big\}$ are Riesz bases for $L_2(\square)$ and $H^t_{\boldsymbol{\sigma}}(\square)$, respectively. The collection that is dual to $\boldsymbol{\Psi}_{\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}}}$ reads as

$$\tilde{\boldsymbol{\Psi}}_{\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}}} := \otimes_{i=1}^n \tilde{\Psi}_{\vec{\sigma}_i,\vec{\tilde{\sigma}}_i} = \big\{\tilde{\boldsymbol{\psi}}^{(\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}})}_{\boldsymbol{\lambda}} := \otimes_{i=1}^n \tilde{\psi}^{(\vec{\sigma}_i,\vec{\tilde{\sigma}}_i)}_{\lambda_i} : \boldsymbol{\lambda} \in \boldsymbol{\nabla}_{\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}}}\big\},$$

and its renormalized version $\big\{\big(\sum_{i=1}^n 4^{|\lambda_i|}\big)^{-\tilde{t}/2} \tilde{\boldsymbol{\psi}}^{(\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}})}_{\boldsymbol{\lambda}} : \boldsymbol{\lambda} \in \boldsymbol{\nabla}_{\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}}}\big\}$ is a Riesz basis for $H^{\tilde{t}}_{\tilde{\boldsymbol{\sigma}}}(\square)$.

For $\boldsymbol{\lambda} \in \boldsymbol{\nabla}_{\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}}}$, we set $|\boldsymbol{\lambda}| := (|\lambda_1|, \dots, |\lambda_n|)$. As usual, for $\boldsymbol{j}, \boldsymbol{\jmath} \in \mathbb{N}_0^n$, $|\boldsymbol{j}| \le |\boldsymbol{\jmath}|$ will mean that $|\boldsymbol{j}|_i \le |\boldsymbol{\jmath}|_i$ $(1 \le i \le n)$, whereas $|\boldsymbol{j}| \ge |\boldsymbol{\jmath}|$ or $|\boldsymbol{j}| = |\boldsymbol{\jmath}|$ will mean that $|\boldsymbol{\jmath}| \le |\boldsymbol{j}|$ or $|\boldsymbol{j}| \le |\boldsymbol{\jmath}|$ and $|\boldsymbol{j}| \ge |\boldsymbol{\jmath}|$, respectively.

For $\theta \ge 0$, the *weighted Sobolev space* $\mathcal{H}^d_\theta(\mathcal{I})$ is defined as the space of all measurable functions u on $\mathcal{I}$ for which the norm

$$\|u\|_{\mathcal{H}^d_\theta(\mathcal{I})} := \left[\sum_{j=0}^d \int_{\mathcal{I}} |x^\theta (1-x)^\theta u^{(j)}(x)|^2 \,\mathrm{d}x\right]^{\frac{1}{2}}$$

is finite. For

$$m \in \{0, \dots, \lfloor t \rfloor\},$$

we will consider the *weighted Sobolev space*

$$\mathcal{H}^d_{m,\theta}(\square) := \cap_{p=1}^n \otimes_{i=1}^n \mathcal{H}^d_{\theta-\delta_{ip}\min(m,\theta)}(\mathcal{I}),$$

equipped with a squared norm that is the sum over $p = 1, \dots, n$ of the squared norms on $\otimes_{i=1}^n \mathcal{H}^d_{\theta-\delta_{ip}\min(m,\theta)}(\mathcal{I})$.

Theorem 5.3.2 ([47, Thm. 4.3])**.** *For any $\theta \in [0, d)$, there exist a (nested) sequence $(\boldsymbol{\nabla}_M^{(\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}})})_{M\in\mathbb{N}} \subset \boldsymbol{\nabla}_{\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}}}$ with $\#\boldsymbol{\nabla}_M^{(\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}})} \eqsim M$, such that*

$$\inf_{v\in\operatorname{span}\{\boldsymbol{\psi}_{\boldsymbol{\lambda}}^{(\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}})}:\boldsymbol{\lambda}\in\boldsymbol{\nabla}_M^{(\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}})}\}} \|u - v\|_{H^m(\square)} \lesssim M^{-(d-m)}\|u\|_{\mathcal{H}^d_{m,\theta}(\square)}, \quad (u \in \mathcal{H}^d_{m,\theta}(\square) \cap H^m_{\boldsymbol{\sigma}}(\square)),$$

where for $m = 0$, $M^{-(d-m)}$ should be read as $(\log \#M)^{(n-1)(\frac{1}{2}+d)}M^{-d}$.

The index sets $\boldsymbol{\nabla}_M^{(\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}})}$ can be chosen to have the following multiple tree property*: For any $\boldsymbol{\lambda} \in \boldsymbol{\nabla}_M^{(\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}})}$ and any $\boldsymbol{j} \in \mathbb{N}_0^n$ with $\boldsymbol{j} \le |\boldsymbol{\lambda}|$, there exists a $\boldsymbol{\mu} \in \boldsymbol{\nabla}_M^{(\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}})}$ with $|\boldsymbol{\mu}| = \boldsymbol{j}$, and $\operatorname{supp}\boldsymbol{\psi}_{\boldsymbol{\lambda}}^{(\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}})} \cap \operatorname{supp}\boldsymbol{\psi}_{\boldsymbol{\mu}}^{(\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}})} \neq \emptyset$.*

With the notations $u \in H^t_{\boldsymbol{\sigma}}(\alpha + \square)$ and $u \in \mathcal{H}^d_{m,\theta}(\alpha + \square)$, we will mean that $u(\cdot + \alpha) \in H^t_{\boldsymbol{\sigma}}(\square)$ or $u(\cdot + \alpha) \in \mathcal{H}^d_{m,\theta}(\square)$, respectively.

5.4 Construction of Riesz bases by extension

Let $\{\square_0, \dots, \square_N\}$ be a set of hypercubes from $\{\tau + \square : \tau \in \mathbb{Z}^n\}$, and let $\hat{\Omega}$ be a (reference) domain (i.e., open and connected) in $\mathbb{R}^n$ with $\cup_{k=0}^N \square_k \subset \hat{\Omega} \subset (\cup_{k=0}^N \overline{\square}_k)^{\text{int}}$, and such that $\partial\hat{\Omega}$ is the union of (closed) facets of the $\square_k$'s. The case $\hat{\Omega} \subsetneq (\cup_{k=0}^N \overline{\square}_k)^{\text{int}}$ corresponds to the situation that $\hat{\Omega}$ has one or more cracks. We will describe a construction of Riesz bases for Sobolev spaces on $\hat{\Omega}$ from Riesz bases for corresponding Sobolev spaces on the subdomains $\square_k$ using extension operators. We start with giving sufficient conditions $(\mathcal{D}_1)$–$(\mathcal{D}_5)$ such that suitable extension operators exist. At the end of this section, we will consider domains given as the parametric image of $\hat{\Omega}$.

We assume that there exists a sequence $(\{\hat{\Omega}_k^{(q)} : q \le k \le N\})_{0\le q\le N}$ of sets of polytopes, such that $\hat{\Omega}_k^{(0)} = \square_k$ and where each next term in the sequence is created from its predecessor by joining two of its polytopes. More precisely, we assume that for any $1 \le q \le N$, there exists a $q \le \bar{k} = \bar{k}^{(q)} \le N$ and $q-1 \le k_1 = k_1^{(q)} \neq k_2 = k_2^{(q)} \le N$ such that

$\mathcal{D}_1$. $\hat{\Omega}_{\bar{k}}^{(q)} = \left(\overline{\hat{\Omega}_{k_1}^{(q-1)} \cup \hat{\Omega}_{k_2}^{(q-1)}} \setminus \partial\hat{\Omega}\right)^{\text{int}}$ is connected, and the interface $J := \hat{\Omega}_{\bar{k}}^{(q)} \setminus (\hat{\Omega}_{k_1}^{(q-1)} \cup \hat{\Omega}_{k_2}^{(q-1)})$ is part of a hyperplane,

$\mathcal{D}_2$. $\{\hat{\Omega}_k^{(q)} : q \le k \le N, k \neq \bar{k}\} = \{\hat{\Omega}_k^{(q-1)} : q-1 \le k \le N, k \neq \{k_1, k_2\}\}$,

$\mathcal{D}_3$. $\hat{\Omega}_N^{(N)} = \hat{\Omega}$.

For some

$$t \in [0,\infty) \setminus (\mathbb{N}_0 + \{\tfrac{1}{2}\}),$$

to each of the *closed* facets of all the hypercubes $\square_k$, we associate a number in $\{0,\dots,\lfloor t+\frac{1}{2}\rfloor\}$ indicating the order of the Dirichlet boundary condition on that facet (where a Dirichlet boundary condition of order 0 means no boundary condition). On facets on the boundary of $\hat{\Omega}$, this number can be chosen at one's convenience (it is dictated by the boundary conditions of the boundary value problem that one wants to solve on $\hat{\Omega}$), and, as will follow from the conditions imposed below, on the other facets it should be either 0 or $\lfloor t+\frac{1}{2}\rfloor$.

By construction, each facet of any $\hat{\Omega}_k^{(q)}$ is a union of some facets of the $\square_{k'}$'s, that will be referred to as subfacets. Letting each of these subfacets inherit the Dirichlet boundary conditions imposed on the $\square_{k'}$'s, we define

$$\mathring{H}^t(\hat{\Omega}_k^{(q)}),$$

and so for $k=q=N$ in particular $\mathring{H}^t(\hat{\Omega}) = \mathring{H}^t(\hat{\Omega}_N^{(N)})$, to be the closure in $H^t(\hat{\Omega}_k^{(q)})$ of the smooth functions on $\hat{\Omega}_k^{(q)}$ that satisfy these boundary conditions. Note that for $0 \le k \le N$, for some $\boldsymbol{\sigma}(k) \in (\{0,\dots,\lfloor t+\frac{1}{2}\rfloor\}^2)^n$,

$$\mathring{H}^t(\hat{\Omega}_k^{(0)}) = \mathring{H}^t(\square_k) = H^t_{\boldsymbol{\sigma}(k)}(\square_k).$$

Remark 5.4.1. On the intersection of facets of hypercubes $\square_{k'}$, the natural interpretation of the boundary conditions is the minimal one such that the boundary conditions on each of these facets is not violated.

The boundary conditions on the hypercubes that determine the spaces $\mathring{H}^t(\hat{\Omega}_k^{(q)})$, and the order in which polytopes are joined should be chosen such that

$\mathcal{D}_4$. on the $\hat{\Omega}_{k_1}^{(q-1)}$ and $\hat{\Omega}_{k_2}^{(q-1)}$ sides of J, the boundary conditions are of order 0 and $\lfloor t+\frac{1}{2}\rfloor$, respectively,

and, w.l.o.g. assuming that $J = \{0\}\times\breve{J}$ and $(0,1)\times\breve{J} \subset \Omega_{k_1}^{(q-1)}$,

$\mathcal{D}_5$. for any function in $\mathring{H}^t(\hat{\Omega}_{k_1}^{(q-1)})$ that vanishes near $\{0,1\}\times\breve{J}$, its reflection in $\{0\}\times\mathbb{R}^{n-1}$ (extended with zero, and then restricted to $\hat{\Omega}_{k_2}^{(q-1)}$) is in $\mathring{H}^t(\hat{\Omega}_{k_2}^{(q-1)})$.

The condition ($\mathcal{D}_5$) can be formulated by saying that the order of the boundary condition at any subfacet of $\hat{\Omega}_{k_1}^{(q-1)}$ adjacent to J should not be less than this order at its reflection in J, where in case this reflection is not part of $\partial\hat{\Omega}_{k_2}^{(q-1)}$ the latter order should be read as the highest possible one $\lfloor t+\frac{1}{2}\rfloor$; and furthermore, that the order of the boundary condition at any subfacet of $\hat{\Omega}_{k_2}^{(q-1)}$ adjacent to J should not be larger than this order at its reflection in J, where in case this reflection is not part of $\partial\hat{\Omega}_{k_1}^{(q-1)}$, the latter order should be read as the lowest possible one 0. See Figure 5.3 for an illustration.

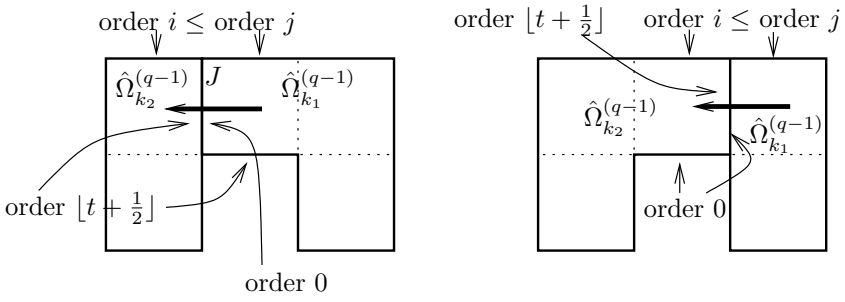


Figure 5.3: Two illustrations with $(\mathcal{D}_1)$–$(\mathcal{D}_5)$. The fat arrow indicates the action of the extension $E_1^{(q)}$.

Given $1 \le q \le N$, for $i \in \{1,2\}$, let $R_i^{(q)}$ be the *restriction* of functions on $\hat\Omega_{\bar k}^{(q)}$ to $\hat\Omega_{k_i}^{(q-1)}$, let $\eta_2^{(q)}$ be the *extension* of functions on $\hat\Omega_{k_2}^{(q-1)}$ to $\hat\Omega_{\bar k}^{(q)}$ by zero, and let $E_1^{(q)}$ be some *extension* of functions on $\hat\Omega_{k_1}^{(q-1)}$ to $\hat\Omega_{\bar k}^{(q)}$.

Proposition 5.4.2. *Assume that*

$$E_1^{(q)} \in B(L_2(\hat\Omega_{k_1}^{(q-1)}), L_2(\hat\Omega_{\bar k}^{(q)})), \quad E_1^{(q)} \in B(\mathring{H}^t(\hat\Omega_{k_1}^{(q-1)}), \mathring{H}^t(\hat\Omega_{\bar k}^{(q)})).$$

Then for $s \in [0,1]$

$$E^{(q)} := [E_1^{(q)} \quad \eta_2^{(q)}] \in B\Big(\prod_{i=1}^{2}[L_2(\hat\Omega_{k_i}^{(q-1)}), \mathring{H}^t(\hat\Omega_{k_i}^{(q-1)})]_{s,2}, [L_2(\hat\Omega_{\bar k}^{(q)}), \mathring{H}^t(\hat\Omega_{\bar k}^{(q)})]_{s,2}\Big)$$

is boundedly invertible.

Proof. Taking $V^{(q)} = L_2(\hat\Omega_{\bar k}^{(q)})$, $V_i^{(q)} = L_2(\hat\Omega_{k_i}^{(q-1)})$, $\underline{V}^{(q)} = \mathring{H}^t(\hat\Omega_{\bar k}^{(q)})$, $\underline{V}_i^{(q)} = \mathring{H}^t(\hat\Omega_{k_i}^{(q-1)})$, and noting that $\Im(\mathrm{Id}-E_1^{(q)}R_1^{(q)}) \subset \{u \in \mathring{H}^t(\hat\Omega_{\bar k}^{(q)}) : u = 0 \text{ on } \hat\Omega_{k_1}^{(q-1)}\} = \Im(\eta_2^{(q)}|_{\mathring{H}^t(\hat\Omega_{k_2}^{(q-1)})})$, the result follows from an application of Proposition 5.2.5(a). □

Corollary 5.4.3. *For E being the composition for $q = 1,\dots,N$ of the mappings $E^{(q)}$ from Proposition 5.4.2, trivially extended with identity operators in coordinates $k \in \{q-1,\dots,N\} \setminus \{k_1^{(q)}, k_2^{(q)}\}$, it holds that*

$$E \in B\Big(\prod_{k=0}^{n}[L_2(\square_k), \mathring{H}^t(\square_k)]_{s,2}, [L_2(\hat\Omega), \mathring{H}^t(\hat\Omega)]_{s,2}\Big). \tag{5.4.1}$$

is boundedly invertible.

Under the conditions $(\mathcal{D}_1)$–$(\mathcal{D}_5)$, the extensions $E_1^{(q)}$ can be constructed (essentially) as tensor products of *univariate extensions* with identity operators in the other Cartesian directions.

Proposition 5.4.4. *W.l.o.g. let* $J = \{0\} \times \breve{J}$ *and* $(0,1) \times \breve{J} \subset \hat{\Omega}_{k_1}^{(q-1)}$. *Let* G_1 *be an* extension *operator of functions on* $(0,1)$ *to functions on* $(-1,1)$ *such that*

$$G_1 \in B(L_2(0,1), L_2(-1,1)), \quad G_1 \in B(H^t(0,1), H^t_{(\lfloor t+\frac{1}{2} \rfloor, 0)}(-1,1)).$$

Then $E_1^{(q)}$ *defined by* $R_2^{(q)} E_1^{(q)}$ *being the composition of the restriction to* $(0,1) \times \breve{J}$, *followed by an application of*

$$G_1 \otimes \mathrm{Id} \otimes \cdots \otimes \mathrm{Id},$$

followed by an extension by 0 *to* $\hat{\Omega}_{k_2}^{(q-1)} \setminus (-1,0) \times \breve{J}$, *satisfies the assumptions made in Proposition 5.4.2.*

Remark 5.4.5. The condition that an extension by G_1 vanishes up to order $\lfloor t + \frac{1}{2} \rfloor$ at -1 is fully harmless since it can easily be enforced by multiplying an extension by some smooth cut-off function. The scale-dependent extension that we will discuss in Subsection 5.5.1 satisfies this boundary condition automatically.

Our main interest of Corollary 5.4.3 lies in the following:

Corollary 5.4.6. *For* $0 \leq k \leq N$, *let* $\mathbf{\Psi}_k$ *be a Riesz basis for* $L_2(\square_k)$, *that renormalized in* $H^t(\square_k)$, *is a Riesz basis for* $\overset{\circ}{H}{}^t(\square_k) = H^t_{\boldsymbol{\sigma}(k)}(\square)$. *Then for* $s \in [0,1]$, *and with* E *from Corollary 5.4.3, the collection* $E(\prod_{k=0}^N \mathbf{\Psi}_k)$, *normalized in the corresponding norm, is a Riesz basis for* $[L_2(\hat{\Omega}), \overset{\circ}{H}{}^t(\hat{\Omega})]_{s,2}$.

Remark 5.4.7. Although we allow for $t \in (0, \frac{1}{2})$, for these values of t our exposition is not very relevant. Indeed, for those t, a piecewise tensor product basis can simply be constructed as the union of the tensor product bases on the hypercubes.

To find the corresponding *dual* basis, we follow Section 5.2. Taking for $q = 1, \ldots, N$,

$$H^{(q)} = L_2(\hat{\Omega}_{\bar{k}}^{(q)}), \quad H_i^{(q)} = L_2(\hat{\Omega}_{k_i}^{(q)}),$$

and with $\eta_1^{(q)}$ being the extension of functions on $\hat{\Omega}_{k_1}^{(q-1)}$ to $\hat{\Omega}_{\bar{k}}^{(q)}$ by zero, Proposition 5.2.2 shows that

$$(E^{(q)})^{-*} = [\eta_1^{(q)} \quad (\mathrm{Id} - \eta_1^{(q)} (E_1^{(q)})^*) \eta_2^{(q)}].$$

Corollary 5.4.8. *In the situation of Corollary 5.4.6, let* $\tilde{\mathbf{\Psi}}_k$ *the Riesz basis for* $L_2(\square_k)$ *that is dual to* $\mathbf{\Psi}_k$. *Then* $E^{-*}(\prod_{k=0}^N \tilde{\mathbf{\Psi}}_k)$ *is the Riesz basis for* $L_2(\hat{\Omega})$ *that is dual to* $E(\prod_{k=0}^N \mathbf{\Psi}_k)$. *The operator* E^{-*} *is the composition for* $q = 1, \ldots, N$ *of the mappings* $(E^{(q)})^{-*}$ *trivially extended with identity operators in coordinates* $k \in \{q-1, \ldots, N\} \setminus \{k_1^{(q)}, k_2^{(q)}\}$.

