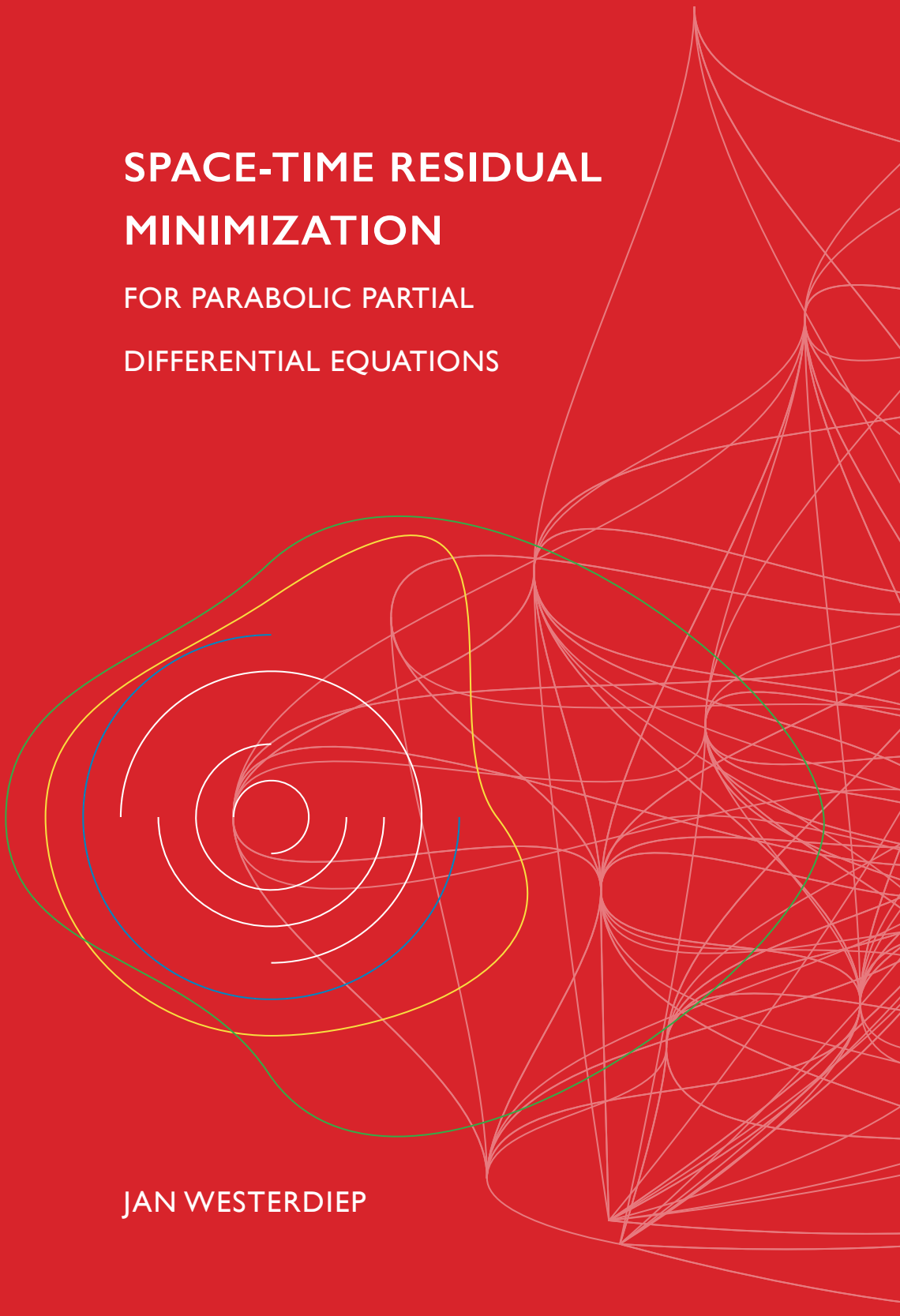
Space-time residual minimization for parabolic partial differential equations

Westerdiep, J.H.

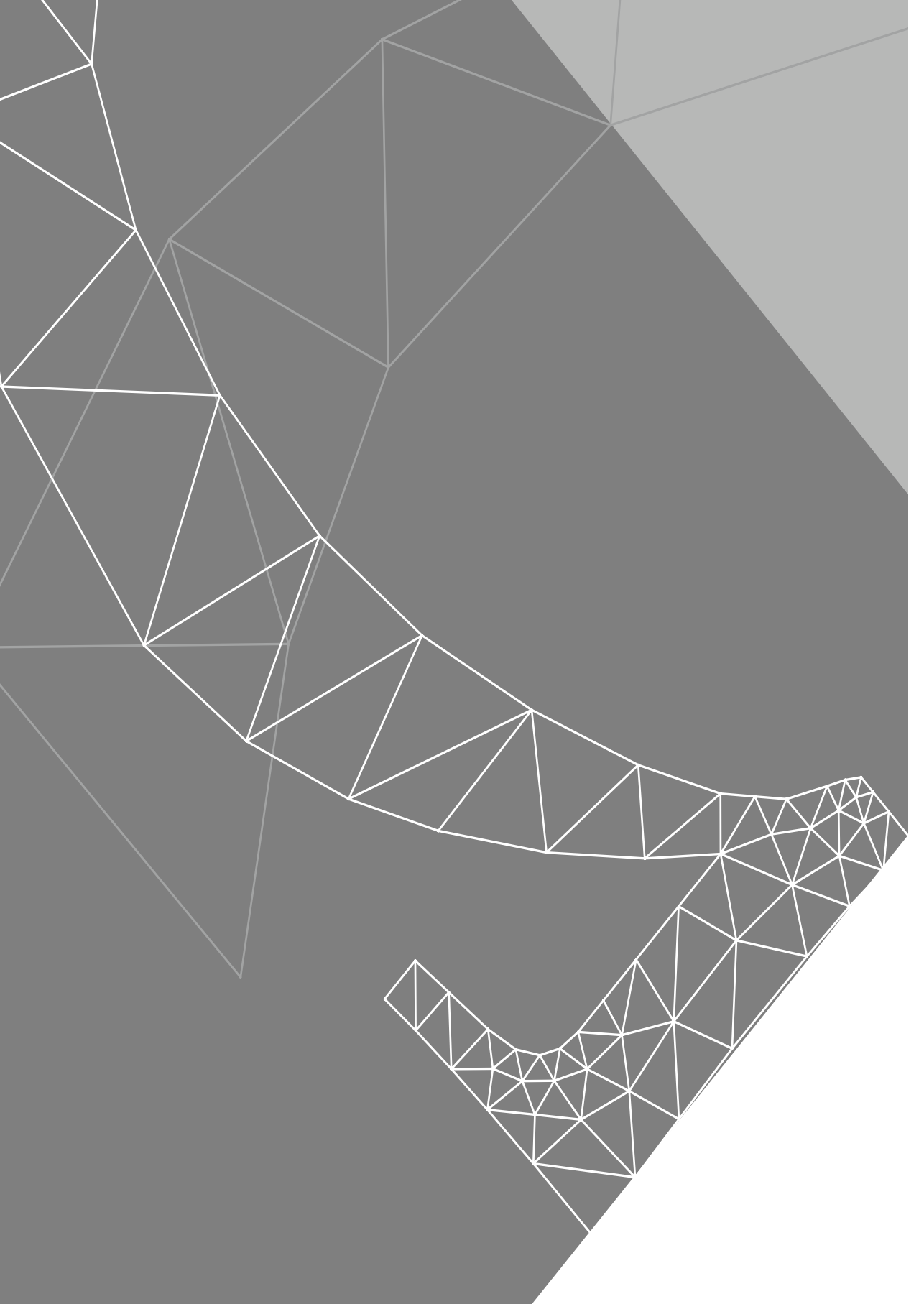
Publication date
2021
Document Version
Final published version

SPACE-TIME RESIDUAL MINIMIZATION

FOR PARABOLIC PARTIAL DIFFERENTIAL EQUATIONS



JAN WESTERDIEP



Space-time residual minimization
for parabolic partial differential equations

Jan Westerdiep

Cover design and chapter illustrations: Fenna Westerdiep
Layout: Fenna Westerdiep, Jan Westerdiep
Typesetting: LaTeX using the `memoir` class
Paper: CyclusPrint 90g/m^2 (made from 100% recycled pulp)
Reproduction: GVO Printers & Designers, Ede

This thesis is produced on FSC®-certified materials.

The research for this doctoral thesis received financial assistance from the Netherlands Organization for Scientific Research (NWO) under contract no. 613.001.652. The work was carried out at the Korteweg–de Vries Institute for Mathematics (KdVI) at the University of Amsterdam.

ISBN 978-94-6332-779-4

Space-time residual minimization
for parabolic partial differential equations

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op donderdag 23 september 2021, te 16.00 uur

door Jan Hindrik Westerdiep
geboren te Amsterdam

Promotiecommissie

Promotor:	prof. dr. R.P. Stevenson	Universiteit van Amsterdam
Copromotor:	dr. C.C. Stolk	Universiteit van Amsterdam
Overige leden:	prof. dr. J.J.O.O. Wiegerinck	Universiteit van Amsterdam
	dr. J.H. Brandts	Universiteit van Amsterdam
	dr. G. Gantner	Universiteit van Amsterdam
	prof. dr. W.A. Dahmen	University of South Carolina
	prof. dr. A.A. Reusken	RWTH Aachen University
	prof. dr. D.T. Crommelin	Universiteit van Amsterdam

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

Preface

September 2021 marks my ten-year anniversary at Science Park. My first encounter with my future advisor Rob Stevenson was during my second-year project, and somehow, I've been his student ever since. It feels weird that our official collaboration is now coming to a close. Rob, your ever-critical view distraught me at times, but it did push me to never settle for mediocrity. Your ability to debug code by just looking at its output still baffles me. Thank you for allowing me to pursue a Ph.D. under your wing, and for your continuous support and supervision this past decade. Thanks to my copromotor Chris Stolk for being there when I had questions, and the members of my committee for reading my thesis—I look forward to discussing its results with you.

I spent the spring of 2018 in the London office of Google, working on a research project in the field of reinforcement learning. Claudio, thanks for taking me in as an intern, even though I had no formal education on the topic; I had an absolute blast. Good luck in Amsterdam, I know we'll keep in touch.

It was a strange feeling when in March 2020, we suddenly had to do our work from home. Still, aside from teaching over Zoom, I have very fond memories of this period. Carmen and I had just moved in together and were spending practically every moment of every day together with our two cats. At some point during a lunch break, we realised our Ph.D. projects had much more in common than we previously thought. With lots of uncertainty on the funding of her final year, we set out to apply for a grant together. We never got the money, but perhaps the real treasure was the process itself. My gratitude goes out to everyone that helped us during this rollercoaster.

Zoë, David, Pjotr, Raymond, Gregor, Lenny, Sonja, Anita, Tiny, and the other colleagues: thank you for *overzomeren in Coronatijd*. Vera, it was fun supervising your thesis—good luck with the rest of your studies. Marieke: thank you for the homemade jams and inkle weaves, our many discussions, and for running the institute. Evelien: thank you for everything you do.

After a good four years at KdVI, and some six before that at Science Park in general, it is time to leave the place I now call home. Although I've been away for a couple of months here and there, this is the first time I'll leave without a definitive plan on *what's next*. I look forward to being a little bored.

Raymond, sinds wij elkaar leerden kennen, hebben we praktisch elk studieproject samen gedaan, en ik vond het geweldig. Je bent simpelweg een programmeur van wereldklasse. Je haalt m'n meest competitieve zelf naar boven,

maar bent natuurlijk ook een enorm goede vriend, en ik kijk met plezier terug op ons avontuur in Amerika. Bedankt dat je m'n paranimf wilde zijn, en ook bedankt voor alle gekke programmeerwedstrijdjes met Mees en Ruben; hoe leuk dat *unsigned long long* nu vereeuwigd is in ons papertje [BVVW20]!

Gregor, sinds jouw komst op het instituut voelt onze groep als een groep, en inmiddels lunchen we zelfs samen wanneer je er niet bij bent. Ik leerde je kennen als collega, maar inmiddels zijn Olivia en jij zo veel meer. Dank jullie.

Proeflezers Gregor, Eric, Thomas, Ray, Okke, en Carmen: enorm bedankt.

Mees: elke keer dat ik iets van je lees, leer ik weer een nieuw Engels woord. Ik ben trots dat *Training BAPC* blijft bestaan, zelfs nadat we alle drie weg zijn.

Eric and Rhiannon, thank you in advance for the many boat rides to come.

Daan, Djera, Tom, Michael, Okke, Pom, Nina, Erik, Floris, Menno, Deniz: *friendship is making memories together*, and boy, do we have some. Zonder jullie had mijn afgelopen decennium er flink anders uitgezien. Dank jullie voor alle geweldige, vreemde, zwierige, zangerige, zweterige, maar vooral mooie herinneringen tot nu toe, en alvast bedankt voor alle die we nog gaan maken.

Lieve Raoul, Nikki, Willem, en alle andere Unityers: jullie maken m'n weekenden (en de tweede dinsdag van de maand) spannend. Dank voor alles.

Mark and Jenny, thank you for taking a gamble and welcoming me into your home. The California adventure grew me immensely as a person. I look back on an amazing time, all the more thanks to you as my 'overseas family'.

Lieve Peter en Tineke, Thomas en Anne, Lennart en Tinke: dankzij jullie heb ik er een familie bij, en leer ik dat het ook gezellig kan zijn *buiten* de ring!

Awi, er was geen twijfel mogelijk over wie ik zou vragen m'n paranimf te zijn. We kennen elkaar al bijna twintig jaar en zijn praktisch samen opgegroeid. We zijn hartstikke verschillend maar juist daardoor leer ik weer iets elke keer dat we elkaar zien. Ik kijk enorm naar je op, en zal nooit *niet* lachen om je ad rem woordgrappen. Je bent geen vriend, je bent een lotgenoot. <4lyff

Lieve papa. Je woont dan misschien ver weg in Gobabis, maar gelukkig kunnen we elkaar altijd spreken. Het betekent enorm veel voor me dat je er bij bent op 23 september. Lieve Jona en Tessel: we verschillen veel van elkaar, maar toch is het altijd fijn met jullie. In Marcus' woorden—Marcus, Ziggy, Stefan, Lasse: jullie zijn *bloedtoppers*. Lieve Dior, ik ben heel erg benieuwd wat voor mens je gaat worden en kijk nu al uit naar ons gedeelde verhaal.

Lieve mama, dank je voor onvermoeibare hulp. Je bent er altijd voor me. Zelfs nu het niet goed gaat met Flow, heb je enorm veel energie gestoken in de opmaak van dit proefschrift. We mogen trots zijn op het resultaat. Ik zeg het niet genoeg, dus dan maar voor altijd zwart op wit: ik hou van je.

The ideal of calm exists in a sitting cat. Doron en Jonah, dank jullie voor deze kalmte. Eli, bedankt voor het vertrouwen; er wordt van je jongens gehouden.

Liefste Carmen. Het afgelopen jaar was een geweldige rollercoaster samen, maar we hebben het (bijna) overleefd! Bedankt voor de steun op momenten dat ik het niet meer zag zitten, en bedankt voor de geweldige momenten tot nu toe. Samen met Doron en Jonah laat je me zien hoe een huis een *thuis* wordt. Ik ben heel benieuwd wat voor avonturen we samen nog gaan beleven. Op dat we nog vele vuurtjes mogen bouwen, thuistheater mogen spelen, en samen mogen verdwalen in kronkelende dorpsstraten. Je bent mijn favoriete mens.

Contents

1 Introduction

Many processes in nature and engineering are governed by partial differential equations (PDEs). We focus on *parabolic* PDEs, that describe time-dependent phenomena like heat conduction, chemical concentration, and fluid flow.

Parabolic evolution equations describe how a function evolves from a given initial state as governed by the PDE. On a time interval $I := [0, T]$ and spatial domain $\Omega \subset \mathbb{R}^d$, given a linear elliptic spatial partial differential operator A, an *initial state* $u_0 : \Omega \to \mathbb{R}$ and *source term* $g : I \times \Omega \to \mathbb{R}$, we aim to find $u : I \times \Omega \to \mathbb{R}$, subject to appropriate boundary conditions, that solves

$$\begin{cases} \frac{\partial}{\partial t} u + Au = g & (\text{on } I \times \Omega), \\ u = u_0 & (\text{on } \{0\} \times \Omega). \end{cases} \tag{1.1}$$

Even if we know that a unique solution u *exists*, we can express it in closed form only under very strict circumstances[1]; see also [Eva10, §7]. To understand what u looks like, we turn to *numerical approximation*. Similar to the idea of using many tiny line segments to approximate a circle, numerical methods for PDEs couple many simple equations to approximate a complex equation.

Historically, parabolic evolution equations are solved using *time-stepping*; see [Tho06]. One first discretizes the equation in space as to obtain a system of coupled *ordinary* differential equations in time. This system is then solved using the vast theory for ODEs. While efficient in terms of memory and computational cost, time-stepping schemes take *global* time steps, which are independent of spatial position. As a result, these methods cannot efficiently resolve details of u in localized regions of space and time. Moreover, being inherently sequential, they have limited possibilities for parallel computation.

In this thesis, we take a different approach and reformulate the parabolic evolution equation as a linear operator equation posed in space and time simultaneously. These *space-time methods* are able to increase the resolution of discretizations locally, parallelize gracefully, and moreover produce approximations to the unknown solution that are *uniformly quasi-optimal*. Returning to our circle analogy: out of all n-sided polygons inscribed in a circle, the regular polygon is its uniformly optimal approximation, for any choice of n.

[1] One is reminded of the following metaphore. *Milk production at a dairy farm was low, so the farmer wrote to the local university asking for help. A multidisciplinary team of professors was assembled, headed by a theoretical physicist, and weeks of intensive on-site investigation took place. The scholars returned to the university, notebooks crammed with data. Shortly thereafter, the report came, stating that "**we found a solution, but it works only for spherical cows in a vacuum**."*

1.1 Residual minimization

Following the pioneering works [BJ89, BJ90], recent years have seen a rapidly growing interest in space-time methods for parabolic evolution equations. By starting from a well-posed variational formulation of (1.1), one aims to transfer fundamental properties enjoyed by numerical methods for *elliptic* PDEs (like error control, a posteriori error analysis, adaptive discretization, efficient solution) to the parabolic case. See [SY19] for a review of the recent literature.

The focal point of this thesis is the *space-time minimal residual (MR) method* introduced by R. Andreev [And12]. Informally, with δ some discretization parameter and (X^δ, Y^δ) its discrete *trial- and test spaces*, the MR method finds

$$u^\delta_{\mathrm{MR}} := \operatorname*{arg\,min}_{w^\delta \in X^\delta} \Big[\underbrace{\| \tfrac{\partial}{\partial t} w^\delta + A w^\delta - g \|^2_{Y^{\delta\prime}}}_{\text{PDE residual}} + \underbrace{\| w^\delta|_{t=0} - u_0 \|^2}_{\text{initial residual}} \Big].$$

Considering a *family* $(X^\delta, Y^\delta)_{\delta\in\Delta}$ of discrete trial- and test spaces, we would like to have guarantees on the *error* $\|u - u^\delta_{\mathrm{MR}}\|$ of MR solutions. An appealing property is *uniform quasi-optimality*, meaning that MR solutions are, up to some constant C, the best-possible approximation to u from X^δ:

$$\|u - u^\delta_{\mathrm{MR}}\| \le C \min_{w\in X^\delta} \|u - w^\delta\| \quad \text{uniformly in} \quad \delta \in \Delta.$$

This raises two important questions.
(**Q**1) What are the conditions on $(X^\delta, Y^\delta)_{\delta\in\Delta}$ for *uniform quasi-optimality*?
(**Q**2) Can we estimate the *error* $\|u - u^\delta_{\mathrm{MR}}\|$ robustly, even if u is unknown?

As we will see, the conditions of (**Q**1) are met when Y^δ is *stable*, i.e., large enough relative to X^δ. At the same time, for the sake of efficiency, we want the dimension of Y^δ to be as small as possible. With this, our goal becomes finding X^δ that allows an accurate approximation of u, and a corresponding stable Y^δ that is also *efficient*, in that $\dim Y^\delta / \dim X^\delta$ is bounded uniformly in $\delta \in \Delta$.

The highest approximation power is held by trial spaces that *adapt* to the solution at hand, with increased resolution in those regions of the space-time cylinder where u has local details. Moreover, in view of a practical algorithm, we want to compute (good approximations to) the MR solutions efficiently. This raises two further questions.
(**Q**3) Is there a stable and efficient family $(X^\delta, Y^\delta)_{\delta\in\Delta}$ that allows *adaptivity*?
(**Q**4) How do we construct quasi-optimal solutions as efficiently as possible?

In areas like weather prediction and measuring blood flow in humans, we are supplied with vast amounts of data as (noisy) measurements of some unknown function u that abides by a physical law. The process of recovering u by fusing observations with an underlying mathematical model is known as *data assimilation*. It raises the two final questions.
(**Q**5) How do our findings on residual minimization for parabolic evolution equations extend to *parabolic data assimilation problems*?
(**Q**6) Are there other problems governed by parabolic PDEs that can be understood using residual minimization?

1.2 Outline and contributions

This thesis is based on the following works.

[SW21a] R. Stevenson and J. Westerdiep
Stability of Galerkin discretizations of a mixed space-time variational formulation of parabolic evolution equations
IMA Journal of Numerical Analysis, 41(1):28–47, 2021.

[SW21b] R. Stevenson and J. Westerdiep
Minimal residual space-time discretizations of parabolic equations: Asymmetric spatial operators
Submitted to *Computers & Mathematics with Applications*, arXiv:2106.01090, 2021.

[vVW21a] R. van Venetië and J. Westerdiep
Efficient space-time adaptivity for parabolic evolution equations using wavelets in time and finite elements in space
Submitted to *Numerical Linear Algebra with Applications*, arXiv:2104.08143, 2021.

[vVW20a] R. van Venetië and J. Westerdiep
A parallel algorithm for solving linear parabolic evolution equations
To appear in *Parallel-in-Time Integration Methods*, arXiv:2009.08875, 2020.

[DSW21] W. Dahmen, R. Stevenson, and J. Westerdiep
Accuracy controlled data assimilation for parabolic problems
To appear in *Mathematics of Computation*, arXiv:2105.05836, 2021.

[Wes20] J. Westerdiep
On p-Robust Saturation on Quadrangulations
Computational Methods in Applied Mathematics, 20(1):169–186, 2020.

Each chapter is essentially one of these papers, is self-contained, and can be read independently. Notation is roughly consistent between chapters. On average, the authors contributed equally to all works. We give a short outline.

Chapter 2: introduction on residual minimization

In this chapter, we discuss the preliminaries from an abstract point of view. Starting from some linear operator equation posed on Hilbert spaces, we introduce equivalent conditions for *well-posedness* of the problem, discuss its *minimal residual approximation* and derive conditions for *uniform quasi-optimality* (**Q**1). We will find the focal point to be *uniform inf-sup stability* of trial- and test spaces, and shed some light on estimating the error (**Q**2).

We then discuss solving the arising linear systems efficiently (**Q**4), and finally apply our ideas to the linear parabolic evolution equation (1.1).

Chapter 3: symmetric spatial operators and stable discretizations [SW21a]

As our first exploration into the minimal residual method for parabolic problems, we restrict ourselves to *linear parabolic evolution equations with symmetric spatial operators*, so we consider (1.1) where A is self-adjoint.

We formulate sufficient conditions for uniform inf-sup stability of trial- and test spaces, and with it, uniform quasi-optimality of MR solutions (**Q**1).

Moreover, we show that these conditions are satisfied when discretizing the space-time cylinder into *time slabs*, where the spatial discretization can

be adapted on each slab (**Q**3). While this family of discretizations allows for refinements localized in space *or* time, it can't refine locally in space *and* time. In [SvVW21], we find a family that *can*, constructed as certain spans of *wavelets in time* tensorized with *(locally refined) finite element functions in space*.

Chapter 4: asymmetric spatial operators and error estimation [SW21b]

Next, we take on evolution equations where the spatial partial differential operator A is not necessarily symmetric. We find that for these equations, MR solutions are also quasi-optimal (**Q**1) but that the optimality constant degrades for an increasing relative size of the antisymmetric part $\frac{1}{2}(A - A')$ of A.

On the other hand, in the *energy norm* induced by the parabolic operator, this optimality constant is robust not only in the discretization parameter, but also the relative size of $\frac{1}{2}(A - A')$. We derive sufficient conditions for robust error bounds (**Q**2), and find the *Fortin interpolant* to be the leading actor.

We apply this theory to *convection-diffusion-reaction equations*, with A as

$$A = \underbrace{-\varepsilon\Delta_{\boldsymbol{x}}}_{\text{diffusion}} + \underbrace{\boldsymbol{b}\cdot\nabla_{\boldsymbol{x}}}_{\text{convection}} + \underbrace{c.}_{\text{reaction}}$$

We think of the convection- and reaction-terms as fixed, and of the *diffusion* $\varepsilon > 0$ as a problem parameter. When convection dominates diffusion, so for ε small, the solution has boundary- and interior layers that we need to resolve.

Chapter 5: adaptive refinement in linear complexity [vVW21a]

In the symmetric setting, we consider the algorithm from [SvVW21] that aims for *optimal convergence* using adaptive refinement locally in space and time (**Q**3). There, we construct our discretizations as the span of wavelets in time tensorized with finite element spaces in space. The theory is elegant, but its implementation requires diligence: the resulting system matrix is *dense*, so naive matrix-vector products have quadratic complexity.

In this chapter, we discuss its implementation at *optimal linear cost* (**Q**4). We overcome the problem of matrix-vector multiplication, and use a matrix-free iterative solver to produce approximate MR solutions $\hat{u}^\delta_{\mathrm{MR}}$ efficiently. We highlight the interplay between *algebraic error* $\|u^\delta_{\mathrm{MR}} - \hat{u}^\delta_{\mathrm{MR}}\|$ (**Q**4) and *discretization error* $\|u - u^\delta_{\mathrm{MR}}\|$ (**Q**2) to achieve quasi-optimality, and see that these considerations translate to a highly efficient method in practice.

Chapter 6: parallel complexity and parallel implementation [vVW20a]

Next, we explore the minimal residual method in parallel computation (**Q**4). Define *parallel complexity* as the asymptotic runtime of an algorithm given *sufficiently many* parallel processors with access to a shared memory on which communication is free. We show that on tensor-product discretizations, the proposed algorithm runs in polylogarithmic parallel complexity. This is on par with the best-known algorithms for *elliptic* problems.

This parallel complexity translates to a highly scalable algorithm in practice, and we produce a quasi-optimal solution for the *heat equation* with over 4 billion unknowns using over 2 thousand parallel cores in under 2 minutes.

Although stated for *parabolic evolution equations with symmetric spatial operators*, the results extend naturally to the setting of Chapters 4 and 7.

Chapter 7: accuracy controlled data assimilation [DSW21]

Abandoning our utopia of *well-posed* problems, we now turn to the ill-posed parabolic problem where initial data is missing, but we have observational data instead (**Q**5). Suppose we are given a source term $g : I \times \Omega \to \mathbb{R}$ and observational data $f : I \times \omega \to \mathbb{R}$ on a subdomain $\omega \subsetneq \Omega$ (possibly much smaller than Ω). The *data assimilation* problem is to recover $u : I \times \Omega \to \mathbb{R}$ s.t.

$$\begin{cases} \frac{\partial}{\partial t} u + Au = g & (\text{on } I \times \Omega), \\ u = f & (\text{on } I \times \omega). \end{cases} \tag{1.2}$$

The catch is that (g, f) need not be *consistent*, in that (1.2) may not be solvable, and we instead aim for some u that fits both data and PDE *as closely as possible*.

Based on our knowledge of well-posed operator equations and residual minimization, we introduce a regularization term to arrive at a a *regularized* minimal residual problem, and study the properties of its discretizations.

A crucial tool for deriving tight error bounds (**Q**2) is the *Carleman estimate*. It allows bounding the norm of a function in our *trial space* away from zero by that of its image under the problem operator. More explicitly, say $A = -\Delta_x$ is the negative Laplacian and u is subject to $u|_{\partial\Omega} = 0$. Take trial- and test spaces

$$X := L_2(I; H_0^1(\Omega)) \cap H^1(I; H^{-1}(\Omega)), \quad Y := L_2(I; H_0^1(\Omega)).$$

Given $\eta \in (0, T)$, define $X_\eta := L_2([\eta, T]; H_0^1(\Omega)) \cap H^1([\eta, T]; H^{-1}(\Omega))$. Then the Carleman estimate states that there is a constant $C_{\eta,\omega}$ for which

$$\|w\|_{X_\eta} \leq C_{\eta,\omega} \Big(\|\tfrac{\partial}{\partial t} w + Aw\|_{Y'} + \|w\|_{L_2(I \times \omega)} \Big) \quad (w \in X).$$

We formulate a highly efficient practical algorithm, and finish this chapter with extensive numerical experiments to showcase its efficacy.

Chapter 8: p-robust saturation for elliptic problems [Wes20]

In this chapter, we change gears completely and look at a topic related to (**Q**2)–(**Q**3), but in the setting of hp-adaptivity for *elliptic boundary value problems*.

Classical (h-)adaptive finite element methods for these problems work as follows. Starting from some initial partition of the problem domain into, say, triangles, we find the best approximation of the solution in a space of globally continuous elementwise polynomials of fixed degree. We estimate the error in this best approximation locally on every element, and refine only those elements that carry large error. Iterating this process guarantees error convergence at the best *algebraic* rate possible: with N the number of degrees of freedom, the error decays as N^{-s} for the best possible $s > 0$; cf. [Ste07].

Allowing the polynomial degree to vary between elements, it is possible to achieve *exponential* error decay, so proportional to $\exp(-\alpha N^\beta)$ for some $\alpha, \beta > 0$. In [CNSV17a], Canuto, Nochetto, Stevenson, and Verani were the first to prove *instance-optimality* of an hp-adaptive loop running in polynomial time by alternating h-adaptive refinement with hp-adaptive *coarsening*. The former increases the accuracy of the approximation, and the latter increases its efficiency by removing (nearly) redundant degrees of freedom, at the expense of some accuracy. This produces a saw-tooth graph of the error, and the sequence of solutions after each coarsening step converges exponentially.

While the hp-coarsening step of [Bin18] is a highly interesting result—in fact, it featured heavily in both my bachelor's and master's theses—in this chapter, we focus on the refinement step instead. In [CNSV17a], refinement was driven by an error estimator with bounds sensitive to the polynomial degree, resulting in a potentially exponential runtime. Instead, in [CNSV17b], refinement is driven by a p-robust estimator. The central question to refinement in polynomial complexity is *which increase in local polynomial degrees ensures a reduction of the error in energy norm*. We discuss this problem in the setting of quadrilateral partitions that allow one hanging node per edge.

1.3 Outlook

In [SvVW21], we show uniform stability (**Q**3) for a particular family of spans of wavelets in time tensorized with (locally refined) finite elements in space. In Chapter 5, we see this can yield a very efficient algorithm, but also requires careful implementation. It would be interesting to look at function spaces over locally refined (prismatic) partitions of the space-time cylinder directly, so without wavelets in the time-axis, and see if uniform stability holds here too.

We finish this chapter with an outlook of other parabolic problems that could be solved numerically through space-time residual minimization (**Q**6).

Semilinear evolution equation

A natural next step is the *semi*linear problem of finding u that solves

$$\begin{cases} \frac{\partial}{\partial t} u + Au + N(u) = g & (\text{on } I \times \Omega), \\ u = u_0 & (\text{on } \{0\} \times \omega), \end{cases}$$

where N is some nonlinear operator, typically involving no derivatives of u.

Exciting applications are *reaction-diffusion systems* like the Gray–Scott model of [GS84, Pea93] for autocatalytic reactions, or the Allen–Cahn equation [AC75] for phase separation in metal alloys. These exhibit the *Turing patterns* that are postulated to underlie the formation of patterns in animal fur and desert sand; see also the pioneering work [Tur52] by Turing.[2] Another application lies in

[2] Interestingly, Turing seems to have known about the *spherical cows* metaphore of footnote 1 as well, as [Tur52, p.41] reads: *"a system which has spherical symmetry, and whose state is changing because of chemical reactions and diffusion, will remain spherically symmetrical for ever. (The same would hold true if the state were changing according to the laws of electricity and magnetism, or of quantum mechanics.) It certainly cannot result in an organism such as a horse, which is not spherically symmetrical."*

the *diffusion-drift model*, used in semiconduction [MRS90] and in neutron- or electron transport for nuclear fusion; see [DL00, Ch. XXI.§5] and [Bla19].

First steps for the analysis were taken in [And12, §3.4, §4.5] and [Ste14]: As it turns out, for $\|(g, u_0)\|$ small enough, this problem behaves essentially like a linear problem, and admits a unique solution we can find using fixed-point iteration. It would be interesting to continue the analysis, and relate the ideas of [GHPS18, HPSV21] for elliptic problems to parabolic evolution equations.

Goal-oriented adaptivity

In *goal-oriented* problems, we are interested in some *quantity of interest* $J(u)$ rather than the function u itself. Problems like these are prevalent in engineering, where one might be interested in, say, the drag of an airplane, or blood flow dynamics, where estimates of shear stress on vessel walls are used in clinical decision-making for patients with cardiovascular disease [BFB+18].

MR solutions u^δ_{MR} yield good approximations to $J(u)$, as for $J \in X'$,

$$|J(u^\delta_{\mathrm{MR}}) - J(u)| \leq \|J\|_{X'} \|u^\delta_{\mathrm{MR}} - u\|_X.$$

In *goal-oriented adaptivity*, we can essentially double the rate of convergence by 'cutting out the middleman' and running an adaptive loop on $J(u)$ directly. It would be interesting to relate the ideas of [MS09, BGIP21] for elliptic problems, and [DC07, MMCP+19] for parabolic problems, to our simultaneous solution. Our approach is especially welcoming to so-called *time-domain goals* that require the entire time evolution, as it simplifies the treatment of the *adjoint problem*, and in any case allows higher approximation rates as we are not restricted to global time steps.

Optimal control constrained by a parabolic evolution equation

Another type of problem that requires the full time evolution is optimal control constrained by a parabolic PDE; see [Lio71, §III] for an introduction on the theory. The prototypical example is placing heaters in a room as to reach a comfortable temperature everywhere, but many more applications exist.

Using ideas from [BG09, §11] and [GK11, BRU20, LSTY21], it would be interesting to try adapting our space-time method to this problem.

Data assimilation for human blood flow

A powerful tool in the clinical evaluation of cardiovascular disease is non-invasive blood flow quantification. Existing methods that can capture blood flow patterns in three-dimensional space and time either rely on direct measurements (4D flow MRI) or fluid simulations (CFD).

Neither is perfect: 4D flow MRI is progressively gaining ground in clinical applications, but its output is still relatively noisy and low in resolution. Typical CFD simulations are based on invasively acquired and/or noisy, defective initial- and anatomical data, with large errors as a result.

In an effort to increase the quality of the MRI data, assimilating flow measurements with a model for fluid flow like the Navier–Stokes equations has become a highly researched problem; see [DV13, SD18, TZA+20, ZBF+20, BBFV20], and the recent review [KKL+21] on *physics-informed machine learning* for data-driven problems governed by PDEs. It would be highly relevant to apply our space-time method to this task, and relate its performance to existing solutions.

$\|Bw\delta - 1\|$

2 Preliminaries

Before diving into the results of this thesis, let's introduce the relevant topics at a more gentle pace. We discuss the abstract concept of *residual minimization* for well-posed operator equations, formulate sufficient and necessary conditions for *uniform quasi-optimality* of discrete solutions, and say a little about *robust error estimation*. We then turn to solving the arising linear systems efficiently, and finish with the application to parabolic evolution problems.

Notation

In this thesis, by $C \lesssim D$ we mean that C can be bounded by a multiple of D, independently of parameters which C and D may depend on. Obviously, $C \gtrsim D$ is defined as $D \lesssim C$, and $C \eqsim D$ as $C \lesssim D$ and $C \gtrsim D$.

For Hilbert spaces $\mathbb{U}$ and $\mathbb{V}$, by $\mathcal{L}(\mathbb{U}, \mathbb{V}')$ we denote the Hilbert space of bounded linear mappings from $\mathbb{U}$ to the dual $\mathbb{V}'$ of $\mathbb{V}$, and by $\mathcal{L}\mathrm{is}(\mathbb{U}, \mathbb{V}')$ its subset of boundedly invertible linear mappings $\mathbb{U} \to \mathbb{V}'$. When $\mathcal{B} \in \mathcal{L}(\mathbb{U}, \mathbb{V}')$ is invertible with bounded inverse, we write $\mathcal{B} \in \mathcal{L}\mathrm{is}(\mathbb{U}, \mathbb{V}')$. The adjoint of $\mathcal{B}$, denoted $\mathcal{B}' \in \mathcal{L}\mathrm{is}(\mathbb{V}, \mathbb{U}')$, is the unique linear operator that satisfies $(\mathcal{B}w)(v) = (\mathcal{B}'v)(w)$ for all $w \in \mathbb{U}$ and $v \in \mathbb{V}$.

2.1 Well-posed linear operator equations

Given Hilbert spaces $\mathbb{U}$ and $\mathbb{V}$, for convenience over $\mathbb{R}$, a bounded bilinear form $b : \mathbb{U} \times \mathbb{V} \to \mathbb{R}$, and a linear form $l \in \mathbb{V}'$, find a $u \in \mathbb{U}$ such that

$$b(u, v) = l(v) \quad (v \in \mathbb{V}). \tag{2.1.1}$$

The central question that arises is that of *well-posedness* of this problem:

Existence Is there a solution for any $l \in \mathbb{V}'$?

Uniqueness Is there *only a single* u for any $l \in \mathbb{V}'$?

Continuity Do small changes in l correspond with small changes in u?

Because b is bounded, **existence** and **uniqueness** together imply **continuity** (by the Open Mapping Theorem).

To answer this question, we take a functional-analytic perspective. Define the linear operator $\mathcal{B} : \mathbb{U} \to \mathbb{V}' : w \mapsto b(w, \cdot)$. Then $\mathcal{B}$ is bounded with $\|\mathcal{B}\| = \|b\|$, and (2.1.1) can be recast as seeking $u \in \mathbb{U}$ so that $\mathcal{B}u = l$ in $\mathbb{V}'$.

Theorem ([Bra07, Thm. 4.3]). *For $\mathcal{B} \in \mathcal{L}(\mathbb{U}, \mathbb{V}')$, the following are equivalent:*

1. *Problem* (2.1.1) *is well-posed;*
2. *$\mathcal{B}$ is invertible with bounded inverse; $\mathcal{B} \in \mathcal{L}\mathrm{is}(\mathbb{U}, \mathbb{V}')$;*
3. *With $R_{\mathbb{V}} : \mathbb{V} \to \mathbb{V}' : v \mapsto \langle \cdot, v \rangle_{\mathbb{V}}$ the Riesz map, for any $l \in \mathbb{V}'$, u is the $\mathbb{U}$-component of the solution $(\mu, u) \in \mathbb{V} \times \mathbb{U}$ of the saddle point problem*

$$\begin{bmatrix} R_{\mathbb{V}} & \mathcal{B} \\ \mathcal{B}' & 0 \end{bmatrix} \begin{bmatrix} \mu \\ u \end{bmatrix} = \begin{bmatrix} l \\ 0 \end{bmatrix} \quad \text{in} \quad \mathbb{V}' \times \mathbb{U}'. \tag{2.1.2}$$

Remark. The formulation in (2.1.2) is called a *saddle point problem* as (μ, u) is a *saddle point* of its Lagrange functional

$$\mathcal{L}(v, w) := \tfrac{1}{2}(R_{\mathbb{V}} v)(v) + [(\mathcal{B}w)(v) - l(v)]$$

which is to say that

$$\mathcal{L}(\mu, w) \leq \mathcal{L}(\mu, u) \leq \mathcal{L}(v, u) \quad ((v, w) \in \mathbb{V} \times \mathbb{U});$$

(μ, u) minimizes $\mathcal{L}$ in the $\mathbb{V}$-component but maximizes it in the $\mathbb{U}$-component. The name derives from the fact that the prototypical example in 2 dimensions is a surface that curves *up* in one axis and *down* in the other, resembling the shape of a riding saddle. ♢

2.2 Minimal residual approximation

When $\mathbb{U}$ is infinite-dimensional, we cannot be expected to construct u from l. For that reason, the next-best thing is to find its *best approximation* from some finite-dimensional subspace $\mathbb{U}^\delta \subset \mathbb{U}$, i.e. the minimizer of $\|u - w^\delta\|_{\mathbb{U}}$ over $w^\delta \in \mathbb{U}^\delta$. As u is unknown, this is generally not feasible either.

Instead, we aim for a *quasi-optimal* $u^\delta \in \mathbb{U}^\delta$ that satisfies

$$\|u - u^\delta\|_{\mathbb{U}} \lesssim \min_{w^\delta \in \mathbb{U}^\delta} \|u - w^\delta\|_{\mathbb{U}}. \tag{2.2.1}$$

Usually, δ denotes some *discretization parameter* and we are not interested in a single δ, but rather a whole family $\Delta = \{\delta\}$. In that case, we aim for *uniform quasi-optimal* solutions that satisfy (2.2.1) with a constant independent of the discretization parameter $\delta \in \Delta$.

Remark. When $\mathbb{V} = \mathbb{U}$ and $\mathcal{B}$ is *coercive* in that $(\mathcal{B}\cdot)(\cdot) \geq \alpha \|\cdot\|_{\mathbb{U}}^2$—cf. (2.4.2)—finding quasi-best approximations is easy: *Céa's Lemma* states that the unique $u^\delta \in \mathbb{U}^\delta$ so that $\mathcal{B}u^\delta = l$ in $\mathbb{U}^{\delta\prime}$—the *Galerkin approximation* u^δ of u—is quasi-optimal with constant $\frac{\|\mathcal{B}\|}{\alpha}$ (even $\frac{\|\mathcal{B}\|^{1/2}}{\alpha^{1/2}}$ when $\mathcal{B}$ is also symmetric); cf. [BS08, §2.8]. For $\mathbb{V} \neq \mathbb{U}$, our focus, the problem is more nuanced. ♢

The minimizer u^δ of $\|l - \mathcal{B}w^\delta\|_{\mathbb{V}'}$ over $w^\delta \in \mathbb{U}^\delta$ is uniformly quasi-optimal:

$$\begin{aligned} \|u - u^\delta\|_{\mathbb{U}} &\leq \|\mathcal{B}^{-1}\| \|l - \mathcal{B}u^\delta\|_{\mathbb{V}'} = \|\mathcal{B}^{-1}\| \min_{w^\delta \in \mathbb{U}^\delta} \|l - \mathcal{B}w^\delta\|_{\mathbb{V}'} \\ &\leq \|\mathcal{B}^{-1}\| \|\mathcal{B}\| \min_{w^\delta \in \mathbb{U}^\delta} \|u - w^\delta\|_{\mathbb{U}}, \end{aligned}$$

with constant $\|\mathcal{B}\|\|\mathcal{B}^{-1}\|$. However, due to the the dual norm $\|\cdot\|_{\mathbb{V}'}$, in general, neither u^δ nor its residual $\|\mathcal{B}u^\delta - l\|_{\mathbb{V}'}$ can be computed.

There is hope, though. The idea is to find a family $(\mathbb{V}^\delta)_{\delta\in\Delta}$ of closed (finite-dimensional) *test spaces* in $\mathbb{V}$, ideally with a dimension that is proportional to that of its corresponding trial space $\mathbb{U}^\delta$, for which $\|\mathcal{B}\cdot\|_{\mathbb{V}'}$ can be controlled by the computable quantity $\|\mathcal{B}\cdot\|_{\mathbb{V}^{\delta\prime}}$ uniformly in $\delta\in\Delta$, i.e.

$$\inf_{\delta\in\Delta}\inf_{w\in\mathbb{U}^\delta}\sup_{0\neq v\in\mathbb{V}^\delta}\frac{(\mathcal{B}w)(v)}{\|\mathcal{B}w\|_{\mathbb{V}'}\|v\|_{\mathbb{V}}}=\inf_{\delta\in\Delta}\inf_{w\in\mathbb{U}^\delta}\frac{\|\mathcal{B}w\|_{\mathbb{V}^{\delta\prime}}}{\|\mathcal{B}w\|_{\mathbb{V}'}}=:\gamma_\Delta^{\mathcal{B}}>0. \tag{2.2.2}$$

We call such families $(\mathbb{U}^\delta,\mathbb{V}^\delta)_{\delta\in\Delta}$ *uniformly inf-sup stable*; the equivalent

$$\inf_{\delta\in\Delta}\inf_{w\in\mathbb{U}^\delta}\sup_{0\neq v\in\mathbb{V}^\delta}\frac{(\mathcal{B}w)(v)}{\|w\|_{\mathbb{U}}\|v\|_{\mathbb{V}}}>0 \tag{2.2.3}$$

is called the *Ladyzhenskaya–Babuška–Brezzi (LBB) condition* in literature.

Remark. The *optimal test space* $\mathbb{V}^\delta_{\mathrm{opt}}:=R_{\mathbb{V}}^{-1}\mathcal{B}\mathbb{U}^\delta$ satisfies $\dim\mathbb{V}^\delta=\dim\mathbb{U}^\delta$ and the family $(\mathbb{U}^\delta,\mathbb{V}^\delta_{\mathrm{opt}})_{\delta\in\Delta}$ satisfies $\gamma_\Delta^{\mathcal{B}}=1$. The resulting solutions equal the minimizer of $\|\mathcal{B}w-l\|_{\mathbb{V}'}$ over $w\in\mathbb{U}^\delta$, and so are quasi-optimal with constant $\|\mathcal{B}\|\|\mathcal{B}^{-1}\|$. However, the construction of $\mathbb{V}^\delta_{\mathrm{opt}}$ is difficult unless $\mathbb{V}$ is an L_2-type function space. Solutions exist by constructing *projected optimal test spaces*; see the proof below for an example. ◊

In this discrete dual norm, consider the *minimal residual approximation*

$$u^\delta_{\mathrm{MR}}:=\operatorname*{arg\,min}_{w^\delta\in\mathbb{U}^\delta}\|\mathcal{B}w^\delta-l\|_{\mathbb{V}^{\delta\prime}}. \tag{2.2.4}$$

This minimizer exists uniquely thanks to (2.2.2) and unique solvability of the saddle-point system (2.1.2). We discuss its practical computation in §2.3. First, we have the following result.

Theorem 2.2.1 (Inf-sup stability $\iff$ quasi-optimality [§3.3.1])**.** *Assume $\mathcal{B}\in\mathcal{L}\mathrm{is}(\mathbb{U},\mathbb{V}')$. A family $(\mathbb{U}^\delta,\mathbb{V}^\delta)_{\delta\in\Delta}$ of proper nontrivial closed subspaces of $\mathbb{U}\times\mathbb{V}$ satisfies* (2.2.2) *exactly when its minimal residual approximations are uniformly quasi-optimal. In that case,*

$$\|u-u^\delta_{\mathrm{MR}}\|_{\mathbb{U}}\leq\frac{\|\mathcal{B}\|\|\mathcal{B}^{-1}\|}{\gamma_\Delta^{\mathcal{B}}}\inf_{w^\delta\in\mathbb{U}^\delta}\|u-w^\delta\|_{\mathbb{U}}\quad(u\in\mathbb{U},\delta\in\Delta). \tag{2.2.5}$$

Proof. Defining the trial-to-test map $T^\delta\in\mathcal{L}(\mathbb{U},\mathbb{V}^\delta)$ by $\langle T^\delta w,v\rangle_{\mathbb{V}}=(\mathcal{B}w)(v)$, the function $T^\delta w\in\mathbb{V}^\delta$ is the $\mathbb{V}$-orthogonal projection onto $\mathbb{V}^\delta$ of the *optimal test function* $R_{\mathbb{V}}^{-1}\mathcal{B}w$. Define the *projected optimal test space* $\overline{\mathbb{V}}^\delta:=\operatorname{ran}T^\delta|_{\mathbb{U}^\delta}$. Then $\overline{\mathbb{V}}^\delta\subset\mathbb{V}^\delta$ and $\|\mathcal{B}w\|_{\mathbb{V}^{\delta\prime}}=\|\mathcal{B}w\|_{\overline{\mathbb{V}}^{\delta\prime}}$ for $w\in\mathbb{U}^\delta$, so (2.2.2) still holds upon restriction to $v\in\overline{\mathbb{V}}^\delta$. Moreover, $\dim\overline{\mathbb{V}}^\delta=\dim\mathbb{U}^\delta$ and $E_{\overline{\mathbb{V}}}^{\delta\prime}\mathcal{B}\mathbb{U}^\delta\in\mathcal{L}\mathrm{is}(\mathbb{U}^\delta,\overline{\mathbb{V}}^{\delta\prime})$. The result follows from Remark 3.3.2 (which stability constant is the LBB-constant $\kappa_\Delta^{\mathcal{B}}$ from (2.2.3)) after estimating $\kappa_\Delta^{\mathcal{B}}\geq\frac{\gamma_\Delta^{\mathcal{B}}}{\|\mathcal{B}^{-1}\|}$. □

In words, uniform quasi-optimality of minimal residual approximations is *equivalent to* uniform inf-sup stability of the corresponding family of discrete trial- and test spaces, so the game becomes showing (2.2.2). In the sequel, we will often show inf-sup stability conditions using yet *another* equivalence.

2.2.1 Fortin interpolators

A linear operator $Q^\delta : \mathbb{V} \to \mathbb{V}$ is a *Fortin interpolator* [For77] when

$$\operatorname{ran} Q^\delta \subset \mathbb{V}^\delta, \quad \|Q^\delta\| < \infty, \quad (\mathcal{B}w^\delta)(Q^\delta v) = (\mathcal{B}w^\delta)(v) \quad ((w^\delta, v) \in \mathbb{U}^\delta \times \mathbb{V}).$$

It is well-known that the existence of uniformly bounded Fortin interpolators is a sufficient condition for uniform inf-sup stability. Indeed: for all $v \in V$ we have $\|Q^\delta v\|_{\mathbb{V}} \leq \|Q^\delta\| \|v\|_{\mathbb{V}}$, so that for $v \neq 0$, $\frac{1}{\|v\|_{\mathbb{V}}} \leq \frac{\|Q^\delta\|}{\|Q^\delta v\|_{\mathbb{V}}}$. Using this, we find that for any $w^\delta \in \mathbb{U}^\delta$,

$$\begin{aligned}\|\mathcal{B}w^\delta\|_{\mathbb{V}'} &= \sup_{0\neq v\in\mathbb{V}} \frac{(\mathcal{B}w^\delta)(v)}{\|v\|_{\mathbb{V}}} = \sup_{0\neq v\in\mathbb{V}} \frac{(\mathcal{B}w^\delta)(Q^\delta v)}{\|v\|_{\mathbb{V}}} \leq \|Q^\delta\| \sup_{0\neq v\in\mathbb{V}} \frac{(\mathcal{B}w^\delta)(Q^\delta v)}{\|Q^\delta v\|_{\mathbb{V}}} \\ &\leq \|Q^\delta\| \sup_{0\neq v^\delta\in\mathbb{V}^\delta} \frac{(\mathcal{B}w^\delta)(v^\delta)}{\|v^\delta\|_{\mathbb{V}}} = \|Q^\delta\| \|\mathcal{B}w^\delta\|_{\mathbb{V}^{\delta\prime}}\end{aligned}$$

so that $\frac{\|\mathcal{B}w^\delta\|_{\mathbb{V}^{\delta\prime}}}{\|\mathcal{B}w^\delta\|_{\mathbb{V}'}} \geq \|Q^\delta\|^{-1}$, uniformly in $w^\delta \in \mathbb{U}^\delta$ and $\delta \in \Delta$. It turns out that the existence of such interpolators is also *necessary*:[1]

Theorem (Fortin interpolator $\iff$ inf-sup stability [Thm. 7.3.11])**.** *Assume $\mathcal{B} \in \mathcal{L}\mathrm{is}(\mathbb{U}, \mathbb{V}')$, let $(\mathbb{U}^\delta, \mathbb{V}^\delta)_{\delta\in\Delta}$ be a family of closed subspaces of $\mathbb{U} \times \mathbb{V}$. Then there exists a family $(Q^\delta)_{\delta\in\Delta}$ of uniformly bounded Fortin interpolators exactly when $(\mathbb{U}^\delta, \mathbb{V}^\delta)_{\delta\in\Delta}$ is uniformly inf-sup stable as in* (2.2.2)*.*

2.2.2 A posteriori error estimation

To get any idea about the $\mathbb{U}$-norm error $\|w^\delta - u\|_{\mathbb{U}}$ in some $w^\delta \in \mathbb{U}^\delta$. we have to use the information available to us. The *norm of the residual*

$$\|\mathcal{B}w^\delta - l\|_{\mathbb{V}^{\delta\prime}}$$

sounds promising, as its continuous counterpart has $\|\mathcal{B}w^\delta - l\|_{\mathbb{V}'} \eqsim \|w^\delta - u\|_{\mathbb{U}}$. In fact, boundedness of $\mathcal{B}$ immediately yields *efficiency*:

$$\|\mathcal{B}w^\delta - l\|_{\mathbb{V}^{\delta\prime}} = \|\mathcal{B}(w^\delta - u)\|_{\mathbb{V}^{\delta\prime}} \leq \|\mathcal{B}(w^\delta - u)\|_{\mathbb{V}'} \leq \|\mathcal{B}\| \|w^\delta - u\|_{\mathbb{U}}.$$

This means that the residual norm doesn't *overestimate* the error too much. The challenge is to show *reliability*, i.e., that the residual norm doesn't *underestimate* the error too much either.

[1] The theorem is actually stated for $\mathcal{B} \in \mathcal{L}(\mathbb{U}, \mathbb{V}')$, with condition (2.2.2) defined over $\inf_{\{w\in\mathbb{U}^\delta : \mathcal{B}w\neq 0\}}$. In current form, the equivalence is also found in [Bra07, 4.8–4.9].

Property 3 of the Fortin interpolator Q^δ guaranteed by §2.2.1 asserts that

$$(\mathrm{Id} - Q^{\delta\prime})(\mathcal{B}\mathbb{U}^\delta) = 0,$$

which, combined with boundedness of $\mathcal{B}^{-1}$ and a triangle inequality, shows

$$\begin{aligned}
\|w^\delta - u\|_{\mathbb{U}} &\le \|\mathcal{B}^{-1}\|\|\mathcal{B}w^\delta - l\|_{\mathbb{V}'} \\
&\le \|\mathcal{B}^{-1}\| \left(\|Q^{\delta\prime}(\mathcal{B}w^\delta - l)\|_{\mathbb{V}^{\delta\prime}} + \|(\mathrm{Id} - Q^{\delta\prime})(\mathcal{B}w^\delta - l)\|_{\mathbb{V}'} \right) \\
&= \|\mathcal{B}^{-1}\| \left(\|Q^{\delta\prime}(\mathcal{B}w^\delta - l)\|_{\mathbb{V}^{\delta\prime}} + \|(\mathrm{Id} - Q^{\delta\prime})l\|_{\mathbb{V}'} \right) \\
&\le \|\mathcal{B}^{-1}\|\|Q^\delta\|\|\mathcal{B}w^\delta - l\|_{\mathbb{V}^{\delta\prime}} + \|\mathcal{B}^{-1}\|\|(\mathrm{Id} - Q^{\delta\prime})l\|_{\mathbb{V}'}.
\end{aligned}$$

We see that the residual norm is reliable *up to* a quantity $\|(\mathrm{Id} - Q^{\delta\prime})l\|_{\mathbb{V}'}$ we call *data oscillation*. For this result to be satisfactory, $Q^{\delta\prime}$ should have *approximation properties* in that, for sufficiently smooth data l, the data oscillation is of equal or higher order than the discretization error $\inf_{w^\delta \in \mathbb{U}^\delta} \|w^\delta - u\|_{\mathbb{U}}$.

This is out of reach from the abstract viewpoint taken here, but we continue the argument in §4.7.2 for the parabolic evolution problem, and in §7.3.2 for the parabolic data assimilation problem.

2.3 Solving the discretized system

Writing $E^\delta_{\mathbb{U}} : \mathbb{U}^\delta \to \mathbb{U}$ and $E^\delta_{\mathbb{V}} : \mathbb{V}^\delta \to \mathbb{V}$ for the trivial embeddings, the minimal residual solution u^δ_{MR} from (2.2.4) equals the second component of the discrete saddle point problem of finding $(\mu^\delta_{\mathrm{MR}}, u^\delta_{\mathrm{MR}}) \in \mathbb{V}^\delta \times \mathbb{U}^\delta$ such that

$$\begin{bmatrix} {E^\delta_{\mathbb{V}}}' R_{\mathbb{V}} E^\delta_{\mathbb{V}} & {E^\delta_{\mathbb{V}}}' \mathcal{B} E^\delta_{\mathbb{U}} \\ {E^\delta_{\mathbb{U}}}' \mathcal{B}' E^\delta_{\mathbb{V}} & 0 \end{bmatrix} \begin{bmatrix} \mu^\delta_{\mathrm{MR}} \\ u^\delta_{\mathrm{MR}} \end{bmatrix} = \begin{bmatrix} {E^\delta_{\mathbb{V}}}' l \\ 0 \end{bmatrix} \quad \text{in} \quad \mathbb{V}^{\delta\prime} \times \mathbb{U}^{\delta\prime}. \tag{2.3.1}$$

Equipping the pair $(\mathbb{U}^\delta, \mathbb{V}^\delta)$ with appropriate bases $(\Phi^\delta, \Psi^\delta)$, we define

$$\boldsymbol{B} := (\mathcal{B}\Phi^\delta)(\Psi^\delta), \quad \boldsymbol{R} := (R_{\mathbb{V}}\Psi^\delta)(\Psi^\delta), \quad \boldsymbol{l} := l(\Psi^\delta).$$

We call $\boldsymbol{B}$ the *system matrix* and $\boldsymbol{l}$ the *load vector*. The matrix representation of our saddle point problem is to find $(\boldsymbol{\mu}, \boldsymbol{u}) \in \mathbb{R}^{\dim \mathbb{V}^\delta} \times \mathbb{R}^{\dim \mathbb{U}^\delta}$ such that

$$\boldsymbol{M} \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{u} \end{bmatrix} = \begin{bmatrix} \boldsymbol{l} \\ \boldsymbol{0} \end{bmatrix} \quad \text{where} \quad \boldsymbol{M} := \begin{bmatrix} \boldsymbol{R} & \boldsymbol{B} \\ \boldsymbol{B}^\top & \boldsymbol{0} \end{bmatrix}.$$

For vectors $\boldsymbol{v} \in \mathbb{R}^{\dim \mathbb{V}^\delta}$ and $\boldsymbol{w} \in \mathbb{R}^{\dim \mathbb{U}^\delta}$, we can write $v := \boldsymbol{v}^\top \Psi^\delta \in \mathbb{V}^\delta$ and $w := \boldsymbol{w}^\top \Phi^\delta \in \mathbb{U}^\delta$. In this notation, $\mu^\delta_{\mathrm{MR}} = \boldsymbol{\mu}^\top \Psi^\delta$ and $u^\delta_{\mathrm{MR}} = \boldsymbol{u}^\top \Phi^\delta$.

Remark. The proof of Theorem 2.2.1 reveals that with the projected optimal test space $\overline{\mathbb{V}}^\delta \subset \mathbb{V}^\delta$, the resulting system matrix is square and invertible, and

thus can be used to solve for u^δ_{MR}. This idea is used extensively in *discontinuous Petrov–Galerkin methods*, where $\mathbb{V}$ is a 'broken' (discontinuous) space and constructing $\overline{\mathbb{V}}^\delta$ is efficient. In our case, its construction is infeasible, as $R_{\mathbb{V}}^{-1}$ is a *nonlocal* operator so the resulting matrix would be dense. We instead solve the saddle point system (2.3.1) or the normal equations (2.3.2) below. ♢

2.3.1 An important insight

As $\boldsymbol{R}$ is symmetric positive definite (indeed: $\boldsymbol{v}^\top \boldsymbol{R}\boldsymbol{v} = \langle \boldsymbol{v}^\top\Psi^\delta, \boldsymbol{v}^\top\Psi^\delta\rangle_{\mathbb{V}} = \|\boldsymbol{v}^\top\Psi^\delta\|^2_{\mathbb{V}}$), we can take the Schur complement of $\boldsymbol{M}$ w.r.t. the $\mathbb{V}^\delta$-block to isolate the $\mathbb{U}^\delta$-block, yielding the equivalent problem of finding $\boldsymbol{u} \in \mathbb{R}^{\dim \mathbb{U}^\delta}$ that solves

$$\boldsymbol{B}^\top \boldsymbol{R}^{-1}\boldsymbol{B}\boldsymbol{u} = \boldsymbol{B}^\top \boldsymbol{R}^{-1}\boldsymbol{l}. \tag{2.3.2}$$

Usually the exact evaluation of the action of $\boldsymbol{R}^{-1}$ will not be feasible. However, $\boldsymbol{R}^{-1}$ can be replaced by any *spectrally equivalent* $\boldsymbol{K}$, i.e. one for which there are $0 < r_\Delta \leq R_\Delta < \infty$ such that

$$\sigma(\boldsymbol{K}\boldsymbol{R}) \subset [r_\Delta, R_\Delta] \quad (\delta \in \Delta). \tag{2.3.3}$$

Matrices like these are called *optimal preconditioners* and play an important role in solving linear systems efficiently as well, as discussed below in §2.3.3.

As $\boldsymbol{R}^{-1}$ is the matrix inducing the norm on $\mathbb{V}^{\delta\prime}$, this amounts to replacing the norm on $\mathbb{V}^{\delta\prime}$ by the *uniformly equivalent* norm $|||\cdot|||_{\mathbb{V}^{\delta\prime}}$ induced by $\boldsymbol{K}$, hence

$$\frac{|||h^\delta|||_{\mathbb{V}^{\delta\prime}}}{\|h^\delta\|_{\mathbb{V}^{\delta\prime}}} \in [R_\Delta^{-1}, r_\Delta^{-1}] \quad (h^\delta \in \mathbb{V}^{\delta\prime}, \delta \in \Delta).$$

Replacing $\boldsymbol{R}^{-1}$ by $\boldsymbol{K}$ in (2.3.2) yields the problem of finding $\hat{\boldsymbol{u}} \in \mathbb{R}^{\dim \mathbb{U}^\delta}$ s.t.

$$\boldsymbol{S}\hat{\boldsymbol{u}} = \boldsymbol{b} \quad \text{where} \quad \boldsymbol{S} := \boldsymbol{B}^\top\boldsymbol{K}\boldsymbol{B} \quad \text{and} \quad \boldsymbol{b} := \boldsymbol{B}^\top\boldsymbol{K}\boldsymbol{l}.$$

Note that in general, the solutions $\hat{\boldsymbol{u}}$ and $\boldsymbol{u}$ do not coincide. However, we deduce that $\hat{u}^\delta_{\mathrm{MR}} := \hat{\boldsymbol{u}}^\top\Phi^\delta$ is the unique minimizer of $|||\mathcal{B}w^\delta - l|||_{\mathbb{V}^{\delta\prime}}$ over $w^\delta \in \mathbb{U}^\delta$. As a result, $\hat{u}^\delta_{\mathrm{MR}}$ is uniformly quasi-optimal, with

$$\|u - \hat{u}^\delta_{\mathrm{MR}}\|_{\mathbb{U}} \leq \frac{\|\mathcal{B}\|\|\mathcal{B}^{-1}\|}{\gamma^{\mathcal{B}}_\Delta} \frac{\max(R_\Delta, 1)}{\min(r_\Delta, 1)} \inf_{w^\delta \in \mathbb{U}^\delta} \|u - w^\delta\|_{\mathbb{U}} \quad (\delta \in \Delta).$$

The importance of this reformulation is that the matrix $\boldsymbol{S}$ is *symmetric positive definite*, which allows us to use an iterative solver such as Conjugate Gradients [HS52] to obtain a good approximation to $\hat{u}^\delta_{\mathrm{MR}}$ very efficiently.

Alternatively, one can use MINRES [PS75] to solve the saddle-point system, or Bramble–Pasciak CG [BP88], which allows nonsymmetric preconditioners and offers a quantitative advantage over MINRES. When a preconditioner $\boldsymbol{K}$ to $\boldsymbol{R}$ is not available, a variant of LSQR [PS82, Ben99] can be used to solve the minimal residual problem directly.

2.3.2 Matrix-free iterative solvers

One important hurdle in applications is that the dimension of trial- and test spaces can become immense, in the order of billions. In this case, performing a direct solve on the monolithic matrix $\boldsymbol{S}$ is prohibitively expensive.

In fact, even storing the system matrix $\boldsymbol{B}$ in memory may be out of reach. Luckily, efficient methods exist for the *matrix-free* solution of linear systems. Requiring only the *action* $\boldsymbol{x} \mapsto \boldsymbol{S}\boldsymbol{x}$ (and so those of $\boldsymbol{B}^\top$, $\boldsymbol{K}$, and $\boldsymbol{B}$), we solve the linear system using an *iterative* method: Starting from some *initial guess* $\hat{\boldsymbol{u}}_0$, we generate a sequence $(\hat{\boldsymbol{u}}_k)_{k\in\mathbb{N}}$ of improving approximate solutions.

For $k \in \mathbb{N}$, define $\hat{u}_k^\delta := \hat{\boldsymbol{u}}_k^\top \Phi^\delta \in \mathbb{U}^\delta$. Stopping after $\underline{k}$ iterations and incurring an *algebraic error* $\|\hat{u}_{\mathrm{MR}}^\delta - \hat{u}_{\underline{k}}^\delta\|_{\mathbb{U}}$ is acceptable, as long as it is bounded by the *discretization error* $\inf_{w^\delta\in\mathbb{U}^\delta} \|u - w^\delta\|_{\mathbb{U}}$ with some fixed factor λ_Δ. Then, the resulting iterative solution is still quasi-optimal:

$$\begin{aligned}\|u - \hat{u}_{\underline{k}}^\delta\|_{\mathbb{U}} &\le \|u - \hat{u}_{\mathrm{MR}}^\delta\|_{\mathbb{U}} + \|\hat{u}_{\mathrm{MR}}^\delta - \hat{u}_{\underline{k}}^\delta\|_{\mathbb{U}} \le (1+\lambda_\Delta)\|u - \hat{u}_{\mathrm{MR}}^\delta\|_{\mathbb{U}} \\ &\le (1+\lambda_\Delta)\tfrac{\|\mathcal{B}\|\|\mathcal{B}^{-1}\|}{\gamma_\Delta^{\mathcal{B}}}\tfrac{\max(R_\Delta,1)}{\min(r_\Delta,1)} \inf_{w^\delta\in\mathbb{U}^\delta} \|u - w^\delta\|_{\mathbb{U}} \qquad (\delta\in\Delta).\end{aligned}$$

Of course, computing the algebraic error is not straight-forward, as $\hat{u}_{\mathrm{MR}}^\delta$ is unavailable. For Conjugate Gradients, there are methods to get accurate bounds on the algebraic error using information already available in the loop; see the recent work [MPT21] and references therein.

2.3.3 Optimal preconditioning

The performance of the iterative method depends on the quality of the initial guess, but more so on the (spectral) *condition number* of $\boldsymbol{S}$ defined as

$$\kappa_2(\boldsymbol{S}) := \|\boldsymbol{S}\|\|\boldsymbol{S}^{-1}\|.$$

This condition number typically explodes for increasing problem size.

A *preconditioner* $\boldsymbol{P}$ is a matrix such that $\kappa_2(\boldsymbol{P}\boldsymbol{S}) \ll \kappa_2(\boldsymbol{S})$. Preconditioners are an essential tool in solving large sparse systems, as they increase the rate of convergence of an iterative method dramatically. In fact, $\boldsymbol{S}$ is derived from a linear operator $S \in \mathcal{L}(X^\delta, X^{\delta\prime})$, so any linear operator $P \in \mathcal{L}(X^{\delta\prime}, X^\delta)$ satisfying $\cdot(P\cdot) \eqsim \|\cdot\|_{X^{\delta\prime}}^2$ yields an *optimal preconditioner* $\boldsymbol{P}$, for which $\kappa_2(\boldsymbol{P}\boldsymbol{S}) \lesssim 1$ uniformly in $\delta \in \Delta$; cf. [Hip06].

In the most extreme case $P = S^{-1}$, any iterative method is finished after a single iteration, but computing the action $\boldsymbol{x} \mapsto \boldsymbol{S}^{-1}\boldsymbol{x}$ is prohibitively expensive. Therefore, the next best thing is an optimal preconditioner we can apply efficiently (at cost proportional to $\dim X^\delta$). In our parabolic problems, these are built using wavelets in time and multigrid in space; see [And16, SvVW21].

With an optimal preconditioner, the iterative solver reduces the algebraic error with a fixed factor in each iteration. In this case, also estimating the algebraic error becomes feasible by measuring $\|\boldsymbol{S}\hat{\boldsymbol{u}}_k - \boldsymbol{b}\|_{\boldsymbol{P}}$. Indeed:

$$\|\boldsymbol{S}\hat{\boldsymbol{u}}_k - \boldsymbol{b}\|_{\boldsymbol{P}} = \|\boldsymbol{S}(\hat{\boldsymbol{u}}_k - \hat{\boldsymbol{u}})\|_{\boldsymbol{P}} \eqsim \|\boldsymbol{S}(\hat{\boldsymbol{u}}_k - \hat{\boldsymbol{u}})\|_{\boldsymbol{S}^{-1}} = \|\hat{\boldsymbol{u}}_k - \hat{\boldsymbol{u}}\|_{\boldsymbol{S}} = \|\hat{u}_{\mathrm{MR}}^\delta - \hat{u}_k^\delta\|_{\mathbb{U}}.$$

In fact, $\|\boldsymbol{S}\hat{\boldsymbol{u}}_k - \boldsymbol{b}\|_{\boldsymbol{P}}$ is already available within a preconditioned Conjugate Gradients loop (as its variable β).

2.4 Parabolic evolution problems

Heuristically, parabolic PDEs describe time-dependent physical processes that dissipate[2] over time towards some steady state described by an *elliptic* PDE. We give a brief introduction on these elliptic PDEs.

Elliptic boundary-value problems

Let $\Omega \subset \mathbb{R}^d$ be a *Lipschitz* domain as in [BS08, Def. 1.46]. For given *forcing data* $f : \Omega \to \mathbb{R}$, we want to find $u : \Omega \to \mathbb{R}$ solving the *boundary-value problem*

$$\begin{cases} -\Delta u = f & (\text{on } \Omega), \\ u = 0 & (\text{on } \partial\Omega). \end{cases}$$

This is the *Poisson equation*, and it is the prototyptical elliptic PDE. Describing the negative Laplacian $-\Delta u$ by its action when integrated against functions is known as a *weak formulation*. Integration by parts yields the bilinear form

$$a : V \times V \to \mathbb{R} : (\eta, \zeta) \mapsto \int_\Omega \nabla\eta \cdot \nabla\zeta \, \mathrm{d}\boldsymbol{x}. \tag{2.4.1}$$

Then for $V := H^1_0(\Omega)$, a is continuous, and *coercive* in that for some $\alpha > 0$

$$a(\eta, \eta) \geq \alpha \|\eta\|_V^2 \quad (\eta \in V), \tag{2.4.2}$$

(with $\|a\| = \alpha = 1$). For $f \in V' = H^{-1}(\Omega)$, the problem to find $u \in V$ that solves $a(u, \zeta) = f(\zeta)$ for all $\zeta \in V$ is well-posed in the sense of §2.1.

Example. Many more elliptic PDEs exist. The equation underlying many chemical processes replaces the Laplace operator $\Delta = \operatorname{div} \nabla$ with

$$-\operatorname{div} \boldsymbol{K}\nabla + \boldsymbol{b} \cdot \nabla + c,$$

where $\boldsymbol{K}$ models *diffusion*, $\boldsymbol{b}$ models *convective transport*, and c is a *reaction term* that can also depend nonlinearly on u. We touched briefly on these reaction-diffusion systems in the Outlook of Chapter 1, and discuss a linear convection-dominated diffusion problem in Chapter 4. ◊

Example. The fourth-order *biharmonic equation* $\Delta^2 u = f$ models the equilibrium position of a thin elastic plate subject to a vertical force f. Complemented with clamped boundary conditions, the corresponding bilinear form is coercive on $V := H^2_0(\Omega)$, and one can use Argyris elements for its conforming discretization. See [BS08, §5.9] for an overview.

This equation has a rich history, and Boris Galerkin (better romanized as Gal**yo**rkin) first described his now-ubiquitous *Galerkin method* within this context. See the excellent reviews [GW12, GK12]. ◊

[2] Dissipative here means that some notion of 'energy' is lost over time. This should be contrasted with *hyperbolic* equations, that *conserve* energy.

2.4.1 Parabolic evolution problems

Let V and H be Hilbert spaces (of functions over some spatial domain) such that the embedding $V \subset H$ is dense and continuous. Applying the Riesz representation theorem, we identify H with its dual, and find $H' \subset V'$ with dense and continuous embedding; we say $V \subset H \eqsim H' \subset V'$ is a *Gelfand triple*. Denote with $\langle \cdot, \cdot \rangle$ the inner product on $H \times H$ and its unique extension to $V' \times V$, and denote the norm on H by $\| \cdot \|$.

Let $I := (0, T)$ be a time interval. Let $a(t; \cdot, \cdot) : V \times V \to \mathbb{R}$ denote a bilinear form such that $t \mapsto a(t; \cdot, \cdot)$ is measurable and $a(t; \cdot, \cdot)$ is bounded and coercive—cf. (2.4.2)—uniformly in t. Define a family of linear operators by $(A(t)\cdot)(\cdot) := a(t; \cdot, \cdot)$. For an *initial value* u_0 and *forcing data* g, we seek u that solves the *parabolic evolution problem*

$$\begin{cases} \frac{\mathrm{d}u}{\mathrm{d}t}(t) + A(t)u(t) &= g(t) \quad (\text{for } t \in I), \\ u(0) &= u_0. \end{cases} \tag{2.4.3}$$

Example (Heat equation)**.** Taking $H := L_2(\Omega)$, $V := H_0^1(\Omega)$, and a from (2.4.1) yields the prototypical parabolic *heat equation* $\frac{\partial}{\partial t} u - \Delta_x u = g$. Its evolution problem models the temperature evolution of some object Ω subject to a *heat source* g, as the result of thermal conductivity. ◊

In simultaneous space-time variational form, its first equation reads as finding u from some space X of functions of time and space such that

$$(Bu)(v) := \int_I \langle \tfrac{\mathrm{d}u}{\mathrm{d}t}(t), v(t) \rangle + \langle A(t)u(t), v(t) \rangle \, \mathrm{d}t = \int_I \langle g(t), v(t) \rangle \, \mathrm{d}t =: g(v)$$

for all v from another space Y. With the trace map $\gamma_t : w \mapsto w(t, \cdot)$, we can enforce the initial condition *weakly* by testing $\gamma_0 u$ against additional functions. Defining A through $(Au)(v) := \int_I \langle A(t)u(t), v(t) \rangle \, \mathrm{d}t$, we get the following.

Theorem ([SS09])**.** *With $X := L_2(I; V) \cap H^1(I; V')$ and $Y := L_2(I; V)$,*

$$\begin{bmatrix} B \\ \gamma_0 \end{bmatrix} u \in \mathcal{L}\mathrm{is}(X, Y' \times H).$$

In other words, for any $(g, u_0) \in Y' \times H$, the problem of finding $u \in X$ s.t.

$$\begin{bmatrix} B \\ \gamma_0 \end{bmatrix} u = \begin{bmatrix} g \\ u_0 \end{bmatrix} \tag{2.4.4}$$

is a well-posed simultaneous space-time variational formulation of (2.4.3)*.*

Define the symmetric part of A as $A_s := \frac{1}{2}(A + A') \in \mathcal{L}\mathrm{is}(Y, Y')$ and $\partial_t := B - A \in \mathcal{L}(X, Y')$. We equip Y with the energy norm $\| \cdot \|_Y^2 := (A_s \cdot)(\cdot)$, so A_s becomes the Riesz map on Y. Equip X with the norm

$$\| \cdot \|_X^2 := \| \partial_t \cdot \|_{Y'}^2 + \| \cdot \|_Y^2 + \| \gamma_0 \cdot \|^2.$$

Both norms are equivalent to their natural norms.

We are now in the situation of the abstract introduction, with

$$\mathbb{U} := X, \quad \mathbb{V} := Y \times H, \quad \mathcal{B} := \begin{bmatrix} B \\ \gamma_0 \end{bmatrix}, \quad \text{and} \quad l := \begin{bmatrix} g \\ u_0 \end{bmatrix}.$$

Let's apply its results. The first insight is that (2.4.4) is equivalent to the saddle point problem of finding $(\mu, \sigma, u) \in Y \times H \times X$ such that

$$\begin{bmatrix} A_s & 0 & B \\ 0 & \mathrm{Id} & \gamma_0 \\ B' & \gamma_0' & 0 \end{bmatrix} \begin{bmatrix} \mu \\ \sigma \\ u \end{bmatrix} = \begin{bmatrix} g \\ u_0 \\ 0 \end{bmatrix} \quad \text{in} \quad Y' \times H \times X' \tag{2.4.5}$$

and we can look at discrete subspaces of X and $Y \times H$, respectively. In Chapter 3, we will see that it is harmless to *not* discretize H at all, which is equivalent to taking the Schur-complement of the H-block in (2.4.5). The resulting saddle point problem is to find $(\mu, u) \in Y \times X$ such that

$$\begin{bmatrix} A_s & B \\ B' & -\gamma_0'\gamma_0 \end{bmatrix} \begin{bmatrix} \mu \\ u \end{bmatrix} = \begin{bmatrix} g \\ -\gamma_0' u_0 \end{bmatrix} \quad \text{in} \quad Y' \times X'. \tag{2.4.6}$$

2.4.2 The minimal residual method of Andreev

In [And12, And13, And16], R. Andreev looked at the properties of minimal residual (MR) solutions of this parabolic problem and its efficient solution. Take some family $(X^\delta, Y^\delta)_{\delta\in\Delta}$ of subspaces of $X \times Y$. In current notation, the MR method amounts to finding

$$\operatorname*{arg\,min}_{w^\delta \in X^\delta} \|Bw^\delta - g\|^2_{Y^{\delta\prime}} + \|\gamma_0 w^\delta - u_0\|^2.$$

Our abstract introduction shows that MR solutions are *uniformly quasi-optimal* exactly when the family $(X^\delta, Y^\delta)_{\delta\in\Delta}$ is *uniformly inf-sup stable*, i.e.

$$\inf_{\delta\in\Delta} \inf_{w^\delta \in X^\delta} \frac{\|Bw^\delta\|^2_{Y^{\delta\prime}} + \|\gamma_0 w^\delta\|^2}{\|Bw^\delta\|^2_{Y'} + \|\gamma_0 w^\delta\|^2} > 0. \tag{2.4.7}$$

Andreev was able to prove the following important result that reduces inf-sup stability of the entire operator (B, γ_0) to that of the time-derivative ∂_t only.

Theorem ([And13, Thm. 4.1])**.** *Let $X^\delta \subseteq Y^\delta$. When A is symmetric and*

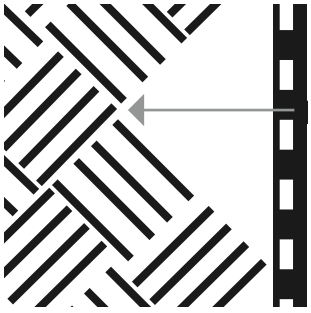
$$\gamma_\Delta^{\partial_t} := \inf_{\delta\in\Delta} \inf_{w^\delta \in X^\delta} \frac{\|\partial_t w^\delta\|_{Y^{\delta\prime}}}{\|\partial_t w^\delta\|_{Y'}} > 0, \tag{2.4.8}$$

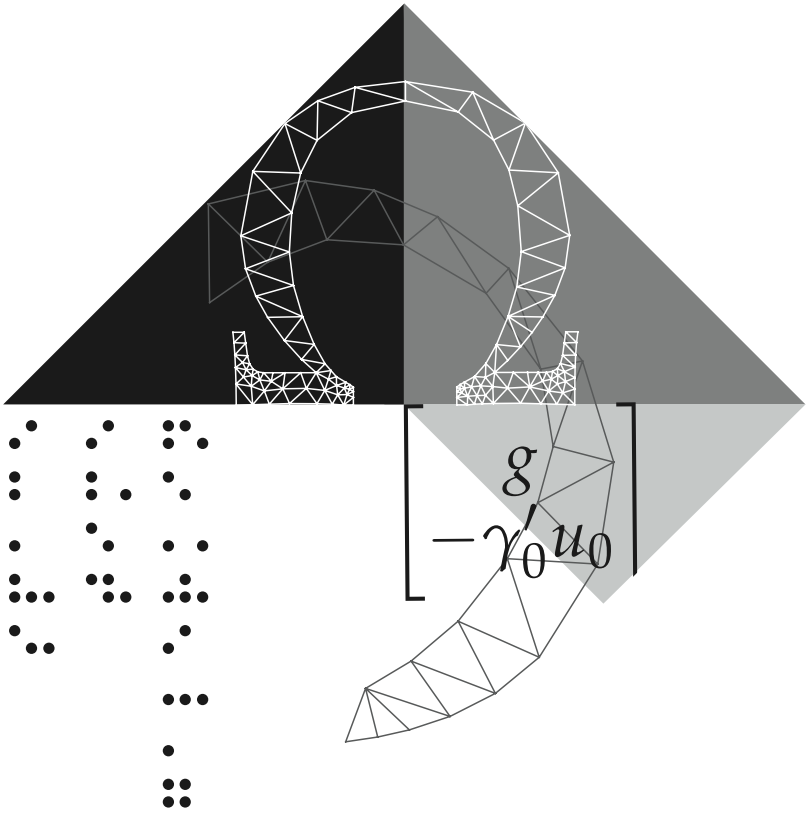
then the family $(X^\delta, Y^\delta)_{\delta\in\Delta}$ is uniformly inf-sup stable, i.e., (2.4.7) *holds.*

In Chapter 4, we show that the symmetry condition on A can be dropped.

Stable trial- and test spaces It is clear by now that the main game is to construct, for every trial space X^δ, a test space $Y^\delta \subset Y$ that guarantees uniform inf-sup stability of the time derivative (2.4.8). To ensure that the resulting linear system can be solved efficiently, we aim for a test space with dimension proportional to its trial space, i.e. $\dim Y^\delta \lesssim \dim X^\delta$ uniformly in $\delta \in \Delta$.

A full explanation is outside the scope of this introduction. We give an overview of families of stable trial- and test spaces in §4.5.




g
$-\gamma_0' u_0$

3 Stable space-time Galerkin discretizations of parabolic PDEs

Abstract We analyze Galerkin discretizations of a new well-posed mixed space-time variational formulation of parabolic PDEs. For suitable pairs of finite element trial spaces, the resulting Galerkin operators are shown to be uniformly stable. The method is compared to two related space-time discretization methods introduced in [*IMA J. Numer. Anal.*, 33(1) (2013), pp. 242–260] by R. Andreev and in [*Comput. Methods Appl. Math.*, 15(4) (2015), pp. 551–566] by O. Steinbach.

3.1 Introduction

In recent years one witnesses a rapidly growing interest in simultaneous space-time methods for solving parabolic evolution equations originally introduced in [BJ89, BJ90], see e.g. [GK11, And13, UP13, Ste15, GN16, LMN16, SS17, DS18, NS19, RS19, VR18, SZ20, FK21]. Compared to classical time marching methods, space-time methods are much better suited for a massively parallel implementation, and have the potential to drive adaptivity simultaneously in space and time.

Apart from the first order system least squares formulation recently introduced in [FK21], the known well-posed simultaneous space-time variational formulations of parabolic equations in terms of partial differential operators only, so not involving non-local operators, are not coercive. As a consequence, it is non-trivial to find families of pairs of discrete trial- and test-spaces for which the resulting Petrov–Galerkin discretizations are uniformly stable. The latter is a sufficient and, as we will see, necessary condition for the Petrov–Galerkin approximations to be *quasi-optimal*, i.e., to yield an up to a constant factor best approximation to the solution from the trial space. This concept has to be contrasted to rate optimality that, for quasi-uniform temporal and spatial partitions, has been shown for any reasonable numerical scheme under the assumption of sufficient regularity of the solution.

If one allows different spatial meshes at different times, then for the classical time marching schemes quasi-optimality of the numerical approximations

This chapter is a minor modification of **Stability of Galerkin discretizations of a mixed space-time variational formulation of parabolic evolution equations**, R. Stevenson and J. Westerdiep, *IMA Journal of Numerical Analysis*, 41(1):28–47, 2021.

is known not to be guaranteed as demonstrated in [Dup82, Sect. 4].

In view of the difficulty in constructing stable pairs of trial- and test-spaces, in [And13] Andreev considered minimal residual Petrov–Galerkin discretizations. They have an equivalent interpretation as Galerkin discretizations of an extended self-adjoint mixed system, with the Riesz lift of the residual of the primal variable being the secondary variable. This is our point of view.

A different path was followed by Steinbach in [Ste15]. Assuming a homogenous initial condition, for equal test and trial finite element spaces w.r.t. fully general finite element meshes, there stability was shown w.r.t. a weaker mesh-dependent norm on the trial space. As we will see, however, this has the consequence that for some solutions of the parabolic problem these Galerkin approximations are far from being quasi-optimal w.r.t. the (mesh-independent) natural norm on the trial space.

In the current work, we modify Andreev's approach by considering an equivalent but simpler mixed system that we construct from a space-time variational formulation that follows from applying the Brézis–Ekeland–Nayroles principle [BE76, Nay76]. With the same trial space for the primal variable, we show stability of the Galerkin discretization of this mixed system whilst utilizing a smaller trial space for the secondary variable. In addition, the stiffness matrix resulting from this mixed system is more sparse. In experiments, the errors in the Galerkin solutions are nevertheless very comparable.

3.1.1 Organization

In Sect. 3.2 we derive the two self-adjoint mixed system formulations of the parabolic problem that are central in this work. In Sect. 3.3 we give sufficient conditions for stability of Galerkin discretizations for both systems. We provide an a priori error bound for the Galerkin discretization of the newly introduced system, and improved a priori error bounds for the methods from [And13] and [Ste15]. In Sect 3.4, we show that the crucial condition for stability (being the only condition for the newly introduced mixed system) is satisfied for prismatic space-time finite elements whenever the generally non-uniform partition in time is independent of the spatial location, and the generally non-uniform spatial mesh in each time slab is such that the corresponding L_2-orthogonal projection is uniformly H^1-stable. In Sect. 3.5 we present some first simple numerical experiments for a one-dimensional spatial domain and uniform meshes. Conclusions are presented in Sect. 3.6.

3.1.2 Notations

In this work, by $C \lesssim D$ we will mean that C can be bounded by a multiple of D, independently of parameters which C and D may depend on. Obviously, $C \gtrsim D$ is defined as $D \lesssim C$, and $C \eqsim D$ as $C \lesssim D$ and $C \gtrsim D$.

For normed linear spaces E and F, by $\mathcal{L}(E,F)$ we will denote the normed linear space of bounded linear mappings $E \to F$, and by $\mathcal{L}\mathrm{is}(E,F)$ its subset of boundedly invertible linear mappings $E \to F$. We write $E \hookrightarrow F$ to denote

that E is continuously embedded into F. For simplicity only, we exclusively consider linear spaces over the scalar field $\mathbb{R}$.

For linear spaces E and F, sequences $\Phi = (\phi_j)_{j\in J} \subset E$, $\Psi = (\psi_i)_{i\in I} \subset F$, $f \in F'$, and a linear $A\colon E \to F'$, we define the column vector $f(\Psi) := [f(\psi_i)]_{i\in I}$ and matrix $(A\Phi)(\Psi) := [(A\phi_j)(\psi_i)]_{i\in I, j\in J}$. If $E = F$ is an inner product space, then with $R\colon E \to E'$ denoting the Riesz map, we set $\langle \Psi, \Phi\rangle := (R\Phi)(\Psi) = [(R\phi_j)(\psi_i)]_{i\in I, j\in J} = [\langle \psi_i, \phi_j\rangle]_{i\in I, j\in J}$.

3.2 Space-time formulations of the parabolic evolution problem

Let V, H be separable Hilbert spaces of functions on some "spatial domain" such that $V \hookrightarrow H$ with dense and compact embedding. Identifying H with its dual, we obtain the Gelfand triple $V \hookrightarrow H \simeq H' \hookrightarrow V'$.

We use the notation $\langle\cdot,\cdot\rangle$ to denote both the scalar product on $H \times H$, and its unique extension by continuity to the duality pairing on $V' \times V$. Correspondingly, the norm on H will be denoted by $\|\cdot\|$.

For a.e.

$$t \in I := (0, T),$$

let $a(t;\cdot,\cdot)$ denote a bilinear form on $V \times V$ such that for any $\eta, \zeta \in V$, $t \mapsto a(t;\eta,\zeta)$ is measurable on I, and such that for a.e. $t \in I$,

$$|a(t;\eta,\zeta)| \lesssim \|\eta\|_V \|\zeta\|_V \quad (\eta,\zeta \in V) \quad (\textit{boundedness}), \tag{3.2.1}$$

$$a(t;\eta,\eta) \gtrsim \|\eta\|_V^2 \quad (\eta \in V) \quad (\textit{coercivity}). \tag{3.2.2}$$

With $A(t) \in \mathcal{L}\mathrm{is}(V, V')$ being defined by $(A(t)\eta)(\zeta) = a(t;\eta,\zeta)$, we are interested in solving the *parabolic initial value problem* to finding u such that

$$\begin{cases} \frac{\mathrm{d}u}{\mathrm{d}t}(t) + A(t)u(t) = g(t) & (t \in I), \\ u(0) = u_0. \end{cases} \tag{3.2.3}$$

Remark 3.2.1. With $\tilde{u}(t) := u(t)e^{-\varrho t}$, (3.2.3) is equivalent to $\frac{d\tilde{u}}{dt}(t) + (A(t) + \varrho\mathrm{Id})\tilde{u}(t) = g(t)e^{-\varrho t}$ $(t \in I)$, $\tilde{u}(0) = u_0$. So if initially $a(t;\eta,\eta)$ is not coercive but only satisfies a *Gårding inequality* $a(t;\eta,\eta) + \varrho\langle\eta,\eta\rangle \gtrsim \|\eta\|_V^2$ $(\eta \in V)$, then one can consider a transformed problem such that (3.2.2) is valid. ◊

In a simultaneous space-time variational formulation, the parabolic PDE reads as finding u from a suitable space of functions of time and space s.t.

$$(Bw)(v) := \int_I \langle \tfrac{\mathrm{d}w}{\mathrm{d}t}(t), v(t)\rangle + a(t; w(t), v(t))\,\mathrm{d}t = \int_I \langle g(t), v(t)\rangle =: g(v) \tag{3.2.4}$$

for all v from another suitable space of functions of time and space. One possibility to enforce the initial condition is by testing it against additional test functions. A proof of the following result can be found in [SS09], cf. [DL92, Ch.XVIII, §3] and [Wlo82, Ch. IV, §26] for slightly different statements.

Theorem 3.2.2. *With* $X := L_2(I;V) \cap H^1(I;V')$, $Y := L_2(I;V)$, *under conditions* (3.2.1) *and* (3.2.2) *it holds that*

$$\begin{bmatrix} B \\ \gamma_0 \end{bmatrix} \in \mathcal{L}\mathrm{is}(X, Y' \times H), \tag{3.2.5}$$

where for $t \in \bar{I}$, $\gamma_t \colon u \mapsto u(t,\cdot)$ *denotes the trace map. That is, assuming* $g \in Y'$ *and* $u_0 \in H$, *finding* $u \in X$ *such that*

$$(Bu)(v_1) + \langle u(0,\cdot), v_2\rangle = g(v_1) + \langle u_0, v_2\rangle \quad ((v_1,v_2) \in Y \times H), \tag{3.2.6}$$

is a well-posed variational formulation of (3.2.3).

One ingredient of the proof of this theorem is the continuity of the embedding $X \hookrightarrow C(\bar{I}, H)$, in particular implying that for any $t \in \bar{I}$, $\gamma_t \in \mathcal{L}(X,H)$.

Defining $A, A_s \in \mathcal{L}\mathrm{is}(Y,Y')$ (here (3.2.2) is used), $A_a \in \mathcal{L}(Y,Y')$, and $C, \partial_t \in \mathcal{L}(X,Y')$ by

$$(Au)(v) := \int_I a(t;u(t),v(t))\,\mathrm{d}t, \quad A_s := \tfrac{1}{2}(A+A'), \quad A_a := \tfrac{1}{2}(A-A'),$$
$$C := B - A_s, \quad \partial_t := B - A,$$

an equivalent well-posed variational formulation of the parabolic PDE is obtained by applying the so-called Brézis–Ekeland–Nayroles variational principle [BE76, Nay76], cf. also [And12, §3.2.4]. It reads as

$$(C'A_s^{-1}C + A_s + \gamma_T'\gamma_T)u = (\mathrm{Id} + C'A_s^{-1})g + \gamma_0'u_0, \tag{3.2.7}$$

where the operator on the left is in $\mathcal{L}\mathrm{is}(X,X')$, is self-adjoint and coercive.

We provide a direct proof of these facts. Since $\begin{bmatrix} A_s & 0 \\ 0 & \mathrm{Id} \end{bmatrix} \in \mathcal{L}\mathrm{is}(Y \times H, Y' \times H)$, an equivalent formulation of (3.2.5) as a self-adjoint saddle point equation reads as finding $(\mu,\sigma,u) \in Y \times H \times X$ (where μ and σ will be zero) such that

$$\begin{bmatrix} A_s & 0 & B \\ 0 & \mathrm{Id} & \gamma_0 \\ B' & \gamma_0' & 0 \end{bmatrix} \begin{bmatrix} \mu \\ \sigma \\ u \end{bmatrix} = \begin{bmatrix} g \\ u_0 \\ 0 \end{bmatrix}, \tag{3.2.8}$$

or

$$(B'A_s^{-1}B + \gamma_0'\gamma_0)u = B'A_s^{-1}g + \gamma_0'u_0. \tag{3.2.9}$$

Thanks to (3.2.5), this Schur complement $B'A_s^{-1}B + \gamma_0'\gamma_0$ is in $\mathcal{L}\mathrm{is}(X,X')$, is self-adjoint and coercive.

We show that (3.2.9) and (3.2.7) are equal. Recalling the definitions of C and ∂_t, note that the right-hand sides of both equations are the same, and that

$$\begin{aligned} B'A_s^{-1}B + \gamma_0'\gamma_0 &= C'A_s^{-1}C + A_s + C + C' + \gamma_0'\gamma_0 \\ &= C'A_s^{-1}C + A_s + \partial_t + \partial_t' + \gamma_0'\gamma_0 \end{aligned}$$

thanks to $A_a' = -A_a$. Our claim is proven after noting that for $w, v \in X$,

$$\begin{aligned}((\partial_t + \partial_t' + \gamma_0'\gamma_0)w)(v) &= \int_I \langle \tfrac{\mathrm{d}w}{\mathrm{d}t}(t), v(t)\rangle + \langle w(t), \tfrac{\mathrm{d}v}{\mathrm{d}t}(t)\rangle \,\mathrm{d}t + \langle w(0), v(0)\rangle \\ &= \int_I \tfrac{\mathrm{d}}{\mathrm{d}t}\langle w(t), v(t)\rangle \,\mathrm{d}t + \langle w(0), v(0)\rangle = (\gamma_T'\gamma_T w)(v).\end{aligned}$$

As (3.2.9) was obtained as the Schur complement equation of (3.2.8), in its form (3.2.7) it is naturally obtained as the Schur complement of the problem of finding $(\lambda, u) \in Y \times X$ such that

$$\begin{bmatrix} A_s & C \\ C' & -(A_s + \gamma_T'\gamma_T) \end{bmatrix} \begin{bmatrix} \lambda \\ u \end{bmatrix} = \begin{bmatrix} g \\ -(g + \gamma_0' u_0) \end{bmatrix}. \tag{3.2.10}$$

Knowing that its Schur complement is in $\mathcal{L}\mathrm{is}(X, X')$, $A_s \in \mathcal{L}\mathrm{is}(Y, Y')$, and $C \in \mathcal{L}(X, Y')$, we infer that the self-adjoint operator at the left hand side of (3.2.10) is in $\mathcal{L}\mathrm{is}(Y \times X, Y' \times X')$.

Substituting $C = B - A_s$ and $Bu = g$, we find the secondary variable to be

$$\lambda = u.$$

Remark 3.2.3. When reading $\gamma_T'\gamma_T$ as $\partial_t + \partial_t' + \gamma_0'\gamma_0$, the system (3.2.10) has remarkable similarities to a certain preconditioned version presented in [NS19] of a discretized parabolic PDE using the implicit Euler method in time. Ideas concerning optimal preconditioning developed in that paper, as well as those in [And16], can be expected to work for Galerkin discretizations of (3.2.10). ♢

Remark 3.2.4. In equations (3.2.8) and (3.2.9), the operator A_s can be replaced by a general self-adjoint $\tilde{A}_s \in \mathcal{L}\mathrm{is}(Y, Y')$. With $\tilde{C} := B - \tilde{A}_s$, the equivalent equation (3.2.7) then reads as

$$(\tilde{C}'\tilde{A}_s^{-1}\tilde{C} + 2A_s - \tilde{A}_s + \gamma_T'\gamma_T)u = (\mathrm{Id} + \tilde{C}'\tilde{A}_s^{-1})g + \gamma_0' u_0,$$

and (3.2.10) as

$$\begin{bmatrix} \tilde{A}_s & \tilde{C} \\ \tilde{C}' & -(2A_s - \tilde{A}_s + \gamma_T'\gamma_T) \end{bmatrix} \begin{bmatrix} \lambda \\ u \end{bmatrix} = \begin{bmatrix} g \\ -(g + \gamma_0' u_0) \end{bmatrix},$$

with solution $\lambda = u$. ♢

In the next section, we study Galerkin discretizations of equations (3.2.8) and (3.2.10), which then are no longer equivalent.

Since the secondary variables μ and σ in (3.2.8) are zero, the subspaces for their approximation do not have to satisfy any approximation properties. Since the secondary variable λ in (3.2.10) is non-zero, the subspace of Y for its approximation has to satisfy approximation properties, and the error in its best approximation enters the upper bound for that of its primal variable.

On the other hand, (uniform) stability will be easier to realize with equation (3.2.10) and will also be proven to hold true for $A_a \neq 0$; the system matrix will be more sparse; and the number of unknowns will be smaller.

In order to facilitate the derivation of some quantitative results, we will equip the spaces Y and X with the 'energy-norms' defined by

$$\|v\|_Y^2 := (A_s v)(v), \quad \|u\|_X^2 := \|u\|_Y^2 + \|\partial_t u\|_{Y'}^2 + \|u(T)\|^2,$$

which are equivalent to the standard norms on these spaces. Correspondingly, we measure orthogonality in Y w.r.t. the 'energy scalar product' $(A_s \cdot)(\cdot)$.

3.3 Stable discretizations of the parabolic problem

3.3.1 Uniformly stable Petrov–Galerkin discretizations, quasi-best approximations

This subsection is devoted to proving the following theorem.

Theorem 3.3.1. *Let W and Z be Hilbert spaces, and $F \in \mathcal{L}\mathrm{is}(Z, W')$. Let $(W^\delta, Z^\delta)_{\delta\in\Delta}$ be a family of closed subspaces of $W \times Z$ such that for each $\delta \in \Delta$ it holds that $E_W^{\delta\,\prime} F E_Z^\delta \in \mathcal{L}\mathrm{is}(Z^\delta, W^{\delta\,\prime})$, where $E_W^\delta \colon W^\delta \to W$, $E_Z^\delta \colon Z^\delta \to Z$ denote the trivial embeddings. Then the collection $(z^\delta)_{\delta\in\Delta}$ of Petrov–Galerkin approximations to $z \in Z$, determined by $E_W^{\delta\,\prime} F E_Z^\delta z^\delta = E_W^{\delta\,\prime} F z$, is* quasi-optimal, *i.e. $\|z - z^\delta\|_Z \lesssim \inf_{0\neq\bar z^\delta\in Z} \|z - \bar z^\delta\|_Z$, uniformly in $z \in Z$ and $\delta \in \Delta$, iff*

$$\inf_{\delta\in\Delta} \inf_{0\neq z\in Z^\delta} \sup_{0\neq w\in W^\delta} \frac{|(Fz)(w)|}{\|z\|_Z \|w\|_W} > 0 \qquad \text{(uniform stability)}.$$

Proof. The mapping $P^\delta := z \mapsto z^\delta = E_Z^\delta (E_W^{\delta\,\prime} F E_Z^\delta)^{-1} E_W^{\delta\,\prime} F z$ is a projector. For $\{0\} \subsetneq Z^\delta \subsetneq Z$, it holds that $P^\delta \notin \{0, \mathrm{Id}\}$, and consequently, $\|\mathrm{Id} - P^\delta\|_{\mathcal{L}(Z,Z)} = \|P^\delta\|_{\mathcal{L}(Z,Z)}$ (see [Kat60, XZ03]). We obtain that

$$\begin{aligned} \sup_{z\in Z\setminus Z^\delta} \frac{\|z - z^\delta\|_Z}{\inf_{\bar z^\delta\in Z^\delta} \|z - \bar z^\delta\|_Z} &= \sup_{z\in Z\setminus Z^\delta} \sup_{\bar z^\delta\in Z^\delta} \frac{\|(I - P^\delta)z\|_Z}{\|z - \bar z^\delta\|_Z} \\ &= \sup_{0\neq\bar z\in Z} \frac{\|(I - P^\delta)\bar z\|_Z}{\|\bar z\|_Z} = \|P^\delta\|_{\mathcal{L}(Z,Z)}. \end{aligned} \tag{3.3.1}$$

It remains to show uniform boundedness of $\|P^\delta\|_{\mathcal{L}(Z,Z)}$ if and only if uniform stability is valid.

The definition of P^δ shows that

$$\|F^{-1}\|_{\mathcal{L}(W',Z)}^{-1} \leq \frac{\|P^\delta\|_{\mathcal{L}(Z,Z)}}{\|E_Z^\delta (E_W^{\delta\,\prime} F E_Z^\delta)^{-1} E_W^{\delta\,\prime}\|_{\mathcal{L}(W',Z)}} \leq \|F\|_{\mathcal{L}(Z,W')}.$$

Further, we have that

$$\begin{aligned} \|E_Z^\delta (E_W^{\delta\,\prime} F E_Z^\delta)^{-1} E_W^{\delta\,\prime}\|_{\mathcal{L}(W',Z)} &= \|(E_W^{\delta\,\prime} F E_Z^\delta)^{-1} E_W^{\delta\,\prime}\|_{\mathcal{L}(W',Z^\delta)} \\ &= \|(E_W^{\delta\,\prime} F E_Z^\delta)^{-1}\|_{\mathcal{L}(W^{\delta\prime},Z^\delta)} \end{aligned}$$

where the last equality follows from $\|{E_W^\delta}'\|_{\mathcal{L}(W',W^{\delta\prime})} \leq 1$ and, for the other direction, from the fact that for given $f^\delta \in W^{\delta\prime}$ the function $f \in W'$ defined by $f|_{W^\delta} := f^\delta$ and $f|_{(W^\delta)^\perp} := 0$ satisfies $\|f\|_{W'} = \|f^\delta\|_{W^{\delta\prime}}$ and $f^\delta = {E_W^\delta}' f$.

Then seeing that

$$\|({E_W^\delta}' F E_Z^\delta)^{-1}\|^{-1}_{\mathcal{L}(W^{\delta\prime},Z^\delta)} = \inf_{0\neq z\in Z^\delta} \sup_{0\neq w\in W^\delta} \frac{|(Fz)(w)|}{\|z\|_Z \|w\|_W} \tag{3.3.2}$$

completes the proof. □

Remark 3.3.2. In particular above analysis provides a short self-contained proof of the quantitative results

$$\|F^{-1}\|^{-1}_{\mathcal{L}(W',Z)} \leq \frac{\sup_{z\in Z\setminus Z^\delta} \frac{\|z-z^\delta\|_Z}{\inf_{\bar z^\delta\in Z^\delta}\|z-\bar z^\delta\|_Z}}{\inf_{0\neq z\in Z^\delta}\sup_{0\neq w\in W^\delta}\frac{|(Fz)(w)|}{\|z\|_Z\|w\|_W}} \leq \|F\|_{\mathcal{L}(Z,W')},$$

that were established earlier in [TV16, §2.1, in particular (2.12)]. ◊

3.3.2 Uniformly stable Galerkin discretizations of (3.2.10)

Let $Y^\delta \times X^\delta$ be a closed subspace of $Y \times X$, and let $E_Y^\delta \colon Y^\delta \to Y$ and $E_X^\delta \colon X^\delta \to X$ denote the trivial embeddings. Since ${E_Y^\delta}' A_s E_Y^\delta \in \mathcal{L}\mathrm{is}(Y^\delta, Y^{\delta\prime})$ (as well as being an isometry), the Galerkin operator resulting from (3.2.10) can be factorized as

$$\begin{aligned}
&\begin{bmatrix} {E_Y^\delta}' A_s E_Y^\delta & {E_Y^\delta}' C E_X^\delta \\ ({E_Y^\delta}' C E_X^\delta)' & -{E_X^\delta}'(A_s+\gamma_T'\gamma_T)E_X^\delta \end{bmatrix} = \\
&\begin{bmatrix} \mathrm{Id} & 0 \\ ({E_Y^\delta}' C E_X^\delta)'({E_Y^\delta}' A_s E_Y^\delta)^{-1} & \mathrm{Id} \end{bmatrix} \circ \\
&\begin{bmatrix} {E_Y^\delta}' A_s E_Y^\delta & 0 \\ 0 & -{E_X^\delta}'(A_s+\gamma_T'\gamma_T)E_X^\delta - ({E_Y^\delta}' C E_X^\delta)'({E_Y^\delta}' A_s E_Y^\delta)^{-1}{E_Y^\delta}' C E_X^\delta \end{bmatrix} \circ \\
&\begin{bmatrix} \mathrm{Id} & ({E_Y^\delta}' A_s E_Y^\delta)^{-1}{E_Y^\delta}' C E_X^\delta \\ 0 & \mathrm{Id} \end{bmatrix}.
\end{aligned} \tag{3.3.3}$$

We conclude that this Galerkin operator is invertible if and only if the Schur complement

$${E_X^\delta}'(A_s+\gamma_T'\gamma_T)E_X^\delta + ({E_Y^\delta}' C E_X^\delta)'({E_Y^\delta}' A_s E_Y^\delta)^{-1}{E_Y^\delta}' C E_X^\delta \tag{3.3.4}$$

is invertible, which holds true for any $X^\delta \neq \{0\}$.

Theorem 3.3.3. *Let $(Y^\delta, X^\delta)_{\delta\in\Delta}$ be a family of closed subspaces of $Y \times X$ such that*

$$\gamma_\Delta := \inf_{\delta\in\Delta} \inf_{\{u\in X^\delta\colon \partial_t u\neq 0\}} \sup_{0\neq v\in Y^\delta} \frac{(\partial_t u)(v)}{\|\partial_t u\|_{Y'}\|v\|_Y} > 0.^1 \tag{3.3.5}$$

Let $\rho = \rho_\Delta$ be the root in $[0,1)$ of

$$\gamma_\Delta^2(\rho^2-\rho) + \|A_a\|^2_{\mathcal{L}(Y,Y')}(\rho-1) + \rho = 0,$$

and let

$$C_\Delta := \frac{(3+\|A_a\|^2_{\mathcal{L}(Y,Y')})(\sqrt{3}+\|A_a\|_{\mathcal{L}(Y,Y')})}{(1-\rho_\Delta)\gamma_\Delta^2},$$

so that $C_\Delta = 3\sqrt{3}\,\gamma_\Delta^{-2}$ when $\|A_a\|_{\mathcal{L}(Y,Y')} = 0$, and $\lim_{\|A_a\|_{\mathcal{L}(Y,Y')}\to\infty} C_\Delta = \infty$. Then with $\lambda = u$ and $(\lambda^\delta, u^\delta)$ denoting the solutions of (3.2.10) *and its Galerkin discretization, respectively, it holds that*

$$\sqrt{\|\lambda-\lambda^\delta\|_Y^2 + \|u-u^\delta\|_X^2} \le C_\Delta \inf_{(\bar\lambda^\delta,\bar u^\delta)\in Y^\delta\times X^\delta} \sqrt{\|\lambda-\bar\lambda^\delta\|_Y^2 + \|u-\bar u^\delta\|_X^2}. \tag{3.3.6}$$

Proof. In view of the second inequality presented in Remark 3.3.2, we start with bounding the norm of the continuous operator. Using Young's inequality, for $(\lambda, u) \in Y \times X$ we have

$$\begin{aligned}
&\|A_s\lambda + \partial_t u\|^2_{Y'} + \|\partial_t'\lambda - (A_s + \gamma_T'\gamma_T)u\|^2_{X'} \\
&\le \tfrac{3}{2}\|A_s\lambda\|^2_{Y'} + 3\|\partial_t u\|^2_{Y'} + \tfrac{3}{2}\|\partial_t'\lambda\|^2_{X'} + 3\|(A_s+\gamma_T'\gamma_T)u\|^2_{X'} \\
&\le \tfrac{3}{2}(\|\lambda\|_Y^2 + \|\lambda\|_Y^2) + 3(\|\partial_t u\|^2_{Y'} + \|u\|_Y^2 + \|u(T)\|^2) = 3(\|\lambda\|_Y^2 + \|u\|_X^2).
\end{aligned}$$

Together with $\|A_a u\|^2_{Y'} + \|A_a'\lambda\|^2_{X'} \le \|A_a\|^2_{\mathcal{L}(Y,Y')}(\|\lambda\|_Y^2 + \|u\|_X^2)$, it shows that

$$\begin{aligned}
&\left\| \begin{bmatrix} A_s & C \\ C' & -(A_s+\gamma_T'\gamma_T) \end{bmatrix} \right\|_{\mathcal{L}(Y\times X, Y'\times X')} \\
&\le \left\| \begin{bmatrix} A_s & \partial_t \\ \partial_t' & -(A_s+\gamma_T'\gamma_T) \end{bmatrix} \right\|_{\mathcal{L}(Y\times X, Y'\times X')} + \left\| \begin{bmatrix} 0 & A_a \\ A_a' & 0 \end{bmatrix} \right\|_{\mathcal{L}(Y\times X, Y'\times X')} \\
&\le \sqrt{3} + \|A_a\|_{\mathcal{L}(Y,Y')}.
\end{aligned}$$

To bound, in view of (3.3.2), the norm of the inverse of the Galerkin operator, we use the block-LDU factorization (3.3.3). With $r := (1 + \|A_a\|^2_{\mathcal{L}(Y,Y')})$, for $u \in X$ it holds that

$$\|Cu\|_{Y'} \le \|\partial_t u\|_{Y'} + \|A_a\|_{\mathcal{L}(Y,Y')}\|u\|_Y \le \sqrt{r}\,\|u\|_X.$$

[1] Here and in the following, $\inf_{\{u\in X^\delta\colon \partial_t u\neq 0\}} \sup_{0\neq v\in Y^\delta} \frac{(\partial_t u)(v)}{\|\partial_t u\|_{Y'}\|v\|_Y}$ should be read as 1 in the case that $\{u \in X^\delta\colon \partial_t u \neq 0\} = \varnothing$.

Together with the fact that $E_Y^{\delta\prime} A_s E_Y^\delta \in \mathcal{L}\mathrm{is}(Y^\delta, Y^{\delta\prime})$ is an isometry and again Young's inequality, it shows that for $(\lambda, u) \in Y^\delta \times X^\delta$,

$$\begin{aligned}\|\lambda - (E_Y^{\delta\prime} A_s E_Y^\delta)^{-1} E_Y^{\delta\prime} C E_X^\delta u\|_Y^2 + \|u\|_X^2 \\ \leq (1+r)\|\lambda\|_Y^2 + (1+r^{-1})r\|u\|_X^2 + \|u\|_X^2 \\ \leq (2+r)(\|\lambda\|_Y^2 + \|u\|_X^2),\end{aligned}$$

or

$$\left\| \begin{bmatrix} \mathrm{Id} & (E_Y^{\delta\prime} A_s E_Y^\delta)^{-1} E_Y^{\delta\prime} C E_X^\delta \\ 0 & \mathrm{Id} \end{bmatrix}^{-1} \right\|_{\mathcal{L}(Y^\delta \times X^\delta, Y^\delta \times X^\delta)} \leq \sqrt{3 + \|A_a\|^2_{\mathcal{L}(Y,Y')}}.$$

Obviously, the $\mathcal{L}(Y^{\delta\prime} \times X^{\delta\prime}, Y^{\delta\prime} \times X^{\delta\prime})$-norm of the inverse of the first factor at the right-hand side of (3.3.3) satisfies the same bound.

For the 2nd factor, we consider the Schur complement operator. From $(E_Y^{\delta\prime} A_s E_Y^\delta \lambda)(\lambda) = \|\lambda\|_Y^2$ for $\lambda \in Y^\delta$, we have for $f \in Y^{\delta\prime}$, $f((E_Y^{\delta\prime} A_s E_Y^\delta)^{-1} f) = \|(E_Y^{\delta\prime} A_s E_Y^\delta)^{-1} f\|_Y^2 = \|f\|^2_{Y^{\delta\prime}}$, and so for $u \in X^\delta$

$$\left((E_Y^{\delta\prime} C E_X^\delta)' (E_Y^{\delta\prime} A_s E_Y^\delta)^{-1} E_Y^{\delta\prime} C E_X^\delta u \right)(u) = \|E_Y^{\delta\prime} C E_X^\delta u\|^2_{Y^{\delta\prime}}.$$

Using that for $u \in X^\delta$,

$$\|E_Y^{\delta\prime} \partial_t E_X^\delta u\|^2_{Y^{\delta\prime}} = \left(\sup_{0 \neq v \in Y^\delta} \frac{(\partial_t u)(v)}{\|v\|_Y} \right)^2 \geq \gamma_\Delta^2 \|\partial_t u\|^2_{Y'}$$

and

$$\|E_Y^{\delta\prime} A_a E_X^\delta u\|^2_{Y^{\delta\prime}} \leq \|A_a\|^2_{\mathcal{L}(Y,Y')} \|u\|_Y^2,$$

Young's inequality shows that

$$\|E_Y^{\delta\prime} C E_X^\delta u\|^2_{Y^{\delta\prime}} \geq (1-\rho_\Delta)\gamma_\Delta^2 \|\partial_t u\|^2_{Y'} + (1-\rho_\Delta^{-1}) \|A_a\|^2_{\mathcal{L}(Y,Y')} \|u\|_Y^2,$$

where we assumed that $\rho_\Delta > 0$ i.e. $A_a \neq 0$. It follows that

$$\begin{aligned}((A_s + \gamma_T' \gamma_T)u)(u) + \|E_Y^{\delta\prime} C E_X^\delta u\|^2_{Y^{\delta\prime}} \\ \geq (1 + (1-\rho_\Delta^{-1})\|A_a\|^2_{\mathcal{L}(Y,Y')})\|u\|_Y^2 + \|u(T)\|^2 + (1-\rho_\Delta)\gamma_\Delta^2 \|\partial_t u\|^2_{Y'} \\ \geq (1-\rho_\Delta)\gamma_\Delta^2 \|u\|_X^2 \qquad (3.3.7)\end{aligned}$$

where we used that $1 + (1-\rho_\Delta^{-1})\|A_a\|^2_{\mathcal{L}(Y,Y')} = (1-\rho_\Delta)\gamma_\Delta^2$ by definition of ρ_Δ. One easily verifies (3.3.7) also in the case that $A_a = 0$ i.e. $\rho_\Delta = 0$.

Since $E_Y^{\delta\prime} A_s E_Y^\delta \in \mathcal{L}\mathrm{is}(Y^\delta, Y^{\delta\prime})$ is an isometry, and $0 < (1-\rho_\Delta)\gamma_\Delta^2 \leq \gamma_\Delta^2 \leq 1$, we conclude that the $\mathcal{L}(Y^{\delta\prime} \times X^{\delta\prime}, Y^\delta \times X^\delta)$-norm of the inverse of the second factor is bounded by $(1-\rho_\Delta)^{-1}\gamma_\Delta^{-2}$.

In view of the second inequality presented in Remark 3.3.2 in combination with (3.3.2), the proof is completed by collecting the derived bounds. □

3.3.3 Galerkin discretizations of (3.2.8)

In Chapter 4, we generalize results to the case of $A_a \neq 0$ but as in [And13, Ste15], in this section we operate under the condition that

$$A = A_s. \tag{3.3.8}$$

Following [Ste15], for a given closed subspace $Y^\delta \subseteq Y$ we define the 'mesh-dependent' norm on X by

$$\|u\|^2_{X,Y^\delta} := \|u\|^2_Y + \sup_{0\neq v\in Y^\delta} \frac{(\partial_t u)(v)^2}{\|v\|^2_Y} + \|u(T)\|^2.$$

Note that $\|\cdot\|_{X,Y} = \|\cdot\|_X$.

The following result generalizes the 'inf-sup identity', known for $Y^\delta = Y$, see e.g. [ESV17], to mesh-dependent norms.

Lemma 3.3.4. *Assuming* (3.3.8), *then for* $u \in Y^\delta \cap X$,

$$\|u\|^2_{X,Y^\delta} = \sup_{0\neq v\in Y^\delta} \frac{(Bu)(v)^2}{\|v\|^2_Y} + \|u(0)\|^2.$$

If additionally $\gamma_0 u \in H^\delta$, *then*

$$\|u\|^2_{X,Y^\delta} = \sup_{0\neq (v_1,v_2)\in Y^\delta\times H^\delta} \frac{((Bu)(v_1) + \langle u(0), v_2\rangle)^2}{\|v_1\|^2_Y + \|v_2\|^2}. \tag{3.3.9}$$

Proof. Define $y \in Y^\delta$ by $(A_s y)(v) = (\partial_t u)(v)$ $(v \in Y^\delta)$. Then $(A_s y)(y) = \sup_{0\neq v\in Y^\delta} \frac{(\partial_t u)(v)^2}{\|v\|^2_Y}$. Furthermore, for $v \in Y^\delta$, $(Bu)(v) = (A_s(y+u))(v)$ and so, thanks to $u \in Y^\delta$,

$$\begin{aligned}\sup_{0\neq v\in Y^\delta} \frac{(Bu)(v)^2}{\|v\|^2_Y} &= (A_s(y+u))(y+u) = (A_s y)(y) + 2(A_s y)(u) + (A_s u)(u)\\ &= (A_s y)(y) + 2(\partial_t u)(u) + (A_s u)(u) = \|u\|^2_{X,Y^\delta} - \|u(0)\|^2\end{aligned}$$

where we used that $2\int_I \langle \partial_t u(t), u(t)\rangle \,\mathrm{d}t = \|u(T)\|^2 - \|u(0)\|^2$.

The second statement follows from

$$\sup_{0\neq (v_1,v_2)\in Y^\delta\times H^\delta} \frac{((A_s(y+u))(v_1) + \langle u(0), v_2\rangle)^2}{\|v_1\|^2_Y + \|v_2\|^2} = (A_s(y+u))(y+u) + \|u(0)\|^2,$$

thanks to $u(0) \in H^\delta$. □

The next theorem gives sufficient conditions for existence and uniqueness of solutions of the Galerkin discretization of (3.2.8), and provides a suboptimal error estimate.

Theorem 3.3.5. *Assuming* (3.3.8)*, for closed subspaces* $Y^\delta \times H^\delta \times X^\delta \subset Y \times H \times X$ *with* $X^\delta \subseteq Y^\delta$ *and* $\operatorname{ran} \gamma_0|_{X^\delta} \subseteq H^\delta$*, the Galerkin discretization of* (3.2.8) *has a unique solution* $(\mu^\delta, \sigma^\delta, u^\delta) \in Y^\delta \times H^\delta \times X^\delta$*, and with* u *the solution of* (3.2.6)*,*

$$\|u - u^\delta\|_{X,Y^\delta} \leq 2 \inf_{\bar{u}^\delta \in X^\delta} \|u - \bar{u}^\delta\|_X.$$

Proof. Thanks to the assumptions $X^\delta \subseteq Y^\delta$ and $\operatorname{ran} \gamma_0|_{X^\delta} \subseteq H^\delta$, the inf-sup identity (3.3.9) quarantees the unique solvability of the Galerkin system.

For any $u \in X^\delta$, there exist unique $y_u \in Y^\delta$, $h_u \in H^\delta$ such that

$$(A_s y_u)(v_1) + \langle h_u, v_2 \rangle = (Bu)(v_1) + \langle \gamma_0 u, v_2 \rangle \quad ((v_1, v_2) \in Y^\delta \times H^\delta).$$

We decompose $Y^\delta \times H^\delta$ into $Z^\delta := \operatorname{clos}\{(y_u, h_u) \colon u \in X^\delta\}$[2] and its orthogonal complement W^δ. Using that for any $u \in X^\delta$ and $(v_1, v_2) \in W^\delta$, $(Bu)(v_1) + \langle u(0), v_2 \rangle = 0$, one infers that for any $u \in X^\delta$, the inf-sup identity (3.3.9) remains valid when the supremum is restricted to $0 \neq (v_1, v_2) \in Z^\delta$. Furthermore, since for any $(v_1, v_2) \in Z^\delta$ there exists a $z \in X^\delta$ with $(Bz)(v_1) + \langle z(0), v_2 \rangle \neq 0$, we infer that u^δ is the unique solution of the Petrov–Galerkin discretization of finding $u^\delta \in X^\delta$ such that

$$(Bu^\delta)(v_1) + \langle u^\delta(0), v_2 \rangle = g(v_1) + \langle u_0, v_2 \rangle \quad ((v_1, v_2) \in Z^\delta). \tag{3.3.10}$$

By applying these observations consecutively, we infer that for any $\bar{u}^\delta \in X^\delta$,

$$\begin{aligned} \|u^\delta - \bar{u}^\delta\|^2_{X,Y^\delta} &= \sup_{0 \neq (v_1,v_2) \in Z^\delta} \frac{((B(u^\delta - \bar{u}^\delta))(v_1) + \langle u^\delta(0) - \bar{u}^\delta(0), v_2 \rangle)^2}{\|v_1\|_Y^2 + \|v_2\|^2} \\ &= \sup_{0 \neq (v_1,v_2) \in Z^\delta} \frac{((B(u - \bar{u}^\delta))(v_1) + \langle u(0) - \bar{u}^\delta(0), v_2 \rangle)^2}{\|v_1\|_Y^2 + \|v_2\|^2} \leq \|u - \bar{u}^\delta\|^2_X, \end{aligned} \tag{3.3.11}$$

where we again applied (3.3.9) now for $Y^\delta = Y$. A triangle-inequality completes the proof. □

Theorem 3.3.5 can be used to demonstrate optimal rates for the error in u^δ in the $\|\cdot\|_{X,Y^\delta}$-norm, and hence also in the Y-norm. Yet, for doing so one needs to control the error of best approximation in the generally strictly stronger $\|\cdot\|_X$-norm, which requires regularity conditions on the solution u that exceeds those that are needed for optimal rates of the best approximation in the $\|\cdot\|_{X,Y^\delta}$-norm. In other words, this theorem does not show that u^δ is a quasi-best approximation from X^δ in the $\|\cdot\|_{X,Y^\delta}$-norm, or in any other norm.

Remark 3.3.6. Theorem 3.3.5 provides a generalization, with an improved constant, of Steinbach's result [Ste15, Theorem 3.2]. There the case was considered that the initial value $u_0 = 0$, $\operatorname{ran} \gamma_0|_{X^\delta} = \{0\}$, $H^\delta = \{0\}$, and $Y^\delta =$

[2] In the (discontinuous) Petrov–Galerkin community, $Y^\delta \times H^\delta$ and Z^δ are known under the names test search space (or search test space), and projected optimal test space (or approximate optimal test space), respectively.

X^δ. In that case the Galerkin discretization of (3.2.8) means solving $u^\delta \in X^\delta$ from $(Bu^\delta)(v) = g(v)$ $(v \in X^\delta)$ (indeed, Z^δ in the proof of Theorem 3.3.5 is $X^\delta \times \{0\}$). So with this approach the forming of 'normal equations' as in (3.2.9) is avoided.

In case of an inhomogeneous initial value $u_0 \in H$, one may approximate the solution as $\bar{u} + w^\delta$, where $\bar{u} \in X$ is such that $\gamma_0 \bar{u} = u_0$, and $w^\delta \in X^\delta$ solves $(Bw^\delta)(v) = g(v) - (B\bar{u})(v)$ $(v \in X^\delta)$. Although such a $\bar{u} \in X$ always exists, its practical construction becomes inconvenient for $u_0 \notin V$. For $u_0 \in V$, $\bar{u}$ can be taken as its constant extension in time.

To investigate in the setting of [Ste15] the relation between the $\|\cdot\|_{X,Y^\delta}$- and $\|\cdot\|_X$-norms, we consider X^δ of the form $X_t^\delta \otimes X_x^\delta$, where X_t^δ is the space of continuous piecewise linears, zero at $t = 0$, w.r.t. a uniform partition of I with mesh-size $h_\delta = \frac{T}{2N_\delta}$ for some $N_\delta \in \mathbb{N}$, and $X_x^\delta \subset V$ with $\cap_{\delta\in\Delta} X_x^\delta \neq \{0\}$. Given $z^\delta \in X^\delta$, Lemma 3.3.4 shows that

$$\sup_{0\neq v\in X^\delta} \frac{|(Bz^\delta)(v)|}{\|z^\delta\|_X \|v\|_Y} = \frac{\|z^\delta\|_{X,Y^\delta}}{\|z^\delta\|_X}. \tag{3.3.12}$$

For some arbitrary, fixed $0 \neq z_x \in \cap_{\delta\in\Delta} X_x^\delta$, we take $z^\delta = z_t^\delta \otimes z_x \in X^\delta$, where $z_t^\delta \in X_t^\delta$ is defined by $\frac{d}{dt} z_t^\delta = (-1)^{i-1}$ on $[(i-1)h_\delta, ih_\delta]$. Since $z_t^\delta(0) = 0$, also $z_t^\delta(T) = 0$. We have $\|z_t^\delta\|_{L_2(I)} \eqsim h_\delta$, $\|\frac{dz_t^\delta}{dt}\|_{L_2(I)} \eqsim 1$, $\sup_{0\neq v\in Y} \frac{(\partial_t z^\delta)(v)}{\|v\|_Y} = \|\frac{dz_t^\delta}{dt}\|_{L_2(I)} \|z_x\|_{V'} \eqsim 1$, $\|z^\delta\|_Y = \|z_t^\delta\|_{L_2(I)} \|z_x\|_V \eqsim h_\delta$, and

$$\begin{aligned}\sup_{0\neq v\in X^\delta} \frac{(\partial_t z^\delta)(v)}{\|v\|_Y} &= \sup_{0\neq v\in X_t^\delta} \frac{\langle \frac{\mathrm{d}z_t^\delta}{\mathrm{d}t}, v\rangle_{L_2(I)}}{\|v\|_{L_2(I)}} \sup_{0\neq v\in X_x^\delta} \frac{\langle z_x, v\rangle}{\|v\|_V} \\ &\leq \sup_{0\neq v\in X_t^\delta} \frac{\langle \frac{\mathrm{d}z_t^\delta}{\mathrm{d}t}, v\rangle_{L_2(I)}}{\|v\|_{L_2(I)}} \|z_x\|_{V'}.\end{aligned}$$

Let us equip the space of piecewise constants w.r.t. the aforementioned uniform partition with the $L_2(I)$-normalized basis $\{\chi_i^\delta\}$ of characteristic functions of the subintervals, and X_t^δ with the set of nodal basis functions $\{\phi_i^\delta\}$ normalized so that their maximal value is $h_\delta^{-\frac{1}{2}}$. Then with $G := [\langle \chi_j, \phi_i\rangle_{L_2(I)}]_{ij} = \frac{1}{2}\begin{bmatrix} 1 & 1 & & \\ & \ddots & \ddots & \\ & & 1 & 1 \\ & & & 1 \end{bmatrix}$, and $\boldsymbol{x} := \sqrt{h_\delta}\,[(-1)^{i-1}]_{1\leq i\leq 2N_\delta}$, from the uniform $L_2(I)$-stability of $\{\phi_i^\delta\}$ one infers that

$$\sup_{0\neq v\in X_t^\delta} \frac{\langle \frac{\mathrm{d}z_t^\delta}{\mathrm{d}t}, v\rangle_{L_2(I)}}{\|v\|_{L_2(I)}} \eqsim \sup_{0\neq \boldsymbol{y}} \frac{\langle G\boldsymbol{x}, \boldsymbol{y}\rangle}{\|\boldsymbol{y}\|} = \|G\boldsymbol{x}\| = \frac{1}{2}\sqrt{h_\delta}.$$

By substituting these estimates in the right-hand side of (3.3.12), we find that its value is $\eqsim \sqrt{h_\delta}$, so that $\inf_{0\neq z^\delta\in X^\delta} \sup_{0\neq v\in X^\delta} \frac{|(Bz^\delta)(v)|}{\|z^\delta\|_X\|v\|_Y} \lesssim \sqrt{h_\delta}$. As follows

from the first inequality in Remark 3.3.2, this means that there exist solutions $u \in X$ of the parabolic problem for which the errors in X-norm in these Galerkin approximations from X^δ are a factor $\gtrsim h_\delta^{-\frac{1}{2}}$ larger than these errors in the best approximations from X^δ.

Numerical evidence provided by [Ste15, Table 6] indicate that in general these Galerkin approximations are not quasi-optimal in the Y-norm either. ♢

Returning to the general setting of Theorem 3.3.5, in the following theorem it will be shown that under an *additional* assumption quasi-optimal error estimates are valid.

Theorem 3.3.7. *Assuming* (3.3.8), *let* $(Y^\delta, H^\delta, X^\delta)_{\delta\in\Delta}$ *be a family of closed subspaces of* $Y \times H \times X$ *such that in addition to* $X^\delta \subseteq Y^\delta$ *and* $\operatorname{ran}\gamma_0|_{X^\delta} \subseteq H^\delta$, *also* (3.3.5) *is valid. Then for the Galerkin solutions* $(\mu^\delta, \sigma^\delta, u^\delta) \in Y^\delta \times H^\delta \times X^\delta$ *of* (3.2.8) *it holds that*

$$\|u - u^\delta\|_X \leq \gamma_\Delta^{-1} \inf_{\bar{u}^\delta \in X^\delta} \|u - \bar{u}^\delta\|_X.$$

Proof. As we have seen in the proof of Theorem 3.3.5, thanks to the assumptions $X^\delta \subseteq Y^\delta$ and $\operatorname{ran}\gamma_0|_{X^\delta} \subseteq H^\delta$, the component $u^\delta \in X^\delta$ of the Galerkin solution of (3.2.8) is the Petrov–Galerkin solution of (3.2.6) with test space $Z^\delta \subset Y^\delta \times H^\delta$.

By (3.3.11), the projector $P^\delta\colon u \mapsto u^\delta$ satisfies $\|P^\delta u\|_{X,Y^\delta} \leq \|u\|_X$. The proof is completed by $\|\cdot\|_X \leq \gamma_\Delta^{-1}\|\cdot\|_{X,Y^\delta}$ on X^δ by assumption (3.3.5), in combination with (3.3.1). □

In [And13], Andreev studied minimal residual Petrov–Galerkin discretizations of $\begin{bmatrix} B \\ \gamma_0 \end{bmatrix} u = \begin{bmatrix} g \\ \gamma_0' u_0 \end{bmatrix}$. They can equivalently be interpreted as Galerkin discretizations of (3.2.8) (cf. [CDW12], [BS14, Prop. 2.2]). In view of this, Theorem 3.3.7 reproduces, though here with a clear-cut constant, the results from [And13, Thms. 3.1 & 4.1].

Remark 3.3.8. As was pointed out earlier in [And13], for practical computations it can be attractive to modify the Galerkin discretization of (3.2.8) by replacing ${E_Y^\delta}' A_s E_Y^\delta$ by some $\tilde{A}_s^\delta = \tilde{A}_s^{\delta\prime} \in \mathcal{L}\mathrm{is}(Y^\delta, {Y^\delta}')$ whose inverse can be applied cheaply (a preconditioner) [3], so that for constants $0 < c_{\mathcal{N}} \leq C_{\mathcal{N}} < \infty$,

$$\frac{(\tilde{A}_s^\delta u)(u)}{(A_s u)(u)} \in [c_{\mathcal{N}}^2, C_{\mathcal{N}}^2] \quad (\delta \in \Delta,\, u \in Y^\delta).$$

Indeed, in that case one can solve the then explicitly available Schur complement equation with precondition CG, instead of applying the preconditioned

[3] For Galerkin discretizations of (3.2.10), such a replacement of ${E_Y^\delta}' A_s E_Y^\delta$ by an equivalent operator will result in an inconsistent discretization.

MINRES iteration. By redefining $Z^\delta := \mathrm{clos}_{Y^\delta \times H^\delta} \operatorname{ran} \begin{bmatrix} (\tilde{A}_s^\delta)^{-1} {E_Y^\delta}' B \\ \gamma_0 \end{bmatrix} \Big|_{X^\delta}$ in the proof of Theorem 3.3.5, and by taking W^δ to be its orthogonal complement in $Y^\delta \times H^\delta$ with Y^δ now being equipped with inner product $(\tilde{A}_s^\delta \cdot)(\cdot)$, instead of (3.3.11) we now estimate for any $\bar{u}^\delta \in X^\delta$,

$$\begin{aligned}
\|u^\delta - \bar{u}^\delta\|_{X,Y^\delta}^2 &= \sup_{0 \neq (v_1,v_2) \in Y^\delta} \frac{((B(u^\delta - \bar{u}^\delta))(v_1) + \langle u^\delta(0) - \bar{u}^\delta(0), v_2 \rangle)^2}{\|v_1\|_Y^2 + \|v_2\|^2} \\
&\leq \tfrac{1}{\min(c_{\mathcal{N}}^2,1)} \sup_{0 \neq (v_1,v_2) \in Y^\delta} \frac{((B(u^\delta - \bar{u}^\delta))(v_1) + \langle u^\delta(0) - \bar{u}^\delta(0), v_2 \rangle)^2}{(\tilde{A}_s^\delta v_1)(v_1)^2 + \|v_2\|^2} \\
&= \tfrac{1}{\min(c_{\mathcal{N}}^2,1)} \sup_{0 \neq (v_1,v_2) \in Z^\delta} \frac{((B(u^\delta - \bar{u}^\delta))(v_1) + \langle u^\delta(0) - \bar{u}^\delta(0), v_2 \rangle)^2}{(\tilde{A}_s^\delta v_1)(v_1)^2 + \|v_2\|^2} \\
&= \tfrac{1}{\min(c_{\mathcal{N}}^2,1)} \sup_{0 \neq (v_1,v_2) \in Z^\delta} \frac{((B(u - \bar{u}^\delta))(v_1) + \langle u(0) - \bar{u}^\delta(0), v_2 \rangle)^2}{(\tilde{A}_s^\delta v_1)(v_1)^2 + \|v_2\|^2} \\
&\leq \tfrac{\max(C_{\mathcal{N}}^2,1)}{\min(c_{\mathcal{N}}^2,1)} \sup_{0 \neq (v_1,v_2) \in Z^\delta} \frac{((B(u - \bar{u}^\delta))(v_1) + \langle u(0) - \bar{u}^\delta(0), v_2 \rangle)^2}{\|v_1\|_Y^2 + \|v_2\|^2} \\
&\leq \tfrac{\max(C_{\mathcal{N}}^2,1)}{\min(c_{\mathcal{N}}^2,1)} \|u - \bar{u}^\delta\|_X^2.
\end{aligned}$$

Consequently, a generalization of the statement of Theorem 3.3.5 reads as

$$\|u - u^\delta\|_{X,Y^\delta} \leq \left(1 + \sqrt{\tfrac{\max(C_{\mathcal{N}}^2,1)}{\min(c_{\mathcal{N}}^2,1)}}\right) \inf_{\bar{u}^\delta \in X^\delta} \|u - \bar{u}^\delta\|_X,$$

and that of Theorem 3.3.7 as

$$\|u - u^\delta\|_X \leq \gamma_\Delta^{-1} \sqrt{\tfrac{\max(C_{\mathcal{N}}^2,1)}{\min(c_{\mathcal{N}}^2,1)}} \inf_{\bar{u}^\delta \in X^\delta} \|u - \bar{u}^\delta\|_X. \quad \diamondsuit$$

Remark 3.3.9. As we have seen in the previous section, under the condition that (3.3.5) is valid, Galerkin discretizations of (3.2.10) yield quasi-optimal approximations. Assuming $A = A'$, in the current section we have seen that the same holds true for Galerkin discretizations of (3.2.8) when in addition $X^\delta \subseteq Y^\delta$ and $\operatorname{ran} \gamma_0|_{X^\delta} \subseteq H^\delta$. For the latter discretization, however, a still suboptimal error bound is valid without assuming (3.3.5). This raises the question whether this is also true for Galerkin discretizations of (3.2.10).

We saw that the Galerkin operator resulting from (3.2.10) is invertible when $X^\delta \neq \{0\}$. Moreover, when equipping X^δ with the 'mesh-dependent' norm $\|\cdot\|_{X,Y^\delta}$, by adapting the proof of Theorem 3.3.3 one can show that the operator is in $\mathcal{L}\mathrm{is}(Y^\delta \times X^\delta, {Y^\delta}' \times {X^\delta}')$ with both the operator and its inverse having a uniformly bounded norm. Despite this result, we could not establish, however, a suboptimal error estimate similar to Theorem 3.3.5. $\diamondsuit$

Finally in this section we comment on the implementation of the Galerkin discretization of (3.2.8). This system reads as

$$\begin{bmatrix} {E_Y^\delta}' A_s E_Y^\delta & 0 & {E_Y^\delta}' B E_X^\delta \\ 0 & {E_H^\delta}' E_H^\delta & {E_H^\delta}' \gamma_0 E_X^\delta \\ {E_X^\delta}' B' E_Y^\delta & {E_X^\delta}' \gamma_0' E_H^\delta & 0 \end{bmatrix} \begin{bmatrix} \mu^\delta \\ \sigma^\delta \\ u^\delta \end{bmatrix} = \begin{bmatrix} {E_Y^\delta}' g \\ {E_H^\delta}' u_0 \\ 0 \end{bmatrix}, \tag{3.3.13}$$

By eliminating σ^δ, it is equivalent to

$$\begin{bmatrix} {E_Y^\delta}' A_s E_Y^\delta & {E_Y^\delta}' B E_X^\delta \\ {E_X^\delta}' B' E_Y^\delta & -{E_X^\delta}' \gamma_0' E_H^\delta \left({E_H^\delta}' E_H^\delta \right)^{-1} {E_H^\delta}' \gamma_0 E_X^\delta \end{bmatrix} \begin{bmatrix} \mu^\delta \\ u^\delta \end{bmatrix} = \begin{bmatrix} {E_Y^\delta}' g \\ -{E_X^\delta}' \gamma_0' u_0 \end{bmatrix}. \tag{3.3.14}$$

The operator $E_H^\delta \left({E_H^\delta}' E_H^\delta \right)^{-1} {E_H^\delta}'$ is the H-orthogonal projector onto H^δ. So under the assumption that

$$\operatorname{ran} \gamma_0|_{X^\delta} \subseteq H^\delta$$

which was made in Theorem 3.3.7, it can be omitted, or equivalently, it can be pretended that $H^\delta = H$, without changing the solution (μ^δ, u^δ). The implementation of the resulting system

$$\begin{bmatrix} {E_Y^\delta}' A_s E_Y^\delta & {E_Y^\delta}' B E_X^\delta \\ {E_X^\delta}' B' E_Y^\delta & -{E_X^\delta}' \gamma_0' \gamma_0 E_X^\delta \end{bmatrix} \begin{bmatrix} \mu^\delta \\ u^\delta \end{bmatrix} = \begin{bmatrix} {E_Y^\delta}' g \\ -{E_X^\delta}' \gamma_0' u_0 \end{bmatrix}. \tag{3.3.15}$$

is easier, and it runs more efficiently than (3.3.13).

Remark 3.3.10. We can view (3.3.15) as a Galerkin discretisation of

$$\begin{bmatrix} A_s & B \\ B' & -\gamma_0' \gamma_0 \end{bmatrix} \begin{bmatrix} \mu \\ u \end{bmatrix} = \begin{bmatrix} g \\ -\gamma_0' u_0 \end{bmatrix}, \tag{3.3.16}$$

but for the analysis of the discretization error in (μ^δ, u^δ) it is still useful to view (3.3.15) before elimination of σ^δ, as a Galerkin discretization of (3.2.8) which yielded the sharp bound on this error presented in Theorem 3.3.7. ♢

3.4 Realization of the uniform inf-sup stability (3.3.5)

In Theorem 3.3.3 we showed that Galerkin discretizations of (3.2.10) are quasi-optimal when (3.3.5) holds, and in Theorem 3.3.7 for Galerkin discretizations of (3.2.8) when in addition $X^\delta \subseteq Y^\delta$ and $\operatorname{ran} \gamma_0|_{X^\delta} \subseteq H^\delta$ (and $A = A_s$).

In this section we realize the condition (3.3.5) for finite element spaces w.r.t. partitions of the space-time domain into prismatic elements. In §3.4.1 generally non-uniform partitions are considered for which the partition in time is independent of the spatial location, and the spatial mesh in each time slab is such that the corresponding H-orthogonal projection is uniformly V-stable. In §3.4.2 we revisit the special case, already studied in [And13], of trial spaces that are tensor products of temporal and spatial trial spaces.

3.4.1 Non-uniform approximation in space *local* in time, non-uniform approximation in time *global* in space

Theorem 3.4.1. *Let $\mathcal{O}$ be a collection of closed subspaces X_x of V such that the H-orthogonal projector Q_{X_x} onto X_x is in $\mathcal{L}(V,V)$, with*

$$\mu_{\mathcal{O}} := \inf_{X_x \in \mathcal{O}} \|Q_{X_x}\|^{-1}_{\mathcal{L}(V,V)} > 0.$$

For $N \in \mathbb{N}$, $0 = t_0 < t_1 < \cdots < t_N = T$, $(q_n)_{n=0}^N \subset \mathbb{N}$, $X_x^0, \dots, X_x^{N-1} \in \mathcal{O}$, let

$$X^\delta := \{u \in C(\bar{I}; V) \colon u|_{(t_i,t_{i+1})} \in P_{q_i} \otimes X_x^i\}$$
$$Y^\delta := \{v \in L_2(I;V) \colon v|_{(t_i,t_{i+1})} \in P_{q_i-1} \otimes X_x^i\}$$

Then with Δ being the collection of all $\delta = \delta(N, (t_i)_i, (q_i)_i, (X_x^i)_i)$, it holds that

$$\inf_{\delta\in\Delta} \inf_{\{u\in X^\delta \colon \partial_t u \neq 0\}} \sup_{0\neq v\in Y^\delta} \frac{(\partial_t u)(v)}{\|\partial_t u\|_{Y'}\|v\|_Y} \geq \mu_{\mathcal{O}}, \tag{3.4.1}$$

i.e. (3.3.5) *is valid.*

Proof. [And13, Lemma 6.2] yields $\inf_{0\neq u\in X_x} \sup_{0\neq v\in X_x} \frac{\langle u,v\rangle}{\|u\|_{V'}\|v\|_V} = \|Q_{X_x}\|^{-1}_{\mathcal{L}(V,V)}$.

With P_n denoting the Legendre polynomial of degree n, extended with zero outside $(-1,1)$, for any $u \in X^\delta$, $\partial_t u$ can be written as the $L_2(I;H)$-orthogonal expansion $(t,x) \mapsto \sum_{i=0}^{N-1}\sum_{n=0}^{q_i-1} P_n\Big(\frac{2t-(t_{i+1}+t_i)}{t_{i+1}-t_i}\Big)u_{i,n}(x)$ for some $u_{i,n} \in X_x^i$. Fixing $\varepsilon \in (0, \mu_{\mathcal{O}})$, for each (i,n) there is a $v_{i,n} \in X_x^i$ with $\|v_{i,n}\|_V = \|u_{i,n}\|_{V'}$ and $\langle u_{i,n}, v_{i,n}\rangle \geq (\mu_{\mathcal{O}} - \varepsilon)\|u_{i,n}\|_{V'}\|v_{i,n}\|_V$. Taking $v := (t,x) \mapsto \sum_{i=0}^{N-1}\sum_{n=0}^{q_i-1} P_n\Big(\frac{2t-(t_{i+1}+t_i)}{t_{i+1}-t_i}\Big)v_{i,n}(x)$, we conclude that

$$\begin{aligned}(\partial_t u)(v) &\geq (\mu_{\mathcal{O}} - \varepsilon)\sum_{i=0}^{N-1}\sum_{n=0}^{q_i-1} \Big\|P_n\Big(\tfrac{2\cdot-(t_{i+1}+t_i)}{t_{i+1}-t_i}\Big)\Big\|^2_{L_2(I)}\|u_{i,n}\|^2_{V'} \\ &= (\mu_{\mathcal{O}} - \varepsilon)\|u\|_{Y'}\|v\|_Y,\end{aligned}$$

which implies the result. □

Remark 3.4.2. In view of Theorem 3.3.7, note that both $X^\delta \subset Y^\delta$ and (3.3.5) are valid by taking $Y^\delta := \{v \in L_2(I;V) \colon v|_{(t_i,t_{i+1})} \in P_{q_i} \otimes X_x^i\}$. ♢

Considering the condition on the collection $\mathcal{O}$ of spatial trial spaces X_x, let us consider the typical situation that $H = L_2(\Omega)$, $V = H^1_{0,\gamma}(\Omega) = \{u \in H^1(\Omega) \colon u = 0 \text{ on } \gamma\}$ where $\Omega \subset \mathbb{R}^d$ is a bounded polytopal domain, and γ is a measurable, closed, possibly empty subset of $\partial\Omega$. We consider $X_x \subset V$ to be finite element spaces of some degree w.r.t. a family of uniformly shape regular, and, say, conforming partitions $\mathcal{T}$ of Ω into, say, d-simplices, where γ is the union of some $(d-1)$-faces of $S \in \mathcal{T}$. When the partitions in this family are

quasi-uniform, then using e.g. the Scott–Zhang quasi-interpolator ([SZ90]), it is easy to demonstrate the (uniform) *simultaneous approximation property*

$$\sup_{X_x\in\mathcal{O}}\sup_{0\neq u\in V}\frac{\inf_{v\in X_x}\{\|v\|_V+\left(\sup_{0\neq w\in X_x}\frac{\|w\|_V}{\|w\|_H}\right)\|u-v\|_H\}}{\|u\|_V}<\infty.$$

Writing for $u\in V$ and any $v\in X_x$, $Qu=v+Q(u-v)$, one easily infers that $\sup_{X_x\in\mathcal{O}}\|Q_x\|_{\mathcal{L}(V,V)}<\infty$.

The uniform boundedness of $\|Q_x\|_{\mathcal{L}(V,V)}$ is, however, by no means restricted to families of finite element spaces w.r.t. quasi-uniform partitions, and it has been demonstrated for families of locally refined partitions, for $d=2$ including those that are generated by the newest vertex bisection algorithm. We refer to [Car01, GHS16].

3.4.2 Non-uniform approximation in space *global* in time, non-uniform approximation in time *global* in space

If in Theorem 3.4.1, the spatial trial spaces X_x^i are independent of the temporal interval (t_i,t_{i+1}), then X^δ is a tensor product of trial spaces in space and time. In that case, one shows inf-sup stability for general temporal trial spaces, e.g. spline spaces with more global smoothness than continuity.

Theorem 3.4.3. *Let $\mathcal{O}$ be as in Theorem 3.4.1. Given closed subspaces $X_t\subset H^1(I)$, $\frac{\mathrm{d}}{\mathrm{d}t}X_t\subseteq Y_t\subset L_2(I)$ and $X_x\in\mathcal{O}$, let $X^\delta:=X_t\otimes X_x$, $Y^\delta:=Y_t\otimes X_x$. Then with Δ being the collection of all $\delta=\delta(X_t,Y_t,X_x)$, (3.4.1) is valid.*

The proof of this result follows from the fact that thanks to the Kronecker product structure of $\partial_t\in\mathcal{L}(X,Y')$, for such trial spaces we have

$$\begin{aligned}&\inf_{\{u\in X^\delta:\,\partial_t u\neq 0\}}\sup_{0\neq v\in Y^\delta}\frac{(\partial_t u)(v)}{\|\partial_t u\|_{Y'}\|v\|_Y}\\&=\inf_{\{u\in X_t:\,\frac{\mathrm{d}u}{\mathrm{d}t}\neq 0\}}\sup_{0\neq v\in Y_t}\frac{\int_I\frac{\mathrm{d}u}{\mathrm{d}t}v\,\mathrm{d}t}{\|\frac{\mathrm{d}u}{\mathrm{d}t}\|_{L_2(I)}\|v\|_{L_2(I)}}\times\inf_{0\neq u\in X_x}\sup_{0\neq v\in X_x}\frac{\langle u,v\rangle}{\|u\|_{V'}\|v\|_V}\qquad(3.4.2)\\&=\inf_{0\neq u\in X_x}\sup_{0\neq v\in X_x}\frac{\langle u,v\rangle}{\|u\|_{V'}\|v\|_V}.\end{aligned}$$

(Indeed: for U and V Hilbert, $T\in\mathcal{L}(U,V')$, and Riesz mappings $R_U\colon U\to U'$, $R_V\colon V\to V'$, we find $\inf_{0\neq u\in U}\sup_{0\neq v\in V}\frac{(Tu)(v)}{\|u\|_U\|v\|_V}=\min\sigma(R_U^{-1}T'R_V^{-1}T)$, with $R_U^{-1}T'R_V^{-1}T\in\mathcal{L}(U,U)$ being self-adjoint and non-negative. In the above setting, it is a Kronecker product of corresponding operators acting in the 'time' and 'space' direction, respectively.)

Remark 3.4.4 (Sparse tensor products)**.** Instead of considering the 'full' tensor product trial spaces from Theorem 3.4.3, more efficient approximations can be found by the application of 'sparse' tensor products. Let $X_x^{(0)}\subset X_x^{(1)}\subset\cdots$ be

a sequence of spaces from $\mathcal{O}$, $X_t^{(0)} \subset X_t^{(1)} \subset \cdots \subset H^1(I)$, and $Y_t^{(0)} \subset Y_t^{(1)} \subset \cdots \subset L_2(I)$ such that $Y_t^{(k)} \supseteq \frac{d}{dt} X_t^{(k)}$. Then for $X^{(\ell)} := \sum_{k=0}^{\ell} X_t^{(k)} \otimes X_x^{(\ell-k)}$, $Y^{(\ell)} := \sum_{k=0}^{\ell} Y_t^{(k)} \otimes X_x^{(\ell-k)}$ inf-sup stability holds true uniformly in ℓ with inf-sup constant $\mu_{\mathcal{O}}$.

Although this result follows as a special case from the analysis in [And13], for convenience we include the argument. With $W_t^{(k)} := Y_t^{(k)} \cap (Y_t^{(k-1)})^{\perp_{L_2(I)}}$ for $k > 0$, and $W_t^{(0)} := Y_t^{(0)}$, from the nestings of $(Y_t^{(i)})_i$ and $(X_x^{(i)})_i$ one infers that $Y^{(\ell)} = \oplus_{k=0}^{\ell} W_t^{(k)} \otimes X_x^{(\ell-k)}$ is an $(L_2(I) \otimes H)$-orthogonal decomposition. Given $y \in Y^{(\ell)}$, let $y = \sum_{k=0}^{\ell} y_k$ be the corresponding expansion. Fixing $\varepsilon \in (0, \mu_{\mathcal{O}})$, there exist $\tilde{y}_k \in W_t^{(k)} \otimes X_x^{(\ell-k)}$ with $\langle y_k, \tilde{y}_k \rangle_{L_2(I)\otimes H} \geq (\mu_{\mathcal{O}} - \varepsilon)\|y_k\|_Y \|\tilde{y}_k\|_{Y'}$ and $\|\tilde{y}_k\|_{Y'} = \|y_k\|_Y$, and so $\langle \sum_{k=0}^{\ell} y_k, \sum_{k=0}^{\ell} \tilde{y}_k \rangle_{L_2(I)\otimes H} \geq (\mu_{\mathcal{O}} - \varepsilon)\| \sum_{k=0}^{\ell} y_k\|_Y \| \sum_{k=0}^{\ell} \tilde{y}_k\|_{Y'}$. The result follows from $\partial_t X^{(\ell)} \subseteq Y^{(\ell)}$. ◊

Remark 3.4.5. In view of (3.4.2), it is obvious that Theorem 3.4.3 remains valid when $\frac{\mathrm{d}}{\mathrm{d}t} X_t \subseteq Y_t$ is relaxed to $\inf_{\{u \in X_t \colon \frac{\mathrm{d}u}{\mathrm{d}t} \neq 0\}} \sup_{0 \neq v \in Y_t} \frac{\int_I \frac{\mathrm{d}u}{\mathrm{d}t} v \, \mathrm{d}t}{\|\frac{\mathrm{d}u}{\mathrm{d}t}\|_{L_2(I)} \|v\|_{L_2(I)}} > 0$ uniformly in the pairs (X_t, Y_t) that are applied. As shown in [And13], the same holds true in the sparse tensor product case. For X_t being the space of continuous piecewise linears w.r.t. some partition $\mathcal{T}$ of I, and Y_t being the space of continuous piecewise linears w.r.t. the once dyadically refined partition, an easy computation shows that the inf-sup constant is not less than $\sqrt{3/4}$.

Since in our experiments with the method from [And13], with this alternative choice of Y_t the numerical results are slightly better than when taking Y_t to be the space of discontinuous piecewise linears w.r.t. $\mathcal{T}$, we will report on results obtained with this alternative choice for Y_t. ◊

3.5 Numerical experiments

For the simplest possible case of the heat equation in one space dimension discretized using as 'primal' trial space X^δ the space of continuous piecewise bilinears w.r.t. a uniform partition into squares, we compare the accuracy of approximations provided by the newly proposed method (i.e. the Galerkin discretization of (3.2.10) with trial space here denoted by $Y^\delta_{\mathrm{new}} \times X^\delta$) with those obtained with the method from [And13] (i.e. the Galerkin discretization of (3.2.8)). We implement the latter method in the form (3.3.15), i.e. after eliminating σ^δ. The remaining trial space is denoted here by $Y^\delta_{\mathrm{Andr.}} \times X^\delta$. So we take $T = 1$, i.e. $I = (0,1)$, and with $\Omega := (0,1)$, $H := L_2(\Omega)$, $V := H^1_0(\Omega)$, $a(t; \eta, \zeta) := \int_\Omega \eta' \zeta' \, dx$. With $\frac{1}{h_t} = \frac{1}{h_x} =: \frac{1}{h} \in \mathbb{N}$, we set

$$
\begin{aligned}
X^\delta &:= \{v \in H^1(I) \colon v|_{(ih,(i+1)h)} \in P_1\} &&\otimes \{v \in H^1_0(\Omega) \colon v|_{(ih,(i+1)h)} \in P_1\},\\
Y^\delta_{\mathrm{new}} &:= \{v \in L_2(I) \colon v|_{(ih,(i+1)h)} \in P_0\} &&\otimes \{v \in H^1_0(\Omega) \colon v|_{(ih,(i+1)h)} \in P_1\},\\
Y^\delta_{\mathrm{Andr}} &:= \{v \in H^1(I) \colon v|_{(ih/2,(i+1)h/2)} \in P_1\} &&\otimes \{v \in H^1_0(\Omega) \colon v|_{(ih,(i+1)h)} \in P_1\},
\end{aligned}
$$

Note that $\dim Y^\delta_{\text{new}} \approx \dim X^\delta$ and $\dim Y^\delta_{\text{Andr}} \approx 2 \dim X^\delta$. The total number of non-zeros in the whole system matrix of the new method is asymptotically a factor 2 smaller than this number for Andreev's method.

Prescribing both a smooth exact solution $u(t,x) = e^{-2t} \sin \pi x$ and a singular one $u(t,x) = e^{-2t}|t-x| \sin \pi x$, Figure 3.1 shows the errors $e^\delta := u - u^\delta$ in X-norm as a function of $\dim X^\delta$. The norms of the errors in the Galerkin

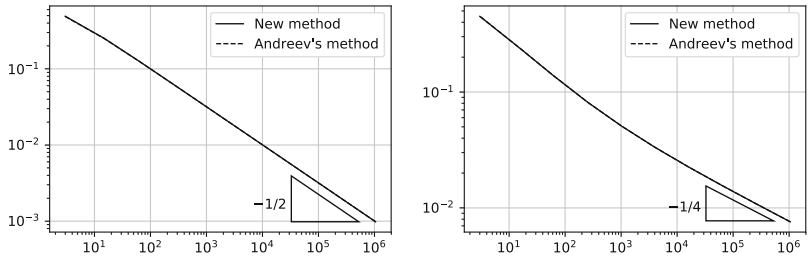


Figure 3.1 $\|e^\delta\|_X$ vs. $\dim X^\delta$ for both numerical methods. Left: $u(t,x) = e^{-2t} \sin \pi x$. Right: $u(t,x) = e^{-2t}|t-x| \sin \pi x$.

solutions found by the two methods are nearly indistinguishable from one another. Furthermore, the observed convergence rates $1/2$ and $1/4$, respectively, are the best possible ones that in view of the polynomial degrees of X^δ and Y^δ (new method) or that of X^δ (Andreev's method) and the regularity of the solutions can be expected with the application of uniform meshes. (For any $\varepsilon > 0$, $e^{-2t}|t-x| \sin \pi x \in H^{\frac{3}{2}-\varepsilon}(I \times \Omega) \setminus H^{\frac{3}{2}}(I \times \Omega)$).

For both solutions and both numerical methods, the errors $e^\delta(T,\cdot)$ measured in $L_2(\Omega)$ converge with the better rate 1, i.e., these errors are asymptotically proportional to h^2, see left picture in Figure 3.2. To illustrate that the two methods yield different Galerkin solutions, we show $e^\delta(0,\cdot)$, measured in $L_2(\Omega)$-norm in the right of Figure 3.2.

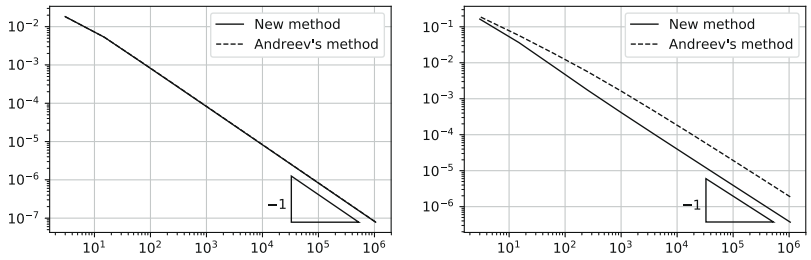


Figure 3.2 Singular solution $u(t,x) = e^{-2t}|t-x| \sin \pi x$. Left: $\|e^\delta(T,\cdot)\|_{L_2(\Omega)}$ vs. $\dim X^\delta$. Right: $\|e^\delta(0,\cdot)\|_{L_2(\Omega)}$ vs. $\dim X^\delta$.

The new method actually yields two approximations for u, viz. u^δ and λ^δ. This secondary approximation is not in X, but it is in $Y = L_2(I;V)$. For both

solutions, the errors in λ^δ measured in Y-norm are slightly larger than in those in u^δ, see left picture in Figure 3.3.

Finally, we replaced the symmetric spatial diffusion operator by a nonsymmetric convection-diffusion operator $a(t;\eta,\zeta) := \int_\Omega \eta'\zeta' + \beta\eta'\zeta dx$. Letting $\beta := 100$ and again taking the singular solution $u(t,x) = e^{-2t}|t-x|\sin \pi x$, the errors e^δ in X-norm of both Galerkin solutions vs. $\dim X^\delta$ are given in Figure 3.3. We once again see that the two methods show very comparable convergence behaviour.