Below we give conditions such that $E^{-*}(\prod_{k=0}^{N}\tilde{\mathbf{\Psi}}_k)$, properly scaled, is a Riesz basis for a range of Sobolev spaces with positive smoothness indices, and so, equivalently, $E(\prod_{k=0}^{N}\mathbf{\Psi}_k)$ to be a Riesz basis for the corresponding dual spaces.

For some $\tilde{t}\in[0,\infty)\setminus(\mathbb{N}_0+\{\frac{1}{2}\})$, to each of the closed facets of all the hypercubes $\square_k$, we associate a number in $\{0,\dots,\lfloor\tilde{t}+\frac{1}{2}\rfloor\}$ indicating the order of the dual Dirichlet boundary condition on that facet. On facets on the boundary of $\hat{\Omega}$, this number can be chosen arbitrarily, where on the interior facets it is 0 or $\lfloor\tilde{t}+\frac{1}{2}\rfloor$.

We define $\overset{\circ}{H}{}^{\tilde{t}}(\hat{\Omega}_k^{(q)})$, and so for $k=q=N$ in particular $\overset{\circ}{H}{}^{\tilde{t}}(\hat{\Omega})=\overset{\circ}{H}{}^{\tilde{t}}(\hat{\Omega}_N^{(N)})$, to be the closure in $H^{\tilde{t}}(\hat{\Omega}_k^{(q)})$ of the smooth functions on $\hat{\Omega}_k^{(q)}$ that on any of its facets satisfy the boundary conditions that were imposed on each of its subfacets. Note that with some abuse of notation, even when $\tilde{t}=t$ generally $\overset{\circ}{H}{}^{\tilde{t}}(\hat{\Omega}_k^{(q)})\neq\overset{\circ}{H}{}^{t}(\hat{\Omega}_k^{(q)})$, and that for $0\leq k\leq N$,

$$\overset{\circ}{H}{}^{\tilde{t}}(\hat{\Omega}_k^{(0)})=\overset{\circ}{H}{}^{\tilde{t}}(\square_k)=H^{\tilde{t}}_{\tilde{\boldsymbol{\sigma}}(k)}(\square_k),$$

for some $\tilde{\boldsymbol{\sigma}}(k)\in(\{0,\dots,\lfloor\tilde{t}+\frac{1}{2}\rfloor\}^2)^n$.

We make the following assumptions on the selection of the boundary conditions that determine the dual spaces $\overset{\circ}{H}{}^{\tilde{t}}(\hat{\Omega}_k^{(q)})$:

$\mathcal{D}'_4$. on the $\hat{\Omega}_{k_1}^{(q-1)}$ and $\hat{\Omega}_{k_2}^{(q-1)}$ sides of J, the boundary conditions are of order $\lfloor\tilde{t}+\frac{1}{2}\rfloor$ and 0, respectively,

and, w.l.o.g. assuming that $J=\{0\}\times\breve{J}$ and $(0,1)\times\breve{J}\subset\Omega_{k_1}^{(q-1)}$,

$\mathcal{D}'_5$. for any function in $\overset{\circ}{H}{}^{\tilde{t}}(\hat{\Omega}_{k_2}^{(q-1)})$ that vanishes near $\{-1,0\}\times\breve{J}$, its reflection in $\{0\}\times\mathbb{R}^{n-1}$ (extended with zero, and then restricted to $\hat{\Omega}_{k_1}^{(q-1)}$) is in $\overset{\circ}{H}{}^{\tilde{t}}(\hat{\Omega}_{k_1}^{(q-1)})$.

Proposition 5.4.9. *For $1\leq q\leq N$, let the extension $E_1^{(q)}$ be of tensor product form as in Proposition 5.4.4 with $G_1^*\in B(H^{\tilde{t}}_{(0,\lfloor\tilde{t}+\frac{1}{2}\rfloor)}(-1,1),H^{\tilde{t}}_{(\lfloor\tilde{t}+\frac{1}{2}\rfloor,\lfloor\tilde{t}+\frac{1}{2}\rfloor)}(0,1))$, and let $\tilde{\mathbf{\Psi}}_k$, properly scaled, be a Riesz basis for $\overset{\circ}{H}{}^{\tilde{t}}(\square_k)$. Then for $s\in[0,1]$, $E^{-*}(\prod_{k=0}^{N}\tilde{\mathbf{\Psi}}_k)$ is, properly scaled, a Riesz basis for $[L_2(\hat{\Omega}),\overset{\circ}{H}{}^{\tilde{t}}(\hat{\Omega})]_{s,2}$.*

Remark 5.4.10. The boundary conditions imposed on G_1^*u at 1 are fully harmless. The scale-dependent extension G_1 that we will discuss in Subsection 5.5.1 satisfies these boundary conditions automatically. On the other hand, thinking of $t\geq\tilde{t}$, the boundary conditions at 0 are, when $\tilde{t}>\frac{1}{2}$, the only properties that are not already implied by the conditions on G_1 from Proposition 5.4.4.

Proof. The conditions $(\mathcal{D}'_4)$, $(\mathcal{D}'_5)$ imply both that $R_2^{(q)}$ has a right-inverse which is in $B(\overset{\circ}{H}{}^{\tilde{t}}(\Omega_{k_2}^{(q-1)}),\overset{\circ}{H}{}^{\tilde{t}}(\Omega_{\tilde{k}}^{(q)}))$ and $(E_1^{(q)})^*\in B(\overset{\circ}{H}{}^{\tilde{t}}(\Omega_{\tilde{k}}^{(q)}),\overset{\circ}{H}{}^{\tilde{t}}(\Omega_{k_1}^{(q-1)}))$, by the assumption on G_1^*. Since $R_2^{(q)}\in B(\overset{\circ}{H}{}^{\tilde{t}}(\Omega_{\tilde{k}}^{(q)}),\overset{\circ}{H}{}^{\tilde{t}}(\Omega_{k_2}^{(q-1)}))$, $\eta_1^{(q)}\in B(\overset{\circ}{H}{}^{\tilde{t}}(\Omega_{k_1}^{(q-1)}),\overset{\circ}{H}{}^{\tilde{t}}(\Omega_{\tilde{k}}^{(q)}))$ directly follow from $(\mathcal{D}'_4)$, an N-fold application of Proposition 5.2.4 together with the assumption on the bases $\tilde{\mathbf{\Psi}}_k$ completes the proof. □

To end the discussion about the stability of $E(\Pi_{k=0}^N \mathbf{\Psi}_k)$ in dual norms, we note that for $\tilde{t} < \frac{1}{2}$, which suffices for our application for solving PDEs, the conditions $(\mathcal{D}_4')$, $(\mathcal{D}_5')$, and those from Proposition 5.4.9 are void, with the exception of the very mild condition of $\tilde{\mathbf{\Psi}}_k$, properly scaled, being a Riesz basis for $\mathring{H}^{\tilde{t}}(\square_k)$.

The construction of Riesz bases on the reference domain $\hat{\Omega}$ extends to more general domains in a standard fashion. Let Ω be the image of $\hat{\Omega}$ under a homeomorphism κ. We define the *pull-back* κ^* by $\kappa^* w = w \circ \kappa$, and so its inverse κ^{-*}, known as the *push-forward*, satisfies $\kappa^{-*} v = v \circ \kappa^{-1}$.

Proposition 5.4.11. *Let κ^* be boundedly invertible as a mapping both from $L_2(\Omega)$ to $L_2(\hat{\Omega})$ and from $H^t(\Omega)$ to $H^t(\hat{\Omega})$. Setting $\mathring{H}^t(\Omega) := \Im\kappa^{-*}|_{\mathring{H}^t(\hat{\Omega})}$, we have that $\kappa^{-*} \in B([L_2(\hat{\Omega}), \mathring{H}^t(\hat{\Omega})]_{s,2}, [L_2(\Omega), \mathring{H}^t(\Omega)]_{s,2})$ is boundedly invertible ($s \in [0,1]$). So if $\mathbf{\Psi}$ is a Riesz basis for $L_2(\hat{\Omega})$ and, properly scaled, for $\mathring{H}^t(\hat{\Omega})$, then for $s \in [0,1]$, $\kappa^{-*}\mathbf{\Psi}$ is, properly scaled, a Riesz basis for $[L_2(\Omega), \mathring{H}^t(\Omega)]_{s,2}$.*

If $\tilde{\mathbf{\Psi}}$ is the collection dual to $\mathbf{\Psi}$, then $|\det D\kappa^{-1}(\cdot)|\kappa^{-*}\tilde{\mathbf{\Psi}}$ is the collection dual to $\kappa^{-*}\mathbf{\Psi}$.

5.5 Approximation by –piecewise– tensor product wavelets

In the setting of Proposition 5.4.4, Corollary 5.4.6 and Proposition 5.4.9, writing $\square_k = \square + \alpha_k$, where $\alpha_k \in \mathbb{Z}^n$, we select the the primal and dual bases for $L_2(\square_k)$ to be

$$\mathbf{\Psi}_{\boldsymbol{\sigma}(k),\tilde{\boldsymbol{\sigma}}(k)}(\cdot - \alpha_k), \quad \tilde{\mathbf{\Psi}}_{\boldsymbol{\sigma}(k),\tilde{\boldsymbol{\sigma}}(k)}(\cdot - \alpha_k)$$

as constructed in Section 5.3, which, properly scaled, are Riesz bases for $H^t_{\boldsymbol{\sigma}(k)}(\square_k)$ and $H^{\tilde{t}}_{\tilde{\boldsymbol{\sigma}}(k)}(\square_k)$, respectively.

In the setting of Proposition 5.4.11, for $m \in \{0, \dots, \lfloor t \rfloor\}$ and $u \in \mathring{H}^m(\Omega) := [L_2(\Omega), \mathring{H}^t(\Omega)]_{m/t,2}$, with additionally

$$u \in \kappa^{-*}\Big(\prod_{k=0}^N \mathcal{H}^d_{m,\theta}(\square_k)\Big) := \{v : \Omega \to \mathbb{R} : v \circ \kappa \in \prod_{k=0}^N \mathcal{H}^d_{m,\theta}(\square_k)\}, \tag{5.5.1}$$

we study approximation rates from $\kappa^{-*} E\big(\prod_{k=0}^N \mathbf{\Psi}_{\boldsymbol{\sigma}(k),\tilde{\boldsymbol{\sigma}}(k)}(\cdot - \alpha_k)\big)$ in the $H^m(\Omega)$-norm. Since, as is assumed in Proposition 5.4.11, $\kappa^* \in B(\mathring{H}^m(\Omega), \mathring{H}^m(\hat{\Omega}))$ is boundedly invertible, it is sufficient to study this issue for the case that $\kappa = \mathrm{Id}$ and so $\Omega = \hat{\Omega}$.

We will apply extension operators $E_1^{(q)}$ that are built from univariate extension operators. The latter will be chosen such that the resulting primal and dual wavelets on $\hat{\Omega}$ are, restricted to each $\square_k \subset \hat{\Omega}$, tensor products of collections of univariate functions that are local and locally finite (cf. parts (1) and (2) of the forthcoming Proposition 5.5.4).

5.5.1 Construction of scale-dependent extension operators

We make the following additional assumptions on the univariate wavelets. For $\vec{\sigma} = (\sigma_\ell, \sigma_r) \in \{0, \dots, \lfloor t + \frac{1}{2} \rfloor\}^2$, $\vec{\tilde{\sigma}} = (\tilde{\sigma}_\ell, \tilde{\sigma}_r) \in \{0, \dots, \lfloor \tilde{t} + \frac{1}{2} \rfloor\}^2$, and with $\vec{0} := (0,0)$,

$\mathcal{W}_7$. $V_j^{(\vec{\sigma})} := \operatorname{span}\{\psi_\lambda^{(\vec{\sigma},\vec{\tilde{\sigma}})} : \lambda \in \nabla_{\vec{\sigma},\vec{\tilde{\sigma}}}, |\lambda| \le j\}$ is independent of $\vec{\tilde{\sigma}}$, and $V_j^{(\vec{\sigma})} = V_j^{(\vec{0})} \cap H^t_{\vec{\sigma}}(\mathcal{I})$,

$\mathcal{W}_8$. $\nabla_{\vec{\sigma},\vec{\tilde{\sigma}}}$ is the disjoint union of $\nabla^{(\ell)}_{\sigma_\ell,\tilde{\sigma}_\ell}$, $\nabla^{(I)}$, $\nabla^{(r)}_{\sigma_r,\tilde{\sigma}_r}$ such that

i) $$\sup_{\lambda \in \nabla^{(\ell)}_{\vec{\sigma},\vec{\tilde{\sigma}}}, x \in \operatorname{supp}\psi_\lambda^{(\vec{\sigma},\vec{\tilde{\sigma}})}} 2^{|\lambda|}|x| \lesssim \varrho, \qquad \sup_{\lambda \in \nabla^{(r)}_{\vec{\sigma},\vec{\tilde{\sigma}}}, x \in \operatorname{supp}\psi_\lambda^{(\vec{\sigma},\vec{\tilde{\sigma}})}} 2^{|\lambda|}|1-x| \lesssim \varrho,$$

ii) for $\lambda \in \nabla^{(I)}$, $\psi_\lambda^{(\vec{\sigma},\vec{\tilde{\sigma}})} = \psi_\lambda^{(\vec{0},\vec{0})}$, $\tilde{\psi}_\lambda^{(\vec{\sigma},\vec{\tilde{\sigma}})} = \tilde{\psi}_\lambda^{(\vec{0},\vec{0})}$, and the extensions of $\psi_\lambda^{(\vec{0},\vec{0})}$ and $\tilde{\psi}_\lambda^{(\vec{0},\vec{0})}$ by zero are in $H^t(\mathbb{R})$ and $H^{\tilde{t}}(\mathbb{R})$, respectively.

$\mathcal{W}_9$.
$$\begin{cases} \operatorname{span}\{\psi_\lambda^{(\vec{0},\vec{0})}(1-\cdot) : \lambda \in \nabla^{(I)}, |\lambda| = j\} = \operatorname{span}\{\psi_\lambda^{(\vec{0},\vec{0})} : \lambda \in \nabla^{(I)}, |\lambda| = j\}, \\ \operatorname{span}\{\psi_\lambda^{(\sigma_\ell,\sigma_r),(\tilde{\sigma}_\ell,\tilde{\sigma}_r)}(1-\cdot) : \lambda \in \nabla^{(\ell)}_{\sigma_\ell,\tilde{\sigma}_\ell}, |\lambda| = j\} = \\ \qquad\qquad \operatorname{span}\{\psi_\lambda^{(\sigma_r,\sigma_\ell),(\tilde{\sigma}_r,\tilde{\sigma}_\ell)} : \lambda \in \nabla^{(r)}_{\sigma_r,\tilde{\sigma}_r}, |\lambda| = j\}, \end{cases}$$

$\mathcal{W}_{10}$.
$$\begin{cases} \psi_\lambda^{(\vec{\sigma},\vec{\tilde{\sigma}})}(2^l\cdot) \in \operatorname{span}\{\psi_\mu^{(\vec{\sigma},\vec{\tilde{\sigma}})} : \mu \in \nabla^{(\ell)}_{\sigma_\ell,\tilde{\sigma}_\ell}\} \quad (l \in \mathbb{N}_0,\ \lambda \in \nabla^{(\ell)}_{\sigma_\ell,\tilde{\sigma}_\ell}), \\ \psi_\lambda^{(\vec{0},\vec{0})}(2^l\cdot) \in \operatorname{span}\{\psi_\mu^{(\vec{0},\vec{0})} : \mu \in \nabla^{(I)}\} \quad (l \in \mathbb{N}_0,\ \lambda \in \nabla^{(I)}). \end{cases}$$

As $(\mathcal{W}_1)$–$(\mathcal{W}_6)$, these conditions are satisfied by following the biorthogonal wavelet constructions on the interval from [100, 50] ($(\mathcal{W}_7)$ is not satisfied by the construction from [41], but the following exposition can be adapted to apply to these wavelets as well).

Remark 5.5.1. In view of the boundary conditions that are imposed on the interfacets, see $(\mathcal{D}_4)$ and $(\mathcal{D}'_4)$, it is actually sufficient to impose $(\mathcal{W}_7)$–$(\mathcal{W}_{10})$ for $(\sigma_\ell, \tilde{\sigma}_\ell)$, $(\sigma_r, \tilde{\sigma}_r) \in \{(\lfloor t + \frac{1}{2} \rfloor, 0), (0, \lfloor \tilde{t} + \frac{1}{2} \rfloor)\}$.

We consider the setting of Proposition 5.4.4. W.l.o.g. we assume that $J = \{0\} \times \breve{J}$, and $(0,1) \times \breve{J} \subset \hat{\Omega}^{(q-1)}_{k_1}$. We assume to have available a univariate extension operator

$$\breve{G}_1 \in B(L_2(0,1), L_2(-1,1)) \text{ with } \begin{cases} \breve{G}_1 \in B(H^t(0,1), H^t(-1,1)), \\ \breve{G}_1^* \in B(H^{\tilde{t}}(-1,1), H^{\tilde{t}}_{(\lfloor \tilde{t} + \frac{1}{2} \rfloor, 0)}(0,1)). \end{cases} \tag{5.5.2}$$

Let η_1 and η_2 denote the extensions by zero of functions on $(0,1)$ or on $(-1,0)$ to functions on $(-1,1)$, with R_1 and R_2 denoting their adjoints. We assume that $\breve{G}_1$ and its "adjoint extension"

$$\breve{G}_2 := (\mathrm{Id} - \eta_1 \breve{G}_1^*)\eta_2$$

(cf. Remark 5.2.3) are local in the sense that

$$\begin{cases} \operatorname{diam}(\operatorname{supp} R_2\breve{G}_1 u) \lesssim \operatorname{diam}(\operatorname{supp} u) & (u \in L_2(0,1)), \\ \operatorname{diam}(\operatorname{supp} R_1\breve{G}_2 u) \lesssim \operatorname{diam}(\operatorname{supp} u) & (u \in L_2(-1,0)), \end{cases} \tag{5.5.3}$$

see Figure 5.4 for an illustration.

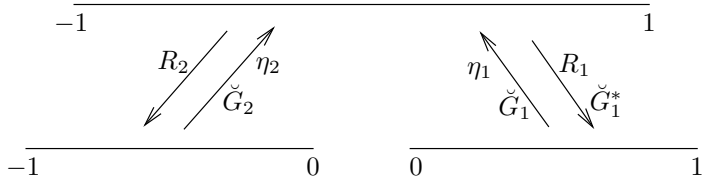


Figure 5.4: Univariate extensions and restrictions.

Examples are given by Hestenes extensions ([71, 44, 82]), which are of the form

$$\breve{G}_1 v(-x) = \sum_{l=0}^{L} \gamma_l (\zeta v)(\beta_l x) \quad (v \in L_2(\mathcal{I}),\, x \in \mathcal{I}), \tag{5.5.4}$$

(and, being an extension, $\breve{G}_1 v(x) = v(x)$ for $x \in \mathcal{I}$), where $\gamma_l \in \mathbb{R}$, $\beta_l > 0$, and $\zeta : [0,\infty) \to [0,\infty)$ is some smooth cut-off function with $\zeta \equiv 1$ in a neighborhood of 0, and $\operatorname{supp}\zeta \subset [0, \min_l(\beta_l, \beta_l^{-1})]$. Its adjoint reads as

$$\breve{G}_1^* w(x) = w(x) + \zeta(x) \sum_{l=0}^{L} \frac{\gamma_l}{\beta_l} w\Big(\frac{-x}{\beta_l}\Big) \quad (w \in L_2(-1,1),\, x \in \mathcal{I}).$$

A Hestenes extension satisfies (5.5.2) if and only if

$$\sum_{l=0}^{L} \gamma_l \beta_l^i = (-1)^i \,(\mathbb{N}_0 \ni i \le \lfloor t - \tfrac{1}{2} \rfloor), \quad \sum_{l=0}^{L} \gamma_l \beta_l^{-(j+1)} = (-1)^{j+1} \,(\mathbb{N}_0 \ni j \le \lfloor \tilde{t} - \tfrac{1}{2} \rfloor).$$

With a univariate extension $\breve{G}_1$ as in (5.5.2) at hand, the obvious approach is to define $E_1^{(q)}$ according to Proposition 5.4.4 with $G_1 = \breve{G}_1$. A problem with the choice $G_1 = \breve{G}_1$ is that generally (5.5.3) does *not* imply the desirable property that $\operatorname{diam}(\operatorname{supp} G_1 u) \lesssim \operatorname{diam}(\operatorname{supp} u)$. Indeed, think of the application of a Hestenes extension to a u with a small support that is not located near the interface.