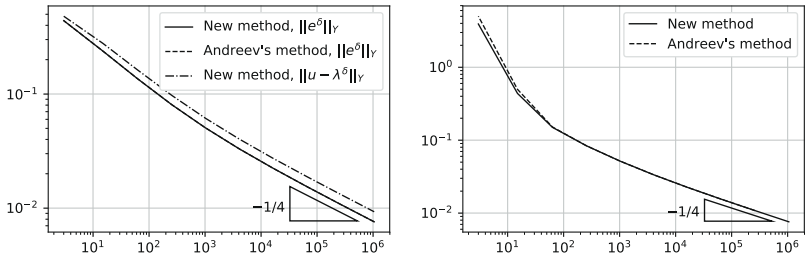
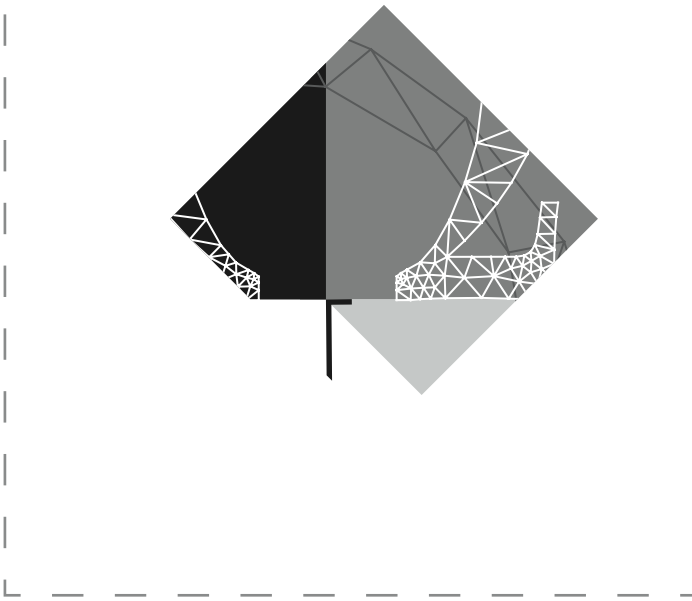


Figure 3.3 Singular solution $u(t,x) = e^{-2t}|t-x|\sin \pi x$. Left: $\|e^\delta\|_Y$ and $\|u-\lambda^\delta\|_Y$ vs. $\dim X^\delta$ for the symmetric problem. Right: $\|e^\delta\|_X$ vs. $\dim X^\delta$ for the nonsymmetric problem.

3.6 Conclusion

Three related (Petrov-) Galerkin discretizations of space-time variational formulations were analyzed. The scheme introduced by Steinbach in [Ste15] has the lowest computational cost, and applies on general space-time meshes, but depending on the exact solution, the numerical solutions can be far from quasi-optimal in the natural mesh-independent norm. The minimal residual Petrov–Galerkin discretization introduced by Andreev in [And13] yields for suitable trial and test pairs quasi-optimal approximations from the trial space. For suitable pairs of trial spaces, Galerkin discretizations of a newly introduced mixed space-time variational formulation also yield quasi-optimal approximations, but for the same accuracy at a lower computational cost than with the method from [And13].



$\sqrt{\|B\cdot\|_{Y'}^2+\beta\|\gamma_0\cdot\|^2}$
$Z^\delta := \operatorname{ran} T^\delta|_{X^\delta}$

4 Minimal residual discretizations of asymmetric parabolic PDEs

Abstract We consider a minimal residual discretization of a simultaneous space-time variational formulation of parabolic evolution equations. Under the usual 'LBB' stability condition on pairs of trial- and test spaces we show quasi-optimality of the numerical approximations without assuming symmetry of the spatial part of the differential operator. Under a stronger LBB condition we show error estimates in an energy-norm which are independent of this spatial differential operator.

4.1 Introduction

This chapter is about the numerical solution of parabolic evolution equations in a simultaneous space-time variational formulation. Compared to classical time-stepping schemes, simultaneous space-time methods are much better suited for massively parallel computation (e.g. [NS19, vVW20a]), allow for local refinements in space and time (e.g. [SY18, GS19, SvVW21, vVW21a]), and produce numerical approximations from the trial spaces that are quasi-best.

The standard bilinear form that results from a space-time variational formulation is non-coercive, which makes it difficult to construct pairs of discrete trial and test spaces that inherit the stability of the continuous formulation. For this reason, in [And13] R. Andreev proposed to use minimal residual discretizations. They have an equivalent interpretation as Galerkin discretizations of an extended self-adjoint, but indefinite, mixed system having as secondary variable the Riesz lift of the PDE-residual of the primal variable.

For pairs of trial spaces that satisfy a Ladyshenskaja–Babuška–Brezzi (LBB) condition, it was shown that w.r.t. the norm on the natural solution space, being an intersection of Bochner spaces, the Galerkin solutions are quasi-best approximations from the selected trial spaces. This LBB condition was verified in [And13] for 'full' and 'sparse' tensor products of various finite elements spaces in space and time. The sparse tensor product setting was then generalized in [SvVW21, Proposition 5.1] to allow for local refinements in space and time whilst retaining (uniform) LBB stability.

This chapter is a minor modification of **Minimal residual space-time discretizations of parabolic equations: Asymmetric spatial operators**, R. Stevenson and J. Westerdiep, submitted to *Computers & Mathematics with Applications*, arXiv:2106.01090.

A different minimal residual formulation of first order system type was introduced in [FK21], see also [GS21]. Here the various residuals are all measured in L_2-norms, meaning that they do not have to be introduced as separate variables, and the resulting bilinear form is coercive.

Closer in spirit to [And13] are the space-time methods of [Ste15, LMN16, BEEN19], in which error bounds are presented w.r.t. mesh-dependent norms. In [Dev20, SZ20] space-time variational methods are presented that lead to coercive bilinear forms based on fractional Sobolev norms of order $\frac{1}{2}$. A first order space-time DPG formulation of the heat equation is presented in [DS20].

A restriction imposed in [And13], as well as in the other mentioned references apart from [BEEN19, GS21], is that the spatial part of the PDO is not only coercive but also symmetric. In Chapter 3 we could remove the symmetry condition for the analysis of a related Brézis–Ekeland–Nayroles (BEN) ([BE76, Nay76]) formulation of the parabolic PDE. In the current work, we prove that also for the minimal residual (MR) method the symmetry condition can be dropped. So for both MR and BEN we show that under the aforementioned LBB condition the Galerkin approximations are quasi-optimal, where the bound on the error in the numerical approximation for BEN improves upon the one from Chapter 3.

The error bounds for both MR and BEN degrade for increasing asymmetry. This is not an artefact of the theory but is confirmed by numerical experiments. Under a stronger LBB condition on the pair of trial spaces, however, we will prove that the MR and BEN approximations are quasi-best w.r.t. a continuous, i.e., mesh-*in*dependent, energy-norm, uniformly in the spatial PDO.

We present numerical tests for the evolution problem governed by the simple PDE $\partial_t - \varepsilon\partial_x^2 + \partial_x + e\mathrm{Id}$ on $(0,1)^2$ with initial and boundary conditions, where e is either 0 or 1. For the case that homogeneous Dirichlet boundary conditions are prescribed at the outflow boundary $x = 1$, the results for very small ε illustrate that *quasi-optimal approximations* do not necessarily mean *accurate approximations*. Indeed the error in the computed solution is large because of the unresolved boundary layer. Minimization of the error in least-squares energy norm causes a global spread of the error along the streamlines. We tackle this problem by imposing these boundary conditions weakly.

4.1.1 Organization

In Sect. 4.2 we recall the well-posed space-time variational formulation of the parabolic problem and study its conditioning. Under the usual LBB condition, in Sect. 4.3 we show quasi-optimality of the MR method without assuming symmetry of the spatial differential operator. A similar result is shown for BEN in Sect. 4.4. Known results concerning the verification of this LBB condition are summarized in Sect. 4.5, together with results about optimal preconditioning.

In Sect. 4.6 we equip the solution space with an energy norm, and, under a stronger LBB condition, show error estimates for MR and BEN which are independent of the spatial differential operator. We present an a posteriori

error estimator which, under an even stronger LBB condition, is efficient and, modulo a data-oscillation term, is reliable.

In Sect. 4.7 we apply the general theory to the example of the convection-diffusion problem. We give pairs of trial- and test spaces which satisfy the 2nd and 3rd mentioned LBB conditions. Finally, in Sect. 4.8 we present numerical results for the MR method in the simple case of having a one-dimensional spatial domain. To solve the problems caused by an unresolved boundary layer, we modify the method by imposing a boundary condition weakly.

4.1.2 Notations

In this work, by $C \lesssim D$ we will mean that C can be bounded by a multiple of D, independently of parameters which C and D may depend on. Obviously, $C \gtrsim D$ is defined as $D \lesssim C$, and $C \eqsim D$ as $C \lesssim D$ and $C \gtrsim D$.

For normed linear spaces E and F, by $\mathcal{L}(E,F)$ we will denote the normed linear space of bounded linear mappings $E \to F$, and by $\mathcal{L}\mathrm{is}(E,F)$ its subset of boundedly invertible linear mappings $E \to F$. We write $E \hookrightarrow F$ to denote that E is continuously embedded into F. For simplicity only, we exclusively consider linear spaces over the scalar field $\mathbb{R}$.

4.2 Well-posed variational formulation

Let V, H be separable Hilbert spaces of functions on some "spatial domain" such that $V \hookrightarrow H$ with dense embedding. Identifying H with its dual, we obtain the Gelfand triple $V \hookrightarrow H \simeq H' \hookrightarrow V'$. We use $\langle \cdot, \cdot \rangle$ to denote both the scalar product on $H \times H$ as well as its unique extension to the duality pairing on $V' \times V$ or $V \times V'$, and denote the norm on H by $\| \cdot \|$.

For a.e.

$$t \in I := (0,T),$$

let $a(t;\cdot,\cdot)$ denote a bilinear form on $V \times V$ such that for any $\eta, \zeta \in V$, $t \mapsto a(t;\eta,\zeta)$ is measurable on I, and such that for some $\varrho \in \mathbb{R}$, for a.e. $t \in I$,

$$|a(t;\eta,\zeta)| \lesssim \|\eta\|_V \|\zeta\|_V \quad (\eta,\zeta \in V) \quad (\textit{boundedness}), \tag{4.2.1}$$

$$a(t;\eta,\eta) + \varrho\langle \eta,\eta\rangle \gtrsim \|\eta\|_V^2 \quad (\eta \in V) \quad (\textit{Gårding inequality}). \tag{4.2.2}$$

With $A(t) \in \mathcal{L}\mathrm{is}(V,V')$ being defined by $(A(t)\eta)(\zeta) := a(t;\eta,\zeta)$, given a forcing function g and an initial value u_0, we are interested in solving the *parabolic initial value problem* to finding u such that

$$\begin{cases} \frac{\mathrm{d}u}{\mathrm{d}t}(t) + A(t)u(t) = g(t) & (t \in I), \\ u(0) = u_0. \end{cases} \tag{4.2.3}$$

In a simultaneous space-time variational formulation, the parabolic PDE reads as finding u from a suitable space of functions X of time and space s.t.

$$(Bw)(v) := \int_I \langle \tfrac{\mathrm{d}w}{\mathrm{d}t}(t), v(t)\rangle + a(t;w(t),v(t))\,\mathrm{d}t = \int_I \langle g(t), v(t)\rangle\,\mathrm{d}t =: g(v)$$

for all v from another suitable space of functions Y of time and space. One possibility to enforce the initial condition is by testing it against additional test functions. A proof of the following result can be found in [SS09], cf. [DL92, Ch.XVIII, §3] and [Wlo82, Ch. IV, §26] for slightly different statements.

Theorem 4.2.1. *With $X := L_2(I;V) \cap H^1(I;V')$, $Y := L_2(I;V)$, under conditions* (4.2.1) *and* (4.2.2) *it holds that*

$$\begin{bmatrix} B \\ \gamma_0 \end{bmatrix} \in \mathcal{L}\mathrm{is}(X, Y' \times H),$$

where for $t \in \bar{I}$, $\gamma_t \colon u \mapsto u(t,\cdot)$ denotes the trace map. That is, assuming $g \in Y'$ and $u_0 \in H$, finding $u \in X$ such that

$$\begin{bmatrix} B \\ \gamma_0 \end{bmatrix} u = \begin{bmatrix} g \\ u_0 \end{bmatrix} \tag{4.2.4}$$

is a well-posed simultaneous space-time variational formulation of (4.2.3).

With $\tilde{u}(t) := u(t)e^{-\varrho t}$, (4.2.3) is equivalent to $\frac{\mathrm{d}\tilde{u}}{\mathrm{d}t}(t) + (A(t) + \varrho\mathrm{Id})\tilde{u}(t) = g(t)e^{-\varrho t}$ ($t \in I$), $\tilde{u}(0) = u_0$. Since $((A(t) + \varrho\mathrm{Id})\eta)(\eta) \gtrsim \|\eta\|_V^2$, w.l.o.g. we will always assume that, besides (4.2.1), (4.2.2) is valid for $\varrho = 0$, i.e., for a.e. $t \in I$,

$$a(t;\eta,\eta) \gtrsim \|\eta\|_V^2 \quad (\eta \in V) \quad (\textit{coercivity}). \tag{4.2.5}$$

We define $A, A_s \in \mathcal{L}\mathrm{is}(Y, Y')$, $A_a \in \mathcal{L}(Y, Y')$, and $C, \partial_t \in \mathcal{L}(X, Y')$ by

$$(Aw)(v) := \int_I a(t; w(t), v(t))\,\mathrm{d}t, \quad A_s := \tfrac{1}{2}(A + A'), \quad A_a := \tfrac{1}{2}(A - A'),$$
$$\partial_t := B - A, \quad C := B - A_s = \partial_t + A_a,$$

and equip Y with 'energy'-scalar product $\langle\cdot,\cdot\rangle_Y := (A_s\cdot)(\cdot)$, and norm

$$\|v\|_Y := \sqrt{(A_s v)(v)}.$$

being, thanks to (4.2.1) and (4.2.5), equivalent to the standard norm on Y. Equipping Y' with the resulting dual norm, $A_s \in \mathcal{L}\mathrm{is}(Y, Y')$ is an isometric isomorphism, and so for $f \in Y'$ we have

$$f(A_s^{-1}f) = (A_s A_s^{-1} f)(A_s^{-1} f) = \|A_s^{-1} f\|_Y^2 = \|f\|_{Y'}^2.$$

For some constant $\beta \geq 1$, we equip X with norm

$$\|\cdot\|_X := \sqrt{\|\cdot\|_Y^2 + \|\partial_t \cdot\|_{Y'}^2 + \|\gamma_T \cdot\|^2 + (\beta - 1)\|\gamma_0 \cdot\|^2},$$

being, thanks to $X \hookrightarrow C(\bar{I};H)$, equivalent to the standard norm on X. In addition, we define the energy-norm on X by

$$|\!|\!| \cdot |\!|\!|_X := \sqrt{\|B \cdot\|_{Y'}^2 + \beta\|\gamma_0 \cdot\|^2},$$

which, thanks to Theorem 4.2.1, is indeed a norm on X.

Proposition 4.2.2. *With $\alpha := \|A_a\|_{\mathcal{L}(Y,Y')}$, for $0 \neq w \in X$ it holds that*

$$\left(1+\frac{\alpha}{2}\left(\alpha+\sqrt{\alpha^2+4}\right)\right)^{-1} \leq \frac{|||w|||_X^2}{\|w\|_X^2} \leq 1+\frac{\alpha}{2}\left(\alpha+\sqrt{\alpha^2+4}\right),$$

so that, in particular, both norms are equal when $A_a = 0$.

Proof. Using that for $w, v \in X$,

$$\begin{aligned}((\partial_t+\partial_t'+\gamma_0'\gamma_0)w)(v) &= \int_I \langle \tfrac{\mathrm{d}w}{\mathrm{d}t}(t), v(t)\rangle + \langle w(t), \tfrac{\mathrm{d}v}{\mathrm{d}t}(t)\rangle \,\mathrm{d}t + \langle w(0), v(0)\rangle \\ &= \int_I \tfrac{\mathrm{d}}{\mathrm{d}t}\langle w(t), v(t)\rangle \,\mathrm{d}t + \langle w(0), v(0)\rangle = (\gamma_T'\gamma_T w)(v),\end{aligned}$$

we find that

$$\begin{aligned}B'A_s^{-1}B + \beta\gamma_0'\gamma_0 &= (C'+A_s)A_s^{-1}(C+A_s) + \beta\gamma_0'\gamma_0 \\ &= C'A_s^{-1}C + A_s + C' + C + \beta\gamma_0'\gamma_0 \\ &= C'A_s^{-1}C + A_s + \partial_t' + \partial_t + \beta\gamma_0'\gamma_0 \\ &= C'A_s^{-1}C + A_s + \gamma_T'\gamma_T + (\beta-1)\gamma_0'\gamma_0.\end{aligned} \tag{4.2.6}$$

For $w \in X$,

$$(C'A_s^{-1}Cw)(w) = (Cw)(A_s^{-1}Cw) = \|(\partial_t+A_a)w\|_{Y'}^2 \leq (\|\partial_t w\|_{Y'} + \alpha\|w\|_Y)^2,$$

and so, for any $\eta \neq 0$, Young's inequality shows that

$$\begin{aligned}\|Bw\|_{Y'}^2 + \beta\|\gamma_0 w\|^2 &= \Big((C'A_s^{-1}C + A_s + \gamma_T'\gamma_T + (\beta-1)\gamma_0'\gamma_0)(w)\Big)(w) \\ &\leq (1+\eta^2)\|\partial_t w\|_{Y'}^2 + ((1+\eta^{-2})\alpha^2+1)\|w\|_Y^2 + \|\gamma_T w\|^2 + (\beta-1)\|\gamma_0 w\|^2.\end{aligned}$$

Solving $(1+\eta^2) = (1+\eta^{-2})\alpha^2+1$ gives $1+\eta^2 = 1+\frac{\alpha}{2}\left(\alpha+\sqrt{\alpha^2+4}\right)$, showing one of the bounds of the statement.

From

$$\|(\partial_t+A_a)w\|_{Y'}^2 \geq (\|\partial_t w\|_{Y'} - \alpha\|w\|_Y)^2 \geq (1-\eta^2)\|\partial_t w\|_{Y'}^2 + (1-\eta^{-2})\alpha^2\|w\|_Y^2$$

again by Young's inequality, by solving η^2 from $1-\eta^2 = (1-\eta^{-2})\alpha^2+1$ the other bound follows. □

Remark 4.2.3. As $\|\cdot\|_Y$ is defined in terms of the symmetric part A_s of the spatial differential operator A, $\alpha = \|A_a\|_{\mathcal{L}(Y,Y')}$ measures the *relative* asymmetry of the operator A. Indeed $\|A_a\|_{\mathcal{L}(Y,Y')} = \|A_s^{-\frac{1}{2}}A_aA_s^{-\frac{1}{2}}\|_{\mathcal{L}(L_2(I;H),L_2(I;H))} = \rho(A_s^{-\frac{1}{2}}A_a'A_s^{-1}A_aA_s^{-\frac{1}{2}})^{\frac{1}{2}} = \rho(A_s^{-1}A_aA_s^{-1}A_a)^{\frac{1}{2}}$, where we used $A_a' = -A_a$. ◊

4.3 Minimal residual (MR) method

Let $(X^\delta, Y^\delta)_{\delta\in\Delta}$ a family of closed, proper, non-zero subspaces of X and Y, respectively. For $\delta \in \Delta$, let E_X^δ and E_Y^δ denote the trivial embeddings $X^\delta \to X$ and $Y^\delta \to Y$ that we often write for clarity. We assume that

$$X^\delta \subseteq Y^\delta \quad (\delta \in \Delta), \tag{4.3.1}$$

$$\gamma_\Delta^{\partial_t} := \inf_{\delta\in\Delta} \inf_{\{w\in X^\delta:\, \partial_t E_X^\delta w \neq 0\}} \frac{\|{E_Y^\delta}'\partial_t E_X^\delta w\|_{Y^{\delta\prime}}}{\|\partial_t E_X^\delta w\|_{Y'}} > 0. \tag{4.3.2}$$

Furthermore, for efficiency, we assume to have a $K_Y^\delta = {K_Y^\delta}' \in \mathcal{L}\mathrm{is}(Y^{\delta\prime}, Y^\delta)$ (a 'preconditioner'), such that for some constants $0 < r_\Delta \leq R_\Delta < \infty$,

$$\frac{((K_Y^\delta)^{-1}v)(v)}{({E_Y^\delta}' A_s E_Y^\delta v)(v)} \in [r_\Delta, R_\Delta] \quad (\delta \in \Delta,\, v \in Y^\delta), \tag{4.3.3}$$

or, equivalently, $\frac{f(K_Y^\delta f)}{f(({E_Y^\delta}' A_s E_Y^\delta)^{-1} f)} \in [R_\Delta^{-1}, r_\Delta^{-1}]$ $(\delta \in \Delta, f \in Y^{\delta\prime})$.

Noticing that $\|f\|_{Y^{\delta\prime}}^2 = f(({E_Y^\delta}' A_s E_Y^\delta)^{-1} f)$, the expression

$$\|\cdot\|_{K_Y^\delta} := \sqrt{(\cdot)(K_Y^\delta \cdot)}$$

defines an equivalent norm on $Y^{\delta\prime}$, and our Minimal Residual approximation $u^\delta \in X^\delta$ of the solution $u \in X$ of (4.2.4) is defined as

$$u^\delta := \operatorname*{arg\,min}_{w\in X^\delta} \|{E_Y^\delta}'(BE_X^\delta w - g)\|_{K_Y^\delta}^2 + \beta\|\gamma_0 E_X^\delta w - u_0\|^2, \tag{4.3.4}$$

for some constant $\beta \geq 1$. Later we will see that, thanks to (4.3.2) and (4.3.3),

$$\inf_{0\neq w\in X^\delta} \sup_{(v_1,v_2)\in Y^\delta\times H} \frac{(BE_X^\delta w)(E_Y^\delta v_1) + \beta\langle \gamma_0 E_X^\delta w, v_2\rangle}{\sqrt{((K_Y^\delta)^{-1}v_1)(v_1) + \beta\|v_2\|^2}} > 0 \tag{4.3.5}$$

(even uniformly in $\delta \in \Delta$)[1] which implies that (4.3.4) has a unique solution. The numerical approximation (4.3.4) was proposed in [And13][2], and further investigated in Chapter 3. In both these references the analysis of the MR method was restricted to the case that $A_a = 0$. The parameter $\beta \geq 1$ allows us to appropriately weight both terms in the least squares minimization.

The solution u^δ of the MR problem is the solution of the resulting Euler–Lagrange equations, which read as

$$({E_X^\delta}' B' E_Y^\delta K_Y^\delta {E_Y^\delta}' B E_X^\delta + {E_X^\delta}'\beta\gamma_0'\gamma_0 E_X^\delta)u^\delta = {E_X^\delta}' B' E_Y^\delta K_Y^\delta {E_Y^\delta}' g + {E_X^\delta}'\beta\gamma_0' u_0, \tag{4.3.6}$$

[1] This follows by combining (4.3.13), (4.3.15), and (4.3.16).

[2] In [And13], the norm $\|\gamma_0 E_X^\delta w - u_0\|$ reads as $\sup_{0\neq z\in Z^\delta} \frac{\langle \gamma_0 E_X^\delta w - u_0, z\rangle}{\|z\|}$ for some $H \supseteq Z^\delta \supseteq \operatorname{ran}\gamma_0|_{X^\delta}$ which generalization seems not very helpful.

as also the second component of the solution $(\mu^\delta, u^\delta) \in Y^\delta \times X^\delta$ of

$$\begin{bmatrix} (K_Y^\delta)^{-1} & {E_Y^\delta}' B E_X^\delta \\ {E_X^\delta}' B' E_Y^\delta & -{E_X^\delta}' \beta \gamma_0' \gamma_0 E_X^\delta \end{bmatrix} \begin{bmatrix} \mu^\delta \\ u^\delta \end{bmatrix} = \begin{bmatrix} {E_Y^\delta}' g \\ -{E_X^\delta}' \beta \gamma_0' u_0 \end{bmatrix}, \tag{4.3.7}$$

being a useful representation when no efficient preconditioner is available and one has to resort to $(K_Y^\delta)^{-1} = {E_Y^\delta}' A_s E_Y^\delta$.

With the "projected" or "approximate" (because generally $Y^\delta \neq Y$) *trial-to-test operator* $T^\delta \in \mathcal{L}(X, Y^\delta \times H)$ defined by

$$\begin{aligned} &((K_Y^\delta)^{-1} T^\delta w)(v_1) + \beta \langle T^\delta w, v_2 \rangle \\ &\qquad = (Bw)(E_Y^\delta v_1) + \beta \langle \gamma_0 u, v_2 \rangle \end{aligned} \qquad ((v_1, v_2) \in Y^\delta \times H), \tag{4.3.8}$$

and "projected" or "approximate" *optimal test space* $Z^\delta := \operatorname{ran} T^\delta|_{X^\delta}$, a third equivalent formulation of (4.3.4) (see e.g. [DG11], [BS14, Prop. 2.2], [DG14]) is finding $u^\delta \in X^\delta$ which solves the Petrov-Galerkin system

$$(BE_X^\delta u^\delta)(E_Y^\delta v_1) + \beta \langle \gamma_0 E_X^\delta u^\delta, v_2 \rangle = g(E_Y^\delta v_1) + \beta \langle u_0, v_2 \rangle \quad ((v_1, v_2) \in Z^\delta). \tag{4.3.9}$$

Note that (4.3.9) avoids the 'normal equations' (4.3.6). It will allow us to derive a quantitatively sharp estimate for the error in u^δ. From (4.3.3) and (4.3.5), one infers that $\sup_{0 \neq w \in X^\delta} \frac{\|T^\delta w\|_{Y \times H}}{\|w\|_X} > 0$, so that, thanks to X^δ being closed, Z^δ is a closed subspace of $Y^\delta \times H$. We orthogonally decompose $Y^\delta \times H$ into Z^δ and $(Z^\delta)^\perp$, where here we equip Y^δ with inner product $((K_Y^\delta)^{-1} \cdot)(\cdot)$. From (4.3.8) one infers that for $w \in X^\delta$ and $(v_1, v_2) \in (Z^\delta)^\perp$, it holds that $(Bw)(v_1) + \beta \langle \gamma_0 u, v_2 \rangle = 0$, and so

$$\begin{aligned} &\sup_{(v_1,v_2) \in Y^\delta \times H} \frac{(BE_X^\delta w)(E_Y^\delta v_1) + \beta \langle \gamma_0 E_X^\delta w, v_2 \rangle}{\sqrt{((K_Y^\delta)^{-1} v_1)(v_1) + \beta \|v_2\|^2}} \\ &\qquad = \sup_{(v_1,v_2) \in Z^\delta} \frac{(BE_X^\delta w)(E_Y^\delta v_1) + \beta \langle \gamma_0 E_X^\delta w, v_2 \rangle}{\sqrt{((K_Y^\delta)^{-1} v_1)(v_1) + \beta \|v_2\|^2}}. \end{aligned} \tag{4.3.10}$$

Theorem 4.3.1. *Under conditions* (4.3.1), (4.3.2), *and* (4.3.3), *the solution* $u^\delta \in X^\delta$ *of* (4.3.6) *exists uniquely, and satisfies*

$$\|u - u^\delta\|_X \leq \sqrt{\frac{\max(R_\Delta, 1)\left(1 + \frac{1}{2}\left(\alpha^2 + \alpha\sqrt{\alpha^2+4}\right)\right)}{\min(r_\Delta, 1)\frac{1}{2}\left((\gamma_\Delta^{\partial_t})^2 + \alpha^2 + 1 - \sqrt{((\gamma_\Delta^{\partial_t})^2 + \alpha^2 + 1)^2 - 4(\gamma_\Delta^{\partial_t})^2}\right)}} \inf_{w \in X^\delta} \|u - w\|_X.$$

(In particular, u^δ *is the best approximation to* u *from* X^δ *when* $\gamma_\Delta^{\partial_t} = r_\Delta = R_\Delta = 1$ *and* $\alpha = 0$.*)*

Remarks 4.3.2. One can verify that

$$\sqrt{\frac{\left(1+\frac{1}{2}\left(\alpha^2+\alpha\sqrt{\alpha^2+4}\right)\right)}{\frac{1}{2}\left((\gamma_\Delta^{\partial_t})^2+\alpha^2+1-\sqrt{((\gamma_\Delta^{\partial_t})^2+\alpha^2+1)^2-4(\gamma_\Delta^{\partial_t})^2}\right)}} \Big/ \frac{1+\frac{1}{2}\left(\alpha^2+\alpha\sqrt{\alpha^2+4}\right)}{\gamma_\Delta^{\partial_t}} \in \left[\tfrac{1}{2}\sqrt{3},1\right],$$

which clarifies the dependence of the bound from Theorem 4.3.1 on α and $\gamma_\Delta^{\partial_t}$. For $\alpha = 0$ (and $\beta = 1$), the estimate equals that of Thm. 3.3.7 & Rem. 3.3.8. $\Diamond$

Proof. Let u solve (4.2.4), i.e., $g = Bu$ and $u_0 = \gamma_0 u$. The mapping $P^\delta \in \mathcal{L}(X,X)$ from u to the solution $u^\delta \in X^\delta$ of (4.3.4) or, equivalently, (4.3.6) or (4.3.9), is a projector onto X^δ which, by assumption $X^\delta \notin \{0, X\}$, is unequal to 0 or Id. Consequently $\|P^\delta\|_{\mathcal{L}(X,X)} = \|\mathrm{Id} - P^\delta\|_{\mathcal{L}(X,X)}$ ([Kat60, XZ03]), and

$$\begin{aligned}\|u-u^\delta\|_X &= \|(\mathrm{Id}-P^\delta)u\|_X = \inf_{w\in X^\delta}\|(\mathrm{Id}-P^\delta)(u-w)\|_X \\ &\le \|P^\delta\|_{\mathcal{L}(X,X)} \inf_{w\in X^\delta}\|u-w\|_X.\end{aligned} \tag{4.3.11}$$

To bound $\|P^\delta\|_{\mathcal{L}(X,X)} = \sup_{0\neq w\in X}\frac{\|P^\delta w\|_X}{\|w\|_X}$, given $w \in X$, let $E_X^\delta w^\delta := P^\delta w$. Using (4.3.3), (4.3.10), (4.3.9), and Proposition 4.2.2 we estimate

$$\begin{aligned}
&\sup_{(v_1,v_2)\in Y^\delta\times H}\frac{\left((BE_X^\delta w^\delta)(E_Y^\delta v_1)+\beta\langle\gamma_0 E_X^\delta w^\delta, v_2\rangle\right)^2}{\|E_Y^\delta v_1\|_Y^2+\beta\|v_2\|^2} \\
&\le \tfrac{1}{\min(r_\Delta,1)}\sup_{(v_1,v_2)\in Y^\delta\times H}\frac{\left((BE_X^\delta w^\delta)(E_Y^\delta v_1)+\beta\langle\gamma_0 E_X^\delta w^\delta, v_2\rangle\right)^2}{((K_Y^\delta)^{-1}v_1)(v_1)+\beta\|v_2\|^2} \\
&= \tfrac{1}{\min(r_\Delta,1)}\sup_{(v_1,v_2)\in Z^\delta}\frac{\left((BE_X^\delta w^\delta)(E_Y^\delta v_1)+\beta\langle\gamma_0 E_X^\delta w^\delta, v_2\rangle\right)^2}{((K_Y^\delta)^{-1}v_1)(v_1)+\beta\|v_2\|^2} \\
&= \tfrac{1}{\min(r_\Delta,1)}\sup_{(v_1,v_2)\in Z^\delta}\frac{\left((Bw)(E_Y^\delta v_1)+\beta\langle\gamma_0 w, v_2\rangle\right)^2}{((K_Y^\delta)^{-1}v_1)(v_1)+\beta\|v_2\|^2} \\
&\le \tfrac{\max(R_\Delta,1)}{\min(r_\Delta,1)}\sup_{(v_1,v_2)\in Y\times H}\frac{\left((Bw)(v_1)+\beta\langle\gamma_0 w, v_2\rangle\right)^2}{\|v_1\|_Y^2+\beta\|v_2\|^2} \\
&= \tfrac{\max(R_\Delta,1)}{\min(r_\Delta,1)}|||w|||_X^2 \le \tfrac{\max(R_\Delta,1)}{\min(r_\Delta,1)}\left(1+\tfrac{1}{2}\left(\alpha^2+\alpha\sqrt{\alpha^2+4}\right)\right)\|w\|_X^2.
\end{aligned} \tag{4.3.12}$$

On the other hand,

$$
\begin{aligned}
&\sup_{(v_1,v_2)\in Y^\delta\times H}\frac{\Big((BE_X^\delta w^\delta)(E_Y^\delta v_1)+\beta\langle\gamma_0E_X^\delta w^\delta,v_2\rangle\Big)^2}{\|E_Y^\delta v_1\|_Y^2+\beta\|v_2\|^2}\\
&=\sup_{(v_1,v_2)\in Y^\delta\times H}\frac{\Big((A_sE_Y^\delta(E_Y^{\delta\prime}A_sE_Y^\delta)^{-1}E_Y^{\delta\prime}BE_X^\delta w^\delta)(E_Y^\delta v_1)+\beta\langle\gamma_0E_X^\delta w^\delta,v_2\rangle\Big)^2}{\|E_Y^\delta v_1\|_Y^2+\beta\|v_2\|^2}\\
&=\sup_{(v_1,v_2)\in Y^\delta\times H}\frac{\Big(\langle E_Y^\delta(E_Y^{\delta\prime}A_sE_Y^\delta)^{-1}E_Y^{\delta\prime}BE_X^\delta w^\delta,E_Y^\delta v_1\rangle_Y+\beta\langle\gamma_0E_X^\delta w^\delta,v_2\rangle\Big)^2}{\|E_Y^\delta v_1\|_Y^2+\beta\|v_2\|^2}\\
&=\|E_Y^\delta(E_Y^{\delta\prime}A_sE_Y^\delta)^{-1}E_Y^{\delta\prime}BE_X^\delta w^\delta\|_Y^2+\beta\|\gamma_0E_X^\delta w^\delta\|^2\\
&=(A_sE_Y^\delta(E_Y^{\delta\prime}A_sE_Y^\delta)^{-1}E_Y^{\delta\prime}BE_X^\delta w^\delta)(E_Y^\delta(E_Y^{\delta\prime}A_sE_Y^\delta)^{-1}E_Y^{\delta\prime}BE_X^\delta w^\delta)\\
&\qquad\qquad+\beta(E_X^{\delta\prime}\gamma_0'\gamma_0E_X^\delta w^\delta)(w^\delta)\\
&=\Big((E_Y^{\delta\prime}B'E_Y^\delta(E_Y^{\delta\prime}A_sE_Y^\delta)^{-1}E_Y^{\delta\prime}BE_X^\delta+\beta E_X^{\delta\prime}\gamma_0'\gamma_0E_X^\delta)w^\delta\Big)(w^\delta).
\end{aligned}
\tag{4.3.13}
$$

Using (4.3.1), we write $E_X^\delta=E_Y^\delta F^\delta$ with F^δ the trivial embedding $X^\delta\to Y^\delta$. Using $B=C+A_s$ and $C+C'+\gamma_0'\gamma_0=\gamma_T'\gamma_T$, similar to (4.2.6) we find

$$
\begin{aligned}
&E_X^{\delta\prime}B'E_Y^\delta(E_Y^{\delta\prime}A_sE_Y^\delta)^{-1}E_Y^{\delta\prime}BE_X^\delta+E_X^{\delta\prime}\beta\gamma_0'\gamma_0E_X^\delta\\
&=F^{\delta\prime}\Big(E_Y^{\delta\prime}B'E_Y^\delta(E_Y^{\delta\prime}A_sE_Y^\delta)^{-1}E_Y^{\delta\prime}BE_Y^\delta+E_Y^{\delta\prime}\beta\gamma_0'\gamma_0E_Y^\delta\Big)F^\delta\\
&=F^{\delta\prime}\Big(E_Y^{\delta\prime}\Big(C'E_Y^\delta(E_Y^{\delta\prime}A_sE_Y^\delta)^{-1}E_Y^{\delta\prime}C+A_s+(\gamma_T'\gamma_T+(\beta-1)\gamma_0'\gamma_0)\Big)E_Y^\delta\Big)F^\delta\\
&=E_X^{\delta\prime}\Big(C'E_Y^\delta(E_Y^{\delta\prime}A_sE_Y^\delta)^{-1}E_Y^{\delta\prime}C+A_s+(\gamma_T'\gamma_T+(\beta-1)\gamma_0'\gamma_0)\Big)E_X^\delta.
\end{aligned}
\tag{4.3.14}
$$

We conclude that for any $\eta\in(0,1]$,

$$
\begin{aligned}
&\Big((E_X^{\delta\prime}B'E_Y^\delta(E_Y^{\delta\prime}A_sE_Y^\delta)^{-1}E_Y^{\delta\prime}BE_X^\delta+E_X^{\delta\prime}\beta\gamma_0'\gamma_0E_X^\delta)w^\delta\Big)(w^\delta)\\
&\quad=\|E_Y^{\delta\prime}CE_X^\delta w^\delta\|_{Y^{\delta\prime}}^2+\|E_X^\delta w^\delta\|_Y^2+\|\gamma_TE_X^\delta w^\delta\|^2+(\beta-1)\|\gamma_0E_X^\delta w^\delta\|^2\\
&\quad\geq(\|E_Y^{\delta\prime}\partial_tE_X^\delta w^\delta\|_{Y^{\delta\prime}}-\alpha\|E_X^\delta w^\delta\|_Y)^2+\|E_X^\delta w^\delta\|_Y^2+\|\gamma_TE_X^\delta w^\delta\|^2\\
&\qquad\qquad+(\beta-1)\|\gamma_0E_X^\delta w^\delta\|^2\\
&\quad\geq(1-\eta^2)\|E_Y^{\delta\prime}\partial_tE_X^\delta w^\delta\|_{Y^{\delta\prime}}^2+\Big((1-\eta^{-2})\alpha^2+1\Big)\|E_X^\delta w^\delta\|_Y^2+\|\gamma_TE_X^\delta w^\delta\|^2\\
&\qquad\qquad+(\beta-1)\|\gamma_0E_X^\delta w^\delta\|^2\\
&\quad\overset{(4.3.2)}{\geq}(1-\eta^2)(\gamma_\Delta^{\partial_t})^2\|\partial_tE_X^\delta w^\delta\|_{Y'}^2+\Big((1-\eta^{-2})\alpha^2+1\Big)\|E_X^\delta w^\delta\|_Y^2+\|\gamma_TE_X^\delta w^\delta\|^2\\
&\qquad\qquad+(\beta-1)\|\gamma_0E_X^\delta w^\delta\|^2\\
&\quad\geq\min\Big((1-\eta^2)(\gamma_\Delta^{\partial_t})^2,\Big((1-\eta^{-2})\alpha^2+1\Big)\Big)\|E_X^\delta w^\delta\|_X^2,
\end{aligned}
\tag{4.3.15}
$$

with Young's inequality. Solving $(1-\eta^2)(\gamma_\Delta^{\partial_t})^2 = (1-\eta^{-2})\alpha^2 + 1$ for η yields

$$(1-\eta^2)(\gamma_\Delta^{\partial_t})^2 = \tfrac{1}{2}\left((\gamma_\Delta^{\partial_t})^2 + \alpha^2 + 1 - \sqrt{((\gamma_\Delta^{\partial_t})^2 + \alpha^2 + 1)^2 - 4(\gamma_\Delta^{\partial_t})^2}\right) > 0. \tag{4.3.16}$$

Recalling (4.3.11) and $\|P^\delta\|_{\mathcal{L}(X,X)} = \sup_{0\neq w\in X} \frac{\|w^\delta\|_X}{\|w\|_X}$, the proof is completed by combining (4.3.12), (4.3.13), and (4.3.15). □

4.4 Brézis–Ekeland–Nayroles (BEN) formulation

The minimizer $u \in X$ of $\left\| \begin{bmatrix} B \\ \sqrt{\beta}\,\gamma_0 \end{bmatrix} w - \begin{bmatrix} g \\ \sqrt{\beta}\,u_0 \end{bmatrix} \right\|^2_{Y'\times H}$, that is equal to the unique solution of (4.2.4), is the unique solution of

$$(B' A_s^{-1} B + \beta \gamma_0' \gamma_0) u = B' A_s^{-1} g + \beta \gamma_0' u_0. \tag{4.4.1}$$

As we have seen in (4.2.6), this system is equivalent to

$$(C' A_s^{-1} C + A_s + \gamma_T' \gamma_T + (\beta-1)\gamma_0'\gamma_0) u = (\mathrm{Id} + C' A_s^{-1}) g + \beta \gamma_0' u_0, \tag{4.4.2}$$

showing that u is the second component of the pair $(\lambda, u) \in Y \times X$ solving

$$\begin{bmatrix} A_s & C \\ C' & -(A_s + \gamma_T'\gamma_T + (\beta-1)\gamma_0'\gamma_0) \end{bmatrix} \begin{bmatrix} \lambda \\ u \end{bmatrix} = \begin{bmatrix} g \\ -(g + \beta\gamma_0' u_0) \end{bmatrix}. \tag{4.4.3}$$

Notice that $\lambda = u$.

The formulation (4.4.2) of the parabolic equation can alternatively be derived from the application of the Brézis–Ekeland–Nayroles variational principle ([BE76, Nay76], cf. also [And12, §3.2.4]), which generalizes beyond the linear, Hilbert space setting.

Given $\delta \in \Delta$, we consider the Galerkin discretization of (4.4.3), i.e.,

$$\begin{bmatrix} {E_Y^\delta}' A_s E_Y^\delta & {E_Y^\delta}' C E_X^\delta \\ ({E_Y^\delta}' C E_X^\delta)' & -{E_X^\delta}'(A_s + \gamma_T'\gamma_T + (\beta-1)\gamma_0'\gamma_0) E_X^\delta \end{bmatrix} \begin{bmatrix} \lambda^\delta \\ \bar{u}^\delta \end{bmatrix} = \begin{bmatrix} {E_Y^\delta}' g \\ -{E_X^\delta}'(g + \beta\gamma_0' u_0) \end{bmatrix} \tag{4.4.4}$$

or, equivalently

$$\begin{aligned} {E_X^\delta}'\Big(C' E_Y^\delta ({E_Y^\delta}' A_s E_Y^\delta)^{-1} {E_Y^\delta}' C + A_s + \gamma_T'\gamma_T + (\beta-1)\gamma_0'\gamma_0\Big) E_X^\delta \bar{u}^\delta \\ = {E_X^\delta}'\Big(C' E_Y^\delta ({E_Y^\delta}' A_s E_Y^\delta)^{-1} {E_Y^\delta}' g + g + \beta\gamma_0' u_0\Big). \end{aligned} \tag{4.4.5}$$

Remark 4.4.1. Assuming $X^\delta \subseteq Y^\delta$ ((4.3.1)) and $K_Y^\delta = ({E_Y^\delta}' A_s E_Y^\delta)^{-1}$, it holds that $\bar{u}^\delta = u^\delta$, i.e., the solutions of BEN and MR *are equal*. Indeed, (4.3.14) shows that in this case the operator at the left-hand side of (4.4.5) equals the

operator in (4.3.6), and from $E_X^{\delta\,\prime} A_s E_Y^\delta (E_Y^{\delta\,\prime} A_s E_Y^\delta)^{-1} E_Y^{\delta\,\prime} = E_X^{\delta\,\prime}$ when $X^\delta \subseteq Y^\delta$ one deduces that also the right-hand sides agree.

In contrast to MR, with BEN, it is *not* possible to replace $(E_Y^{\delta\,\prime} A_s E_Y^\delta)^{-1}$ by a general preconditioner as in (4.3.7)–(4.3.6) and still obtain a quasi-best approximation to (λ, u) from $Y^\delta \times X^\delta$. This can be understood by noticing that replacing A_s^{-1} in (4.4.2) by another operator changes the solution, whereas this is not the case in (4.4.1). So for the iterative solution of BEN one has to operate on the saddle point system (4.4.4) instead of on a symmetric positive definite system as with MR, see (4.3.6).

On the other hand, BEN doesn't require $X^\delta \subseteq Y^\delta$, as we will see below. ♢

The applicability of BEN for the case $A_a \neq 0$ was already demonstrated in Chapter 3. The following result gives a quantitatively better error bound.

Theorem 4.4.2. *Under the sole condition* (4.3.2)*, the solution* $\bar{u}^\delta \in X^\delta$ *of* (4.4.5) *exists uniquely, and satisfies*

$$\|u - \bar{u}^\delta\|_X \leq \frac{\left(1+\frac{1}{2}\left(\alpha^2+\alpha\sqrt{\alpha^2+4}\right)\right) \inf_{w\in X^\delta} \|u-w\|_X + \sqrt{1+\alpha^2} \inf_{v\in Y^\delta} \|u-v\|_Y}{\frac{1}{2}\left((\gamma_\Delta^{\partial_t})^2+\alpha^2+1-\sqrt{((\gamma_\Delta^{\partial_t})^2+\alpha^2+1)^2-4(\gamma_\Delta^{\partial_t})^2}\right)}.$$

Proof. With $g = Bu$ and $u_0 = \gamma_0 u$, using $B = C + A_s$ and $\gamma_0'\gamma_0 = \gamma_T'\gamma_T - (C' + C)$, the right-hand side of (4.4.5) reads as

$$\begin{aligned} &E_X^{\delta\,\prime}\Big(C' E_Y^\delta (E_Y^{\delta\,\prime} A_s E_Y^\delta)^{-1} E_Y^{\delta\,\prime}(C + A_s) + A_s + \gamma_T'\gamma_T + (\beta-1)\gamma_0'\gamma_0 - C'\Big)u = \\ &E_X^{\delta\,\prime}\Big(C' E_Y^\delta (E_Y^{\delta\,\prime} A_s E_Y^\delta)^{-1} E_Y^{\delta\,\prime} C + A_s + \gamma_T'\gamma_T + (\beta-1)\gamma_0'\gamma_0 \\ &\qquad + C'\Big[E_Y^\delta (E_Y^{\delta\,\prime} A_s E_Y^\delta)^{-1} E_Y^{\delta\,\prime} A_s - \mathrm{Id}\Big]\Big)u. \end{aligned}$$

So with $G(\delta) := C' E_Y^\delta (E_Y^{\delta\,\prime} A_s E_Y^\delta)^{-1} E_Y^{\delta\,\prime} C + A_s + \gamma_T'\gamma_T + (\beta-1)\gamma_0'\gamma_0$, we have

$$u \mapsto E_X^\delta \bar{u}^\delta = E_X^\delta (E_X^{\delta\,\prime} G(\delta) E_X^\delta)^{-1} E_X^{\delta\,\prime}\Big(G(\delta) + C'\Big[E_Y^\delta (E_Y^{\delta\,\prime} A_s E_Y^\delta)^{-1} E_Y^{\delta\,\prime} A_s - \mathrm{Id}\Big]\Big)u,$$

where we already used that $E_X^{\delta\,\prime} G(\delta) E_X^\delta$ is invertible, which we verify below. Since $E_X^\delta (E_X^{\delta\,\prime} G(\delta) E_X^\delta)^{-1} E_X^{\delta\,\prime} G(\delta) \in \mathcal{L}(X,X)$ and $E_Y^\delta (E_Y^{\delta\,\prime} A_s E_Y^\delta)^{-1} E_Y^{\delta\,\prime} A_s \in \mathcal{L}(Y,Y)$ are projectors onto X^δ and Y^δ, respectively, the latter being orthogonal, for any $v \in Y^\delta$ and $w \in X^\delta$ it holds that

$$\begin{aligned} u - \bar{u}^\delta =& (\mathrm{Id} - E_X^\delta (E_X^{\delta\,\prime} G(\delta) E_X^\delta)^{-1} E_X^{\delta\,\prime} G(\delta))(u - E_X^\delta w) \\ &+ E_X^\delta (E_X^{\delta\,\prime} G(\delta) E_X^\delta)^{-1} E_X^{\delta\,\prime} C'\Big[\mathrm{Id} - E_Y^\delta (E_Y^{\delta\,\prime} A_s E_Y^\delta)^{-1} E_Y^{\delta\,\prime} A_s\Big](u - E_Y^\delta v) \end{aligned}$$

and so, also using $Y^\delta \notin \{0, Y\}$,

$$\begin{aligned} \|u - \bar{u}^\delta\|_X \leq \|(E_X^{\delta\,\prime} G(\delta) E_X^\delta)^{-1}\|_{\mathcal{L}(X^{\delta\prime}, X^\delta)} \Big\{ &\|G(\delta)\|_{\mathcal{L}(X,X')} \inf_{w\in X^\delta} \|u-w\|_X \\ &+ \|C\|_{\mathcal{L}(X,Y')} \inf_{v\in Y^\delta} \|u-v\|_Y \Big\}. \end{aligned}$$

For $w \in X$, we have

$$\begin{aligned}(G(\delta)w)(w) &= \|{E_Y^\delta}' C w\|^2_{Y^{\delta\prime}} + \|w\|_Y^2 + \|\gamma_T w\|^2 + (\beta-1)\|\gamma_0 w\|^2 \\ &\leq \|Cw\|^2_{Y'} + \|w\|_Y^2 + \|\gamma_T w\|^2 + (\beta-1)\|\gamma_0 w\|^2 \\ &= ((C' A_s^{-1} C + A_s + \gamma_T' \gamma_T + (\beta-1)\gamma_0'\gamma_0)w)(w) = \|Bw\|^2_{Y'} + \beta\|\gamma_0 w\|^2 \\ &\leq \left(1 + \tfrac{1}{2}\left(\alpha^2 + \alpha\sqrt{\alpha^2+4}\right)\right)\|w\|_X^2\end{aligned}$$

by Proposition 4.2.2. Since $(G(\delta)\cdot)(\cdot)$ is symmetric semi-positive-definite, we conclude that $\|G(\delta)\|_{\mathcal{L}(X,X')} \leq 1 + \frac{1}{2}\left(\alpha^2 + \alpha\sqrt{\alpha^2+4}\right)$.

For $w \in X^\delta$, one deduces

$$\begin{aligned}&(G(\delta)E_X^\delta w)(E_X^\delta w) \\ &\qquad = \|{E_Y^\delta}' C E_X^\delta w\|^2_{Y^{\delta\prime}} + \|E_X^\delta w\|_Y^2 + \|\gamma_T E_X^\delta w\|^2 + (\beta-1)\|\gamma_0 E_X^\delta w\|^2 \\ &\qquad \geq \tfrac{1}{2}\left((\gamma_\Delta^{\partial_t})^2 + \alpha^2 + 1 - \sqrt{((\gamma_\Delta^{\partial_t})^2 + \alpha^2 + 1)^2 - 4(\gamma_\Delta^{\partial_t})^2}\right)\|E_X^\delta w\|_X^2\end{aligned}$$

by following the lines starting at the second line of (4.3.15), in particular showing that ${E_X^\delta}' G(\delta) E_X^\delta$ is invertible.

Finally, for $w \in X$, $\|Cw\|_{Y'} \leq \|\partial_t w\|_{Y'} + \alpha\|w\|_Y \leq \sqrt{1+\alpha^2}\,\|w\|_X$. The theorem follows by combining the above estimates. □

4.5 Stable subspaces and preconditioners

By the boundedness and coercivity assumptions (4.2.1) and (4.2.5), it holds that $\|\cdot\|_Y \eqsim \|\cdot\|_{L_2(I;V)}$. Since with

$$\gamma^\delta := \gamma^\delta(X^\delta, Y^\delta) := \inf_{\{w \in X^\delta :\, \partial_t w \neq 0\}} \sup_{0 \neq v \in Y^\delta} \frac{\int_I \langle \partial_t w, v\rangle \,\mathrm{d}t}{\|\partial_t w\|_{L_2(I;V')}\|v\|_{L_2(I;V)}}, \tag{4.5.1}$$

consequently it holds that $\gamma_\Delta^{\partial_t} \eqsim \inf_{\delta\in\Delta}\gamma^\delta$, we will summarize some known results about settings for which $\inf_{\delta\in\Delta}\gamma^\delta > 0$ has been demonstrated.

In the final subsection of this section we will briefly comment on the construction of preconditioners at the Y-side, i.e. condition (4.3.3), and the X-side. The preconditioner K_Y^δ has its application for the reduction of the saddle-point system (4.3.7) (reading $(K_Y^\delta)^{-1}$ as ${E_Y^\delta}' A_s E_Y^\delta$) to the elliptic system (4.3.6), and as an ingredient for building a preconditioner for the saddle-point system (4.4.4), whereas K_X^δ can be applied for preconditioning (4.3.6), and as the other ingredient to construct a preconditioner for (4.4.4).

Since inf-sup conditions of the type $\gamma^\delta > 0$, or $\inf_{\delta\in\Delta}\gamma^\delta > 0$, will be encountered more often in this work, in an abstract setting we recall their relation with existence of certain Fortin interpolators, see Theorem 7.3.11.

Proposition 4.5.1. *Let $\widetilde{\mathcal{A}}$ and $\widetilde{\mathcal{B}}$ be closed subspaces of Hilbert spaces $\mathcal{A}$ and $\mathcal{B}$, and let $F \in \mathcal{L}(\mathcal{A}, \mathcal{B}')$. Let*

$$Q \in \mathcal{L}(\mathcal{B}, \mathcal{B}) \text{ with } \operatorname{ran} Q \subset \widetilde{\mathcal{B}} \text{ and } (\mathrm{Id} - Q')F\widetilde{\mathcal{A}} = 0. \tag{4.5.2}$$

Then $\mathfrak{G} := \inf_{\{a \in \widetilde{\mathcal{A}}\colon Fa \neq 0\}} \sup_{0 \neq b \in \widetilde{\mathcal{B}}} \frac{(Fa)(b)}{\|Fa\|_{\mathcal{B}'}\|b\|_{\mathcal{B}}} \geq \|Q\|^{-1}_{\mathcal{L}(\mathcal{B},\mathcal{B})}$. Conversely, if $\mathfrak{G} > 0$, then there exists a projector Q as in (4.5.2)*, and $\|Q\|_{\mathcal{L}(\mathcal{B},\mathcal{B})} \leq 2 + 1/\mathfrak{G}$.*

4.5.1 'Full' tensor product case

Concerning the verification of $\inf_{\delta \in \Delta} \gamma^\delta > 0$, we start with the easy case of X^δ and Y^δ being *'full' tensor products* of approximation spaces in time and space (as opposed to sparse tensor products, see below). With $Y_t := L_2(I)$ and $X_t := H^1(I)$, for $Z \in \{X, Y\}$ let $(Z_t^\delta)_{\delta \in \Delta}$ and $(Z_x^\delta)_{\delta \in \Delta}$ be families of closed subspaces of Z_t and V, respectively, and let $Z^\delta := Z_t^\delta \otimes Z_x^\delta$. Assuming that

$$\gamma_t^\delta := \inf_{\{w \in X_t^\delta\colon w' \neq 0\}} \sup_{0 \neq v \in Y_t^\delta} \frac{\int_I w' v \,\mathrm{d}t}{\|w'\|_{L_2(I)}\|v\|_{L_2(I)}} \gtrsim 1, \tag{4.5.3}$$

$$\gamma_x^\delta := \inf_{0 \neq w \in X_x^\delta} \sup_{0 \neq v \in Y_x^\delta} \frac{\langle w, v\rangle}{\|w\|_{V'}\|v\|_V} \gtrsim 1, \tag{4.5.4}$$

a tensor product argument shows that

$$\gamma^\delta = \gamma_t^\delta \gamma_x^\delta \gtrsim 1.$$

Obviously, (4.5.3) is true when $\frac{\mathrm{d}}{\mathrm{d}t} X_t^\delta \subseteq Y_t^\delta$, which however is not a necessary condition. For example, when X_t^δ is the space of continuous piecewise linears w.r.t. some partition of I, and Y_t^δ is the space of continuous piecewise linears w.r.t. a once dyadically refined partition, an easy computation ([And13, Prop. 6.1]) shows that $\gamma_t^\delta \geq \sqrt{3/4}$.

Considering, for a domain $\Omega \subset \mathbb{R}^d$ and $\Gamma \subset \partial\Omega$, $H = L_2(\Omega)$ and $V = H^1_{0,\Gamma}(\Omega) := \{v \in H^1(\Omega)\colon v|_\Gamma = 0\}$, $H^1(\Omega)$-stability of the $L_2(\Omega)$-orthogonal projector onto Lagrange finite element spaces $X_x^\delta = Y_x^\delta$ is an extensively studied subject. In view of Proposition 4.5.1, taking F to be the Riesz map $H \to H'$ viewed as a mapping $V \to V'$, this stability implies (4.5.4). For finite element spaces w.r.t. shape regular quasi-uniform partitions into, say, d-simplices, where Γ is the union of faces of $T \in \mathcal{T}$, stability follows easily from direct and inverse estimates. It is known that this stability holds also true for (shape regular) locally refined partitions when they are sufficiently mildly graded. In [GHS16], it is shown that in two space dimensions the meshes generated by newest vertex bisection satisfy this requirement, see also [DST20] for extensions.

4.5.2 Sparse tensor product case

As shown in [And13, Prop. 4.2], these results for full tensor products extend to *sparse tensor products*. When $(Z_t^\delta)_{\delta \in \Delta}$ and $(Z_x^\delta)_{\delta \in \Delta}$ are nested sequences of

closed subspaces $Z_t^{\delta_0} \subset Z_t^{\delta_1} \subset \cdots \subset Z_t$, $Z_x^{\delta_0} \subset Z_x^{\delta_1} \subset \cdots \subset V$ which satisfy (4.5.3)–(4.5.4), then for $Z^{\delta_n} := \sum_{\{0 \le n_t + n_x \le n\}} Z_t^{\delta_{n_t}} \otimes Z_x^{\delta_{n_x}}$ we have

$$\gamma^{\delta_n} \ge \min_{0 \le n_t \le n} \gamma_t^{\delta_{n_t}} \min_{0 \le n_x \le n} \gamma_t^{\delta_{n_x}} \gtrsim 1.$$

4.5.3 Time-slab partition case

Another extension of the full tensor product case is given by the following. Let $(\bar{X}^\delta, \bar{Y}^\delta)_{\delta \in \bar{\Delta}}$ be a family of pairs of closed subspaces of X and Y for which

$$\gamma_{\bar{\Delta}} := \inf_{\delta \in \bar{\Delta}} \inf_{\{w \in \bar{X}^\delta \colon \partial_t w \neq 0\}} \sup_{0 \neq v \in \bar{Y}^\delta} \frac{\int_I \langle \partial_t w, v \rangle \, \mathrm{d}t}{\|w\|_{L_2(I;V')} \|v\|_{L_2(I;V)}} > 0.$$

Then if, for $\delta \in \Delta$, X^δ and Y^δ are such that for some finite partition $I^\delta = ([t_{i-1}^\delta, t_i^\delta])_i$ of I, with $G_i^\delta(t) := t_{i-1}^\delta + \frac{t}{T}(t_i^\delta - t_{i-1}^\delta)$ and arbitrary $\delta_i \in \bar{\Delta}$ we have

$$X^\delta \subseteq \{u \in X \colon u|_{(t_{i-1}^\delta, t_i^\delta)} \circ G_i^\delta \in \bar{X}^{\delta_i}\},$$
$$Y^\delta \supseteq \{v \in L_2(I;V) \colon v|_{(t_{i-1}^\delta, t_i^\delta)} \circ G_i^\delta \in \bar{Y}^{\delta_i}\},$$

then $\gamma^\delta \ge \gamma_{\bar{\Delta}} > 0$ as one easily verifies using $\int_I \langle \frac{\mathrm{d}u}{\mathrm{d}t}, v \rangle \, \mathrm{d}t = \sum_i \int_{t_{i-1}}^{t_i} \langle \frac{\mathrm{d}u}{\mathrm{d}t}, v \rangle \, \mathrm{d}t$. An example of this *'time-slab partition'* setting will be given in Sect. 4.7. Thinking of the $\bar{X}^\delta$ as being finite element spaces, notice that the condition $X^\delta \subset X$ will require that possible 'hanging nodes' on the interface between different time slabs do not carry degrees of freedom.

4.5.4 Generalized sparse tensor product case

Finally, we informally describe a *'generalized' sparse tensor product* setting that allows for *local refinements* driven by an a posteriori error estimator. For $Z \in \{X, Y\}$, let the nested sequences of closed subspaces $Z_t^{\delta_0} \subset Z_t^{\delta_1} \subset \cdots \subset Z_t$, $Z_x^{\delta_0} \subset Z_x^{\delta_1} \subset \cdots \subset V$ be equipped with hierarchical bases, meaning that the basis for $Z_t^{\delta_i}$ (analogously $Z_x^{\delta_i}$) is inductively defined as the basis for $Z_t^{\delta_{i-1}}$ plus a basis for a complement space of $Z_t^{\delta_{i-1}}$ in $Z_t^{\delta_i}$. The *level* of the functions in the latter basis is defined as i.

Let us consider the usual case that the diameter of the support of a hierarchical basis function with level i is $\eqsim 2^{-i}$, and let us assign to each basis function ϕ on level $i > 0$ one (or a few) parents with level $i-1$ whose supports intersect the support of ϕ. We now let $(Z^\delta)_{\delta \in \Delta}$ be the collection of all spaces that are spanned by sets of product hierarchical basis functions, which sets are *downward closed* (or *lower*) in the sense that if a product of basis functions is in the set, then so are all their parents in both directions. Note that the sparse tensor product spaces $\sum_{\{0 \le n_t + n_x \le n\}} Z_t^{\delta_{n_t}} \otimes Z_x^{\delta_{n_x}}$ are included in this collection, but that it contains many more spaces.

Under conditions on the hierarchical bases for $Z_t^{\delta_0} \subset Z_t^{\delta_1} \subset \cdots \subset Z_t$ for $Z \in \{X, Y\}$, which should be of *wavelet-type*, in [SvVW21] it is shown that to any X^δ one can assign a Y^δ with $\dim Y^\delta \lesssim \dim X^\delta$, such that $\gamma^\delta \gtrsim 1$ holds.

4.5.5 Preconditioners

Moving to condition (4.3.3), obviously we would like to construct K_Y^δ such that it is not only a uniform preconditioner, i.e., it satisfies (4.3.3), but also that its application can be performed in $\mathcal{O}(\dim Y^\delta)$ operations. In the full-tensor product case, after selecting bases for Y_t^δ and Y_x^δ, the construction of K_Y^δ boils down to tensorizing approximate inverses of the 'mass matrix' in time, which does not pose any problems, and the 'stiffness matrix' in space. For $V = H^1(\Omega)$ (or a subspace of aforementioned type), it is well-known that by taking a multi-grid preconditioner as the approximate inverse of the stiffness matrix the resulting K_Y^δ satisfies our needs. A straightforward generalization of this construction of K_Y^δ applies to spaces Y^δ that correspond to the time-slab partitioning approach.

Finally, for the efficient iterative solution of (4.3.6) or (4.4.4), one needs a $K_X^\delta = {K_X^\delta}' \in \mathcal{L}\mathrm{is}(X^{\delta\prime}, X^\delta)$ whose norm and norm of its inverse are uniformly bounded, and whose application can be performed in $\mathcal{O}(\dim X^\delta)$ operations. For the full and generalized sparse tensor product setting such preconditioners have been constructed in [And16] and [SvVW21], respectively.

4.6 Robustness

The quasi-optimality results presented in Theorems 4.3.1 and 4.4.2 for MR and BEN degenerate when $\alpha = \|A_a\|_{\mathcal{L}(Y,Y')} \to \infty$. Aiming at results that are robust for $\alpha \to \infty$, we now study convergence w.r.t. the energy-norm $|||\cdot|||_X$ on X. On its own this change of norms turns out not to be helpful. By replacing $\|\cdot\|_X$ by $|||\cdot|||_X$ in Theorems 4.3.1 and 4.4.2, and adapting their proofs in an obvious way yields for MR the same upper bound for $\frac{|||u-u^\delta|||_X}{\inf_{w\in X^\delta}|||u-w|||_X}$ as we found for $\frac{\|u-u^\delta\|_X}{\inf_{w\in X^\delta}\|u-w\|_X}$ (for $u \notin X^\delta$), whereas instead of Theorem 4.4.2 we arrive at the only slightly more favourable bound

$$|||u-\bar{u}^\delta|||_X \leq \frac{2+\alpha^2+\alpha\sqrt{\alpha^2+4}}{(\gamma_\Delta^{\partial_t})^2+\alpha^2+1-\sqrt{((\gamma_\Delta^{\partial_t})^2+\alpha^2+1)^2-4(\gamma_\Delta^{\partial_t})^2}} \inf_{w\in X^\delta, v\in Y^\delta} |||u-w|||_X + \|u-v\|_Y,$$

which is, however, still far from being robust.

In order to obtain robust bounds, instead of the condition $\gamma_\Delta^{\partial_t} > 0$ ((4.3.2)) we now impose

$$\gamma_\Delta^{\mathrm{C}} := \inf_{\delta\in\Delta} \inf_{\{0\neq w\in X^\delta:\, CE_X^\delta w\neq 0\}} \frac{\|{E_Y^\delta}' CE_X^\delta w\|_{Y^{\delta\prime}}}{\|CE_X^\delta w\|_{Y'}} > 0, \tag{4.6.1}$$

which, when considering a family of operators A, we would like to hold uniformly for $\alpha \to \infty$.

Theorem 4.6.1. *Under conditions* (4.3.1), (4.6.1), *and* (4.3.3), *the solution* $u^\delta \in X^\delta$ *of* (4.3.6) *satisfies*

$$|||u-u^\delta|||_X \le \sqrt{\tfrac{\max(R_\Delta,1)}{\min(r_\Delta,1)}}\,(\gamma_\Delta^{\mathrm{C}})^{-1}\inf_{w\in X^\delta}|||u-w|||_X; \tag{4.6.2}$$

and under condition (4.6.1), *the solution* $\bar{u}^\delta \in X^\delta$ *of* (4.4.5) *satisfies*

$$|||u-\bar{u}^\delta|||_X \le (\gamma_\Delta^{\mathrm{C}})^{-2}\Big\{\inf_{w\in X^\delta}|||u-w|||_X + \inf_{v\in Y^\delta}\|u-v\|_Y\Big\}. \tag{4.6.3}$$

Proof. The first estimate follows from ignoring the last inequality in (4.3.12), and by replacing the first inequality in (4.3.15) by

$$\begin{aligned}
&\|{E_Y^\delta}'CE_X^\delta w^\delta\|^2_{Y^{\delta\prime}} + \|E_X^\delta w^\delta\|_Y^2 + \|\gamma_T E_X^\delta w^\delta\|^2 + (\beta-1)\|\gamma_0 E_X^\delta w^\delta\|^2\\
&\ge (\gamma_\Delta^{\mathrm{C}})^2\Big(\|CE_X^\delta w^\delta\|^2_{Y'} + \|E_X^\delta w^\delta\|_Y^2 + \|\gamma_T E_X^\delta w^\delta\|^2 + (\beta-1)\|\gamma_0 E_X^\delta w^\delta\|^2\Big)\\
&= (\gamma_\Delta^{\mathrm{C}})^2\Big({E_X^\delta}'B'A_s^{-1}BE_X^\delta + {E_X^\delta}'\beta\gamma_0'\gamma_0E_X^\delta)w^\delta\Big)(w^\delta) = (\gamma_\Delta^{\mathrm{C}})^2|||w^\delta|||_X^2.
\end{aligned}$$

Following the proof of Theorem 4.4.2, but now equipping X with $|||\cdot|||_X$, from $\|C\|_{\mathcal{L}(X,Y')} \le 1$, $\|G(\delta)\|_{\mathcal{L}(X,X')} \le 1$, and $\|({E_X^\delta}'G(\delta)E_X^\delta)^{-1}\|_{\mathcal{L}(X^{\delta\prime},X^\delta)} \le (\gamma_\Delta^{\mathrm{C}})^{-2}$, one infers the estimate for BEN. □

We conclude that for a family of operators A robustness w.r.t. $|||\cdot|||_X$ is obtained when $(\gamma_\Delta^{\mathrm{C}})^{-1}$ is uniformly bounded for $\alpha = \|A_a\|_{\mathcal{L}(Y,Y')} \to \infty$. A family for which this will be realized is presented in Sect. 4.7.

4.6.1 A posteriori error estimation

In particular because for $\alpha = \|A_a\|_{\mathcal{L}(Y,Y')} \to \infty$ meaningful a priori error bounds for $\inf_{w\in X^\delta}|||u-w|||_X$ will be hard to derive, it is important to have (robust) a posteriori error bounds.

Let $Q_B^\delta \in \mathcal{L}(Y,Y)$ be such that $\operatorname{ran} Q_B^\delta \subset Y^\delta$ and $(\mathrm{Id}-{Q_B^\delta}')BX^\delta = 0$. Then, with $e^\delta_{\mathrm{osc}}(g) := \|(\mathrm{Id}-{Q_B^\delta}')g\|_{Y'}$, for $w \in X^\delta$ and u the solution of (4.2.4) it holds that

$$\begin{aligned}
r_\Delta\|{E_Y^\delta}'(g-Bw)\|^2_{K_Y^\delta} + \beta\|u_0-\gamma_0 w\|^2 \le |||u-w|||_X^2 \le\\
\Big(\|Q_B^\delta\|_{\mathcal{L}(Y,Y)}\sqrt{R_\Delta}\|{E_Y^\delta}'(g-Bw)\|_{K_Y^\delta} + e^\delta_{\mathrm{osc}}(g)\Big)^2 + \beta\|u_0-\gamma_0 w\|^2,
\end{aligned}$$

which follows from $\|g-Bw\|_{Y^{\delta\prime}} \le \|g-Bw\|_{Y'} \le \|{Q_B^\delta}'(g-Bw)\|_{Y'} + e^\delta_{\mathrm{osc}}(g)$.

Therefore, if $\sup_{\delta\in\Delta}\|Q_B^\delta\|_{\mathcal{L}(Y,Y)} < \infty$, then the a posteriori error estimator

$$\mathcal{E}^\delta(w;g,u_0,\beta) := \sqrt{\|{E_Y^\delta}'(g-Bw)\|^2_{K_Y^\delta} + \beta\|u_0-\gamma_0 w\|^2} \tag{4.6.4}$$

is an efficient and, modulo the *data oscillation term* $e^\delta_{\mathrm{osc}}(g)$, reliable estimator of the error $|||u-w|||_X$. If $\sup_{\delta\in\Delta}\|Q_B^\delta\|_{\mathcal{L}(Y,Y)}$ and $\frac{\max(R_\Delta,1)}{\min(r_\Delta,1)}$ are bounded uniformly in $\alpha\to\infty$, then this estimator is even robust.

Remark. In view of Prop. 4.5.1, the aforementioned assumptions $\operatorname{ran} Q_B^\delta \subset Y^\delta$, $(\mathrm{Id}-{Q_B^\delta}')BX^\delta=0$, and $\sup_{\delta\in\Delta}\|Q_B^\delta\|_{\mathcal{L}(Y,Y)}<\infty$ are equivalent to

$$\gamma_\Delta^B := \inf_{\delta\in\Delta}\inf_{\{0\neq w\in X^\delta:\, BE_X^\delta w\neq 0\}}\frac{\|{E_Y^\delta}'BE_X^\delta w\|_{Y^{\delta\prime}}}{\|BE_X^\delta w\|_{Y'}}>0.$$

In applications the conditions $\gamma_\Delta^{\partial_t}>0$, $\gamma_\Delta^C>0$, and $\gamma_\Delta^B>0$ are increasingly more difficult to fulfill. ♢

To have a meaningful reliability result, in addition we would like to find above Q_B^δ such that, for sufficiently smooth g, the term $e^\delta_{\mathrm{osc}}(g)$ is asymptotically, i.e. for the 'mesh-size' tending to zero, of equal or higher order than the approximation error $\inf_{w\in X^\delta}|||u-w|||_X$. We will realize this in the setting that will be discussed in Sect. 4.7.2.

4.7 Spatial PDOs with dominating asymmetric part

For some domain $\Omega\subset\mathbb{R}^d$, and $\Gamma\subset\partial\Omega$, let

$$\begin{aligned}&H:=L_2(\Omega),\ V:=H^1_{0,\Gamma}(\Omega):=\{v\in H^1(\Omega)\colon v|_\Gamma=0\},\\ &a(t;\eta,\zeta):=\int_\Omega \varepsilon\nabla\eta\cdot\nabla\zeta+(\boldsymbol{b}\cdot\nabla\eta+e\eta)\zeta\,d\boldsymbol{x},\quad \varepsilon>0,\\ &\boldsymbol{b}\in L_\infty(I;L_\infty(\mathrm{div};\Omega)),\ e\in L_\infty(I\times\Omega),\ \operatorname{ess\,inf}(e-\tfrac12\operatorname{div}_{\boldsymbol{x}}\boldsymbol{b})\geq 0,\end{aligned}\tag{4.7.1}$$

and $|\Gamma|>0$ when the latter ess inf is zero, so that (4.2.1) and (4.2.5) are valid. In this setting, the operators A_a, $A_s=A_s(\varepsilon)$, and so $A=A(\varepsilon)=A_s(\varepsilon)+A_a$, are given by

$$\begin{aligned}(A_aw)(v)&=\int_I\int_\Omega(\boldsymbol{b}\cdot\nabla_{\boldsymbol{x}}w+\tfrac12 w\operatorname{div}_{\boldsymbol{x}}\boldsymbol{b})v\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t,\\ (A_s(\varepsilon)w)(v)&=\int_I\int_\Omega\varepsilon\nabla_{\boldsymbol{x}}w\cdot\nabla_{\boldsymbol{x}}v+(e-\tfrac12\operatorname{div}_{\boldsymbol{x}}\boldsymbol{b})wv\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t.\end{aligned}$$

Thinking of $\boldsymbol{b}$ and e fixed, and variable $\varepsilon>0$, one infers that $\alpha=\alpha(\varepsilon)\to\infty$ when $\varepsilon\downarrow 0$ (cf. Remark 4.2.3).

In the next subsection we will construct $(X^\delta)_{\delta\in\Delta}\subset X$ and $(Y^\delta)_{\delta\in\Delta}\subset Y$ that (essentially) satisfy $\inf_{\varepsilon>0}\gamma_\Delta^C(\varepsilon)>0$ as families of finite element spaces w.r.t. subdivisions of $I\times\Omega$ into time-slabs with prismatic elements in each slab w.r.t. generally different partitions of Ω. Notice that although $C=\partial_t+A_a$ is independent of ε, $\gamma_\Delta^C(\varepsilon)$ depends on ε because it is defined in terms of the ε-dependent energy-norm $\|\cdot\|_Y=\sqrt{(A_s(\varepsilon)\cdot)(\cdot)}$.

As a consequence of $\gamma_\Delta^C(\varepsilon)$ being uniformly positive, for $K_Y^\delta\eqsim({E_Y^\delta}'A_sE_Y^\delta)^{-1}$ uniformly in ε and δ, i.e., $\sup_{\varepsilon>0}\frac{\max(R_\Delta,1)}{\min(r_\Delta,1)}<\infty$, Theorem 4.6.1 gives ε-robust quasi-optimality for MR and BEN w.r.t. the ε- and β-dependent $|||\cdot|||_X$-norm.

4.7.1 Realization of $\inf_\varepsilon \gamma_\Delta^{\mathsf{C}}(\varepsilon) > 0$

Given a conforming partition $\mathcal{T}$ of a polytopal $\overline{\Omega}$ into (essentially disjoint) closed d-simplices, we define $\mathcal{S}_{\mathcal{T}}^{-1,q}$ as the space of all (discontinuous) piecewise polynomials of degree q w.r.t. $\mathcal{T}$, and, for $q \geq 1$, set

$$\mathcal{S}_{\mathcal{T},0}^{0,q} := \mathcal{S}_{\mathcal{T}}^{-1,q} \cap H^1_{0,\Gamma}(\Omega),$$

where we assume that Γ is the union of faces of $T \in \mathcal{T}$.

Let $(\mathcal{T}^\delta)_{\delta\in\bar\Delta}$, $(\mathcal{T}_S^\delta)_{\delta\in\bar\Delta}$ be a families of such partitions of $\overline{\Omega}$, which are uniformly shape regular (which for $d = 1$ should be read as to satisfy a uniform K-mesh property), and where $\mathcal{T}_S^\delta$ is a refinement of $\mathcal{T}^\delta$ of some *fixed* maximal depth in the sense that $|T| \gtrsim |T'|$ for $\mathcal{T}_S^\delta \ni T \subset T' \in \mathcal{T}^\delta$, so that $\dim \mathcal{T}_S^\delta \lesssim \dim \mathcal{T}^\delta$. On the other hand, fixing a $q \geq 1$, we require that the refinement from $\mathcal{T}^\delta$ to $\mathcal{T}_S^\delta$ is sufficiently deep that it permits the construction of a projector P_q^δ for which

$$\operatorname{ran} P_q^\delta \subseteq \mathcal{S}_{\mathcal{T}_S^\delta,0}^{0,q}, \quad \operatorname{ran}(\mathrm{Id} - P_q^\delta) \perp_{L_2(\Omega)} \left(\mathcal{S}_{\mathcal{T}^\delta,0}^{0,q} + \mathcal{S}_{\mathcal{T}^\delta}^{-1,q-1}\right), \tag{4.7.2}$$

$$\|P_q^\delta w\|_{L_2(T)} \lesssim \|w\|_{L_2(T)} \quad (T \in \mathcal{T}^\delta,\, w \in L_2(\Omega)). \tag{4.7.3}$$

As shown in Lemma 7.5.1 and Remark 7.5.2, regardless of the refinement rule (e.g. red-refinement of newest vertex bisection) that is (recursively) applied to create $(\mathcal{T}_S^\delta)_{\delta\in\bar\Delta}$ from $(\mathcal{T}^\delta)_{\delta\in\bar\Delta}$, there is a refinement of some fixed depth that suffices to satisfy (4.7.3) as well as

$$\operatorname{ran} P_q^\delta \subseteq \{w \in \mathcal{S}_{\mathcal{T}_S^\delta,0}^{0,q} \colon w|_{\cup_{T\in\mathcal{T}}\partial T} = 0\}, \quad \operatorname{ran}(\mathrm{Id} - P_q^\delta) \perp_{L_2(\Omega)} \mathcal{S}_{\mathcal{T}^\delta,0}^{-1,q}. \tag{4.7.4}$$

Condition (4.7.4) is stronger than (4.7.2), and will be relevant in Sect. 4.7.2 on robust a posteriori error estimation.

For $d \in \{1,2,3\}$ and $q \in \{1,2,3\}$, and both newest vertex bisection and red-refinement it was verified that it is sufficient that the aforementioned depth creates in the space $\mathcal{S}_{\mathcal{T}_S^\delta,0}^{0,q}$ an additional number of degrees of freedom interior to any $T \in \mathcal{T}^\delta$ that is greater or equal to $\binom{q+d}{q}$.

Remark 4.7.1. To satisfy condition (4.7.2)–(4.7.3) generally a smaller number of degrees of freedom interior to any $T \in \mathcal{T}^\delta$ suffices. For $d = 2 = q$, in Appendix 7.A it was shown that in order to satisfy (4.7.2)–(4.7.3) it is sufficient to create $\mathcal{T}_s^\delta$ from $\mathcal{T}^\delta$ by one red-refinement, which creates only three of such degrees of freedom, whereas to satisfy (4.7.3)–(4.7.4) six additional interior degrees of freedom are needed. ◊

We show robustness of MR and BEN in a time-slab partition setting.

Theorem 4.7.2. *Let H, V, and $a(\cdot;\cdot,\cdot)$ be as in* (4.7.1), *with* constant $\boldsymbol{b}$ *and* constant $e \geq 0$. *Let $(\mathcal{T}^\delta)_{\delta\in\bar\Delta}$ and $(\mathcal{T}_S^\delta)_{\delta\in\bar\Delta}$ be as specified above. Then if, for $\delta \in \Delta$, X^δ*

and Y^δ satisfy, for some finite partition $I^\delta = ([t_{i-1}^\delta, t_i^\delta])_i$ of I, and arbitrary $\delta_i \in \bar{\Delta}$,

$$X^\delta \subseteq \{w \in C(I; H^1_{0,\Gamma}(\Omega)) \colon w|_{(t_{i-1}^\delta, t_i^\delta)} \in \mathcal{P}_q(t_{i-1}^\delta, t_i^\delta) \otimes \mathcal{S}^{0,q}_{\mathcal{T}^{\delta_i},0}\}, \tag{4.7.5}$$
$$Y^\delta \supseteq \{v \in L_2(I; H^1_{0,\Gamma}(\Omega)) \colon v|_{(t_{i-1}^\delta, t_i^\delta)} \in \mathcal{P}_q(t_{i-1}^\delta, t_i^\delta) \otimes \mathcal{S}^{0,q}_{\mathcal{T}_S^{\delta_i},0}\},$$

then $\inf_{\varepsilon>0} \gamma_\Delta^C(\varepsilon) > 0$. *Consequently the bounds* (4.6.2) *and* (4.6.3) *show quasi-optimality of MR and BEN w.r.t. the (ε- and β-dependent) norm $|||\cdot|||_X$, uniformly in $\varepsilon > 0$ and $\beta \geq 1$.*

Proof. As follows from Proposition 4.5.1 the statement $\inf_{\varepsilon>0} \gamma_\Delta^C(\varepsilon) > 0$ is equivalent to existence of $Q_C^\delta \in \mathcal{L}(Y, Y)$ with

$$\sup_{\varepsilon>0,\, \delta\in\Delta} \|Q_C^\delta\|_{\mathcal{L}(Y,Y)} < \infty,\ \operatorname{ran} Q_C^\delta \subset Y^\delta,\ \int_I \int_\Omega ((\partial_t + \boldsymbol{b}\cdot\nabla_{\boldsymbol{x}})X^\delta)(I - Q_C^\delta)Y \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t = 0, \tag{4.7.6}$$

using that, thanks to constant $\boldsymbol{b}$, $Y = L_2(I; H^1_{0,\Gamma}(\Omega))$ is equipped with norm

$$\begin{aligned}\sqrt{(A_s(\varepsilon)v)(v)} &= \sqrt{\int_I \varepsilon\|\nabla_x v\|^2_{L_2(\Omega)^d} + e\|v\|^2_{L_2(\Omega)}\,\mathrm{d}t} \\ &\eqsim \sqrt{\varepsilon}\|\nabla_x v\|_{L_2(I\times\Omega)^d} + \sqrt{e}\|v\|_{L_2(I;L_2(\Omega))}.\end{aligned}$$

It holds that

$$(\partial_t + \boldsymbol{b}\cdot\nabla_{\boldsymbol{x}})X^\delta \subseteq \Big\{v \in L_2(I\times\Omega) \colon v|_{(t_{i-1}^\delta, t_i^\delta)} \in \mathcal{P}_q(t_{i-1}^\delta, t_i^\delta) \otimes (\mathcal{S}^{0,q}_{\mathcal{T}^{\delta_i},0} + \mathcal{S}^{-1,q-1}_{\mathcal{T}^{\delta_i}})\Big\}. \tag{4.7.7}$$

Let $(Q_{\boldsymbol{x}}^\delta)_{\delta\in\bar{\Delta}}$ denote a family of projectors such that

$$\sup_{\delta\in\Delta} \max\Big(\|Q_{\boldsymbol{x}}^\delta\|_{\mathcal{L}(L_2(\Omega),L_2(\Omega))}, \|Q_{\boldsymbol{x}}^\delta\|_{\mathcal{L}(H^1_{0,\Gamma}(\Omega),H^1_{0,\Gamma}(\Omega))}\Big) < \infty, \tag{4.7.8}$$
$$\operatorname{ran} Q_{\boldsymbol{x}}^\delta \subset \mathcal{S}^{0,q}_{\mathcal{T}_S^\delta,0},\quad \operatorname{ran}(\mathrm{Id} - Q_{\boldsymbol{x}}^\delta) \perp_{L_2(\Omega)} \Big(\mathcal{S}^{0,q}_{\mathcal{T}^\delta,0} + \mathcal{S}^{-1,q-1}_{\mathcal{T}^\delta}\Big), \tag{4.7.9}$$

and let $Q^{\delta,i}$ be the $L_2(t_{i-1}^\delta, t_i^\delta)$-orthogonal projector onto $\mathcal{P}_q(t_{i-1}^\delta, t_i^\delta)$. Then, the operator Q_C^δ, defined by

$$(Q_C^\delta v)|_{(t_{i-1}^\delta, t_i^\delta)\times\Omega} = (Q^{\delta,i} \otimes Q_{\boldsymbol{x}}^{\delta_i})v|_{(t_{i-1}^\delta, t_i^\delta)\times\Omega},$$

satisfies (4.7.6). Indeed its uniform boundedness w.r.t. the energy-norm on Y follows by the boundedness of Q_x^δ w.r.t. both the $L_2(\Omega)$- and $H^1(\Omega)$-norms. By writing $\mathrm{Id} - Q^{\delta,i} \otimes Q_{\boldsymbol{x}}^{\delta_i} = (\mathrm{Id} - Q^{\delta,i}) \otimes \mathrm{Id} + Q^{\delta,i} \otimes (\mathrm{Id} - Q_{\boldsymbol{x}}^{\delta_i})$, and using (4.7.7) one verifies the third condition in (4.7.6).

We seek $Q_{\boldsymbol{x}}^\delta$ of the form $Q_{\boldsymbol{x}}^\delta = \check{Q}_{\boldsymbol{x}}^\delta + \hat{Q}_{\boldsymbol{x}}^\delta + \hat{Q}_{\boldsymbol{x}}^\delta \check{Q}_{\boldsymbol{x}}^\delta$ where

$$\operatorname{ran}\check{Q}_{\boldsymbol{x}}^\delta, \operatorname{ran}\hat{Q}_{\boldsymbol{x}}^\delta \subset \mathcal{S}^{0,q}_{\mathcal{T}_S^\delta,0},\quad \operatorname{ran}(\mathrm{Id} - \hat{Q}_{\boldsymbol{x}}^\delta) \perp_{L_2(\Omega)} (\mathcal{S}^{0,q}_{\mathcal{T}^\delta,0} + \mathcal{S}^{-1,q-1}_{\mathcal{T}^\delta}). \tag{4.7.10}$$

Then from $\mathrm{Id} - Q_{\boldsymbol{x}}^\delta = (\mathrm{Id} - \hat{Q}_{\boldsymbol{x}}^\delta)(\mathrm{Id} - \check{Q}_{\boldsymbol{x}}^\delta)$, we infer that (4.7.9) is satisfied.

We take $\hat{Q}_x^\delta = P_q^\delta$ from (4.7.2)–(4.7.3). It satisfies the properties required in (4.7.10). With $\hbar_\delta$ being the piecewise constant function defined by $\hbar_\delta|_T = \operatorname{diam} T$ ($T \in \mathcal{T}^\delta$), thanks to the uniform K-mesh property of $\mathcal{T} \in (\mathcal{T}^\delta)_{\delta\in\bar{\Delta}}$, (4.7.3) implies both $\|\hbar_\delta^{-1} P_q^\delta \hbar_\delta\|_{\mathcal{L}(L_2(\Omega),L_2(\Omega))} \lesssim 1$ and $\|P_q^\delta\|_{\mathcal{L}(L_2(\Omega),L_2(\Omega))} \lesssim 1$.

Take $\breve{Q}_x^\delta$ as a modified Scott–Zhang quasi-interpolator onto $\mathcal{S}^{0,q}_{\mathcal{T}_S^\delta,0}$ ([GL01, Appendix]). The modification consists in setting the degrees of freedom on Γ to zero. When applied to a function from $H^1_{0,\Gamma}(\Omega)$ it equals the original Scott–Zhang interpolator ([SZ90]), but thanks to the modification it is uniformly bounded w.r.t. $L_2(\Omega)$, and so $\|Q_x^\delta\|_{\mathcal{L}(L_2(\Omega),L_2(\Omega))}$ is uniformly bounded.

Writing $Q_x^\delta = \breve{Q}_x^\delta + P_q^\delta(\mathrm{Id} - \breve{Q}_x^\delta)$, from $\hbar_\delta^{-1}(\mathrm{Id} - \breve{Q}_x^\delta) \in \mathcal{L}(H^1_{0,\Gamma}(\Omega), L_2(\Omega))$, $\hbar_\delta^{-1} P_q^\delta \hbar_\delta \in \mathcal{L}(L_2(\Omega), L_2(\Omega))$, and $\breve{Q}_x^\delta \in \mathcal{L}(H^1_{0,\Gamma}(\Omega), H^1_{0,\Gamma}(\Omega))$ all being uniformly bounded, and $\|\cdot\|_{H^1(\Omega)} \lesssim \|\hbar_\delta^{-1}\cdot\|_{L_2(\Omega)}$ on $\mathcal{S}^{0,q}_{\mathcal{T}_S^\delta,0}$, we infer the uniform boundedness of $\|Q_x^\delta\|_{\mathcal{L}(H^1_{0,\Gamma}(\Omega),H^1_{0,\Gamma}(\Omega))}$. □

Next under the condition that $\operatorname{ess\,inf}(e - \frac{1}{2}\operatorname{div}_x \boldsymbol{b}) > 0$, we consider the case of *variable* $\boldsymbol{b}$ and e. The scaling argument that was applied directly below Theorem 4.2.1 shows that it is no real restriction to assume that $\operatorname{ess\,inf}(e - \frac{1}{2}\operatorname{div}_x \boldsymbol{b}) > 0$. Although we will not be able to show $\inf_{\varepsilon>0} \gamma_\Delta^C(\varepsilon) > 0$, this inf-sup condition will be valid modulo a perturbation which can be dealt with using Young's inequality similarly as in the proofs of Theorems 4.3.1 and 4.4.2. It will result in ε- (and β-) robust quasi-optimality results for MR and BEN similar as for constant $\boldsymbol{b}$ and constant $e \geq 0$.

Theorem 4.7.3. *Take the situation of Theorem 4.7.2, but now without assuming that $\boldsymbol{b}$ and e are constants. Assume $\boldsymbol{b} \in W^1_\infty(I\times\Omega)^d$, $\operatorname{ess\,inf}(e - \frac{1}{2}\operatorname{div}_x \boldsymbol{b}) > 0$, and, only for the case that $\boldsymbol{b}$ is time-dependent,*

$$|t_{i-1}^\delta - t_i^\delta| \lesssim \max_{T\in\mathcal{T}^{\delta_i}} \operatorname{diam}(T). \tag{4.7.11}$$

Then for MR and BEN it holds

$$\begin{aligned}
|\!|\!| u - u^\delta |\!|\!|_X &\lesssim \tfrac{\max(R_\Delta,1)}{\min(r_\Delta,1)} \inf_{w\in X^\delta} |\!|\!| u - w |\!|\!|_X,\\
|\!|\!| u - \bar{u}^\delta |\!|\!|_X &\lesssim \inf_{w\in X^\delta} |\!|\!| u - w |\!|\!|_X + \inf_{v\in Y^\delta} \|u - v\|_Y,
\end{aligned}$$

uniformly *in* $\varepsilon > 0$ *and* $\beta \geq 1$.

Proof. As in the proof of Theorem 4.6.1, we follow the proofs of Theorems 4.3.1 (MR) and 4.4.2 (BEN). We only need to adapt the derivation of a lower bound for the expression in the second line of (4.3.15).

With $\xi = \operatorname{ess\,inf}(e - \frac{1}{2}\operatorname{div}_x \boldsymbol{b})$, it holds that

$$\sqrt{\xi}\|\cdot\|_{Y'} \leq \|\cdot\|_{L_2(I\times\Omega)} \leq \tfrac{1}{\sqrt{\xi}}\|\cdot\|_Y.$$

Let $\boldsymbol{b}_\delta$ be the piecewise constant vector field defined by taking the average of $\boldsymbol{b}$ over each prismatic element $(t_{i-1}^\delta, t_i^\delta) \times T$ for $T \in \mathcal{T}^{\delta_i}$. We use $w \mapsto \boldsymbol{b}_\delta \cdot \nabla_{\boldsymbol{x}} w$ to approximate A_a; then $\|\boldsymbol{b} - \boldsymbol{b}_\delta\|_{L_\infty((t_{i-1}^\delta, t_i^\delta)\times T)^d} \lesssim \operatorname{diam}(T) \|\boldsymbol{b}\|_{W^1_\infty((t_{i-1}^\delta, t_i^\delta)\times T)^d}$ by (4.7.11). An application of the inverse inequality on the family of spaces $(\mathcal{S}^{0,q}_{\mathcal{T},0})_{\mathcal{T}\in\bar{\Delta}}$ shows that for some constant $L > 0$, for $w \in X^\delta$ it holds that

$$\|(\boldsymbol{b} - \boldsymbol{b}_\delta) \cdot \nabla_{\boldsymbol{x}} w + \tfrac{1}{2} w \operatorname{div}_{\boldsymbol{x}} \boldsymbol{b}\|_{L_2(I\times\Omega)} \leq L \|w\|_{L_2(I\times\Omega)}.$$

Because (4.7.7) is also valid for *piecewise* constant $\boldsymbol{b}$, and

$$\sqrt{(A_{\mathrm{s}}(\varepsilon)v)(v)} \eqsim \sqrt{\varepsilon} \|\nabla_x v\|_{L_2(I\times\Omega)^d} + \sqrt{\xi} \|v\|_{L_2(I;L_2(\Omega))},$$

only dependent on $\|e - \frac{1}{2} \operatorname{div}_{\boldsymbol{x}} \boldsymbol{b}\|_{L_\infty(I\times\Omega)} / \xi$, the proof of Theorem 4.7.2 shows that for some constant $\gamma > 0$, for $w \in X^\delta$ it holds that

$$\|{E_Y^\delta}'(\partial_t + \boldsymbol{b}_\delta \cdot \nabla_{\boldsymbol{x}}) E_Y^\delta w\|_{Y^{\delta\prime}} \geq \gamma \|(\partial_t + \boldsymbol{b}_\delta \cdot \nabla_{\boldsymbol{x}}) E_Y^\delta w\|_{Y'}.$$

By combining these estimates, we find that for $w \in X^\delta$ it holds that

$$\begin{aligned}
\|{E_Y^\delta}' C E_Y^\delta w\|_{Y^{\delta\prime}} &\geq \gamma \|(\partial_t + \boldsymbol{b}_\delta \cdot \nabla_{\boldsymbol{x}}) E_Y^\delta w\|_{Y'} - \tfrac{L}{\sqrt{\xi}} \|E_Y^\delta w\|_{L_2(I\times\Omega)} \\
&\geq \gamma \|C E_Y^\delta w\|_{Y'} - (\gamma + 1) \tfrac{L}{\sqrt{\xi}} \|E_Y^\delta w\|_{L_2(I\times\Omega)} \\
&\geq \gamma \|C E_Y^\delta w\|_{Y'} - (\gamma + 1) \tfrac{L}{\xi} \|E_Y^\delta w\|_Y,
\end{aligned}$$

and so

$$\begin{aligned}
&\|{E_Y^\delta}' C E_Y^\delta w\|^2_{Y^{\delta\prime}} + \|E_X^\delta w\|^2_Y + \|\gamma_T E_X^\delta w\|^2 + (\beta - 1)\|\gamma_0 E_X^\delta w\|^2 \\
&\geq \Big(\gamma \|C E_Y^\delta w\|_{Y'} - (\gamma + 1) \tfrac{L}{\xi} \|E_Y^\delta w\|_Y\Big)^2 + \|E_X^\delta w\|^2_Y \\
&\qquad\qquad + \|\gamma_T E_X^\delta w\|^2 + (\beta - 1)\|\gamma_0 E_X^\delta w\|^2 \\
&\geq (1 - \eta^2)\gamma^2 \|C E_Y^\delta w\|^2_{Y'} + \Big\{(1 - \eta^{-2})(\gamma + 1)^2 \tfrac{L^2}{\xi^2} + 1\Big\} \|E_X^\delta w\|^2_Y \\
&\qquad\qquad + \|\gamma_T E_X^\delta w\|^2 + (\beta - 1)\|\gamma_0 E_X^\delta w\|^2.
\end{aligned}$$

Minimizing over η shows that, with $\alpha^2 := (\gamma + 1)^2 \frac{L^2}{\xi^2}$, the last expression is greater than or equal to

$$\tfrac{1}{2}\left(\gamma^2 + \alpha^2 + 1 - \sqrt{(\gamma^2 + \alpha^2 + 1)^2 - 4\gamma^2}\right) |||E_X^\delta w|||^2_X,$$

which completes the proof. □

4.7.2 Robust a posteriori error estimation

A robust error estimator will be realized in the following limited setting.

Consider the spaces and bilinear form a as in (4.7.1), where $\boldsymbol{b}$ is constant, $e = 0$, and the polytope $\Omega \subset \mathbb{R}^d$ is convex. For families of quasi-uniform partitions $(I^\delta)_{\delta\in\Delta}$ of $\overline{I}$, and $(\mathcal{T}^\delta)_{\delta\in\Delta}$ and $(\mathcal{T}_S^\delta)_{\delta\in\Delta}$ of $\overline{\Omega}$ as before, where $\mathcal{T}_S^\delta$ is a sufficiently deep refinement of $\mathcal{T}^\delta$ that permits the construction of a projector P_1^δ that satisfies (4.7.3)–(4.7.4), and for some $h_\delta > 0$, $\operatorname{diam} T \eqsim h_\delta \eqsim \operatorname{diam} J$ ($T \in \mathcal{T}^\delta$, $J \in I^\delta$), let $X^\delta := S^{0,1}_{I^\delta} \otimes S^{0,1}_{\mathcal{T}^\delta,0}$ and $Y^\delta := S^{-1,1}_{I^\delta} \otimes S^{0,1}_{\mathcal{T}_S^\delta,0}$. For completeness, $S^{-1,1}_{I^\delta}$ denotes the space of piecewise linears w.r.t. I^δ, and $S^{0,1}_{I^\delta}$ the space of continuous piecewise linears w.r.t. I^δ.

In this setting, in Theorem 7.5.4 projectors $Q_B^\delta \in \mathcal{L}(Y,Y)$ have been constructed with $\operatorname{ran} Q_B^\delta \subset Y^\delta$ and $(I - {Q_B^\delta}')BX^\delta = 0$. Moreover, these Q_B^δ are uniformly bounded in $Y = L_2(I; H^1_{0,\Gamma}(\Omega))$ equipped with the standard Bochner norm, with $H^1_{0,\Gamma}(\Omega)$ being equipped with $\|\nabla \cdot\|_{L_2(\Omega)^d}$. Since for the current bilinear form a, the energy-norm $\|\cdot\|_Y$ is equal to $\sqrt{\varepsilon}\|\cdot\|_{L_2(I;H^1_{0,\Gamma}(\Omega))}$, it holds that $\sup_{\delta\in\Delta,\varepsilon>0}\|Q_B^\delta\|_{\mathcal{L}(Y,Y)} < \infty$, and so

$$\inf_{\varepsilon>0} \gamma_\Delta^B(\varepsilon) > 0.$$

Let $((\hat{K}_Y^\delta)^{-1}v)(v) \eqsim \int_I\int_\Omega |\nabla_x v|^2 \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t$ ($\delta \in \Delta$, $v \in Y^\delta$), then $(\varepsilon^{-1}\hat{K}_Y^\delta)^{-1} \eqsim {E_Y^\delta}' A_s E_Y^\delta$, i.e., using preconditioner $K_Y^\delta := \varepsilon^{-1}\hat{K}_Y^\delta$, $\sup_{\varepsilon>0}\frac{\max(R_\Delta,1)}{\min(r_\Delta,1)} < \infty$.

We show that *data-oscillation* is asymptotically of higher or equal order as the approximation error in $|||\cdot|||_X = \sqrt{\|B\cdot\|^2_{Y'} + \beta\|\gamma_0\cdot\|^2}$. Noting that $\|\cdot\|_{Y'} = \frac{1}{\sqrt{\varepsilon}}\|\cdot\|_{L_2(I;H^1_{0,\Gamma}(\Omega)')}$, it is natural to select $\beta = \varepsilon^{-1}$. Then $\sqrt{\varepsilon}|||\cdot|||_X$ equals

$$\sqrt{\|(\partial_t + \boldsymbol{b}\cdot\nabla_x)\cdot\|^2_{L_2(I;H^1_{0,\Gamma}(\Omega)')} + \varepsilon^2\|\cdot\|^2_{L_2(I;H^1_{0,\Gamma}(\Omega))} + \varepsilon\|\gamma_T\cdot\|^2 + (1-\varepsilon)\|\gamma_0\cdot\|^2},$$

and so even for a general smooth u, $\sqrt{\varepsilon}$ times the approximation error cannot be expected to be smaller than $\eqsim h_\delta^2$. Since for $g \in L_2(I; H^1(\Omega)) \cap H^2(I; H^{-1}(\Omega))$ it holds that $\sqrt{\varepsilon}\|(I - {Q_B^\delta}')g\|_{Y'} = \|(I - {Q_B^\delta}')g\|_{L_2(I;H^1_{0,\Gamma}(\Omega)')} \lesssim h_\delta^2$ (Theorem 7.5.4), we conclude that $\mathcal{E}^\delta(w; g, u_0, \beta)$ from (4.6.4) is an efficient and, modulo above satisfactory data-oscillation term, reliable a posteriori estimator of the error in w in $|||\cdot|||_X$-norm.

4.8 Numerical experiments

We tested the minimal residual (MR) method applied to the parabolic initial value problem with the singularly perturbed 'spatial component' as given in (4.7.1). We considered the simplest case where $I = \Omega = (0,1)$, $\boldsymbol{b} = 1$, and e is either 0 or 1, and $X^\delta = S^{0,1}_{I^\delta} \otimes S^{0,1}_{\mathcal{T}^\delta,0}$, where $I^\delta = \mathcal{T}^\delta$ is a uniform partition of I with mesh size h_δ. Taking always $(K_Y^\delta)^{-1} = {E_Y^\delta}' A_s E_Y^\delta$, we took either

(i) $Y^\delta = S^{-1,1}_{I^\delta} \otimes S^{0,1}_{\mathcal{T}^\delta,0}(\supseteq X^\delta \cup \partial_t X^\delta)$ which for any *fixed* $\varepsilon > 0$ gives $\gamma^{\partial_t}_\Delta > 0$ (Sect. 4.5.1), so that the MR approximations are quasi-optimal approximations from the trial space w.r.t. $\|\cdot\|_X$ (Thm. 4.3.1), or

(ii) $Y^\delta := S^{-1,1}_{I^\delta} \otimes S^{0,1}_{\mathcal{T}^\delta_s,0}$ where $\mathcal{T}^\delta_s$ is a uniform partition with mesh-size $h_\delta/3$ which even gives $\inf_{\varepsilon>0} \gamma^C_\Delta(\varepsilon) > 0$ (Thm. 4.7.2), so that the MR approximations are quasi-optimal approximations from the trial space w.r.t. the energy-norm $|||\cdot|||_X$ also *uniformly in* $\varepsilon > 0$ (Thm. 4.6.1).

With Remark 4.4.1, in these cases the BEN and MR methods are equivalent.

As discussed in Sect. 4.7.2, for the case that $e = 0$ it is natural to take the weight $\beta = \varepsilon^{-1}$. Unlike with $e = 0$, for $e = 1$ and $0 \neq v \in Y$ the energy-norm $\sqrt{(A_s v)(v)}$ does not tend to zero for $\varepsilon \downarrow 0$ but converges to $\|v\|_{L_2(I\times\Omega)}$, so there is no reason to let β tend to infinity for $\varepsilon \downarrow 0$, and we took $\beta = 1$.

For Y^δ as in (ii), in Sect. 4.7.2 it was shown that for $(e, \beta) = (0, \varepsilon^{-1})$ it holds that $\inf_{\varepsilon>0} \gamma^B_\Delta(\varepsilon) > 0$, and more specifically that the a posteriori error estimator $\mathcal{E}^\delta(w; g, u_0, \beta)$ from (4.6.4) is an efficient and, modulo a data-oscillation term which is at least of equal order, reliable estimator of the error $|||u - w|||_X$. Therefore to assess our numerical results, we used Y^δ as in Option (ii) for error estimation, even when solving with Y^δ as in (i).

For $(e, \beta) = (1, 1)$, we numerically *observed* that for our model problems the a posteriori error estimator $\mathcal{E}^\delta(w; g, u_0, \beta)$ computed with Y^δ as in (ii) is efficient and reliable as, knowing that the estimator *equals* $|||u - w|||_X$ for $Y^\delta = Y$, we saw that further overrefinement of the test space Y^δ never increased the estimated error by more than a percent. So again, regardless of whether we took Y^δ as in Option (i) or (ii), we used Y^δ as in (ii) to compute $\mathcal{E}^\delta(w; g, u_0, \beta)$.

In experiments below, we choose $\varepsilon = 1, 10^{-1}, 10^{-3}, 10^{-6}$; to compare different values of ε, we show the estimated error divided by an accurate approximation for $\sqrt{\|g\|^2_{Y'} + \beta\|u_0\|^2} = |||u|||_X$.

4.8.1 Smooth problem

We take (homogeneous) Dirichlet boundary conditions at left- and right boundary, i.e. $\Gamma = \partial\Omega$, select $(e, \beta) = (0, \varepsilon^{-1})$, and prescribe the solution $u(t, x) := (t^2 + 1)\sin(\pi x)$ with derived data u_0 and g. For this problem, the best possible error in $|||\cdot|||_X$-norm, divided by $|||u|||_X$, decays proportionally to $(\dim X^\delta)^{-1/2}$.

Figure 4.1 shows this relative estimated error as a function of $\dim X^\delta$. In accordance with Theorem 4.3.1, for this parabolic problem with non-symmetric spatial part, both Option (i) and Option (ii) give solutions that converge at the expected rate. For Option (i), however, this convergence is not uniform in ε, but in accordance with Theorem 4.6.1, for Option (ii) it is.

4.8.2 Internal layer problem

We choose $u_0 := 0$ and $g(t, x) := \mathbb{1}_{\{x>t\}}$, select $(e, \beta) = (0, \varepsilon^{-1})$, and prescribe a homogeneous Dirichlet boundary condition only at the left boundary $x = 0$, i.e. $\Gamma := \{0\}$, and so have a Neumann boundary condition at the 'outflow'

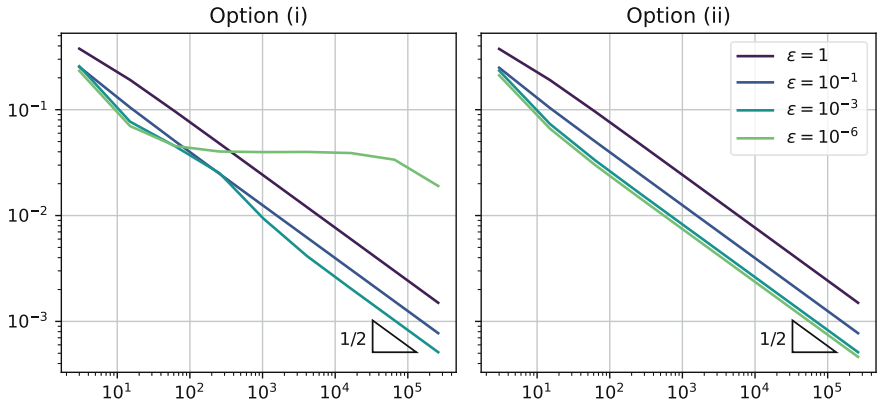


Figure 4.1 Relative estimated error progression for the *smooth problem* as function of $\dim X^\delta$ for different diffusion rates ε. Left: test space Y^δ as in Option (i); right: Y^δ as in (ii).

boundary $x = 1$. Due to the jump in the forcing data, in the limit $\varepsilon \downarrow 0$, the solution $t \cdot \mathbb{1}_{\{x>t\}}$ is discontinuous along the diagonal $x = t$.

The left of Figure 4.2 shows the relative estimated error progression of Option (ii) as a function of $\dim X^\delta$; as Option (i) again suffers from degradation for small ε (with results very similar to the left of Figure 4.1), we omit a graph of its error progression. Its right shows the discrete solution at $h_\delta = \frac{1}{512}$ and $\varepsilon = 10^{-6}$. The solution resembles the *pure transport* solution quite well, with the exception of a small artefact near $x = t = 0$.

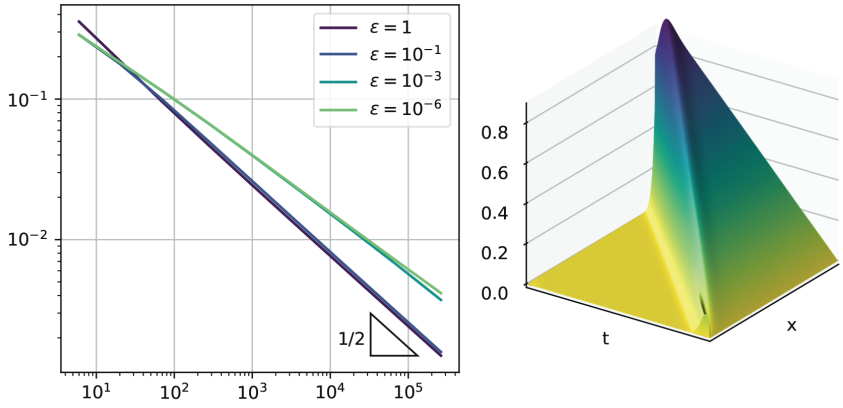


Figure 4.2 Solving the *internal layer problem* with Option (ii). Left: relative estimated error progression as function of $\dim X^\delta$ for different diffusion rates ε. Right: solution at $h_\delta = \frac{1}{512}$ and $\varepsilon = 10^{-6}$.

4.8.3 Boundary layer problem

We choose $u_0(x) := \sin(\pi x)$ and $g = 0$, select $(e, \beta) = (1, 1)$, and set homogenous Dirichlet boundary conditions on $\partial\Omega$, i.e. $\Gamma = \{0, 1\}$. Due to the condition on the outflow boundary, the problem is ill-posed in the limit $\varepsilon = 0$, hence for ε small, the solution has a boundary layer at $x = 1$.

Figure 4.3 shows that the method fails to make progress until the boundary layer is resolved at $h_\delta \lesssim \varepsilon$. Figure 4.4 shows two discrete solutions at $h_\delta = \frac{1}{512}$ computed for Option (ii). We see that for $\varepsilon = 10^{-3}$, the boundary layer is resolved and the solution resembles the *pure transport* solution quite well, with the exception of a small artefact near $x = t = 1$. For $\varepsilon = 10^{-6}$ though, the boundary layer cannot be resolved with the current (uniform) mesh, and the solution is completely wrong. For $\varepsilon \downarrow 0$, the energy-norm of the error in an approximation w approaches $\sqrt{\|(\partial_t + \boldsymbol{b} \cdot \nabla_x) w\|^2_{L_2(I\times\Omega)} + \|u_0 - \gamma_0 w\|^2_{L_2(\Omega)}}$. As a result, for streamlines that hit the outflow boundary, the method 'chooses' to smear the unavoidably large error as a consequence of the layer along the whole streamline resulting in a globally bad approximation. This is a well-known phenomenon when using a least squares method to approximate a solution that has a sharp layer or a shock.

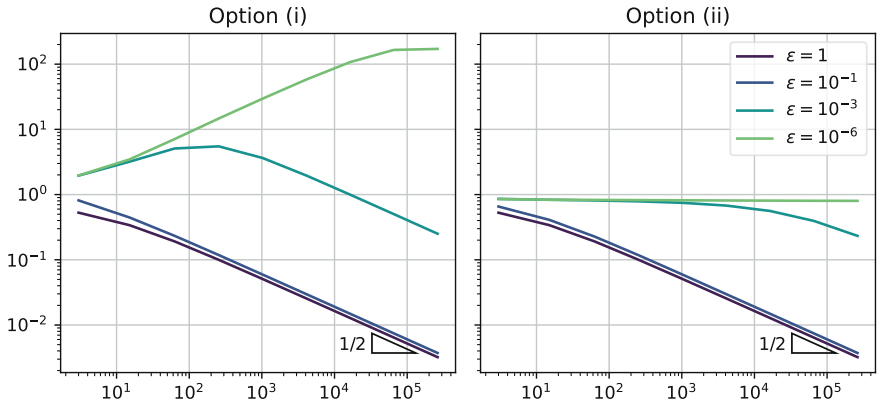


Figure 4.3 Relative estimated error progression for the *boundary layer problem* as function of $\dim X^\delta$ for different diffusion rates ε. Left: test space Y^δ as in Option (i); right: Y^δ as in (ii).

4.8.4 Imposing outflow boundary conditions weakly

One common work-around is to resolve the problem caused by the boundary layer is to refine the mesh strongly towards this layer. An alternative is to impose at the outflow boundary the Dirichlet boundary condition only weakly, see e.g. the references [CDW12, BS14, CEQ14, CFLQ14] where this approach has been applied with least squares methods for stationary convection dominated convection-diffusion methods. Without having a rigorous analysis we

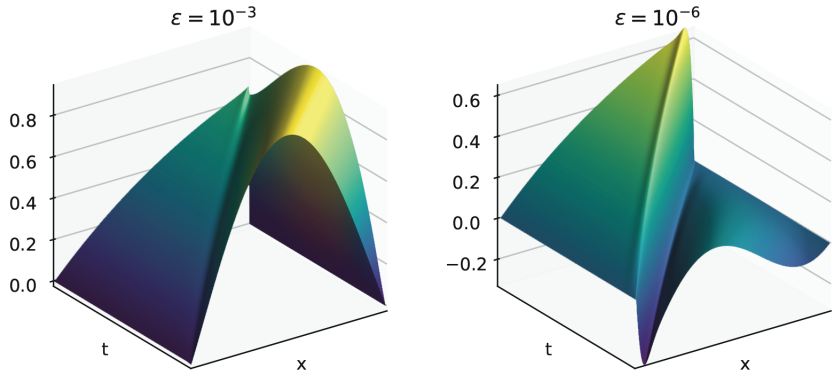


Figure 4.4 Solutions of the *boundary layer problem* with Option (ii) at $h_\delta = \frac{1}{512}$. Left: diffusion $\varepsilon = 10^{-3}$; right: $\varepsilon = 10^{-6}$.

tried this second approach by computing, with Y^δ as in Option (ii),

$$u^\delta := \operatorname*{arg\,min}_{w \in \hat{X}^\delta} \|E_Y^{\delta\prime}(BE_X^\delta w - g)\|^2_{Y^{\delta\prime}} + \beta\|\gamma_0 E_X^\delta w - u_0\|^2 + \varepsilon\|w(\cdot,1)\|^2_{L_2(I)}.$$

Here, $\hat{X}^\delta$ denotes the space X^δ after removing the Dirichlet boundary condition at $x = 1$. Figure 4.5 shows the resulting error progression, which is robust in ε, as well as the minimal residual solution at $h_\delta = \frac{1}{512}$ and $\varepsilon = 10^{-6}$; it resembles the *pure transport* solution quite well, and does not suffer from the artifact present at the right of Figure 4.4.

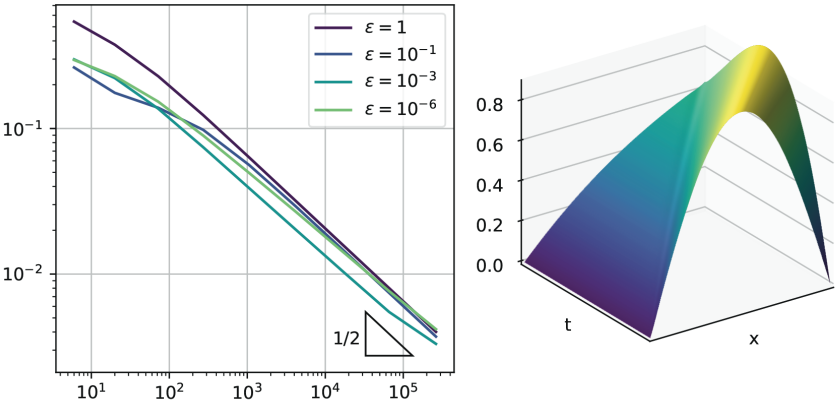
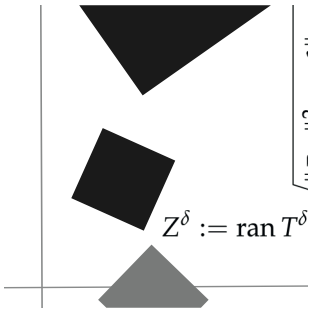
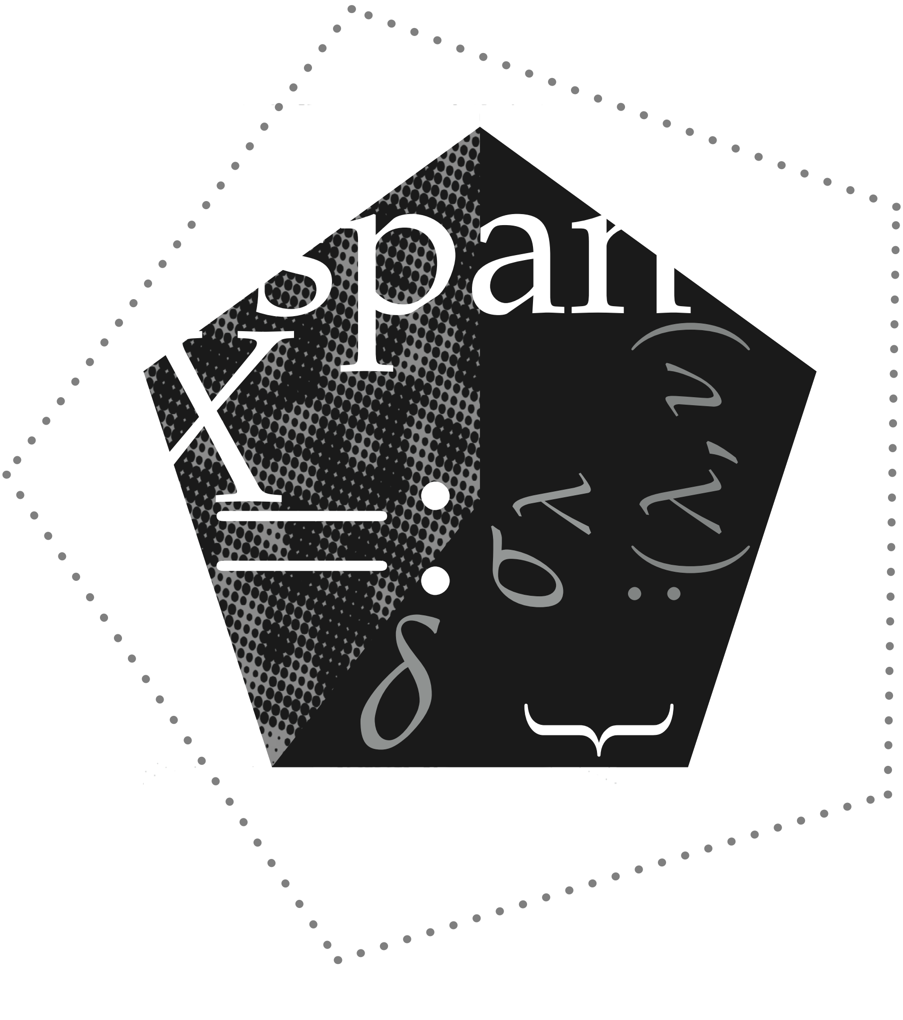


Figure 4.5 Solving the *boundary layer problem* with Option (ii) imposing the outflow boundary condition *weakly*. Left: error progression in $\dim \hat{X}^\delta$ for different ε. Right: solution at $h_\delta = \frac{1}{512}$ and $\varepsilon = 10^{-6}$.


$Z^\delta := \operatorname{ran} T^\delta$


Xspan

5 Efficient space-time adaptivity for parabolic PDEs

Abstract Considering the space-time adaptive method for parabolic evolution equations introduced in [arXiv:2101.03956 [math.NA]], this work discusses an implementation of the method in which every step is of linear complexity. Exploiting the product structure of the space-time cylinder, the method allows for a family of trial spaces given as the spans of wavelets-in-time tensorized with (locally refined) finite element spaces-in-space. On spaces whose bases are indexed by *double-trees*, we derive an algorithm that applies the resulting bilinear forms in linear complexity. We provide extensive numerical experiments to demonstrate the linear runtime of the resulting adaptive loop.

Source code is available at [vVW21b].

5.1 Introduction

This chapter deals with the adaptive numerical solution of parabolic evolution equations using a simultaneous space-time variational formulation. Compared to the more classical time-stepping schemes, these space-time methods are very flexible. Among other things, they are especially well-suited for massively parallel computation ([NS19, vVW20a]), and some can guarantee quasi-best approximations from the trial space ([And13, FK21, SZ20]).

We are interested in those space-time methods that permit adaptive refinement locally in space *and* time. Within this class, wavelet-based methods (see [SS09, GK11, KSU15]) are attractive, as they can be shown to be *quasi-optimal*: they produce a sequence of solutions that converges at the best possible rate, at optimal linear computational cost. Moreover, they can overcome the *curse of dimensionality* using a form of *sparse tensor-product approximation*, solving the whole time evolution at a runtime proportional to that of solving the corresponding *stationary* problem.

In [SvVW21], we constructed an r-linearly converging space-time adaptive solver for parabolic evolution equations exploiting the product structure of

This chapter is a minor modification of **Efficient space-time adaptivity for parabolic evolution equations using wavelets in time and finite elements in space**, R. van Venetië and J. Westerdiep, submitted to *Numerical Linear Algebra & Applications*, arXiv:2104.08143.

the space-time cylinder for a family of trial spaces as the spans of wavelets-in-time tensorized with (locally refined) finite element spaces-in-space.

The principal difference between this and other wavelet-based methods is that we use wavelets in time *only*, and standard finite elements in space. This eases implementation, and alleviates the need for a suitable spatial wavelet basis, a difficulty for general domains ([RS18a]). Unfortunately, there is no free lunch: a proof of optimal convergence is, for our method, not yet available.

In this work we discuss an implementation of [SvVW21] in which the different steps (each iteration of the linear algebraic solver, the error estimation, Dörfler marking, and refinement of trial- and test spaces) of the adaptive algorithm are of linear complexity.

Special care has to be taken for matrix-vector products. For a bilinear form that is 'local' and equals (a sum of) tensor-product(s) of bilinear forms in time and space, and 'trial' and 'test' spaces spanned by tensor-product multi-level bases with *double-tree* index sets, the resulting system matrix w.r.t. both bases can be applied in linear complexity, even though this matrix is not sparse. The algorithm that realizes this complexity makes a clever use of multi- to single-scale transformations alternately in time and space. This *unidirectional principle* was introduced in [BZ96] for 'uniform' sparse grids, so without 'local refinements', and it was later extended to general *downward closed* or *lower sets*, also called *adaptive sparse grids*, in [KS14]. The definition of a lower set in [KS14], there called multi-tree, is more restrictive than our current definition that allows more localized refinements.

To the best of our knowledge, other implementations for the efficient evaluation of tensor-product bilinear forms (see [Pfl10, KS14, Pab15, Rek18]) are based on the concept of hash maps. There, a hash function is used to map basis functions to array indices. In an adaptive loop, the final set of basis functions is unknown in advance so it is impossible to construct a hash function that guarantees an upper bound on the number of hash collisions. Aiming at true linear complexity, we implement these operations by traversing *trees* and *double-trees*, so without the use of hash maps.

Organization

In §5.2, we look at the abstract parabolic problem, its stable discretization, and the adaptive routine. In §5.3, we provide an abstract algorithm for the efficient evaluation of tensor-product bilinear forms w.r.t. multilevel bases indexed on *double-trees*. In §5.4, we take the *heat equation* as a model problem, and provide a concrete family of trial- and test spaces with bases indexed by double-trees that permits local space-time adaptivity. In §5.5, we discuss the practical implementation of the adaptive algorithm. Finally, in §5.6, we provide extensive numerical experiments to demonstrate linear runtime.

Notation

In this work, by $C \lesssim D$ we will mean that C can be bounded by a multiple of D, independently of parameters which C and D may depend on. Obviously,

$C \gtrsim D$ is defined as $D \lesssim C$, and $C \eqsim D$ as $C \lesssim D$ and $C \gtrsim D$.

For normed linear spaces E and F, by $\mathcal{L}(E,F)$ we will denote the normed linear space of bounded linear mappings $E \to F$, and by $\mathcal{L}\mathrm{is}(E,F)$ its subset of boundedly invertible linear mappings $E \to F$. We write $E \hookrightarrow F$ to denote that E is continuously embedded into F. For simplicity only, we exclusively consider linear spaces over the scalar field $\mathbb{R}$.

5.2 Space-time adaptivity for a parabolic model problem

In this section, we summarize the relevant parts of [SvVW21, §2–5].

Let V, H be separable Hilbert spaces of functions on some "spatial domain" such that $V \hookrightarrow H$ with dense and compact embedding. Identifying H with its dual, we obtain the Gelfand triple $V \hookrightarrow H \simeq H' \hookrightarrow V'$.

For a.e.

$$t \in I := (0,T),$$

let $a(t;\cdot,\cdot)$ denote a bilinear form on $V \times V$ so that for any $\eta, \zeta \in V$, $t \mapsto a(t;\eta,\zeta)$ is measurable on I, and such that for a.e. $t \in I$,

$$\begin{aligned} |a(t;\eta,\zeta)| &\lesssim \|\eta\|_V \|\zeta\|_V \quad (\eta,\zeta \in V) \quad (\textit{boundedness}),\\ a(t;\eta,\eta) &\gtrsim \|\eta\|_V^2 \qquad (\eta \in V) \qquad (\textit{coercivity}). \end{aligned}$$

With $(A(t)\cdot)(\cdot) := a(t;\cdot,\cdot) \in \mathcal{L}\mathrm{is}(V,V')$, given a forcing function g and initial value u_0, we want to solve the *parabolic initial value problem* of

$$\text{finding } u : I \to V \text{ such that } \quad \begin{cases} \frac{\mathrm{d}u}{\mathrm{d}t}(t) + A(t)u(t) = g(t) & (t \in I),\\ u(0) = u_0. \end{cases} \tag{5.2.1}$$

Example 5.2.1. For the *heat equation*, on some spatial domain $\Omega \subset \mathbb{R}^d$ we select $V := H_0^1(\Omega)$, $H := L_2(\Omega)$, and $a(t;\eta,\zeta) := \int_\Omega \nabla_{\boldsymbol{x}}\eta \cdot \nabla_{\boldsymbol{x}}\zeta \,\mathrm{d}\boldsymbol{x}$. ♢

In our simultaneous space-time variational formulation, the parabolic problem is to find u that solves

$$(Bu)(v) := \int_I \langle \tfrac{\mathrm{d}u}{\mathrm{d}t}(t), v(t)\rangle_H + a(t;u(t),v(t))\,\mathrm{d}t = \int_I \langle g(t), v(t)\rangle_H =: g(v)$$

for all v from some suitable space of functions of time and space. We can enforce the initial condition by testing against additional test functions.

Theorem ([SS09]). *With $X := L_2(I;V) \cap H^1(I;V')$, $Y := L_2(I;V)$, we have*

$$\begin{bmatrix} B \\ \gamma_0 \end{bmatrix} \in \mathcal{L}\mathrm{is}(X, Y' \times H),$$

where for $t \in \bar{I}$, $\gamma_t \colon u \mapsto u(t,\cdot)$ denotes the trace map. In other words,

$$\textit{finding } u \in X \textit{ s.t. } \quad (Bu, \gamma_0 u) = (g, u_0) \quad \textit{given} \quad (g,u_0) \in Y' \times H \tag{5.2.2}$$

is a well-posed simultaneous space-time variational formulation of (5.2.1).

We define $A \in \mathcal{L}\mathrm{is}(Y, Y')$ and $\partial_t \in \mathcal{L}\mathrm{is}(X, Y')$ as

$$(Au)(v) := \int_I a(t; u(t), v(t))\,\mathrm{d}t, \quad \text{and} \quad \partial_t := B - A.$$

Following [SvVW21], we assume that A is *self-adjoint*. Morever, in view of an efficient implementation, we assume that A is a finite sum of tensor-product operators. If A does not have this structure, one may alternatively consider (low-rank) tensor-product approximations of A, see e.g. [Hac12] for an overview.

We equip Y and X with *'energy'-norms*

$$\|\cdot\|_Y^2 := (A\cdot)(\cdot), \quad \|\cdot\|_X^2 := \|\partial_t \cdot\|_{Y'}^2 + \|\cdot\|_Y^2 + \|\gamma_T \cdot\|_H^2,$$

which are equivalent to the canonical norms on Y and X.

The solution u of (5.2.2) equals the solution of the minimization problem

$$u = \operatorname*{arg\,min}_{w \in X} \|Bw - g\|_{Y'}^2 + \|\gamma_0 w - u_0\|_H^2, \tag{5.2.3}$$

which in turn is the second component of the solution of

$$\text{finding } (\mu, u) \in Y \times X \text{ s.t.} \quad \begin{bmatrix} A & B \\ B' & -\gamma_0'\gamma_0 \end{bmatrix} \begin{bmatrix} \mu \\ u \end{bmatrix} = \begin{bmatrix} g \\ -u_0 \end{bmatrix}. \tag{5.2.4}$$

Indeed, taking the Schur complement of (5.2.4) w.r.t. the Y-block results in the Euler–Lagrange equations of (5.2.3).

5.2.1 Discretizations

Take a family $(X^\delta)_{\delta \in \Delta}$ of closed subspaces of X, and define

$$u_\delta = \operatorname*{arg\,min}_{w \in X^\delta} \|Bw - g\|_{Y'}^2 + \|\gamma_0 w - u_0\|_H^2, \tag{5.2.5}$$

being the best approximation to u from X^δ w.r.t. $\|\cdot\|_X$. Solving this problem, however, is not feasible because of the presence of the dual norm. Therefore, take $(Y^\delta)_{\delta \in \Delta}$ to be a family of closed subspaces of Y such that

$$X^\delta \subseteq Y^\delta \quad (\delta \in \Delta), \quad \text{and} \quad \gamma_\Delta := \inf_{\delta \in \Delta} \inf_{0 \neq w \in X^\delta} \sup_{0 \neq v \in Y^\delta} \frac{(\partial_t w)(v)}{\|\partial_t w\|_{Y'} \|v\|_Y} > 0. \tag{5.2.6}$$

For $\underline{\delta} \in \Delta$ with $Y^{\underline{\delta}} \supseteq Y^\delta$, we replace Y' by ${Y^{\underline{\delta}}}'$ in (5.2.5) yielding

$$u^{\underline{\delta}\delta} = \operatorname*{arg\,min}_{w \in X^\delta} \|Bw - g\|_{{Y^{\underline{\delta}}}'}^2 + \|\gamma_0 w - u_0\|_H^2.$$

Notice that $u^{\underline{\delta}\delta}$ approximates u_δ in that $u^{\underline{\delta}\delta} = u_\delta$ when $Y^{\underline{\delta}} = Y$.

With $E_Y^{\underline{\delta}} : Y^{\underline{\delta}} \to Y$ and $E_X^\delta : X^\delta \to X$ denoting the trivial embeddings, $u^{\underline{\delta}\delta}$ is the second component of the solution of

$$\begin{bmatrix} {E_Y^{\underline{\delta}}}' A E_Y^{\underline{\delta}} & {E_Y^{\underline{\delta}}}' B E_X^\delta \\ {E_X^\delta}' B' E_Y^{\underline{\delta}} & -{E_X^\delta}' \gamma_0' \gamma_0 E_X^\delta \end{bmatrix} \begin{bmatrix} \mu^{\underline{\delta}\delta} \\ u^{\underline{\delta}\delta} \end{bmatrix} = \begin{bmatrix} {E_Y^{\underline{\delta}}}' g \\ -{E_X^\delta}' \gamma_0' u_0 \end{bmatrix}.$$

Taking the Schur complement w.r.t. the $Y^{\underline{\delta}}$-block then leads to the equation

$$\begin{aligned} {E_X^\delta}'(B'E_Y^\delta({E_Y^\delta}' A E_Y^\delta)^{-1}{E_Y^\delta}' B + \gamma_0'\gamma_0)E_X^\delta u^{\underline{\delta}\delta} \\ = {E_X^\delta}'(B'E_Y^\delta({E_Y^\delta}' A E_Y^\delta)^{-1}{E_Y^\delta}' g + \gamma_0' u_0), \end{aligned} \tag{5.2.7}$$

which has a unique solution (cf. [SvVW21, Lem. 3.3]) that satisfies $\|u - u^{\underline{\delta}\delta}\|_X \leq \gamma_\Delta^{-1}\|u - u_\delta\|_X$ whenever $Y^{\underline{\delta}} \supseteq Y^\delta$; cf. Theorem 3.3.7. For now, we assume the right-hand side of (5.2.7) to be evaluated exactly. Later, in §5.4.5, we will discuss approximation of the right-hand side.

In view of obtaining an efficient solver, we want to replace the inverses in (5.2.7) while aiming to preserve quasi-optimality of the solution. To this end, let $K_Y^\delta = {K_Y^\delta}' \in \mathcal{L}\mathrm{is}(Y^{\underline{\delta}'}, Y^{\underline{\delta}})$ be a uniformly optimal preconditioner for ${E_Y^\delta}' A E_Y^\delta$ that can be applied in linear complexity. Then, for some $\kappa_\Delta \geq 1$,

$$\frac{((K_Y^\delta)^{-1}v)(v)}{(Av)(v)} \in [\kappa_\Delta^{-1}, \kappa_\Delta] \quad (\delta \in \Delta, v \in Y^{\underline{\delta}}).$$

Replacing $({E_Y^\delta}' A E_Y^\delta)^{-1}$ by K_Y^δ, we denote the solution of (5.2.7) again by $u^{\underline{\delta}\delta}$. It is quasi-optimal with $\|u - u^{\underline{\delta}\delta}\|_X \leq \frac{\kappa_\Delta}{\gamma_\Delta}\|u - u_\delta\|_X$; cf. Remark 3.3.8.

5.2.2 Adaptive refinement loop

Our adaptive loop, given in Algorithm 5.1, takes the familiar Solve, Estimate, Mark and refine steps, and is driven by an efficient and reliable 'hierarchical basis' a posteriori error estimator.

The adaptive loop below requires a saturation assumption. Define a *partial order* on Δ by $\tilde{\delta} \succeq \delta$ whenever $X^{\tilde{\delta}} \supseteq X^\delta$. Let $\delta \mapsto \underline{\delta} \succeq \delta$ be a mapping providing *saturation* in that for some $\zeta < 1$,

$$\|u - u_{\underline{\delta}}\|_X \leq \zeta\|u - u_\delta\|_X \quad (\delta \in \Delta). \tag{5.2.8}$$

With this choice of $\underline{\delta}$, we are interested in finding $u^\delta := u^{\underline{\delta}\delta} \in X^\delta$ that solves

$$\underbrace{{E_X^\delta}'(B'E_Y^{\underline{\delta}}K_Y^{\underline{\delta}}{E_Y^{\underline{\delta}}}' B + \gamma_0'\gamma_0)E_X^\delta}_{S^{\underline{\delta}\delta}:=} u^\delta = \underbrace{{E_X^\delta}'(B'E_Y^{\underline{\delta}}K_Y^{\underline{\delta}}{E_Y^{\underline{\delta}}}' g + \gamma_0' u_0)}_{f^\delta:=}. \tag{5.2.9}$$

Notice that (5.2.9) is uniquely solvable even with $X^{\underline{\delta}}$ as 'trial space', and we use this 'room' between X^δ and $X^{\underline{\delta}}$ to our advantage. Expanding X^δ to some intermediate space $X^\delta \subset X^{\tilde{\delta}} \subset X^{\underline{\delta}}$ yields a $u^{\tilde{\delta}}$ that is a better approximation to u than u^δ; cf. [SvVW21, Prop. 4.2]. This function will be the successor of u^δ in our loop, and we will show that the resulting sequence of functions converges r-linearly to u; see Algorithm 5.1 and Theorem 5.2.3.

Solving Instead of solving the symmetric positive definite system (5.2.9) exactly, we construct an approximate solution $\hat{u}^\delta$ using Preconditioned Conjugate Gradients (PCG). To this end, let $K_X^\delta = K_X^{\delta\prime} \in \mathcal{L}\mathrm{is}(X^{\delta\prime}, X^\delta)$ be a uniformly optimal preconditioner for $S^{\underline{\delta}\underline{\delta}}$. Then $((K_X^\delta)^{-1}w)(w) \eqsim \|w\|_X^2 \eqsim \|K_X^\delta S^{\underline{\delta}\underline{\delta}} w\|_X^2$ for $w \in X^\delta$. Writing $w = K_X^\delta S^{\underline{\delta}\underline{\delta}}(u^\delta - v^\delta)$ leads to an algebraic error estimator

$$\beta^\delta(v^\delta) := \sqrt{(f^\delta - S^{\underline{\delta}\underline{\delta}} v^\delta)(K_X^\delta(f^\delta - S^{\underline{\delta}\underline{\delta}} v^\delta))} \eqsim \|u^\delta - v^\delta\|_X \quad (v^\delta \in X^\delta, \delta \in \Delta). \tag{5.2.10}$$

With $\hat{u}_k^\delta$ denoting the approximant at iteration k of the PCG loop, $\beta^\delta(\hat{u}_k^\delta)$ is already available as the variable β_k used in computing the next search direction.

Error estimation Let $\Theta_\delta := \{\theta_\lambda\}_{\lambda\in J_\delta}$ with $X^\delta \oplus \mathrm{span}\Theta_\delta = X^{\underline{\delta}}$ be uniformly X-stable, i.e.,

$$\|z + \boldsymbol{c}^\top \Theta_\delta\|_X^2 \eqsim \|z\|_X^2 + \|\boldsymbol{c}\|^2 \quad (\boldsymbol{c} \in \ell_2(J_\delta), z \in X^\delta, \delta \in \Delta). \tag{5.2.11}$$

Define the trivial embedding $P^\delta : X^\delta \to X^{\underline{\delta}}$. Akin to (5.2.9), we set $S^{\underline{\delta}\underline{\delta}}$, $f^{\underline{\delta}\underline{\delta}}$, and then the residual-based a posteriori error estimator $\boldsymbol{r}^\delta : X^\delta \to \ell_2(J_\delta)$, as

$$\begin{aligned} &S^{\underline{\delta}\underline{\delta}} := E_X^{\underline{\delta}\prime}(B' E_Y^{\underline{\delta}} K_Y^\delta E_Y^{\delta\prime} B + \gamma_0'\gamma_0) E_X^{\underline{\delta}}, \quad f^{\underline{\delta}\underline{\delta}} := E_X^{\underline{\delta}\prime}(B' E_Y^{\underline{\delta}} K_Y^\delta E_Y^{\delta\prime} g + \gamma_0' u_0), \\ &\boldsymbol{r}^\delta(\hat{u}^\delta) := (f^{\underline{\delta}\underline{\delta}} - S^{\underline{\delta}\underline{\delta}} P^\delta \hat{u}^\delta)(\Theta_\delta). \end{aligned} \tag{5.2.12}$$

For $\hat{u}^\delta$ close to u^δ, the error estimator $\|\boldsymbol{r}^\delta(\hat{u}^\delta)\|$ is reliable and efficient:

Lemma 5.2.2. *Assume* (5.2.8) *and* (5.2.11), $\frac{\kappa_\Delta}{\gamma_\Delta} < \frac{1}{\xi}$, *and fix some* $\xi > 0$ *small enough. For* $\hat{u}^\delta \in X^\delta$ *satisfying* $\beta(\hat{u}^\delta) \le \frac{\xi}{1-\xi}\|\boldsymbol{r}^\delta(\hat{u}^\delta)\|$, *we have*

$$\|\boldsymbol{r}^\delta(\hat{u}^\delta)\| \eqsim \|u - \hat{u}^\delta\|_X \quad \textit{and} \quad \|u - \hat{u}^\delta\|_X \lesssim \|u - u^\delta\|_X \quad (\delta \in \Delta).$$

Proof. For convenience, we write $\hat{\boldsymbol{r}}^\delta := \boldsymbol{r}^\delta(\hat{u}^\delta)$ and $\boldsymbol{r}^\delta := \boldsymbol{r}^\delta(u^\delta)$.

By (5.2.8), (5.2.11) and $\frac{\kappa_\Delta}{\gamma_\Delta} < \frac{1}{\xi}$, [SvVW21, Prop. 4.4] shows that

$$\|\boldsymbol{r}^\delta\| \eqsim \|u - u^\delta\|_X \quad (\delta \in \Delta). \tag{5.2.13}$$

From (5.2.11) one deduces that $\|\boldsymbol{r}^\delta - \hat{\boldsymbol{r}}^\delta\| \lesssim \|u^\delta - \hat{u}^\delta\|_X$; cf. [SvVW21, (4.13)]. By assumption, for $\xi < 1$, we find $\beta^\delta(\hat{u}^\delta) \lesssim \xi\|\hat{\boldsymbol{r}}^\delta\|$, revealing that

$$\|\boldsymbol{r}^\delta - \hat{\boldsymbol{r}}^\delta\| \overset{(5.2.11)}{\lesssim} \|u^\delta - \hat{u}^\delta\|_X \overset{(5.2.10)}{\eqsim} \beta^\delta(\hat{u}^\delta) \lesssim \xi\|\hat{\boldsymbol{r}}^\delta\|. \tag{5.2.14}$$

Using this, we can show reliability of the estimator by

$$\begin{aligned} \|u - \hat{u}^\delta\|_X &\le \|u - u^\delta\|_X + \|u^\delta - \hat{u}^\delta\|_X \\ &\overset{(5.2.13),(5.2.10)}{\eqsim} \|\boldsymbol{r}^\delta\| + \beta^\delta(\hat{u}^\delta) \le \|\hat{\boldsymbol{r}}^\delta\| + \|\boldsymbol{r}^\delta - \hat{\boldsymbol{r}}^\delta\| + \beta^\delta(\hat{u}^\delta) \overset{(5.2.14)}{\lesssim} \|\hat{\boldsymbol{r}}^\delta\|. \end{aligned}$$

For efficiency of the estimator, we deduce

$$\|\hat{\boldsymbol{r}}^\delta\| \overset{(5.2.13)}{\lesssim} \|u-u^\delta\|_X + \|\boldsymbol{r}^\delta - \hat{\boldsymbol{r}}^\delta\| \leq \|u-\hat{u}^\delta\|_X + \|u^\delta - \hat{u}^\delta\|_X + \|\boldsymbol{r}^\delta - \hat{\boldsymbol{r}}^\delta\|$$
$$\overset{(5.2.14)}{\lesssim} \|u-\hat{u}^\delta\|_X + \xi\|\hat{\boldsymbol{r}}^\delta\|,$$

so taking ξ sufficiently small and kicking back $\|\hat{\boldsymbol{r}}^\delta\|$ yields

$$\|\hat{\boldsymbol{r}}^\delta\| \lesssim \|u-\hat{u}^\delta\|_X. \tag{5.2.15}$$

Similarly, from (5.2.13) and (5.2.14) it follows that

$$\|\hat{\boldsymbol{r}}^\delta\| \lesssim \|u-u^\delta\|_X. \tag{5.2.16}$$

We infer quasi-optimality of $\hat{u}^\delta$ from

$$\|u-\hat{u}^\delta\|_X \overset{(5.2.14)}{\lesssim} \|u-u^\delta\|_X + \xi\|\hat{\boldsymbol{r}}^\delta\| \overset{(5.2.16)}{\lesssim} \|u-u^\delta\|_X. \qquad \square$$

In the solve step, we iterate PCG until $\beta^\delta(\hat{u}_k^\delta)/\|\boldsymbol{r}^\delta(\hat{u}_k^\delta)\|$ is small enough. In the algorithm below, this is ensured by the do-while loop which also avoids the (expensive) recomputation of the residual at every PCG iteration.

Marking and refinement Denoting the output of the solve step by $\hat{u}^\delta$, we drive the adaptive loop by performing Dörfler marking on the residual $\hat{\boldsymbol{r}}^\delta := \boldsymbol{r}^\delta(\hat{u}^\delta)$, i.e., for some $\theta \in (0,1]$, we mark the smallest set $J \subset J_\delta$ for which $\|\hat{\boldsymbol{r}}^\delta|_J\| \geq \theta\|\hat{\boldsymbol{r}}^\delta\|$. We then construct the smallest $\tilde{\delta} \succeq \delta$ such that $X^{\tilde{\delta}}$ contains $\operatorname{span}\Theta_\delta|_J$.

Data: $\theta \in (0,1]$, $\xi \in (0,1)$, $\delta := \delta_{\text{init}} \in \Delta$;
$t_\delta := \mathcal{E}^\delta(0) = \sqrt{(E_Y^{\delta\prime} g)(K_Y^\delta E_Y^{\delta\prime} g) + \|u_0\|_H^2}$;
repeat
  Solve:
    do
      Compute $\hat{u}_*^\delta \in X^\delta$ with $\beta^\delta(\hat{u}_*^\delta) \leq t_\delta/2$;
      $t_\delta := \beta^\delta(\hat{u}_*^\delta)$;
      $e_\delta := \|\boldsymbol{r}^\delta(\hat{u}_*^\delta)\| + t_\delta$;
    while $t_\delta > \xi e_\delta$;
    $\hat{u}^\delta := \hat{u}_*^\delta$;
  Estimate: Set $\hat{\boldsymbol{r}}^\delta := \boldsymbol{r}^\delta(\hat{u}^\delta)$;
  Mark: Mark a smallest $J \subset J_\delta$ for which $\|\hat{\boldsymbol{r}}^\delta|_J\| \geq \theta\|\hat{\boldsymbol{r}}^\delta\|$;
  Refine: Determine smallest $\tilde{\delta} \in \Delta$ such that $X^{\tilde{\delta}} \supset X^\delta \oplus \operatorname{span}\Theta_\delta|_J$;
  $t_{\tilde{\delta}} := e_\delta$, $\delta := \tilde{\delta}$;

Algorithm 5.1 Space-time adaptive refinement loop.

Theorem 5.2.3 ([SvVW21, Thm. 4.9 with $\eta = 0$])**.** *Assume* (5.2.8), (5.2.11). *For ξ and $\frac{\kappa_\Delta}{\gamma_\Delta} - 1$ sufficiently small with $\frac{\kappa_\Delta}{\gamma_\Delta} - 1 \downarrow 0$ when $\theta \downarrow 0$, the sequence of approximations produced by Algorithm 5.1 converges r-linearly to u, in that after every iteration, $\|u - \hat{u}^\delta\|_X$ decreases with a factor at least $\rho < 1$.*

Remark 5.2.4. In practice, to ensure termination, Algorithm 5.1 has to be complemented by an appropriate stopping criterium; cf. [SvVW21, Alg. 4.8]. ♢

Proof. For convenience, we denote $\boldsymbol{r}^\delta := \boldsymbol{r}^\delta(u^\delta)$ and $\hat{\boldsymbol{r}}^\delta := \boldsymbol{r}^\delta(\hat{u}^\delta)$. The stopping criterium of the solve step ensures that $\beta^\delta(\hat{u}^\delta) \leq \xi\Big(\|\hat{\boldsymbol{r}}^\delta\| + \beta^\delta(\hat{u}^\delta)\Big)$, so for $\xi < 1$ we are in the situation of Lemma 5.2.2.

We have

$$\|\hat{\boldsymbol{r}}^\delta - \boldsymbol{r}^\delta\| \overset{(5.2.14)}{\lesssim} \xi\|\hat{\boldsymbol{r}}^\delta\| \leq \xi\Big(\|\boldsymbol{r}^\delta\| + \|\hat{\boldsymbol{r}}^\delta - \boldsymbol{r}^\delta\|\Big),$$

so taking ξ sufficiently small and kicking back $\|\hat{\boldsymbol{r}}^\delta - \boldsymbol{r}^\delta\|$ yields

$$\|\hat{\boldsymbol{r}}^\delta - \boldsymbol{r}^\delta\| \lesssim \xi\|\boldsymbol{r}^\delta\|. \tag{5.2.17}$$

After marking, we have $\|\hat{\boldsymbol{r}}^\delta\| \leq \theta^{-1}\|\hat{\boldsymbol{r}}^\delta|_J\|$, which shows that

$$\|\boldsymbol{r}^\delta\| \overset{(5.2.14)}{\lesssim} \|\hat{\boldsymbol{r}}^\delta\| \lesssim \|\hat{\boldsymbol{r}}^\delta|_J\| \leq \|\boldsymbol{r}^\delta|_J\| + \|\boldsymbol{r}^\delta - \hat{\boldsymbol{r}}^\delta\| \overset{(5.2.17)}{\lesssim} \|\boldsymbol{r}^\delta|_J\| + \xi\|\boldsymbol{r}^\delta\|;$$

for ξ small enough, kicking back $\|\boldsymbol{r}^\delta\|$ we find for a $\hat{\theta} > 0$ dependent on θ,

$$\|\boldsymbol{r}^\delta|_J\| \geq \hat{\theta}\|\boldsymbol{r}^\delta\|.$$

From [SvVW21, Prop. 4.3] we now find that, for $\frac{\kappa_\Delta}{\gamma_\Delta} - 1 \downarrow 0$ when $\theta \downarrow 0$, there is a $\bar{\rho} < 1$ for which

$$\|u - u^{\tilde{\delta}}\|_X \leq \bar{\rho}\|u - u^\delta\|_X. \tag{5.2.18}$$

Combining the results shows that

$$\begin{aligned}
\|u - \hat{u}^{\tilde{\delta}}\|_X &\leq \|u - u^{\tilde{\delta}}\|_X + \|u^{\tilde{\delta}} - \hat{u}^{\tilde{\delta}}\|_X \\
&\overset{(5.2.14),(5.2.16)}{\leq} (1 + \mathcal{O}(\xi))\|u - u^{\tilde{\delta}}\|_X \\
&\overset{(5.2.18)}{\leq} (1 + \mathcal{O}(\xi))\bar{\rho}\|u - u^\delta\|_X \\
&\leq (1 + \mathcal{O}(\xi))\bar{\rho}(\|u - \hat{u}^\delta\|_X + \|u^\delta - \hat{u}^\delta\|_X) \\
&\overset{(5.2.14),(5.2.15)}{\leq} \underbrace{(1 + \mathcal{O}(\xi))\bar{\rho}}_{=:\rho}\|u - \hat{u}^\delta\|_X,
\end{aligned}$$

so for ξ small enough, $\rho < 1$, completing the proof of r-linear convergence. □

5.2.3 Adaptive trial- and test spaces

The convergence rate of our adaptive loop is determined by the approximation properties of the family $(X^\delta)_{\delta\in\Delta}$. We want to construct a family that allows for *local* refinements. Here, the crucial problem is guaranteeing the inf-sup stability condition (5.2.6). It is known that inf-sup stability is satisfied for *full* tensor-products of (non-uniform) finite element spaces, and in [And13, Prop. 4.2], this result was generalized to families of *sparse* tensor-products. Unfortunately, neither family allows for adaptive refinements both locally in time and space.

In §5.4 we will solve this by first equipping X with a tensor-product of (infinite) bases: a wavelet basis Σ in time, and a hierarchical basis in space. We then construct X^δ as the span of a (finite) subset of this tensor-product basis, which we grow by adding particular functions.

By imposing a *double-tree* constraint on the index set of the basis of X^δ, we can apply tensor-product operators in linear complexity; see §5.3. Moreover, this constraint implies that for our model problem the inf-sup condition (5.2.6) is satisfied and we can construct optimal preconditioners K_Y^δ and K_X^δ.

5.3 Applying linear operators in linear complexity

An efficient implementation of our adaptive loop requires the efficient application of the operators ${E_Y^\delta}' B E_X^\delta$ and ${E_X^\delta}' \gamma_0' \gamma_0 E_X^\delta$ appearing in (5.2.9). Both terms are finite sums of tensor-products of operators in time and space. When we equip our trial and test spaces with tensor-products of multilevel bases, it turns out that we can evaluate these operators in linear complexity.

More precisely, this section will show the abstract result that given

- tensor-products $\Psi := \Psi^0 \times \Psi^1$, $\breve{\Psi} := \breve{\Psi}^0 \times \breve{\Psi}^1$ of multilevel bases Ψ^0, Ψ^1, $\breve{\Psi}^0$, $\breve{\Psi}^1$ indexed by $\vee^0$, $\vee^1$, $\breve{\vee}^0$, $\breve{\vee}^1$, and
- (finite) subsets $\boldsymbol{\Lambda} \subset \vee^0 \times \vee^1$, $\breve{\boldsymbol{\Lambda}} \subset \breve{\vee}^0 \times \breve{\vee}^1$ that are *double-trees*, and
- linear operators $A_i : \operatorname{span}\Psi^i \to (\operatorname{span}\breve{\Psi}^i)'$ that are *local* ($i \in \{0,1\}$),

we can apply the matrix $((A_0 \otimes A_1)\Psi|_{\boldsymbol{\Lambda}})(\breve{\Psi}|_{\breve{\boldsymbol{\Lambda}}})$ in $\mathcal{O}(\#\boldsymbol{\Lambda} + \#\breve{\boldsymbol{\Lambda}})$ operations even though this matrix is not uniformly sparse.

Example. For our model problem, Ψ^0 and $\breve{\Psi}^0$ will be wavelets for $H^1(I)$ or $L_2(I)$ in time, and $\Psi^1 = \breve{\Psi}^1$ will be a hierarchical finite element basis for $H_0^1(\Omega)$ in space. We will apply the result of this section to the operators $\gamma_0'\gamma_0$ and $B = \partial_t + A$. ◊

We will achieve this complexity using a variant of the *unidirectional principle*. Denote with $I_{\boldsymbol{\Lambda}}$ the extension with zeros of a vector supported on $\boldsymbol{\Lambda}$ to one on $\vee^0 \times \vee^1$, and with $R_{\boldsymbol{\Lambda}}$ its adjoint; define $I_{\breve{\boldsymbol{\Lambda}}}$ and $R_{\breve{\boldsymbol{\Lambda}}}$ analogously. Define $\boldsymbol{A}_i := (A_i\Psi^i)(\breve{\Psi}^i)$. We will split $\boldsymbol{A}_0$ in its upper and strictly lower triangular parts $\boldsymbol{U}_0$ and $\boldsymbol{L}_0$, so that

$$R_{\breve{\boldsymbol{\Lambda}}}(\boldsymbol{A}_0 \otimes \boldsymbol{A}_1) I_{\boldsymbol{\Lambda}} = R_{\breve{\boldsymbol{\Lambda}}}(\boldsymbol{L}_0 \otimes \mathrm{Id})(\mathrm{Id} \otimes \boldsymbol{A}_1) I_{\boldsymbol{\Lambda}} + R_{\breve{\boldsymbol{\Lambda}}}(\boldsymbol{U}_0 \otimes \mathrm{Id})(\mathrm{Id} \otimes \boldsymbol{A}_1) I_{\boldsymbol{\Lambda}}.$$

This in itself is useless, as $(\mathrm{Id}\otimes \boldsymbol{A}_1)I_{\boldsymbol{\Lambda}}$ maps into a space which dimension we cannot control. However, the restriction $R_{\check{\boldsymbol{\Lambda}}}$ gives us elbow room: in Theorem 5.3.7 we construct double-trees $\boldsymbol{\Sigma}, \boldsymbol{\Theta}$ with $\#\boldsymbol{\Sigma} + \#\boldsymbol{\Theta} \lesssim \#\check{\boldsymbol{\Lambda}} + \#\boldsymbol{\Lambda}$ s.t.

$$\begin{cases} R_{\check{\boldsymbol{\Lambda}}}(\boldsymbol{L}_0\otimes \mathrm{Id})(\mathrm{Id}\otimes \boldsymbol{A}_1)I_{\boldsymbol{\Lambda}} = R_{\check{\boldsymbol{\Lambda}}}(\boldsymbol{L}_0\otimes \mathrm{Id})R_{\boldsymbol{\Sigma}}I_{\boldsymbol{\Sigma}}(\mathrm{Id}\otimes \boldsymbol{A}_1)I_{\boldsymbol{\Lambda}}, \\ R_{\check{\boldsymbol{\Lambda}}}(\boldsymbol{U}_0\otimes \mathrm{Id})(\mathrm{Id}\otimes \boldsymbol{A}_1)I_{\boldsymbol{\Lambda}} = R_{\check{\boldsymbol{\Lambda}}}(\boldsymbol{U}_0\otimes \mathrm{Id})R_{\boldsymbol{\Theta}}I_{\boldsymbol{\Theta}}(\mathrm{Id}\otimes \boldsymbol{A}_1)I_{\boldsymbol{\Lambda}}. \end{cases} \tag{5.3.1}$$

These right hand sides we *can* apply efficiently, and their application boils down to applications of $\boldsymbol{L}_0$, $\boldsymbol{U}_0$, and $\boldsymbol{A}_1$ in a *single* coordinate direction only. Simple matrix-vector products are inefficient though, as these matrices are again not uniformly sparse. However, by using the properties of a double-tree and the sparsity of the operator in *single scale*, we can evaluate $\boldsymbol{U}_0$, $\boldsymbol{L}_0$ and $\boldsymbol{A}_1$ in linear time; see §5.3.1.

We follow the structure of [KS14, §3], which applies the aforementioned idea to *multi-trees* though with a slightly more restrictive definition of a *tree*. For readability, we defer the proofs of Theorems 5.3.3–5.3.7 to Appendix 5.A.

5.3.1 Applying linear operators on trees

Let Ψ be a (multilevel) collection of functions on some domain Q.

Example 5.3.1. In our application, Q will be either the time interval I with Ψ being a collection of wavelets, or the spatial domain Ω, in which case Ψ is a collection of hierarchical basis functions. ♢

Writing $\Psi = \{\psi_\lambda : \lambda \in \vee\}$, we assume that the ψ_λ are *locally supported* in the sense that with $|\lambda| \in \mathbb{N}_0$ denoting the *level* of λ,

$$\sup_{\lambda\in\vee} 2^{|\lambda|}\operatorname{diam}\operatorname{supp}\psi_\lambda < \infty, \tag{5.3.2}$$

$$\sup_{\ell\in\mathbb{N}_0}\sup_{x\in Q} \#\{\lambda\in\vee : |\lambda| = \ell \wedge \operatorname{supp}\psi_\lambda \cap B(x;2^{-\ell}) \neq \varnothing\} < \infty. \tag{5.3.3}$$

We will refer to the functions ψ_λ as being *wavelets*, although not necessarily they have vanishing moments or other specific wavelet properties.

For $\ell \in \mathbb{N}_0$, and any $\Lambda \subset \vee$, we set $\Lambda_\ell := \{\lambda \in \Lambda : |\lambda| = \ell\}$ and $\Lambda_{\ell\uparrow} := \{\lambda\in\Lambda : |\lambda| \geq \ell\}$, and write $\Psi_\ell := \Psi|_{\vee_\ell}$.

For $\ell \in \mathbb{N}_0$, we assume a collection $\Phi_\ell = \{\phi_\lambda : \lambda \in \Delta_\ell\}$, whose members will be referred to as being *scaling functions*, with

$$\operatorname{span}\Phi_{\ell+1} \supseteq \operatorname{span}\Phi_\ell \cup \Psi_{\ell+1}, \quad \Phi_0 = \Psi_0 \quad (\Delta_0 := \vee_0), \tag{5.3.4}$$

$$\sup_{\ell\in\mathbb{N}_0}\sup_{\lambda\in\Delta_\ell} 2^{\ell}\operatorname{diam}\operatorname{supp}\phi_\lambda < \infty, \tag{5.3.5}$$

$$\sup_{\ell\in\mathbb{N}_0}\sup_{x\in Q} \#\{\lambda\in\Delta_\ell : \operatorname{supp}\phi_\lambda \cap B(x;2^{-\ell}) \neq \varnothing\} < \infty, \tag{5.3.6}$$

$$\{\phi_\lambda|_\Sigma : \lambda\in\Delta_\ell,\ \phi_\lambda|_\Sigma \not\equiv 0\} \text{ is independent } (\Sigma\subset Q \text{ open}, \ell\in\mathbb{N}_0). \tag{5.3.7}$$

W.l.o.g. we assume that the index sets Δ_ℓ for different ℓ are mutually disjoint, and set $\Phi := \cup_{\ell\in\mathbb{N}_0}\Phi_\ell$ with index set $\Delta := \cup_{\ell\in\mathbb{N}_0}\Delta_\ell$. For $\lambda \in \Delta$, we set $|\lambda| := \ell$ when $\lambda \in \Delta_\ell$.