To solve this and the corresponding problem for the adjoint extension, in any case for u being any primal or dual wavelet, respectively, following [44] we will apply our construction using the modified, *scale-dependent* univariate extension operator

$$G_1 : u \mapsto \sum_{\lambda \in \nabla_{0,0}^{(\ell)}} \langle u, \tilde{\psi}_\lambda^{(\vec{0},\vec{0})} \rangle_{L_2(\mathcal{I})} \breve{G}_1 \psi_\lambda^{(\vec{0},\vec{0})} + \sum_{\lambda \in \nabla^{(I)} \cup \nabla_{0,0}^{(r)}} \langle u, \tilde{\psi}_\lambda^{(\vec{0},\vec{0})} \rangle_{L_2(\mathcal{I})} \eta_1 \psi_\lambda^{(\vec{0},\vec{0})}. \tag{5.5.5}$$

Taking $\breve{G}_1$ to be a Hestenes extension, under the condition of ϱ being sufficiently small, its first advantage is that its application in (5.5.5) does not "see" the cut-off function ζ, which prevents potential quadrature problems.

Proposition 5.5.2. *Assuming ϱ to be sufficiently small, the scale-dependent extension G_1 from* (5.5.5) *satisfies, for $\vec{\sigma} \in \{0,\dots,\lfloor t+\frac{1}{2}\rfloor\}^2$, $\vec{\tilde{\sigma}} \in \{0,\dots,\lfloor \tilde{t}+\frac{1}{2}\rfloor\}^2$*

$$G_1\psi_\mu^{(\vec{\sigma},\vec{\tilde{\sigma}})} = \begin{cases} \eta_1\psi_\mu^{(\vec{\sigma},\vec{\tilde{\sigma}})} & \text{when } \mu \in \nabla^{(I)} \cup \nabla^{(r)}_{\sigma_r,\tilde{\sigma}_r}, \\ \breve{G}_1\psi_\mu^{(\vec{\sigma},\vec{\tilde{\sigma}})} & \text{when } \mu \in \nabla^{(\ell)}_{\sigma_\ell,\tilde{\sigma}_\ell}. \end{cases} \tag{5.5.6}$$

Assuming, additionally, $\breve{G}_1$ to be a Hestenes extension with $\beta_l = 2^l$, *the resulting adjoint extension $G_2 := (\mathrm{Id} - \eta_1 G_1^*)\eta_2$ satisfies*

$$G_2(\tilde{\psi}_\mu^{(\vec{\sigma},\vec{\tilde{\sigma}})}(1+\cdot)) = \begin{cases} \eta_2(\tilde{\psi}_\mu^{(\vec{\sigma},\vec{\tilde{\sigma}})}(1+\cdot)) & \text{when } \mu \in \nabla^{(I)} \cup \nabla^{(\ell)}_{\sigma_\ell,\tilde{\sigma}_\ell}, \\ \breve{G}_2(\tilde{\psi}_\mu^{(\vec{\sigma},\vec{\tilde{\sigma}})}(1+\cdot)) & \text{when } \mu \in \nabla^{(r)}_{\sigma_r,\tilde{\sigma}_r}. \end{cases} \tag{5.5.7}$$

We have $G_1 \in B(L_2(0,1),L_2(-1,1))$, $G_1 \in B(H^t(0,1),H^t(-1,1))$, and $G_1^* \in B(H^{\tilde{t}}(-1,1),H^{\tilde{t}}_{(\lfloor\tilde{t}+\frac{1}{2}\rfloor,0)}(0,1))$.

Finally, for $\mu \in \nabla_{\vec{\sigma},\vec{\tilde{\sigma}}}$, it holds that

$$\begin{aligned} \operatorname{diam}(\operatorname{supp}\breve{G}_1\psi_\mu^{(\vec{\sigma},\vec{\tilde{\sigma}})}) &\lesssim \operatorname{diam}(\operatorname{supp}\psi_\mu^{(\vec{\sigma},\vec{\tilde{\sigma}})}), \\ \operatorname{diam}(\operatorname{supp}\breve{G}_2\tilde{\psi}_\mu^{(\vec{\sigma},\vec{\tilde{\sigma}})}) &\lesssim \operatorname{diam}(\operatorname{supp}\tilde{\psi}_\mu^{(\vec{\sigma},\vec{\tilde{\sigma}})}). \end{aligned} \tag{5.5.8}$$

Proof. By $(\mathcal{W}_8)$(ii), for $\mu \in \nabla^{(I)} \cup \nabla^{(r)}_{\sigma_r,\tilde{\sigma}_r}$, $\lambda \in \nabla^{(\ell)}_{0,0}$, one has $\langle\psi_\mu^{(\vec{\sigma},\vec{\tilde{\sigma}})},\tilde{\psi}_\lambda^{(\vec{0},\vec{0})}\rangle_{L_2(\mathcal{I})} = 0$, and so $G_1\psi_\mu^{(\vec{\sigma},\vec{\tilde{\sigma}})} = \sum_{\lambda\in\nabla_{\vec{0},\vec{0}}}\langle\psi_\mu^{(\vec{\sigma},\vec{\tilde{\sigma}})},\tilde{\psi}_\lambda^{(\vec{0},\vec{0})}\rangle_{L_2(\mathcal{I})}\eta_1\psi_\lambda^{(\vec{0},\vec{0})} = \eta_1\psi_\mu^{(\vec{\sigma},\vec{\tilde{\sigma}})}$, the last equality from $\Psi^{(\vec{0},\vec{0})}$ being a Riesz basis for $L_2(\mathcal{I})$, and η_1 being L_2-bounded.

Similarly, for $\mu \in \nabla^{(\ell)}_{\sigma_\ell,\tilde{\sigma}_\ell}$, $\lambda \in \nabla^{(I)} \cup \nabla^{(r)}_{0,0}$, it holds that $\langle\psi_\mu^{(\vec{\sigma},\vec{\tilde{\sigma}})},\tilde{\psi}_\lambda^{(\vec{0},\vec{0})}\rangle_{L_2(\mathcal{I})} = 0$, and so $G_1\psi_\mu^{(\vec{\sigma},\vec{\tilde{\sigma}})} = \sum_{\lambda\in\nabla_{(\vec{0},\vec{0})}}\langle\psi_\mu^{(\vec{\sigma},\vec{\tilde{\sigma}})},\tilde{\psi}_\lambda^{(\vec{0},\vec{0})}\rangle_{L_2(\mathcal{I})}\breve{G}_1\psi_\lambda^{(\vec{0},\vec{0})} = \breve{G}_1\psi_\mu^{(\vec{\sigma},\vec{\tilde{\sigma}})}$.

If $\breve{G}_1$ is a Hestenes extension with $\beta_l = 2^l$, then for $v \in L_2(\mathcal{I})$,

$$\begin{aligned} G_1^*\eta_2(v(1+\cdot)) &= \sum_{\lambda\in\nabla_{\vec{0},\vec{0}}}\langle G_1^*\eta_2(v(1+\cdot)),\psi_\lambda^{(\vec{0},\vec{0})}\rangle_{L_2(\mathcal{I})}\tilde{\psi}_\lambda^{(\vec{0},\vec{0})} \\ &= \sum_{\lambda\in\nabla_{\vec{0},\vec{0}}}\langle v(1-\cdot),(R_2G_1\psi_\lambda^{(\vec{0},\vec{0})})(-\cdot)\rangle_{L_2(\mathcal{I})}\tilde{\psi}_\lambda^{(\vec{0},\vec{0})} \\ &= \sum_{\lambda\in\nabla^{(\ell)}_{0,0}}\langle v(1-\cdot),(R_2\breve{G}_1\psi_\lambda^{(\vec{0},\vec{0})})(-\cdot)\rangle_{L_2(\mathcal{I})}\tilde{\psi}_\lambda^{(\vec{0},\vec{0})} \\ &= \sum_{\lambda\in\nabla^{(\ell)}_{0,0}}\Big\langle v(1-\cdot),\sum_{l=0}^{L}\gamma_l\psi_\lambda^{(\vec{0},\vec{0})}(2^l\cdot)\Big\rangle_{L_2(\mathcal{I})}\tilde{\psi}_\lambda^{(\vec{0},\vec{0})}. \end{aligned} \tag{5.5.9}$$

For $v = \tilde{\psi}_\mu^{(\vec{\sigma},\vec{\tilde{\sigma}})}$ and $\mu \in \nabla^{(I)} \cup \nabla^{(\ell)}_{\sigma_\ell,\tilde{\sigma}_\ell}$, (5.5.9) is zero by $(\mathcal{W}_9)$, $(\mathcal{W}_{10})$, and $(\mathcal{W}_8)$(ii). For $v = \tilde{\psi}_\mu^{(\vec{\sigma},\vec{\tilde{\sigma}})}$ and $\mu \in \nabla^{(r)}_{\sigma_r,\tilde{\sigma}_r}$, one has $\left\langle v(1-\cdot), \sum_{l=0}^{L} \gamma_l (\zeta \psi_\lambda^{(\vec{0},\vec{0})})(2^l \cdot) \right\rangle_{L_2(\mathcal{I})} = 0$ for $\lambda \in \nabla^{(I)} \cup \nabla^{(r)}_{0,0}$ by $(\mathcal{W}_9)$, $(\mathcal{W}_{10})$, and $(\mathcal{W}_8)$(ii). So for those μ, one has $G_1^* \eta_2 \tilde{\psi}_\mu^{(\vec{\sigma},\vec{\tilde{\sigma}})} = \breve{G}_1^* \eta_2 \tilde{\psi}_\mu^{(\vec{\sigma},\vec{\tilde{\sigma}})}$, which completes the proof of (5.5.7).

Since $\operatorname{span}\{\psi_\mu^{(\vec{0},\vec{0})} : \mu \in \nabla^{(I)} \cup \nabla^{(r)}_{0,0}\} + \operatorname{span}\{\psi_\mu^{(\vec{0},\vec{0})} : \mu \in \nabla^{(\ell)}_{0,0}\}$ defines a stable splitting of both $L_2(\mathcal{I})$ and $H^t(\mathcal{I})$ into two subspaces, the statements about the boundedness of G_1 follow from (5.5.6) with $(\vec{\sigma},\vec{\tilde{\sigma}}) = (\vec{0},\vec{0})$, (5.5.2), and $(\mathcal{W}_8)$(iii).

The mapping $P : u \mapsto \sum_{\mu \in \nabla^{(I)} \cup \nabla^{(\ell)}_{\sigma_\ell,\tilde{\sigma}_\ell}} \langle u, \psi_\mu^{(\vec{\sigma},\vec{\tilde{\sigma}})}(1+\cdot)\rangle_{L_2(-1,0)} \eta_2(\tilde{\psi}_\mu^{(\vec{\sigma},\vec{\tilde{\sigma}})}(1+\cdot))$ is in $B(H^{\tilde{t}}(-1,1), H^{\tilde{t}}(-1,1))$ by the assumption on $\tilde{\Psi}^{(\vec{\sigma},\vec{\tilde{\sigma}})}$ and $(\mathcal{W}_8)$(ii). Since $\Psi^{(\vec{\sigma},\vec{\tilde{\sigma}})}(1+\cdot)$ is a Riesz basis for $L_2(-1,0)$,

$$R_2(I-P)u = \sum_{\mu \in \nabla^{(r)}_{\sigma_r,\tilde{\sigma}_r}} \langle u, \psi_\mu^{(\vec{\sigma},\vec{\tilde{\sigma}})}(1+\cdot)\rangle_{L_2(-1,0)} \tilde{\psi}_\mu^{(\vec{\sigma},\vec{\tilde{\sigma}})}(1+\cdot).$$

We conclude that $(G_1^* - \breve{G}_1^*)\eta_2 R_2(I-P) = 0$. Since G_1 and $\breve{G}_1$ are extensions, we also have $G_1^* \eta_1 = \mathrm{Id} = \breve{G} - 1^* \eta_1$, and so $G_1^*(I-P) = G_1^*(\eta_1 R_1 + \eta_2 R_2)(I-P) = \breve{G}_1^*(I-P)$. Together with $G_1^* P = 0$, from (5.5.2) we conclude that $G_1^* \in B(H^{\tilde{t}}(-1,1), H^{\tilde{t}}_{(\lfloor \tilde{t}+\frac{1}{2} \rfloor,0)}(0,1))$.

The last statement is a direct consequence of (5.5.6) and (5.5.7). □

Remark 5.5.3. Although implicitly claimed otherwise in [44, (4.3.12)], we note that (5.5.7), and so (5.5.8), *cannot* be expected for $\breve{G}_1$ being a general Hestenes extension as given by (5.5.4), so without assuming that $\beta_l = 2^l$.

Moreover, (5.5.7), and so (5.5.8), are only guaranteed when, for $(\sigma_\ell, \tilde{\sigma}_\ell) = (0, \lfloor \tilde{t} + \frac{1}{2} \rfloor)$, for any $\lambda \in \nabla_{\vec{\sigma},\vec{\tilde{\sigma}}}$ for which either $\psi_\lambda^{(\vec{\sigma},\vec{\tilde{\sigma}})}$ *or* $\tilde{\psi}_\lambda^{(\vec{\sigma},\vec{\tilde{\sigma}})}$ "depends on" the boundary conditions imposed at the left boundary, the primal wavelet $\psi_\lambda^{(\vec{\sigma},\vec{\tilde{\sigma}})}$ is extended by the application of $\breve{G}_1$. The reason to emphasize this is that with common biorthogonal wavelet constructions on the interval, the number of dual wavelets that depend on the boundary conditions is larger than that of the primal ones. Note that even if dual wavelets may not enter the computations, their locality as given by (5.5.8) will be used to prove the forthcoming Theorem 5.5.6 about the approximation rates provided by the primal piecewise tensor product wavelets.

Some examples of relevant Hestenes extensions with $\beta_l = 2^l$ are:

- $L = 0$, $\gamma_0 = 1$ (reflection). Satisfies (5.5.2) for $t < \frac{3}{2}$, $\tilde{t} < \frac{1}{2}$,
- $L = 1$, $\gamma_0 = 3$, $\gamma_1 = -2$. Satisfies (5.5.2) for $t < \frac{5}{2}$, $\tilde{t} < \frac{1}{2}$,
- $L = 1$, $\gamma_0 = -3$, $\gamma_1 = 4$. Satisfies (5.5.2) for $t < \frac{3}{2}$, $\tilde{t} < \frac{3}{2}$,
- $L = 2$, $\gamma_0 = -5$, $\gamma_1 = 10$, $\gamma_2 = -4$. Satisfies (5.5.2) for $t < \frac{5}{2}$, $\tilde{t} < \frac{3}{2}$.

In order to identify individual wavelets from the collections constructed by the applications of the extension operators, we have to introduce some more notations. For $0 \le q \le N$, we set the index sets

$$\begin{aligned}
\boldsymbol{\nabla}_k^{(0)} &:= \boldsymbol{\nabla}_{\boldsymbol{\sigma}(k),\tilde{\boldsymbol{\sigma}}(k)} \times \{k\} \text{ and, for } q > 0,\\
\boldsymbol{\nabla}_k^{(q)} &:= \begin{cases} \boldsymbol{\nabla}_{k_1}^{(q-1)} \cup \boldsymbol{\nabla}_{k_2}^{(q-1)} & \text{if } k = \bar{k},\\ \boldsymbol{\nabla}_{\hat{k}}^{(q-1)} & \text{if } k \in \{q,\dots,N\} \setminus \{\bar{k}\} \text{ and } \Omega_k^{(q)} = \Omega_{\hat{k}}^{(q-1)}, \end{cases}
\end{aligned}$$

and, for $(\lambda, p) \in \boldsymbol{\nabla}_k^{(q)}$, the primal and dual wavelets,

$$\boldsymbol{\psi}_{\boldsymbol{\lambda},p}^{(0,k)} := \boldsymbol{\psi}_{\boldsymbol{\lambda}}^{(\boldsymbol{\sigma}(p),\tilde{\boldsymbol{\sigma}}(p))}(\cdot - \alpha_p), \quad \tilde{\boldsymbol{\psi}}_{\boldsymbol{\lambda},p}^{(0,k)} := \tilde{\boldsymbol{\psi}}_{\boldsymbol{\lambda}}^{(\boldsymbol{\sigma}(p),\tilde{\boldsymbol{\sigma}}(p))}(\cdot - \alpha_p),$$

and, for $q > 0$,

$$\begin{aligned}
\boldsymbol{\psi}_{\boldsymbol{\lambda},p}^{(q,k)} &:= \begin{cases} \left\{ \begin{matrix} E_1^{(q)} \boldsymbol{\psi}_{\boldsymbol{\lambda},p}^{(q-1,k_1)} & (\boldsymbol{\lambda},p) \in \boldsymbol{\nabla}_{k_1}^{(q-1)} \\ \eta_2^{(q)} \boldsymbol{\psi}_{\boldsymbol{\lambda},p}^{(q-1,k_2)} & (\boldsymbol{\lambda},p) \in \boldsymbol{\nabla}_{k_2}^{(q-1)} \end{matrix} \right\} & \text{if } k = \bar{k},\\ \boldsymbol{\psi}_{\boldsymbol{\lambda},p}^{(q-1,\hat{k})} & \text{if } k \in \{q,\dots,N\} \setminus \{\bar{k}\} \text{ and } \Omega_k^{(q)} = \Omega_{\hat{k}}^{(q-1)}, \end{cases}\\
\tilde{\boldsymbol{\psi}}_{\boldsymbol{\lambda},p}^{(q,k)} &:= \begin{cases} \left\{ \begin{matrix} \eta_1^{(q)} \tilde{\boldsymbol{\psi}}_{\boldsymbol{\lambda},p}^{(q-1,k_1)} & (\boldsymbol{\lambda},p) \in \boldsymbol{\nabla}_{k_1}^{(q-1)} \\ (\mathrm{Id} - \eta_1^{(q)}(E_1^{(q)})^*)\eta_2^{(q)} \tilde{\boldsymbol{\psi}}_{\boldsymbol{\lambda},p}^{(q-1,k_2)} & (\boldsymbol{\lambda},p) \in \boldsymbol{\nabla}_{k_2}^{(q-1)} \end{matrix} \right\} & \text{if } k = \bar{k},\\ \tilde{\boldsymbol{\psi}}_{\boldsymbol{\lambda},p}^{(q-1,\hat{k})} & \text{if } k \in \{q,\dots,N\} \setminus \{\bar{k}\} \text{ and } \Omega_k^{(q)} = \Omega_{\hat{k}}^{(q-1)}, \end{cases}
\end{aligned}$$

Then, as we have seen,

$$(\boldsymbol{\Psi}_k^{(q)}, \tilde{\boldsymbol{\Psi}}_k^{(q)}) := (\{\boldsymbol{\psi}_{\boldsymbol{\lambda},p}^{(q,k)}) : (\boldsymbol{\lambda},p) \in \boldsymbol{\nabla}_k^{(q)}\}, \{\tilde{\boldsymbol{\psi}}_{\boldsymbol{\lambda},p}^{(q,k)} : (\boldsymbol{\lambda},p) \in \boldsymbol{\nabla}_k^{(q)}\})$$

is a pair of biorthogonal Riesz bases for $L_2(\hat{\Omega}_k^{(q)})$, and for $s \in [0,1]$, $\boldsymbol{\Psi}_k^{(q)}$ or $\tilde{\boldsymbol{\Psi}}_k^{(q)}$ are, properly scaled, Riesz bases for $[L_2(\hat{\Omega}_k^{(q)}), \overset{\circ}{H}{}^t(\hat{\Omega}_k^{(q)})]_{s,2}$ and $[L_2(\hat{\Omega}_k^{(q)}), \overset{\circ}{H}{}^{\tilde{t}}(\hat{\Omega}_k^{(q)})]_{s,2}$, respectively.