Viewing Ψ_ℓ, Φ_ℓ as column vectors, the assumptions we made so far guarantee the existence of matrices $\mathfrak{p}_\ell$, $\mathfrak{q}_\ell$ such that

$$\begin{bmatrix}(\Phi_{\ell-1})^\top & (\Psi_\ell)^\top\end{bmatrix} = (\Phi_\ell)^\top \begin{bmatrix}\mathfrak{p}_\ell & \mathfrak{q}_\ell\end{bmatrix},$$

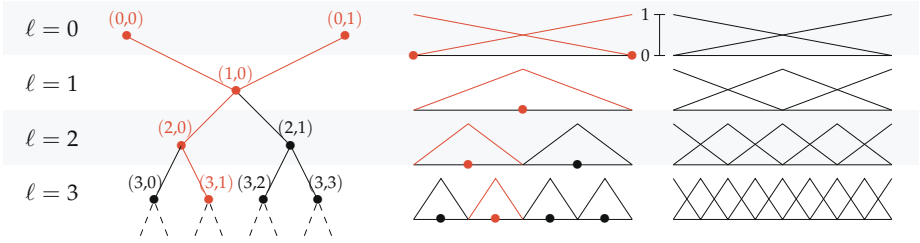
where the number of non-zeros per row and column of $\mathfrak{p}_\ell$ and $\mathfrak{q}_\ell$ is finite, uniformly in the rows and columns and in $\ell \in \mathbb{N}$ (here also (5.3.7) has been used). We refer to $\mathfrak{p}_\ell$ as the *prolongation matrix*. Columns of $\mathfrak{p}_\ell$ contain the *mask* of the scaling functions, and those of $\mathfrak{q}_\ell$ contain the mask of the wavelets.

To each $\lambda \in \vee$ with $|\lambda| > 0$, we associate one or more $\mu \in \vee$ with $|\mu| = |\lambda| - 1$ and $|\operatorname{supp}\psi_\lambda \cap \operatorname{supp}\psi_\mu| > 0$. We call μ a *parent* of λ, and so λ a *child* of μ. To each $\lambda \in \vee$, we associate some neighbourhood $S(\lambda)$ of $\operatorname{supp}\psi_\lambda$, with diameter $\lesssim 2^{-|\lambda|}$, such that for $|\lambda| > 0$, $S(\lambda) \subset \cup_{\mu\in\mathtt{parent}(\lambda)} S(\mu)$.

Remark. Such a neighborhood always exists even when a child has only one parent. With $C := \sup_{\lambda\in\vee} 2^{|\lambda|}\operatorname{diam}\operatorname{supp}\psi_\lambda$ and $S(\lambda) := \{x \in Q : \operatorname{dist}(x, \operatorname{supp}\psi_\lambda) < C2^{-|\lambda|}\}$, for μ a parent of λ and $x \in S(\lambda)$, $\operatorname{dist}(x, \operatorname{supp}\psi_\mu) \le \operatorname{dist}(x, \operatorname{supp}\psi_\lambda) + \operatorname{diam}\operatorname{supp}\psi_\lambda < 2C2^{-|\lambda|} = C2^{-|\mu|}$, i.e., $x \in S(\mu)$. ◊

Definition 5.3.2 (Tree)**.** A finite $\Lambda \subset \vee_{\ell\uparrow}$ is called an *ℓ-tree*, or simply a *tree* when $\ell = 0$, when for any $\lambda \in \Lambda$ its parents in $\vee_{\ell\uparrow}$ are in Λ. This is not a tree in the graph-theoretical sense, but rather in the family history sense. ◊

Example (Hierarchical basis in 1D)**.** Figure 5.1 shows an example multilevel collection Ψ of functions defined on the interval $[0,1]$. Its index set $\vee_\mathfrak{I}$ with parent-child relations is shown left, with a tree $\Lambda \subset \vee_\mathfrak{I}$ visualised in red. This collection is called the *hierarchical basis*. With $S(\lambda) := \operatorname{supp}\psi_\lambda$ for $\lambda \in \vee_\mathfrak{I}$, the hierarchical basis satisfies conditions mentioned above. ◊



Index set $\vee_\mathfrak{I}$ and tree $\Lambda \subset \vee_\mathfrak{I}$ Multilevel functions Ψ Scaling functions Φ

Figure 5.1 Hierarchical basis for the interval $[0,1]$.

A routine `eval` Let (Ψ, Φ) and $(\breve{\Psi}, \breve{\Phi})$ satisfy the conditions of the previous subsection, and let $A\colon \operatorname{span}\Phi \to (\operatorname{span}\breve{\Phi})'$ be *local* in that $(Au)(v) = (Au|_{\operatorname{supp} v})(v)$. Typically, A is a (partial) differential operator in variational form; e.g. $A \in \mathcal{L}(H^1(I), L_2(I)')$ with $(Au)(v) = \int_I \frac{\mathrm{d}u}{\mathrm{d}t} v \,\mathrm{d}t$. For trees $\Lambda \subset \vee$ and $\breve{\Lambda} \in \breve{\vee}$, we want to apply the matrix $(A\Psi|_\Lambda)(\breve{\Psi}|_{\breve{\Lambda}})$ efficiently.

Just for brevity of the following argument, assume $\Psi = \breve{\Psi}$ and $\Phi = \breve{\Phi}$. The matrix $(A\Psi|_\Lambda)(\Psi|_\Lambda)$ is not uniformly sparse, so a straight-forward matrix-vector product is not of linear complexity. However, for Λ a uniform tree up to level ℓ, i.e. $\Lambda = \{\lambda \in \vee : |\lambda| \le \ell\}$, a solution is provided by the multi- to single-scale transform T characterized by $\Psi|_\Lambda = T^\top \Phi_\ell$ through the equality $(A\Psi|_\Lambda)(\Psi|_\Lambda) = T^\top (A\Phi_\ell)(\Phi_\ell) T$, as the transforms can be applied in linear complexity and the single-scale matrix is uniformly sparse.

For general trees however, we don't have $\dim \Phi_\ell \lesssim \dim \Psi|_\Lambda$ so the previous approach is not of linear complexity. Clever level-by-level multi-to-singlescale transformations and the prolongation of *only* relevant functions *does* allow applying $(A\Psi|_\Lambda)(\breve{\Psi}|_{\breve{\Lambda}})$ at linear cost; see Algorithm 5.2 below.

On several places the restriction of a vector (of scalars or of functions) to its indices in some subset of the index set should be read as the vector of full length where the entries with indices outside this subset are replaced by zeros. For index sets Δ and $\breve{\Delta}$, matrix $\mathtt{m} \in \mathbb{R}^{\#\breve{\Delta} \times \#\Delta}$, and subset $\Pi \subset \Delta$, we write $\operatorname{supp}(\mathtt{m}, \Pi) \subset \breve{\Delta}$ for the index set corresponding to the image of $\mathtt{m}$ under $\{\boldsymbol{x}|_\Pi : \boldsymbol{x} \in \mathbb{R}^{\#\Delta}\}$.

Data: $\ell \in \mathbb{N}$, $\breve{\Pi} \subset \breve{\Delta}_{\ell-1}$, $\Pi \subset \Delta_{\ell-1}$, ℓ-trees $\breve{\Lambda} \subset \breve{\vee}_{\ell\uparrow}$ and $\Lambda \subset \vee_{\ell\uparrow}$,
$\boldsymbol{d} \in \mathbb{R}^{\#\Pi}$, $\boldsymbol{c} \in \mathbb{R}^{\#\Lambda}$.
Result: $[\boldsymbol{e}, \boldsymbol{f}]$ where $\boldsymbol{e} = (Au)(\breve{\Phi}|_{\breve{\Pi}})$, $\boldsymbol{f} = (Au)(\breve{\Psi}|_{\breve{\Lambda}})$, with
$u := \boldsymbol{d}^\top \Phi|_\Pi + \boldsymbol{c}^\top \Psi|_\Lambda$.
if $\breve{\Pi} \cup \breve{\Lambda} \neq \varnothing$ **then**
$\breve{\Pi}_B := \{\lambda \in \breve{\Pi} : \left| \operatorname{supp} \breve{\phi}_\lambda \cap \cup_{\mu \in \Lambda_\ell} S(\mu) \right| > 0\}$, $\breve{\Pi}_A := \breve{\Pi} \setminus \breve{\Pi}_B$
$\Pi_B := \{\lambda \in \Pi : \left| \operatorname{supp} \phi_\lambda \cap \left(\cup_{\mu \in \breve{\Lambda}_\ell} \breve{S}(\mu) \cup_{\gamma \in \breve{\Pi}_B} \operatorname{supp} \breve{\phi}_\gamma \right) \right| > 0\}$
$\Pi_A := \Pi \setminus \Pi_B$
$\underline{\breve{\Pi}} := \operatorname{supp}(\breve{\mathtt{p}}_\ell, \breve{\Pi}_B) \cup \operatorname{supp}(\breve{\mathtt{q}}_\ell, \breve{\Lambda}_\ell)$
$\underline{\Pi} := \operatorname{supp}(\mathtt{p}_\ell, \Pi_B) \cup \operatorname{supp}(\mathtt{q}_\ell, \Lambda_\ell)$
$\underline{\boldsymbol{d}} := \mathtt{p}_\ell \boldsymbol{d}|_{\Pi_B} + \mathtt{q}_\ell \boldsymbol{c}|_{\Lambda_\ell}$
$[\underline{\boldsymbol{e}}, \underline{\boldsymbol{f}}] := \mathtt{eval}(A)(\ell+1, \underline{\breve{\Pi}}, \breve{\Lambda}_{\ell+1\uparrow}, \underline{\Pi}, \Lambda_{\ell+1\uparrow}, \underline{\boldsymbol{d}}, \boldsymbol{c}|_{\Lambda_{\ell+1\uparrow}})$
$\boldsymbol{e} = \begin{bmatrix} \boldsymbol{e}|_{\breve{\Pi}_A} \\ \boldsymbol{e}|_{\breve{\Pi}_B} \end{bmatrix} := \begin{bmatrix} (A\Phi|_\Pi)(\breve{\Phi}|_{\breve{\Pi}_A}) \boldsymbol{d} \\ (\breve{\mathtt{p}}_\ell^\top \underline{\boldsymbol{e}})|_{\breve{\Pi}_B} \end{bmatrix}$
$\boldsymbol{f} = \begin{bmatrix} \boldsymbol{f}|_{\breve{\Lambda}_\ell} \\ \boldsymbol{f}|_{\breve{\Lambda}_{\ell+1\uparrow}} \end{bmatrix} := \begin{bmatrix} (\breve{\mathtt{q}}_\ell^\top \underline{\boldsymbol{e}})|_{\breve{\Lambda}_\ell} \\ \underline{\boldsymbol{f}} \end{bmatrix}$

Algorithm 5.2 Function $\mathtt{eval}(A)$.

Remark. Let $\check{\Lambda} \subset \check{\vee}$, $\Lambda \subset \vee$ be trees, and $\boldsymbol{c} \in \ell_2(\Lambda)$, then

$$(A\Psi|_\Lambda)(\check{\Psi}|_{\check{\Lambda}})\boldsymbol{c} = \texttt{eval}(A)(1, \check{\Lambda}_0, \check{\Lambda}_{1\uparrow}, \Lambda_0, \Lambda_{1\uparrow}, \boldsymbol{c}|_{\Lambda_0}, \boldsymbol{c}|_{\Lambda_{1\uparrow}}). \qquad \Diamond$$

Theorem 5.3.3. *A call of* `eval` *yields the output as specified, at the cost of* $\mathcal{O}(\#\check{\Pi}\check{\mathrm{I}} + \#\check{\Lambda} + \#\Pi + \#\Lambda)$ *operations.*

Proof. See Appendix 5.A. □

Routines `evalupp`, `evallow` For $A\colon \operatorname{span}\Phi \to (\operatorname{span}\check{\Phi})'$ local and *linear*, set

$$\boldsymbol{A} := (A\Psi)(\check{\Psi}) = [(A\psi_\mu)(\check{\psi}_\lambda)]_{(\lambda,\mu)\in\check{\vee}\times\vee}$$

as well as $\boldsymbol{U} := [(A\psi_\mu)(\check{\psi}_\lambda)]_{|\lambda|\le|\mu|}$ and $\boldsymbol{L} := [(A\psi_\mu)(\check{\psi}_\lambda)]_{|\lambda|>|\mu|}$ so $\boldsymbol{A} = \boldsymbol{L} + \boldsymbol{U}$. As sketched in the introduction of this section, this splitting is going to be necessary for the application of system matrices in the tensor-product setting; cf. (5.3.1). Algorithms 5.3 and 5.4 below evaluate $\boldsymbol{U}$ and $\boldsymbol{L}$ in linear complexity.

Data: $\ell \in \mathbb{N}$, $\check{\Pi}\check{\mathrm{I}} \subset \check{\Delta}_{\ell-1}$, $\Pi \subset \Delta_{\ell-1}$, ℓ-trees $\check{\Lambda} \subset \check{\vee}_{\ell\uparrow}$ and $\Lambda \subset \vee_{\ell\uparrow}$,
$\boldsymbol{d} \in \mathbb{R}^{\#\Pi}$, $\boldsymbol{c} \in \mathbb{R}^{\#\Lambda}$.
Result: $[\boldsymbol{e}, \boldsymbol{f}]$ where $\boldsymbol{e} = (Au)(\check{\Phi}|_{\check{\Pi}\check{\mathrm{I}}})$, $\boldsymbol{f} = \boldsymbol{U}|_{\check{\Lambda}\times\Lambda}\boldsymbol{c}$, with
$u := \boldsymbol{d}^\top \Phi|_\Pi + \boldsymbol{c}^\top \Psi|_\Lambda$.
if $\check{\Pi}\check{\mathrm{I}} \cup \check{\Lambda} \neq \varnothing$ **then**
$\check{\Pi}\check{\mathrm{I}}_B := \{\lambda \in \check{\Pi}\check{\mathrm{I}} : \left|\operatorname{supp}\check{\phi}_\lambda \cap \cup_{\mu\in\Lambda_\ell} S(\mu)\right| > 0\}$, $\check{\Pi}\check{\mathrm{I}}_A := \check{\Pi}\check{\mathrm{I}} \setminus \check{\Pi}\check{\mathrm{I}}_B$
$\underline{\check{\Pi}\check{\mathrm{I}}} := \operatorname{supp}(\check{\mathfrak{p}}_\ell, \check{\Pi}\check{\mathrm{I}}_B) \cup \operatorname{supp}(\check{\mathfrak{q}}_\ell, \check{\Lambda}_\ell)$
$\underline{\Pi} := \operatorname{supp}(\mathfrak{q}_\ell, \Lambda_\ell)$
$\underline{\boldsymbol{d}} := \mathfrak{q}_\ell \boldsymbol{c}|_{\Lambda_\ell}$
$[\underline{\boldsymbol{e}}, \underline{\boldsymbol{f}}] := \texttt{evalupp}(A)(\ell+1, \underline{\check{\Pi}\check{\mathrm{I}}}, \check{\Lambda}_{\ell+1\uparrow}, \underline{\Pi}, \Lambda_{\ell+1\uparrow}, \underline{\boldsymbol{d}}, \boldsymbol{c}|_{\Lambda_{\ell+1\uparrow}})$
$\boldsymbol{e} = \begin{bmatrix} \boldsymbol{e}|_{\check{\Pi}\check{\mathrm{I}}_A} \\ \boldsymbol{e}|_{\check{\Pi}\check{\mathrm{I}}_B} \end{bmatrix} := \begin{bmatrix} (A\Phi|_\Pi)(\check{\Phi}|_{\check{\Pi}\check{\mathrm{I}}_A})\boldsymbol{d} \\ (A\Phi|_\Pi)(\check{\Phi}|_{\check{\Pi}\check{\mathrm{I}}_B})\boldsymbol{d} + (\check{\mathfrak{p}}_\ell^\top \underline{\boldsymbol{e}})|_{\check{\Pi}\check{\mathrm{I}}_B} \end{bmatrix}$
$\boldsymbol{f} = \begin{bmatrix} \boldsymbol{f}|_{\check{\Lambda}_\ell} \\ \boldsymbol{f}|_{\check{\Lambda}_{\ell+1\uparrow}} \end{bmatrix} := \begin{bmatrix} (\check{\mathfrak{q}}_\ell^\top \underline{\boldsymbol{e}})|_{\check{\Lambda}_\ell} \\ \underline{\boldsymbol{f}} \end{bmatrix}$

Algorithm 5.3 Function `evalupp`(A).

Remark. Let $\check{\Lambda} \subset \check{\vee}$, $\Lambda \subset \vee$ be trees, and $\boldsymbol{c} \in \ell_2(\Lambda)$, then

$$\boldsymbol{U}|_{\check{\Lambda}\times\Lambda}\boldsymbol{c} = \texttt{evalupp}(A)(1, \check{\Lambda}_0, \check{\Lambda}_{1\uparrow}, \Lambda_0, \Lambda_{1\uparrow}, \boldsymbol{c}|_{\Lambda_0}, \boldsymbol{c}|_{\Lambda_{1\uparrow}}). \qquad \Diamond$$

Theorem 5.3.4. *A call of* `evalupp` *yields the output as specified, at the cost of* $\mathcal{O}(\#\check{\Pi}\check{\mathrm{I}} + \#\check{\Lambda} + \#\Pi + \#\Lambda)$ *operations.*

Proof. See Appendix 5.A. □

Data: $\ell \in \mathbb{N}$, $\Pi \subset \Delta_{\ell-1}$, ℓ-trees $\breve{\Lambda} \subset \breve{\vee}_{\ell\uparrow}$ and $\Lambda \subset \vee_{\ell\uparrow}$, $\boldsymbol{d} \in \mathbb{R}^{\#\Pi}$, $\boldsymbol{c} \in \mathbb{R}^{\#\Lambda}$.
Result: $\boldsymbol{f} = (A\Phi|_{\Pi})(\breve{\Psi}|_{\breve{\Lambda}})\boldsymbol{d} + \boldsymbol{L}|_{\breve{\Lambda}\times\Lambda}\boldsymbol{c}$.
if $\breve{\Pi} \cup \breve{\Lambda} \neq \varnothing$ **then**
$\Pi_B := \{\lambda \in \Pi : \left|\operatorname{supp}\phi_\lambda \cap \cup_{\mu\in\breve{\Lambda}_\ell}\breve{S}(\mu)\right| > 0\}$,
$\underline{\Pi} := \operatorname{supp}(\mathfrak{p}_\ell, \Pi_B) \cup \operatorname{supp}(\mathfrak{q}_\ell, \Lambda_\ell)$
$\underline{\Pi}_B := \operatorname{supp}(\mathfrak{p}_\ell, \Pi_B)$
$\underline{\breve{\Pi}} := \operatorname{supp}(\breve{\mathfrak{q}}_\ell, \breve{\Lambda}_\ell)$
$\underline{\boldsymbol{d}} := \mathfrak{p}_\ell \boldsymbol{d}|_{\Pi_B} + \mathfrak{q}_\ell \boldsymbol{c}|_{\Lambda_\ell}$
$\underline{\boldsymbol{e}} := (A\Phi|_{\underline{\Pi}_B})(\breve{\Phi}|_{\underline{\breve{\Pi}}})\mathfrak{p}_\ell \boldsymbol{d}|_{\Pi_B}$
$\boldsymbol{f} = \begin{bmatrix}\boldsymbol{f}|_{\breve{\Lambda}_\ell} \\ \boldsymbol{f}|_{\breve{\Lambda}_{\ell+1\uparrow}}\end{bmatrix} := \begin{bmatrix}(\breve{\mathfrak{q}}_\ell^\top \underline{\boldsymbol{e}})|_{\breve{\Lambda}_\ell} \\ \texttt{evallow}(A)(\ell+1, \breve{\Lambda}_{\ell+1\uparrow}, \underline{\Pi}, \Lambda_{\ell+1\uparrow}, \underline{\boldsymbol{d}}, \boldsymbol{c}|_{\Lambda_{\ell+1\uparrow}})\end{bmatrix}$

Algorithm 5.4 Function `evallow`(A).

Remark. Let $\breve{\Lambda} \subset \breve{\vee}$, $\Lambda \subset \vee$ be trees, and $\boldsymbol{c} \in \ell_2(\Lambda)$, then

$$\boldsymbol{L}|_{\breve{\Lambda}\times\Lambda}\boldsymbol{c} = \texttt{evallow}(A)(1, \breve{\Lambda}_{1\uparrow}, \Lambda_0, \Lambda_{1\uparrow}, \boldsymbol{c}|_{\Lambda_0}, \boldsymbol{c}|_{\Lambda_{1\uparrow}}). \qquad \lozenge$$

Theorem 5.3.5. *A call of* `evallow` *yields the output as specified, at the cost of* $\mathcal{O}(\#\breve{\Lambda} + \#\Pi + \#\Lambda)$ *operations.*

Proof. See Appendix 5.A. □

5.3.2 Applying tensor-product operators on double-trees

For $i \in \{0,1\}$, let $A_i \colon \operatorname{span}\Phi_i \to \operatorname{span}\breve{\Phi}'_i$ be local and linear and let

$$\boldsymbol{A}_i = (A\Psi_i)(\breve{\Psi}_i) = [(A\psi^i_\mu)(\breve{\psi}^i_\lambda)]_{\lambda\in\breve{\vee}^i,\mu\in\vee^i} = \boldsymbol{L}_i + \boldsymbol{U}_i,$$

where $\boldsymbol{U}_i := [(\boldsymbol{A}_i)_{\lambda,\mu}]_{|\lambda|\leq|\mu|}$ and $\boldsymbol{L}_i := [(\boldsymbol{A}_i)_{\lambda,\mu}]_{|\lambda|>|\mu|}$. Set $\neg i := 1 - i$.

Definition 5.3.6 (Double-tree)**.** Define the coordinate projector $P_i(b_0, b_1) := b_i$. We call $\boldsymbol{\Lambda} \subset \{\breve{\vee}^0 \times \breve{\vee}^1, \vee^0 \times \breve{\vee}^1, \breve{\vee}^0 \times \vee^1, \vee^0 \times \vee^1\}$ a *double-tree* when for $i \in \{0,1\}$ and any $\mu \in P_{\neg i}\boldsymbol{\Lambda}$, the *fiber*

$$\boldsymbol{\Lambda}_{i,\mu} := P_i(P_{\neg i}|_{\boldsymbol{\Lambda}})^{-1}\{\mu\}$$

is a tree (in $\breve{\vee}^i$ or $\vee^i$), i.e., $\boldsymbol{\Lambda}$ is a double-tree when 'frozen' in each of its coordinates, at any value of that coordinate, it is a tree in the other coordinate. $\lozenge$

From $\boldsymbol{\Lambda} = \cup_{\mu\in P_{\neg i}\boldsymbol{\Lambda}}(P_{\neg i}|_{\boldsymbol{\Lambda}})^{-1}\{\mu\}$, we have $P_i\boldsymbol{\Lambda} = \cup_{\mu\in P_{\neg i}\boldsymbol{\Lambda}}\boldsymbol{\Lambda}_{i,\mu}$, which, being a union of trees, is a tree itself. See also Figure 5.2.

For a subset $\triangleleft$ of a (double) index set $\lozenge$, let $I^{\lozenge}_{\triangleleft}$ denote the extension operator with zeros of a vector supported on $\triangleleft$ to one on $\lozenge$, and let $R^{\lozenge}_{\triangleleft}$ denotes its

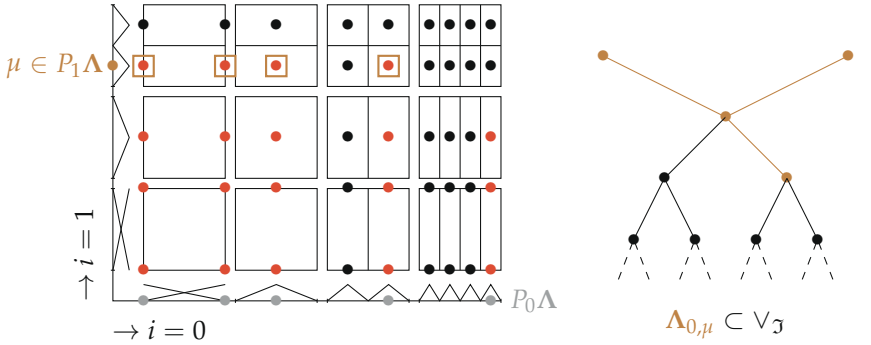


Figure 5.2 With $\vee_{\mathfrak{J}}$ from Fig. 5.1: $\vee_{\mathfrak{J}} \times \vee_{\mathfrak{J}}$ in black; a double-tree $\mathbf{\Lambda} \subset \vee_{\mathfrak{J}} \times \vee_{\mathfrak{J}}$ in red; the projection $P_0\mathbf{\Lambda}$ in gray, and a fiber $\mathbf{\Lambda}_{0,\mu}$ for $\mu \in P_1\mathbf{\Lambda}$ in brown.

(formal) adjoint, being the restriction operator of a vector supported on $\Diamond$ to one on $\triangleleft$. Since the set $\Diamond$ will always be clear from the context, we will denote these operators simply by $I_{\triangleleft}$ and $R_{\triangleleft}$.

As sketched in the introduction of this section, the pieces are now in place to apply $R_{\check{\mathbf{\Lambda}}}(A_0 \otimes A_1)I_{\mathbf{\Lambda}}$ in linear complexity.

Theorem 5.3.7. *Let $\check{\mathbf{\Lambda}} \subset \check{\vee}^0 \times \check{\vee}^1$, $\mathbf{\Lambda} \subset \vee^0 \times \vee^1$ be finite double-trees. Then*

$$\mathbf{\Sigma} := \bigcup_{\lambda \in P_0\mathbf{\Lambda}} \left(\{\lambda\} \times \bigcup_{\left\{\mu \in P_0\check{\mathbf{\Lambda}} : |\mu| = |\lambda|+1,\, |\check{S}^0(\mu) \cap S^0(\lambda)| > 0\right\}} \check{\mathbf{\Lambda}}_{1,\mu} \right),$$

$$\mathbf{\Theta} := \bigcup_{\lambda \in P_1\mathbf{\Lambda}} \left(\{\mu \in P_0\check{\mathbf{\Lambda}} : \exists \gamma \in \mathbf{\Lambda}_{0,\lambda} \text{ s.t. } |\gamma| = |\mu|,\, |\check{S}^0(\mu) \cap S^0(\gamma)| > 0\} \times \{\lambda\} \right),$$

are double-trees with $\#\mathbf{\Sigma} \lesssim \#\check{\mathbf{\Lambda}}$ and $\#\mathbf{\Theta} \lesssim \#\mathbf{\Lambda}$, and

$$\begin{aligned} R_{\check{\mathbf{\Lambda}}}(A_0 \otimes A_1)I_{\mathbf{\Lambda}} = {} & R_{\check{\mathbf{\Lambda}}}(\boldsymbol{L}_0 \otimes \mathrm{Id})I_{\mathbf{\Sigma}}R_{\mathbf{\Sigma}}(\mathrm{Id} \otimes A_1)I_{\mathbf{\Lambda}} + \\ & R_{\check{\mathbf{\Lambda}}}(\mathrm{Id} \otimes A_1)I_{\mathbf{\Theta}}R_{\mathbf{\Theta}}(\boldsymbol{U}_0 \otimes \mathrm{Id})I_{\mathbf{\Lambda}}. \end{aligned}$$

Proof. See Appendix 5.A. □

Applying $R_{\check{\mathbf{\Lambda}}}(\boldsymbol{L}_0 \otimes \mathrm{Id})I_{\mathbf{\Sigma}}$ boils down to applying $R_{\check{\mathbf{\Lambda}}_{0,\mu}}\boldsymbol{L}_0 I_{\mathbf{\Sigma}_{0,\mu}}$ for every $\mu \in P_1\mathbf{\Sigma} \cap P_1\check{\mathbf{\Lambda}}$. Such an application can be performed in $\mathcal{O}(\#\check{\mathbf{\Lambda}}_{0,\mu} + \#\mathbf{\Sigma}_{0,\mu})$ operations by means of a call of $\texttt{evallow}(A_0)$; see also Algorithm 5.9. Since $\sum_{\mu \in \check{\vee}_1} \#\check{\mathbf{\Lambda}}_{0,\mu} + \#\mathbf{\Sigma}_{0,\mu} = \#\check{\mathbf{\Lambda}} + \#\mathbf{\Sigma}$, we conclude that the application of $R_{\check{\mathbf{\Lambda}}}(\boldsymbol{L}_0 \otimes \mathrm{Id})I_{\mathbf{\Sigma}}$ can be performed in $\mathcal{O}(\#\check{\mathbf{\Lambda}} + \#\mathbf{\Sigma})$ operations.

Similarly, applications of $R_{\mathbf{\Sigma}}(\mathrm{Id} \otimes A_1)I_{\mathbf{\Lambda}}$, $R_{\check{\mathbf{\Lambda}}}(\mathrm{Id} \otimes A_1)I_{\mathbf{\Theta}}$, and $R_{\mathbf{\Theta}}(\boldsymbol{U}_0 \otimes \mathrm{Id})I_{\mathbf{\Lambda}}$ using calls of $\texttt{eval}(A_1)$, $\texttt{eval}(A_1)$, and $\texttt{evalupp}(A_0)$ respectively, can be done in $\mathcal{O}(\#\mathbf{\Sigma} + \#\mathbf{\Lambda})$, $\mathcal{O}(\#\check{\mathbf{\Lambda}} + \#\mathbf{\Theta})$, and $\mathcal{O}(\#\mathbf{\Theta} + \#\mathbf{\Lambda})$ operations. From $\#\mathbf{\Sigma} \lesssim \#\check{\mathbf{\Lambda}}$ and $\#\mathbf{\Theta} \lesssim \#\mathbf{\Lambda}$ we conclude the following.

Corollary 5.3.8. *Let $\check{\mathbf{\Lambda}} \subset \check{\vee}^0 \times \check{\vee}^1$, $\mathbf{\Lambda} \subset \vee^0 \times \vee^1$ be finite double-trees, then $R_{\check{\mathbf{\Lambda}}}(A_0 \otimes A_1)I_{\mathbf{\Lambda}}$ can be applied in $\mathcal{O}(\#\check{\mathbf{\Lambda}} + \#\mathbf{\Lambda})$ operations.*

5.4 The heat equation and practical realization

In this section, we consider the numerical approximation of the *heat equation*

$$\begin{cases} \frac{\mathrm{d}u}{\mathrm{d}t}(t) - (\Delta_{\boldsymbol{x}} u)(t) &= g(t) \quad (t \in I), \\ u(0) &= u_0. \end{cases} \tag{5.4.1}$$

For some bounded domain $\Omega \subset \mathbb{R}^2$, we take $H := L_2(\Omega)$ and $V := H_0^1(\Omega)$, so that $X = L_2(I; H_0^1(\Omega)) \cap H^1(I; H^{-1}(\Omega))$ and $Y = L_2(I; H_0^1(\Omega))$. We define

$$a(t; \eta, \zeta) := \int_\Omega \nabla\eta \cdot \nabla\zeta \,\mathrm{d}\boldsymbol{x},$$

and aim to solve the parabolic initial value problem (5.2.1) numerically. The bilinear forms present in our variational formulation (5.2.4) satisfy

$$A = M_t \otimes A_{\boldsymbol{x}}, \quad B = D_t \otimes M_{\boldsymbol{x}} + A, \quad \text{and} \quad \gamma_0'\gamma_0 = G_t \otimes M_{\boldsymbol{x}}$$

where

$$\begin{aligned} (M_t v)(w) &:= \int_I vw \,\mathrm{d}t, \quad (D_t v)(w) := \int_I v'w \,\mathrm{d}t, \quad (G_t v)(w) := v(0)w(0), \\ (A_{\boldsymbol{x}}\eta)(\zeta) &:= \int_\Omega \nabla\eta \cdot \nabla\zeta \,\mathrm{d}\boldsymbol{x}, \quad (M_{\boldsymbol{x}}\eta)(\zeta) := \int_\Omega \eta\zeta \,\mathrm{d}\boldsymbol{x}. \end{aligned} \tag{5.4.2}$$

In this section, we first construct suitable tensor-product bases for X and Y which functions are wavelets in time and hierarchical finite element functions in space. We then build our discrete 'trial' and 'test' spaces $(X^\delta, Y^\delta)_{\delta\in\Delta}$ as the span of subsets of these tensor-product bases. We finish with concrete preconditioners K_X^δ and K_Y^δ, the basis necessary for error estimation in the adaptive loop, and evaluation of the right-hand side of (5.2.9) using interpolants.

5.4.1 Wavelets in time

We construct piecewise linear wavelet bases Σ for $H^1(I)$ and Ξ for $L_2(I)$.

Basis on the trial side For Σ, we choose the three-point wavelet from [Ste98]; for completeness, we include its construction. For $\ell \geq 0$, define the scaling functions as the nodal continuous piecewise linears w.r.t. a uniform partition into 2^ℓ subintervals, i.e., $\Phi_\ell^\Sigma := \{\phi_{(\ell,n)} : 0 \leq n \leq 2^\ell\}$ with $\phi_{(\ell,n)}(k2^{-\ell}) = \delta_{kn}$ for $0 \leq k \leq 2^\ell$. Set $\Sigma_0 := \Phi_0^\Sigma$, and $\Sigma_\ell := \{\sigma_\lambda : \lambda := (\ell, n) \text{ with } 0 \leq n < 2^{\ell-1}\}$ with $\sigma_\lambda = \sigma_{(\ell,n)}$ for $\ell \geq 1$ as in the right of Figure 5.3. Note that each *tree-point wavelet* σ_λ is a linear combination of three nodal functions from Φ_ℓ^Σ.

By imposing the parent-child structure

$$\tilde\lambda \lhd_\Sigma \lambda \iff |\tilde\lambda| + 1 = |\lambda| \text{ and } |\operatorname{supp}\sigma_\lambda \cap \operatorname{supp}\sigma_{\tilde\lambda}| > 0, \tag{5.4.3}$$

on any two indices $\tilde\lambda, \lambda$, we get the tree shown left in Figure 5.3.

Define $\Sigma := \cup_{\ell\geq 0}\Sigma_\ell$, $\vee_\Sigma := \{\lambda : \sigma_\lambda \in \Sigma\}$, and $S(\sigma_\lambda) := \operatorname{supp}\sigma_\lambda$. We see that Σ satisfies (5.3.2)–(5.3.3) and that the Φ_ℓ^Σ satisfy (5.3.4)–(5.3.7). Moreover, one can show that Σ is a Riesz basis for $L_2(I)$ (cf. [Ste98, Thm. 4.2]), and that $\{2^{-|\lambda|}\sigma_\lambda\}$ is a Riesz basis for $H^1(I)$ (cf. [Ste98, Thm. 4.3]).

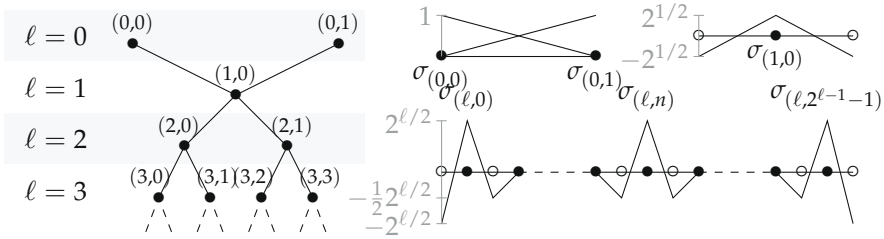


Figure 5.3 Left: three-point wavelet index set $\vee_\Sigma$ with parent-child relations; right: three-point wavelets.

Basis on the test side We construct an $L_2(I)$-orthonormal basis Ξ.

For $\ell \geq 0$, define the (discontinuous) piecewise linear scaling functions w.r.t. a uniform partition into 2^ℓ subintervals by $\Phi_\ell^\Xi := \{\phi_{(\ell,n)} : 0 \leq n < 2^{\ell+1}\}$ where $\phi_{(0,0)}(t) := \mathbb{1}_{[0,1]}(t)$ and $\phi_{(0,1)}(t) := \sqrt{3}(2t-1)\mathbb{1}_{[0,1]}$, and for $\ell \geq 1$, $\phi_{(\ell,2k)}(t) := \phi_{(0,0)}(2^\ell t - k)$ and $\phi_{(\ell,2k+1)}(t) := \phi_{(0,1)}(2^\ell t - k)$. Let $\Xi_0 := \Phi_0^\Xi$, and define $\Xi_1 := \{\xi_{(1,0)}, \xi_{(1,1)}\}$ as in the right of Figure 5.4. For $\ell \geq 2$, we take $\Xi_\ell := \{\xi_{(\ell,n)} : 0 \leq n < 2^\ell\}$ with

$$\xi_{(\ell,2k)}(t) := 2^{(\ell-1)/2}\xi_{(1,0)}(2^{\ell-1}t - k), \quad \xi_{(\ell,2k+1)}(t) := 2^{(\ell-1)/2}\xi_{(1,1)}(2^{\ell-1}t - k).$$

The resulting $\Xi := \cup_{\ell\geq 0}\Xi_\ell$ is an *orthonormal* basis for $L_2(I)$, and together with its scaling functions $\cup_\ell \Phi_\ell^\Xi$, the conditions from §5.3.1 are satisfied with $S(\xi_\mu) := \operatorname{supp}\xi_\mu$. We impose a parent-child relation analogously to (5.4.3); see the left of Figure 5.4.

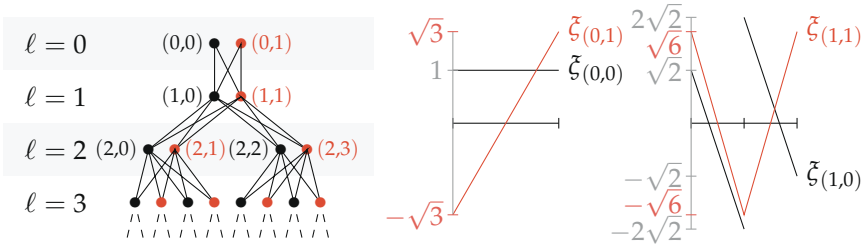


Figure 5.4 Left: orthonormal wavelet index set $\vee_\Xi$ with parent-child relations; right: the wavelets at levels 0 and 1.

5.4.2 Finite elements in space

Let $\mathbb{T}$ be the family of all *conforming* partitions of Ω into triangles that can be created by Newest Vertex Bisection from some given conforming initial triangulation $\mathcal{T}_\perp$ with an assignment of newest vertices satisfying the matching condition; cf. [Ste08].

Define $\mathfrak{T} := \cup_{\mathcal{T}\in\mathbb{T}}\{T : T \in \mathcal{T}\}$. For $T \in \mathfrak{T}$, set $\mathrm{gen}(T)$ as the number of bisections needed to create T from its 'ancestor' $T' \in \mathcal{T}_\perp$. With $\mathfrak{N}$ the set of all vertices of all $T \in \mathfrak{T}$, for $\nu \in \mathfrak{N}$ we set $\mathrm{gen}(\nu) := \min\{\mathrm{gen}(T) : \nu \text{ is a vertex of } T \in \mathfrak{T}\}$.

Any $\nu \in \mathfrak{N}$ with $\mathrm{gen}(\nu) > 0$ is the midpoint of an edge e_ν of one or two $T \in \mathfrak{T}$ with $\mathrm{gen}(T) = \mathrm{gen}(\nu) - 1$. The set of newest vertices $\tilde{\nu}$ of these T, so those vertices of T with $|\tilde{\nu}| = \mathrm{gen}(\nu) - 1$, are defined as the parents of ν, denoted $\tilde{\nu} \lhd_{\mathfrak{N}} \nu$. The set of *godparents* of ν, denoted $\mathtt{gp}(\nu)$, are defined as the two endpoints of e_ν. Vertices with $\mathrm{gen}(\nu) = 0$ have no parents or godparents.

Example 5.4.1. In Fig. 5.5, the parents of ν_4 are ν_1 and ν_3 and its godparents are ν_0, ν_2; the sole parent of ν_5 is ν_4, and its godparents are ν_0 and ν_3. ◊

Proposition ([DKS16]). *An (essentially) non-overlapping partition $\mathcal{T}$ of $\overline{\Omega}$ into triangles is in $\mathbb{T}$ if and only if the set $N_\mathcal{T}$ of vertices of all $T \in \mathcal{T}$ forms a* tree *in the sense of §5.3.1, meaning that it contains every vertex of generation zero as well as all parents of any $\nu \in N_\mathcal{T}$; see also Figure 5.5.*

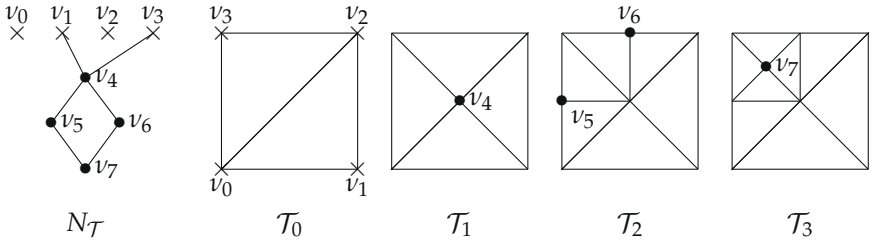


Figure 5.5 Vertex tree $N_\mathcal{T}$ and its triangulation $\mathcal{T}$ shown level-by-level.

Let $\mathcal{O}$ be the collection of spaces $W_\mathcal{T}$ of continuous piecewise linears over $\mathcal{T} \in \mathbb{T}$ vanishing on $\partial\Omega$. For $\nu \in \mathfrak{N}$, we set ψ_ν as that continuous piecewise linear function on the *uniform partition* $\mathcal{T}_\nu := \{T \in \mathfrak{T} : \mathrm{gen}(T) = \mathrm{gen}(\nu)\} \in \mathbb{T}$ for which $\psi_\nu(\tilde{\nu}) = \delta_{\nu\tilde{\nu}}$ for $\tilde{\nu} \in \mathcal{T}_\nu$. Setting $\mathfrak{N}_0 := \mathfrak{N} \setminus \partial\Omega$, the collection $\{\psi_\nu : \nu \in \mathfrak{N}_0\}$ is known as the *hierarchical basis*. For $\mathcal{T} \in \mathbb{T}$, write $N_{\mathcal{T},0} := N_\mathcal{T} \setminus \partial\Omega$ and $\Psi_\mathcal{T} := \{\psi_\nu : \nu \in N_{\mathcal{T},0}\}$; it holds that $W_\mathcal{T} = \mathrm{span}\Psi_\mathcal{T}$.

Applying stiffness matrices The hierarchical basis satisfies conditions (5.3.2) and (5.3.3), and so, the application of stiffness matrices $(A\Psi_\mathcal{T})(\Psi_\mathcal{T})$ for $A \in \{A_x, M_x\}$ can be done through $\mathtt{eval}(A)$.[1] However, the computation in Theorem 5.3.7 does not involve the lower and upper parts of A. This crucial insight allows for a faster and easier approach using standard finite element techniques: $\mathrm{span}\Psi_\mathcal{T}$ is a continuous piecewise linear finite element space, so it has a *canonical* single-scale basis $\Phi_\mathcal{T} := \mathrm{span}\{\phi_{\mathcal{T},\nu}\}$ characterized by $\phi_{\mathcal{T},\nu}(\tilde{\nu}) = \delta_{\nu\tilde{\nu}}$ for $\tilde{\nu} \in N_{\mathcal{T},0}$, for which the application of $(A\Phi_\mathcal{T})(\Phi_\mathcal{T})$ at linear cost using local element matrices is standard. This is different from the

[1] This would require the definition of a suitable single-scale basis.

general setting in §5.3.1, in that $\dim \Phi_{\mathcal{T}} = \dim \Psi_{\mathcal{T}}$ also for locally refined triangulations. With T the transformation defined by $\Psi_{\mathcal{T}} = T^\top \Phi_{\mathcal{T}}$, we find

$$(A\Psi_{\mathcal{T}})(\Psi_{\mathcal{T}}) = T^\top (A\Phi_{\mathcal{T}})(\Phi_{\mathcal{T}})T. \tag{5.4.4}$$

We can apply T in linear complexity by iterating over the vertices bottom-up while applying elementary local transformations in which not parent-child, but *godparent*-child relations play a role.

5.4.3 Inf-sup stable family of trial- and test spaces

With Σ and Ξ from §5.4.1 and $\Psi_{\mathfrak{N}_0} := \{\psi_\nu : \nu \in \mathfrak{N}_0\}$ from §5.4.2, we find that $X = \overline{\operatorname{span}(\Sigma \otimes \Psi_{\mathfrak{N}_0})}$ and $Y = \overline{\operatorname{span}(\Xi \otimes \Psi_{\mathfrak{N}_0})}$. We now turn to the construction of X^δ and Y^δ.

Definition 5.4.2. For a double-tree $\boldsymbol{\Lambda}^\delta \subset \vee_\Sigma \times \mathfrak{N}$, define $\boldsymbol{\Lambda}_0^\delta := \boldsymbol{\Lambda}^\delta \setminus \vee_\Sigma \times \partial\Omega$. We construct our 'trial' space as

$$X^\delta := \operatorname{span}\{\sigma_\lambda \otimes \psi_\nu : (\lambda, \nu) \in \boldsymbol{\Lambda}_0^\delta\}.$$

Defining the double-tree $\boldsymbol{\Lambda}_{Y,0}^\delta \subset \vee_\Xi \times \mathfrak{N}_0$ as

$$\boldsymbol{\Lambda}_{Y,0}^\delta := \{(\mu, \nu) : \exists (\lambda, \nu) \in \boldsymbol{\Lambda}_0^\delta,\ \mu \in \vee_\Xi,\ |\mu| = |\lambda|,\ |\operatorname{supp} \xi_\mu \cap \operatorname{supp} \sigma_\lambda| > 0\},$$

we set the 'test' space to be $Y^\delta = Y^\delta(X^\delta) := \operatorname{span}\{\xi_\mu \otimes \psi_\nu : (\mu, \nu) \in \boldsymbol{\Lambda}_{Y,0}^\delta\}$. ◊

Theorem 5.4.3 ([SvVW21, Props. 5.2, 5.3])**.** *Define* $\Delta := \{\delta : \boldsymbol{\Lambda}^\delta \subset \vee_\Sigma \times \mathfrak{N}$ *is a double-tree*$\}$ *equipped with the partial ordering* $\delta \preceq \tilde{\delta} \iff \boldsymbol{\Lambda}^\delta \subseteq \boldsymbol{\Lambda}^{\tilde{\delta}}$. *With* X^δ *and* Y^δ *as above, uniform inf-sup stability holds; cf.* (5.2.6)*.*

Definition 5.4.4. Given a double-tree $\boldsymbol{\Lambda}^\delta \subset \vee_\Sigma \times \mathfrak{N}$, we define $\boldsymbol{\Lambda}^{\underline{\delta}} \supset \boldsymbol{\Lambda}^\delta$ by adding, for $(\lambda, \nu) \in \boldsymbol{\Lambda}^\delta$ and any child $\tilde{\lambda}$ of λ and descendant $\tilde{\nu}$ of ν up to generation 2, all pairs $(\tilde{\lambda}, \nu)$ and $(\lambda, \tilde{\nu})$. We expect this choice of $X^{\underline{\delta}}$ to provide saturation; cf. (5.2.8). ◊

5.4.4 Preconditioners

We follow [SvVW21, §5.6] for the construction of optimal preconditioners K_Y^δ for $E_Y^{\delta\prime} A E_Y^\delta$ and K_X^δ for $S^{\underline{\delta}\delta}$ necessary for solving (5.2.9). With notation from Definition 5.3.6, we equip X^δ and Y^δ with bases

$$\begin{cases} \bigcup\limits_{\lambda \in P_0 \boldsymbol{\Lambda}_0^\delta} \sigma_\lambda \otimes \Psi_\lambda^\delta \quad \text{with} \quad \Psi_\lambda^\delta := \{\psi_\nu : \nu \in (\boldsymbol{\Lambda}_0^\delta)_{1,\lambda}\}, \\ \bigcup\limits_{\mu \in P_0 \boldsymbol{\Lambda}_{Y,0}^\delta} \xi_\mu \otimes \Psi_\mu^\delta \quad \text{with} \quad \Psi_\mu^\delta := \{\psi_\nu : \nu \in (\boldsymbol{\Lambda}_{Y,0}^\delta)_{1,\mu}\}. \end{cases}$$

In matrix form, the preconditioners from [SvVW21, §5.6] then satisfy

$$\begin{cases} \boldsymbol{K}_Y^\delta := \operatorname{blockdiag}[\boldsymbol{K}_\mu^\delta]_{\mu\in P_0\boldsymbol{\Lambda}_{Y,0}^\delta} & \text{where} \quad \boldsymbol{K}_\mu^\delta \eqsim (\boldsymbol{A}_\mu^\delta)^{-1}, \\ \boldsymbol{K}_X^\delta := \operatorname{blockdiag}[\boldsymbol{K}_\lambda^\delta \boldsymbol{A}_\lambda^\delta \boldsymbol{K}_\lambda^\delta]_{\lambda\in P_0\boldsymbol{\Lambda}_0^\delta} & \text{where} \quad \boldsymbol{K}_\lambda^\delta \eqsim (\boldsymbol{A}_\lambda^\delta + 2^{|\lambda|}\boldsymbol{M}_\lambda^\delta)^{-1} \end{cases}$$

with $\boldsymbol{A}_\mu^\delta := (A_{\boldsymbol{x}}\Psi_\mu^\delta)(\Psi_\mu^\delta)$, $\boldsymbol{A}_\lambda^\delta := (A_{\boldsymbol{x}}\Psi_\lambda^\delta)(\Psi_\lambda^\delta)$, and $\boldsymbol{M}_\lambda^\delta := (M_{\boldsymbol{x}}\Psi_\lambda^\delta)(\Psi_\lambda^\delta)$. Suitable spatial preconditioners $\boldsymbol{K}_\mu^\delta$ are provided by multigrid methods. In [OR00] it was shown that for quasi-uniform meshes, under a 'full-regularity' assumption, a multiplicative multigrid method yields suitable $\boldsymbol{K}_\lambda^\delta$, and we assume these results to transfer to our locally refined triangulations $\mathcal{T} \in \mathbb{T}$. In §5.5.1, we detail our linear-complexity multigrid implementation following [WZ17].

5.4.5 Right-hand side

We follow [SvVW21, §6.4]. For $g \in C(\overline{I\times\Omega})$, $u_0 \in C(\overline{\Omega})$, we can approximate the right-hand side of (5.2.9) by interpolants, avoiding quadrature issues.

The procedure of §5.4.2 for constructing the hierarchical basis $\Psi_{\mathfrak{N}} := \{\psi_\nu : \nu \in \mathfrak{N}\}$ can be applied in time as well, yielding the basis $\{\psi_\lambda : \lambda \in \vee_{\mathfrak{I}}\}$ from Figure 5.1 which index set $\vee_{\mathfrak{I}}$ coincides with $\vee_\Sigma$. We construct $\{\tilde\psi_\nu : \nu \in \mathfrak{N}\} \subset C(\overline{\Omega})'$ biorthogonal to $\Psi_{\mathfrak{N}}$, with $\tilde\psi_\nu := \delta_\nu - \sum_{\tilde\nu\in\mathbf{gp}(\nu)} \delta_{\tilde\nu}/2$. In time, define $\{\tilde\psi_\lambda : \lambda \in \vee_{\mathfrak{I}}\} \subset C(\bar I)'$ analogously. Define the vectors $\boldsymbol{g} := [(\tilde\psi_\lambda \otimes \tilde\psi_\nu)(g)]_{(\lambda,\nu)\in\boldsymbol{\Lambda}^\delta}$ and $\boldsymbol{u}_0 := [\tilde\psi_\nu(u_0)]_{\nu\in P_1\boldsymbol{\Lambda}^\delta}$. Upon replacing (g, u_0) in (5.2.9) by the interpolants

$$ {}^\delta g := \sum_{(\lambda,\nu)\in\boldsymbol{\Lambda}^\delta} \boldsymbol{g}_{(\lambda,\nu)}\psi_\lambda\otimes\psi_\nu, \quad {}^\delta u_0 := \sum_{\nu\in P_1\boldsymbol{\Lambda}^\delta} \boldsymbol{u}_{0,\nu}\psi_\nu, $$

we can evaluate its right-hand side in linear complexity through the quantities

$$\begin{aligned} [\langle \xi_\mu\otimes\psi_\nu, {}^\delta g\rangle_{L_2(I\times\Omega)}]_{(\mu,\nu)\in\boldsymbol{\Lambda}_{Y,0}^\delta} &= R_{\boldsymbol{\Lambda}_{Y,0}^\delta}(M_t\otimes M_{\boldsymbol{x}})I_{\boldsymbol{\Lambda}^\delta}\boldsymbol{g}, \\ [\sigma_\lambda(0)\langle\psi_\nu, {}^\delta u_0\rangle_{L_2(\Omega)}]_{(\lambda,\nu)\in\boldsymbol{\Lambda}_0^\delta} &= [\sigma_\lambda(0)\boldsymbol{w}_\nu]_{(\lambda,\nu)\in\boldsymbol{\Lambda}_0^\delta} \\ \text{where} \quad \boldsymbol{w} &= (M_{\boldsymbol{x}}\Psi_{\mathfrak{N}}|_{P_1\boldsymbol{\Lambda}^\delta})(\Psi_{\mathfrak{N}}|_{P_1\boldsymbol{\Lambda}^\delta})\boldsymbol{u}_0. \end{aligned}$$

5.4.6 Two-level basis

We now discuss the construction of a uniformly X-stable basis Θ_δ, needed in the local error estimator $\boldsymbol{r}^\delta$ of (5.2.12). Following [SvVW21, §6.3], define a *modified hierarchical basis* $\{\hat\psi_\nu : \nu \in \mathfrak{N}_0\}$ by

$$\hat\psi_\nu = \psi_\nu \ \text{ when } \ \operatorname{gen}(\nu) = 0, \quad \text{else} \quad \hat\psi_\nu := \psi_\nu - \frac{\sum_{\{\tilde\nu\in\mathfrak{N}:\tilde\nu\lhd_{\mathfrak{N}}\nu\}} \frac{\int_\Omega \psi_\nu\,\mathrm{d}\boldsymbol{x}}{\int_\Omega \psi_{\tilde\nu}\,\mathrm{d}\boldsymbol{x}}\psi_{\tilde\nu}}{\#\{\tilde\nu\in\mathfrak{N} : \tilde\nu\lhd_{\mathfrak{N}}\nu\}}.$$

For any $\mathcal{T}\in\mathbb{T}$, $W_\mathcal{T} = \operatorname{span}\{\hat\psi_\nu : \nu \in N_{\mathcal{T},0}\} = \operatorname{span}\Psi_\mathcal{T}$ and the transformation from modified to unmodified hierarchical basis can be performed in linear

complexity. For $\underline{\mathcal{T}} \succeq \mathcal{T} \in \mathbb{T}$, $\boldsymbol{d} \in \ell_2(N_{\underline{\mathcal{T}},0} \setminus N_{\mathcal{T},0})$ and $v \in W_{\mathcal{T}}$, [SvVW21, Lem. 6.7] shows that

$$\begin{cases} \|v + \sum_\nu \boldsymbol{d}_\nu \hat{\psi}_\nu\|^2_{H^1(\Omega)} \eqsim \|v\|^2_{H^1(\Omega)} + \|\boldsymbol{d}\|^2, \\ \|v + \sum_\nu \boldsymbol{d}_\nu \hat{\psi}_\nu\|^2_{H^{-1}(\Omega)} \eqsim \|v\|^2_{H^{-1}(\Omega)} + \sum_\nu 4^{-\operatorname{gen}(\nu)} |\boldsymbol{d}_\nu|^2, \end{cases} \tag{5.4.5}$$

where the constants in the $\eqsim$-symbols depend on $\max_{\{\underline{\mathcal{T}} \ni \underline{T} \subset T \in \mathcal{T}\}} \{\operatorname{gen}(\underline{T}) - \operatorname{gen}(T)\}$ only. We then construct a basis for $X^{\underline{\delta}} \ominus X^\delta$ as

$$\Theta_\delta := \{e_{\lambda\nu}\sigma_\lambda \otimes \hat{\psi}_\nu : (\lambda,\nu) \in \boldsymbol{\Lambda}_0^{\underline{\delta}} \setminus \boldsymbol{\Lambda}_0^\delta\} \quad \text{with} \quad \frac{1}{e_{\lambda\nu}} = \sqrt{1 + 4^{|\lambda| - \operatorname{gen}(\nu)}}.$$

Define the *gradedness* of a double-tree $\boldsymbol{\Lambda}^\delta \subset \vee_\Sigma \times \mathfrak{N}$ as the smallest $L_\delta \in \mathbb{N}$ for which every $(\lambda,\nu) \in \boldsymbol{\Lambda}^\delta$ with $\tilde{\nu}$ an ancestor of ν with $\operatorname{gen}(\nu) - \operatorname{gen}(\tilde{\nu}) = L_\delta$, it holds that $(\check{\lambda},\tilde{\nu}) \in \boldsymbol{\Lambda}^\delta$ for all $\check{\lambda} \lhd_\Sigma \lambda$. Thanks to Σ being a (scaled) Riesz basis for $L_2(I)$ and $H^1(I)$, together with the $H^1(\Omega)$- and $H^{-1}(\Omega)$-stable splittings of (5.4.5), it holds that

$$\|z + \boldsymbol{c}^\top \Theta_\delta\|^2_X \eqsim \|z\|^2_X + \|\boldsymbol{c}\|^2 \quad (\boldsymbol{c} \in \ell_2(\boldsymbol{\Lambda}_0^{\underline{\delta}} \setminus \boldsymbol{\Lambda}_0^\delta), z \in X^\delta),$$

with the constant in the $\eqsim$-symbol dependent on L_δ only, so when L_δ is uniformly bounded, condition (5.2.11) is satisfied.

5.5 Implementation

A tree-based implementation of the aforementioned adaptive algorithm in C++ can be found at [vVW21b]. In this section, we describe our design choices for a linear complexity implementation.

5.5.1 Trees and linear operators in one axis

In §5.3, we consider an abstract multilevel collection Ψ indexed on $\vee_\Psi$. Endowed with a parent-child relation, $\vee_\Psi$ has a tree-like structure that we call a *mother tree*; see also Figures 5.3 and 5.4.

In our applications, the support of a wavelet ψ_λ is a union of simplices of generation $|\lambda|$. In time, these simplices are subintervals of I found by dyadic refinement. In space, they are elements of $\mathfrak{T}$, the collection of all triangles found by newest vertex bisection. Endowed with the natural parent-child relation, both collections of simplices have a tree structure we call the *domain mother tree*. Every wavelet ψ_λ stores references to the simplices T of generation $|\lambda|$ that make up its support; conversely, every T stores a reference to ψ_λ.

Every mother tree $\vee$ is stored once in memory, and every node $\lambda \in \vee$ stores references to its parents, children, and siblings. We treat the mother tree as infinite by *lazy initialization*, constructing new nodes as they are needed.

Trees We store a tree $\Lambda \subset \vee$ using the parent-child relation, and additionally, at each $\lambda \in \Lambda$ store a reference to the corresponding node in $\vee$. This allows us to compare different trees subject to the same mother tree. This tree-like representation does not allow direct access of arbitrary nodes: in any operation, we traverse Λ from its roots in breadth-first, or level-wise, order.

Tree operations One important operation is the `union` of one tree Λ into another $\check{\Lambda}$. This can be implemented by traversing both trees simultaneously in breadth-first order. The union allows us to easily perform high-level operations, such as vector addition: given two vectors $\boldsymbol{c} \in \ell_2(\Lambda)$, $\boldsymbol{d} \in \ell_2(\check{\Lambda})$ on the same mother tree $\vee$, we use the union to perform $\boldsymbol{c} := \boldsymbol{c} + \boldsymbol{d}$. See Figure 5.6 for an example.

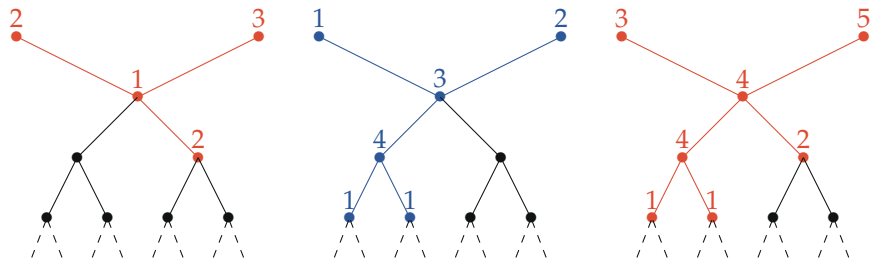


Figure 5.6 Left: $\boldsymbol{c} \in \ell_2(\Lambda)$ for $\Lambda \subset \vee_{\mathfrak{I}}$; Middle: $\boldsymbol{d} \in \ell_2(\check{\Lambda})$ for $\check{\Lambda} \subset \vee_{\mathfrak{I}}$; Right: in-place sum $\boldsymbol{c} := \boldsymbol{c} + \boldsymbol{d}$.

Tree operations in time The routines `eval`, `evalupp`, and `evallow` from §5.3.1 involve various level-wise index sets (represented as arrays of references into their mother trees) like $\check{\Pi}_B = \{\lambda \in \check{\Pi} : |\operatorname{supp} \check{\phi}_\lambda \cap \cup_{\mu \in \Lambda_\ell} S(\mu)| > 0\}$. We constructed such sets efficiently using the domain mother tree; see Algorithm 5.5.

We can apply the linear operators appearing in the routines of §5.3.1 efficiently by again traversing the domain mother tree; for example, Algorithm 5.6 details a matrix-free application of $(A\Phi|_\Pi)(\check{\Phi}|_{\check{\Pi}})$.

Operations in space We can construct a triangulation $\mathcal{T}$ from a vertex tree $N_\mathcal{T}$ in linear complexity. First mark every $\nu \in N_\mathcal{T}$ in its mother tree, then traverse the domain mother tree $\mathfrak{T}$. A triangle T visited in this traversal is in $\mathcal{T}$ exactly when the newest vertex of its children is not marked.

For the preconditioners $\boldsymbol{K}_\mu^\delta$ and $\boldsymbol{K}_\lambda^\delta$ from §5.4.4 we use multigrid. We apply multiplicative V-cycle multigrid, in each cycle applying one pre- and one post Gauss–Seidel smoother with reversed ordering of the unknowns.

To obtain a linear cost algorithm, at level k we restrict smoothing to the vertices of generation k and their godparents, cf. [WZ17]. For $\mathcal{T} \in \mathbb{T}$ consider $W_\mathcal{T}$, the space of continuous piecewise linears w.r.t. $\mathcal{T}$, zero on $\partial\Omega$, equipped with single-scale basis $\Phi_\mathcal{T}$. With $L = L(\mathcal{T}) := \max_{T \in \mathcal{T}} \operatorname{gen}(T)$, define

$$\mathcal{T}_\perp = \mathcal{T}_0 \prec \mathcal{T}_1 \prec \cdots \prec \mathcal{T}_L = \mathcal{T} \subset \mathbb{T}$$

Data: $\ell \in \mathbb{N}$, $\check{\Pi} \subset \check{\Delta}_{\ell-1}$, $\Lambda_\ell \subset \vee_\ell$.
Result: $[\check{\Pi}_A, \check{\Pi}_B]$ where $\check{\Pi}_A = \check{\Pi} \setminus \check{\Pi}_B$,
$\check{\Pi}_B = \{\lambda \in \check{\Pi} : |\operatorname{supp} \check{\phi}_\lambda \cap \cup_{\mu\in\Lambda_\ell} S(\mu)| > 0\}$.
$\check{\Pi}_A := \check{\Pi}_B := \varnothing$;
for $\mu \in \Lambda_\ell$ **do**
  for $T \in \psi_\mu$.`support` **do** `// We have` $S(\psi_\mu) = \operatorname{supp}\psi_\mu$.
    T.`parent.marked` := true;
for $\lambda \in \check{\Pi}$ **do**
  if $\exists T \in \check{\phi}_\lambda$.`support` *with* T.`marked` = *true* **then**
    $\check{\Pi}_B$.insert(λ);
  else $\check{\Pi}_A$.insert(λ);
for $\mu \in \Lambda_\ell$ **do**
  for $T \in \psi_\mu$.`support` **do** T.`parent.marked` := false;

Algorithm 5.5 The construction of $\check{\Pi}_B$.

Data: Index sets $\Pi \subset \Delta_\ell$, $\check{\Pi} \subset \check{\Delta}_\ell$, $\boldsymbol{d} \in \ell_2(\Pi)$, local and linear
$A\colon \operatorname{span}\Phi \to \operatorname{span}\check{\Phi}'$.
Result: $\boldsymbol{e} = (A\Phi|_\Pi)(\check{\Phi}|_{\check{\Pi}})\boldsymbol{d}$
for $\lambda \in \Pi$ **do** ϕ_λ.`data` := $\boldsymbol{d}_\lambda$;
for $\mu \in \check{\Pi}$ **do**
  $\boldsymbol{e}_\lambda := 0$;
  for $T \in \check{\phi}_\mu$.`support` **do**
    for $\phi_\lambda \in T$.`functions`(Δ_ℓ) **do**
      `//` $\{\phi_\lambda : \lambda \in \Delta_\ell, |\operatorname{supp}\phi_\lambda \cap T| > 0\}$
      $\boldsymbol{e}_\lambda := \boldsymbol{e}_\lambda + A(\phi_\lambda)(\check{\phi}_\mu|_T)\cdot \phi_\lambda$.`data`;
for $\lambda \in \Pi$ **do** ϕ_λ.`data` := 0;

Algorithm 5.6 The computation of $\boldsymbol{e} = (A\Phi|_\Pi)(\check{\Phi}|_{\check{\Pi}})\boldsymbol{d}$.

where $\mathcal{T}_{k-1}$ is constructed from $\mathcal{T}_k$ by removing all vertices $\nu \in N_{\mathcal{T}_k}$ for which $\operatorname{gen}(\nu) = k$. For $1 \le k \le L$, let M_k be the set of new vertices and their godparents, i.e., $M_k := \bigcup_{\nu\in N_{\mathcal{T}_k}\setminus N_{\mathcal{T}_{k-1}}} \{\nu\} \cup \operatorname{gp}(\nu)$, and let $M_{k,0} := M_k \setminus \partial\Omega$ be the set of interior vertices. We consider the multilevel decomposition, cf. [WZ17],

$$W_{\mathcal{T}_L} = W_{\mathcal{T}_0} + \sum_{k=1}^{L} \sum_{\nu\in M_{k,0}} \operatorname{span}\phi_{k,\nu}, \quad \text{where} \quad \phi_{k,\nu} := \phi_{\mathcal{T}_k,\nu}. \tag{5.5.1}$$

For $1 \le k \le L$, let $\boldsymbol{P}_k$ be the *prolongation matrix*, i.e., the matrix representation of the embedding $W_{\mathcal{T}_{k-1}} \to W_{\mathcal{T}_k}$, and enumerate the vertices $M_{k,0}$ as $(\nu_k^i)_{i=1}^{n_k}$. Algorithm 5.7 details a (non-recursive) implementation of a single multiplicative V-cycle for the multilevel decomposition (5.5.1) using Gauss–Seidel smoothing. We assume the availability of an efficient coarse-grid solver;

in our application, a direct solve suffices. For linear complexity, we use in-place vector updates restricted to non-zeros.

Note that this multigrid method is given in terms of the single-scale basis $\Phi_{\mathcal{T}}$; it can be transformed to the hierarchical basis $\Psi_{\mathcal{T}}$ using (5.4.4). Multiple V-cycles are done by setting $u_0 := 0$ and iterating $u_k := \mathrm{MG}(A, f - Au_{k-1})$.

Data: Some $f \in W'_{\mathcal{T}}$ and a linear operator $A\colon W_{\mathcal{T}} \to W'_{\mathcal{T}}$.
Result: $u = \boldsymbol{u}^\top \Phi_{\mathcal{T}} \in W_{\mathcal{T}}$, the result of a single V-cycle applied to f.
$\boldsymbol{r} := f(\Phi_{\mathcal{T}})$;
for $L \geq k \geq 1$ **do**
  for $\nu = \nu_k^1, \dots, \nu_k^{n_k}$ **do**
    $r_{k,\nu} := \boldsymbol{r}_\nu$;
    $e_{k,\nu} := r_{k,\nu} / (A\phi_{k,\nu})(\phi_{k,\nu})$;
    $\boldsymbol{r} := \boldsymbol{r} - e_{k,\nu}(A\phi_{k,\nu})(\Phi_{\mathcal{T}_k})$;
  $\boldsymbol{r} := \boldsymbol{P}_k^\top \boldsymbol{r}$;
Solve $(A\Phi_{\mathcal{T}_0})(\Phi_{\mathcal{T}_0})\boldsymbol{u} = \boldsymbol{r}$;

for $1 \leq k \leq L$ **do**
  $\boldsymbol{u} := \boldsymbol{P}_k \boldsymbol{u}$;
  for $\nu = \nu_k^{n_k}, \dots, \nu_k^1$ **do**
    $\boldsymbol{u}_\nu := \boldsymbol{u}_\nu + e_{k,\nu}$;
    $\boldsymbol{u}_\nu := \boldsymbol{u}_\nu + (r_{k,\nu} - (A\phi_{k,\nu})(\boldsymbol{u}^\top \Phi_{\mathcal{T}_k}))/(A\phi_{k,\nu})(\phi_{k,\nu})$;

Algorithm 5.7 Single multiplicative V-cycle multigrid $\mathrm{MG}(A, f)$.

5.5.2 Double-trees and tensor-product operators

For every node in a double-tree $\boldsymbol{\Lambda} \subset \vee^0 \times \vee^1$, we store a reference to the underlying pair of nodes in their mother trees. This allows growing double-trees intuitively, and allows comparing different double-trees over the same pair of mother trees. C++ templates allow us to re-use much of the tree code without runtime performance loss.

In §5.3.2 we saw how to apply a tensor-product operator using the double-trees $\boldsymbol{\Sigma}$ and $\boldsymbol{\Theta}$. Construction of $\boldsymbol{\Sigma}$ is illustrated in Algorithm 5.8, and evaluation of the operator then reduces to the four simple steps of Algorithm 5.9.

Memory optimizations As the memory consumption of a double-tree is significant, at around 280 bytes per node, we want to have as few double-trees in memory as possible. By storing the nodes of $\boldsymbol{\Lambda}$ in a persistent container, every node is uniquely identified with its index in the container. This induces a bijection $\mathbb{R}^{\#\boldsymbol{\Lambda}} \leftrightarrow \ell_2(\boldsymbol{\Lambda})$ and allows us to overlay multiple vectors on the same underlying double-tree in a memory-friendly way.

The $\boldsymbol{\Sigma}$ generated by Algorithm 5.8 for the application of a tensor-product operator can play the role of $\boldsymbol{\Theta}$ necessary for the application of its transpose

Data: $\check{\mathbf{\Lambda}} \subset \check{\vee}^0 \times \check{\vee}^1$, $\mathbf{\Lambda} \subset \vee^0 \times \vee^1$
Result: $\mathbf{\Sigma}$ for application of Theorem 5.3.7 with $\check{\mathbf{\Lambda}}$ and $\mathbf{\Lambda}$.
$\mathbf{\Sigma} := P_0\mathbf{\Lambda} \times \{\nu \in P_1\check{\mathbf{\Lambda}} : |\nu| = 0\}$;
for $\lambda \in \mathbf{\Sigma}$.`project`$(0)$ **do**
  for $T \in \phi_\lambda$.`support` **do**
    for $\mu \in T$.`functions`$(\check{\vee}^0_{|\lambda|})$ **do**
      $\mathbf{\Sigma}$.`fiber`$(1,\lambda)$.`union`$(\check{\mathbf{\Lambda}}$.`fiber`$(1,\mu))$;

Algorithm 5.8 Function `GenerateSigma`$(\check{\mathbf{\Lambda}}, \mathbf{\Lambda})$.

Data: $\mathbf{\Lambda} \subset \vee^0 \times \vee^1$, $\check{\mathbf{\Lambda}} \subset \check{\vee}^0 \times \check{\vee}^1$, $\boldsymbol{c} \in \ell_2(\mathbf{\Lambda})$, $\boldsymbol{d} \in \ell_2(\check{\mathbf{\Lambda}})$.
$\mathbf{\Sigma}$:= `GenerateSigma`$(\check{\mathbf{\Lambda}}, \mathbf{\Lambda})$;
$\mathbf{\Theta}$:= `GenerateTheta`$(\check{\mathbf{\Lambda}}, \mathbf{\Lambda})$;
$\boldsymbol{s} := \mathbf{0} \in \ell_2(\mathbf{\Sigma})$;
$\boldsymbol{t} := \mathbf{0} \in \ell_2(\mathbf{\Theta})$;
$\boldsymbol{l} := \mathbf{0} \in \ell_2(\check{\mathbf{\Lambda}})$;
for $\lambda \in \boldsymbol{s}$.`project`$(0)$ **do** `eval`$(A_1)(\boldsymbol{s}$.`fiber`$(1,\lambda), \boldsymbol{c}$.`fiber`$(1,\lambda))$;
for $\mu \in \boldsymbol{l}$.`project`(1) **do** `evallow`$(A_0)(\boldsymbol{l}$.`fiber`$(0,\mu), \boldsymbol{s}$.`fiber`$(0,\mu))$;
for $\mu \in \boldsymbol{t}$.`project`(1) **do** `evalupp`$(A_0)(\boldsymbol{t}$.`fiber`$(0,\mu), \boldsymbol{c}$.`fiber`$(0,\mu))$;
for $\lambda \in \boldsymbol{d}$.`project`$(0)$ **do** `eval`$(A_1)(\boldsymbol{d}$.`fiber`$(1,\lambda), \boldsymbol{t}$.`fiber`$(1,\lambda))$;
$\boldsymbol{d} := \boldsymbol{d} + \boldsymbol{l}$;

Algorithm 5.9 Algorithm to evaluate $\boldsymbol{d} = R_{\check{\mathbf{\Lambda}}}(A_0 \otimes A_1) I_{\mathbf{\Lambda}} \boldsymbol{c}$.

operator (and vice versa). This allows tensor-product operators and their transposes to share the double-trees $\mathbf{\Sigma}$ and $\mathbf{\Theta}$.

With these insights, our implementation of the heat equation has at most 5 different double-trees in memory.

5.5.3 The adaptive loop

In the refine step of the adaptive loop, we first mark a set J of nodes in $\mathbf{\Lambda}^{\underline{\delta}} \setminus \mathbf{\Lambda}^\delta$ using Dörfler marking (possible in linear complexity; cf. [PP20]). We then refine $\mathbf{\Lambda}^\delta$ to the smallest double-tree containing J:

1. mark all nodes in $\mathbf{\Lambda}^{\underline{\delta}}$ that are also present in $\mathbf{\Lambda}^\delta$ ((ii) in Fig. 5.7);
2. from every node in J, traverse top-down in level-wise order, until hitting an already marked node. Mark all nodes along the way ((iii)–(iv));
3. `union` the marked nodes of $\mathbf{\Lambda}^{\underline{\delta}}$ into $\mathbf{\Lambda}^\delta$ ((v) in Fig. 5.7).

As $\#\mathbf{\Lambda}^{\underline{\delta}} \lesssim \#\mathbf{\Lambda}^\delta$ and we visit every node of $\mathbf{\Lambda}^{\underline{\delta}}$ at most twice, the traversal is linear in $\#\mathbf{\Lambda}^\delta$. See also Figure 5.7.

Figure 5.7 Refinement of a double-tree with underlying *unary* mother trees. Left to right: (i) $\boldsymbol{\Lambda}^\delta$; (ii) $\boldsymbol{\Lambda}^{\underline{\delta}}$ with nodes in $\boldsymbol{\Lambda}^{\underline{\delta}} \setminus \boldsymbol{\Lambda}^\delta$ in white; (iii) nodes in J marked in red; (iv) nodes marked in the top-down traversal; (v) refined $\boldsymbol{\Lambda}^\delta$.

5.6 Numerical experiments

We consider the heat equation (5.4.1), and assess our implementation of the adaptive Algorithm 5.1 for its numerical solution. Complementing the convergence results gathered in [SvVW21, §7], here we provide results on the practical performance of the adaptive loop. Results were gathered on a multicore 2.2 GHz machine, provided by the Dutch national e-infrastructure with the support of SURF Cooperative.

5.6.1 The adaptive loop

We summarize the main results from [SvVW21, §7]. We run Algorithm 5.1 with $\theta = \frac{1}{2}$ and $\xi = \frac{1}{2}$. We consider four problems.

In the *smooth problem*, we select $\Omega := [0,1]^2$ and prescribe the solution

$$u(t,x,y) := (1+t^2)x(1-x)y(1-y).$$

In the *moving peak* problem, we select $\Omega := [0,1]^2$ with prescribed solution

$$u(t,x,y) := x(1-x)y(1-y)\exp(-100[(x-t)^2+(y-t)^2]);$$

u is essentially zero outside a small strip along the diagonal $(0,0,0)$–$(1,1,1)$.

In the *cylinder* problem, we select $\Omega := [-1,1]^2 \setminus [-1,0]^2$ with data

$$u_0 \equiv 0, \quad \text{and} \quad g(t,x,y) := t \cdot \mathbb{1}_{\{x^2+y^2<1/4\}}.$$

The solution has singularities in the re-entrant corner and along the wall of the cylinder $\{(t,x,y) : x^2+y^2 = 1/4\}$.

In the *singular* problem, we select $\Omega := [-1,1]^2 \setminus [-1,0]^2$ with data $u_0 \equiv 1$ and $g \equiv 0$; the solution then has singularities along $\{0\} \times \partial\Omega$ and $I \times \{(0,0)\}$. This problem especially is interesting, as uniform refinement makes almost no progress. The adaptive algorithm performs very well. Looking at Figure 5.8, we see strong adaptivity towards the singularities, and observe basis functions with a barycenter at $t = 2^{-14} \approx 10^{-4}$.

Convergence To estimate the error $\|u - \hat{u}^\delta\|_X$, we measure the residual error estimator $\|\boldsymbol{r}^\delta(\hat{u}^\delta)\|$ from (5.2.12); see also Lemma 5.2.2. In the left pane of Figure 5.9, for the first three problems, we observe a convergence rate of $1/2$,

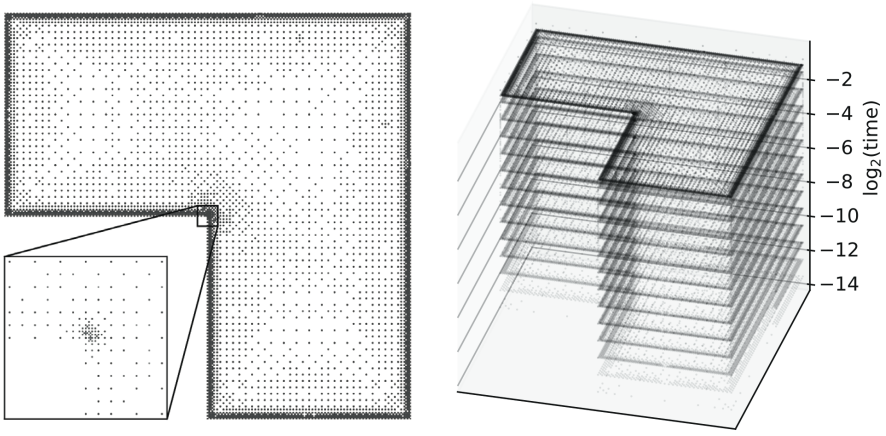


Figure 5.8 Barycenters of supports of basis functions $\sigma_\lambda \otimes \phi_\nu$ spanning X^δ of dimension 81 074 for the *singular* problem. Left: a top-down view, with a 10× zoom to the origin; right: centers in spacetime, logarithmic in time.

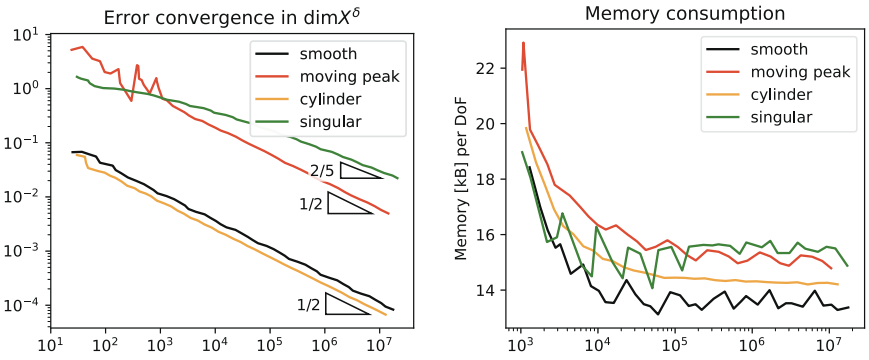


Figure 5.9 Error convergence and peak memory usage of the adaptive loop for the four problems of §5.6.1.

which is the best that can be expected from our family of trial spaces $(X^\delta)_{\delta\in\Delta}$. For the singular problem, the reduced rate 0.4 is found; it is unknown if a better rate can be expected.

Memory The right pane of Figure 5.9 shows peak memory consumption after every iteration of the adaptive algorithm. We see that it is linear in dim X^δ, stabilizing to around 15kB per degree of freedom. This is relatively high due to our implementation based on double-trees. In fact, the double-trees together make up around 85% of the total memory. For the singular problem, the largest double-tree $\mathbf{\Lambda}^\delta_Y$ occupies around 40% of the total memory.

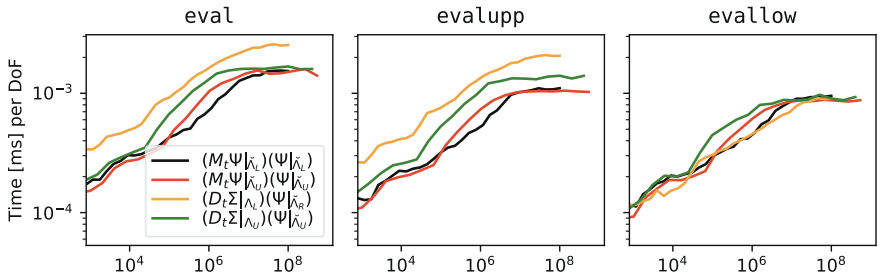


Figure 5.10 Time (in ms) per DoF of bilinear form evaluations in time.

5.6.2 Linearity of operations

The majority of our runtime is spent in the application of bilinear forms. In this section, we measure the application times to assert their linear complexity.

In time We select three sequences $\{\Lambda_U\}$, $\{\Lambda_L\}$, $\{\Lambda_R\}$ of trees in $\vee_\Sigma$, one uniformly refined and two graded towards the left and right respectively. For each such tree $\Lambda \subset \vee_\Sigma$, we define a corresponding tree $\check{\Lambda} := \{\mu \in \vee_\Xi : \exists \lambda \in \Lambda, |\lambda| = |\mu|, |\operatorname{supp} \xi_\mu \cap \operatorname{supp} \sigma_\lambda| > 0\} \subset \vee_\Xi$.

We select the bilinear forms M_t and D_t from (5.4.2), and run the algorithms from §5.3.1. We see in Figure 5.10 that the runtime per degree of freedom stabilizes to 10^{-3} ms, essentially independent of the bilinear form and the trees. We suspect the increase until 10^7 degrees of freedom has to do with cache locality.

In space On the L-shaped domain $\Omega := [-1,1]^2 \setminus [-1,0]^2$, we select two sequences of hierarchical basis trees, one uniformly refined and the other refined by a standard adaptive loop on $-\Delta u = 1$, $u|_{\partial\Omega} = 0$.

For a hierarchical basis tree $\Psi_\mathcal{T} = \{\psi_\nu : \nu \in N_{\mathcal{T},0}\}$, we denote the stiffness matrix $\langle \nabla\Psi_\mathcal{T}, \nabla\Psi_\mathcal{T}\rangle_{L_2(\Omega)}$ as $\boldsymbol{A}_\mathcal{T}$. We measure the runtime of the conversion from vertex tree $N_\mathcal{T}$ to triangulation $\mathcal{T}$ (cf. §5.5.1), the application time of $\boldsymbol{A}_\mathcal{T}$ through (5.4.4), and that of multigrid on $\boldsymbol{A}_\mathcal{T}$ (with 1 and 3 V-cycles) through Algorithm 5.7. Figure 5.11 confirms that the relative runtime of every operation is essentially independent of the refinement strategy. Interesting is again the increase until 10^5 degrees of freedom.

In space-time Solving (5.2.9) using PCG requires the application of the four linear operators $E_Y^\delta B E_X^\delta$, ${E_X^\delta}' \gamma_0' \gamma_0 E_X^\delta$, K_X^δ, and K_Y^δ. For the first two, Corollary 5.3.8 asserts that their application time is of linear complexity, while for the preconditioners K_X^δ and K_Y^δ, this follows from block-diagonality of their matrix representation.

We run the adaptive algorithm on the four problems of §5.6.1. Figure 5.12 shows that the application time of the aforementioned operators is essentially

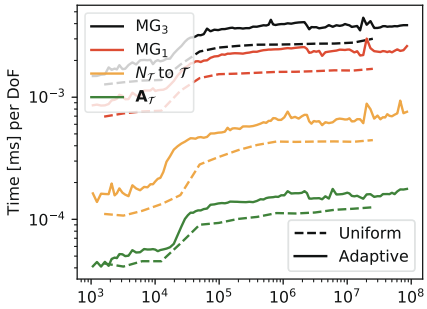


Figure 5.11 Time (in ms) per DoF of important operations in space, for uniform and adaptive refinements.

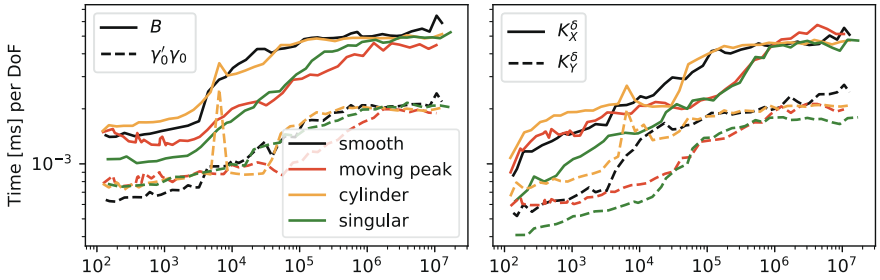


Figure 5.12 Time (in ms) per DoF of the four bilinear forms applied in the solve step of the adaptive algorithm.

independent of the problem, even though the underlying double-trees differ vastly. Note the increase in relative runtime until 10^6 degrees of freedom.

Figure 5.13 shows the runtimes of the solve, estimate, mark and refine steps of the adaptive loop. We confirm that each step is of linear complexity, and that the total runtime is governed by the solve and estimate steps.

5.6.3 Shared-memory parallelism

Most of our execution time is spent applying the linear operators from Figure 5.12. We can obtain a significant speedup with multithreading. In Algorithm 5.9, all fibers inside each of the four for-loops are disjoint, and we can easily parallelize each loop using OpenMP.

We run the parallel code on the smooth and singular problems. The right pane of Figure 5.14 shows decent parallel performance for the singular problem, with $10\times$ speedup at 16 cores. The left pane however reveals a load balancing issue: as u is smooth, the two fibers $(\mathbf{\Lambda}_0^\delta)_{1,\lambda}$ with $|\lambda| = 0$ contain the majority of the degrees of freedom. This results in poor parallel efficiency for the first and fourth loop in Algorithm 5.9.

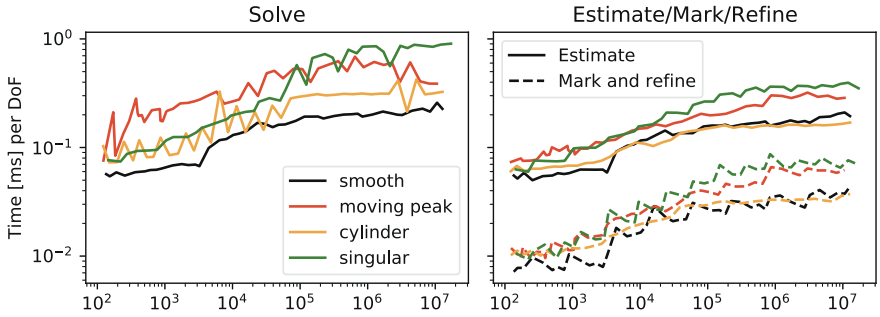


Figure 5.13 Time (in ms) per DoF of the steps in the adaptive loop.

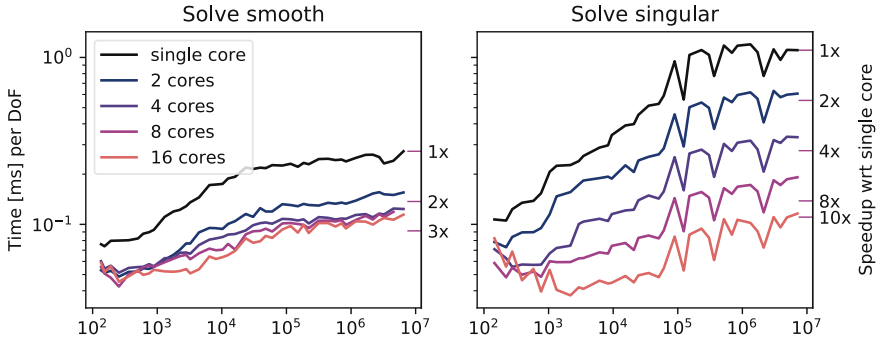


Figure 5.14 Speedup and time (in ms) per DoF of the solve step in the adaptive loop, for different number of parallel processors.

5.7 Conclusion

We discussed an implementation of a space-time adaptive solver for parabolic evolution equations where every step is of linear complexity.

We constructed a family of trial spaces spanned by tensor-products of wavelets in time and hierarchical basis functions in space. The resulting adaptive loop is able to resolve singularities locally in space and time, and we proved its r-linear convergence.

After imposing a *double-tree* constraint on the index set of the trial spaces, we devised an abstract algorithm that is able to apply the system matrices in linear complexity. We achieve this complexity in practice by a *tree-based* implementation. The numerical results show high performance of the adaptive loop as a whole.

5.A Proofs of Theorems in §5.3

Theorem 5.3.3. *A call of* `eval` *yields the output as specified, at the cost of* $\mathcal{O}(\#\check{\Pi}\check{\mathrm{I}} + \#\check{\Lambda} + \#\Pi + \#\Lambda)$ *operations.*

Proof. By locality of the collections $\check{\Phi}$ and $\check{\Psi}$, and sparsity of the matrices $\check{\mathfrak{p}}_\ell$ and $\check{\mathfrak{q}}_\ell$, we see that $\#\underline{\check{\Pi}\check{\mathrm{I}}} \lesssim \#\check{\Pi}\check{\mathrm{I}}_B + \#\check{\Lambda}_\ell \lesssim \#\Lambda_\ell + \#\check{\Lambda}_\ell$. So after sufficiently many recursive calls, the current set $\check{\Pi}\check{\mathrm{I}} \cup \check{\Lambda}$ will be empty. For use later, we note that similarly $\#\underline{\Pi\Pi} \lesssim \#\Pi\mathrm{I}_B + \#\Lambda_\ell \lesssim \#\check{\Lambda}_\ell + \#\check{\Pi}\check{\mathrm{I}}_B + \#\Lambda_\ell \lesssim \#\Lambda_\ell + \#\check{\Lambda}_\ell$.

For $\check{\Pi}\check{\mathrm{I}} \cup \check{\Lambda} = \varnothing$, the call produces nothing, which is correct.

Now let $\check{\Pi}\check{\mathrm{I}} \cup \check{\Lambda} \neq \varnothing$. From Λ being an ℓ-tree, the definitions of $S(\cdot)$ and $\check{\Pi}\check{\mathrm{I}}_A$, and the locality of A, one has

$$\boldsymbol{e}|_{\check{\Pi}\check{\mathrm{I}}_A} = (Au)(\check{\Phi}|_{\check{\Pi}\check{\mathrm{I}}_A}) = (A(\boldsymbol{d}^\top \Phi|_\Pi))(\check{\Phi}|_{\check{\Pi}\check{\mathrm{I}}_A}).$$

By choice of $\underline{\Pi}$ we have

$$\underline{u} := \underline{\boldsymbol{d}}^\top \Phi|_{\underline{\Pi}} + \boldsymbol{c}|^\top_{\Lambda_{\ell+1\uparrow}} \Psi|_{\Lambda_{\ell+1\uparrow}} = (\boldsymbol{d}|_{\Pi_B})^\top \Phi|_{\Pi_B} + \boldsymbol{c}^\top \Psi|_\Lambda = u - (\boldsymbol{d}|_{\Pi_A})^\top \Phi|_{\Pi_A}.$$

The recursive call yields $\underline{\boldsymbol{e}} = (A\underline{u})(\check{\Phi}|_{\underline{\check{\Pi}\check{\mathrm{I}}}})$, and $\underline{\boldsymbol{f}} = (A\underline{u})(\check{\Psi}|_{\check{\Lambda}_{\ell+1\uparrow}})$. From $\check{\Lambda}$ being an ℓ-tree, the definitions of $\check{S}(\cdot)$ and Π_A, and the locality of A, we have

$$(Au)(\check{\Psi}|_{\check{\Lambda}_{\ell\uparrow}}) = (A\underline{u})(\check{\Psi}|_{\check{\Lambda}_{\ell\uparrow}}),$$

and so in particular $\boldsymbol{f}|_{\Lambda_{\ell+1,\uparrow}} = \underline{\boldsymbol{f}}$.

The definition of $\underline{\check{\Pi}\check{\mathrm{I}}}$ shows that

$$\check{\Phi}|_{\check{\Pi}\check{\mathrm{I}}_B} = (\check{\mathfrak{p}}_\ell^\top \check{\Phi}|_{\underline{\check{\Pi}\check{\mathrm{I}}}})|_{\check{\Pi}\check{\mathrm{I}}_B}, \quad \check{\Psi}|_{\check{\Lambda}_\ell} = (\check{\mathfrak{q}}_\ell^\top \check{\Phi}|_{\underline{\check{\Pi}\check{\mathrm{I}}}})|_{\check{\Lambda}_\ell}.$$

We conclude that

$$\boldsymbol{f}|_{\check{\Lambda}_\ell} = (Au)(\check{\Psi}|_{\check{\Lambda}_\ell}) = (A\underline{u})(\check{\Psi}|_{\check{\Lambda}_\ell}) = \left(\check{\mathfrak{q}}_\ell^\top \underline{\boldsymbol{e}}\right)|_{\check{\Lambda}_\ell},$$

and from $|\operatorname{supp}\phi_\lambda \cap \operatorname{supp}\check{\phi}_\mu| = 0$ for $(\lambda, \mu) \in \Pi_A \times \check{\Pi}\check{\mathrm{I}}_B$, that

$$\boldsymbol{e}|_{\check{\Pi}\check{\mathrm{I}}_B} = (Au)(\check{\Phi}|_{\check{\Pi}\check{\mathrm{I}}_B}) = (A\underline{u})(\check{\Phi}|_{\check{\Pi}\check{\mathrm{I}}_B}) = \left(\check{\mathfrak{p}}_\ell^\top \underline{\boldsymbol{e}}\right)|_{\check{\Pi}\check{\mathrm{I}}_B}.$$

From the assumptions on the collections Φ, $\check{\Phi}$, $\check{\Psi}$, and Ψ, and their consequences on the sparsity of the matrices $\mathfrak{p}_\ell$, $\check{\mathfrak{p}}_\ell$, $\mathfrak{q}_\ell$, and $\check{\mathfrak{q}}_\ell$, one infers that the total cost of the evaluations of the statements in `eval` is $\mathcal{O}(\#\check{\Pi}\check{\mathrm{I}} + \#\check{\Lambda}_\ell + \#\Pi + \#\Lambda_\ell)$ plus the cost of the recursive call. Using $\#\underline{\check{\Pi}\check{\mathrm{I}}} + \#\underline{\Pi\Pi} \lesssim \#\check{\Lambda}_\ell + \#\Lambda_\ell$ and induction, we conclude the second statement of the theorem. □

Theorem 5.3.4. *A call of* `evalupp` *yields the output as specified, at the cost of* $\mathcal{O}(\#\check{\Pi}\check{\mathrm{I}} + \#\check{\Lambda} + \#\Pi + \#\Lambda)$ *operations.*

Proof. By locality of the collections $\check{\Phi}$ and $\check{\Psi}$, and sparsity of the matrices $\check{\mathfrak{p}}_\ell$ and $\check{\mathfrak{q}}_\ell$, we see that $\#\underline{\check{\Pi}} \lesssim \#\check{\Pi}_B + \#\check{\Lambda}_\ell \lesssim \#\Lambda_\ell + \#\check{\Lambda}_\ell$. So after sufficiently many recursive calls, the current set $\check{\Pi} \cup \check{\Lambda}$ will be empty. Notice that $\#\underline{\Pi} \lesssim \#\Lambda_\ell$.

For $\check{\Pi} \cup \check{\Lambda} = \varnothing$, the call produces nothing, which is correct.

Now let $\check{\Pi} \cup \check{\Lambda} \neq \varnothing$. From Λ being an ℓ-tree, the definitions of $S(\cdot)$ and $\check{\Pi}_A$, and the locality of A, one has

$$\boldsymbol{e}|_{\check{\Pi}_A} = (Au)(\check{\Phi}|_{\check{\Pi}_A}) = (A(\boldsymbol{d}^\top \Phi|_\Pi))(\check{\Phi}|_{\check{\Pi}_A}).$$

By definition of $\underline{\Pi}$ we have

$$\underline{u} := \underline{\boldsymbol{d}}^\top \Phi|_{\underline{\Pi}} + \boldsymbol{c}|_{\Lambda_{\ell+1\uparrow}}^\top \Psi|_{\Lambda_{\ell+1\uparrow}} = \boldsymbol{c}^\top \Psi|_\Lambda = u - \boldsymbol{d}^\top \Phi|_\Pi.$$

The recursive call yields $\underline{\boldsymbol{e}} = (A\underline{u})(\check{\Phi}|_{\underline{\check{\Pi}}})$ and $\underline{\boldsymbol{f}} = \boldsymbol{U}_{\check{\Lambda}_{\ell+1\uparrow} \times \Lambda_{\ell+1\uparrow}} \boldsymbol{c}|_{\Lambda_{\ell+1\uparrow}} = \boldsymbol{f}|_{\check{\Lambda}_{\ell+1\uparrow}}$.

The definition of $\underline{\check{\Pi}}$ shows that

$$\check{\Phi}|_{\check{\Pi}_B} = (\check{\mathfrak{p}}_\ell^\top \check{\Phi}|_{\underline{\check{\Pi}}})|_{\check{\Pi}_B}, \quad \check{\Psi}|_{\check{\Lambda}_\ell} = (\check{\mathfrak{q}}_\ell^\top \check{\Phi}|_{\underline{\check{\Pi}}})|_{\check{\Lambda}_\ell}.$$

We conclude that

$$\boldsymbol{f}|_{\check{\Lambda}_\ell} = (A(\boldsymbol{c}^\top \Psi|_\Lambda))(\check{\Psi}|_{\check{\Lambda}_\ell}) = (A\underline{u})(\check{\Psi}|_{\check{\Lambda}_\ell}) = \left(\check{\mathfrak{q}}_\ell^\top \underline{\boldsymbol{e}}\right)|_{\check{\Lambda}_\ell},$$

and

$$\begin{aligned}\boldsymbol{e}|_{\check{\Pi}_B} = (Au)(\check{\Phi}|_{\check{\Pi}_B}) &= (A\underline{u})(\check{\Phi}|_{\check{\Pi}_B}) + (A(\boldsymbol{d}^\top \Phi|_\Pi))(\check{\Phi}|_{\check{\Pi}_B}) \\ &= \left(\mathfrak{p}_\ell^\top \underline{\boldsymbol{e}}\right)|_{\check{\Pi}_B} + (A(\boldsymbol{d}^\top \Phi|_\Pi))(\check{\Phi}|_{\check{\Pi}_B}).\end{aligned}$$

From the assumptions on the collections Φ, $\check{\Phi}$, $\check{\Psi}$, and Ψ, and their consequences on the sparsity of the matrices $\mathfrak{p}_\ell$, $\check{\mathfrak{p}}_\ell$, $\mathfrak{q}_\ell$, and $\check{\mathfrak{q}}_\ell$, one infers that the total cost of the evaluations of the statements in `eval` is $\mathcal{O}(\#\check{\Pi} + \#\check{\Lambda}_\ell + \#\Pi + \#\Lambda_\ell)$ plus the cost of the recursive call. Using $\#\underline{\check{\Pi}} + \#\underline{\Pi} \lesssim \#\check{\Lambda}_\ell + \#\Lambda_\ell$ and induction, we conclude the second statement of the theorem. □

Theorem 5.3.5. *A call of* `evallow` *yields the output as specified, at the cost of* $\mathcal{O}(\#\check{\Lambda} + \#\Pi + \#\Lambda)$ *operations.*

Proof. Notice that $\#\underline{\Pi} \lesssim \#\Lambda_\ell + \#\Pi_B \lesssim \#\Lambda_\ell + \#\check{\Lambda}_\ell$.

For $\check{\Pi} \cup \check{\Lambda} = \varnothing$, the call produces nothing, which is correct.

Now let $\check{\Pi} \cup \check{\Lambda} \neq \varnothing$. The definitions of $\underline{\check{\Pi}}$ and $\underline{\Pi}_B$ show that

$$\begin{aligned}\boldsymbol{f}|_{\check{\Lambda}_\ell} &= (A\Phi|_\Pi)(\check{\Psi}|_{\check{\Lambda}_\ell})\boldsymbol{d} = (A\Phi|_\Pi)(\check{\Psi}|_{\check{\Lambda}_\ell})\boldsymbol{d}|_{\Pi_B} \\ &= \left(\check{\mathfrak{q}}_\ell^\top (A\Phi|_{\underline{\Pi}_B})(\check{\Phi}|_{\underline{\check{\Pi}}})\mathfrak{p}_\ell \boldsymbol{d}|_{\Pi_B}\right)|_{\check{\Lambda}_\ell} = (\check{\mathfrak{q}}_\ell^\top \underline{\boldsymbol{e}})|_{\check{\Lambda}_\ell}.\end{aligned}$$

From $\check{\Lambda}$ being an ℓ-tree, the definitions of $\check{S}(\cdot)$ and Π_B, and the locality of a, and for the third equality, the definition of $\underline{\Pi}$, one has

$$\begin{aligned} f|_{\check{\Lambda}_{\ell+1\uparrow}} &= a(\check{\Psi}|_{\check{\Lambda}_{\ell+1\uparrow}}, \Phi|_{\Pi})\boldsymbol{d} + \boldsymbol{L}|_{\check{\Lambda}_{\ell+1\uparrow}\times\Lambda_\ell}\boldsymbol{c}|_{\Lambda_\ell} + \boldsymbol{L}|_{\check{\Lambda}_{\ell+1\uparrow}\times\Lambda_{\ell+1\uparrow}}\boldsymbol{c}|_{\Lambda_{\ell+1\uparrow}} \\ &= (A\Phi|_{\Pi})(\check{\Psi}|_{\check{\Lambda}_{\ell+1\uparrow}})\boldsymbol{d}|_{\Pi_B} + (A\Psi|_{\Lambda_\ell})(\check{\Psi}|_{\check{\Lambda}_{\ell+1\uparrow}})\boldsymbol{c}|_{\Lambda_\ell} + \boldsymbol{L}|_{\check{\Lambda}_{\ell+1\uparrow}\times\Lambda_{\ell+1\uparrow}}\boldsymbol{c}|_{\Lambda_{\ell+1\uparrow}} \\ &= (A\Phi|_{\underline{\Pi}})(\check{\Psi}|_{\check{\Lambda}_{\ell+1\uparrow}})\underline{\boldsymbol{d}} + \boldsymbol{L}|_{\check{\Lambda}_{\ell+1\uparrow}\times\Lambda_{\ell+1\uparrow}}\boldsymbol{c}|_{\Lambda_{\ell+1\uparrow}} \\ &= \texttt{evallow}(A)(\ell+1, \check{\Lambda}_{\ell+1\uparrow}, \underline{\Pi}, \Lambda_{\ell+1\uparrow}, \underline{\boldsymbol{d}}, \boldsymbol{c}|_{\Lambda_{\ell+1\uparrow}}) \end{aligned}$$

by induction.

From the assumptions on the collections Φ, $\check{\Psi}$, and Ψ, and their consequences on the sparsity of the matrices $\mathfrak{p}_\ell$, $\mathfrak{q}_\ell$, and $\check{\mathfrak{q}}_\ell$, one easily infers that the total cost of the evaluations of the statements in `evallow` is $\mathcal{O}(\#\check{\Lambda}_\ell + \#\Pi + \#\Lambda_\ell)$ plus the cost of the recursive call. Using $\#\underline{\Pi} \lesssim \#\check{\Lambda}_\ell + \#\Lambda_\ell$ and induction, we conclude the second statement of the theorem. □

Theorem 5.3.7. *Let $\check{\boldsymbol{\Lambda}} \subset \check{\vee}^0 \times \check{\vee}^1$, $\boldsymbol{\Lambda} \subset \vee^0 \times \vee^1$ be finite double-trees. Then*

$$\boldsymbol{\Sigma} := \bigcup_{\lambda\in P_0\boldsymbol{\Lambda}} \Bigg(\{\lambda\} \times \bigcup_{\left\{\mu\in P_0\check{\boldsymbol{\Lambda}}:|\mu|=|\lambda|+1,\, |\check{S}^0(\mu)\cap S^0(\lambda)|>0\right\}} \check{\boldsymbol{\Lambda}}_{1,\mu} \Bigg),$$

$$\boldsymbol{\Theta} := \bigcup_{\lambda\in P_1\boldsymbol{\Lambda}} \Big(\{\mu \in P_0\check{\boldsymbol{\Lambda}} : \exists\gamma \in \boldsymbol{\Lambda}_{0,\lambda} \text{ s.t. } |\gamma| = |\mu|,\, |\check{S}^0(\mu)\cap S^0(\gamma)| > 0\} \times \{\lambda\} \Big),$$

are double-trees with $\#\boldsymbol{\Sigma} \lesssim \#\check{\boldsymbol{\Lambda}}$ and $\#\boldsymbol{\Theta} \lesssim \#\boldsymbol{\Lambda}$, and

$$\begin{aligned} R_{\check{\boldsymbol{\Lambda}}}(\boldsymbol{A}_0 \otimes \boldsymbol{A}_1) I_{\boldsymbol{\Lambda}} = {} & R_{\check{\boldsymbol{\Lambda}}}(\boldsymbol{L}_0 \otimes \mathrm{Id}) I_{\boldsymbol{\Sigma}} R_{\boldsymbol{\Sigma}} (\mathrm{Id} \otimes \boldsymbol{A}_1) I_{\boldsymbol{\Lambda}} + \\ & R_{\check{\boldsymbol{\Lambda}}}(\mathrm{Id} \otimes \boldsymbol{A}_1) I_{\boldsymbol{\Theta}} R_{\boldsymbol{\Theta}} (\boldsymbol{U}_0 \otimes \mathrm{Id}) I_{\boldsymbol{\Lambda}}. \end{aligned}$$

Proof. We write

$$\begin{aligned} R_{\check{\boldsymbol{\Lambda}}}(\boldsymbol{A}_0 \otimes \boldsymbol{A}_1) I_{\boldsymbol{\Lambda}} = {} & R_{\check{\boldsymbol{\Lambda}}}((\boldsymbol{L}_0 + \boldsymbol{U}_0) \otimes \boldsymbol{A}_1) I_{\boldsymbol{\Lambda}} \\ = {} & R_{\check{\boldsymbol{\Lambda}}}(\boldsymbol{L}_0 \otimes \mathrm{Id})(\mathrm{Id} \otimes \boldsymbol{A}_1) I_{\boldsymbol{\Lambda}} + \qquad (5.A.1) \\ & R_{\check{\boldsymbol{\Lambda}}}(\mathrm{Id} \otimes \boldsymbol{A}_1)(\boldsymbol{U}_0 \otimes \mathrm{Id}) I_{\boldsymbol{\Lambda}}. \qquad (5.A.2) \end{aligned}$$

Considering (5.A.1), the range of $(\mathrm{Id} \otimes \boldsymbol{A}_1) I_{\boldsymbol{\Lambda}}$ consists of vectors whose entries with first index outside $P_0\boldsymbol{\Lambda}$ are zero. In view of the subsequent application of $\boldsymbol{L}_0 \otimes \mathrm{Id}$, furthermore only those indices $(\lambda, \gamma) \in P_0\boldsymbol{\Lambda} \times \check{\vee}^1$ of these vectors might be relevant for which $\exists(\mu, \gamma) \in \check{\boldsymbol{\Lambda}}$, i.e. $\gamma \in \boldsymbol{\Lambda}_{1,\mu}$, with $|\mu| > |\lambda|$ and $|\check{S}^0(\mu) \cap S^0(\lambda)| > 0$. Indeed $|\check{S}^0(\mu) \cap S^0(\lambda)| = 0$ implies $|\operatorname{supp} \check{\psi}^0_\mu \cap \operatorname{supp} \psi^0_\lambda| = 0$, and so $A_0(\check{\psi}^0_\mu, \psi^0_\lambda) = 0$. If for given (λ, γ) such a pair (μ, γ) exists for $|\mu| > |\lambda|$, then such a pair exists for $|\mu| = |\lambda| + 1$ as well, because $\check{\boldsymbol{\Lambda}}_{0,\gamma}$ is a tree, and $\check{S}^0(\mu') \supset \check{S}^0(\mu)$ for any ancestor μ' of μ. In order words, the condition $|\mu| > |\lambda|$ can be read as $|\mu| = |\lambda| + 1$. The set of (λ, γ) that we just described is given by the set $\boldsymbol{\Sigma}$, and so we infer that

$$R_{\check{\boldsymbol{\Lambda}}}(\boldsymbol{L}_0 \otimes \mathrm{Id})(\mathrm{Id} \otimes \boldsymbol{A}_1) I_{\boldsymbol{\Lambda}} = R_{\check{\boldsymbol{\Lambda}}}(\boldsymbol{L}_0 \otimes \mathrm{Id}) I_{\boldsymbol{\Sigma}} R_{\boldsymbol{\Sigma}} (\mathrm{Id} \otimes \boldsymbol{A}_1) I_{\boldsymbol{\Lambda}}.$$

Now let $(\lambda,\gamma)\in\mathbf{\Sigma}$. Using that $P_0\mathbf{\Lambda}$ is a tree, and $S^0(\lambda)\subset S^0(\lambda')$ for any ancestor λ' of λ, we infer that $(\lambda',\gamma)\in\mathbf{\Sigma}$. Using that for any $\mu\in P_0\check{\mathbf{\Lambda}}$, $\check{\mathbf{\Lambda}}_{1,\mu}$ is a tree, we infer that for any ancestor γ' of γ, $(\lambda,\gamma')\in\mathbf{\Sigma}$, so that $\mathbf{\Sigma}$ is a double-tree.

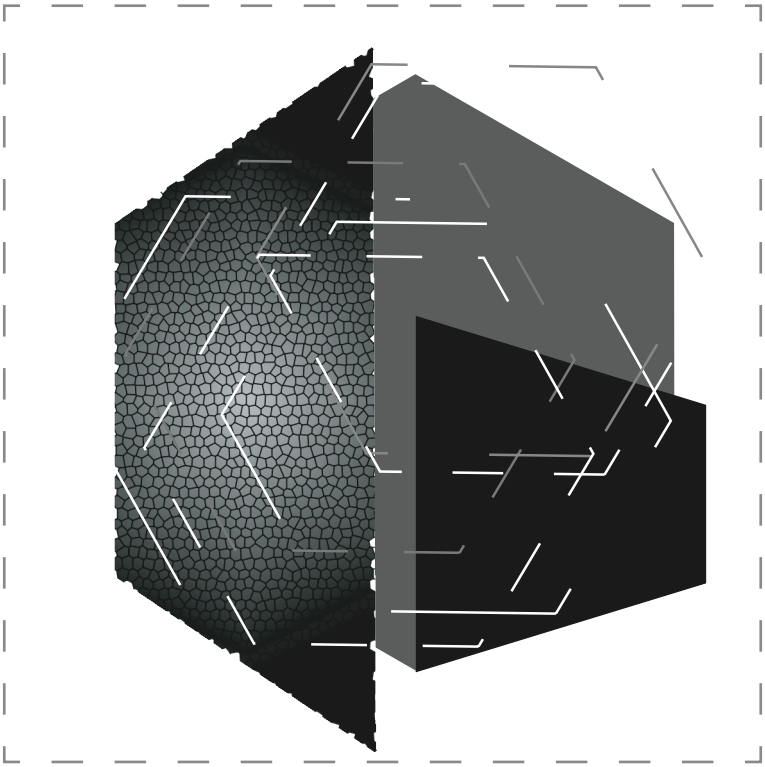
For $\mu\in\check{\vee}^0$, the number of $\lambda\in\vee^0$ with $|\mu|=|\lambda|+1$ and $|\check{S}^0(\mu)\cap S^0(\lambda)|>0$ is uniformly bounded, from which we infer that $\#\mathbf{\Sigma}\lesssim\sum_{\mu\in P_0\check{\mathbf{\Lambda}}}\#\check{\mathbf{\Lambda}}_{1,\mu}=\#\check{\mathbf{\Lambda}}$.

Considering (5.A.2), the range of $(\boldsymbol{U}_0\otimes\mathrm{Id})I_{\mathbf{\Lambda}}$ consists of vectors that can only have non-zero entries for indices $(\mu,\lambda)\in\check{\vee}^0\times P_1\mathbf{\Lambda}$ for which there exists a $\gamma\in\mathbf{\Lambda}_{0,\lambda}$ with $|\gamma|\geq|\mu|$ and $|\check{S}^0(\mu)\cap S^0(\gamma)|>0$. Since $\mathbf{\Lambda}_{0,\lambda}$ is a tree, and $S^0(\gamma')\supset S^0(\gamma)$ for any ancestor γ' of γ, equivalently $|\gamma|\geq|\mu|$ can be read as $|\gamma|=|\mu|$. Furthermore, in view of the subsequent application of $R_{\check{\mathbf{\Lambda}}}(\mathrm{Id}\otimes\boldsymbol{A}_1)$, it suffices to consider those indices (μ,λ) with $\mu\in P_0\check{\mathbf{\Lambda}}$. The set of (μ,λ) that we just described is given by the set $\mathbf{\Theta}$, and so we infer that

$$R_{\check{\mathbf{\Lambda}}}(\mathrm{Id}\otimes\boldsymbol{A}_1)(\boldsymbol{U}_0\otimes\mathrm{Id})I_{\mathbf{\Lambda}}=R_{\check{\mathbf{\Lambda}}}(\mathrm{Id}\otimes\boldsymbol{A}_1)I_{\mathbf{\Theta}}R_{\mathbf{\Theta}}(\boldsymbol{U}_0\otimes\mathrm{Id})I_{\mathbf{\Lambda}}.$$

Now let $(\mu,\lambda)\in\mathbf{\Theta}$. If λ' is an ancestor of λ, then from $P_0\mathbf{\Lambda}$ being a tree, and $\mathbf{\Lambda}_{0,\lambda}\subset\mathbf{\Lambda}_{0,\lambda'}$, we have $(\mu,\lambda')\in\mathbf{\Theta}$. If μ' is an ancestor of μ, then from $P_0\check{\mathbf{\Lambda}}$ being a tree, and $\check{S}^0(\mu')\supset\check{S}^0(\mu)$, we infer that $(\mu',\lambda)\in\mathbf{\Theta}$, and thus that $\mathbf{\Theta}$ is a double-tree.

For $\gamma\in\vee^0$, the number of $\mu\in\check{\vee}^0$ with $|\mu|=|\gamma|$ and $|\check{S}^0(\mu)\cap S^0(\gamma)|>0$ is uniformly bounded, from which we infer that $\#\mathbf{\Theta}\lesssim\sum_{\lambda\in P_1\mathbf{\Lambda}}\#\mathbf{\Lambda}_{0,\lambda}=\#\mathbf{\Lambda}$. □



6 A space-time parallel algorithm for parabolic PDEs

Abstract We present an algorithm for the solution of a simultaneous space-time discretization of linear parabolic evolution equations with a symmetric differential operator in space. Building on earlier work, we recast this discretization into a Schur-complement equation whose solution is a quasi-optimal approximation to the weak solution of the equation at hand. Choosing a tensor-product discretization, we arrive at a remarkably simple linear system. Using wavelets in time and standard finite elements in space, we solve the resulting system in linear complexity on a single processor, and in polylogarithmic complexity when parallelized in both space and time. We complement these theoretical findings with large-scale parallel computations showing the effectiveness of the method.

Source code is available at [vVW20b].

6.1 Introduction

This chapter deals with solving parabolic evolution equations numerically in a time-parallel fashion using tensor-product discretizations. Time-parallel algorithms for solving parabolic evolution equations have become a focal point following the enormous increase in parallel computing power. Spatial parallelism is a ubiquitous component in large-scale computations, but when spatial parallelism is exhausted, parallelization of the time axis is of interest.

Time-stepping methods first discretize the problem in space, and then solve the arising system of coupled ODEs sequentially. This immediately reveals a primary source of difficulty for time-parallel computation.

Alternatively, one can solve simultaneously in space *and* time. Originally introduced in [BJ89, BJ90], these space-time methods are very flexible: some can guarantee quasi-best approximations, meaning that their error is proportional to that of the best approximation from the trial space [And13, DS18, FK21, SZ20], or drive adaptive routines [SY18, RS19]. Many are especially well-suited for time-parallel computation [GN16, NS19]. Since the first contribution to time-parallel algorithms [Nie64] in 1964, many methods suitable for parallel computation have surfaced; see the review [Gan15].

This chapter is a minor modification of **A parallel algorithm for solving linear parabolic evolution equations**, R. van Venetië and J. Westerdiep, to appear in Parallel-in-Time Integration Methods, arXiv:2009.08875.

Parallel complexity The (serial) complexity of an algorithm measures asymptotic runtime on a single processor in terms of the input size. *Parallel complexity* measures asymptotic runtime given *sufficiently many* parallel processors having access to a shared memory, i.e., assuming communication is free.

In the current context of tensor-product discretizations of parabolic PDEs, we write N_t and N_x for the number of unknowns in time and space.

The parareal method [LMT01] aims at time-parallelism by alternating a serial coarse-grid solve with fine-grid computations in parallel. Each iteration has a time-parallel complexity of $\mathcal{O}(\sqrt{N_t}N_x)$, and combined with parallel multigrid in space, a parallel complexity of $\mathcal{O}(\sqrt{N_t}\log N_x)$. The popular MGRIT algorithm extends these ideas to multiple levels in time; cf. [FFK$^+$14].

Recently, Neumüller and Smears proposed an iterative algorithm that uses a Fast Fourier Transform in time. Each iteration runs in $\mathcal{O}(N_t\log(N_t)N_x)$ on a serial computer, and parallel in time, in $\mathcal{O}(\log(N_t)N_x)$. By incorporating parallel multigrid in space, its parallel runtime is reduced to $\mathcal{O}(\log N_t + \log N_x)$.

Our contribution We study the variational formulation introduced in Chapter 3 which was based on work by Andreev [And13, And16]. In [SvVW21] and Chapter 5, we studied this formulation in the context of space-time adaptivity and its efficient implementation in serial and on shared-memory parallel computers. The current chapter instead focuses on its massively parallel implementation and time-parallel performance.

Our method has remarkable similarities with the approach of [NS19], and the most essential difference is the substitution of their Fast Fourier Transform by our Fast Wavelet Transform. The strengths of both methods include a solid inf-sup theory that enables quasi-optimal approximate solutions from the trial space, ease of implementation, and excellent parallel performance in practice.

Our method has another strength: based on a wavelet transform, for fixed algebraic tolerance, it runs serially in linear complexity. Parallel in time, in complexity $\mathcal{O}(\log(N_t)N_x)$; parallel in *space and time*, in $\mathcal{O}(\log(N_tN_x))$. Moreover, when solving to an algebraic error proportional to the discretization error, incorporating a *nested iteration* (cf. [Hac85, Ch. 5]) results in complexities $\mathcal{O}(N_tN_x)$, $\mathcal{O}(\log(N_t)N_x)$, and $\mathcal{O}(\log^2(N_tN_x))$ respectively. This is on par with best-known results on parallel complexity for elliptic problems; cf. [Bra81].

Organization of this chapter In §6.2, we introduce the problem, derive a saddle-point formulation, and provide sufficient conditions for quasi-optimality of discrete solutions. In §6.3, we detail on the efficient computation of these discrete solutions. In §6.4 we take a concrete example—the reaction-diffusion equation—and analyze the serial and parallel complexity of our algorithm. In §6.5, we test these theoretical findings in practice. We conclude in §6.6.

Notations For normed linear spaces U and V, in this chapter for convenience over $\mathbb{R}$, $\mathcal{L}(U,V)$ will denote the space of bounded linear mappings $U \to V$ endowed with the operator norm $\|\cdot\|_{\mathcal{L}(U,V)}$. The subset of invertible operators in $\mathcal{L}(U,V)$ with inverses in $\mathcal{L}(V,U)$ will be denoted as $\mathcal{L}\mathrm{is}(U,V)$.

Given a finite-dimensional subspace U^δ of a normed linear space U, we denote the trivial embedding $U^\delta \to U$ by E_U^δ. For a basis Φ^δ—viewed formally as a column vector—of U^δ, we define the *synthesis operator* as

$$\mathcal{F}_{\Phi^\delta} : \mathbb{R}^{\dim U^\delta} \to U^\delta : \boldsymbol{c} \mapsto \boldsymbol{c}^\top \Phi^\delta =: \sum_{\phi \in \Phi^\delta} c_\phi \phi.$$

Equip $\mathbb{R}^{\dim U^\delta}$ with the Euclidean inner product and identify $(\mathbb{R}^{\dim U^\delta})'$ with $\mathbb{R}^{\dim U^\delta}$ using the corresponding Riesz map. We find the adjoint of $\mathcal{F}_{\Phi^\delta}$, the *analysis operator*, to satisfy

$$(\mathcal{F}_{\Phi^\delta})' : (U^\delta)' \to \mathbb{R}^{\dim U^\delta} : f \mapsto f(\Phi^\delta) := [f(\phi)]_{\phi \in \Phi^\delta}.$$

For quantities f and g, by $f \lesssim g$, we mean that $f \leq C \cdot g$ with a constant that does not depend on parameters that f and g may depend on. By $f \eqsim g$, we mean that $f \lesssim g$ and $g \lesssim f$. For matrices $\boldsymbol{A}$ and $\boldsymbol{B} \in \mathbb{R}^{N \times N}$, by $\boldsymbol{A} \eqsim \boldsymbol{B}$ we will denote *spectral equivalence*, i.e. $\boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x} \eqsim \boldsymbol{x}^\top \boldsymbol{B} \boldsymbol{x}$ for all $\boldsymbol{x} \in \mathbb{R}^N$.

6.2 Quasi-best approximations to the parabolic problem

Let V, H be separable Hilbert spaces of functions on a spatial domain so that V is continuously embedded in H, i.e. $V \hookrightarrow H$, with dense compact embedding. Identifying H with its dual yields the Gelfand triple $V \hookrightarrow H \simeq H' \hookrightarrow V'$.

For a.e.

$$t \in I := (0, T),$$

let $a(t; \cdot, \cdot)$ denote a bilinear form on $V \times V$ so that for any $\eta, \zeta \in V$, $t \mapsto a(t; \eta, \zeta)$ is measurable on I, and such that for a.e. $t \in I$,

$$\begin{aligned} |a(t; \eta, \zeta)| &\lesssim \|\eta\|_V \|\zeta\|_V \quad (\eta, \zeta \in V) \quad (\textit{boundedness}), \\ a(t; \eta, \eta) &\gtrsim \|\eta\|_V^2 \qquad (\eta \in V) \qquad (\textit{coercivity}). \end{aligned}$$

With $(A(t)\cdot)(\cdot) := a(t; \cdot, \cdot) \in \mathcal{L}\mathrm{is}(V, V')$, given a forcing function g and initial value u_0, we want to solve the *parabolic initial value problem* of

$$\text{finding } u : I \to V \text{ s.t.} \quad \begin{cases} \frac{\mathrm{d}u}{\mathrm{d}t}(t) + A(t)u(t) = g(t) & (t \in I), \\ u(0) = u_0. \end{cases} \tag{6.2.1}$$

6.2.1 An equivalent self-adjoint saddle-point system

In a simultaneous space-time variational formulation, the parabolic problem reads as finding u from a suitable space of functions of time and space s.t.

$$(Bw)(v) := \int_I \langle \tfrac{\mathrm{d}w}{\mathrm{d}t}(t), v(t) \rangle_H + a(t; w(t), v(t)) \, \mathrm{d}t = \int_I \langle g(t), v(t) \rangle_H =: g(v)$$

for all v from another suitable space of functions of time and space. We can enforce the initial condition by testing against additional test functions.

Theorem 6.2.1 ([SS09]). *With $Y := L_2(I;V)$, $X := Y \cap H^1(I;V')$, we have*

$$\begin{bmatrix} B \\ \gamma_0 \end{bmatrix} \in \mathcal{L}\mathrm{is}(X, Y' \times H),$$

where for $t \in \bar{I}$, $\gamma_t : u \mapsto u(t,\cdot)$ denotes the trace map. In other words,

$$\text{finding } u \in X \text{ s.t. } \quad (Bu, \gamma_0 u) = (g, u_0) \quad \text{given} \quad (g, u_0) \in Y' \times H \tag{6.2.2}$$

is a well-posed simultaneous space-time variational formulation of (6.2.1).

We define $A \in \mathcal{L}\mathrm{is}(Y, Y')$ and $\partial_t \in \mathcal{L}\mathrm{is}(X, Y')$ as

$$(Au)(v) := \int_I a(t; u(t), v(t))\,\mathrm{d}t, \quad \text{and} \quad \partial_t := B - A.$$

Following Chapter 3, we assume that A is *symmetric*. This is however not a necessary assumption, and the results extend naturally to the nonsymmetric case discussed in Chapter 4. We can reformulate (6.2.2) as the self-adjoint saddle point problem

$$\text{finding } (v, \sigma, u) \in Y \times H \times X \text{ s.t. } \quad \begin{bmatrix} A & 0 & B \\ 0 & \mathrm{Id} & \gamma_0 \\ B' & \gamma_0' & 0 \end{bmatrix} \begin{bmatrix} v \\ \sigma \\ u \end{bmatrix} = \begin{bmatrix} g \\ u_0 \\ 0 \end{bmatrix}. \tag{6.2.3}$$

By taking a Schur complement w.r.t. the H-block, we can reformulate this as

$$\text{finding } (v, u) \in Y \times X \text{ s.t. } \quad \begin{bmatrix} A & B \\ B' & -\gamma_0'\gamma_0 \end{bmatrix} \begin{bmatrix} v \\ u \end{bmatrix} = \begin{bmatrix} g \\ -\gamma_0' u_0 \end{bmatrix}. \tag{6.2.4}$$

We equip Y and X with *'energy'-norms*

$$\|\cdot\|_Y^2 := (A\cdot)(\cdot), \quad \|\cdot\|_X^2 := \|\partial_t \cdot\|_{Y'}^2 + \|\cdot\|_Y^2 + \|\gamma_T \cdot\|_H^2,$$

which are equivalent to the canonical norms on Y and X.

6.2.2 Uniformly quasi-optimal Galerkin discretizations

Our numerical scheme is based on the saddle-point formulation (6.2.4). Let $(Y^\delta, X^\delta)_{\delta\in\Delta}$ be a collection of closed subspaces of $Y \times X$ satisfying

$$X^\delta \subset Y^\delta, \quad \partial_t X^\delta \subset Y^\delta \quad (\delta \in \Delta), \tag{6.2.5}$$

and

$$1 \geq \gamma_\Delta := \inf_{\delta\in\Delta} \inf_{0\neq u\in X^\delta} \sup_{0\neq v\in Y^\delta} \frac{(\partial_t u)(v)}{\|\partial_t u\|_{Y'}\|v\|_Y} > 0. \tag{6.2.6}$$

Remark 6.2.2. In §3.4, we verify these conditions for X^δ and Y^δ being tensor-products of (locally refined) finite element spaces in time and space. In [SvVW21], we verify them for X_t^δ and Y^δ being *adaptive sparse grids*, allowing adaptive refinement locally in space *and* time simultaneously. ♢

For $\delta \in \Delta$, let $(v^\delta, \overline{u}^\delta) \in Y^\delta \times X^\delta$ solve the Galerkin discretization of (6.2.4):

$$\begin{bmatrix} {E_Y^\delta}' A E_Y^\delta & {E_Y^\delta}' B E_X^\delta \\ {E_X^\delta}' B' E_Y^\delta & -{E_X^\delta}' \gamma_0' \gamma_0 E_X^\delta \end{bmatrix} \begin{bmatrix} v^\delta \\ \overline{u}^\delta \end{bmatrix} = \begin{bmatrix} {E_Y^\delta}' g \\ -{E_X^\delta}' \gamma_0' u_0 \end{bmatrix}. \tag{6.2.7}$$

The solution $(v^\delta, \overline{u}^\delta)$ of (6.2.7) exists uniquely, and is *uniformly quasi-optimal* in that $\|u - \overline{u}^\delta\|_X \leq \gamma_\Delta^{-1} \inf_{u_\delta \in X^\delta} \|u - u_\delta\|_X$ for all $\delta \in \Delta$.

Instead of solving a matrix representation of (6.2.7) using e.g. preconditioned MINRES, we opt for a computationally more attractive method. Taking the Schur complement w.r.t. the Y^δ-block in (6.2.7), and replacing $({E_Y^\delta}' A E_Y^\delta)^{-1}$ in the resulting formulation by a *preconditioner* K_Y^δ that can be applied cheaply, we arrive at the *Schur complement formulation* of finding $u^\delta \in X^\delta$ s.t.

$$\underbrace{{E_X^\delta}' (B' E_Y^\delta K_Y^\delta {E_Y^\delta}' B + \gamma_0' \gamma_0) E_X^\delta}_{=:S^\delta} u^\delta = \underbrace{{E_X^\delta}' (B' E_Y^\delta K_Y^\delta {E_Y^\delta}' g + \gamma_0' u_0)}_{=:f^\delta}. \tag{6.2.8}$$

The resulting operator $S^\delta \in \mathcal{L}\mathrm{is}(X^\delta, {X^\delta}')$ is self-adjoint and elliptic. Given a self-adjoint operator $K_Y^\delta \in \mathcal{L}({Y^\delta}', Y^\delta)$ satisfying, for some $\kappa_\Delta \geq 1$,

$$\frac{((K_Y^\delta)^{-1} v)(v)}{(Av)(v)} \in [\kappa_\Delta^{-1}, \kappa_\Delta] \quad (\delta \in \Delta,\ v \in Y^\delta), \tag{6.2.9}$$

the solution u^δ of (6.2.8) exists uniquely as well. In fact, the following holds.

Theorem 6.2.3 (Remark 3.3.8)**.** *Take* $(Y^\delta \times X^\delta)_{\delta \in \Delta}$ *satisfying* (6.2.5)–(6.2.6)*, and* K_Y^δ *satisfying* (6.2.9)*. Solutions* $u^\delta \in X^\delta$ *of* (6.2.8) *are uniformly quasi-optimal:*

$$\|u - u^\delta\|_X \leq \tfrac{\kappa_\Delta}{\gamma_\Delta} \inf_{u_\delta \in X^\delta} \|u - u_\delta\|_X \quad (\delta \in \Delta).$$

6.3 Solving efficiently on tensor-product discretizations

We assume that $X^\delta := X_t^\delta \otimes X_{\boldsymbol{x}}^\delta$ and $Y^\delta := Y_t^\delta \otimes Y_{\boldsymbol{x}}^\delta$ are *tensor-products*, and for ease of presentation, we assume that the spatial discretizations on X^δ and Y^δ coincide, i.e. $X_{\boldsymbol{x}}^\delta = Y_{\boldsymbol{x}}^\delta$, reducing (6.2.5) to $X_t^\delta \subset Y_t^\delta$ and $\frac{\mathrm{d}}{\mathrm{d}t} X_t^\delta \subset Y_t^\delta$.

We equip X_t^δ with a basis Φ_t^δ, $X_{\boldsymbol{x}}^\delta$ with $\Phi_{\boldsymbol{x}}^\delta$, and Y_t^δ with Ξ^δ.

6.3.1 Construction of K_Y^δ

With $\boldsymbol{O} := \langle \Xi^\delta, \Xi^\delta \rangle_{L_2(I)}$, $\boldsymbol{A_x} := \langle \Phi_{\boldsymbol{x}}^\delta, \Phi_{\boldsymbol{x}}^\delta \rangle_V$, and $\boldsymbol{K_x} \eqsim \boldsymbol{A_x}^{-1}$ uniformly in δ, take

$$\boldsymbol{K}_Y := \boldsymbol{O}^{-1} \otimes \boldsymbol{K_x}.$$

Then $K_Y^\delta := \mathcal{F}_{\Xi^\delta \otimes \Phi_{\boldsymbol{x}}^\delta} \boldsymbol{K}_Y (\mathcal{F}_{\Xi^\delta \otimes \Phi_{\boldsymbol{x}}^\delta})' \in \mathcal{L}({Y^\delta}', Y^\delta)$ satisfies (6.2.9); cf. [SvVW21, §5.6.1]. When Ξ^δ is orthogonal, $\boldsymbol{O}$ is diagonal and can be inverted exactly. For standard finite element bases $\Phi_{\boldsymbol{x}}^\delta$, suitable $\boldsymbol{K_x}$ that can be applied efficiently (at linear cost) are provided by symmetric multigrid methods.

6.3.2 Preconditioning the Schur complement formulation

We will solve a matrix representation of (6.2.8) with an iterative solver, thus requiring a preconditioner. Inspired by the constructions of [And16, NS19], we build an *optimal* self-adjoint coercive preconditioner $K_X^\delta \in \mathcal{L}(X^{\delta\prime}, X^\delta)$ as a wavelet-in-time block-diagonal matrix with multigrid-in-space blocks.

Let U be a separable Hilbert space of functions over some domain. A given collection $\Psi = \{\psi_\lambda\}_{\lambda\in\vee_\Psi}$ is a *Riesz basis* for U when

$$\overline{\operatorname{span}\Psi} = U, \quad \text{and} \quad \|\boldsymbol{c}\|_{\ell_2(\vee_\Psi)} \eqsim \|\boldsymbol{c}^\top\Psi\|_U \quad \text{for all} \quad \boldsymbol{c} \in \ell_2(\vee_\Psi).$$

Thinking of Ψ being a basis of wavelet-type, for indices $\lambda \in \vee_\Psi$, its *level* is denoted $|\lambda| \in \mathbb{N}_0$. We call Ψ *uniformly local* when for all $\lambda \in \vee_\Psi$,

$$\operatorname{diam}(\operatorname{supp}\psi_\lambda) \lesssim 2^{-|\lambda|}, \ \ \#\{\mu \in \vee_\Psi : |\mu| = |\lambda|, |\operatorname{supp}\psi_\mu \cap \operatorname{supp}\psi_\lambda| > 0\} \lesssim 1.$$

Assume $\Sigma := \{\sigma_\lambda : \lambda \in \vee_\Sigma\}$ is a uniformly local Riesz basis for $L_2(I)$ with $\{2^{-|\lambda|}\sigma_\lambda : \lambda \in \vee_\Sigma\}$ Riesz for $H^1(I)$. Writing $w \in X$ as $\sum_{\lambda\in\vee_\Sigma} \sigma_\lambda \otimes w_\lambda$ for some $w_\lambda \in V$, we define the bounded, symmetric, and coercive bilinear form

$$(D_X \sum_{\lambda\in\vee_\Sigma} \sigma_\lambda \otimes w_\lambda)(\sum_{\mu\in\vee_\Sigma} \sigma_\mu \otimes v_\mu) := \sum_{\lambda\in\vee_\Sigma} \langle w_\lambda, v_\lambda\rangle_V + 4^{|\lambda|}\langle w_\lambda, v_\lambda\rangle_{V'}.$$

The operator $D_X^\delta := {E_X^\delta}' D_X E_X^\delta$ is in $\mathcal{L}\mathrm{is}(X^\delta, X^{\delta\prime})$. Its norm and that of its inverse are bounded uniformly in $\delta \in \Delta$. When $X^\delta = \operatorname{span}\Sigma^\delta \otimes \Phi_{\boldsymbol{x}}^\delta$ for some $\Sigma^\delta := \{\sigma_\lambda : \lambda \in \vee_{\Sigma^\delta}\} \subset \Sigma$, the matrix representation of D_X^δ w.r.t. $\Sigma^\delta \otimes \Phi_{\boldsymbol{x}}^\delta$ is

$$(\mathcal{F}_{\Sigma^\delta\otimes\Phi^\delta})' D_X^\delta \mathcal{F}_{\Sigma^\delta\otimes\Phi^\delta} =: \boldsymbol{D}_X^\delta = \operatorname{blockdiag}[\boldsymbol{A}_{\boldsymbol{x}} + 4^{|\lambda|}\langle\Phi_{\boldsymbol{x}}^\delta, \Phi_{\boldsymbol{x}}^\delta\rangle_{V'}]_{\lambda\in\vee_{\Sigma^\delta}}.$$

Theorem 6.3.1 ([SvVW21, §5.6.2])**.** *Define* $\boldsymbol{M}_{\boldsymbol{x}} := \langle\Phi_{\boldsymbol{x}}^\delta, \Phi_{\boldsymbol{x}}^\delta\rangle_H$. *With matrices* $\boldsymbol{K}_j \eqsim (\boldsymbol{A}_{\boldsymbol{x}} + 2^j\boldsymbol{M}_{\boldsymbol{x}})^{-1}$ *uniformly in* $\delta \in \Delta$ *and* $j \in \mathbb{N}_0$, *it follows that*

$$\boldsymbol{D}_X^{-1} \eqsim \boldsymbol{K}_X := \operatorname{blockdiag}[\boldsymbol{K}_{|\lambda|}\boldsymbol{A}_{\boldsymbol{x}}\boldsymbol{K}_{|\lambda|}]_{\lambda\in\vee_{\Sigma^\delta}}.$$

We find the optimal preconditioner $K_X^\delta := \mathcal{F}_{\Sigma^\delta\otimes\Phi^\delta}\boldsymbol{K}_X(\mathcal{F}_{\Sigma^\delta\otimes\Phi^\delta})' \in \mathcal{L}\mathrm{is}(X^{\delta\prime}, X^\delta)$.

In [OR00] it was shown that under a 'full-regularity' assumption, for quasi-uniform meshes, a multiplicative multigrid method yields $\boldsymbol{K}_j$ satisfying the conditions of Thm. 6.3.1, which can moreover be applied in linear time.

6.3.3 Wavelets in time

The preconditioner $\boldsymbol{K}_X$ requires X_t^δ to be equipped with a *wavelet* basis Σ^δ, whereas one typically uses a different (single-scale) basis Φ_t^δ on X_t^δ. To bridge this gap, a basis transformation from Σ^δ to Φ_t^δ is required. We define the wavelet transform as $\boldsymbol{W}_t := (\mathcal{F}_{\Phi_t^\delta})^{-1}\mathcal{F}_{\Sigma^\delta}$.[1]

[1] In literature, this transform is typically called an *inverse wavelet transform*.

Define $V_j := \operatorname{span}\{\sigma_\lambda \in \Sigma : |\lambda| \leq j\}$. Equip each V_j with a (single-scale) basis Φ_j, and assume that $\Phi_t^\delta := \Phi_J$ for some J, so that $X_t^\delta := V_J$. Since $V_{j+1} = V_j \oplus \operatorname{span}\Sigma_j$ where $\Sigma_j := \{\sigma_\lambda : |\lambda| = j\}$, there exist matrices $\boldsymbol{P}_j$ and $\boldsymbol{Q}_j$ such that $\Phi_j^\top = \Phi_{j+1}^\top \boldsymbol{P}_j$ and $\Psi_j^\top = \Phi_{j+1}^\top \boldsymbol{Q}_j$, with $\boldsymbol{M}_j := [\boldsymbol{P}_j | \boldsymbol{Q}_j]$ invertible.

Writing $v \in V_J$ in both forms $v = \boldsymbol{c}_0^\top \Phi_0 + \sum_{j=0}^{J-1} \boldsymbol{d}_j^\top \Psi_j$ and $v = \boldsymbol{c}_J^\top \Phi_J$, the transformation $\boldsymbol{W}_t := \boldsymbol{W}_J$ mapping wavelet coordinates $(\boldsymbol{c}_0^\top, \boldsymbol{d}_0^\top, \dots, \boldsymbol{d}_{J-1}^\top)$ to single-scale coordinates $\boldsymbol{c}_J$ satisfies

$$\boldsymbol{W}_J = \boldsymbol{M}_{J-1} \begin{bmatrix} \boldsymbol{W}_{J-1} & \boldsymbol{0} \\ \boldsymbol{0} & \mathbf{Id} \end{bmatrix}, \quad \text{and} \quad \boldsymbol{W}_0 := \mathbf{Id}. \tag{6.3.1}$$

Uniform locality of Σ implies *uniform sparsity* of the $\boldsymbol{M}_j$, i.e. with $\mathcal{O}(1)$ nonzeros per row and column. Then, assuming a geometrical increase in $\dim V_j$ in terms of j, which is true in the concrete setting below, matrix-vector products $\boldsymbol{x} \mapsto \boldsymbol{W}_t \boldsymbol{x}$ can be performed (serially) in linear complexity; cf. [Ste03].

6.3.4 Solving the system

The matrix representation of S^δ,f^δ from (6.2.8) w.r.t. a basis $\Phi_t^\delta \otimes \Phi_{\boldsymbol{x}}^\delta$ of X^δ is

$$\boldsymbol{S} := (\mathcal{F}_{\Phi_t^\delta \otimes \Phi_{\boldsymbol{x}}^\delta})' S^\delta \mathcal{F}_{\Phi_t^\delta \otimes \Phi_{\boldsymbol{x}}^\delta} \quad \text{and} \quad \boldsymbol{f} := (\mathcal{F}_{\Phi_t^\delta \otimes \Phi_{\boldsymbol{x}}^\delta})' f^\delta.$$

Envisioning an iterative solver, using §6.3.2 we have a preconditioner in terms of the wavelet-in-time basis $\Sigma^\delta \otimes \Phi_{\boldsymbol{x}}^\delta$. In this basis, we define

$$\hat{\boldsymbol{S}} := (\mathcal{F}_{\Sigma^\delta \otimes \Phi_{\boldsymbol{x}}^\delta})' S^\delta F_{\Sigma^\delta \otimes \Phi_{\boldsymbol{x}}^\delta} \quad \text{and} \quad \hat{\boldsymbol{f}} := (\mathcal{F}_{\Sigma^\delta \otimes \Phi_{\boldsymbol{x}}^\delta})' f^\delta. \tag{6.3.2}$$

These two forms are related: with the transform $\boldsymbol{W} := \boldsymbol{W}_t \otimes \mathbf{Id}_{\boldsymbol{x}}$, we have $\hat{\boldsymbol{S}} = \boldsymbol{W}^\top \boldsymbol{S} \boldsymbol{W}$ and $\hat{\boldsymbol{f}} = \boldsymbol{W}^\top \boldsymbol{f}$, and the matrix representation of (6.2.8) becomes

$$\text{finding } \boldsymbol{w} \quad \text{s.t.} \quad \hat{\boldsymbol{S}} \boldsymbol{w} = \hat{\boldsymbol{f}}. \tag{6.3.3}$$

We can then recover the solution in single-scale coordinates as $\boldsymbol{u} = \boldsymbol{W}\boldsymbol{w}$.

We use Preconditioned Conjugate Gradients (PCG), with preconditioner $\boldsymbol{K}_X$, to solve (6.3.3). Given an algebraic error tolerance $\varepsilon > 0$ and current guess $\boldsymbol{w}_k$, we monitor $\boldsymbol{r}_k^\top \boldsymbol{K}_X \boldsymbol{r}_k \leq \varepsilon^2$ where $\boldsymbol{r}_k := \hat{\boldsymbol{f}} - \hat{\boldsymbol{S}} \boldsymbol{w}_k$. This data is available within PCG, and is a stopping criterium: with $u_k^\delta := \mathcal{F}_{\Sigma^\delta \otimes \Phi_{\boldsymbol{x}}^\delta} \boldsymbol{w}_k \in X^\delta$, we see

$$\boldsymbol{r}_k^\top \boldsymbol{K}_X \boldsymbol{r}_k = (f^\delta - S^\delta u_k^\delta)(K_X^\delta (f^\delta - S^\delta u_k^\delta)) \eqsim \|u^\delta - u_k^\delta\|_X^2 \tag{6.3.4}$$

with $\eqsim$ following from [SvVW21, (4.12)]. Then $\|u^\delta - u_k^\delta\|_X \lesssim \varepsilon$.

6.4 A concrete setting: the reaction-diffusion equation

On a bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$, take $H := L_2(\Omega)$, $V := H_0^1(\Omega)$, and

$$a(t; \eta, \zeta) := \int_\Omega \boldsymbol{D} \nabla \eta \cdot \nabla \zeta + c \eta \zeta \, \mathrm{d}\boldsymbol{x}$$

where $\boldsymbol{D} = \boldsymbol{D}^\top \in \mathbb{R}^{d\times d}$ is positive definite, and $c \geq 0$.[2] We note that $A(t)$ is symmetric and coercive. W.l.o.g. we take $I := (0,1)$, i.e. $T := 1$.

Fix $p_t, p_{\boldsymbol{x}} \in \mathbb{N}$. With $\{\mathcal{T}_I\}$ the family of quasi-uniform partitions of I into subintervals, and $\{\mathcal{T}_\Omega\}$ that of conforming quasi-uniform triangulations of Ω, we take Δ as the collection of pairs $(\mathcal{T}_I, \mathcal{I}_\Omega)$. We construct trial- and test spaces

$$X^\delta := X_t^\delta \otimes X_{\boldsymbol{x}}^\delta, \quad Y^\delta := Y_t^\delta \otimes X_{\boldsymbol{x}}^\delta,$$

where, with $\mathbb{P}_p^{-1}(\mathcal{T})$ the space of piecewise degree-p polynomials on $\mathcal{T}$,

$$X_t^\delta := H^1(I) \cap \mathbb{P}_{p_t}^{-1}(\mathcal{T}_I), \quad X_{\boldsymbol{x}}^\delta := H_0^1(\Omega) \cap \mathbb{P}_{p_{\boldsymbol{x}}}^{-1}(\mathcal{T}_\Omega), \quad Y_t^\delta := \mathbb{P}_{p_t}^{-1}(\mathcal{T}_I).$$

These spaces satisfy condition (6.2.5), and the spatial discretizations on X^δ and Y^δ coincide. For this choice of Δ, condition (6.2.6) follows from Theorem 3.4.3.

For X_t^δ, we choose Φ_t^δ to be the Lagrange basis of degree p_t on $\mathcal{T}_I$; for $X_{\boldsymbol{x}}^\delta$, we choose $\Phi_{\boldsymbol{x}}^\delta$ to be that of degree $p_{\boldsymbol{x}}$ on $\mathcal{T}_\Omega$. An orthogonal basis Ξ^δ for Y_t^δ may be built as piecewise shifted Legendre polynomials of degree p_t w.r.t. $\mathcal{T}_I$.

For $p_t = 1$, one finds a suitable wavelet basis Σ in [Ste98]. For $p_t > 1$, one can split the basis into lowest- and higher-order parts and do the transform on the lowest-order part only, or build higher-order wavelets directly; cf. [Dij09].

Owing to the tensor-product structure of X^δ and Y^δ and of the operators A and ∂_t, the matrix of our formulation becomes remarkably simple.

Lemma 6.4.1. *Define* $\boldsymbol{g} := (\mathcal{F}_{\Xi^\delta \otimes \Phi_{\boldsymbol{x}}^\delta})' g$, $\boldsymbol{u}_0 := \Phi_t^\delta(0) \otimes \langle u_0, \Phi_{\boldsymbol{x}}^\delta \rangle_{L_2(\Omega)}$, *and*

$$\begin{aligned} \boldsymbol{T} &:= \langle \tfrac{\mathrm{d}}{\mathrm{d}t}\Phi_t^\delta, \Xi^\delta \rangle_{L_2(I)}, \quad \boldsymbol{N} := \langle \Phi_t^\delta, \Xi^\delta \rangle_{L_2(I)}, \quad \boldsymbol{\Gamma}_0 := \Phi_t^\delta(0)[\Phi_t^\delta(0)]^\top \\ \boldsymbol{M}_{\boldsymbol{x}} &:= \langle \Phi_{\boldsymbol{x}}^\delta, \Phi_{\boldsymbol{x}}^\delta \rangle_{L_2(\Omega)}, \quad \boldsymbol{A}_{\boldsymbol{x}} := \langle \Phi_{\boldsymbol{x}}^\delta, \Phi_{\boldsymbol{x}}^\delta \rangle_V, \quad \boldsymbol{B} := \boldsymbol{T} \otimes \boldsymbol{M}_{\boldsymbol{x}} + \boldsymbol{N} \otimes \boldsymbol{A}_{\boldsymbol{x}}. \end{aligned}$$

With $\boldsymbol{K}_Y := \boldsymbol{O}^{-1} \otimes \boldsymbol{K}_{\boldsymbol{x}}$ *from §6.3.1, we can write* $\boldsymbol{S}$ *and* $\boldsymbol{f}$ *from §6.3.4 as*

$$\boldsymbol{S} = \boldsymbol{B}^\top \boldsymbol{K}_Y \boldsymbol{B} + \boldsymbol{\Gamma}_0 \otimes \boldsymbol{M}_{\boldsymbol{x}}, \quad \boldsymbol{f} = \boldsymbol{B}^\top \boldsymbol{K}_Y \boldsymbol{g} + \boldsymbol{u}_0.$$

Note that $\boldsymbol{N}$ *and* $\boldsymbol{T}$ *are non-square,* $\boldsymbol{\Gamma}_0$ *is very sparse, and* $\boldsymbol{T}$ *is bidiagonal.*

In fact, assumption (6.2.5) allows us to write $\boldsymbol{S}$ in an even simpler form.

Lemma 6.4.2. *The matrix* $\boldsymbol{S}$ *can be written as*

$$\begin{aligned} \boldsymbol{S} = \boldsymbol{A}_t \otimes (\boldsymbol{M}_{\boldsymbol{x}} \boldsymbol{K}_{\boldsymbol{x}} \boldsymbol{M}_{\boldsymbol{x}}) + \boldsymbol{M}_t \otimes (\boldsymbol{A}_{\boldsymbol{x}} \boldsymbol{K}_{\boldsymbol{x}} \boldsymbol{A}_{\boldsymbol{x}}) + \boldsymbol{L}^\top \otimes (\boldsymbol{M}_{\boldsymbol{x}} \boldsymbol{K}_{\boldsymbol{x}} \boldsymbol{A}_{\boldsymbol{x}}) \\ + \boldsymbol{L} \otimes (\boldsymbol{A}_{\boldsymbol{x}} \boldsymbol{K}_{\boldsymbol{x}} \boldsymbol{M}_{\boldsymbol{x}}) + \boldsymbol{\Gamma}_0 \otimes \boldsymbol{M}_{\boldsymbol{x}} \end{aligned}$$

where

$$\boldsymbol{L} := \langle \tfrac{\mathrm{d}}{\mathrm{d}t}\Phi_t^\delta, \Phi_t^\delta \rangle_{L_2(I)}, \quad \boldsymbol{M}_t := \langle \Phi_t^\delta, \Phi_t^\delta \rangle_{L_2(I)}, \quad \boldsymbol{A}_t := \langle \tfrac{\mathrm{d}}{\mathrm{d}t}\Phi_t^\delta, \tfrac{\mathrm{d}}{\mathrm{d}t}\Phi_t^\delta \rangle_{L_2(I)}.$$

This matrix representation does not depend on Y_t^δ *or* Ξ^δ *at all.*

[2] This is easily generalized to variable coefficients, but notation becomes more obtuse.

Proof. The expansion of $\boldsymbol{B} := \boldsymbol{T} \otimes \boldsymbol{M_x} + \boldsymbol{N} \otimes \boldsymbol{A_x}$ in $\boldsymbol{S}$ yields a sum of five Kronecker products, one of which is

$$(\boldsymbol{T}^\top \otimes \boldsymbol{M_x})\boldsymbol{K}_Y(\boldsymbol{T} \otimes \boldsymbol{A_x}) = (\boldsymbol{T}^\top \boldsymbol{O}^{-1}\boldsymbol{N}) \otimes (\boldsymbol{M_x}\boldsymbol{K_x}\boldsymbol{A_x}).$$

We show that $\boldsymbol{T}^\top \boldsymbol{O}^{-1}\boldsymbol{N} = \boldsymbol{L}^\top$; similar arguments hold for the other terms. Thanks to $X_t^\delta \subset Y_t^\delta$, we can take the trivial embedding $F_t^\delta : X_t^\delta \to Y_t^\delta$. With

$$T^\delta \colon X_t^\delta \to {Y_t^\delta}', \quad (T^\delta u)(v) := \langle \tfrac{\mathrm{d}}{\mathrm{d}t} u, v \rangle_{L_2(I)},$$
$$M^\delta \colon Y_t^\delta \to {Y_t^\delta}', \quad (M^\delta u)(v) := \langle u, v \rangle_{L_2(I)},$$

we find $\boldsymbol{O} = (\mathcal{F}_{\Xi^\delta})' M^\delta \mathcal{F}_{\Xi^\delta}$, $\boldsymbol{N} = (\mathcal{F}_{\Xi^\delta})' M^\delta \mathcal{F}_t^\delta \mathcal{F}_{\Phi_t^\delta}$ and $\boldsymbol{T} = (\mathcal{F}_{\Xi^\delta})' T^\delta \mathcal{F}_{\Phi_t^\delta}$, so

$$\boldsymbol{T}^\top \boldsymbol{O}^{-1}\boldsymbol{N} = (\mathcal{F}_{\Phi_t^\delta})' {T^\delta}' F_t^\delta \mathcal{F}_{\Phi_t^\delta} = \langle \Phi_t, \tfrac{\mathrm{d}}{\mathrm{d}t}\Phi_t \rangle_{L_2(I)} = \boldsymbol{L}^\top. \qquad \square$$

6.4.1 Parallel complexity

The *parallel complexity* of our algorithm is the asymptotic runtime of solving (6.3.3) for $\boldsymbol{u} \in \mathbb{R}^{N_t N_x}$ in terms of $N_t := \dim X_t^\delta$ and $N_x := \dim X_x^\delta$, given sufficiently many parallel processors and assuming no communication cost.

We understand the serial (resp. parallel) cost of a matrix $\boldsymbol{B}$, denoted $C_{\boldsymbol{B}}^s$ (resp. $C_{\boldsymbol{B}}^p$), as the asymptotic runtime of performing $\boldsymbol{x} \mapsto \boldsymbol{B}\boldsymbol{x} \in \mathbb{R}^N$ in terms of N, on a single (resp. sufficiently many) processors at no communication cost. For *uniformly sparse* matrices, i.e. with $\mathcal{O}(1)$ nonzeros per row and column, the serial cost is $\mathcal{O}(N)$; the parallel cost is $\mathcal{O}(1)$ (compute cells of $\boldsymbol{B}\boldsymbol{x}$ concurrently).

From Theorem 6.3.1, we see that $\boldsymbol{K}_X$ is such that $\kappa_2(\boldsymbol{K}_X \hat{\boldsymbol{S}}) \lesssim 1$ uniformly in $\delta \in \Delta$. Therefore, for a given algebraic error tolerance ε, we require $\mathcal{O}(\log \varepsilon^{-1})$ PCG iterations. Assuming that the parallel cost of matrices dominates that of vector addition and inner products, the parallel complexity of a single PCG iteration is dominated by the cost of applying $\boldsymbol{K}_X$ and $\hat{\boldsymbol{S}}$. As $\hat{\boldsymbol{S}} = \boldsymbol{W}^\top \boldsymbol{S}\boldsymbol{W}$, our algorithm runs in complexity

$$\mathcal{O}(\log \varepsilon^{-1}[C_{\boldsymbol{K}_X}^\circ + C_{\boldsymbol{W}^\top}^\circ + C_{\boldsymbol{S}}^\circ + C_{\boldsymbol{W}}^\circ]) \quad (\circ \in \{s, p\}). \tag{6.4.1}$$

Theorem 6.4.3. *For fixed algebraic error tolerance $\varepsilon > 0$, our solver runs in*

- *serial complexity $\mathcal{O}(N_t N_x)$;*
- *time-parallel complexity $\mathcal{O}(\log(N_t) N_x)$;*
- *space-time-parallel complexity $\mathcal{O}(\log(N_t N_x))$.*

Proof. We absorb the constant factor $\log \varepsilon^{-1}$ of (6.4.1) into $\mathcal{O}$. We analyse the cost of every matrix separately.

The (inverse) wavelet transform As $\boldsymbol{W} = \boldsymbol{W}_t \otimes \mathbf{Id}_{\boldsymbol{x}}$, its serial cost is $\mathcal{O}(C_{\boldsymbol{W}_t}^s N_x)$. For this wavelet, $\boldsymbol{x} \mapsto \boldsymbol{W}_t \boldsymbol{x}$ has linear serial cost (cf. §6.3.3), so $C_{\boldsymbol{W}}^s = \mathcal{O}(N_t N_x)$.

Using (6.3.1), we write $\boldsymbol{W}_t$ as the composition of J matrices, each uniformly sparse and hence at parallel cost $\mathcal{O}(1)$. Because the mesh in time is quasi-uniform, we have $J \eqsim \log N_t$. We find $C_{\boldsymbol{W}_t}^p = \mathcal{O}(J) = \mathcal{O}(\log N_t)$, so the time-parallel cost of $\boldsymbol{W}$ equals $\mathcal{O}(\log(N_t) N_x)$. By exploiting spatial parallelism as well, we find $C_{\boldsymbol{W}}^p = \mathcal{O}(\log N_t)$. Analogous arguments hold for $\boldsymbol{W}_t^\top$ and $\boldsymbol{W}^\top$.

The preconditioner Recall that $\boldsymbol{K}_X := \operatorname{blockdiag}[\boldsymbol{K}_{|\lambda|}\boldsymbol{A}_{\boldsymbol{x}}\boldsymbol{K}_{|\lambda|}]_\lambda$. Since the cost of $\boldsymbol{K}_j$ is independent of j, we see that

$$C^s_{\boldsymbol{K}_X} = \mathcal{O}(N_t \cdot (2C^s_{\boldsymbol{K}_j} + C^s_{\boldsymbol{A}_{\boldsymbol{x}}})) = \mathcal{O}(2N_t C^s_{\boldsymbol{K}_j} + N_t N_{\boldsymbol{x}}).$$

Implementing the $\boldsymbol{K}_j$ as typical multiplicative multigrid solvers with linear serial cost, we find $C^s_{\boldsymbol{K}_X} = \mathcal{O}(N_t N_{\boldsymbol{x}})$.

Through temporal parallelism, we can apply each block of $\boldsymbol{K}_X$ concurrently, resulting in a time-parallel cost of $\mathcal{O}(2C^s_{\boldsymbol{K}_j} + C^s_{\boldsymbol{A}_{\boldsymbol{x}}}) = \mathcal{O}(N_{\boldsymbol{x}})$.

By parallelizing in space too, we reduce the cost of the uniformly sparse $\boldsymbol{A}_{\boldsymbol{x}}$ to $\mathcal{O}(1)$. The parallel cost of multiplicative multigrid on quasi-uniform triangulations is $\mathcal{O}(\log N_{\boldsymbol{x}})$; cf. [MFL+91]. It follows that $C^p_{\boldsymbol{K}_X} = \mathcal{O}(\log N_{\boldsymbol{x}})$.