Proposition 5.5.4. *With $E_1^{(q)}$ being defined using the scale-dependent extension operator as in Proposition 5.5.2, for $0 \le q \le k \le N$, we have*

1. $\operatorname{supp} \boldsymbol{\psi}_{\boldsymbol{\lambda},p}^{(q,k)}$, $\operatorname{supp} \tilde{\boldsymbol{\psi}}_{\boldsymbol{\lambda},p}^{(q,k)}$ *are contained in a hyperrectangle aligned with the Cartesian coordinates with sides in length of order $2^{-|\boldsymbol{\lambda}|_1}, \dots, 2^{-|\boldsymbol{\lambda}|_n}$,*
2. *for any $y \in \mathbb{R}^n$ and $\boldsymbol{j} \in \mathbb{N}_0^n$, the hyperrectangle $y + \prod_{i=1}^n [0, 2^{-j_i}]$ is intersected by the supports at most a uniformly bounded number of primal or dual wavelets $\boldsymbol{\psi}_{\boldsymbol{\lambda},p}^{(q,k)}$, $\tilde{\boldsymbol{\psi}}_{\boldsymbol{\lambda},p}^{(q,k)}$ with $|\boldsymbol{\lambda}| = \boldsymbol{j}$,*
3. *let*

$$\begin{aligned}
V_{\mathbf{j}}(\hat{\Omega}_k^{(q)}) &:= \{u \in \overset{\circ}{H}{}^t(\hat{\Omega}_k^{(q)}) : u(\cdot + \alpha_{k'})|_{\square} \in \otimes_{i=1}^n V_{j_i}^{(\vec{0})} \ (\square_{k'} \subset \hat{\Omega}_k^{(q)})\}\\
Z_{\mathbf{j}}(\hat{\Omega}_k^{(q)}) &:= \operatorname{span}\{\boldsymbol{\psi}_{\boldsymbol{\lambda},p}^{(q,k)} : (\boldsymbol{\lambda},p) \in \boldsymbol{\nabla}_k^{(q)}, |\boldsymbol{\lambda}| \le \boldsymbol{j}\},
\end{aligned}$$

and $\mathbf{e} := (1,\dots,1)^\top \in \mathbb{R}^n$. *Then for some constants* m_q, $M_q \in \mathbb{N}_0$, *for all* $\mathbf{j} \in \{m_q, m_q+1, \dots\}^n$,

$$V_{\mathbf{j}-m_q\mathbf{e}}(\hat\Omega_k^{(q)}) \subset Z_{\mathbf{j}}(\hat\Omega_k^{(q)}) \subset V_{\mathbf{j}+M_q\mathbf{e}}(\hat\Omega_k^{(q)}).$$

Proof. Parts (1) and (2) follow from the locality and the locally finiteness of the univariate primal and dual wavelets ($(\mathcal{W}_5)$ and $(\mathcal{W}_6)$), and the locality of the extension and the adjoint extension given by (5.5.6) and (5.5.7).

By construction of the wavelet basis, the second inclusion in (3) follows from (5.5.6) and $\breve G_1$ being a Hestenes extension with $\beta_l = 2^l$, $(\mathcal{W}_{10})$, $(\mathcal{W}_9)$, and $(\mathcal{W}_7)$. The constant M_q can be taken to be less than or equal to $2L$, or to L when the domain has no cracks.

The first inclusion in (3) holds true for $q = 0$ with $m_0 = 0$ by $(\mathcal{W}_7)$. Suppose, for some m_{q-1}, it is true for $q-1$ and $q-1 \le k \le N$. For some constant $m_q \ge m_{q-1}$ that will be determined below, let $v \in V_{\mathbf{j}-m_q\mathbf{e}}(\hat\Omega_{\bar k}^{(q)})$. Then $R_1^{(q)}v \in V_{\mathbf{j}-m_q\mathbf{e}}(\hat\Omega_{k_1}^{(q-1)}) \subset Z_{\mathbf{j}+(m_{q-1}-m_q)\mathbf{e}}(\hat\Omega_{k_1}^{(q-1)})$, and so

$$E_1^{(q)}R_1^{(q)}v \in Z_{\mathbf{j}+(m_{q-1}-m_q)\mathbf{e}}(\hat\Omega_{\bar k}^{(q)}) \subset Z_{\mathbf{j}}(\hat\Omega_{\bar k}^{(q)}) \tag{5.5.10}$$

by definition of $\mathbf{\Psi}_{\bar k}^{(q)}$.

From (5.5.10), we have $E_1^{(q)}R_1^{(q)}v \in V_{\mathbf{j}+(m_{q-1}+M_q-m_q)\mathbf{e}}(\hat\Omega_{\bar k}^{(q)})$, and so $(I - E_1^{(q)}R_1^{(q)})v \in V_{\mathbf{j}+(m_{q-1}+M_q-m_q)\mathbf{e}}(\hat\Omega_{\bar k}^{(q)})$, and therefore

$$\begin{aligned} &R_2^{(q)}(\mathrm{Id} - E_1^{(q)}R_1^{(q)})v \\ &\qquad \in \{u \in L_2(\hat\Omega_{k_2}^{(q-1)}) : u(\cdot+\alpha_{k'})|_\square \in \otimes_{i=1}^n V_{j_i+m_{q-1}+M_q-m_q}^{(\vec 0)} \ (\square_{k'} \subset \hat\Omega_{k_2}^{(q-1)})\} \end{aligned}$$

Since, as shown in the proof of Proposition 5.4.2, $(\mathrm{Id} - E_1^{(q)}R_1^{(q)})v \in \Im(\eta_2^{(q)}|_{\mathring H^t(\hat\Omega_{k_2}^{(q-1)})})$, and so $R_2^{(q)}(\mathrm{Id} - E_1^{(q)}R_1^{(q)})v \in \mathring H^t(\hat\Omega_{k_2}^{(q-1)})$, we infer that

$$R_2^{(q)}(\mathrm{Id} - E_1^{(q)}R_1^{(q)})v \in V_{\mathbf{j}+(m_{q-1}+M_q-m_q)\mathbf{e}}(\hat\Omega_{k_2}^{(q-1)}) \subset Z_{\mathbf{j}+(2m_{q-1}+M_q-m_q)\mathbf{e}}(\hat\Omega_{k_2}^{(q-1)}).$$

By taking $m_q = 2m_{q-1} + M_q$, we conclude that $(\mathrm{Id} - E_1^{(q)}R_1^{(q)})v = \eta_2^{(q)}R_2^{(q)}(\mathrm{Id} - E_1^{(q)}R_1^{(q)})v \in Z_{\mathbf{j}}(\hat\Omega_{\bar k}^{(q)})$ by definition of $\mathbf{\Psi}_{\bar k}^{(q)}$. Together with (5.5.10), this completes the proof. □

Remark 5.5.5. The above proof shows that for $L = 0$ (reflection), $V_{\mathbf{j}}(\hat\Omega_k^{(q)}) = Z_{\mathbf{j}}(\hat\Omega_k^{(q)})$.

Now we are ready to study the question, raised at the beginning of this section, about the rate of approximation in $\mathring H^m(\hat\Omega)$ from the span of $\mathbf{\Psi} := \mathbf{\Psi}_N^{(N)}$.

Theorem 5.5.6. *Let the $E_1^{(q)}$ be defined using the scale-dependent extension operators as in Proposition 5.5.2. Then for any $\theta \in [0,d)$, and any $0 \le q \le k \le N$, there exists a (nested) sequence $(\boldsymbol{\nabla}_{k,M}^{(q)})_{M\in\mathbb{N}} \subset \boldsymbol{\nabla}_k^{(q)}$ with $\#\boldsymbol{\nabla}_{k,M}^{(q)} \eqsim M$, such that*

$$\inf_{v\in\operatorname{span}\{\boldsymbol{\psi}_{\boldsymbol{\lambda},p}^{(q,k)}:(\boldsymbol{\lambda},p)\in\boldsymbol{\nabla}_{k,M}^{(q)}\}} \|u-v\|_{H^m(\hat\Omega_k^{(q)})} \lesssim M^{-(d-m)} \sqrt{\sum_{\square_{k'}\subset\hat\Omega_k^{(q)}} \|u\|^2_{\mathcal{H}^d_{m,\theta}(\square'_k)}}, \tag{5.5.11}$$

for any $u \in \mathring{H}^m(\hat\Omega_k^{(q)})$ for which the right-hand side is finite (for $q=k=N$, i.e., for $\hat\Omega_k^{(q)} = \hat\Omega$, this is equivalent to saying that u satisfies (5.5.1) *(with $\kappa = \mathrm{Id}$)).*

For $m=0$, the factor $M^{-(d-m)}$ in (5.5.11) *has to be read as $(\log M)^{(n-1)(\frac12+d)}M^{-d}$.*

Proof. We prove the statement with the additional property that the index sets $\boldsymbol{\nabla}_{k,M}^{(q)}$ have the *multiple tree property* introduced in Theorem 5.3.2 for subsets of $\boldsymbol{\nabla}_{(\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}})}$, and that in the current generalized setting reads as: For any $(\boldsymbol{\lambda},p) \in \boldsymbol{\nabla}_{k,M}^{(q)}$ and any $\boldsymbol{j} \in \mathbb{N}_0^n$ with $\boldsymbol{j} \le |\boldsymbol{\lambda}|$, there exists a $(\boldsymbol{\lambda}',p') \in \boldsymbol{\nabla}_{k,M}^{(q)}$ with $|\boldsymbol{\lambda}'| = \boldsymbol{j}$, and $\operatorname{supp}\boldsymbol{\psi}_{\boldsymbol{\lambda},p}^{(q,k)} \cap \operatorname{supp}\boldsymbol{\psi}_{\boldsymbol{\lambda}',p'}^{(q,k)} \neq \emptyset$.

For $q=0$, the so extended statement is equal to that of Theorem 5.3.2. Let us assume that the statement is valid for some $0 \le q-1 \le N-1$.

To prove the statement for q, it is sufficient to consider $k=\bar k$. Let ϱ be a smooth function on $\mathbb{R}^n$ such that for some sufficiently small $\varepsilon_2 > \varepsilon_1 > 0$, $\varrho \equiv 1$ within distance ε_1 of the interface J between $\hat\Omega_{k_1}^{(q-1)}$ and $\hat\Omega_{k_2}^{(q-1)}$, and vanishes outside distance ε_2 of J. Writing any function v on $\hat\Omega_{\bar k}^{(q)}$ as $\varrho v + (1-\varrho)v$ induces a stable splitting of $\mathring{H}^m(\hat\Omega_{\bar k}^{(q)}) \cap \prod_{\square_{k'}\subset\hat\Omega_{\bar k}^{(q)}} \mathcal{H}^d_{m,\theta}(\square_{k'})$ into two subspaces.

For functions u of type $(1-\varrho)v$, one has $u|_{\hat\Omega_{k_2}^{(q-1)}} \in \mathring{H}^m(\hat\Omega_{k_2}^{(q-1)})$, and, assuming ϱ to be sufficiently small, $\langle u|_{\hat\Omega_{k_1}^{(q-1)}}, \tilde{\boldsymbol{\psi}}_{\boldsymbol{\lambda},p}^{(q-1,k_1)}\rangle_{L_2(\hat\Omega_{k_1}^{(q-1)})} = 0$ for all $(\boldsymbol{\lambda},p) \in \boldsymbol{\nabla}_{k_1}^{(q-1)}$ with $\eta_1^{(q)}\boldsymbol{\psi}_{\boldsymbol{\lambda},p}^{(q-1,k_1)} \neq E_1^{(q)}\boldsymbol{\psi}_{\boldsymbol{\lambda},p}^{(q-1,k_1)}$. We conclude that for such functions (5.5.11) is valid when

$$\boldsymbol{\nabla}_{\bar k,M}^{(q)} \supseteq \boldsymbol{\nabla}_{k_1,M}^{(q-1)} \cup \boldsymbol{\nabla}_{k_2,M}^{(q-1)}.$$

In the remainder of this proof, we consider functions of type $u = \varrho v$, so with support inside some sufficiently small neighborhood of J. For $q-1 \le k \le N$, we set the biorthogonal projectors

$$P_{k,M}^{(q-1)} : v \mapsto \sum_{(\boldsymbol{\lambda},p)\in\boldsymbol{\nabla}_{k,M}^{(q-1)}} \langle v, \tilde{\boldsymbol{\psi}}_{\boldsymbol{\lambda},p}^{(q-1,k)}\rangle_{L_2(\hat\Omega_k^{(q-1)})} \boldsymbol{\psi}_{\boldsymbol{\lambda},p}^{(q-1,k)}.$$

W.l.o.g. we assume $J = \{0\}\times\breve{J}$ and define the (scale-independent) extension $\hat E_1^{(q)}$ as $E_1^{(q)}$ with G_1 reading as $\hat G_1$, defined by $\hat G_1 v(-x) = \sum_{l=0}^L \gamma_l v(2^l x)$ and $\hat G_1 v(x) = v(x)$ $(x \in \mathcal{I})$. So $\hat G_1$ is the Hestenes extension $\breve G_1$ without the smooth cut-off function which is not needed here because of the assumption on $\operatorname{supp} u$.

It holds that $R_2^{(q)}(\mathrm{Id}-\hat{E}_1^{(q)}R_1^{(q)})u \in \mathring{H}^m(\hat{\Omega}_{k_2}^{(q-1)})$ and $R_1^{(q)}u \in \mathring{H}^m(\hat{\Omega}_{k_1}^{(q-1)})$. Since $\hat{E}_1^{(q)}$ preserves the piecewise weighted Sobolev smoothness of a function supported near the interface, we have

$$\begin{aligned}&\sum_{\square_{k'}\subset\hat{\Omega}_{k_2}^{(q-1)}}\|R_2^{(q)}(\mathrm{Id}-\hat{E}_1^{(q)}R_1^{(q)})u\|^2_{\mathcal{H}^d_{m,\theta}(\square_{k'})}+\sum_{\square_{k'}\subset\hat{\Omega}_{k_1}^{(q-1)}}\|R_1^{(q)}u\|^2_{\mathcal{H}^d_{m,\theta}(\square_{k'})}\\&\lesssim\sum_{\square_{k'}\subset\hat{\Omega}_{\bar{k}}^{(q)}}\|u\|^2_{\mathcal{H}^d_{m,\theta}(\square_{k'})}.\end{aligned}\tag{5.5.12}$$

Setting $u_1 := P_{k_1,M}^{(q-1)}R_1^{(q)}u$, $u_2 := P_{k_2,M}^{(q-1)}R_2^{(q)}(\mathrm{Id}-\hat{E}_1^{(q)}R_1^{(q)})u$, from $[\hat{E}_1^{(q)}\quad \eta_2^{(q)}] \in B(\mathring{H}^m(\hat{\Omega}_{k_1}^{(q-1)})\times\mathring{H}^m(\hat{\Omega}_{k_2}^{(q-1)}),\mathring{H}^m(\hat{\Omega}_{\bar{k}}^{(q)}))$ (see Proposition 5.4.2), we conclude that

$$\begin{aligned}&\|u-(\hat{E}_1^{(q)}u_1+\eta_2^{(q)}u_2)\|_{H^m(\hat{\Omega}_{\bar{k}}^{(q)})}\\&\quad=\left\|[\hat{E}_1^{(q)}\quad\eta_2^{(q)}]\left(\begin{bmatrix}R_1^{(q)}\\R_2^{(q)}(\mathrm{Id}-\hat{E}_1^{(q)}R_1^{(q)})\end{bmatrix}u-\begin{bmatrix}u_1\\u_2\end{bmatrix}\right)\right\|_{H^m(\hat{\Omega}_{\bar{k}}^{(q)})}\\&\quad\lesssim\sqrt{\|R_1^{(q)}u-u_1\|^2_{H^m(\hat{\Omega}_{k_1}^{(q-1)})}+\|R_2^{(q)}(\mathrm{Id}-\hat{E}_1^{(q)}R_1^{(q)})u-u_2\|^2_{H^m(\hat{\Omega}_{k_2}^{(q-1)})}}\\&\quad\lesssim M^{-(d-m)}\sqrt{\sum_{\square_{k'}\subset\hat{\Omega}_{\bar{k}}^{(q)}}\|u\|^2_{\mathcal{H}^d_{m,\theta}(\square'_k)}},\end{aligned}\tag{5.5.13}$$

the last inequality by the induction hypothesis and (5.5.12).

Next, we write

$$u-(E_1^{(q)}u_1+\eta_2^{(q)}u_2)=u-(\hat{E}_1^{(q)}u_1+\eta_2^{(q)}u_2)+(\hat{E}_1^{(q)}-E_1^{(q)})u_1.\tag{5.5.14}$$

By construction of G_1 from $\breve{G}_1$, we have that $(\mathrm{Id}-\eta_2^{(q)}R_2^{(q)})(\hat{E}_1^{(q)}-E_1^{(q)})u_1=0$, and $R_2^{(q)}(\hat{E}_1^{(q)}-E_1^{(q)})u_1\in\mathring{H}^m(\hat{\Omega}_{k_2}^{(q-1)})$, and so

$$(\hat{E}_1^{(q)}-E_1^{(q)})u_1=\sum_{(\hat{\boldsymbol{\lambda}},\hat{p})\in\boldsymbol{\nabla}_{k_2}^{(q-1)}}\langle R_2^{(q)}(\hat{E}_1^{(q)}-E_1^{(q)})u_1,\tilde{\boldsymbol{\psi}}_{\hat{\boldsymbol{\lambda}},\hat{p}}^{(q-1,k_2)}\rangle_{L_2(\hat{\Omega}_{k_2}^{(q-1)})}\eta_2^{(q)}\boldsymbol{\psi}_{\hat{\boldsymbol{\lambda}},\hat{p}}^{(q-1,k_2)}.$$

We set

$$\begin{aligned}\hat{\boldsymbol{\nabla}}_{k_2,M}^{(q-1)}:=\{(\hat{\boldsymbol{\lambda}},\hat{p})\in\boldsymbol{\nabla}_{k_2}^{(q-1)}:\langle R_2^{(q)}(\hat{E}_1^{(q)}-E_1^{(q)})u_1,\tilde{\boldsymbol{\psi}}_{\hat{\boldsymbol{\lambda}},\hat{p}}^{(q-1,k_2)}\rangle_{L_2(\hat{\Omega}_{k_2}^{(q-1)})}\neq 0,\\ \text{for some } u_1\in\Im P_{k_1,M}^{(q-1)}\}.\end{aligned}$$

Below we will show that, even after a possible enlargement to ensure the multiple tree property, it holds that $\#\hat{\boldsymbol{\nabla}}_{k_2,M}^{(q-1)}\lesssim\#\boldsymbol{\nabla}_{k_1,M}^{(q-1)}$. Defining $\boldsymbol{\nabla}_{\bar{k},M}^{(q)}:=\boldsymbol{\nabla}_{k_1,M}^{(q-1)}\cup\boldsymbol{\nabla}_{k_2,M}^{(q-1)}\cup\hat{\boldsymbol{\nabla}}_{k_2,M}^{(q-1)}$, the proof is completed.

If $(\hat{\boldsymbol{\lambda}},\hat{p})\in\hat{\boldsymbol{\nabla}}_{k_2,M}^{(q-1)}$, then $\langle R_2^{(q)}\hat{E}_1^{(q)}\boldsymbol{\psi}_{\boldsymbol{\lambda},p}^{(q-1,k_1)},\tilde{\boldsymbol{\psi}}_{\hat{\boldsymbol{\lambda}},\hat{p}}^{(q-1,k_2)}\rangle_{L_2(\hat{\Omega}_{k_2}^{(q-1)})}\neq 0$ for some $(\boldsymbol{\lambda},p)\in\boldsymbol{\nabla}_{k_1,M}^{(q-1)}$ with $R_2^{(q)}\hat{E}_1^{(q)}\boldsymbol{\psi}_{\boldsymbol{\lambda},p}^{(q-1,k_1)}\in\mathring{H}^m(\hat{\Omega}_{k_2}^{(q-1)})$. Using $Z_{|\boldsymbol{\lambda}|}(\hat{\Omega}_{k_1}^{(q-1)})\subset V_{|\boldsymbol{\lambda}|+M_{q-1}\mathbf{e}}(\hat{\Omega}_{k_1}^{(q-1)})$ and the assumptions on the extension, we have

$$R_2^{(q)}\hat{E}_1^{(q)}\boldsymbol{\psi}_{\boldsymbol{\lambda},p}^{(q-1,k_1)}\in V_{|\boldsymbol{\lambda}|+(M_{q-1}+L)\mathbf{e}}(\hat{\Omega}_{k_2}^{(q-1)})\subset Z_{|\boldsymbol{\lambda}|+(M_{q-1}+m_{q-1}+L)\mathbf{e}}(\hat{\Omega}_{k_2}^{(q-1)})$$

and so $|\hat{\boldsymbol{\lambda}}|\leq|\boldsymbol{\lambda}|+(M_{q-1}+m_{q-1}+L)\mathbf{e}$. Here we applied both inclusions from Proposition 5.5.4(3).

Thanks to the multiple tree property of $\boldsymbol{\nabla}_{k_1,M}^{(q-1)}$, there exists a $(\boldsymbol{\lambda}',p')\in\boldsymbol{\nabla}_{k_1,M}^{(q-1)}$ with $|\boldsymbol{\lambda}'|_i=\min(|\hat{\boldsymbol{\lambda}}|_i,|\boldsymbol{\lambda}|_i)$ $(1\leq i\leq n)$ and $\operatorname{supp}\boldsymbol{\psi}_{\boldsymbol{\lambda},p}^{(q-1,k_1)}\cap\operatorname{supp}\boldsymbol{\psi}_{\boldsymbol{\lambda}',p'}^{(q-1,k_1)}\neq\emptyset$. Note that because of $|\hat{\boldsymbol{\lambda}}|\leq|\boldsymbol{\lambda}|+(M_{q-1}+m_{q-1}+L)\mathbf{e}$, we have $|\boldsymbol{\lambda}'|\leq|\hat{\boldsymbol{\lambda}}|\leq|\boldsymbol{\lambda}'|+(M_{q-1}+m_{q-1}+L)\mathbf{e}$.

The "localness" of $\boldsymbol{\Psi}_{k_1}^{(q-1)}$ as given by Proposition 5.5.4(1), the assumptions on the extension, and the "locally finiteness" of $\tilde{\boldsymbol{\Psi}}_{k_2}^{(q-1)}$ as given by Proposition 5.5.4(2) show that the number of $(\hat{\boldsymbol{\lambda}},\hat{p})\in\hat{\boldsymbol{\nabla}}_{k_2,M}^{(q-1)}$ with $\langle R_2^{(q)}\hat{E}_1^{(q)}\boldsymbol{\psi}_{\boldsymbol{\lambda},p}^{(q-1,k_1)},\tilde{\boldsymbol{\psi}}_{\hat{\boldsymbol{\lambda}},\hat{p}}^{(q-1,k_2)}\rangle_{L_2(\hat{\Omega}_{k_2}^{(q-1)})}\neq 0$ on the *same* level $|\hat{\boldsymbol{\lambda}}|\leq|\boldsymbol{\lambda}|+(M_{q-1}+m_{q-1}+L)\mathbf{e}$ is uniformly bounded. With this, we conclude that with the above mapping $(\hat{\boldsymbol{\lambda}},\hat{p})\mapsto(\boldsymbol{\lambda}',p')$, an at most uniformly bounded number of $(\hat{\boldsymbol{\lambda}},\hat{p})\in\hat{\boldsymbol{\nabla}}_{k_2,M}^{(q-1)}$ is mapped onto any $(\boldsymbol{\lambda}',p')\in\boldsymbol{\nabla}_{k_1,M}^{(q-1)}$, and so that $\#\hat{\boldsymbol{\nabla}}_{k_2,M}^{(q-1)}\lesssim\#\boldsymbol{\nabla}_{k_1,M}^{(q-1)}$.

Finally, to bound $\#\hat{\boldsymbol{\nabla}}_{k_2,M}^{(q-1)}$, we only used that for $(\hat{\boldsymbol{\lambda}},\hat{p})\in\hat{\boldsymbol{\nabla}}_{k_2,M}^{(q-1)}$ there exists a $(\boldsymbol{\lambda},p)\in\boldsymbol{\nabla}_{k_1,M}^{(q-1)}$ with $\operatorname{supp}R_2^{(q)}\hat{E}_1^{(q)}\boldsymbol{\psi}_{\boldsymbol{\lambda},p}^{(q-1,k_1)}\cap\operatorname{supp}\tilde{\boldsymbol{\psi}}_{\hat{\boldsymbol{\lambda}},\hat{p}}^{(q-1,k_1)}\neq\emptyset$ and $|\hat{\boldsymbol{\lambda}}|\leq|\boldsymbol{\lambda}|+(M_{q-1}+m_{q-1}+L)\mathbf{e}$. The same proof would have applied with the condition about the non-empty intersection of the supports reading as the condition that $\operatorname{supp}R_2^{(q)}\hat{E}_1^{(q)}\boldsymbol{\psi}_{\boldsymbol{\lambda},p}^{(q-1,k_1)}$ has non-empty intersection with some hyperrectangle, containing $\operatorname{supp}\tilde{\boldsymbol{\psi}}_{\hat{\boldsymbol{\lambda}},\hat{p}}^{(q-1,k_1)}$, that is aligned with the Cartesian coordinates with sides of lengths of order $2^{-|\hat{\boldsymbol{\lambda}}|_1},\ldots,2^{-|\hat{\boldsymbol{\lambda}}|_n}$. In view of this, if $\hat{\boldsymbol{\nabla}}_{k_2,M}^{(q-1)}$ does not already has the multiple tree property, then it can be enlarged to have this property while retaining $\#\hat{\boldsymbol{\nabla}}_{k_2,M}^{(q-1)}\lesssim\#\boldsymbol{\nabla}_{k_1,M}^{(q-1)}$. □

5.6 Regularity

We study the issue whether we may expect (5.5.1) for u being the solution of an elliptic boundary value problem of order $2m=2$.

5.6.1 Two-dimensional case

Let Ω be a *polygonal* domain. This means that its boundary is the union of a finite number of line segments, knowns as *edges*, with ends known as *corners*. It is not assumed that Ω is a Lipschitz domain, so it may contain cracks. We denote with $\mathcal{E}$

the set of edges, with $\mathcal{C}$ the set of corners, and set for $\mathbf{c} \in \mathcal{C}$,

$$r_{\mathbf{c}}(\mathbf{x}) := \operatorname{dist}(\mathbf{x}, \mathbf{c}).$$

Following [29], for $m \in \mathbb{N}_0$, we define the (non-homogeneous) *weighted Sobolev space* $J^m_\beta(\Omega)$ as the set of $u \in L^{\mathrm{loc}}_2(\Omega)$ that have a finite squared norm

$$\|v\|^2_{J^m_\beta(\Omega)} := \sum_{k=0}^{m} \sum_{|\alpha|=k} \|\{\prod_{\mathbf{c}\in\mathcal{C}} r_{\mathbf{c}}^{\beta+m}\}\partial^\alpha v\|^2_{L_2(\Omega)}.$$

(in [29] the generalization is considered of β being possibly dependent on $\mathbf{c}$).

Let A be a constant, real, symmetric and positive definite 2×2 matrix. Let $\mathcal{E}_D \subset \mathcal{E}$, and

$$V(\Omega) := \begin{cases} \{v \in H^1(\Omega) : v|_{\mathbf{e}} = 0 \ \forall \mathbf{e} \in \mathcal{E}_D\} & \text{when } \mathcal{E}_D \neq \emptyset, \\ \{v \in H^1(\Omega) : \int_\Omega v \, \mathrm{d}x = 0\} & \text{otherwise.} \end{cases}$$

Given $g \in V(\Omega)'$, let $u \in V(\Omega)$ denote the solution of

$$\int_\Omega A\nabla u \cdot \nabla v \, \mathrm{d}x = g(v) \quad (v \in V(\Omega)). \tag{5.6.1}$$

Theorem 5.6.1. *For $m \in \mathbb{N}_0$, there exists a $b^* \in (0, m+2]$ such that for any $b \in [0, b^*)$, the mapping $g \mapsto u \in B(J^m_{-b+1}(\Omega), J^{m+2}_{-b-1}(\Omega))$.*

The proof follows from [29, formula (6.7)]. As stated in [29, Example 6.7], for m sufficiently large, $b^* > \frac{1}{4}$.

We refer to [29, Sect. 7] for generalizations of Theorem 5.6.1 to differential operators with variable coefficients and/or lower order terms.

Concerning the smoothness condition on the right-hand side g, note that for $b \leq m+1$,

$$H^m(\Omega) \hookrightarrow J^m_{-b+1}(\Omega).$$

Let us now consider the situation that $\Omega = \cup_{i=1}^K \Omega_i$ is an essentially disjoint subdivision into subdomains, where $\Omega_i = \kappa_i(\square)$ with κ_i being a regular parametrization. Let R_i denote the restriction of functions on Ω to Ω_i.

Proposition 5.6.2. *If $\kappa_i \in C^{m+2}(\overline{\square})$ and $b \leq m+1$, then*

$$\kappa_i^* R_i \in B(J^{m+2}_{-b-1}(\Omega), J^{m+2}_{-b-1}(\square)).$$

Proof. This follows from the smoothness of κ_i, and from the fact that $\kappa_i^* u|_{\Omega_i}$ localized near corners of $\square$ that do not correspond to corners of Ω is a function in $H^{m+2} \hookrightarrow J^{m+2}_{-b-1}$, the latter by $-b-1+m+2 \geq 0$. □

The following Proposition demonstrates (5.5.1).

Proposition 5.6.3. *For $d \in \mathbb{N}_0$, $\theta \geq \max(1, d - b/2)$, it holds that*

$$J^{2d}_{-b-1}(\square) \hookrightarrow \mathcal{H}^d_{\theta-1}(\mathcal{I}) \otimes \mathcal{H}^d_\theta(\mathcal{I}) \cap \mathcal{H}^d_\theta(\mathcal{I}) \otimes \mathcal{H}^d_{\theta-1}(\mathcal{I}) = \mathcal{H}^d_{1,\theta}(\square).$$

Proof. This follows from $\max(x^{\theta-1}y^\theta, x^\theta y^{\theta-1}) \leq r_{\mathbf{0}}^{2\theta-1} \leq r_{\mathbf{0}}^{2d-b-1}$ when $r_{\mathbf{0}} \in [0,1]$. □

5.6.2 Three-dimensional case

As in the previous section we follow [29] closely. Let Ω be a *polyhedral* domain. This means that its boundary is the union of a finite number of polygons, known as the *faces*; the segments forming their boundaries are the *edges*, and the ends of the edges are the *corners*. It is not assumed that Ω is a Lipschitz domain, so it may contain crack surfaces. We denote with $\mathcal{F}$, $\mathcal{E}$, and $\mathcal{C}$ the set of faces, edges, and corners, respectively, and set for $\mathbf{e} \in \mathcal{E}$ and $\mathbf{c} \in \mathcal{C}$,

$$r_{\mathbf{e}}(\mathbf{x}) := \operatorname{dist}(\mathbf{x}, \mathbf{e}),\ r_{\mathbf{c}}(\mathbf{x}) := \operatorname{dist}(\mathbf{x}, \mathbf{c}),\ r_{\mathcal{C}}(\mathbf{x}) := \min_{\mathbf{c}\in\mathcal{C}} r_{\mathbf{c}}(\mathbf{x}),\ r_{\mathcal{E}}(\mathbf{x}) := \min_{\mathbf{e}\in\mathcal{E}} r_{\mathbf{e}}(\mathbf{x}).$$

There exists an $\varepsilon > 0$ small enough such that if we set

$$\begin{aligned}
\Omega_{\mathbf{e}} &:= \{\mathbf{x} \in \Omega : r_{\mathbf{e}}(\mathbf{x}) < \varepsilon,\ r_{\tilde{\mathbf{e}}}(\mathbf{x}) > r_{\mathbf{e}}(\mathbf{x})\,(\mathbf{e} \neq \tilde{\mathbf{e}} \in \mathcal{E}),\ \text{and } r_{\mathcal{C}}(\mathbf{x}) > \tfrac{\varepsilon}{2}\}\\
\Omega_{\mathbf{c}} &:= \{\mathbf{x} \in \Omega : r_{\mathbf{c}}(\mathbf{x}) < \varepsilon \text{ and } r_{\mathcal{E}}(\mathbf{x}) > \tfrac{\varepsilon}{2} r_{\mathbf{c}}(\mathbf{x})\}\\
\Omega_{\mathbf{ce}} &:= \{\mathbf{x} \in \Omega : r_{\mathbf{c}}(\mathbf{x}) < \varepsilon \text{ and } r_{\mathbf{e}}(\mathbf{x}) < \varepsilon r_{\mathbf{c}}(\mathbf{x})\}\\
\Omega_I &:= \{\mathbf{x} \in \Omega : r_{\mathcal{E}}(\mathbf{x}) > \tfrac{\varepsilon}{2}\}
\end{aligned}$$

we have the following properties

$$\begin{gathered}
\overline{\Omega}_{\mathbf{e}} \cap \overline{\Omega}_{\mathbf{e}'} = \emptyset, \quad \overline{\Omega}_{\mathbf{ce}} \cap \overline{\Omega}_{\mathbf{ce}'} = \{\mathbf{c}\} \quad (\mathbf{e} \neq \mathbf{e}' \in \mathcal{E},\ \mathbf{c} \in \mathcal{C}),\\
\overline{B(\mathbf{c};\varepsilon)} \cap \overline{B(\mathbf{c}';\varepsilon)} = \emptyset \quad (\mathbf{c} \neq \mathbf{c}' \in \mathcal{C}), \quad \Omega = \Omega_I \cup_{\{\mathbf{c}\in\mathcal{C}\}} \Omega_{\mathbf{c}} \cup_{\{\mathbf{e}\in\mathcal{E}\}} \Omega_{\mathbf{e}} \cup_{\{\mathbf{c}\in\mathcal{C},\,\mathbf{e}\in\mathcal{E}\}} \Omega_{\mathbf{ce}}.
\end{gathered}$$

In a neighborhood of any edge $\mathbf{e} \in \mathcal{E}$, we will take partial derivatives in an orthogonal coordinate system with one of the coordinate directions being parallel to e. For a multi-index α in that coordinate system, $|\alpha_\perp|$ will denote the sum of the coordinates in the directions perpendicular to e, and $|\alpha_{||}| := |\alpha| - |\alpha_\perp|$.

For $m \in \mathbb{N}_0$, $\beta > -m$, and $\mathcal{E}_0 \subset \mathcal{E}$, we define the *anisotropic weighted Sobolev space*

$$\begin{aligned}
N^m_\beta(\Omega, \mathcal{C}, \mathcal{E}_0) := \Big\{ u \in L_2^{\mathrm{loc}}(\Omega) : \forall \alpha,\ |\alpha| \le m,\ \partial^\alpha u \in L_2(\Omega_I),\\
r_{\mathbf{c}}(\mathbf{x})^{\beta+|\alpha|} \partial^\alpha u \in L_2(\Omega_{\mathbf{c}})\ \forall \mathbf{c} \in \mathcal{C},\\
r_{\mathbf{e}}(\mathbf{x})^{\beta+|\alpha_\perp|} \partial^\alpha u \in L_2(\Omega_{\mathbf{e}})\ \forall \mathbf{e} \in \mathcal{E}_0,\\
r_{\mathbf{c}}(\mathbf{x})^{\beta+|\alpha|} (r_{\mathbf{e}}(\mathbf{x})/r_{\mathbf{c}}(\mathbf{x}))^{\beta+|\alpha_\perp|} \partial^\alpha u \in L_2(\Omega_{\mathbf{ce}})\ \forall \mathbf{c} \in \mathcal{C},\, \mathbf{e} \in \mathcal{E}_0\\
r_{\mathbf{e}}(\mathbf{x})^{\max(\beta+|\alpha_\perp|,0)} \partial^\alpha u \in L_2(\Omega_{\mathbf{e}})\ \forall \mathbf{e} \in \mathcal{E} \setminus \mathcal{E}_0,\\
r_{\mathbf{c}}(\mathbf{x})^{\beta+|\alpha|} (r_{\mathbf{e}}(\mathbf{x})/r_{\mathbf{c}}(\mathbf{x}))^{\max(\beta+|\alpha_\perp|,0)} \partial^\alpha u \in L_2(\Omega_{\mathbf{ce}})\ \forall \mathbf{c} \in \mathcal{C},\, \mathbf{e} \in \mathcal{E} \setminus \mathcal{E}_0 \Big\},
\end{aligned} \tag{5.6.2}$$

with squared norm being the sum over $|\alpha| \le m$ of the squared L_2-norms over Ω_I, $\Omega_{\mathbf{c}}$, $\Omega_{\mathbf{e}}$, $\Omega_{\mathbf{ce}}$, and $\mathbf{c} \in \mathcal{C}$, $\mathbf{e} \in \mathcal{E}$, respectively. (As in the two-dimensional case, this definition can be generalized to β being possibly dependent on $\mathbf{c}$ and $\mathbf{e}$).

The definition of $N^m_\beta(\Omega, \mathcal{C}, \mathcal{E}_0)$ is a special case of a definition of $N^m_\beta(\Omega, \mathcal{C}_0, \mathcal{E}_0)$ from [29] for general $\mathcal{C}_0 \subseteq \mathcal{C}$. In particular, the definition of the (fully) non-homogeneous

anisotropic weighted Sobolev space $N^m_\beta(\Omega) := N^m_\beta(\Omega, \emptyset, \emptyset)$ is obtained from (5.6.2) by taking $\mathcal{E}_0 = \emptyset$, and by replacing $r_{\mathbf{c}}(\mathbf{x})^{\beta+|\alpha|}$ by $r_{\mathbf{c}}(\mathbf{x})^{\max(\beta+|\alpha|,0)}$ on all three occurrences. Obviously,

$$N^m_\beta(\Omega, \mathcal{C}, \mathcal{E}_0) \hookrightarrow N^m_\beta(\Omega). \tag{5.6.3}$$

Let A be a constant, real, symmetric and positive definite 3×3 matrix. Let $\mathcal{F}_D \subset \mathcal{F}$, and

$$V(\Omega) := \begin{cases} \{v \in H^1(\Omega) : v|_{\mathbf{f}} = 0 \ \forall \mathbf{f} \in \mathcal{F}_D\} & \text{when } \mathcal{F}_D \neq \emptyset, \\ \{v \in H^1(\Omega) : \int_\Omega v \, \mathrm{d}x = 0\} & \text{otherwise.} \end{cases}$$

Given $g \in V(\Omega)'$, let $u \in V(\Omega)$ denote the solution of

$$\int_\Omega A\nabla u \cdot \nabla v \, \mathrm{d}x = g(v) \quad (v \in V(\Omega)). \tag{5.6.4}$$

Theorem 5.6.4. *Let $\mathcal{E}_0$ be the set of all $\mathbf{e} \in \mathcal{E}$ that are an edge of an $\mathbf{f} \in \mathcal{F}_D$. There exists a $b^* \in (0,1]$ such that for $m \in \mathbb{N}$, $m > 1$, and for any $b \in [0, b^*)$, the mapping $g \mapsto u \in B\big(N^m_{1-b}(\Omega, \mathcal{C}, \mathcal{E}_0), N^m_{-1-b}(\Omega, \mathcal{C}, \mathcal{E}_0)\big)$.*

Indeed, with the isotropic weighted Sobolev spaces $J^m_\beta(\Omega)$ as defined in [29, Def. 5.9] (where we consider the value of β to be independent of the edges and corners), [92, Th. 7.1] shows that $g \mapsto u \in B(J^0_{1-b}(\Omega), J^2_{1-b}(\Omega))$, and thus that $g \mapsto u \in B(J^0_{1-b}(\Omega), J^1_{1-b}(\Omega))$. Using that $N^m_{1-b}(\Omega, \mathcal{C}, \mathcal{E}_0) \hookrightarrow J^0_{1-b}(\Omega))$, we conclude the statement of the theorem from the *anisotropic regularity shift theorem* [29, (5.25)(a)]. Here we used that the Assumptions 5.5 and 5.13 from [29] for $\mathbf{e} \in \mathcal{E}_0$ or $\mathbf{e} \in \mathcal{E} \setminus \mathcal{E}_0$, respectively, are satisfied by an application of [92, Th. 7.2].