The Schur matrix Using Lemma 6.4.1, we write $\boldsymbol{S} = \boldsymbol{B}^\top \boldsymbol{K}_Y \boldsymbol{B} + \boldsymbol{\Gamma}_0 \otimes \boldsymbol{M}_{\boldsymbol{x}}$ where $\boldsymbol{B} = \boldsymbol{T} \otimes \boldsymbol{M}_{\boldsymbol{x}} + \boldsymbol{N} \otimes \boldsymbol{A}_{\boldsymbol{x}}$, which immediately reveals that

$$\begin{aligned} C^s_{\boldsymbol{S}} &= C^s_{\boldsymbol{B}^\top} + C^s_{\boldsymbol{K}_Y} + C^s_{\boldsymbol{B}} + C^s_{\boldsymbol{\Gamma}_0} \cdot C^s_{\boldsymbol{M}} = \mathcal{O}(N_t N_{\boldsymbol{x}} + C^s_{\boldsymbol{K}_Y}), \quad \text{and} \\ C^p_{\boldsymbol{S}} &= \max\left\{ C^p_{\boldsymbol{B}^\top} + C^p_{\boldsymbol{K}_Y} + C^p_{\boldsymbol{B}},\ C^p_{\boldsymbol{\Gamma}_0} \cdot C^p_{\boldsymbol{M}} \right\} = \mathcal{O}(C^p_{\boldsymbol{K}_Y}) \end{aligned}$$

as every matrix except $\boldsymbol{K}_Y$ is uniformly sparse. With arguments similar to the previous paragraph, we see that $\boldsymbol{K}_Y$ (and hence $\boldsymbol{S}$) has serial cost $\mathcal{O}(N_t N_{\boldsymbol{x}})$, time-parallel cost $\mathcal{O}(N_{\boldsymbol{x}})$, and space-time-parallel cost $\mathcal{O}(\log N_{\boldsymbol{x}})$. □

6.4.2 Solving to higher accuracy

Instead of *fixing* the algebraic error tolerance, maybe more realistic is is to desire a solution $\tilde{u}^\delta \in X^\delta$ for which the error is proportional to the discretization error, i.e. $\|u - \tilde{u}^\delta\|_X \lesssim \inf_{u_\delta \in X^\delta} \|u - u_\delta\|_X$.

Assuming that this error decays with a (problem-dependent) rate $s > 0$, i.e. $\inf_{u_\delta \in X^\delta} \|u - u_\delta\|_X \lesssim (N_t N_{\boldsymbol{x}})^{-s}$, then the same holds for the solution u^δ of (6.2.8); cf. Thm. 6.2.3. When the algebraic error tolerance decays as $\varepsilon \lesssim (N_t N_{\boldsymbol{x}})^{-s}$, a triangle inequality and (6.3.4) show that the error of our solution $\tilde{u}^\delta$ obtained by PCG decays at rate s too.

In this case, $\log \epsilon^{-1} = \mathcal{O}(\log(N_t N_{\boldsymbol{x}}))$; (6.4.1) and the proof of Theorem 6.4.3 yield that the solver has superlinear serial complexity $\mathcal{O}(N_t N_{\boldsymbol{x}} \log(N_t N_{\boldsymbol{x}}))$, time-parallel complexity $\mathcal{O}(\log^2(N_t) \log(N_{\boldsymbol{x}}) N_{\boldsymbol{x}})$, and polylogarithmic complexity $\mathcal{O}(\log^2(N_t N_{\boldsymbol{x}}))$ parallel in space and time.

For elliptic PDEs, algorithms are available that offer quasi-optimal solutions, serially in linear complexity $\mathcal{O}(N_{\boldsymbol{x}})$—the cost of a serial solve to *fixed* algebraic error—and in parallel in $\mathcal{O}(\log^2 N_{\boldsymbol{x}})$, by combining a *nested iteration* with parallel multigrid; cf. [Hac85, Ch. 5] and [Bra81].

In [HVW95], the question is posed whether "good serial algorithms for parabolic PDEs are intrinsically as parallel as good serial algorithms for elliptic PDEs", basically asking if the lower bound $\mathcal{O}(\log^2(N_t N_{\boldsymbol{x}}))$ can be attained by an algorithm running serially in $\mathcal{O}(N_t N_{\boldsymbol{x}})$; see [Wor91, §2.2] for details.

Nested iteration drives down the serial complexity of our algorithm to a linear $\mathcal{O}(N_t N_x)$, and improves the time-parallel complexity to $\mathcal{O}(\log(N_t)N_x)$.[3] This is on par with the best-known results for elliptic problems, so we answer the question posed in [HVW95] in the affirmative.

6.5 Numerical experiments

We take the simple heat equation, i.e. $D = \mathbf{Id}_x$ and $c = 0$. We select $p_t = p_x = 1$, i.e. lowest order finite elements in space and time. We will use the 3-point wavelet introduced in [Ste98].

We implemented our algorithm in Python using the open source finite element library `NGSolve` [Sch14] for meshing and discretization of the bilinear forms in space and time, `MPI` through `mpi4py` [DPS05] for distributed computations, and `SciPy` [Vir20] for the sparse matrix-vector computations. The source code is available at [vVW20b].

6.5.1 Preconditioner calibration on a 2D problem

Our wavelet-in-time, multigrid-in-space preconditioner is optimal, so we have $\kappa_2(\boldsymbol{K}_X \hat{\boldsymbol{S}}) \lesssim 1$. Here we will investigate this condition number quantitatively.

As a model problem, we partition the temporal interval I uniformly into 2^J subintervals. We consider the domain $\Omega := [0,1]^2$ triangulated uniformly into 4^K triangles. Set $N_t := \dim X_t^\delta = 2^J + 1$ and $N_x := \dim X_x^\delta = (2^K - 1)^2$.

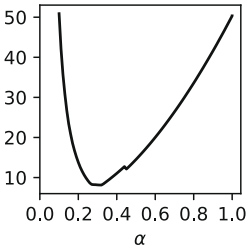
We start by using direct inverses $\boldsymbol{K}_j = (\boldsymbol{A}_x + 2^j \boldsymbol{M}_x)^{-1}$ and $\boldsymbol{K}_x = \boldsymbol{A}_x^{-1}$ to determine the best possible condition numbers. We found that replacing $\boldsymbol{K}_j$ by $\boldsymbol{K}_j^\alpha = (\alpha \boldsymbol{A}_x + 2^j \boldsymbol{M}_x)^{-1}$ for $\alpha = 0.3$ gave better conditioning; see also the left of Table 6.5.1. At the right of Table 6.5.1, we see that the condition numbers are very robust with respect to spatial refinements, but less so for refinements in time. Still, at $N_t = 16\,129$, we observe a modest $\kappa_2(\boldsymbol{K}_X \hat{\boldsymbol{S}})$ of 8.74.

Replacing the direct inverses with multigrid solvers, we found a good balance between speed and conditioning at 2 V-cycles with 3 Gauss–Seidel smoothing steps per grid. We decided to use these for our experiments.

6.5.2 Time-parallel results

We perform computations on Cartesius, the Dutch supercomputer. Each node has 64GB of memory and 12 cores (at 2 threads per core) running at 2.6GHz. Using the preconditioner detailed above, we iterate PCG on (6.3.3) with $\boldsymbol{S}$ computed as in Lemma 6.4.2, until achieving an algebraic error of $\varepsilon = 10^{-6}$; see also §6.3.4. For the spatial multigrid solvers, we use 2 V-cycles with 3 Gauss–Seidel smoothing steps per grid.

[3] Interestingly, nested iteration offers no improvements parallel in space *and* time, with complexity still $\mathcal{O}(\log^2(N_t N_x))$.



$N_t = 65$	129	257	513	1 025	2 049	4 097	8 193
$N_x = 49$	6.34	7.05	7.53	7.89	8.15	8.37	8.60	8.78
225	6.33	6.89	7.55	7.91	8.14	8.38	8.57	8.73
961	6.14	6.89	7.55	7.93	**8**.15	8.38	8.57	8.74
3 969	6.14	7.07	7.56	7.87	8.16	8.38	8.57	8.74
16 129	6.14	6.52	7.55	7.86	8.16	8.37	8.57	8.74

Table 6.1 Condition numbers $\kappa_2(\boldsymbol{K}_X\hat{\boldsymbol{S}})$. Left: fixed $N_t = 1025$, $N_x = 961$ for varying α. Right: fixed $\alpha = 0.3$ for varying N_t and N_x.

Memory-efficient time-parallel implementation For $\boldsymbol{X} \in \mathbb{R}^{N_x \times N_t}$, denote the vector obtained by stacking columns of $\boldsymbol{X}$ vertically by $\mathrm{Vec}(\boldsymbol{X}) \in \mathbb{R}^{N_t N_x}$. For memory efficency, we do not build the Kronecker matrices $\boldsymbol{B}_t \otimes \boldsymbol{B}_x$ of Lemma 6.4.2 directly, but instead perform matrix-vector products using

$$(\boldsymbol{B}_t \otimes \boldsymbol{B}_x)\,\mathrm{Vec}(\boldsymbol{X}) = \mathrm{Vec}(\boldsymbol{B}_x(\boldsymbol{B}_t\boldsymbol{X}^\top)^\top) = (\mathbf{Id}_t \otimes \boldsymbol{B}_x)\,\mathrm{Vec}(\boldsymbol{B}_t\boldsymbol{X}^\top). \tag{6.5.1}$$

Each parallel processor stores only a subset of the temporal degrees of freedom, i.e. a subset of columns of $\boldsymbol{X}$. When $\boldsymbol{B}_t$ is uniformly sparse, which holds true for all our temporal matrices, using (6.5.1) we evaluate $(\boldsymbol{B}_t \otimes \boldsymbol{B}_x)\,\mathrm{Vec}(\boldsymbol{X})$ in $\mathcal{O}(C^{\mathrm{s}}_{\boldsymbol{B}_x})$ operations parallel in time: on each parallel processor, we compute 'our' columns of $\boldsymbol{Y} := \boldsymbol{B}_t\boldsymbol{X}^\top$ by receiving the necessary columns of $\boldsymbol{X}$ from neighbouring processors, and then compute $\boldsymbol{B}_x\boldsymbol{Y}^\top$ without communication.

The preconditioner $\boldsymbol{K}_X$ is block-diagonal, making its time-parallel application trivial. Representing the wavelet transform of §6.3.3 as the composition of J Kronecker products allows a time-parallel implementation using the above.

2D problem We select $\Omega := [0,1]^2$ with a uniform triangulation $\mathcal{T}_\Omega$, and we triangulate I uniformly into $\mathcal{T}_I$. We select the smooth solution

$$u(t,x,y) := \exp(-2\pi^2 t)\sin(\pi x)\sin(\pi y),$$

so the problem has vanishing forcing data g.

Table 6.2 details the strong scaling results, i.e. fixing the problem size and increasing the number of processors P. We triangulate I into 2^{14} time slabs, yielding $N_t = 16\,385$ temporal degrees of freedom, and Ω into 4^8 triangles, yielding a X_x^δ of dimension $N_x = 65\,025$. The resulting system has 1 065 434 625 degrees of freedom and our solver reaches the algebraic error tolerance after 16 iterations. In perfect strong scaling, the total number of CPU-hours remains constant. Even at 2 048 processors, we observe a parallel efficiency of around 92.9%, solving this system in a modest 11.7 CPU-hours. Acquiring strong scaling results on a single node was not possible due to memory limitations.

Table 6.3 details the weak scaling results, i.e. fixing the problem size per processor and increasing the number of processors. In perfect weak scaling, the time per iteration should remain constant. We observe a slight increase in

P	N_t	N_x	$N = N_t N_x$	its	time (s)	time/it (s)	CPU-hrs
1–16	16 385	65 025	1 065 434 625		——— out of memory ———		
32	16 385	65 025	1 065 434 625	16	1224.85	76.55	10.89
64	16 385	65 025	1 065 434 625	16	615.73	38.48	10.95
128	16 385	65 025	1 065 434 625	16	309.81	19.36	11.02
256	16 385	65 025	1 065 434 625	16	163.20	10.20	11.61
512	16 385	65 025	1 065 434 625	16	96.54	6.03	13.73
512	16 385	65 025	1 065 434 625	16	96.50	6.03	13.72
1 024	16 385	65 025	1 065 434 625	16	45.27	2.83	12.88
2 048	16 385	65 025	1 065 434 625	16	20.59	1.29	11.72

Table 6.2 Strong scaling results for the 2D problem.

	P	N_t	N_x	$N = N_t N_x$	its	time (s)	time/it (s)	CPU-hrs
single node	1	9	261 121	2 350 089	8	33.36	4.17	0.01
	2	17	261 121	4 439 057	11	46.66	4.24	0.03
	4	33	261 121	8 616 993	12	54.60	4.55	0.06
	8	65	261 121	16 972 865	13	65.52	5.04	0.15
	16	129	261 121	33 684 609	13	86.94	6.69	0.39
multiple nodes	32	257	261 121	67 108 097	14	93.56	6.68	0.83
	64	513	261 121	133 955 073	14	94.45	6.75	1.68
	128	1 025	261 121	267 649 025	14	93.85	6.70	3.34
	256	2 049	261 121	535 036 929	15	101.81	6.79	7.24
	512	4 097	261 121	1 069 812 737	15	101.71	6.78	14.47
	1 024	8 193	261 121	2 139 364 353	16	108.32	6.77	30.81
	2 048	16 385	261 121	4 278 467 585	16	109.59	6.85	62.34

Table 6.3 Weak scaling results for the 2D problem.

P	N_t	N_x	$N = N_t N_x$	its	time (s)	time/it (s)	CPU-hrs
1–64	16 385	250 047	4 097 020 095		——— out of memory ———		
128	16 385	250 047	4 097 020 095	18	3 308.49	174.13	117.64
256	16 385	250 047	4 097 020 095	18	1 655.92	87.15	117.75
512	16 385	250 047	4 097 020 095	18	895.01	47.11	127.29
1 024	16 385	250 047	4 097 020 095	18	451.59	23.77	128.45
2 048	16 385	250 047	4 097 020 095	18	221.12	12.28	125.80

Table 6.4 Strong scaling results for the 3D problem.

P	N_t	N_x	$N = N_t N_x$	its	time (s)	time/it (s)	CPU-hrs
16	129	250 047	32 256 063	15	183.65	12.24	0.82
32	257	250 047	64 262 079	16	196.26	12.27	1.74
64	513	250 047	128 274 111	16	197.55	12.35	3.51
128	1 025	250 047	256 298 175	17	210.21	12.37	7.47
256	2 049	250 047	512 346 303	17	209.56	12.33	14.90
512	4 097	250 047	1 024 442 559	17	210.14	12.36	29.89
1 024	8 193	250 047	2 048 635 071	18	221.77	12.32	63.08
2 048	16 385	250 047	4 097 020 095	18	221.12	12.28	125.80

Table 6.5 Weak scaling results for the 3D problem.

time per iteration on a single node, but when scaling to multiple nodes, we observe a near-perfect parallel efficiency of around 96.7%, solving the final system with 4 278 467 585 degrees of freedom in a mere 109 seconds.

3D problem We select $\Omega := [0,1]^3$, and prescribe the solution

$$u(t,x,y,z) := \exp(-3\pi^2 t)\sin(\pi x)\sin(\pi y)\sin(\pi z),$$

so the problem has vanishing forcing data g.

Table 6.4 shows the strong scaling results. We triangulate I uniformly into 2^{14} time slabs, and Ω uniformly into 8^6 tetrahedra. The arising system has $N = 4\,097\,020\,095$ unknowns, which we solve to tolerance in 18 iterations. The results are comparable to those in two dimensions, albeit a factor two slower at similar problem sizes.

Table 6.5 shows the weak scaling results for the 3D problem. As in the two-dimensional case, we observe excellent scaling properties, and see that the time per iteration is nearly constant.

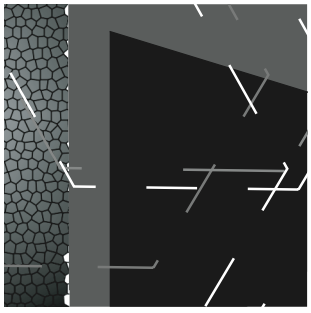
6.6 Conclusion

We have presented a framework for solving linear parabolic evolution equations massively in parallel. Based on earlier ideas [And16, NS19] and Chapter 3, we found a remarkably simple symmetric Schur-complement equation. With a tensor-product discretization of the space-time cylinder using standard finite elements in time and space together with a wavelet-in-time multigrid-in-space preconditioner, we were able to solve the arising systems to fixed accuracy in a uniformly bounded number of PCG steps.

We found that our algorithm runs in linear complexity on a single processor. Moreover, when *sufficiently many* parallel processors are available and communication is free, its runtime scales *logarithmically* in the discretization size. This complexity translates to a highly efficient algorithm in practice.

The numerical experiments serve as a showcase for the described space-time method, and exhibit its excellent time-parallelism by solving a linear system with over four billion unknowns in just 109 seconds, using just over two thousand parallel processors. By incorporating spatial parallelism as well, we expect these results to scale well to much larger problems.

Although performed in the rather restrictive setting of the heat equation discretized using piecewise linear polynomials on uniform triangulations, the parallel framework already allows solving more general linear parabolic PDEs using polynomials of varying degree on locally refined (tensor-product) meshes. In this more general setting, we envision load balancing to become the main hurdle in achieving good scaling results.



Γω
f = 1/2

7 Accuracy controlled data assimilation for parabolic PDEs

Abstract This chapter is concerned with the recovery of (approximate) solutions to parabolic problems from incomplete and possibly inconsistent observational data, given on a time-space cylinder that is a strict subset of the computational domain under consideration. Unlike previous approaches to this and related problems our starting point is a *regularized least squares* formulation in a continuous *infinite-dimensional* setting that is based on stable variational *time-space* formulations of the parabolic PDE. This allows us to derive a priori as well as a posteriori error bounds for the recovered states with respect to a certain reference solution. In these bounds the regularization parameter is disentangled from the underlying discretization. An important ingredient for the derivation of a posteriori bounds is the construction of suitable *Fortin operators* which allow us to control oscillation errors stemming from the discretization of dual norms. Moreover, the variational framework allows us to contrive preconditioners for the discrete problems whose application can be performed in linear time, and for which the condition numbers of the preconditioned systems are uniformly proportional to that of the regularized continuous problem. In particular, we provide suitable stopping criteria for the iterative solvers based on the a posteriori error bounds. The presented numerical experiments quantify the theoretical findings and demonstrate the performance of the numerical scheme in relation with the underlying discretization and regularization.

7.1 Introduction

7.1.1 Background

Ever-increasing computational resources encourage considering more and more complex mathematical models for the simulation or prediction of physical or technological processes. However, striving for increasing quantifiable accuracy such models typically exhibit significant bias or are incomplete in that important model data or accurate constitutive laws are missing. It is all too

This chapter is a minor modification of **Accuracy controlled data assimilation for parabolic problems**, W. Dahmen, R. Stevenson, and J. Westerdiep, to appear in *Mathematics of Computation*, arXiv:2105.05836.

natural to gather complementary information from *data* provided by also ever-improving sensor capabilities. Such a process of fusing models and data is often referred to as *data assimilation* which seems to originate from climatology [Dal94, LLD06]. In this context, streaming data are used to stabilize otherwise chaotic dynamical systems for prediction purposes, typically with the aid of (statistical) filtering techniques. While this is still an expanding and vibrant research area [Maj16], the notion of data assimilation is by now understood in a wider sense referring to efforts of improving quantifiable information extraction from synthesizing model-based and data driven approaches. Incompleteness of underlying models or model deficiencies could come in different forms. For instance, one could lack model data such as initial conditions, or the model involves uncalibrated coefficients represented e.g. by a parameter-dependent family of coefficients.

In this chapter we focus on such a problem scenario where the physical law takes the form of a parabolic partial differential equation (PDE), in the simplest case just the heat equation in combination with a known source term. We then assume that the state of interest, a (near-)solution to this PDE, can be observed on some restricted *time-space cylinder* while its initial conditions are unknown. We are then interested in recovering the partially observed state on the whole time-space domain from the given information.

This problem is known to be (mildly) ill-posed. This or related problems have been treated in numerous articles. In particular, the recent work in [BO18, BIHO18] proposes a finite element method with built-in *mesh-dependent* regularization terms has been a primary motivation for the present chapter. Moreover, similar concepts for an analogous data-assimilation problem associated with the wave equation have been applied in [BFMO21]. Considering first a semi-discretization in [BO18], the main results for a fully discrete scheme in [BIHO18] provide a priori estimates for the recovered state on a domain that excludes a small region around the location of initial data.

The results obtained in the present chapter, although similar in nature, are instead based on a different approach and exhibit a few noteworthy distinctions explained below. In fact, our starting point is the formulation of a *regularized* estimation problem in terms of a *least squares* functional in an *infinite-dimensional* function space setting. We postpone for a moment the particular role of the regularization parameter in the present context and remark first that our approach resembles a number of other prior studies of ill-posed operator equations, that are also based on a *Tikhonov regularization* in terms of similar mixed variational formulations, see e.g. [BBFD15, BR18, BLO18, DHH13, MS17]. These contributions are typically formulated in more general setting (see e.g. [BR18]), covering also problems that exhibit a stronger level of ill-posedness such as the Cauchy problem for second order elliptic equations, or the backward heat equation. Although a direct comparison with these works is therefore difficult, there are noteworthy relevant conceptual links as well as distinctions that we will briefly comment on next.

For instance, in [BR18], one arrives at a similar mixed formulation as in the present chapter exploiting then, however, just coercivity where, for a regularizing term $\varepsilon\|\cdot\|$, the coercivity constant decreases proportionally to the regu-

larization parameter. The results in [BBFD15] for related numerical schemes indeed confirm a corresponding adverse dependence of error bounds on the regularization parameter. Moreover, these bounds are obtained only under *additional regularity* assumptions. In contrast, our approach is based on numerically realizing *inf-sup stability* needed to handle dual norms, resulting in the present context in ε-independent error estimates without any a priori additional regularity assumptions.

This hints at the perhaps main principal distinction from the above prior related work. Our guiding theme is that, how well one can solve the inverse problem, depends on the condition of the corresponding forward problem (already on an infinite-dimensional level), in the present context a parabolic initial-boundary value problem. Specifically, this requires identifying first a suitable pair X, Y of (infinite-dimensional) trial- and test-spaces, for which the forward operator takes X onto the *dual* Y' of Y, without imposing any "excess regularity" assumptions on the solution beyond membership to X. This is tantamount to a *stable variational formulation* of the forward problem in the sense of the Babuska–Necas theorem. We briefly refer to this as "natural mapping properties". Drawing on the works of [And13, RS18b] and Chapter 3, the present approach is solely based on such natural mapping properties. As a consequence, the basic error analysis is independent of any data-consistency assumptions or of the regularity of solutions in the case of consistent data, which in general never occur in practice.

In summary, the guiding "general hope" is that, just exploiting natural mapping properties rather than assuming any excess regularity, should "help" minimizing the necessary amount of regularization in an inverse problem. This in turn, is intimately related to the central motivation of this chapter, namely the development of *efficient* and *certifiable* numerical methods that should not rely on unverifiable assumptions. In a nutshell, for the particular problem type at hand, significant consequences of a stable variational formulation of the forward problem are: the proposed numerical solvers exhibit a favorable quantifiable performance to be commented on further below; regardless of data consistency and without imposing any regularity assumptions we derive sharp a priori error bounds that do not degrade when the regularization parameter tends to zero; there is no need for tuning parameters inside any mesh-dependent stabilization terms; we can derive computable *a posteriori error bounds* that are valid without any excess regularity assumptions, for arbitrary (inconsistent) data, and, in the present particular inverse problem, are independent of the regularization parameter.

However, it should be emphasized that our "general hope" could so far be realized only for the current rather mildly ill-posed problem class. The following remarks elaborate a bit more on some of the related aspects.

(i) Respecting natural mapping properties reveals, in particular, that a unique minimizer of the objective functional exists for *any arbitrarily small regularization parameter and even for a vanishing regularization parameter*. In fact, a least squares formulation by itself turns out to be already a sufficient regularization. However, the *condition number* of corresponding discrete systems may increase with decreasing regularization parameter. Our numerical experiments

will shed some light on this interdependence. We use this insight to develop efficient preconditioners for the discrete problems. In fact, within the limitations of the infinite-dimensional formulation the solvers will be seen to exhibit a quasi-optimal performance for any fixed regularization parameter. Even for the mildly-ill posed problem under consideration this seems to be unprecedented in the literature. In that sense, the primary role of a non-vanishing regularization parameter for us is to facilitate a rigorous performance analysis of the iterative solver in favor of its quantitative improvement.

(ii) A stable infinite dimensional variational formulation is also an essential prerequisite for deriving rigorous *a posteriori* regularity-free - meaning they are valid without any excess regularity assumptions - error bounds for the recovered states. As shown later, such bounds can be used, in particular, to identify suitable stopping criteria for iterative solvers. Finally, we demonstrate some practical consequences of regularity-free computable a posteriori bounds in Section 7.6. We indicate their use for estimating data consistency errors as well as for choosing the regularization parameter in a way that accuracy of the results is not compromized in any essential way while enhancing solver performance.

(iii) Respecting natural mapping properties allows to "disentangle" discretization and regularization by studying the intrinsic necessary "strength" of the regularization in the infinite-dimensional setting. Moreover, it turns out that additional regularization beyond the least squares formulation is not necessary on the infinite-dimensional level, persists to remain true for the proposed inf-sup stable discretizations. Choosing nevertheless a positive regularization parameter in favor of a better and rigorously founded solver performance, still requires a balanced choice so as to warrant optimal achievable accuracy of the state estimate. Our formulation reveals that the relevant balance criterion is then the achievable approximation accuracy of the trial space. Only sufficiently high regularity, typically hard to check in practice, allows one to express this quantity in terms of a uniform mesh-size. Our approach will be seen to offer more flexibility and potentially different choices of regularization parameters than those stemming from the a priori fixed mesh-dependent approach in [BO18, BIHO18] or [DHH13]. This concerns, for instance, adaptively refined meshes or higher order discretizations.

A perhaps more subtle further consequence of exploiting natural mapping properties are somewhat stronger *a priori estimates* than those obtained in previous works.

Of course, the robustness of our results with respect to the regularization parameter reflects the mild degree of ill-posedness of the data-assimilation problem under consideration. This cannot be expected to carry over to less stable problems in exactly the same fashion. We claim though that important elements will persist to hold, for instance, for conditionally stable problems. In particular, non-vanishing regularization parameters will then be essential and regularity-free a posteriori bounds will be all the more important for arriving at properly balanced choices in the spirit of the strategy indicated in Section 7.6. A detailed discussion is beyond the scope of this chapter and is therefore deferred to forthcoming work.

7.1.2 Layout

In Section 7.2 we present a stable weak time-space formulation of a parabolic model problem and introduce the data assimilation problem considered in this work. Based on these findings we propose in Section 7.3 a regularized *least squares formulation* of the state estimation task. This formulation permits model as well as data errors as the recovered states are neither required to satisfy the parabolic equation exactly nor to match the data. We then derive *a priori* as well as *a posteriori* error estimates for the infinite-dimensional minimizer as well as for the minimizer over a finite dimensional trial space revealing the basic interplay between model inconsistencies, data errors, and regularization strength.

Since the "ideal" infinite-dimensional objective functional involves a dual-norm a practical numerical method needs to handle this term. We show that a proper discretization of the dual norm is tantamount to identifying a stable *Fortin operator*. For the given formulation of the parabolic problem this turns out to impose theoretical limitations on discretizations based on a standard second order variational formulation. Therefore we consider in Section 7.4 an equivalent first order system least squares formulation. Section 7.5 is devoted to the construction of Fortin operators for both settings. Moreover, we present in Sections 7.3.4 and 7.4.2 effective preconditioners for the iterative solution of the arising discrete problems along with suitable stopping criteria. Section 7.6 is devoted to numerical experiments that quantify the theoretical findings and illustrate the performance of the numerical schemes, in particular, depending on the choice of the regularization parameter which, in principal, could be chosen as zero. We conclude in Section 7.7 with a brief discussion of several ramifications of the results, including the application of a posteriori bounds for estimating data-consistency errors.

7.1.3 Notations

In this work, by $C \lesssim D$ we will mean that C can be bounded by a multiple of D, independently of parameters which C and D may depend on. Exceptions are given by the parameters η and ω in the Carleman estimate (7.2.7), the polynomial degrees of various finite element spaces, and the dimension d of the spatial domain Ω. Obviously, $C \gtrsim D$ is defined as $D \lesssim C$, and $C \eqsim D$ as $C \lesssim D$ and $C \gtrsim D$.

For normed linear spaces E and F, by $\mathcal{L}(E,F)$ we will denote the normed linear space of bounded linear mappings $E \to F$, and by $\mathcal{L}\mathrm{is}(E,F)$ its subset of boundedly invertible linear mappings $E \to F$. We write $E \hookrightarrow F$ to denote that E is continuously embedded into F. For simplicity only, we exclusively consider linear spaces over the scalar field $\mathbb{R}$.

7.2 Problem formulation and preview

For a given domain $\Omega \subset \mathbb{R}^d$ and time-horizon $T > 0$, let $I := [0,T]$. Let $a(t;\cdot,\cdot)$ denote a bilinear form on $H^1_0(\Omega) \times H^1_0(\Omega)$ such that for any $\theta, \zeta \in$

$H_0^1(\Omega)$, the function $t \mapsto a(t;\theta,\zeta)$ is measurable on I. Moreover, we assume that for almost all $t \in I$, $a(t;\cdot,\cdot)\colon H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ is *bounded* and *coercive*, i.e.

$$|a(t;\theta,\zeta)| \lesssim \|\theta\|_{H^1(\Omega)}\|\zeta\|_{H^1(\Omega)} \quad (\theta,\zeta \in H_0^1(\Omega)), \tag{7.2.1}$$

$$a(t;\theta,\theta) \gtrsim \|\theta\|^2_{H^1(\Omega)} \quad (\theta \in H_0^1(\Omega)), \tag{7.2.2}$$

hold with constants independent of $t \in I$. By Lax Milgram's Theorem, the operator $A(t)$, defined by $(A(t)\theta)(\zeta) := a(t;\theta,\zeta)$, $(\theta,\, \zeta \in H_0^1(\Omega))$, belongs to $\mathcal{L}\mathrm{is}(H_0^1(\Omega), H^{-1}(\Omega))$.

Before discussing the parabolic data assimilation problem, we recall some facts about a time-space variational formulation of the *parabolic initial value problem*—the corresponding forward problem—of the form

$$\begin{cases} \frac{dz}{dt}(t) + A(t)z(t) = h(t) & (t \in I \text{ a.e.}), \\ \gamma_0 z = z_0, \end{cases} \tag{7.2.3}$$

with trace map $\gamma_t\colon z \mapsto z(t)$. With the spaces

$$X := L_2(I; H_0^1(\Omega)) \cap H^1(I; H^{-1}(\Omega)), \quad Y := L_2(I; H_0^1(\Omega)),$$

the operator B defined by

$$(Bw)(v) := \int_I \tfrac{\mathrm{d}w}{\mathrm{d}t}(t)(v(t)) + a(t; w(t), v(t))\,\mathrm{d}t,$$

belongs to $\mathcal{L}(X, Y')$. Recall also that

$$X \hookrightarrow C(I; L_2(\Omega)) \tag{7.2.4}$$

with the latter space being equipped with the norm on $L_\infty(I; L_2(\Omega))$. In particular, this implies that $\gamma_t \in \mathcal{L}(X, L_2(\Omega))$, with a norm that is uniformly bounded in $t \in I$. The resulting weak formulation of (7.2.3) reads as

$$Bz = h, \quad \gamma_0 z = z_0,$$

and it is known (e.g. see [DL92, Ch.XVIII, §3], [Wlo82, Ch. IV, §26], [SS09]) to be well-posed in the sense that

$$\begin{bmatrix} B \\ \gamma_0 \end{bmatrix} \in \mathcal{L}\mathrm{is}(X, Y' \times L_2(\Omega)). \tag{7.2.5}$$

Turning to the data assimilation problem, suppose in what follows that $\omega \subset \Omega$ is a fixed non-empty sub-domain (possibly much smaller than Ω) and that we are given data $f \in L_2(I \times \omega)$ as well as $g \in Y'$. The *data-assimilation* problem considered in this chapter is to seek a state $u \in X$, that approximately

satisfies $Bu = g$, while also closely agreeing with f in $L_2(I \times \omega)$, see [BO18, BIHO18]. To make this precise, ideally one would like to solve

$$\begin{cases} \frac{du}{dt}(t) + A(t)u(t) &= g(t) \quad (t \in I \text{ a.e.}), \\ u(t)|_\omega &= f(t) \quad (t \in I). \end{cases} \tag{7.2.6}$$

However, in general such data (g, f) may be *inconsistent*, i.e., (7.2.6) has no solution and is therefore ill-posed. To put this formally, denoting by Γ_ω the restriction of a function on $I \times \Omega$ to a function on $I \times \omega$, we have $\Gamma_\omega \in \mathcal{L}(X, L_2(I \times \omega))$, i.e., Γ_ω is bounded on X. However, the range of the operator

$$B_\omega := \begin{bmatrix} B \\ \Gamma_\omega \end{bmatrix} \in \mathcal{L}(X, Y' \times L_2(I \times \omega))$$

induced by (7.2.6), is a strict subset of $Y' \times L_2(I \times \omega)$.

Before addressing this issue, it is instructive to understand the case of a *consistent* pair $(g, f) \in \operatorname{ran} B_\omega$, i.e., when there exists a $u \in X$ such that $(g, f) = (Bu, \Gamma_\omega u)$.

Remark 7.2.1. Any data consistent pair $(g, f) \in \operatorname{ran} B_\omega$ determines a unique state $u \in X$ satisfying (7.2.6). ◊

That this is indeed the case can be derived from the following crucial tool that has been employed in prior related studies such as [BO18, BIHO18] and will be heavily used in what follows as well. For $\eta \in (0, T)$ let

$$X_\eta := L_2([\eta, T]; H_0^1(\Omega)) \cap H^1([\eta, T]; H^{-1}(\Omega)).$$

Fixing both η and a subdomain $\omega \subset \Omega$, a version of the so-called *Carleman Estimate* says in the present terms

$$\|w\|_{X_\eta} \lesssim \|\Gamma_\omega w\|_{L_2(I\times\omega)} + \|Bw\|_{Y'} \quad (w \in X). \tag{7.2.7}$$

Remark 7.2.2. The validity of (7.2.7) has been established in [BO18, Thm. 2] for the *heat operator* (i.e., $a(t; \theta, \zeta) = \int_\Omega \nabla\theta(t) \cdot \nabla\zeta(t)\, d\boldsymbol{x}$) and $\Omega \subset \mathbb{R}^d$ being a *convex polytope*. It holds in greater generality though. For instance, the argument in the proof of [BO18, Lemma 7] still works when Ω is star-shaped w.r.t. an $x_0 \in \Omega$ and any open $\omega \subset \Omega$ that contains x_0. In what follows up to this point we will tacitly assume at this point suitable problem specifications that guarantee the validity of (7.2.7) without further mentioning. ◊

Returning to the uniqueness of u given consistent data (g, f), suppose there exist two solutions, then their difference $e \in X$ satisfies $\|\Gamma_\omega e\|_{L_2(I\times\omega)} + \|Be\|_{Y'} = 0$, meaning in view of (7.2.7) that $\|e\|_{X_\eta} = 0$, and so thanks to $X_\eta \hookrightarrow C([\eta, T]; L_2(\Omega))$, that $e(t) = 0$ for $t \in [\eta, T]$. From $X \hookrightarrow C([0, T]; L_2(\Omega))$, and the fact that $\eta > 0$ is arbitrary, it follows that $e = 0$.

However, the nature of the Carleman Estimate indicates that one cannot stably recover the trace $\gamma_0 u$ which would then together with g stably recover u. In fact, one may convince oneself that significantly different initial data (far) outside ω may give rise to homogeneous solutions of (7.2.3) that hardly differ on $I \times \omega$. Thus, even for a state $u \in X$ from (nearly) consistent data $(g, f) \in Y' \times L_2(I \times \omega)$ we cannot expect to find an accurate numerical approximation to u on the whole time-space cylinder $I \times \Omega$. Moreover, any perturbation of the data may land outside ran B_ω.

In practice, neither will the data/measurements $(g, f) \in Y' \times L_2(I \times \omega)$ be exact, nor will the observed state behind f satisfy the model—here a parabolic PDE—exactly. Thus, in general a pair of data $(g, f) \in Y' \times L_2(I \times \omega)$ allows one to recover any hypothetical source $u \in X$ only within some *uncertainty*. A central theme in this article is to quantify this uncertainty (theoretically and numerically) by properly exploiting the information provided by the PDE model, and the data. While any such assimilation attempt rests on the basic hypothesis that the data (g, f) are "close" in $Y' \times L_2(I \times \omega)$ to a consistent pair $(Bu, u|_{I\times\omega}) \in$ ran B_ω, for some $u \in X$, this "closeness" is generally not known beforehand.

To perform such a recovery we formulate in the next section a family of *regularizations* of the ill-posed problem (7.2.6) involving a parameter $\varepsilon \geq 0$, taking data errors and model bias into account. We then show, first on the continuous *infinite-dimensional* level, that for each $\varepsilon \geq 0$ there exists a unique regularized solution $u_\varepsilon \in X$. Letting this precede an actual *discrete* scheme, will be important for a number of issues, such as the design of efficient iterative solvers, the derivation of *a posteriori* error bounds, as well as disentangling regularization and discretization in favor of an overall good balance of uncertainties. Aside from the question what a preferable choice of ε would be in that latter respect, a central issue will be to assess the quality of a regularized state u_ε and of its approximation $u_\varepsilon^{\delta,\delta}$ from a given finite-dimensional trial space $X^\delta \subset X$ provided by our numerical scheme.

To that end, recall that generally (for inconsistent data) the idealized assimilation problem (7.2.6) has no solution. So whatever state $u \in X$ may be used to "explain" the data, should be viewed as a *candidate* or *reference state* that is connected with the recovery task through the *consistency error*

$$e_{\mathrm{cons}}(u) := \sqrt{\|Bu - g\|_{Y'}^2 + \|\Gamma_\omega u - f\|_{L_2(I\times\omega)}^2}. \tag{7.2.8}$$

At the heart of our analysis is then an a priori estimate of the type

$$\|u - u_\varepsilon^{\delta,\delta}\|_{X_\eta} \lesssim e_{\mathrm{cons}}(u) + e_{\mathrm{approx}}^\delta(u) + \varepsilon\|\gamma_0 u\|_{L_2(\Omega)}, \tag{7.2.9}$$

where $e_{\mathrm{approx}}^\delta(u)$ denotes the error of the best approximation to u from X^δ in X, thereby implicitly quantifying the regularity of the state u. Recall that, as always, the constant in this estimate absorbed by the $\lesssim$-symbol may depend on $\eta > 0$, but neither on u nor on ε.

It is important to note that (7.2.9) is valid for *any* $u \in X$, not making use of the assumption that $(Bu, \Gamma_\omega u)$ be close to (g, f). It is of evident value, of

course, for states u with small or at least moderate consistency error. This suggests singling out a particular state

$$u_0 := \operatorname*{arg\,min}_{u \in X} e_{\mathrm{cons}}(u)$$

that minimizes the consistency error. As in the case of consistent data, we will see that u_0 is unique, and, and as is suggested by its notation, it will turn out to be the limit for $\varepsilon \downarrow 0$ of the regularized solutions u_ε that will be defined later. One reason for not confining the error estimates—or perhaps better termed *distance estimates*—to the specific state u_0 is the potential significant model bias. In fact, we view it as a strength to keep (7.2.9) general since this covers automatically various somewhat specialized scenarios. For instance, if the data were exact, i.e., $e_{\mathrm{cons}}(u_0) = 0$, (7.2.9), for $u = u_0$, shows the dependence of the error just on ε and the choice of the discretization. Moreover, if the model is exact (or the model bias is negligible compared with data accuracy) it will later be seen how to get a "nearly-computable" bound for $e_{\mathrm{cons}}(u_0)$ and hence an idea of the model bias (due to g) and measurement errors in f. Another case of interest is $u = u_\varepsilon$ because this is the "compromise-solution" suggested by the chosen regularization and targeted by the numerical scheme.

Finally, while in principle, ε can be chosen as small as we wish (even zero), it will be seen to benefit solving the discrete problems by choosing ε as large as possible so as to remain just dominated, ideally, by $e_{\mathrm{cons}}(u_0)$, in practice, by the announced a posteriori bounds.

7.3 Regularized least squares

Knowing that the data assimilation problem is ill-posed and taking the preceding considerations into account, we consider for some parameter $\varepsilon \geq 0$ the regularized *least squares problem* of finding the minimizer u_ε over X of

$$G_\varepsilon \colon w \mapsto \|Bw - g\|^2_{Y'} + \|\Gamma_\omega w - f\|^2_{L_2(I\times\omega)} + \varepsilon^2 \|\gamma_0 w\|^2_{L_2(\Omega)},^1 \tag{7.3.1}$$

where, as before, Γ_ω is the restriction of a function on $I \times \Omega$ to a function on $I \times \omega$. The resulting Euler–Lagrange equations read as

$$\langle Bu_\varepsilon - g, Bw\rangle_{Y'} + \langle \Gamma_\omega u_\varepsilon - f, \Gamma_\omega w\rangle_{L_2(I\times\omega)} + \varepsilon^2 \langle \gamma_0 u_\varepsilon, \gamma_0 w\rangle_{L_2(\Omega)} = 0 \quad (w \in X). \tag{7.3.2}$$

Since $\Gamma_\omega \in \mathcal{L}(X, L_2(I \times \omega))$, and on account of (7.2.5), for $w \in X$ it holds that

$$\varepsilon^2 \|w\|^2_X \lesssim \|Bw\|^2_{Y'} + \|\Gamma_\omega w\|^2_{L_2(I\times\omega)} + \varepsilon^2 \|\gamma_0 w\|^2_{L_2(\Omega)} \lesssim \max(1, \varepsilon^2)\|w\|^2_X. \tag{7.3.3}$$

By the Lax–Milgram Lemma, we thus know that for $\varepsilon > 0$ the minimizer u_ε exists uniquely, and satisfies

$$\|u_\varepsilon\|_X \lesssim \max(\varepsilon^{-1}, 1)\Big(\|g\|_{Y'} + \|f\|_{L_2(I\times\omega)}\Big). \tag{7.3.4}$$

[1] We could have included additional weights in front of the first terms that could reflect a priori knowledge on model- or data-fidelity. Since this would not affect the subsequent developments we disregard this option for simplicity of exposition.

Selecting any reference state $u \in X$, similar to (7.3.4) one finds for $\varepsilon > 0$

$$\|u - u_\varepsilon\|_X \lesssim \max(\varepsilon^{-1}, 1)\Big(\varepsilon\|\gamma_0 u\|_{L_2(\Omega)} + e_{\rm cons}(u)\Big),$$

see (7.2.8). This result is by no means satisfactory. With the aid of (7.2.7), much better bounds will be established for $\|u - u_\varepsilon\|_{X_\eta}$.

Remark 7.3.1. Also for $\varepsilon = 0$, the minimizer u_0 of $G_0(\cdot) = e_{\rm cons}(\cdot)^2$ over X exists uniquely. Indeed, suppose there are two minimizers. Then, by (7.3.2), their difference e_0 satisfies

$$\langle Be_0, Bw\rangle_{Y'} + \langle \Gamma_\omega e_0, \Gamma_\omega w\rangle_{L_2(I\times\omega)} = 0 \quad (w \in X),$$

and so $\|\Gamma_\omega e_0\|^2_{L_2(I\times\omega)} + \|Be_0\|^2_{Y'} = 0$. As we have seen, using (7.2.4) and (7.2.7) this implies $e_0 = 0$. ♢

A frequently used tool reads as follows.

Lemma 7.3.2. *For any $w \in X$ one has*

$$\|u - w\|_{X_\eta} \lesssim \sqrt{G_0(w)} + e_{\rm cons}(u).$$

Proof. (7.2.7) and a triangle-inequality for the norm $\|\cdot\|_{L_2(I\times\omega)\times Y'}$ yield

$$\|u - w\|_{X_\eta} \lesssim \sqrt{\|\Gamma_\omega(u-w)\|^2_{L_2(I\times\omega)} + \|B(u-w)\|^2_{Y'}} \le \sqrt{G_0(w)} + e_{\rm cons}(u)$$

which confirms the claim. □

Taking as reference state $u = u_\varepsilon$, we obtain the following *a posteriori bound*.

Proposition 7.3.3. *For $\varepsilon \ge 0$ and $w \in X$, one has*

$$\|u_\varepsilon - w\|_{X_\eta} \lesssim \sqrt{G_\varepsilon(w)}.$$

Proof. The proof follows from Lemma 7.3.2 and

$$\sqrt{G_0(w)} + e_{\rm cons}(u_\varepsilon) \le \sqrt{G_\varepsilon(w)} + \sqrt{G_\varepsilon(u_\varepsilon)} \le 2\sqrt{G_\varepsilon(w)}. \quad \square$$

The same arguments, used to show for $\varepsilon \ge 0$ existence and uniqueness of the minimizer u_ε of G_ε over X, show for any closed subspace $X^\delta \subset X$ uniqueness of the minimizer u^δ_ε of G_ε over X^δ. An *a priori bound* for $\|u - u^\delta_\varepsilon\|_{X_\eta}$ for an arbitrary reference state $u \in X$ is given in the next proposition.

Proposition 7.3.4. *It holds that*

$$\|u - u^\delta_\varepsilon\|_{X_\eta} \lesssim e_{\rm cons}(u) + e^\delta_{\rm approx}(u) + \varepsilon\|\gamma_0 u\|_{L_2(\Omega)},$$

where

$$e^\delta_{\rm approx}(u) := \min_{w\in X^\delta} \|u - w\|_X$$

denotes the corresponding approximation error *of the state u.*

Proof. Let P_{X^δ} denote the X-orthogonal projector onto X^δ, then using $B \in \mathcal{L}(X, Y')$ and (7.2.4), we infer that

$$\begin{aligned}\sqrt{G_0(u_\varepsilon^\delta)} \leq & \sqrt{G_\varepsilon(u_\varepsilon^\delta)} \leq \sqrt{G_\varepsilon(P_{X^\delta}u)} \\ \leq & \|B(u - P_{X^\delta}u)\|_{Y'} + \|Bu - g\|_{Y'} + \|f - \Gamma_\omega u\|_{L_2(I\times\omega)} + \\ & \|\Gamma_\omega(u - P_{X^\delta}u)\|_{L_2(I\times\omega)} + \varepsilon\|\gamma_0(u - P_{X^\delta}u)\|_{L_2(\Omega)} + \varepsilon\|\gamma_0 u\|_{L_2(\Omega)} \\ \lesssim & e^\delta_{\mathrm{approx}}(u) + e_{\mathrm{cons}}(u) + \varepsilon\|\gamma_0 u\|_{L_2(\Omega)},\end{aligned}$$

which together with Lemma 7.3.2 completes the proof. □

At this point we note that because of the presence of the dual norm $\|\cdot\|_{Y'}$ in G_ε, neither u_ε^δ nor the a posteriori bound for $\|u_\varepsilon - w\|_{X_\eta}$ from Proposition 7.3.3 for e.g. $w = u_\varepsilon^\delta$ can be computed. Both problems are going to be tackled in the next two subsections.

Remark 7.3.5. Although the upper bound from Proposition 7.3.4 is minimal for $\varepsilon = 0$, a reason for nevertheless taking $\varepsilon > 0$, say of the order of the expected magnitude of $e_{\mathrm{cons}}(u) + e^\delta_{\mathrm{approx}}(u)$, is to enhance the numerical stability of solving the Euler–Lagrange equations. ◊

Remark 7.3.6. Notice that even when $e_{\mathrm{cons}}(u) = 0$ and $\varepsilon = 0$, Proposition 7.3.4 does not show that u_0^δ is a quasi-best approximation to u from X^δ. Indeed the norm $\|\cdot\|_X$ used to define $e^\delta_{\mathrm{approx}}(u)$ differs from the norm $\|\cdot\|_{X_\eta}$ in which $u - u_0^\delta$ is measured. ◊

We conclude this section with a few comments on the behavior of u_ε when ε tends to zero. First, note that the consistency error of u_ε approaches the minimal consistency error $e_{\mathrm{cons}}(u_0)$ when $\varepsilon \to 0$ because

$$e_{\mathrm{cons}}(u_\varepsilon) \leq \sqrt{G_\varepsilon(u_\varepsilon)} \leq \sqrt{G_\varepsilon(u_0)} \leq e_{\mathrm{cons}}(u_0) + \varepsilon\|\gamma_0 u_0\|_{L_2(\Omega)}.$$

In particular, a first trivial consequence of Proposition 7.3.4 is that, for consistent and exact data, i.e., $e_{\mathrm{cons}}(u_0) = 0$, u_ε tends to the state u_0 in X_η for any $\eta > 0$. Even without the assumption $e_{\mathrm{cons}}(u_0) = 0$, a stronger result is derived in the following remark.

Remark 7.3.7. One has $\|u_0 - u_\varepsilon\|_X \to 0$ as $\varepsilon \to 0$. ◊

Proof. We remark first that $\sqrt{G_\varepsilon}$ Γ-converges to $\sqrt{G_0} =: F$. In fact, let $(\varepsilon_n)_{n\in\mathbb{N}}$ tend to zero. The functionals $F_n := \sqrt{G_{\varepsilon_n}} : X \to \mathbb{R}_+$ are uniformly coercive (in the sense of optimization, meaning that $F_n(w) \to \infty$ for $\|w\|_X \to \infty$). Let $(w_n)_{n\in\mathbb{N}}$ be any sequence in X with limit $w \in X$. Then

$$\begin{aligned}F(w) - F_n(w_n) \leq F(w) - F(w_n) \leq & \sqrt{\|B(w - w_n)\|^2_{Y'} + \|\gamma_\omega(w - w_n)\|^2_{L_2(I\times\omega)}} \\ \lesssim & \|w - w_n\|_X, \quad n \in \mathbb{N},\end{aligned}$$

so that $F(w) \le \liminf_{n\to\infty} F_n(w_n)$. Moreover, for any $w \in X$ there exists a sequence $(w_n)_{n\in\mathbb{N}}$ in X such that $F(w) \ge \limsup_{n\to\infty} F_n(w_n)$, as can be seen by simply taking $w_n = w$. Thus, by the main Theorem of Γ-convergence, minimizers of F_n converge to the minimizer of F. □

Thus, trying to solve the regularized problem, with ε as small as possible incidentally favors u_0 as target state. Thus, it is of interest to estimate $e_{\mathrm{cons}}(u_0)$ (see Corollary 7.3.14 below) since a relatively large $e_{\mathrm{cons}}(u_0)$ weakens the relevance of u_0, favoring correspondingly larger regularization parameters.

7.3.1 Discretizing the dual norm

Minimizing G_ε over X^δ does not correspond to a practical method because the dual norm $\|\cdot\|_{Y'}$ cannot be evaluated. Therefore, given a family of finite dimensional subspaces $(X^\delta)_{\delta\in\Delta}$ of X, the idea is to find a family $(Y^\delta)_{\delta\in\Delta}$ of finite dimensional subspaces of Y, ideally with $\dim Y^\delta \lesssim \dim X^\delta$, such that $\|Bw\|_{Y'}$ can be controlled for $w \in X^\delta$ by the computable quantity $\|Bw\|_{Y^{\delta\prime}}$. This is ensured whenever

$$\inf_{\delta\in\Delta} \inf_{\{w\in X^\delta:\, Bw\neq 0\}} \sup_{0\neq\mu\in Y^\delta} \frac{(Bw)(\mu)}{\|Bw\|_{Y'}\|\mu\|_Y} > 0 \tag{7.3.5}$$

is valid.

In the subsequent discussion we make heavy use of the Riesz isometry $R \in \mathcal{L}\mathrm{is}(Y, Y')$, defined by

$$(Rv)(w) := \langle v, w\rangle_Y := \int_I \int_\Omega \nabla_{\boldsymbol{x}} v \cdot \nabla_{\boldsymbol{x}} w \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t, \quad (v, w \in Y).$$

Introducing auxiliary variables for $\mu_\varepsilon = R^{-1}(g - Bu_\varepsilon) \in Y$, $\theta_\varepsilon = f - \Gamma_\omega u_\varepsilon \in L_2(I\times\omega)$, and $\nu_\varepsilon = -\gamma_0 u_\varepsilon \in L_2(\Omega)$ gives rise to a mixed formulation of the problem of finding the minimizer u_ε over X of G_ε defined in (7.3.1) in terms of the saddle point system

$$S_\varepsilon(\mu_\varepsilon, \theta_\varepsilon, \nu_\varepsilon, u_\varepsilon) := \begin{bmatrix} R & 0 & 0 & B \\ 0 & I & 0 & \Gamma_\omega \\ 0 & 0 & I & \varepsilon\gamma_0 \\ B' & \Gamma'_\omega & \varepsilon\gamma'_0 & 0 \end{bmatrix} \begin{bmatrix} \mu_\varepsilon \\ \theta_\varepsilon \\ \nu_\varepsilon \\ u_\varepsilon \end{bmatrix} = \begin{bmatrix} g \\ f \\ 0 \\ 0 \end{bmatrix}. \tag{7.3.6}$$

(see [CDW12, Sect. 2.2]). (Equivalently, (7.3.6) characterizes the critical point of the Lagrangian obtained when inserting in G_ε these variables and appending corresponding constraints by Lagrange multipliers.)

Remark 7.3.8. Eliminating the second and third variable from (7.3.6), one arrives at the equivalent more compact formulation

$$\begin{bmatrix} R & B \\ B' & -(\Gamma'_\omega\Gamma_\omega + \varepsilon^2\gamma'_0\gamma_0) \end{bmatrix} \begin{bmatrix} \mu_\varepsilon \\ u_\varepsilon \end{bmatrix} = \begin{bmatrix} g \\ -\Gamma'_\omega f, \end{bmatrix}$$

It serves in Section 7.3.4 as the starting point for a numerical scheme. ◊

Theorem 7.3.9. *Let* (7.3.5) *be valid. For $u_\varepsilon^{\delta,\delta}$ the (unique) minimizer over X^δ of*

$$G_\varepsilon^\delta := w \mapsto \|Bw - g\|_{Y^{\delta\prime}}^2 + \|\Gamma_\omega w - f\|_{L_2(I\times\omega)}^2 + \varepsilon^2\|\gamma_0 w\|_{L_2(\Omega)}^2,$$

one has

$$\|u - u_\varepsilon^{\delta,\delta}\|_{X_\eta} \lesssim e_{\mathrm{cons}}(u) + e_{\mathrm{approx}}^\delta(u) + \varepsilon\|\gamma_0 u\|_{L_2(\Omega)}.$$

(We recall that, as always, the constant absorbed by the $\lesssim$-symbol may (actually will) depend on ω and η, but not *on $\varepsilon \geq 0$ or $\delta \in \Delta$.)*

Proof. Denoting the block-diagonal operator comprized of the leading 3×3 block in S_ε by D, the operator S_ε can be rewritten as

$$S_\varepsilon = \begin{bmatrix} D & C_\varepsilon \\ C_\varepsilon' & 0 \end{bmatrix},$$

where $C_\varepsilon \in \mathcal{L}\Big(X, Y' \times L_2(I\times\omega) \times L_2(\Omega)\Big)$ is defined by

$$(C_\varepsilon w)(\mu,\theta,\nu) := (Bw)(\mu) + \langle \Gamma_\omega w, \theta\rangle_{L_2(I\times\omega)} + \varepsilon\langle \gamma_0 w, \nu\rangle_{L_2(\Omega)}.$$

With the usual identification of $L_2(I\times\omega)$ and $L_2(\Omega)$ with their duals, D is just the isometric Riesz isomorphism between $Y \times L_2(I\times\omega) \times L_2(\Omega)$ and its dual. Equipping X with the (ε-dependent) "energy"-norm

$$|\!|\!| w |\!|\!|_\varepsilon := \sqrt{\|Bw\|_{Y'}^2 + \|\Gamma_\omega w\|_{L_2(I\times\omega)}^2 + \varepsilon^2\|\gamma_0 w\|_{L_2(\Omega)}^2},$$

one verifies that $\|C_\varepsilon w\|_{Y'\times L_2(I\times\omega)\times L_2(\Omega)} = |\!|\!| w |\!|\!|_\varepsilon$, so that in particular C_ε satisfies an 'inf-sup' condition. Consequently, the operator S_ε on the left hand side of (7.3.6) is a boundedly invertible mapping from $Y \times L_2(I\times\omega) \times L_2(\Omega) \times (X, |\!|\!|\ |\!|\!|_\varepsilon)$ to its dual (uniformly in ε).

Analogously to the continuous case, the minimizer $u_\varepsilon^{\delta,\delta}$ of G_ε^δ equals the fourth component of the solution $(\mu_\varepsilon^{\delta,\delta}, \theta_\varepsilon^{\delta,\delta}, \nu_\varepsilon^{\delta,\delta}, u_\varepsilon^{\delta,\delta})$ of the Galerkin discretization of (7.3.6) with trial space $Y^\delta \times L_2(I\times\omega) \times L_2(\Omega) \times X^\delta$. Thanks to (7.3.5), for $w \in X^\delta$ we have

$$\sup_{0\neq(\tilde\mu,\tilde\theta,\tilde\nu)\in Y^\delta\times L_2(I\times\omega)\times L_2(\Omega)} \frac{(Bw)(\tilde\mu) + \langle\Gamma_\omega w, \tilde\theta\rangle_{L_2(I\times\omega)} + \varepsilon\langle\gamma_0 w, \tilde\nu\rangle_{L_2(\Omega)}}{\sqrt{\|\tilde\mu\|_Y^2 + \|\tilde\theta\|_{L_2(I\times\omega)}^2 + \|\tilde\nu\|_{L_2(\Omega)}^2}} \eqsim |\!|\!| w |\!|\!|_\varepsilon,$$

so that the so-called Ladyzhenskaya–Babuška–Brezzi condition is satisfied. Therefore, the discretization of the saddle-point system is uniformly stable, so

$$\begin{aligned} &\|\mu_\varepsilon - \mu_\varepsilon^{\delta,\delta}\|_Y + \|\theta_\varepsilon - \theta_\varepsilon^{\delta,\delta}\|_{L_2(I\times\omega)} + \|\nu_\varepsilon - \nu_\varepsilon^{\delta,\delta}\|_{L_2(\Omega)} + |\!|\!| u_\varepsilon - u_\varepsilon^{\delta,\delta} |\!|\!|_\varepsilon \\ &\quad\lesssim \min_{(\tilde\mu,\tilde u)\in Y^\delta\times X^\delta} \|\mu_\varepsilon - \tilde\mu\|_Y + |\!|\!| u_\varepsilon - \tilde u |\!|\!|_\varepsilon \\ &\quad\leq \|\mu_\varepsilon\|_Y + \min_{\tilde u\in X^\delta} |\!|\!| u_\varepsilon - \tilde u |\!|\!|_\varepsilon = \|g - Bu_\varepsilon\|_{Y'} + \min_{\tilde u\in X^\delta} |\!|\!| u_\varepsilon - \tilde u |\!|\!|_\varepsilon. \end{aligned} \tag{7.3.7}$$

From (7.2.7), we have

$$\|u - u_\varepsilon^{\delta,\delta}\|_{X_\eta} \lesssim |||u - u_\varepsilon^{\delta,\delta}|||_\varepsilon \le |||u - u_\varepsilon|||_\varepsilon + |||u_\varepsilon - u_\varepsilon^{\delta,\delta}|||_\varepsilon,$$

where, by (7.3.7),

$$\begin{aligned} |||u_\varepsilon - u_\varepsilon^{\delta,\delta}|||_\varepsilon &\lesssim \|g - Bu_\varepsilon\|_{Y'} + |||u - u_\varepsilon|||_\varepsilon + \min_{\tilde{u}\in X^\delta} |||u - \tilde{u}|||_\varepsilon, \\ &\lesssim \|g - Bu_\varepsilon\|_{Y'} + |||u - u_\varepsilon|||_\varepsilon + e^\delta_{\mathrm{approx}}(u), \end{aligned}$$

where we have used $|||\,|||_\varepsilon \lesssim \|\,\|_X$. By applying a triangle-inequality for the norm $\sqrt{\|\cdot\|^2_{Y'} + \|\cdot\|^2_{L_2(I\times\omega)} + \varepsilon^2\|\cdot\|^2_{L_2(\Omega)}}$, one infers that

$$\begin{aligned} \|g - Bu_\varepsilon\|_{Y'} + |||u - u_\varepsilon|||_\varepsilon &\le \sqrt{G_0(u_\varepsilon)} + \sqrt{G_\varepsilon(u)} + \sqrt{G_\varepsilon(u_\varepsilon)} \le 3\sqrt{G_\varepsilon(u)} \\ &\le 3(e_{\mathrm{cons}}(u) + \varepsilon\|\gamma_0 u\|_{L_2(\Omega)}). \end{aligned}$$

The proof is completed by combining the last three displayed formulas. □

Remark 7.3.10. Let $(X^\delta)_{\delta\in\Delta} = (X^{\delta_n})_{n\in\mathbb{N}}$ be so that $\overline{\cup_n X^{\delta_n}} = X$ and $X^{\delta_n} \subset X^{\delta_{n+1}}$ $(\forall n)$. Let $(Y^{\delta_n})_{n\in\mathbb{N}}$ be a corresponding sequence so that (7.3.5) is valid, $\overline{\cup_n Y^{\delta_n}} = Y$ and $Y^{\delta_n} \subset Y^{\delta_{n+1}}$ $(\forall n)$, and let $(\varepsilon_n)_{n\in\mathbb{N}}$ be so that $\lim_{n\to\infty}\varepsilon_n = 0$. Then $\lim_{n\to\infty} G_0(u_{\varepsilon_n}^{\delta_n,\delta_n}) = \lim_{n\to\infty} G_{\varepsilon_n}(u_{\varepsilon_n}^{\delta_n,\delta_n}) = G_0(u_0) = e_{\mathrm{cons}}(u_0)$. ◊

Proof. For convenience writing $(\delta,\varepsilon) = (\delta_n,\varepsilon_n)$, for $\xi \in \{0,\varepsilon\}$ we write

$$G_0(u_0) - G_\xi(u_\varepsilon^{\delta,\delta}) = G_0(u_0) - G_\xi(u_\varepsilon) + G_\xi(u_\varepsilon) - G_\xi(u_\varepsilon^{\delta,\delta}).$$

Since $\lim_{n\to\infty}\|u_0 - u_\varepsilon\|_X = 0$ by Remark 7.3.7, and hence $\lim_{n\to\infty}\|\mu_0 - \mu_\varepsilon\|_Y = \lim_{n\to\infty}\|B(u_0 - u_\varepsilon)\|_{Y'} = 0$, we have $\lim_{n\to\infty} G_\xi(u_\varepsilon) = G_0(u_0)$. As shown in (7.3.7), it holds that

$$\begin{aligned} |G_\xi(u_\varepsilon) - G_\xi(u_\varepsilon^{\delta,\delta})| &\le |||u_\varepsilon - u_\varepsilon^{\delta,\delta}|||_\varepsilon \lesssim \min_{(\tilde{\mu},\tilde{u})\in Y^\delta\times X^\delta} \|\mu_\varepsilon - \tilde{\mu}\|_Y + |||u_\varepsilon - \tilde{u}|||_\varepsilon \\ &\lesssim \min_{(\tilde{\mu},\tilde{u})\in Y^\delta\times X^\delta} \|\mu_\varepsilon - \tilde{\mu}\|_Y + \|u_\varepsilon - \tilde{u}\|_X \\ &\le \|\mu_0 - \mu_\varepsilon\|_Y + \|u_0 - u_\varepsilon\|_X + \min_{(\tilde{\mu},\tilde{u})\in Y^\delta\times X^\delta} \|\mu_0 - \tilde{\mu}\|_Y + \|u_0 - \tilde{u}\|_X \to 0 \end{aligned}$$

for $n \to \infty$. □

The above results hinge on the validity of (7.3.5). When $A(t) \equiv A$ is a spatial (second order elliptic) differential operator with *constant coefficients* on a *convex* polytopal domain Ω, and X^δ is a *lowest order* finite element space w.r.t. *quasi-uniform* prismatic elements, we will be able to verify in §7.5.2 the inf-sup condition (7.3.5).

Since we are able to show (7.3.5) only under such restrictive conditions on Ω and the trial spaces X^δ, we will consider in Sect. 7.4 a *First Order System*

Least Squares formulation of the data assimilation problem, for which we show a corresponding inf-sup condition in more general situations in §7.5.3.

Stability of the discretization, and hence (7.3.5), is in particular intimately connected with *a posteriori accuracy control*. A well-known tool for establishing (7.3.5) is the identification of suitable Fortin operators which also serve to define appropriate notions of *data oscillation* as discussed next.

7.3.2 Fortin operators, error estimation, and data-oscillation

It is well-known that existence of uniformly bounded Fortin interpolators is a *sufficient* condition for the inf-sup condition (7.3.5) to hold. In the next theorem it is shown that existence of such interpolators is also a necessary condition, and quantitative statements are provided.

Theorem 7.3.11. *Let*

$$Q^\delta \in \mathcal{L}(Y,Y) \text{ with } \operatorname{ran} Q^\delta \subset Y^\delta \text{ and } (BX^\delta)\Big((\mathrm{Id}-Q^\delta)Y\Big) = 0. \tag{7.3.8}$$

Then $\gamma^\delta := \inf_{\{w\in X^\delta:\, Bw\neq 0\}} \sup_{0\neq\mu\in Y^\delta} \frac{(Bw)(\mu)}{\|Bw\|_{Y'}\|\mu\|_Y} \geq \|Q^\delta\|^{-1}_{\mathcal{L}(Y,Y)}$.

Conversely, when $\gamma^\delta > 0$, *then there exists a* Q^δ *as in* (7.3.8), *which is a projector, and* $\|Q^\delta\|_{\mathcal{L}(Y,Y)} \leq 2 + 1/\gamma^\delta$.

Proof. If a Q^δ as in (7.3.8) exists, then for $w \in X^\delta$ it holds that

$$\|Bw\|_{Y'} = \sup_{0\neq\mu\in Y} \frac{(Bw)(\mu)}{\|\mu\|_Y} = \sup_{0\neq\mu\in Y} \frac{(Bw)(Q^\delta\mu)}{\|\mu\|_Y} \leq \|Q^\delta\|_{\mathcal{L}(Y,Y)} \sup_{0\neq\mu^\delta\in Y^\delta} \frac{(Bw)(\mu)}{\|\mu\|_Y},$$

or $\gamma^\delta \geq \|Q^\delta\|^{-1}_{\mathcal{L}(Y,Y)}$.

Now let $\gamma^\delta > 0$. Equipping $X^\delta/\ker B$ with $\|B\cdot\|_{(Y^\delta)'}$, given $\mu \in Y$ consider the problem: find $(\mu^\delta, [w^\delta]) \in Y^\delta \times X^\delta/\ker B$ that solves

$$\left(\begin{bmatrix} R & B \\ B' & 0 \end{bmatrix}\begin{bmatrix} \mu^\delta - \mu \\ [w^\delta] \end{bmatrix}\right)\begin{bmatrix} \tilde\mu^\delta \\ [\tilde w^\delta] \end{bmatrix} = 0 \quad ((\tilde\mu^\delta, [\tilde w^\delta]) \in Y^\delta \times X^\delta/\ker B). \tag{7.3.9}$$

One verifies that $Q^\delta := \mu \mapsto \mu^\delta$ is a projector and satisfies (7.3.8), and so what remains is to bound its norm.

Denoting by $I_Y^\delta\colon Y^\delta \to Y$ and $I_X^\delta\colon X^\delta/\ker B \to X/\ker B$ the trivial embeddings, in operator language the above system reads as

$$\begin{bmatrix} (I_Y^\delta)'RI_Y^\delta & (I_Y^\delta)'BI_X^\delta \\ (I_X^\delta)'B'I_Y^\delta & 0 \end{bmatrix}\begin{bmatrix} \mu^\delta \\ [w^\delta] \end{bmatrix} = \begin{bmatrix} (I_Y^\delta)'R\mu \\ (I_X^\delta)'B'\mu \end{bmatrix}.$$

One verifies that $B^\delta := (I_Y^\delta)'BI_X^\delta\colon X^\delta/\ker B \to (Y^\delta)'$ is an isometry, and furthermore that $R^\delta := (I_Y^\delta)'RI_Y^\delta\colon Y^\delta \to (Y^\delta)'$ is an isometric isomorphism.

Therefore, the Schur complement $S^\delta := {B^\delta}' {R^\delta}^{-1} B^\delta$ is an isometric isomorphism. From

$$\mu^\delta = {R^\delta}^{-1}\Big[(I_Y^\delta)' R\mu + B^\delta {S^\delta}^{-1}\Big((I_X^\delta)' B'\mu - {B^\delta}' {R^\delta}^{-1}(I_Y^\delta)' R\mu\Big)\Big],$$

$\|(I_Y^\delta)' R\|_{\mathcal{L}(Y, {Y^\delta}')} \leq 1$, and

$$\begin{aligned}\|(I_X^\delta)' B'\|_{\mathcal{L}(Y,(X^\delta/\ker B)')} &= \|BI^\delta\|_{\mathcal{L}(X^\delta/\ker B, Y')} \\ &= \sup_{\{w\in X^\delta:\, Bw\neq 0\}} \inf_{0\neq\mu\in Y^\delta} \frac{\|Bw\|_{Y'}\|\mu\|_Y}{(Bw)(\mu)} = 1/\gamma^\delta,\end{aligned}$$

we conclude that $\|\mu^\delta\|_Y \leq (2 + 1/\gamma^\delta)\|\mu\|_Y$ which completes the proof. □

Lemma 7.3.12. *Let $(X^\delta, Y^\delta)_{\delta\in\Delta} \subset X \times Y$ satisfy* (7.3.5)*, and let $(Q^\delta)_{\delta\in\Delta}$ be a corresponding family of uniformly bounded Fortin interpolators as in* (7.3.8)*. With*

$$e^\delta_{\mathrm{osc}}(g) := \|(\mathrm{Id} - {Q^\delta}')g\|_{Y'}{}^2$$

one has for any $\varepsilon \geq 0$ and $w \in X^\delta$,

$$\sqrt{G_\varepsilon(w)} \lesssim \sqrt{G^\delta_\varepsilon(w)} + e^\delta_{\mathrm{osc}}(g).$$

Proof. Thanks to $(\mathrm{Id} - {Q^\delta}')BX^\delta = 0$, the proof follows from

$$\begin{aligned}\|Bw - g\|_{Y'} &\leq \|{Q^\delta}'(Bw - g)\|_{Y'} + \|(\mathrm{Id} - {Q^\delta}')g\|_{Y'} \\ &\leq \|Q^\delta\|_{\mathcal{L}(Y,Y)}\|Bw - g\|_{{Y^\delta}'} + e^\delta_{\mathrm{osc}}(g).\end{aligned}$$ □

Together Proposition 7.3.3 and Lemma 7.3.12 show the following bound.

Corollary 7.3.13. *In the situation of Lemma 7.3.12, one has for $\varepsilon \geq 0$ and $w \in X^\delta$,*

$$\|u_\varepsilon - w\|_{X_\eta} \lesssim \sqrt{G^\delta_\varepsilon(w)} + e^\delta_{\mathrm{osc}}(g).$$

Lemma 7.3.12 can also be used to compute an a posteriori upper bound, modulo data-oscillation, for the minimal consistency error.

Corollary 7.3.14. *Adhering to the setting in Lemma 7.3.12, one has for any $w \in X^\delta$*

$$e_{\mathrm{cons}}(u_0) \lesssim \sqrt{G^\delta_0(w)} + e^\delta_{\mathrm{osc}}(g).$$

Proof. The proof follows from $e_{\mathrm{cons}}(u_0) = \sqrt{G_0(u_0)} \leq \sqrt{G_0(w)}$ and an application of Lemma 7.3.12. □

[2] A similar data-oscillation term appears in [CDG14] within the derivation of a posteriori error estimators for minimal residual methods w.r.t. a dual norm (as our norm on Y').

Using Prop. 7.3.4 or Theorem 7.3.9, this upper bound on $e_{\mathrm{cons}}(u_0)$ narrows the range for appropriate regularization parameters balancing accuracy of the state estimator and the condition of corresponding discrete systems.

In the light of the error bound from Theorem 7.3.9 the above observations hint at further desirable properties of the family $(Y^\delta)_{\delta\in\Delta}$ associated with given trial spaces $(X^\delta)_{\delta\in\Delta}$. Namely, they should permit the construction of uniformly bounded Fortin interpolators Q^δ, as in (7.3.8), for which *in addition*,

$$e^\delta_{\mathrm{osc}}(g) = \mathcal{O}(e^\delta_{\mathrm{approx}}(u)), \text{ or even } e^\delta_{\mathrm{osc}}(g) = o(e^\delta_{\mathrm{approx}}(u)) \tag{7.3.10}$$

hold for sufficiently smooth g. For the model case mentioned at the end of §7.3.1, in §7.5.2 we construct $(Y^\delta)_{\delta\in\Delta}$ satisfying both (7.3.8) and (7.3.10).

7.3.3 Comparisons with the Forward Problem

To show that the solution of the least squares problem

$$\arg\min_{w\in X^\delta} \|Bw-h\|^2_{Y^{\delta\prime}} + \|\gamma_0 w - z_0\|^2_{L_2(\Omega)}$$

is a quasi-best approximation from X^δ to the solution of the *initial-value problem* (7.2.3), the corresponding inf-sup condition reads as

$$\inf_{\delta\in\Delta}\inf_{0\neq w\in X^\delta}\sup_{0\neq(\mu,z)\in Y^\delta\times L_2(\Omega)} \frac{(Bw)(\mu)+\langle\gamma_0 w, z\rangle_{L_2(\Omega)}}{\|w\|_X(\|\mu\|_Y+\|z\|_{L_2(\Omega)})} > 0. \tag{7.3.11}$$

The inf-sup condition (7.3.5) which is relevant for our data-assimilation problem implies (7.3.11). The converse is true when $\gamma_0 w = 0$ for all $w \in X^\delta$. If there is no reason to assume that the target solution u of our data-assimilation problem vanishes at $t = 0$, then however this is not a relevant case.

As shown in Chapter 4, sufficient conditions for (7.3.11) are $X^\delta \subset Y^\delta$ and

$$\inf_{\delta\in\Delta}\inf_{0\neq w\in X^\delta}\sup_{0\neq\mu\in Y^\delta} \frac{(\partial_t w)(\mu)}{\|\partial_t w\|_{Y'}\|\mu\|_Y} > 0$$

This latter inf-sup condition can be realized in far more general discretization settings than we are able to show (7.3.5).

Even for the initial-value problem, a benefit of having (7.3.5), i.e. (7.3.8), is that it gives rise to the efficient and, up to a data-oscillation term, reliable a posteriori error bound

$$\sqrt{\|Bw-h\|^2_{Y^{\delta\prime}} + \|\gamma_0 w - z_0\|^2_{L_2(\Omega)}} \lesssim \|z-w\|_X \lesssim \sqrt{\|Bw-h\|^2_{Y^{\delta\prime}} + \|\gamma_0 w - z_0\|^2_{L_2(\Omega)}} + \mathrm{osc}^\delta(h).$$

where z is the solution of (7.2.3), and w is any element of X^δ.

7.3.4 Numerical Solution of the Discrete Problem

As in Remark 7.3.8, by eliminating the second and third variable from the Galerkin discretization of (7.3.6) with trial space $Y^\delta \times L_2(I\times\omega)\times L_2(\Omega)\times X^\delta$, the minimizer $u_\varepsilon^{\delta,\delta}$ of G_ε^δ over X^δ can be found as the second component of the solution $(\mu_\varepsilon^{\delta,\delta}, u_\varepsilon^{\delta,\delta}) \in Y^\delta \times X^\delta$ of

$$\left(\begin{bmatrix} R & B \\ B' & -(\Gamma_\omega'\Gamma_\omega + \varepsilon^2\gamma_0'\gamma_0)\end{bmatrix}\begin{bmatrix}\mu_\varepsilon^{\delta,\delta}\\ u_\varepsilon^{\delta,\delta}\end{bmatrix} - \begin{bmatrix} g \\ -\Gamma_\omega' f\end{bmatrix}\right)\begin{bmatrix}\tilde\mu\\ \tilde u\end{bmatrix} = 0 \quad ((\tilde\mu,\tilde u)\in Y^\delta\times X^\delta).$$

To solve this system, we need to select bases. Let $\Phi^{Y^\delta} = \{\phi_1^{Y^\delta}, \phi_2^{Y^\delta}, \ldots\}$ and $\Phi^{X^\delta} = \{\phi_1^{X^\delta}, \phi_2^{X^\delta}, \ldots\}$ denote ordered bases, formally viewed as column vectors, for Y^δ and X^δ. Write $\mu_\varepsilon^{\delta,\delta} = (\boldsymbol{\mu}_\varepsilon^{\delta,\delta})^\top \Phi^{Y^\delta}$, $u_\varepsilon^{\delta,\delta} = (\boldsymbol{u}_\varepsilon^{\delta,\delta})^\top \Phi^{X^\delta}$, and define the vectors $\boldsymbol{g}^\delta := g(\Phi^{Y^\delta})$, $\boldsymbol{f}_\omega^\delta := f(\Phi^{X^\delta}|_{I\times\omega})$, the matrices $\boldsymbol{R}^\delta := (R\Phi^{Y^\delta})(\Phi^{Y^\delta}) = [\langle \phi_j^{Y^\delta}, \phi_i^{Y^\delta}\rangle_Y]_{ij}$, $\boldsymbol{B}^\delta := (B\Phi^{X^\delta})(\Phi^{Y^\delta})$, $\boldsymbol{M}_{\Gamma_\omega}^\delta := \langle \Gamma_\omega\Phi^{X^\delta}, \Gamma_\omega\Phi^{X^\delta}\rangle_{L_2(I\times\omega)}$, and $\boldsymbol{M}_{\gamma_0}^\delta := \langle \gamma_0\Phi^{X^\delta}, \gamma_0\Phi^{X^\delta}\rangle_{L_2(\Omega)}$. Then $(\boldsymbol{\mu}_\varepsilon^{\delta,\delta}, \boldsymbol{u}_\varepsilon^{\delta,\delta})$ is the solution of

$$\begin{bmatrix}\boldsymbol{R}^\delta & \boldsymbol{B}^\delta \\ \boldsymbol{B}^{\delta\top} & -(\boldsymbol{M}_{\Gamma_\omega}^\delta + \varepsilon^2\boldsymbol{M}_{\gamma_0}^\delta)\end{bmatrix}\begin{bmatrix}\boldsymbol{\mu}_\varepsilon^{\delta,\delta}\\ \boldsymbol{u}_\varepsilon^{\delta,\delta}\end{bmatrix} = \begin{bmatrix}\boldsymbol{g}^\delta \\ -\boldsymbol{f}_\omega^\delta\end{bmatrix}. \tag{7.3.12}$$

Remark 7.3.15. Using that $\|Bu_\varepsilon^{\delta,\delta} - g\|_{Y^{\delta\prime}} = \|\mu_\varepsilon^{\delta,\delta}\|_Y$, one verifies that for any $\tilde\varepsilon \geq 0$, the a posteriori estimate from Corollary 7.3.13 for the deviation of $u_\varepsilon^{\delta,\delta}$ from $u_{\tilde\varepsilon}$ can be evaluated as

$$\begin{aligned} G_{\tilde\varepsilon}^\delta(u_\varepsilon^{\delta,\delta}) =& \langle \boldsymbol{R}^\delta\boldsymbol{\mu}_\varepsilon^{\delta,\delta}, \boldsymbol{\mu}_\varepsilon^{\delta,\delta}\rangle + \langle \boldsymbol{M}_{\Gamma_\omega}^\delta \boldsymbol{u}_\varepsilon^{\delta,\delta}, \boldsymbol{u}_\varepsilon^{\delta,\delta}\rangle \\ &- 2\langle \boldsymbol{u}_\varepsilon^{\delta,\delta}, \boldsymbol{f}_\omega^\delta\rangle + \|f\|_{L_2(I\times\omega)}^2 + \tilde\varepsilon^2\langle \boldsymbol{M}_{\gamma_0}^\delta \boldsymbol{u}_\varepsilon^{\delta,\delta}, \boldsymbol{u}_\varepsilon^{\delta,\delta}\rangle. \end{aligned}$$

This will later be used in the numerical experiments. ◊

For spatial domains with dimension $d > 1$, the realization of any reasonable accuracy gives rise to system sizes that require resorting to an *iterative solver*. When employing a discretization based on a partition of the time-space cylinder into "time slabs", the availability of a uniformly spectrally equivalent preconditioner $\boldsymbol{K}_Y^\delta \eqsim (\boldsymbol{R}^\delta)^{-1}$ that can be applied at linear cost, is actually a mild assumption.