Concerning the smoothness condition on the right-hand side g, note that for $b \leq 1$,

$$H^m(\Omega) \hookrightarrow N^m_{1-b}(\Omega, \mathcal{C}, \mathcal{E}_0).$$

The fact, as proven in Thm. 5.6.4, that for sufficiently smooth right-hand side, the tangential derivatives of sufficiently high order along the edges of Ω of the solution of (5.6.4) are in the (unweighted) $L_2(\Omega)$ space, is essential for our goal of proving approximation rates with piecewise tensor product approximation as for one-dimensional problems.

Let us consider the situation that $\Omega = \cup_{i=1}^K \Omega_i$ is an essentially disjoint conforming subdivision into hexahedra that are images of $\square$ under trilinear diffeomorphisms κ_i, with $\inf_{\mathbf{x}\in\square} |D\kappa_i(\mathbf{x})| > 0$.

Aiming at deriving a three-dimensional analogue of Proposition 5.6.2, care has to be taken that for $\kappa_i^* R_i u$ to be in $N^m_{-1-b}(\square)$, its tangential derivatives along an edge up to order m have to be in $L_2(\square)$. Therefore, we have to ensure that if an edge of $\square$ is mapped onto the boundary of Ω, then lines parallel to this edge are (smoothly) mapped onto lines parallel to the boundary of Ω.

Proposition 5.6.5. *Let for any i, κ_i be such that* if *it maps an edge* $\mathbf{e}$ *of* $\square$ *to an edge of* Ω, then *it maps all three edges that are parallel to* $\mathbf{e}$ *to edges that are parallel to* $\kappa_i(\mathbf{e})$. *Then*

$$\kappa_i^* R_i \in B(N^m_{-1-b}(\Omega), N^m_{-1-b}(\square)).$$

Proof. What has to be shown is that if an edge $\mathbf{e}$ of $\square$ is mapped to the boundary of Ω, then the tangential derivatives along $\mathbf{e}$ of $u \circ \kappa_i$ up to order m are a smooth functions of the tangential derivatives of u along $\kappa_i(\mathbf{e})$ up to order m.

W.l.o.g., let $\mathbf{e}$ be one of the edges $\mathbf{e}^{(1)}, \ldots, \mathbf{e}^{(4)}$ that are parallel to the first unit vector. The vector $\partial_1 \kappa_i(\mathbf{x})$ is a bilinear function of x_2 and x_3, and so in particular constant on each of the $\mathbf{e}^{(j)}$. These constant vectors are the differences of the endpoints of $\kappa_i(\mathbf{e}^{(j)})$, and so, by assumption, multiples of $\partial_1 \kappa_i|_{\mathbf{e}}$. We conclude that $\partial_1 \kappa_i(\mathbf{x})$ is a multiple of a bilinear *scalar* function and $\partial_1 \kappa_i|_{\mathbf{e}}$. □

Next we will show that the condition on the parametrizations imposed in Proposition 5.6.5 can always be satisfied by making some refinement of the initial conforming subdivision into hexahedra: Let us cut each hexahedron in the partition along 6 planes parallel to the 6 faces of the hexahedron on distance $\zeta > 0$, see Figure 5.5. When ζ

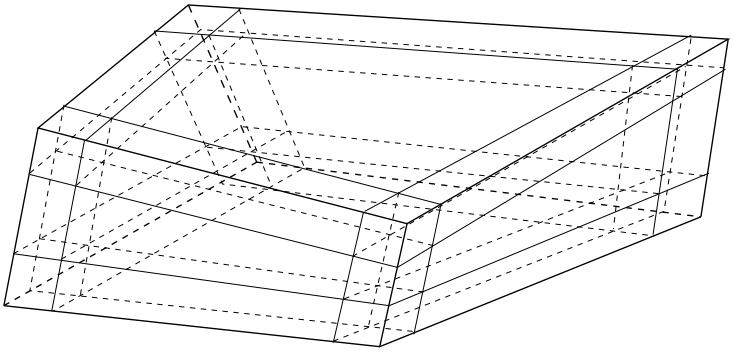


Figure 5.5: Hexahedron cut into 3^3 subhexahedra.

is small enough, then the planes parallel to opposite faces of the hexahedron do not intersect inside the hexahedron, and we obtain a subdivision of the hexahedron in 3^3 hexahedra. Eight of these hexahedra share a corner with the original hexahedron and so have three edges on edges of this hexahedron, and so possibly three edges on edges of Ω. These hexahedra are parallelepipeds and so satisfy the condition from Proposition 5.6.5. Twelve other hexahedra have one edge on an edge of the original hexahedron, and so possibly on an edge of Ω. For each of these hexahedra, the edges opposite to this specific edge are parallel to this edge and so satisfy the condition from Proposition 5.6.5. The remaining seven hexahedra have no edges on edges of the original hexahedron, and thus no edge on an edge of Ω. Six of them have a face on a face of the original hexahedron, whereas the boundary of the remaining "interior" hexahedron has empty intersection with the boundary of the original hexahedron.

Above subdivision of a hexahedron induces a subdivision of each of its faces into 3^2 quadrilaterals; 4 parallelograms at the corners, 4 trapezoids at the edges, and one interior quadrilateral. Conversely, such a subdivision of 3 non-opposite faces of the hexahedron, where the interior quadrilaterals are sufficiently large, determines uniquely the subdivision of the hexahedron into 3^3 subhexahedra by making cuts along planes parallel to the faces. So if we start with a subdivision of one hexahedron and use the resulting subdivision of its faces to induce subdivisions of its neighbors, then by choosing ζ small enough we obtain a refinement of the original conforming decomposition into hexahedra to a conforming decomposition into hexahedra that satisfy the conditions needed for Proposition 5.6.5.

What is left to show is whether the hexahedra in the refined subdivision are images of $\square$ under trilinear diffeomorphisms κ_i, with $\inf_{\mathbf{x}\in\square} |D\kappa_i(\mathbf{x})| > 0$. When the aforementioned parameter ζ tends to zero, the interior hexahedron converges to the hexahedron in the original decomposition, which was assumed to have this property. So for ζ small enough, the interior hexahedra have this property.

The other hexahedra in the refined subdivision have at least two parallel faces, and so are instances of a prismatoid. Let us consider such a hexahedron with its parallel faces, being convex quadrilaterals, on the planes $z = 0$ and $z = 1$. Let $q_1, q_2 : (0,1)^2 \to \mathbb{R}^2$ be bilinear parametrizations of the bottom and top face with $\inf_{(x,y)\in(0,1)^2} |Dq_i(x,y)| > 0$, and such that the images of each corner of $(0,1)^2$ under q_1 and q_2 are connected by an edge in the hexahedron. Then a trilinear parametrization $\square \to \mathbb{R}^3$ is given by

$$\kappa(x,y,z) = (1-z)q_1(x,y) + zq_2(x,y)$$

and so $\inf_{(x,y,z)\in\square} |D\kappa(x,y,z)| = \inf_{(x,y,z)\in\square}(1-z)|Dq_1(x,y)| + z|Dq_2(x,y)| > 0$.

The following Proposition demonstrates (5.5.1).

Proposition 5.6.6. *For $d \in \mathbb{N}_0$, $\theta \geq \max(1, d - \frac{b}{3})$ where $b > 0$, it holds that*

$$N^{3d}_{-1-b}(\square) \hookrightarrow \mathcal{H}^d_{1,\theta}(\square).$$

Proof. It is sufficient to show continuity of the embedding of the spaces restricted to $\Omega_{\mathbf{c}}$, $\Omega_{\mathbf{e}}$, and $\Omega_{\mathbf{ce}}$ intersected with $(0, \frac{1}{2})^3$, where $\mathbf{c} = (0,0,0)$ and $\mathbf{e} = \mathbf{e}_1$.

For $\|\alpha\|_\infty \leq d$, the conditions on θ show that on $\Omega_{\mathbf{c}} \cap (0, \frac{1}{2})^3$,

$$\max(x^{\theta-1}y^\theta z^\theta, x^\theta y^{\theta-1} z^\theta, x^\theta y^\theta z^{\theta-1}) \leq r_{\mathbf{c}}(\mathbf{x})^{3\theta-1} \leq r_{\mathbf{c}}(\mathbf{x})^{\max(-1-b+|\alpha|,0)},$$

and on $\Omega_{\mathbf{e}} \cap (0, \frac{1}{2})^3$,

$$\max(y^\theta z^\theta, y^{\theta-1} z^\theta, y^\theta z^{\theta-1}) \leq r_{\mathbf{e}}(\mathbf{x})^{2\theta-1} \leq r_{\mathbf{e}}(\mathbf{x})^{\max(-1-b+\alpha_2+\alpha_3,0)}.$$

On $\Omega_{\mathbf{ce}} \cap (0, \frac{1}{2})^3$, we have

$$\max(x^{\theta-1}y^\theta z^\theta, x^\theta y^{\theta-1} z^\theta, x^\theta y^\theta z^{\theta-1}) \leq r_{\mathbf{c}}(\mathbf{x})^\theta r_{\mathbf{e}}(\mathbf{x})^{2\theta-1}.$$

To show that this right-hand side can be bounded on

$$r_{\mathbf{c}}(\mathbf{x})^{\max(-1-b+|\alpha|,0)-\max(-1-b+\alpha_2+\alpha_3,0)} r_{\mathbf{e}}(\mathbf{x})^{\max(-1-b+\alpha_2+\alpha_3,0)}$$

we distinguish between 3 cases: For $-1-b+|\alpha| \leq 0$, this results from $\theta \geq 0$ and $2\theta - 1 \geq 0$. For $-1-b+|\alpha| \geq 0 \geq -1-b+\alpha_2+\alpha_3$, we have $r_{\mathbf{c}}(\mathbf{x})^{\theta} r_{\mathbf{e}}(\mathbf{x})^{2\theta-1} \leq r_{\mathbf{c}}(\mathbf{x})^{3\theta-1} \leq r_{\mathbf{c}}(\mathbf{x})^{-1-b+|\alpha|}$ by $\theta \geq d - \frac{b}{3}$. For $-1-b+\alpha_2+\alpha_3 \geq 0$, $r_{\mathbf{c}}(\mathbf{x})^{\theta} r_{\mathbf{e}}(\mathbf{x})^{2\theta-1} \leq r_{\mathbf{c}}(\mathbf{x})^{\theta+\frac{b}{3}} r_{\mathbf{e}}(\mathbf{x})^{2\theta-1-\frac{b}{3}} \leq r_{\mathbf{c}}(\mathbf{x})^{\alpha_1} r_{\mathbf{e}}(\mathbf{x})^{-1-b+\alpha_2+\alpha_3}$ by $\theta \geq d - \frac{b}{3}$. □

5.7 Numerical results

As the univariate building block of the piecewise tensor product wavelet construction, we apply the C^1, piecewise quartic (so $d = 5$) (multi-) wavelets, with (discontinuous) piecewise quartic duals as constructed in [21]. The primal wavelets satisfy Dirichlet boundary conditions of order 1 at both boundaries 0 and 1, i.e., $\vec{\sigma} = (\sigma_\ell, \sigma_r) = (1,1)$, whereas at the dual side no boundary conditions can be imposed, i.e., $\vec{\tilde{\sigma}} = (\tilde{\sigma}_\ell, \tilde{\sigma}_r) = (0,0)$.

For the present work, we generalized this construction to obtain also wavelet collections that satisfy no boundary conditions (at primal side) at either or both boundaries, i.e., $\vec{\sigma} \in \{0,1\}^2 \setminus \{(1,1)\}$. Actually, we also slightly modified the biorthogonal collections $(\Psi_{(1,1),(0,0)}, \tilde{\Psi}_{(1,1),(0,0)})$ from [21] with the aim to minimize, for $\vec{\sigma} \in \{0,1\}^2 \setminus \{(1,1)\}$, the number of $\lambda \in \nabla_{\vec{\sigma},(0,0)}$ for which either $\psi_\lambda^{(\vec{\sigma},(0,0))} \notin \Psi_{(1,1),(0,0)}$ *or* $\tilde{\psi}_\lambda^{(\vec{\sigma},(0,0))} \notin \tilde{\Psi}_{(1,1),(0,0)}$. Indeed, recall from Remark 5.5.3 that the extension operator has to be applied to all primal wavelets with such indices λ (at either left or right boundary). We obtained the result that the number of such λ on each level at left or right boundary is equal to 2. One of them corresponds to a primal wavelet that does not vanish at the boundary and therefore has to be extended to obtain a continuous extension, whereas the primal wavelet corresponding to the other only has to be extended to guarantee locality of the resulting dual wavelets by an application of Proposition 5.5.2.

As extension operator, we apply the simple reflection suited for $\frac{1}{2} < t < \frac{3}{2}$, $0 < \tilde{t} < \frac{1}{2}$.

As domains, we consider the two-dimensional *slit domain* $\Omega = (0,2)^2 \setminus \{1\} \times [1,2)$, whose closure is the union of 4 squares $\tau + [0,1]^2$ ($\tau \in \mathbb{Z}^2$), the three-dimensional *"thick" L-shaped domain* $\Omega = (0,2)^2 \times (0,1) \setminus [1,2)^2 \times (0,1)$, whose closure is the union of 3 cubes $\tau + [0,1]^3$ ($\tau \in \mathbb{Z}^3$), and the three-dimensional *Fichera corner domain* $\Omega = (0,2)^3 \setminus [1,2)^3$, whose closure is the union of 7 cubes $\tau + [0,1]^3$ ($\tau \in \mathbb{Z}^3$). Aiming at constructing Riesz bases for $[H^{\tilde{t}}(\Omega), H_0^t(\Omega)]_{s,2}$ ($s \in [0,1]$), in particular for $H_0^1(\Omega)$, we impose homogeneous Dirichlet boundary conditions of order 1 at $\partial\Omega$.

In the slit domain case, we consider tensor product wavelet bases on $(0,1)^2$, on $\{(1,0)\} + (0,1)^2$ with no boundary conditions on its left edge, and on $\{(0,1)\} + (0,1)^2$ and $\{(1,1)\} + (0,1)^2$ with no boundary conditions on their bottom edges, all with

homogeneous Dirichlet boundary conditions of order 1 on the remaining edges. By applying the scale-dependent extension, first from $\{(1,0)\}+(0,1)^2$ to $(0,2)\times(0,1)$, and then from both top domains $\{(0,1)\}+(0,1)^2$ and $\{(1,1)\}+(0,1)^2$ over their bottom edges to Ω (see Figure 5.6), we end up with a piecewise tensor product basis.

In the thick L-shaped domain case, we consider tensor product wavelet bases on

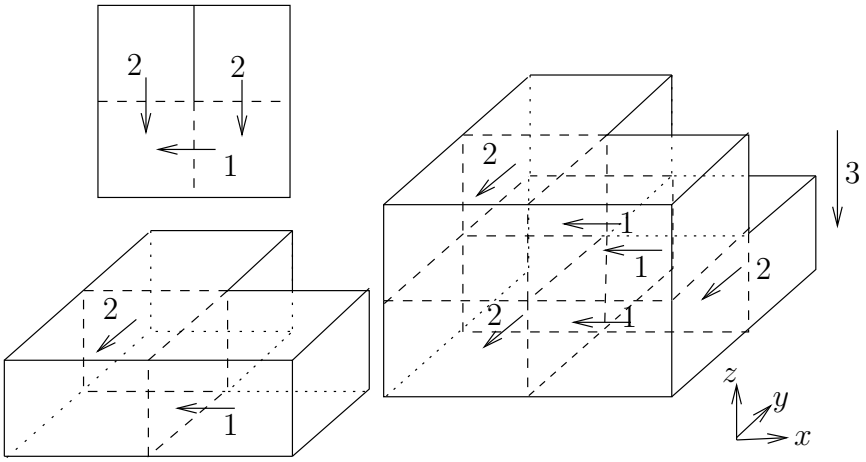


Figure 5.6: The direction and ordering of the extensions.

$(0,1)^3$, and on $\{(1,0,0)\}+(0,1)^3$ and $\{(0,1,0)\}+(0,1)^3$ with no boundary conditions on their interface with $(0,1)^3$, all with homogeneous Dirichlet boundary conditions of order 1 on the remaining faces. By applying the scale-dependent extension from $\{(1,0,0)\}+(0,1)^3$ to $(0,2)\times(0,1)^2$, and then from $\{(0,1,0)\}+(0,1)^3$ to Ω (see Figure 5.6), a piecewise tensor product basis is obtained.

In the Fichera corner domain case, we consider tensor product wavelet bases on $(0,1)^3$, on $\{(1,0,0)\}+(0,1)^3$ with no boundary conditions on its left face, on $\{(1,0,1)\}+(0,1)^3$ with no boundary conditions on its left and bottom faces, on $\{(1,1,0)\}+(0,1)^3$ with no boundary conditions on its left and front faces, on $\{(0,0,1)\}+(0,1)^3$ with no boundary conditions on its bottom face, on $\{(0,1,0)\}+(0,1)^3$ with no boundary conditions on its front face, and on $\{(0,1,1)\}+(0,1)^3$ with no boundary conditions on its front and and bottom faces, all with homogeneous Dirichlet boundary conditions of order 1 on the remaining faces. By applying the scale-dependent extensions in the order as indicated in Figure 5.6, a piecewise tensor product basis is obtained.

Using these piecewise tensor product bases, we solved the Poisson problem of finding $u\in H_0^1(\Omega)$ such that

$$\int_\Omega \nabla u\cdot\nabla v = f(v)\quad (v\in H_0^1(\Omega))$$

by applying the *adaptive wavelet-Galerkin method* ([26, 121]). This method is known to produce a sequence of approximations from the span of the basis that converges in $H^1(\Omega)$-norm with the best possible rate. Assuming a sufficiently smooth right-hand

side, Theorem 5.5.6 together with the regularity results from §5.6.1 or §5.6.2 show that this rate is $d - m = 5 - 1 = 4$ (indeed an even higher rate can generally not be expected).

Furthermore, if the bi-infinite stiffness matrix of the PDE w.r.t. the basis is sufficiently close to a sparse matrix, in the sense that it is s^*-*compressible* for some $s^* > 4$, then this adaptive method has optimal computational complexity. The univariate wavelet basis from [21] was designed such that any second order PDE on $(0,1)^n$ with homogeneous Dirichlet boundary conditions gives rise, w.r.t. the tensor product basis, to a bi-infinite stiffness matrix which is *truly sparse*. By losing the Dirichlet boundary conditions on one side of each interface between subdomains, and by the application of reflections, this sparsity, however, is partly lost in the sense that columns corresponding to wavelets that are non-zero at an interface contain infinitely many non-zero entries. The sizes of these entries, however, decay sufficiently fast as function of the difference in levels of the wavelets involved so that, nevertheless, the stiffness matrix is s^*-compressible with $s^* = \infty$, meaning that indeed the adaptive method has optimal computational complexity.

In all three examples, to avoid approximating an infinite forcing vector, for our convenience we took as right hand side function $f = 1$. As this right-hand side nowhere vanishes on the boundary, it gives rise to *all* singular terms in the solution associated to corners and edges. Our solution method does not take advantage of symmetries in the solution due to those in the right-hand side, or of other special properties of $f = 1$. As such, we expect that our results are representative for those that are obtained for any smooth right-hand side function that nowhere vanishes on the boundary.

To investigate how the application of the extensions, and the incorporation of univariate wavelet bases without boundary conditions at either or both endpoints affects the conditioning of the bi-infinite stiffness matrix, we computed numerically the *condition number* of the stiffness matrix ("preconditioned" by its diagonal) restricted to "full-grid" wavelet index sets. We considered the cases of the slit domain $(0,2)^2 \setminus \{1\} \times [1,2)$ subdivided into 4 squares, the square $(-1,1)^2$ subdivided into 4 squares, and the square $(0,1)^2$ not being subdivided. The results, given in Table 5.1, show the price to be paid for the construction of a *piecewise* tensor product basis, as well as that seemingly a re-entrant corner does not negatively affect the condition number.

Let us now first consider the Poisson problem with $f = 1$ on the two-dimensional slit domain. Its solution is illustrated in Figure 5.7.

In Figure 5.8 we give support lengths of the approximate solutions in piecewise tensor product wavelet coordinates obtained by the adaptive wavelet-Galerkin scheme vs. the (relative) ℓ_2-norm of their residual in the bi-infinite matrix vector system, the latter being equivalent to the $H^1(\Omega)$-norm of the error. The optimal rate -4 indicated by the slope of the hypotenuse of the triangle is accurately approached for the problems sizes near the end of the computation.

At the end of this computation, the cardinality of the set of adaptively selected

J	0	1	2	3	4	5	6	7
$(-1,1)^2$ into 4	790	1180	1288	1816	2335	2827	3263	3650
slit domain into 4	378	634	860	1167	1509	1882	2258	2620
J	1	2	3	4	5	6	7	8
$(0,1)^2$	37	61	96	122	146	167	185	201

Table 5.1: Condition numbers of the diagonally preconditioned stiffness matrix restricted to the square block corresponding to row and column indices $\boldsymbol{\lambda}$ with $\|\|\boldsymbol{\lambda}\|\|_\infty \leq J$. The cardinality of this set of row- or column-indices is (approximately) equal to 9.4^{J+2} (first two cases) and 9.4^{J+1} (last case), respectively.

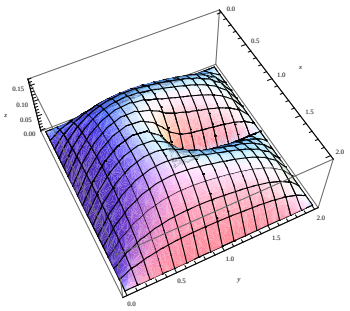


Figure 5.7: The solution of the Poisson problem with $f = 1$ on the slit domain $(0,2)^2 \setminus \{1\} \times [1,2)$.

wavelets was approximately $1.5 \cdot 10^5$. The maximum of $\|\|\boldsymbol{\lambda}\|\|_\infty$ or $\|\|\boldsymbol{\lambda}\|\|_1$ over all $\boldsymbol{\lambda}$ from this set was equal to 39 or 78, respectively, essentially meaning that locally, near the re-entrant corner the approximation space has the character of a "full-grid". The smallest non-adaptive "full-grid" or "sparse-grid" index set that contains all adaptively selected wavelets has cardinality equal to approximately $4.4 \cdot 10^{25}$ and $6.8 \cdot 10^{27}$, respectively, illustrating the strong local refinement.

Centers of supports of the piecewise tensor product wavelets that were selected by the adaptive wavelet-Galerkin scheme are indicated in Figure 5.9.

Next, we give numerical results for the Poisson problem with $f = 1$ on the thick L-shaped domain $\Omega = (0,2)^2 \times (0,1) \setminus [1,2)^2 \times (0,1)$. In Figure 5.10, we give the support lengths, in piecewise tensor product wavelet coordinates, of the approximate solutions obtained by the adaptive wavelet-Galerkin scheme vs. the (relative) ℓ_2-norm of their residual in the bi-infinite matrix vector system, the latter being equivalent to the $H^1(\Omega)$-norm of the error. The optimal rate -4 indicated by the slope of the hypotenuse of the triangle is quite accurately approached for the problems sizes near the end of the computation.

The centers of supports of the piecewise tensor product wavelets that were selected

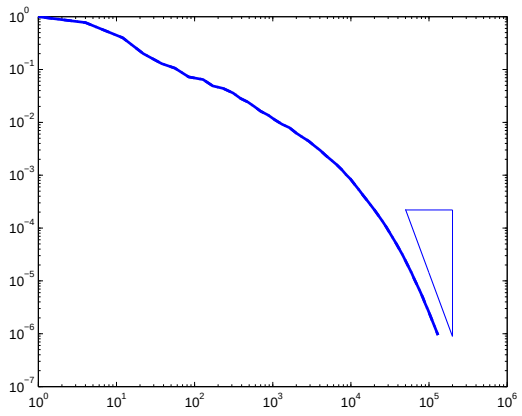


Figure 5.8: Support length vs. relative residual of the approximations produced by the adaptive wavelet-Galerkin scheme for the Poisson problem with $f = 1$ on the slit domain $(0,2)^2 \setminus \{1\} \times [1,2)$ with the piecewise tensor product basis.

by the adaptive wavelet-Galerkin scheme are illustrated in Figure 5.11.

At the end of the computation, the cardinality of the set of adaptively selected wavelets was approximately $3 \cdot 10^6$. The maximum of $\|\|\boldsymbol{\lambda}\|\|_\infty$ or $\|\|\boldsymbol{\lambda}\|\|_1$ over all $\boldsymbol{\lambda}$ from this set was equal to 46 or 92, respectively. The maximum of $\|\|\boldsymbol{\lambda}\|\|_1$ was attained for $\boldsymbol{\lambda}$ with $|\boldsymbol{\lambda}| = (46, 46, 0)$, cf. the clustering of points around $(1, 1, \frac{1}{2})$ in Figure 5.11. The smallest non-adaptive "full-grid" or "sparse-grid" index set that contains all adaptively selected wavelets has cardinality equal to approximately $2.3 \cdot 10^{44}$ and $2.8 \cdot 10^{34}$, respectively.

Finally, we give numerical results for the Poisson problem with $f = 1$ on the Fichera corner domain $\Omega = (0,2)^3 \setminus [1,2)^3$. In Figure 5.12, we give the support lengths, in piecewise tensor product wavelet coordinates, of the approximate solutions obtained by the adaptive wavelet-Galerkin scheme vs. the (relative) ℓ_2-norm of their residual in the bi-infinite matrix vector system, the latter being equivalent to the $H^1(\Omega)$-norm of the error. Due to strong singularities caused by the re-entrant corners and edges, even with a problem size at the end of our computation of approximately $2.5 \cdot 10^6$, the rate is not yet very close to the asymptotic rate -4. Nevertheless, we consider a reduction of the initial error by more than a factor 10^6 to be a convincing result for this notorious hard problem. Recall that a rate -4 in the $H^1(\Omega)$-norm with an isotropic method would require approximation of order 13, if already attainable at all in view of regularity constraints.

The centers of supports of the piecewise tensor product wavelets that were selected by the adaptive wavelet-Galerkin scheme are illustrated in Figure 5.13. The maximum of $\|\|\boldsymbol{\lambda}\|\|_\infty$ or $\|\|\boldsymbol{\lambda}\|\|_1$ over all $\boldsymbol{\lambda}$ from the set of adaptively selected wavelets at the end

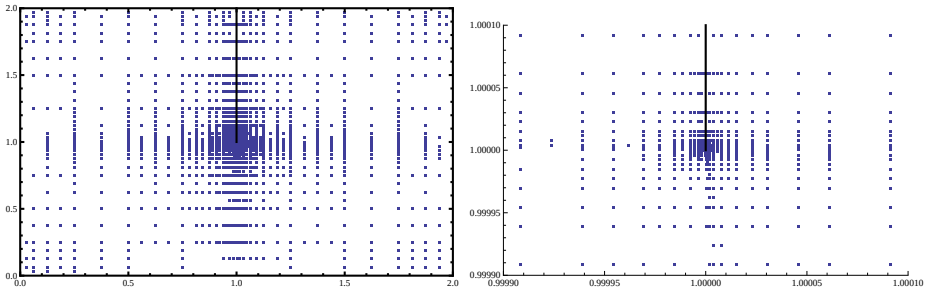


Figure 5.9: Centers of the supports of the piecewise tensor product wavelets that were selected by the adaptive wavelet-Galerkin scheme for the slit domain. The number of wavelets is here 25339. The right picture is a zoom in of the left one.

of our computation was equal to 32 or 64, respectively. The maximum of $\||\boldsymbol{\lambda}|\|_1$ was attained for $\boldsymbol{\lambda}$ with $|\boldsymbol{\lambda}|$ equal to $(32, 32, 0)$, $(32, 0, 32)$ or $(0, 32, 32)$, cf. the clustering of points around $(1, 1, 1) \pm \frac{1}{2} e_i$ $(1 \leq i \leq 3)$ in Figure 5.13. The smallest non-adaptive "full-grid" or "sparse-grid" index set that contains all adaptively selected wavelets has cardinality approximately equal to $1.2 \cdot 10^{32}$ and $3.6 \cdot 10^{25}$, respectively.

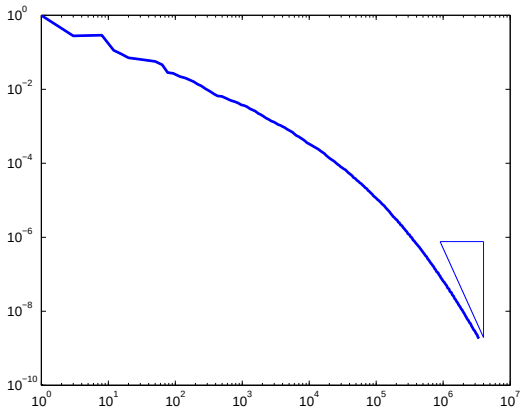


Figure 5.10: Support length vs. relative residual of the approximations produced by the adaptive wavelet-Galerkin scheme for the Poisson problem with $f = 1$ on the thick L-shaped domain $\Omega = (0,2)^2 \times (0,1) \setminus [1,2)^2 \times (0,1)$ with the piecewise tensor product basis.

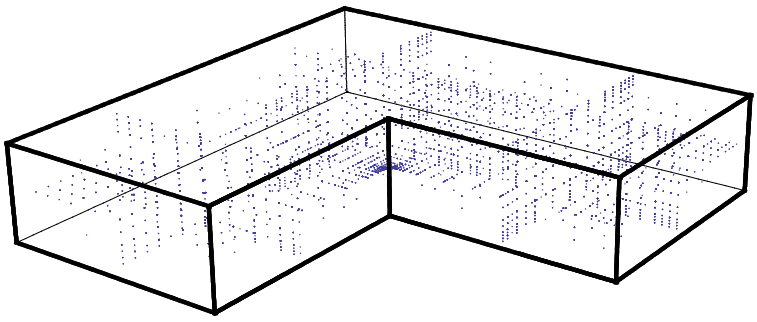


Figure 5.11: Centers of the supports of the piecewise tensor product wavelets that were selected by the adaptive wavelet-Galerkin scheme for the thick L-shaped domain. The number of wavelets is here 20421.

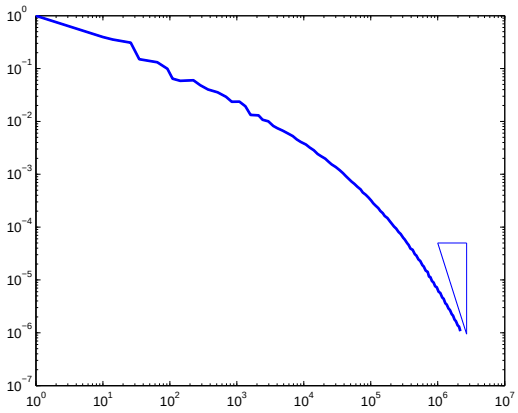


Figure 5.12: Support length vs. relative residual of the approximations produced by the adaptive wavelet-Galerkin scheme for the Poisson problem with $f = 1$ on the Fichera corner domain $\Omega = (0,2)^3 \setminus [1,2)^3$ with the piecewise tensor product basis.

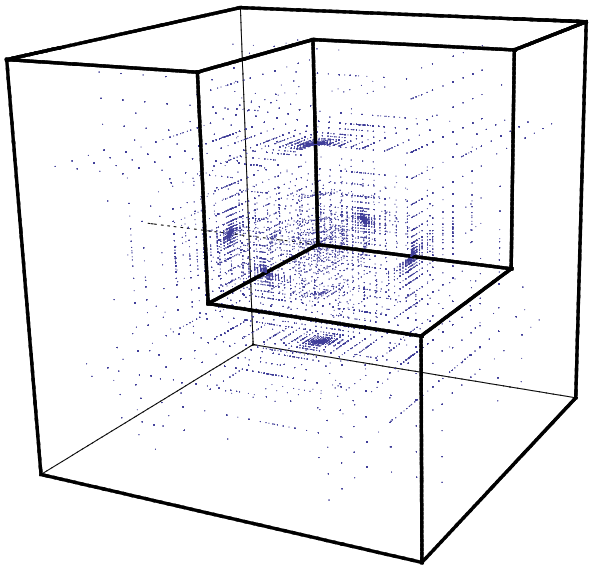


Figure 5.13: Centers of the supports of the piecewise tensor product wavelets that were selected by the adaptive wavelet-Galerkin scheme for the Fichera corner domain. The number of wavelets is here 30104.

Bibliography

[1] H. Amann, *Linear and Quasilinear Parabolic Problems. Vol. 1: Abstract Linear Theory*, Monogr. Math., vol. 189, Birkhäuser, Basel, 1995.

[2] T. Apel, *Anisotropic Finite Elements: Local Estimates and Applications. Advances in Numerical Mathematics*, Teubner, Stuttgart, 1999.

[3] J. Appell and P.P. Zabrejko, *Nonlinear Superposition Operators*, Cambridge Tracts in Math., vol. 95, Cambridge University Press, Cambridge, 1990.

[4] W. Arendt, R. Chill, S. Fornaro, and C. Poupaud, *L_p-maximal regularity for non-autonomous evolution equations*, J. Differential Equations **237** (2007), no. 1, 1–26.

[5] I. Babuška, *Advances in the p and h-p versions of the finite element method. A survey*, Numerical Mathematics Singapore 1988, Internat. Ser. Numer. Math., vol. 86, Birkhäuser, Basel, 1988, pp. 31–46.

[6] I. Babuška and W.C. Rheinboldt, *A survey of a posteriori error estimators and adaptive approaches in the finite element method*, Proceedings of the China-France Symposium on Finite Element Methods (Beijing, 1982), Sci. Press Beijing, Beijing, 1983, pp. 1–56.

[7] M. Bachmayr and W. Dahmen, *Adaptive near-optimal rank tensor approximation for high-dimensional operator equations*, Found. Comput. Math. (2014), 1–60.

[8] R.E. Bank and A. Weiser, *Some a posteriori error estimators for elliptic partial differential equations*, Math. Comp. **44** (1985), no. 170, 283–301.

[9] A. Barinka, *Fast Computation Tools for Adaptive Wavelet Schemes*, Ph.D. thesis, RWTH Aachen, 2005.

[10] S.K. Berberian, *Lectures in Functional Analysis and Operator Theory*, Grad. Texts in Math., vol. 15, Springer-Verlag, Berlin - Heidelberg - New York, 1974.

[11] J. Bergh and J. Löfström, *Interpolation Spaces. An Introduction*, Grundlehren Math. Wiss., vol. 223, Springer-Verlag, Berlin - Heidelberg - New York, 1976.

[12] P. Binev, W. Dahmen, and R.A. DeVore, *Adaptive finite element methods with convergence rates*, Numer. Math. **97** (2004), no. 2, 219–268.

[13] T. Bonesky, K. Bredies, D.A. Lorenz, and P. Maass, *A generalized conditional gradient method for nonlinear operator equations with sparsity constraints*, Inverse Problems **23** (2007), no. 5, 2041–2058.

[14] F.A. Bornemann, B. Erdmann, and R. Kornhuber, *A posteriori error estimates for elliptic problems in two and three space dimensions*, SIAM J. Numer. Anal. **33** (1996), no. 3, 1188–1204.

[15] K. Bredies, D.A. Lorenz, and P. Maass, *A generalized conditional gradient method and its connection to an iterative shrinkage method*, Comput. Optim. Appl. **42** (2009), no. 2, 173–193.

[16] H.-J. Bungartz and M. Griebel, *Sparse grids*, Acta Numer. **13** (2004), 147–269.

[17] C. Canuto, A. Tabacco, and K. Urban, *The wavelet element method. Part II. Realization and additional features in 2D and 3D*, Appl. Comput. Harmon. Anal. **8** (2000), no. 2, 123–165.

[18] A. Chambolle, R.A. DeVore, N. Lee, and B.J. Lucier, *Nonlinear wavelet image processing: Variational problems, compression, and noise removal through wavelet shrinkage*, IEEE Trans. Image Process. **7** (1998), no. 3, 319–335.

[19] N. Chegini and R.P. Stevenson, *Adaptive wavelet schemes for parabolic problems: Sparse matrices and numerical results*, SIAM J. Numer. Anal. **49** (2011), no. 1, 182–212.

[20] N.G. Chegini, S. Dahlke, U. Friedrich, and R.P. Stevenson, *Piecewise tensor product wavelet bases by extensions and approximation rates*, Math. Comp. **82** (2013), no. 284, 2157–2190.

[21] N.G. Chegini and R.P. Stevenson, *The adaptive tensor product wavelet scheme: Sparse matrices and the application to singularly perturbed problems*, IMA J. Numer. Anal. **32** (2012), no. 1, 75–104.

[22] T. Chen, H.L. He, and G.M. Church, *Modeling gene expression with differential equations*, Pac. Symp. Biocomput. **4** (1999), 29–40.

[23] Z. Ciesielski and T. Figiel, *Spline bases in classical function spaces on compact C^∞ manifolds. Part I and II*, Studia Math. **76** (1983), no. 1–2, 1–58, 95–136.

[24] A. Cohen, *Wavelet methods in numerical analysis*, Solution of Equation in $\mathbb{R}$ (Part 3), Techniques of Scientific Computing (Part 3), Handb. Numer. Anal., vol. VII, North-Holland, Amsterdam, 2000, pp. 417–711.

[25] ———, *Numerical Analysis of Wavelet Methods*, Studies in Mathematics and its Applications, vol. 32, North-Holland, Amsterdam, 2003.

[26] A. Cohen, W. Dahmen, and R.A. DeVore, *Adaptive wavelet methods for elliptic operator equations: Convergence rates*, Math. Comp. **70** (2001), no. 233, 27–75.

[27] ———, *Adaptive wavelet methods II — Beyond the elliptic case*, Found. Comput. Math. **2** (2002), no. 3, 203–245.

[28] ———, *Adaptive wavelet schemes for nonlinear variational problems*, SIAM J. Numer. Anal. **41** (2003), no. 5, 1785–1823.

[29] M. Costabel, M. Dauge, and S. Nicaise, *Analytic regularity for linear elliptic systems in polygons and polyhedra*, Tech. report, Cornell University Library, 2011.

[30] M. Crouzeix and V. Thomée, *On the discretization in time of semilinear parabolic equations with nonsmooth initial data*, Math. Comp. **49** (1987), no. 180, 359–377.

[31] S. Dahlke, *Besov regularity for elliptic boundary value problems in polygonal domains*, Appl. Math. Lett. **12** (1999), no. 6, 31–36.

[32] S. Dahlke, W. Dahmen, and R.A. DeVore, *Nonlinear approximation and adaptive techniques for solving elliptic operator equations*, Multiscale Wavelet Methods for Partial Differential Equations, Wavelet Anal. Appl., vol. 6, Academic Press, San Diego, 1997, pp. 237–283.

[33] S. Dahlke, W. Dahmen, R. Hochmuth, and R. Schneider, *Stable multiscale bases and local error estimation for elliptic problems*, Appl. Numer. Math. **23** (1997), no. 1, 21–47.

[34] S. Dahlke and R.A. DeVore, *Besov regularity for elliptic boundary value problems*, Comm. Partial Differential Equations **22** (1997), no. 1–2, 1–16.

[35] S. Dahlke, M. Fornasier, and T. Raasch, *Multilevel preconditioning for adaptive sparse optimization*, Preprint 25, DFG Priority Program 1324, 2009.

[36] ———, *Multilevel preconditioning and adaptive sparse solution of inverse problems*, Math. Comp. **81** (2012), no. 277, 419–446.

[37] S. Dahlke, M. Fornasier, T. Raasch, R.P. Stevenson, and M. Werner, *Adaptive frame methods for elliptic operator equations: the steepest descent approach*, IMA J. Numer. Anal. **27** (2007), no. 4, 717–740.

[38] S. Dahlke, E. Novak, and W. Sickel, *Optimal approximation of elliptic problems by linear and nonlinear mappings. I*, J. Complexity **22** (2006), no. 1, 29–49.

[39] S. Dahlke and W. Sickel, *On Besov regularity of solutions to nonlinear elliptic partial differential equations*, Rev. Mat. Complut. **26** (2013), no. 1, 115–145.

[40] W. Dahmen, *Wavelet and multiscale methods for operator equations*, Acta Numer. **6** (1997), 55–228.

[41] W. Dahmen, A. Kunoth, and K. Urban, *Biorthogonal spline-wavelets on the interval — stability and moment conditions*, Appl. Comput. Harmon. Anal. **6** (1999), no. 6, 132–196.

[42] W. Dahmen and R. Schneider, *Wavelets with complementary boundary conditions — function spaces on the cube*, Results Math. **34** (1998), no. 3–4, 255–293.

[43] ———, *Composite wavelet bases for operator equations*, Math. Comp. **68** (1999), no. 228, 1533–1567.

[44] ———, *Wavelets on manifolds. I. Construction and domain decomposition*, SIAM J. Math. Anal. **31** (1999), no. 1, 184–230.

[45] W. Dahmen, R. Schneider, and Y. Xu, *Nonlinear functionals of wavelet expansions — adaptive reconstruction and fast evaluation*, Numer. Math. **86** (2000), no. 1, 49–101.

[46] I. Daubechies, M. Defrise, and C. De Mol, *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*, Comm. Pure Appl. Math. **57** (2004), no. 11, 1413–1457.

[47] M. Dauge and R.P. Stevenson, *Sparse tensor product wavelet approximation of singular functions*, SIAM J. Math. Anal. **42** (2010), no. 5, 2203–2228.

[48] R.A. DeVore, *Nonlinear approximation*, Acta Numer. **7** (1998), 51–150.

[49] P. D'Haeseleer, X. Wen, S. Fuhrman, and R. Somogyi, *Linear modeling of mRNA expression levels during CNS development and injury*, Pac. Symp. Biocomput. **4** (1999), 41–52.

[50] T.J. Dijkema, *Adaptive Tensor Product Wavelet Methods for Solving PDEs*, Ph.D. thesis, Univ. Utrecht, 2009.

[51] T.J. Dijkema, C. Schwab, and R.P. Stevenson, *An adaptive wavelet method for solving high-dimensional elliptic PDEs*, Constr. Approx. **30** (2009), no. 3, 423–455.

[52] W. Dörfler, *A convergent adaptive algorithm for Poisson's equation*, SIAM J. Numer. Anal. **33** (1996), no. 3, 1106–1124.

[53] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, *Least angle regression*, Ann. Statist. **32** (2004), no. 2, 407–499.

[54] H.W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Mathematics and its Applications, vol. 375, Kluwer Academic Publishers Group, Dordrecht, 1996.

[55] K. Eriksson, *An adaptive finite element method with efficient maximum norm error control for elliptic problems*, Math. Models Methods Appl. Sci. **4** (1994), no. 3, 313–329.

[56] K. Eriksson and C. Johnson, *Adaptive finite element methods for parabolic problems. I. A linear model problem*, SIAM J. Numer. Anal. **28** (1991), no. 1, 43–77.

[57] ———, *Adaptive finite element methods for parabolic problems. II. Optimal error estimates in $L_\infty L_2$ and $L_\infty L_\infty$*, SIAM J. Numer. Anal. **32** (1995), no. 3, 706–740.

[58] K. Eriksson, C. Johnson, and S. Larsson, *Adaptive finite element methods for parabolic problems. VI. Analytic semigroups*, SIAM J. Numer. Anal. **35** (1998), no. 4, 1315–1325.

[59] L.C. Evans, *Partial Differential Equations*, 2nd ed., Grad. Stud. Math., vol. 19, American Mathematical Society, Providence, 2010.

[60] M. Figueiredo and R. Nowak, *An EM algorithm for wavelet-based image restoration*, IEEE Trans. Image Process. **12** (2003), no. 8, 906–916.

[61] Y. Fomekong-Nanfack, J.A. Kaandorp, and J. Blom, *Efficient parameter estimation for spatio-temporal models of pattern formation: Case study of Drosophila melanogaster*, Bioinformatics **23** (2007), no. 24, 3356–3363.

[62] Y. Fomekong-Nanfack, M. Postma, and J.A. Kaandorp, *Inferring Drosophila gap gene regulatory network: A parameter sensitivity and perturbation analysis*, BMC Syst. Biol. **3** (2009), no. 1, 94.

[63] M. Grasmair, M. Haltmeier, and O. Scherzer, *Sparse regularization with l^q penalty term*, Inverse Problems **24** (2008), no. 5, 055020, 13.

[64] R. Griesse and D.A. Lorenz, *A semismooth Newton method for Tikhonov functionals with sparsity constraints*, Inverse Problems **24** (2008), no. 3, 035007, 19.

[65] E. Hairer, S.P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations. I. Nonstiff Problems*, 2nd rev. ed., Springer Ser. Comput. Math., vol. 8, Springer-Verlag, Berlin, 1993.

[66] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations. II. Stiff and Differential–Algebraic Problems*, 2nd rev. ed., Springer Ser. Comput. Math., vol. 14, Springer-Verlag, Berlin, 1996.

[67] R. Haller-Dintelmann and J. Rehberg, *Maximal parabolic regularity for divergence operators including mixed boundary conditions*, J. Differential Equations **247** (2009), no. 5, 1354–1396.

[68] M. Hanke-Bourgeois, *Foundations of Numerical Mathematics and Scientific Computing*, 3rd rev. ed., Vieweg+Teubner, Wiesbaden, 2009.

[69] P. Hansbo and C. Johnson, *Adaptive finite element methods in computational mechanics*, Comput. Methods Appl. Mech. Engrg. **101** (1992), no. 1–3, 143–181.

[70] P.C. Hansen and D. O'Leary, *The use of the L-curve in the regularization of discrete ill-posed problems*, SIAM J. Sci. Comput. **14** (1993), no. 6, 1487–1503.

[71] M.R. Hestenes, *Extension of the range of a differentiable function*, Duke Math. J. **8** (1941), no. 1, 183–192.

[72] K. Ito, B. Jin, and J. Zou, *A two-stage method for inverse medium scattering*, J. Comput. Phys. **237** (2013), 211–223.

[73] K. Ito and K. Kunisch, *Semi-smooth Newton methods for state-constrained optimal control problems*, Systems Control Lett. **50** (2003), no. 3, 221–228.

[74] D. Jerison and C.E. Kenig, *The inhomogeneous Dirichlet problem in Lipschitz domains*, J. Funct. Anal. **130** (1995), no. 1, 161–219.

[75] B. Jin, T. Khan, and P. Maass, *A reconstruction algorithm for electrical impedance tomography based on sparsity regularization*, Internat. J. Numer. Methods Engrg. **89** (2012), no. 3, 337–353.

[76] B. Jin and P. Maass, *Sparsity regularization for parameter identification problems*, Inverse Problems **28** (2012), no. 12, 123001, 70.

[77] C. Johnson, *Numerical Solution of Partial Differential Equations by the Finite Element Method*, Dover Publications Inc., Mineola, 2009.

[78] B. Kaltenbacher, F. Schöpfer, and T. Schuster, *Iterative methods for nonlinear ill-posed problems in Banach spaces: Convergence and applications to parameter identification problems*, Inverse Problems **25** (2009), no. 6, 065003, 19.

[79] J. Kappei, *Adaptive frame methods for nonlinear elliptic problems*, Appl. Anal. **90** (2011), no. 8, 1323–1353.

[80] T. Kato, *Perturbation Theory for Linear Operators*, Reprint of the Corr. 2nd ed., Classics Math., Springer-Verlag, Berlin - Heidelberg - New York, 1995.

[81] A. Kirsch, *An Introduction to the Mathematical Theory of Inverse Problems*, 2nd ed., Appl. Math. Sci., vol. 120, Springer-Verlag, New York, 2011.

[82] A. Kunoth and J. Sahner, *Wavelets on manifolds: An optimized construction*, Math. Comp. **75** (2006), no. 255, 1319–1349.

[83] E.C. Lai, P. Tomancak, R.W. Williams, and G.M. Rubin, *Computational identification of Drosophila microRNA genes*, Genome Biol. **4** (2003), no. 7, R42.

[84] J. Lang, *Adaptive Multilevel Solution of Nonlinear Parabolic PDE Systems. Theory, Algorithm, and Applications*, Lect. Notes Comput. Sci. Eng., vol. 16, Springer-Verlag, Berlin, 2001.

[85] A. Langenbach, *Über L-Diffbarkeit auf metrischen Räumen und implizite Funktionen*, Math. Nachr. **166** (1994), 55–65.

[86] J.L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Grundlehren Math. Wiss., vol. 170, Springer-Verlag, Berlin - Heidelberg - New York, 1971.

[87] D.A. Lorenz, *Convergence rates and source conditions for Tikhonov regularization with sparsity constraints*, J. Inverse Ill-Posed Probl. **16** (2008), no. 5, 463–478.

[88] D.A. Lorenz, P. Maass, and P.Q. Muoi, *Gradient descent for Tikhonov functionals with sparsity constraints: Theory and numerical comparison of step size rules*, Electron. Trans. Numer. Anal. **39** (2012), 437–463.

[89] I. Loris, *On the performance of algorithms for the minimization of l_1-penalized functionals*, Inverse Problems **25** (2009), no. 3, 035008, 16.

[90] A.K. Louis, *Inverse und schlechtgestellte Probleme*, Teubner Studienbüch. Math., Teubner, Stuttgart, 1989.

[91] C. Lubich and A. Ostermann, *Linearly implicit time discretization of non-linear parabolic equations*, IMA J. Numer. Anal. **15** (1995), no. 4, 555–583.

[92] V.G. Maz'ya and J. Roßmann, *Weighted L_p estimates of solutions to boundary value problems for second order elliptic systems in polyhedral domains*, ZAMM Z. Angew. Math. Mech. **83** (2003), no. 7, 435–467.

[93] E. Mjolsness, D. Sharp, and J. Reinitz, *A connectionist model of development*, J. Theor. Biol. **152** (1991), no. 4, 429–453.

[94] P. Morin, R.H. Nochetto, and K.G. Siebert, *Convergence of adaptive finite element methods*, SIAM Rev. **44** (2002), no. 4, 631–658.

[95] Y. Nesterov, *Gradient methods for minimizing composite functions*, Math. Program. **140** (2013), no. 1, Ser. B, 125–161.

[96] P.-A. Nitsche, *Sparse approximation of singularity functions*, Constr. Approx. **21** (2005), no. 1, 63–81.

[97] ———, *Best N-term approximation spaces for tensor product wavelet bases*, Constr. Approx. **24** (2006), no. 1, 49–70.

[98] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Appl. Math. Sci., vol. 44, Springer-Verlag, New York, 1983.

[99] M. Primbs, *Stabile biorthogonale Spline-Waveletbasen auf dem Intervall*, Ph.D. thesis, Univ. Duisburg-Essen, 2006.

[100] ———, *New stable biorthogonal spline-wavelets on the interval*, Results Math. **57** (2010), no. 1–2, 121–162.

[101] T. Raasch, *Adaptive Wavelet and Frame Schemes for Elliptic and Parabolic Equations*, Berlin: Logos Verlag; Marburg: Univ. Marburg (Diss.), 2007.

[102] R. Ramlau and G. Teschke, *A Tikhonov-based projection iteration for nonlinear ill-posed problems with sparsity constraints*, Numer. Math. **104** (2006), no. 2, 177–203.

[103] R. Ramlau, G. Teschke, and M. Zhariy, *A compressive Landweber iteration for solving ill-posed inverse problems*, Inverse Problems **24** (2008), no. 6, 26.

[104] J. Reinitz and D. Sharp, *Mechanism of eve stripe formation*, Mech. Dev. **49** (1995), no. 1–2, 133–158.

[105] R.A. Ressel, *A Parameter Identification Problem Involving a Nonlinear Parabolic Differential Equation*, Ph.D. thesis, Univ. Bremen, 2012.

[106] A. Rieder, *Keine Probleme mit inversen Problemen*, Vieweg, Braunschweig, 2003.

[107] O. Scherzer, M. Grasmair, H. Grossauer, M. Haltmeier, and F. Lenzen, *Variational Methods in Imaging*, Applied Mathematics Sciences, vol. 167, Springer-Verlag, New York, 2009.

[108] T. Schuster, B. Hofmann, and B. Kaltenbacher, *Tackling inverse problems in a Banach space environment: From theory to applications*, Inverse Problems **28** (2012), no. 10, 100201.

[109] T. Schuster, B. Kaltenbacher, B. Hofmann, and K.S. Kazimierski, *Regularization Methods in Banach Spaces*, Radon Ser. Comput. Appl. Math, vol. 10, De Gruyter, Berlin, 2012.

[110] C. Schwab and R.P. Stevenson, *Adaptive wavelet algorithms for elliptic PDEs on product domains*, Math. Comp. **77** (2008), no. 261, 71–92.

[111] ———, *Space-time adaptive wavelet methods for parabolic evolution problems*, Math. Comp. **78** (2009), no. 267, 1293–1318.

[112] ———, *Fast evaluation of nonlinear functionals of tensor product wavelet expansions*, Numer. Math. **119** (2011), no. 4, 765–786.

[113] A. Shapiro, *On concepts of directional differentiability*, J. Optim. Theory Appl. **66** (1990), no. 3, 477–487.

[114] R.E. Showalter, *Monotone Operators in Banach Space and Nonlinear Partial Differential Equations*, Math. Surveys Monogr., vol. 49, American Mathematical Society, Providence, 1997.

[115] W. Sickel and T. Ullrich, *Tensor products of Sobolev-Besov spaces and applications to approximation from the hyperbolic cross*, J. Approx. Theory **161** (2009), no. 2, 748–786.

[116] J.-L. Starck, D.L. Donoho, and E.J. Candès, *Astronomical image representation by the curvelet transform*, Astron. Astrophys. **398** (2003), no. 2, 785–800.

[117] J.-L. Starck, M.K. Nguyen, and F. Murtagh, *Wavelets and curvelets for image deconvolution: A combined approach*, Signal Process. **83** (2003), no. 10, 2279–2283.

[118] R.P. Stevenson, *Adaptive solution of operator equations using wavelet frames*, SIAM J. Numer. Anal. **41** (2003), no. 3, 1074–1100.

[119] ———, *On the compressibility of operators in wavelet coordinates*, SIAM J. Math. Anal. **35** (2004), no. 5, 1110–1132.

[120] ———, *Optimality of a standard adaptive finite element method*, Found. Comput. Math. **7** (2007), no. 2, 245–269.

[121] ———, *Adaptive wavelet methods for solving operator equations: An overview*, Multiscale, Nonlinear and Adaptive Approximation. Dedicated to Wolfgang Dahmen on the Occasion of his 60th Birthday, Springer-Verlag, Berlin, 2009, pp. 543–597.

[122] K. Strehmel and R. Weiner, *Linear-implizite Runge-Kutta-Methoden und ihre Anwendung*, Teubner-Texte Math., vol. 127, Teubner, Stuttgart, 1992.

[123] G. Teschke and C. Borries, *Accelerated projected steepest descent method for nonlinear inverse problems with sparsity constraints*, Inverse Problems **26** (2010), no. 2, 23.

[124] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, 2nd ed., Springer Ser. Comput. Math., vol. 25, Springer-Verlag, Berlin, 2006.

[125] K. Urban, *Wavelet methods for elliptic partial differential equations*, Numerical Mathematics and Scientific Computation, Oxford University Press, Oxford, 2009.

[126] R. Verfürth, *A posteriori error estimation and adaptive mesh-refinement techniques*, J. Comput. Appl. Math. **50** (1994), no. 1-3, 67–83.

[127] ———, *A Review of a Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Chichester: Wiley; Stuttgart: Teubner, 1996.

[128] J.G. Verwer, E.J. Spee, J.G. Blom, and W. Hundsdorfer, *A second-order Rosenbrock method applied to photochemical dispersion problems*, SIAM J. Sci. Comput. **20** (1999), no. 4, 1456–1480.

[129] D.C. Weaver, C.T. Workman, and G.D. Stormo, *Modeling regulatory networks with weight matrices*, Pac. Symp. Biocomput. **4** (1999), 112–123.

[130] M. Werner, *Adaptive wavelet frame domain decomposition methods for elliptic operator equations*, Ph.D. thesis, Univ. Marburg, 2009.

[131] C. Zenger, *Sparse grids*, Parallel Algorithms for Partial Differential Equations. Proceedings of the Sixth GAMM-Seminar, Kiel, Germany, January 19–21, 1990, Notes Numer. Fluid Mech., vol. 31, Vieweg, Braunschweig, 1991, pp. 241–251.

Erklärung

Ich erkläre, dass ich in der Vergangenheit keinen anderen Promotionsversuch unternommen habe.

Ich versichere, dass ich meine Dissertation "Adaptive Wavelet Methods for Inverse Problems: Acceleration Strategies, Adaptive Rothe Method and Generalized Tensor Wavelets" selbstständig, ohne unerlaubte Hilfe angefertigt, keine anderen als der von mir ausdrücklich bezeichneten Quellen und Hilfen benutzt und alle vollständig oder sinngemäß übernommenen Zitate als solche gekennzeichnet habe.

Die Dissertation wurde in der jetzigen oder einer ähnlichen Form noch bei keiner anderen in- oder ausländischen Hochschule anlässlich eines Promotionsgesuchs oder zu anderen Prüfungszwecken eingereicht.

Declaration

I declare that I did not conduct any previous attempt to earn a doctor's degree.

I assure that I wrote my doctoral thesis "Adaptive Wavelet Methods for Inverse Problems: Acceleration Strategies, Adaptive Rothe Method and Generalized Tensor Wavelets" independently and without unauthorized help. I used no other sources or aid than the ones explicitly stated. All the content from literature or other sources that I cited, either completely or in altered form, is clearly marked as such.

This thesis in its current or a similar form has not been used previously to apply for a doctoral degree or to be used for other exams at a university in Germany or any other country.

Marburg, 18. Dezember 2014

Ulrich Friedrich