All properties we have derived for the solution of (7.3.12) remain valid when we replace $\boldsymbol{R}^\delta$ in this system by $(\boldsymbol{K}_Y^\delta)^{-1}$, because this replacement amounts to replacing the Y-norm on Y^δ by an equivalent norm. Therefore, despite this replacement, we continue to denote the solution vector and corresponding function in X^δ by $\boldsymbol{u}_\varepsilon^{\delta,\delta}$ and $u_\varepsilon^{\delta,\delta} = (\boldsymbol{u}_\varepsilon^{\delta,\delta})^\top\Phi^{X^\delta}$, respectively.

To approximate $\boldsymbol{u}_\varepsilon^{\delta,\delta}$ we apply Preconditioned Conjugate Gradients to the Schur complement equation

$$\underbrace{(\boldsymbol{B}^{\delta\top}\boldsymbol{K}_Y^\delta\boldsymbol{B}^\delta + \boldsymbol{M}_{\Gamma_\omega}^\delta + \varepsilon^2\boldsymbol{M}_{\gamma_0}^\delta)}_{\boldsymbol{G}_\varepsilon^\delta :=}\boldsymbol{u}_\varepsilon^{\delta,\delta} = \underbrace{\boldsymbol{f}_\omega^\delta + \boldsymbol{B}^{\delta\top}\boldsymbol{K}_Y^\delta\boldsymbol{g}^\delta}_{\boldsymbol{h}^\delta :=}. \tag{7.3.13}$$

We use a preconditioner $\boldsymbol{K}_X^\delta$ that is the representation of a uniformly boundedly invertible operator $X^{\delta'} \to X^\delta$, with X^δ and $X^{\delta'}$ being equipped with Φ^{X^δ} and the corresponding dual basis. Again, under the time-slab restriction, such preconditioners $\boldsymbol{K}_X^\delta$ of wavelet-in-time, multigrid-in-space type, that can be applied at linear cost, have been constructed in [AT15, SvVW21]. Assuming (7.3.5) (even (7.3.11) suffices), it follows from (7.3.3) that $\lambda_{\max}(\boldsymbol{K}_X^\delta \boldsymbol{G}_\varepsilon^\delta) \lesssim \max(1,\varepsilon^2)$ and $\lambda_{\min}(\boldsymbol{K}_X^\delta \boldsymbol{G}_\varepsilon^\delta) \gtrsim \varepsilon^2$. Consequently, the number of iterations that is sufficient to reduce an initial algebraic error by a factor ρ in the $\|(\boldsymbol{G}_\varepsilon^\delta)^{\frac{1}{2}} \cdot\|$-norm[3] can be bounded by $\lesssim \varepsilon^{-1}\log\rho^{-1}$.

To derive a stopping criterion for the iteration, for $\tilde{\boldsymbol{u}}_\varepsilon^{\delta,\delta} \approx \boldsymbol{u}_\varepsilon^{\delta,\delta}$ let $\boldsymbol{e} := \boldsymbol{u}_\varepsilon^{\delta,\delta} - \tilde{\boldsymbol{u}}_\varepsilon^{\delta,\delta}$, $\boldsymbol{r} := \boldsymbol{h}^\delta - \boldsymbol{G}_\varepsilon^\delta \tilde{\boldsymbol{u}}_\varepsilon^{\delta,\delta}$, $\tilde{u}_\varepsilon^{\delta,\delta} := (\tilde{\boldsymbol{u}}_\varepsilon^{\delta,\delta})^\top \Phi^{X^\delta}$, and the *algebraic error* $e := u_\varepsilon^{\delta,\delta} - \tilde{u}_\varepsilon^{\delta,\delta}$. Then, from (7.3.5) we have that

$$\begin{aligned} G_0^\delta(e) &\le \|Be\|_{Y^{\delta\prime}}^2 + \|\Gamma_\omega e\|_{L_2(I\times\omega)}^2 + \varepsilon^2\|\gamma_0 e\|_{L_2(\Omega)}^2 \\ &= \left\langle (\boldsymbol{B}^{\delta\top}\boldsymbol{R}^{\delta-1}\boldsymbol{B}^\delta + \boldsymbol{M}_{\Gamma_\omega}^\delta + \varepsilon^2\boldsymbol{M}_{\gamma_0}^\delta)\boldsymbol{e}, \boldsymbol{e}\right\rangle \\ &\eqsim \left\langle (\boldsymbol{B}^{\delta\top}\boldsymbol{K}_Y^\delta\boldsymbol{B}^\delta + \boldsymbol{M}_{\Gamma_\omega}^\delta + \varepsilon^2\boldsymbol{M}_{\gamma_0}^\delta)\boldsymbol{e}, \boldsymbol{e}\right\rangle = \langle \boldsymbol{G}_\varepsilon^\delta \boldsymbol{e}, \boldsymbol{e}\rangle. \end{aligned}$$

Moreover, for $\boldsymbol{e} \neq \boldsymbol{0}$, we have[4]

$$\max(1,\varepsilon^2)^{-1} \lesssim \lambda_{\max}(\boldsymbol{K}_X^\delta \boldsymbol{G}_\varepsilon^\delta)^{-1} \le \frac{\langle \boldsymbol{G}_\varepsilon^\delta \boldsymbol{e}, \boldsymbol{e}\rangle}{\langle \boldsymbol{r}, \boldsymbol{K}_X^\delta \boldsymbol{r}\rangle} \le \lambda_{\min}(\boldsymbol{K}_X^\delta \boldsymbol{G}_\varepsilon^\delta)^{-1} \lesssim \varepsilon^{-2}. \quad (7.3.14)$$

Taking u_0 as the reference state, the iteration should ideally be stopped as soon as the algebraic error is dominated by $\|u_0 - u_\varepsilon^{\delta,\delta}\|_{X_\eta}$. Ignoring data-oscillation, as an indication that $\tilde{u}_\varepsilon^{\delta,\delta}$ is indeed close enough to $u_\varepsilon^{\delta,\delta}$ we accept that the respective upper bounds from Corollary 7.3.13 are close enough, i.e., $\sqrt{G_0^\delta(\tilde{u}_\varepsilon^{\delta,\delta})}$ satisfies $\sqrt{G_0^\delta(\tilde{u}_\varepsilon^{\delta,\delta})} \lesssim \sqrt{G_0^\delta(u_\varepsilon^{\delta,\delta})}$. Using $\sqrt{G_0^\delta(\tilde{u}_\varepsilon^{\delta,\delta})} \le \sqrt{G_0^\delta(u_\varepsilon^{\delta,\delta})} + \sqrt{G_0^\delta(e)}$, and the above bound for $G_0^\delta(e)$, we conclude that for the latter to hold true it suffices when for a sufficiently small constant $\mu > 0$,

$$\langle \boldsymbol{r}, \boldsymbol{K}_X^\delta \boldsymbol{r}\rangle \le \mu\varepsilon^2 G_0^\delta(\tilde{u}_\varepsilon^{\delta,\delta}).$$

Since we expect (7.3.14) to be pessimistic, we simply take $\mu = 1$ and thus will stop the iterative solver as soon as $\langle \boldsymbol{r}, \boldsymbol{K}_X^\delta \boldsymbol{r}\rangle \le \varepsilon^2 G_0^\delta(\tilde{u}_\varepsilon^{\delta,\delta})$.

7.4 First order system least squares (FOSLS) form

In view of the difficulty to demonstrate the inf-sup condition (7.3.5) in general settings for the second order weak formulation of the data assimilation prob-

[3] A reduction of the desired factor ρ can be achieved by applying a nested iteration approach.

[4] Instead of the possibly very pessimistic upper bound in (7.3.14), that moreover requires estimating $\lambda_{\min}(\boldsymbol{K}_X^\delta \boldsymbol{G}_\varepsilon^\delta)$, one may consult [GM97, MT13, AK01] for methods to accurately estimate $\langle \boldsymbol{G}_\varepsilon^\delta \boldsymbol{e}, \boldsymbol{e}\rangle$ using data that is obtained in the PCG iteration.

lem, we consider in this section a regularized FOSLS formulation. Its analysis builds to a large extent on the concepts used in Section 7.3.

For $\boldsymbol{b} \in L_\infty(I\times\Omega)^d$, $c \in L_\infty(I\times\Omega)$, and uniformly positive definite $K = K^\top \in L_\infty(I\times\Omega)^{d\times d}$, we consider $a(t;\theta,\zeta)$ as in (7.2.1)–(7.2.2) of the form

$$a(t;\theta,\zeta) = \int_\Omega K\nabla\theta\cdot\nabla\zeta\,\mathrm{d}\boldsymbol{x} + (\boldsymbol{b}\cdot\nabla\theta + c\theta)\zeta\,\mathrm{d}\boldsymbol{x}. \tag{7.4.1}$$

Adhering to the definitions of the spaces X, Y from the previous sections, we abbreviate $Z := L_2(I;L_2(\Omega)^d)$ and define the operator $C \in \mathcal{L}(X\times Z, Y')$ as

$$C(w,\boldsymbol{q})(v) := \int_I\int_\Omega \partial_t w v + \boldsymbol{q}\cdot\nabla_{\boldsymbol{x}} v + (\boldsymbol{b}\cdot\nabla_{\boldsymbol{x}} w + cw)v\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t. \tag{7.4.2}$$

We define the corresponding least squares functional $H_\varepsilon\colon X\times Z\to\mathbb{R}$ as

$$H_\varepsilon(w,\boldsymbol{q}) := \|C(w,\boldsymbol{q}) - g\|_{Y'}^2 + \|\boldsymbol{q} - K\nabla_{\boldsymbol{x}} w\|_Z^2 + \|\Gamma_\omega w - f\|_{L_2(I\times\omega)}^2 + \varepsilon^2\|\gamma_0 w\|_{L_2(\Omega)}^2.$$

The following simple observations allow us to tie the analysis of the corresponding minimization problem to the the concepts of the previous section.

Remark 7.4.1. One has for any $w \in X$

$$C(w, K\nabla_{\boldsymbol{x}} w)(v) = (Bw)(v), \quad v \in Y,$$

and more generally, for any $(w,\boldsymbol{q},\ell) \in X\times Z\times Y'$,

$$C(w,\boldsymbol{q}) - \ell(v) = (Bw)(v) - \ell(v) + \int_{I\times\Omega}(\boldsymbol{q} - K\nabla_{\boldsymbol{x}} w)\cdot\nabla_{\boldsymbol{x}} v\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t.$$

Hence

$$\begin{aligned}\|Bw-\ell\|_{Y'} &\le \|C(w,\boldsymbol{q})-\ell\|_{Y'} + \|v\mapsto\int_I\int_\Omega(\boldsymbol{q}-K\nabla_{\boldsymbol{x}} w)\cdot\nabla_{\boldsymbol{x}} v\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t\|_{Y'}\\ &\le \|C(w,\boldsymbol{q})-\ell\|_{Y'} + \|\boldsymbol{q}-K\nabla_{\boldsymbol{x}} w\|_Z,\end{aligned} \tag{7.4.3}$$

which, with $\ell = g$ in particular implies that

$$H_\varepsilon(w,K\nabla_{\boldsymbol{x}} w) = G_\varepsilon(w) \le 2H_\varepsilon(w,\boldsymbol{q}), \quad \text{for any } \boldsymbol{q}\in Z. \qquad \lozenge \tag{7.4.4}$$

Using (7.4.3), one infers from (7.3.3) that

$$\varepsilon^2 \lesssim \frac{\|C(w,\boldsymbol{q})\|_{Y'}^2 + \|\boldsymbol{q}-K\nabla_{\boldsymbol{x}} w\|_Z^2 + \|\Gamma_\omega w\|_{L_2(I\times\omega)}^2 + \varepsilon^2\|\gamma_0 w\|_{L_2(\Omega)}^2}{\|w\|_X^2 + \|\boldsymbol{q}\|_Z^2} \lesssim \max(1,\varepsilon^2).$$

By an application of the Lax–Milgram Lemma, we conclude that for $\varepsilon > 0$ the minimizer $(\bar{u}_\varepsilon, \boldsymbol{p}_\varepsilon)$ over $X\times Z$ of H_ε exists uniquely, and satisfies

$$\|\bar{u}_\varepsilon\|_X + \|\boldsymbol{p}_\varepsilon\|_Z \lesssim \max(\varepsilon^{-1},1)(\|g\|_{Y'} + \|f\|_{L_2(I\times\omega)}),$$

as well as, for any reference state $u \in X$,

$$\|u - \bar{u}_\varepsilon\|_X + \|K\nabla u - \boldsymbol{p}_\varepsilon\|_Z \lesssim \max(\varepsilon^{-1}, 1)\Big(\varepsilon\|\gamma_0 u\|_{L_2(\Omega)} + e_{\mathrm{cons}}(u)\Big).$$

Again, using (7.2.7), much better bounds will be established for $\|u - \bar{u}_\varepsilon\|_{X_\eta}$.

Remark 7.4.2. Also for $\varepsilon = 0$, the minimizer $(\bar{u}_0, \boldsymbol{p}_0)$ exists uniquely. Indeed, let there be two minimizers of H_0 over $X \times Z$. Then their difference $(e_0, \boldsymbol{e}_0)$ is a homogeneous solution of the corresponding Euler–Lagrange equations, so $\|C(e_0, \boldsymbol{e}_0)\|^2_{Y'} + \|\boldsymbol{e}_0 - K\nabla_{\boldsymbol{x}} e_0\|^2_Z + \|\Gamma_\omega e_0\|^2_{L_2(I\times\omega)} = 0$, and so $\|Be_0\|^2_{Y'} + \|\Gamma_\omega e_0\|^2_{L_2(I\times\omega)} = 0$, which we know implies $e_0 = 0$, and so $\boldsymbol{e}_0 = 0$. ◊

Proposition 7.4.3. *For any $w \in X$, $\boldsymbol{q} \in Z$ one has*

$$\|u - w\|_{X_\eta} \lesssim \sqrt{H_0(w, \boldsymbol{q})} + e_{\mathrm{cons}}(u).$$

In particular, for $\varepsilon \geq 0$, we have the a posteriori bound

$$\|\bar{u}_\varepsilon - w\|_{X_\eta} \lesssim \sqrt{H_\varepsilon(w, \boldsymbol{q})}.$$

Proof. Lemma 7.3.2 gives $\|u - w\|_{X_\eta} \lesssim \sqrt{G_0(w)} + e_{\mathrm{cons}}(u)$, and $G_0(w) \leq 2H_0(w, \boldsymbol{q})$ by (7.4.4). The second result follows from

$$\begin{aligned}\sqrt{H_0(w, \boldsymbol{q})} + e_{\mathrm{cons}}(\bar{u}_\varepsilon) &\leq \sqrt{H_\varepsilon(w, \boldsymbol{q})} + \sqrt{G_\varepsilon(\bar{u}_\varepsilon)} \\ &\leq \sqrt{H_\varepsilon(w, \boldsymbol{q})} + \sqrt{2}\sqrt{H_\varepsilon(\bar{u}_\varepsilon, \boldsymbol{p}_\varepsilon)} \leq (1 + \sqrt{2})\sqrt{H_\varepsilon(w, \boldsymbol{q})}.\end{aligned}$$ □

The same arguments used to show for $\varepsilon \geq 0$ existence and uniqueness of the minimizer $(\bar{u}_\varepsilon, \boldsymbol{p}_\varepsilon)$ of H_ε over $X \times Z$ show for any closed subspace $X^\delta \times Z^\delta \subset X \times Z$ uniqueness of the minimizer $(u^\delta_\varepsilon, \boldsymbol{p}^\delta_\varepsilon)$ of H_ε over $X^\delta \times Z^\delta$. An *a priori bound* for $\|u - \bar{u}^\delta_\varepsilon\|_{X_\eta}$ for an arbitrary reference state $u \in X$ is given in the next proposition.

Proposition 7.4.4. *It holds that*

$$\|u - \bar{u}^\delta_\varepsilon\|_{X_\eta} \lesssim e_{\mathrm{cons}}(u) + \bar{e}^\delta_{\mathrm{approx}}(u) + \varepsilon\|\gamma_0 u\|_{L_2(\Omega)},$$

where $\bar{e}^\delta_{\mathrm{approx}}(u) := \min_{(w,\boldsymbol{q})\in X^\delta \times Z^\delta} \|u - w\|_X + \|K\nabla_{\boldsymbol{x}} u - \boldsymbol{q}\|_Z$.

Proof. Let P_{X^δ} and P_{Z^δ} denote the X- respectively Z-orthogonal projector onto

X^δ respectively Z^δ. Then we have

$$\begin{aligned}
&\sqrt{H_\varepsilon(P_{X^\delta}u, P_{Z^\delta}K\nabla_{\boldsymbol{x}}u)}\\
&\quad\le \|C(P_{X^\delta}u, P_{Z^\delta}K\nabla_{\boldsymbol{x}}u) - C(u, K\nabla_{\boldsymbol{x}}u) + Bu - g\|_{Y'}\\
&\qquad + \|P_{Z^\delta}K\nabla_{\boldsymbol{x}}u - K\nabla_{\boldsymbol{x}}u + K\nabla_{\boldsymbol{x}}u - K\nabla_{\boldsymbol{x}}P_{X^\delta}u\|_Z\\
&\qquad + \|\Gamma_\omega(P_{X^\delta}u - u) + \Gamma_\omega u - f\|_{L_2(I\times\omega)} + \varepsilon\|\gamma_0 P_{X^\delta}u\|_{L_2(\Omega)}\\
&\quad\le \|C(P_{X^\delta}u - u, P_{Z^\delta}K\nabla_{\boldsymbol{x}}u - K\nabla_{\boldsymbol{x}}u)\|_{Y'} + \|Bu - g\|_{Y'}\\
&\qquad + \|P_{Z^\delta}K\nabla_{\boldsymbol{x}}u - K\nabla_{\boldsymbol{x}}u\|_Z + \|K\nabla_{\boldsymbol{x}}(u - P_{X^\delta}u)\|_Z\\
&\qquad + \|\Gamma_\omega(P_{X^\delta}u - u)\|_{L_2(I\times\omega)} + \|\Gamma_\omega u - f\|_{L_2(I\times\omega)}\\
&\qquad + \varepsilon\|\gamma_0(u - P_{X^\delta}u)\|_{L_2(\Omega)} + \varepsilon\|\gamma_0 u\|_{L_2(\Omega)}\\
&\quad\lesssim \varepsilon\|\gamma_0 u\|_{L_2(\Omega)} + e_{\mathrm{cons}}(u) + \bar{e}^\delta_{\mathrm{approx}}(u),
\end{aligned}$$

where we have used $C \in \mathcal{L}(X \times Z, Y')$, $K\nabla_{\boldsymbol{x}} \in \mathcal{L}(Z, X)$, and (7.2.4). Since by Proposition 7.4.3, $\|u - \bar{u}^\delta_\varepsilon\|_{X_\eta} \lesssim \sqrt{H_\varepsilon(\bar{u}^\delta_\varepsilon, \boldsymbol{p}^\delta_\varepsilon)} + e_{\mathrm{cons}}(u)$ and since, by definition, $H_\varepsilon(\bar{u}^\delta_\varepsilon, \boldsymbol{p}^\delta_\varepsilon) \le H_\varepsilon(P_{X^\delta}u, P_{Z^\delta}K\nabla_{\boldsymbol{x}}u)$, the proof is completed. □

Since the definition of H_ε incorporates the dual norm $\|\cdot\|_{Y'}$ neither its minimizer $(\bar{u}^\delta_\varepsilon, \boldsymbol{p}^\delta_\varepsilon)$ over $X^\delta \times Z^\delta$ can be computed, nor the a posteriori error bound from Proposition 7.4.3 can be evaluated. In the next subsection both problems will be tackled by discretizing this dual norm.

Remark 7.4.5. Our FOSLS formulation of the data-assimilation problem has been based on the fact that a well-posed FOSLS formulation of the initial-value problem (7.2.3), with $(A(t)\theta)(\zeta) = a(t;\theta,\zeta)$ of the form (7.4.1), is given by

$$\operatorname*{arg\,min}_{(w,\boldsymbol{q})\in X\times Z} \|C(w,\boldsymbol{q}) - g\|^2_{Y'} + \|\boldsymbol{q} - K\nabla_{\boldsymbol{x}}w\|^2_Z + \|\gamma_0 w - z_0\|^2_{L_2(\Omega)},$$

see [RS18b, Lem. 2.3 and Rem. 2.4]. Notice that with well-posedness we mean that $(w,\boldsymbol{q}) \mapsto (C(w,\boldsymbol{q}), \boldsymbol{q} - K\nabla_{\boldsymbol{x}}w, \gamma_0 w) \in \mathcal{L}\mathrm{is}(X \times Z, Y' \times Z \times L_2(\Omega))$. In the recent work [FK21] it was shown that an alternative well-posed FOSLS formulation for this problem[5] is given by

$$\operatorname*{arg\,min}_{\substack{\{(w,\boldsymbol{q})\in X\times Z:\\ \partial_t w - \mathrm{div}_x\,\boldsymbol{q}\in L_2(I;L_2(\Omega))\}}} \|C(w,\boldsymbol{q}) - g\|^2_{L_2(I;L_2(\Omega))} + \|\boldsymbol{q} - K\nabla_{\boldsymbol{x}}w\|^2_Z + \|\gamma_0 w - z_0\|^2_{L_2(\Omega)}.$$

Applying the latter formulation to the data-assimilation setting would offer the important advantage that there is no need to discretize the dual norm $\|\,\|_{Y'}$. On the other hand, error estimates for such a formulation would be based on the estimate $\|w\|_{X_\eta} \lesssim \|\Gamma_\omega w\|_{L_2(I\times\omega)} + \|Bw\|_{L_2(I;L_2(\Omega))}$, which is a weaker version of the Carleman estimate $\|w\|_{X_\eta} \lesssim \|\Gamma_\omega w\|_{L_2(I\times\omega)} + \|Bw\|_{Y'}$.

[5]The result given in [FK21] for the heat equation immediately generalizes to the more general parabolic problem under consideration, see [GS21]. Surjectivity of $(w,\boldsymbol{q}) \mapsto (C(w,\boldsymbol{q}), \boldsymbol{q} - K\nabla_{\boldsymbol{x}}w, \gamma_0 w)$ has also been shown in the latter work.

Furthermore, in view of an iterative solution process, a likely non-trivial issue is the development of optimal preconditioners for the space $\{(w,\boldsymbol{q}) \in X \times Z\colon \partial_t w - \operatorname{div}_x \boldsymbol{q} \in L_2(I;L_2(\Omega))\}$ equipped with the graph norm. $\Diamond$

7.4.1 Discretizing the dual norm

Given a family of finite dimensional subspaces $(X^\delta \times Z^\delta)_{\delta\in\Delta}$ of $X \times Z$, for each $\delta \in \Delta$ we seek a finite dimensional subspace $\bar{Y}^\delta \subset Y$, with $\dim \bar{Y}^\delta \lesssim \dim X^\delta + \dim Z^\delta$, such that in analogy to (7.3.5)

$$\inf_{\delta\in\Delta} \inf_{0\neq(w,\boldsymbol{q})\in X^\delta\times Z^\delta} \sup_{0\neq\mu\in\bar{Y}^\delta} \frac{C(w,\boldsymbol{q})(\mu)}{\|C(w,\boldsymbol{q})\|_{Y'}\|\mu\|_Y} > 0. \tag{7.4.5}$$

Theorem 7.4.6. *Let* (7.4.5) *be valid. For* $(\bar{u}_\varepsilon^{\delta,\delta}, \boldsymbol{p}_\varepsilon^{\delta,\delta})$ *denoting the (unique) minimizer over* $X^\delta \times Z^\delta$ *of*

$$H_\varepsilon^\delta := (w,\boldsymbol{q}) \mapsto$$
$$\|C(w,\boldsymbol{q}) - g\|^2_{\bar{Y}^{\delta\prime}} + \|\boldsymbol{q} - K\nabla_{\boldsymbol{x}} w\|^2_Z + \|\Gamma_\omega w - f\|^2_{L_2(I\times\omega)} + \varepsilon^2\|\gamma_0 w\|^2_{L_2(\Omega)},$$

it holds that

$$\|u - \bar{u}_\varepsilon^{\delta,\delta}\|_{X_\eta} \lesssim e_{\mathrm{cons}}(u) + \bar{e}^\delta_{\mathrm{approx}}(u) + \varepsilon\|\gamma_0 u\|_X.$$

Proof. Equipping $X \times Z$ with "energy"-norm

$$|\!|\!|(w,\boldsymbol{q})|\!|\!|_\varepsilon := \sqrt{\|C(w,\boldsymbol{q})\|^2_{Y'} + \|\boldsymbol{q} - K\nabla_{\boldsymbol{x}} u\|^2_Z + \|\Gamma_\omega w\|^2_{L_2(I\times\omega)} + \varepsilon^2\|\gamma_0 w\|^2_{L_2(\Omega)}},$$

analogously to the proof of Theorem 7.3.9, in particular following the same reasoning that leads to (7.3.7), one concludes that

$$\begin{aligned}|\!|\!|(\bar{u}_\varepsilon,\boldsymbol{p}_\varepsilon) - (\bar{u}_\varepsilon^{\delta,\delta},\boldsymbol{p}_\varepsilon^{\delta,\delta})|\!|\!|_\varepsilon \lesssim \|g - C(\bar{u}_\varepsilon,\boldsymbol{p}_\varepsilon)\|_{Y'} \\ + \min_{(w,\boldsymbol{q})\in X^\delta\times Z^\delta} |\!|\!|(\bar{u}_\varepsilon,\boldsymbol{p}_\varepsilon) - (w,\boldsymbol{q})|\!|\!|_\varepsilon.\end{aligned} \tag{7.4.6}$$

From (7.2.7), the triangle-inequality, and (7.4.3) we have

$$\begin{aligned}\|u - \bar{u}_\varepsilon^{\delta,\delta}\|_{X_\eta} &\lesssim |\!|\!|u - \bar{u}_\varepsilon|\!|\!|_\varepsilon + |\!|\!|\bar{u}_\varepsilon - \bar{u}_\varepsilon^{\delta,\delta}|\!|\!|_\varepsilon \\ &\leq \sqrt{2}(|\!|\!|(u,K\nabla_{\boldsymbol{x}} u) - (\bar{u}_\varepsilon,\boldsymbol{p}_\varepsilon)|\!|\!|_\varepsilon + |\!|\!|(\bar{u}_\varepsilon,\boldsymbol{p}_\varepsilon) - (\bar{u}_\varepsilon^{\delta,\delta},\boldsymbol{p}_\varepsilon^{\delta,\delta})|\!|\!|_\varepsilon)\end{aligned}$$

From (7.4.6) one infers

$$\begin{aligned}&|\!|\!|(\bar{u}_\varepsilon,\boldsymbol{p}_\varepsilon) - (\bar{u}_\varepsilon^{\delta,\delta},\boldsymbol{p}_\varepsilon^{\delta,\delta})|\!|\!|_\varepsilon \\ &\qquad\lesssim \|g - C(\bar{u}_\varepsilon,\boldsymbol{p}_\varepsilon)\|_{Y'} + |\!|\!|(u,K\nabla_{\boldsymbol{x}} u) - (\bar{u}_\varepsilon,\boldsymbol{p}_\varepsilon)|\!|\!|_\varepsilon + \bar{e}_{\mathrm{approx}}(u),\end{aligned}$$

where we have used that $|\!|\!|\ |\!|\!|_\varepsilon \lesssim \|\ \|_{X\times Z}$. An application of a triangle-inequality for the norm $\sqrt{\|\ \|^2_{Y'} + \|\ \|^2_Z + \|\ \|^2_{L_2(I\times\omega)} + \varepsilon^2\|\ \|^2_{L_2(\Omega)}}$ gives

$$\begin{aligned}&\|g - C(\bar{u}_\varepsilon,\boldsymbol{p}_\varepsilon)\|_{Y'} + |\!|\!|(u,K\nabla_{\boldsymbol{x}} u) - (\bar{u}_\varepsilon,\boldsymbol{p}_\varepsilon)|\!|\!|_\varepsilon \\ &\qquad\leq \sqrt{H_0(\bar{u}_\varepsilon,\boldsymbol{p}_\varepsilon)} + \sqrt{H_\varepsilon(u,K\nabla_{\boldsymbol{x}} u)} + \sqrt{H_\varepsilon(\bar{u}_\varepsilon,\boldsymbol{p}_\varepsilon)} \\ &\qquad\leq 3\sqrt{H_\varepsilon(u,K\nabla_{\boldsymbol{x}} u)} = 3\sqrt{G_\varepsilon(u)} \leq 3(e_{\mathrm{cons}}(u) + \varepsilon\|\gamma_0 u\|_{L_2(\Omega)}).\end{aligned}$$

Combining the last three displayed fomulas completes the proof. □

Similar to Sect. 7.3.2, a necessary and sufficient condition for (7.4.5) to hold is the existence of a family of uniformly bounded Fortin interpolators, i.e.,

$$\bar{Q}^\delta \in \mathcal{L}(Y, \bar{Y}^\delta), \quad (C(X^\delta \times Z^\delta))\Big((\mathrm{Id} - \bar{Q}^\delta)Y\Big) = 0, \quad \sup_{\delta\in\Delta} \|\bar{Q}^\delta\|_{\mathcal{L}(Y,Y)} < \infty. \tag{7.4.7}$$

Similar to Corollary 7.3.13, we have the following a posteriori error bound.

Proposition 7.4.7. *Let $(X^\delta, Z^\delta, \bar{Y}^\delta)_{\delta\in\Delta} \subset X \times Z \times Y$ satisfy* (7.4.5)*, and let $(\bar{Q}^\delta)_{\delta\in\Delta}$ be a corresponding family of Fortin interpolators as in* (7.4.7)*. Then with*

$$\bar{e}^\delta_{\mathrm{osc}}(g) := \|(\mathrm{Id} - \bar{Q}^{\delta\prime})g\|_{Y'},$$

for any $(w, \boldsymbol{q}) \in X^\delta \times Z^\delta$ it holds that

$$\|\bar{u}_\varepsilon - w\|_{X_\eta} \lesssim \sqrt{H^\delta_\varepsilon(w, \boldsymbol{q})} + \bar{e}^\delta_{\mathrm{osc}}(g).$$

Proof. From $\|C(w, \boldsymbol{q}) - g\|_{Y'} \le \|\bar{Q}^\delta\|_{\mathcal{L}(Y,Y)}\|C(w, \boldsymbol{q}) - g\|_{\bar{Y}^{\delta\prime}} + \bar{e}^\delta_{\mathrm{osc}}(g)$ and Proposition 7.4.3 the proof follows. □

Bearing the a priori error bound from Theorem 7.4.6 in mind, this result shows that a desirable additional property of the sequence of spaces $(\bar{Y}^\delta)_{\delta\in\Delta}$, associated with a given sequence of trial spaces $(X^\delta \times Z^\delta)_{\delta\in\Delta}$, gives rise to Fortin interpolators $\bar{Q}^\delta$, as in (7.4.7), warranting for sufficiently smooth g

$$\bar{e}^\delta_{\mathrm{osc}}(g) = \mathcal{O}(\bar{e}^\delta_{\mathrm{approx}}(u)), \text{ or even } \bar{e}^\delta_{\mathrm{osc}}(g) = o(\bar{e}^\delta_{\mathrm{approx}})(u).$$

We conclude by remarking that, in analogy to the second order formulation, condition (7.4.5) is sufficient for the well-posedness of the corresponding forward problem, and gives in addition an a posteriori error bound.

Remark. Concerning the *initial-value problem* (7.2.3), if (7.4.5) holds, then

$$\operatorname*{arg\,min}_{(w,\boldsymbol{q})\in X^\delta\times Z^\delta} \|C(w, \boldsymbol{q}) - h\|^2_{\bar{Y}^{\delta\prime}} + \|\boldsymbol{q} - K\nabla_{\boldsymbol{x}} w\|^2_Z + \|\gamma_0 w - z_0\|^2_{L_2(\Omega)}$$

is a quasi-best approximation to $(z, K\nabla_{\boldsymbol{x}} z) \in X \times Z$ from $X^\delta \times Z^\delta$, and for any $(w, \boldsymbol{q}) \in X^\delta \times Z^\delta$, we have

$$\begin{aligned}
&\sqrt{\|C(w, \boldsymbol{q}) - h\|^2_{\bar{Y}^{\delta\prime}} + \|\boldsymbol{q} - K\nabla_{\boldsymbol{x}} w\|^2_Z + \|\gamma_0 w - z_0\|^2_{L_2(\Omega)}} \\
&\quad\lesssim \|z - w\|_X + \|K\nabla_{\boldsymbol{x}} z - \boldsymbol{q}\|_Z \\
&\quad\lesssim \sqrt{\|C(w, \boldsymbol{q}) - h\|^2_{\bar{Y}^{\delta\prime}} + \|\boldsymbol{q} - K\nabla_{\boldsymbol{x}} w\|^2_Z + \|\gamma_0 w - z_0\|^2_{L_2(\Omega)}} + \bar{e}^\delta_{\mathrm{osc}}(h). \quad \diamondsuit
\end{aligned}$$

7.4.2 Numerical Solution of the Discrete Problem

Recalling the Riesz operator $R \in \mathcal{L}\mathrm{is}(Y, Y')$, we can compute $\bar{u}_\varepsilon^{\delta,\delta}$ as the second component of the solution $(\lambda_\varepsilon^{\delta,\delta}, \bar{u}_\varepsilon^{\delta,\delta}, \boldsymbol{p}_\varepsilon^{\delta,\delta}) \in \bar{Y}^\delta \times X^\delta \times Z^\delta$ of the linear system

$$\left(\begin{bmatrix} R & C_u & C_p \\ C_u' & -(\nabla_x' K^2 \nabla_x + \Gamma_\omega' \Gamma_\omega + \varepsilon^2 \gamma_0' \gamma_0) & \nabla_x' K \\ C_p' & K \nabla_x & -\mathrm{Id} \end{bmatrix} \begin{bmatrix} \lambda_\varepsilon^{\delta,\delta} \\ \bar{u}_\varepsilon^{\delta,\delta} \\ \boldsymbol{p}_\varepsilon^{\delta,\delta} \end{bmatrix} - \begin{bmatrix} g \\ -\Gamma_\omega' f \\ 0 \end{bmatrix} \right) \begin{bmatrix} \tilde{\lambda} \\ \tilde{u} \\ \tilde{\boldsymbol{p}} \end{bmatrix} = 0,$$

$((\tilde{\lambda}, \tilde{u}, \tilde{\boldsymbol{p}}) \in \bar{Y}^\delta \times X^\delta \times Z^\delta)$, where for $C(\cdot,\cdot)$ defined by (7.4.2), $(C_u w)(v) := C(w,0)(v)$ and $(C_p q)(v) := C(0,q)(v)$.

With ordered bases $\Phi^{\bar{Y}^\delta}$, Φ^{X^δ}, and Φ^{Z^δ} for $\bar{Y}^\delta$, X^δ, and Z^δ, and the known or obvious notations $\boldsymbol{\lambda}_\varepsilon^{\delta,\delta}$, $\bar{\boldsymbol{u}}_\varepsilon^{\delta,\delta}$, $\boldsymbol{p}_\varepsilon^{\delta,\delta}$, $\bar{\boldsymbol{g}}^\delta = g(\Phi^{\bar{Y}^\delta})$, $\boldsymbol{f}_\omega^\delta$, $\boldsymbol{R}^\delta$, $\boldsymbol{C}_u^\delta$, $\boldsymbol{C}_p^\delta$, $\boldsymbol{M}_{\Gamma_\omega}^\delta$, and $\boldsymbol{M}_{\gamma_0}^\delta$, and $\boldsymbol{J}^\delta := \langle K\Phi_x^{Z^\delta}, \nabla_x \Phi^{X^\delta} \rangle_{L_2(\Omega)^d}$, $\boldsymbol{L}^\delta := \langle K \nabla_x \Phi^{X^\delta}, K \nabla_x \Phi^{X^\delta} \rangle_{L_2(\Omega)^d}$, and $\boldsymbol{N}^\delta := \langle \Phi^{Z^\delta}, \Phi^{Z^\delta} \rangle_{L_2(\Omega)^d}$, we find $(\boldsymbol{\lambda}_\varepsilon^{\delta,\delta}, \bar{\boldsymbol{u}}_\varepsilon^{\delta,\delta}, \boldsymbol{p}_\varepsilon^{\delta,\delta})$ as the solution of

$$\begin{bmatrix} \boldsymbol{R}^\delta & \boldsymbol{C}_u^\delta & \boldsymbol{C}_p^\delta \\ \boldsymbol{C}_u^{\delta\top} & -(\boldsymbol{L}^\delta + \boldsymbol{M}_{\Gamma_\omega}^\delta + \varepsilon^2 \boldsymbol{M}_{\gamma_0}^\delta) & \boldsymbol{J}^\delta \\ \boldsymbol{C}_p^{\delta\top} & \boldsymbol{J}^{\delta\top} & -\boldsymbol{N}^\delta \end{bmatrix} \begin{bmatrix} \boldsymbol{\lambda}_\varepsilon^{\delta,\delta} \\ \bar{\boldsymbol{u}}_\varepsilon^{\delta,\delta} \\ \boldsymbol{p}_\varepsilon^{\delta,\delta} \end{bmatrix} = \begin{bmatrix} \bar{\boldsymbol{g}}^\delta \\ -\boldsymbol{f}_\omega^\delta \\ 0 \end{bmatrix}. \tag{7.4.8}$$

Similar to Sect. 7.3.4, one expresses the a posteriori error bound $\sqrt{H_\varepsilon(\bar{u}_\varepsilon^{\delta,\delta}, \boldsymbol{p}_\varepsilon^{\delta,\delta})}$ (modulo $\bar{e}_{\mathrm{osc}}^\delta(g)$) in terms of the vectors $\boldsymbol{\lambda}_\varepsilon^{\delta,\delta}$, $\bar{\boldsymbol{u}}_\varepsilon^{\delta,\delta}$, and $\boldsymbol{p}_\varepsilon^{\delta,\delta}$.

As in Sect. 7.3.4, in the above system we replace $\boldsymbol{R}^\delta$ by a uniform preconditioner $(\boldsymbol{K}_Y^\delta)^{-1}$, whilst keeping the notation for the resulting solution vector and corresponding function in $\bar{Y}^\delta \times X^\delta \times Z^\delta$, and apply Preconditioned Conjugate Gradients to the symmetric positive definite Schur complement system

$$\boldsymbol{H}_\varepsilon^\delta \begin{bmatrix} \bar{\boldsymbol{u}}_\varepsilon^{\delta,\delta} \\ \boldsymbol{p}_\varepsilon^{\delta,\delta} \end{bmatrix} = \begin{bmatrix} \boldsymbol{f}_\omega^\delta + \boldsymbol{C}_u^\top \boldsymbol{K}_Y^\delta \bar{\boldsymbol{g}}^\delta \\ \boldsymbol{C}_p^\top \boldsymbol{K}_Y^\delta \bar{\boldsymbol{g}}^\delta \end{bmatrix} \tag{7.4.9}$$

where

$$\boldsymbol{H}_\varepsilon^\delta := \begin{bmatrix} \boldsymbol{L}^\delta + \boldsymbol{M}_{\Gamma_\omega}^\delta + \varepsilon^2 \boldsymbol{M}_{\gamma_0}^\delta + \boldsymbol{C}_u^{\delta\top} \boldsymbol{K}_Y^\delta \boldsymbol{C}_u^\delta & \boldsymbol{C}_u^{\delta\top} \boldsymbol{K}_Y^\delta \boldsymbol{C}_p^\delta - \boldsymbol{J}^\delta \\ \boldsymbol{C}_p^{\delta\top} \boldsymbol{K}_Y^\delta \boldsymbol{C}_u^\delta - \boldsymbol{J}^{\delta\top} & \boldsymbol{N}^\delta + \boldsymbol{C}_p^{\delta\top} \boldsymbol{K}_Y^\delta \boldsymbol{C}_p^\delta \end{bmatrix}.$$

With $\boldsymbol{K}_X^\delta$ from Sect. 7.3.4, and $\boldsymbol{K}_Z^\delta$ being spectrally equivalent to the inverse of the mass matrix of Φ^{Z^δ}, the eigenvalues of the preconditioned system $\begin{bmatrix} \boldsymbol{K}_X^\delta & 0 \\ 0 & \boldsymbol{K}_Z^\delta \end{bmatrix} \boldsymbol{H}_\varepsilon^\delta$ are bounded from above and below, up to constant factors, by $\max(1, \varepsilon^2)$ and ε^2, respectively.

For $\tilde{\boldsymbol{u}}_\varepsilon^{\delta,\delta} \approx \boldsymbol{u}_\varepsilon^{\delta,\delta}$, $\tilde{\boldsymbol{p}}_\varepsilon^{\delta,\delta} \approx \boldsymbol{p}_\varepsilon^{\delta,\delta}$, with $\boldsymbol{e}_u := \boldsymbol{u}_\varepsilon^{\delta,\delta} - \tilde{\boldsymbol{u}}_\varepsilon^{\delta,\delta}$, $\boldsymbol{e}_p := \boldsymbol{p}_\varepsilon^{\delta,\delta} - \tilde{\boldsymbol{p}}_\varepsilon^{\delta,\delta}$, $e_u := (\boldsymbol{e}_u)^\top \Phi^{X^\delta}$, $\boldsymbol{e}_p := (\boldsymbol{e}_p)^\top \Phi^{Z^\delta}$, we apply (7.2.7) and the arguments from the proof

of Proposition 7.4.3 to obtain

$$\begin{aligned}
\|e_u\|_{X_\eta}^2 &\lesssim \|\Gamma_\omega e_u\|_{L_2(I\times\omega)}^2 + \|Be_u\|_{Y'}^2 \\
&\le \|\Gamma_\omega e_u\|_{L_2(I\times\omega)}^2 + \|C(e_u, e_p)\|_{Y'}^2 + \|e_p - K\nabla_x e_u\|_X^2 \\
&\eqsim \|\Gamma_\omega e_u\|_{L_2(I\times\omega)}^2 + \|C(e_u, e_p)\|_{\bar{Y}^{\delta\prime}}^2 + \|e_p - K\nabla_x e_u\|_X^2 \\
&\le \|\Gamma_\omega e_u\|_{L_2(I\times\omega)}^2 + \|C(e_u, e_p)\|_{\bar{Y}^{\delta\prime}}^2 + \|e_p - K\nabla_x e_u\|_X^2 + \varepsilon^2\|\gamma_0 e_u\|_{L_2(\omega)}^2 \\
&\eqsim \left\langle \boldsymbol{H}_\varepsilon^\delta \begin{bmatrix} e_u \\ e_p \end{bmatrix}, \begin{bmatrix} e_u \\ e_p \end{bmatrix} \right\rangle,
\end{aligned}$$

where the last "$\eqsim$"-symbol reads as an equality for $(\boldsymbol{K}_Y^\delta)^{-1} = \boldsymbol{R}^\delta$.

For the residuals $\begin{bmatrix} r_u \\ r_p \end{bmatrix} := \begin{bmatrix} f_\omega^\delta + C_u^\top K_Y^\delta \bar{g}^\delta \\ C_p^\top K_Y^\delta \bar{g}^\delta \end{bmatrix} - \boldsymbol{H}_\varepsilon^\delta \begin{bmatrix} e_u \\ e_p \end{bmatrix}$ it then holds that

$$\max(1,\varepsilon^2)^{-1} \left\langle \begin{bmatrix} r_u \\ r_p \end{bmatrix}, \begin{bmatrix} K_X^\delta r_u \\ K_Z^\delta r_p \end{bmatrix} \right\rangle \lesssim \left\langle \boldsymbol{H}_\varepsilon^\delta \begin{bmatrix} e_u \\ e_p \end{bmatrix}, \begin{bmatrix} e_u \\ e_p \end{bmatrix} \right\rangle \lesssim \varepsilon^{-2} \left\langle \begin{bmatrix} r_u \\ r_p \end{bmatrix}, \begin{bmatrix} K_X^\delta r_u \\ K_Z^\delta r_p \end{bmatrix} \right\rangle,$$

uniformly in δ.

Remark 7.4.8. A reasonable stopping criterion can be determined by the same reasoning as used in Section 7.3.4. Ignoring again data oscillation we use the a posteriori bound from Proposition 7.4.7 to see whether the pair $(\tilde{\boldsymbol{u}}_\varepsilon^{\delta,\delta}, \tilde{\boldsymbol{p}}_\varepsilon^{\delta,\delta})$ is sufficiently close to $(\boldsymbol{u}_0^{\delta,\delta}, \boldsymbol{p}_0^{\delta,\delta})$. Specifically, we stop the iteration as soon as

$$\left\langle \begin{bmatrix} r_u \\ r_p \end{bmatrix}, \begin{bmatrix} K_X^\delta r_u \\ K_Z^\delta r_p \end{bmatrix} \right\rangle \le \varepsilon^2 H_0^\delta(\tilde{\boldsymbol{u}}_\varepsilon^{\delta,\delta}, \tilde{\boldsymbol{p}}_\varepsilon^{\delta,\delta}). \quad \diamondsuit$$

7.5 Construction of a suitable Fortin interpolator

The spaces X^δ and Y^δ, or X^δ, Z^δ and $\bar{Y}^\delta$, that we are going to employ, will be finite element spaces w.r.t. a partition of the time-space cylinder into 'time slabs' with each time-slab being partitioned into prismatic elements. As a preparation for the derivation of a suitable Fortin interpolator for both the standard second order formulation from §7.3 and the first order order formulation from §7.4, we start with constructing certain biorthogonal projectors acting on the spatial domain.

7.5.1 Construction of auxiliary biorthogonal projectors

Let $(\mathcal{T}^\delta)_{\delta\in\Delta}$, $(\mathcal{T}_S^\delta)_{\delta\in\Delta}$ be a families of conforming, uniformly shape regular partitions of $\overline{\Omega} \subset \mathbb{R}^d$ into, say, closed d-simplices, where $\mathcal{T}_S^\delta$ is a refinement of $\mathcal{T}^\delta$ (denoted by $\mathcal{T}^\delta \prec \mathcal{T}_S^\delta$) of some *fixed* maximal depth in the sense that

$|T| \gtrsim |T'|$ for $\mathcal{T}_S^\delta \ni T \subset T' \in \mathcal{T}^\delta$. Thus, one still has $\dim \mathcal{T}_S^\delta \lesssim \dim \mathcal{T}^\delta$. On the other hand, setting

$$\sigma := \sup_{\delta\in\Delta} \sup_{T'\in\mathcal{T}^\delta} \sup_{\{T\in\mathcal{T}_S^\delta:\, T\subset T'\}} \frac{|T|}{|T'|},$$

we assume that σ is sufficiently small, so that refinement is sufficiently fine.

Thanks to the conformity and the uniform shape regularity, for $d > 1$ we know that adjacent $T, T' \in \mathcal{T}^\delta$ (or $\mathcal{T}_S^\delta$) with $T \cap T' \neq \varnothing$ have uniformly comparable sizes. For $d = 1$, we impose this uniform 'K-mesh property' explicitly.

Given a conforming partition $\mathcal{T}$ of $\overline{\Omega}$ into closed d-simplices, we define $\mathcal{S}_\mathcal{T}^{-1,q}$ as the space of all piecewise polynomials of degree q w.r.t. $\mathcal{T}$, and for $q \geq 1$, set $\mathcal{S}_{\mathcal{T},0}^{0,q} := \mathcal{S}_\mathcal{T}^{-1,q} \cap H_0^1(\Omega)$. With $\partial\mathcal{T}$ we denote the mesh skeleton $\cup_{\{T\in\mathcal{T}\}}\partial T$. Next we construct projectors whose range is included in a conforming finite element space of prescribed degree on the refined partition and which vanish on the skeleton of the coarse partition. Moreover, the range of their adjoints contains all piecewise polynomials of the same degree on the coarse partition, as specified next.

Lemma 7.5.1. *Let $q \geq 1$. Then, for a sufficiently small, but* fixed *σ there exists a family of projectors $(P_q^\delta)_{\delta\in\Delta}$ with*

$$\operatorname{ran} P_q^{\delta'} \supseteq \mathcal{S}_{\mathcal{T}^\delta}^{-1,q}, \quad \operatorname{ran} P_q^\delta \subseteq \{w \in \mathcal{S}_{\mathcal{T}_S^\delta,0}^{0,q} : w|_{\partial\mathcal{T}^\delta} = 0\}, \tag{7.5.1}$$

$$\|P_q^\delta w\|_{L_2(T')} \lesssim \|w\|_{L_2(T')} \quad (T' \in \mathcal{T}^\delta,\, w \in L_2(\Omega)). \tag{7.5.2}$$

Proof. Let $T' \in \mathcal{T}^\delta$. Given $p \in \mathcal{P}_q(T')$, let $p_S \in H_0^1(T')$ denote its continuous piecewise polynomial interpolant of degree q w.r.t. to the partition $\mathcal{T}_S^\delta|_{T'}$ using the canonical selection of the interpolation points, where on $\partial T'$ the interpolation values are replaced by zeros.

Obviously, p and p_S coincide on each $T \in \mathcal{T}_S^\delta|_{T'}$ for which $T \cap \partial T' = \varnothing$. Now consider $T \in \mathcal{T}_S^\delta|_{T'}$ with $T \cap \partial T' \neq \varnothing$. Equivalence of norms on finite dimensional spaces, and standard homogeneity arguments show that

$$\|p - p_S\|_{L_2(T)} \eqsim |T|^{\frac{1}{2}}\|p - p_S\|_{L_\infty(T)} \lesssim |T|^{\frac{1}{2}}\|p\|_{L_\infty(T')} \eqsim |T|^{\frac{1}{2}}|T'|^{-\frac{1}{2}}\|p\|_{L_2(T')}.$$

Using the uniform shape regularity of $\mathcal{T}_S^\delta$ and the definition of σ, we arrive at

$$\|p - p_S\|_{L_2(T')}^2 = \sum_{\{T\in\mathcal{T}_S^\delta|_{T'}:\, T\cap\partial T'\neq\varnothing\}} \|p - p_S\|_{L_2(T)}^2 \lesssim \sigma^{1/d}\|p\|_{L_2(T')}^2.$$

From this closeness of p and p_S, one infers that for σ sufficiently small,

$$\inf_{0\neq p\in\mathcal{P}_q(T')} \sup_{0\neq\tilde{p}\in\mathcal{S}_{\mathcal{T}_S^\delta,0}^{0,q}\cap H_0^1(T')} \frac{\langle p, \tilde{p}\rangle_{L_2(T')}}{\|p\|_{L_2(T')}\|\tilde{p}\|_{L_2(T')}} \gtrsim 1, \tag{7.5.3}$$

which implies there is a (uniform) $L_2(T')$-Riesz collection of functions in $\mathcal{S}^{0,q}_{\mathcal{T}^\delta_S,0} \cap H^1_0(T')$ that is biorthogonal to the $L_2(T)$-normalized nodal basis for $\mathcal{P}_q(T')$.

Taking $P_q^{\delta'}$ restricted to T' to be the corresponding biorthogonal projector onto $\mathcal{P}_q(T')$, it has all three stated properties. □

As shown in the above lemma, the projectors P_q^δ exist when $\mathcal{T}^\delta_S$ is a refinement of $\mathcal{T}^\delta$ of sufficient *fixed* depth. Hence, the size of the resulting linear systems remains *uniformly* proportional to $\dim X^\delta$, with a proportionality factor depending on σ. In applications, one needs to know which depth suffices. The usual procedure to construct a partition $\mathcal{T}^\delta$ of the closure of a (polytopal) domain Ω is to recursively apply some fixed 'affine equivalent' refinement rule to each simplex in an initial (conforming) partition of $\overline{\Omega}$. With this approach, the partition of each $T' \in \mathcal{T}^\delta$ formed by its 'descendants' of some fixed generation $\ell \geq 1$ falls into a fixed finite number of classes $\mathcal{T}_{\ell,1}(T'), \ldots, \mathcal{T}_{\ell,N(\ell)}(T')$. By using that the left-hand side of (7.5.3) is invariant under affine transformations, fixing a reference d-simplex T' and a refinement procedure of the above type, given a degree q and a generation ℓ, it suffices to check whether

$$\alpha(q,\ell) := \inf_{1\leq j\leq N(\ell)} \inf_{0\neq p\in\mathcal{P}_q(T')} \sup_{0\neq\tilde{p}\in H^1_0(T')\cap\prod_{T\in\mathcal{T}_{\ell,j}(T')}\mathcal{P}_q(T)} \frac{\langle p,\tilde{p}\rangle_{L_2(T')}}{\|p\|_{L_2(T')}\|\tilde{p}\|_{L_2(T')}} > 0,$$

Remark 7.5.2. For $d \in \{1,2,3\}$, $q \in \{1,2,3,4\}$, and both newest-vertex bisection and red-refinement, we have calculated the minimal ℓ such that $\alpha(q,\ell) > 0$. In all cases but one, this minimal ℓ equals the minimal generation for which $\dim H^1_0(T') \cap \prod_{T\in\mathcal{T}_{\ell,j}(T')} \mathcal{P}_q(T) \geq \dim \mathcal{P}_q(T')$. Only for $d = 3$, $q = 4$, and newest vertex bisection, for one of the three classes it was necessary to increase this generation by one in order to ensure uniform inf-sup stability. ◊

Remark 7.5.3. For the construction of the Fortin interpolator in the FOSLS case, it will be sufficient to replace the conditions (7.5.1)–(7.5.2) on the projectors from Lemma 7.5.1 by the somewhat weaker ones

$$\operatorname{ran} P_q^{\delta'} \supseteq \mathcal{S}^{0,q}_{\mathcal{T}^\delta,0} + \mathcal{S}^{-1,q-1}_{\mathcal{T}^\delta}, \quad \operatorname{ran} P_q^\delta \subseteq \mathcal{S}^{0,q}_{\mathcal{T}^\delta_S,0}, \tag{7.5.4}$$

$$\|\hbar_\delta^{-1} P_q^\delta \hbar_\delta\|_{\mathcal{L}(L_2(\Omega),L_2(\Omega))} \lesssim 1, \tag{7.5.5}$$

where $\hbar_\delta$ is the piecewise constant function defined by $\hbar_\delta|_{T'} = \operatorname{diam} T'$ ($T' \in \mathcal{T}^\delta$). Note that because of the uniform 'K-mesh property', (7.5.5) is implied by local L_2-stability of the form

$$\|P_q^\delta w\|_{L_2(T')} \lesssim \|w\|_{L_2(\{x\in\Omega:\, d(x,T')\lesssim \operatorname{diam} T'\})} \quad (T' \in \mathcal{T}^\delta,\, w \in L_2(\Omega)), \tag{7.5.6}$$

which, in particular, is implied by (7.5.2).

For $d = 1$, the codimension of $\mathcal{S}^{0,q}_{\mathcal{T}^\delta,0} + \mathcal{S}^{-1,q-1}_{\mathcal{T}^\delta}$ in $\mathcal{S}^{-1,q}_{\mathcal{T}^\delta}$ is 1 when $q = 1$, or 0 when $q > 1$. Since we do not expect to benefit from the relaxation of the

condition $\operatorname{ran} P_q^\delta \subseteq \{w \in \mathcal{S}^{0,q}_{\mathcal{T}_S^\delta,0} : w|_{\partial\mathcal{T}^\delta} = 0\}$ to $\operatorname{ran} P_q^\delta \subseteq \mathcal{S}^{0,q}_{\mathcal{T}_S^\delta,0'}$ for $d = 1$, we doubt that the relaxed conditions hold for any less deep refinement $\mathcal{T}_S^\delta$ of $\mathcal{T}^\delta$.

For $d > 1$ and any fixed degree q, however, the aforementioned codimension is $\eqsim \dim \mathcal{S}^{-1,q}_{\mathcal{T}^\delta}$, and we may hope that a less deep refinement $\mathcal{T}_S^\delta$ of $\mathcal{T}^\delta$ suffices to satisfy the relaxed conditions.

So far we have studied this issue in one particular example of $d = 2$, $q = 2$, and the red-refinement rule. For this case, we could show the existence of the projectors from Lemma 7.5.1 when $\mathcal{T}_S^\delta$ is created by applying *two* recursive red-refinements to each triangle from $\mathcal{T}^\delta$. In the appendix, we show that *one* red-refinement already satisfies the relaxed conditions (7.5.4) and (7.5.6). ♢

Remark. When P_q^δ is an L_2-orthogonal projector onto a finite element space, (7.5.5) is known to ensure its H^1-stability (see e.g. [BY14]). ♢

Remark. Spaces of type $\mathcal{S}^{0,q}_{\mathcal{T}^\delta,0} + \mathcal{S}^{-1,q-1}_{\mathcal{T}^\delta}$, or more precisely $\mathcal{S}^{0,q}_{\mathcal{T}^\delta,0} + \mathcal{S}^{-1,0}_{\mathcal{T}^\delta}$, have been used as approximation spaces for the pressure in Stokes solvers to ensure local mass conservation (see e.g. [Che14]). ♢

Remark. Also for the construction of the Fortin interpolator for the standard second order formulation, it suffices when $\operatorname{ran} P_q^{\delta\prime} \supseteq \mathcal{S}^{0,q}_{\mathcal{T}^\delta,0} + \mathcal{S}^{-1,q-1}_{\mathcal{T}^\delta}$ instead of $\operatorname{ran} P_q^{\delta\prime} \supseteq \mathcal{S}^{-1,q}_{\mathcal{T}^\delta,0}$. The second condition in (7.5.1) turns out to be essential. ♢

7.5.2 Standard, second order formulation

For this formulation our construction of a suitable Fortin interpolator will be restricted to second order elliptic spatial differential operators with constant coefficients on convex domains, and *lowest order* finite elements w.r.t. partitions of the time-space cylinder that are *Cartesian products* of a *quasi-uniform* temporal mesh and a *quasi-uniform* conforming, uniformly shape regular spatial mesh into d-simplices.

Consider the families of partitions $(\mathcal{T}^\delta)_{\delta\in\Delta}$ and $(\mathcal{T}_S^\delta)_{\delta\in\Delta}$ of $\overline{\Omega} \subset \mathbb{R}^d$ of §7.5.1. Assuming them to be *quasi-uniform*, we set $h_\delta := \max_{T'\in\mathcal{T}^\delta} \operatorname{diam} T'$ (not to be confused with the piecewise constant function $\hbar_\delta$).

Let $(\mathcal{I}^\delta)_{\delta\in\Delta}$ be a family of *quasi-uniform* partitions of I into subintervals, where the lengths of the subintervals in $\mathcal{I}^\delta$ are $\eqsim h_\delta$. We denote by $\mathcal{S}^{-1,q}_{\mathcal{I}^\delta}$ and $\mathcal{S}^{0,q}_{\mathcal{I}^\delta}$ the space of all piecewise polynomials or continuous piecewise polynomials of degree q w.r.t. $\mathcal{I}^\delta$, respectively.

Theorem 7.5.4. *Let $\Omega \subset \mathbb{R}^d$ be a convex polytope, $a(t;\theta,\zeta)$ be of the form* (7.4.1) *for* constant *K, $\boldsymbol{b}$ and c, and let $X^\delta := \mathcal{S}^{0,1}_{\mathcal{I}^\delta} \otimes \mathcal{S}^{0,1}_{\mathcal{T}^\delta,0} \subset X$ and $Y^\delta := \mathcal{S}^{-1,1}_{\mathcal{I}^\delta} \otimes \mathcal{S}^{0,1}_{\mathcal{T}_S^\delta,0} \subset Y$, where $\mathcal{T}_S^\delta$ is a sufficiently deep refinement of $\mathcal{T}^\delta$ such that a projector P_1^δ as in Lemma 7.5.1 exists. Then, a Fortin interpolator Q^δ as in* (7.3.8) *exists, and for $g \in F := L_2(I) \otimes H^1(\Omega) \cap H^2(I) \otimes H^{-1}(\Omega)$, it holds that $e^\delta_{\mathrm{osc}}(g) \lesssim h_\delta^2$.*

Remark. For this $(X^\delta)_{\delta\in\Delta}$, and a sufficiently smooth u we have $e^\delta_{\mathrm{approx}}(u) \lesssim h_\delta$ where in general an approximation error of higher cannot be expected. So indeed, $e^\delta_{\mathrm{osc}}(g)$ is of higher order as desired, cf. (7.3.10). $\Diamond$

Proof. We will construct uniformly bounded $Q^\delta_t \in \mathcal{L}(L_2(I), L_2(I))$ and $Q^\delta_{\boldsymbol{x}} \in \mathcal{L}(H^1_0(\Omega), H^1_0(\Omega))$ with $\operatorname{ran} Q^\delta_t \subset \mathcal{S}^{-1,1}_{\mathcal{I}^\delta}$, $\operatorname{ran} Q^\delta_{\boldsymbol{x}} \subset \mathcal{S}^{0,1}_{\mathcal{T}^\delta_S,0}$ and

$$\begin{aligned}\Big\langle \mathcal{S}^{0,1}_{\mathcal{I}^\delta}, \operatorname{ran}(\mathrm{Id}-Q^\delta_t)\Big\rangle_{L_2(I)} &= 0 = \Big\langle \tfrac{d}{dt}\mathcal{S}^{0,1}_{\mathcal{I}^\delta}, \operatorname{ran}(\mathrm{Id}-Q^\delta_t)\Big\rangle_{L_2(I)},\\ \Big\langle \mathcal{S}^{-1,0}_{\mathcal{T}^\delta,0}+\mathcal{S}^{0,1}_{\mathcal{T}^\delta,0}, \operatorname{ran}(\mathrm{Id}-Q^\delta_{\boldsymbol{x}})\Big\rangle_{L_2(\Omega)} &= 0 = \Big\langle K\nabla_{\boldsymbol{x}}\mathcal{S}^{0,1}_{\mathcal{T}^\delta,0}, \nabla_{\boldsymbol{x}}\operatorname{ran}(\mathrm{Id}-Q^\delta_{\boldsymbol{x}})\Big\rangle_{L_2(\Omega)^d}.\end{aligned} \tag{7.5.7}$$

Then one verifies that $Q^\delta := Q^\delta_t \otimes Q^\delta_{\boldsymbol{x}}$ satisfies the conditions in (7.3.8).

A valid choice for Q^δ_t is given by the $L_2(I)$-orthogonal projector onto $\mathcal{S}^{-1,1}_{\mathcal{I}^\delta}$. It satisfies in addition

$$\|(\mathrm{Id}-Q^\delta_t)'\|_{\mathcal{L}(H^2(I),L_2(I))} \lesssim h_\delta^2. \tag{7.5.8}$$

We seek $Q^\delta_{\boldsymbol{x}}$ as $Q^{A,\delta}_{\boldsymbol{x}} + Q^{B,\delta}_{\boldsymbol{x}} + Q^{B,\delta}_{\boldsymbol{x}}Q^{A,\delta}_{\boldsymbol{x}}$ with $\operatorname{ran} Q^{A,\delta}_{\boldsymbol{x}}, \operatorname{ran} Q^{B,\delta}_{\boldsymbol{x}} \subset \mathcal{S}^{0,1}_{\mathcal{T}^\delta_S,0}$ s.t.

$$\|Q^{A,\delta}_{\boldsymbol{x}}\|_{\mathcal{L}(H^1_0(\Omega),H^1_0(\Omega))} \lesssim 1, \quad \|\mathrm{Id}-Q^{A,\delta}_{\boldsymbol{x}}\|_{\mathcal{L}(H^1_0(\Omega),L_2(\Omega))} \lesssim h_\delta, \tag{7.5.9}$$

$$\Big\langle K\nabla_{\boldsymbol{x}}\mathcal{S}^{0,1}_{\mathcal{T}^\delta,0}, \nabla_{\boldsymbol{x}}\operatorname{ran}(\mathrm{Id}-Q^{A,\delta}_{\boldsymbol{x}})\Big\rangle_{L_2(\Omega)^d} = 0, \tag{7.5.10}$$

$$\|Q^{B,\delta}_{\boldsymbol{x}}\|_{\mathcal{L}(L_2(\Omega),L_2(\Omega))} \lesssim 1, \quad \|(\mathrm{Id}-Q^{B,\delta}_{\boldsymbol{x}})'\|_{\mathcal{L}(H^1(\Omega),L_2(\Omega))} \lesssim h_\delta, \tag{7.5.11}$$

$$\Big\langle \mathcal{S}^{-1,0}_{\mathcal{T}^\delta,0}+\mathcal{S}^{0,1}_{\mathcal{T}^\delta,0}, \operatorname{ran}(\mathrm{Id}-Q^{B,\delta}_{\boldsymbol{x}})\Big\rangle_{L_2(\Omega)} = 0 = \Big\langle K\nabla_{\boldsymbol{x}}\mathcal{S}^{0,1}_{\mathcal{T}^\delta,0}, \nabla_{\boldsymbol{x}}\operatorname{ran} Q^{B,\delta}_{\boldsymbol{x}}\Big\rangle_{L_2(\Omega)^d}. \tag{7.5.12}$$

One easily verifies that

$$\mathrm{Id}-Q^\delta_{\boldsymbol{x}} = (\mathrm{Id}-Q^{B,\delta}_{\boldsymbol{x}})(\mathrm{Id}-Q^{A,\delta}_{\boldsymbol{x}}),$$

which, together with the first relation in (7.5.12), yields the first relation in (7.5.7). Moreover, from (7.5.10) and the second relation in (7.5.12) one deduces the second relation (7.5.7).

Similarly, observing that $Q^\delta_{\boldsymbol{x}} = Q^{A,\delta}_{\boldsymbol{x}} + Q^{B,\delta}_{\boldsymbol{x}}(\mathrm{Id}-Q^{A,\delta}_{\boldsymbol{x}})$ in combination with (7.5.9), $\|Q^{B,\delta}_{\boldsymbol{x}}\|_{\mathcal{L}(L_2(\Omega),L_2(\Omega))} \lesssim 1$, and the inverse inequality $\|\,\|_{H^1(\Omega)} \lesssim h_\delta^{-1}\|\,\|_{L_2(\Omega)}$ on $\mathcal{S}^{0,1}_{\mathcal{T}^\delta_S,0} \supset \operatorname{ran} Q^{B,\delta}_{\boldsymbol{x}}$, one infers that $\|Q^\delta_{\boldsymbol{x}}\|_{\mathcal{L}(H^1_0(\Omega),H^1_0(\Omega))} \lesssim 1$. Thus, all claimed properties of $Q^\delta_{\boldsymbol{x}}$ have been verified.

Before turning to $Q^{A,\delta}_{\boldsymbol{x}}$ and $Q^{B,\delta}_{\boldsymbol{x}}$, we estimate $e^\delta_{\mathrm{osc}}(g)$. For $g \in F$, we have

$$\|(\mathrm{Id}-Q^\delta)'g\|_{Y'} \le \|(\mathrm{Id}-Q^\delta)'\|_{\mathcal{L}(F,Y')}\|g\|_F = \|\mathrm{Id}-Q^\delta\|_{\mathcal{L}(Y,F')}\|g\|_F.$$

Writing

$$\mathrm{Id}-Q^\delta = (\mathrm{Id}\otimes(\mathrm{Id}-Q^\delta_{\boldsymbol{x}}))(Q^\delta_t\otimes\mathrm{Id}) + (\mathrm{Id}-Q^\delta_t)\otimes\mathrm{Id},$$

from $L_2(I)\otimes H^1(\Omega)' \hookrightarrow F'$, $H^2(I)'\otimes H^1_0(\Omega)\hookrightarrow F'$, and $\|Q_t^\delta\|_{\mathcal{L}(L_2(I),L_2(I))}\lesssim 1$ we infer

$$\begin{aligned}\|\mathrm{Id}-Q^\delta\|_{\mathcal{L}(Y,F')} &\lesssim \|\mathrm{Id}\otimes(\mathrm{Id}-Q_x^\delta)\|_{\mathcal{L}(Y,L_2(I)\otimes H^1(\Omega)')}\\ &\qquad +\|(\mathrm{Id}-Q_t^\delta)\otimes\mathrm{Id}\|_{\mathcal{L}(Y,H^2(I)'\otimes H^1_0(\Omega))}\\ &= \|\mathrm{Id}-Q_x^\delta\|_{\mathcal{L}(H^1_0(\Omega),H^1(\Omega)')}+\|\mathrm{Id}-Q_t^\delta\|_{\mathcal{L}(L_2(I),H^2(I)')}\\ &\le \|\mathrm{Id}-Q_x^{B,\delta}\|_{\mathcal{L}(L_2(\Omega),H^1(\Omega)')}\|\mathrm{Id}-Q_x^{A,\delta}\|_{\mathcal{L}(H^1_0(\Omega),L_2(\Omega))}\\ &\qquad +\|(\mathrm{Id}-Q_t^\delta)'\|_{\mathcal{L}(H^2(I),L_2(I))}\\ &\lesssim h_\delta h_\delta+h_\delta^2,\end{aligned}$$

where we have used (7.5.8), (7.5.9), and (7.5.11).

We now identify the operators $Q_x^{A,\delta}, Q_x^{B,\delta}$. For $Q_x^{A,\delta}$, we take the 'Galerkin' projector onto $\mathcal{S}^{0,1}_{\mathcal{T}^\delta,0}$, i.e. the orthogonal projector w.r.t. $\langle K\nabla_x\cdot,\nabla_x\cdot\rangle_{L_2(\Omega)^d}$. It satisfies (7.5.10), and $\|Q_x^{A,\delta}\|_{\mathcal{L}(H^1_0(\Omega),H^1_0(\Omega))}=1$.

Thanks to Ω being a convex polytope, the homogeneous Dirichlet problem with operator $-\operatorname{div} K\nabla$ is H^2-regular. Indeed, by making a linear coordinate transformation that transforms the convex polytope into another convex polytope ([Ash15]), the operator reads as $-\Delta$ for which the regularity result is well-known. Consequently the usual Aubin–Nitsche duality argument shows

$$\|(\mathrm{Id}-Q_x^{A,\delta})v\|_{L_2(\Omega)}\lesssim h_\delta\|\nabla(\mathrm{Id}-Q_x^{A,\delta})v\|_{L_2(\Omega)^d}\le h_\delta\|\nabla v\|_{L_2(\Omega)^d}$$

for $v\in H^1_0(\Omega)$. This verifies the validity of (7.5.9).

Next, we take $Q_x^{B,\delta}=P_1^\delta$ as constructed in Lemma 7.5.1. It satisfies $\operatorname{ran}P_1^\delta\subset\mathcal{S}^{0,1}_{\mathcal{T}_S^\delta,0}$, $\|P_1^\delta\|_{\mathcal{L}(L_2(\Omega),L_2(\Omega))}\lesssim 1$, and $\operatorname{ran}P_1^{\delta\prime}\supseteq\mathcal{S}^{-1,1}_{\mathcal{T}^\delta}$. The last property shows the first condition in (7.5.12). Using the uniform boundedness, one concludes

$$\|(\mathrm{Id}-P_1^\delta)'w\|_{L_2(\Omega)}\lesssim\inf_{v\in\mathcal{S}^{-1,1}_{\mathcal{T}^\delta}}\|w-v\|_{L_2(\Omega)}\lesssim h_\delta|w|_{H^1(\Omega)},$$

which is the second condition in (7.5.11). The second condition in (7.5.12) follows by an element-wise integration-by-parts from $w|_{\partial\mathcal{T}^\delta}=0$ for any $w\in\operatorname{ran}P_1^\delta$, and the fact that $\mathcal{S}^{0,1}_{\mathcal{T}^\delta,0}$ is a space of continuous piecewise *linears*.[6] □

7.5.3 FOSLS formulation

We construct a suitable Fortin interpolator for the FOSLS formulation of our data assimilation problem. In contrast to the standard second order formulation, we allow now non-convex domains Ω, higher order finite element spaces w.r.t. possibly non-quasi-uniform partitions into prismatic elements. However, the time-space cylinder must be partitioned into time slabs.

[6] This argument is the sole reason the theorem is restricted to lowest order trial spaces X^δ.

Theorem 7.5.5. *As in Theorem 7.5.4, let $a(t;\theta,\zeta)$ be of the form* (7.4.1) *for* constant *K, $\boldsymbol{b}$ and c. For $(I^\delta = (([t_i^\delta, t_{i+1}^\delta])_i)_{\delta\in\Delta}$ being a family of partitions of I, we consider $(X^\delta)_{\delta\in\Delta}$, $(Z^\delta)_{\delta\in\Delta}$, and $(Y^\delta)_{\delta\in\Delta}$ that satisfy*

$$X^\delta \subseteq \{w \in C(I; H_0^1(\Omega)) \colon w|_{(t_i^\delta, t_{i+1}^\delta)} \in \mathcal{P}_q(t_i^\delta, t_{i+1}^\delta) \otimes \mathcal{S}^{0,q}_{\mathcal{T}^{\delta_i},0}\}, \tag{7.5.13}$$

$$Y \supseteq \bar{Y}^\delta \supseteq \{v \in Y \colon v|_{(t_i^\delta, t_{i+1}^\delta)} \in \mathcal{P}_q(t_i^\delta, t_{i+1}^\delta) \otimes \mathcal{S}^{0,q}_{\mathcal{T}_S^{\delta_i},0}\},^7$$

$$Z^\delta \subseteq \{\boldsymbol{q} \in L_2(I; H(\mathrm{div};\Omega)) \colon \boldsymbol{q}|_{(t_i^\delta, t_{i+1}^\delta)} \in \mathcal{P}_{q-1}(t_i^\delta, t_{i+1}^\delta) \otimes \mathcal{Z}^q_{\mathcal{T}^{\delta_i}}\}, \tag{7.5.14}$$

where $\mathrm{div}\, \mathcal{Z}^q_{\mathcal{T}^{\delta_i}} \subset \mathcal{S}^{-1,q-1}_{\mathcal{T}^{\delta_i}}$, *and where for each i, $\mathcal{T}^{\delta_i}$ is some partition from $(\mathcal{T}^\delta)_{\delta\in\Delta}$ with corresponding refinement $\mathcal{T}_S^{\delta_i} \in (\mathcal{T}_S^\delta)_{\delta\in\Delta}$.*

Then for $\mathcal{T}_S^\delta$ being a sufficiently deep refinement of $\mathcal{T}^\delta$ such that a projector P_q^δ as in Remark 7.5.3 exists, a Fortin interpolator $\bar{Q}^\delta$ as in (7.4.7) *exists, and*

$$(\bar{e}^\delta_{\mathrm{osc}}(g))^2 \lesssim \sum_i \sum_{T'\in\mathcal{T}^{\delta_i}} \Big\{ \inf_{p\in\mathcal{P}_q(t_i^\delta,t_{i+1}^\delta)\otimes L_2(T')} \|g-p\|^2_{L_2((t_i^\delta,t_{i+1}^\delta)\times T')} + (\mathrm{diam}\, T')^2 \inf_{p\in L_2(t_i^\delta,t_{i+1}^\delta)\otimes\mathcal{P}_{q-1}(T')} \|g-p\|^2_{L_2((t_i^\delta,t_{i+1}^\delta)\times T')} \Big\}.$$

Remark 7.5.6. In view of balancing the approximation rates for smooth functions by X^δ in X and Z^δ in Z, for X^δ and Z^δ being the spaces on the right-hand side of (7.5.13) or (7.5.14) (in the latter case, possibly with $\boldsymbol{q} \in L_2(I; H(\mathrm{div};\Omega))$ reading as $\boldsymbol{q} \in C(I; H(\mathrm{div};\Omega))$), a natural choice for $\mathcal{Z}^q_{\mathcal{T}^\delta}$ is the Raviart–Thomas space of index q or the Brezzi–Douglas–Marini finite element space of index $\min(1, q-1)$ w.r.t. $\mathcal{T}^\delta$.

Notice that with these definitions of X^δ and Z^δ, for sufficiently smooth g the local oscillation error is of higher order than the expected local approximation error by X^δ in X and Z^δ in Z. ◊

Proof. For $(w,\boldsymbol{q}) \in X^\delta \times Z^\delta$, $v \in Y$, taking $\mathcal{Z}^q_{\mathcal{T}^{\delta_i}} \subset H(\mathrm{div};\Omega)$ into account, integration-by-parts shows

$$C(w,\boldsymbol{q})(v) = \int_I \int_\Omega \Big(\tfrac{\partial w}{\partial t} - \mathrm{div}_{\boldsymbol{x}}\, \boldsymbol{q} + \boldsymbol{b}\cdot\nabla_{\boldsymbol{x}} w + cw\Big) v \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t.$$

Now let $(\bar{Q}^\delta_{\boldsymbol{x}})_{\delta\in\Delta}$ denote a family of operators $\bar{Q}^\delta_{\boldsymbol{x}} \in \mathcal{L}(H_0^1(\Omega), H_0^1(\Omega))$ with uniformly bounded norm, with the properties

$$\mathrm{ran}\, \bar{Q}^\delta_{\boldsymbol{x}} \subset \mathcal{S}^{0,q}_{\mathcal{T}_S^\delta,0}, \quad \langle \mathcal{S}^{0,q}_{\mathcal{T}^\delta,0} + \mathcal{S}^{-1,q-1}_{\mathcal{T}^\delta}, \mathrm{ran}(\mathrm{Id} - \bar{Q}^\delta_{\boldsymbol{x}})\rangle_{L_2(\Omega)} = 0. \tag{7.5.15}$$

Moreover, let Q_q^i be the $L_2(I)$-orthogonal projector onto $\mathcal{P}_q(t_i^\delta, t_{i+1}^\delta)$. Then, the operator $\bar{Q}^\delta$ defined by

$$(\bar{Q}^\delta v)|_{(t_i^\delta,t_{i+1}^\delta)\times\Omega} = (Q_q^i \otimes \bar{Q}^{\delta_i}_{\boldsymbol{x}}) v|_{(t_i^\delta,t_{i+1}^\delta)\times\Omega},$$

[7] If it were not for guaranteeing an oscillation error of higher order, then the polynomial degree in the time direction could be reduced to $q-1$.

satisfies the conditions of (7.4.7).

We again seek $\bar{Q}_x^\delta$ of the form $\bar{Q}_x^\delta = \bar{Q}_x^{A,\delta} + \bar{Q}_x^{B,\delta} + \bar{Q}_x^{B,\delta}\bar{Q}_x^{A,\delta}$ where

$$\operatorname{ran}\bar{Q}_x^{A,\delta}, \operatorname{ran}\bar{Q}_x^{B,\delta} \subset \mathcal{S}^{0,q}_{\mathcal{T}_S^\delta,0}, \quad \Big\langle \mathcal{S}^{0,q}_{\mathcal{T}^\delta,0} + \mathcal{S}^{-1,q-1}_{\mathcal{T}^\delta}, \operatorname{ran}(\mathrm{Id} - \bar{Q}_x^{B,\delta})\Big\rangle_{L_2(\Omega)} = 0.$$

Then from $\mathrm{Id} - \bar{Q}_x^\delta = (\mathrm{Id} - \bar{Q}_x^{B,\delta})(\mathrm{Id} - \bar{Q}_x^{A,\delta})$, we infer that (7.5.15) is satisfied.

We take $\bar{Q}_x^{A,\delta}$ to be the Scott–Zhang quasi-interpolator onto $\mathcal{S}^{0,q}_{\mathcal{T}_S^\delta,0}$, and $\bar{Q}_x^{B,\delta} = P_q^\delta$ from Remark 7.5.3. Writing $\bar{Q}_x^\delta = \bar{Q}_x^{A,\delta} + P_q^\delta(\mathrm{Id} - \bar{Q}_x^{A,\delta})$, uniform boundedness of $\bar{Q}_x^{A,\delta} \in \mathcal{L}(H_0^1(\Omega), H_0^1(\Omega))$, $\hbar_\delta^{-1}(\mathrm{Id} - \bar{Q}_x^{A,\delta}) \in \mathcal{L}(H_0^1(\Omega), L_2(\Omega))$, as well as $\hbar_\delta^{-1}P_q^\delta\hbar_\delta \in \mathcal{L}(L_2(\Omega), L_2(\Omega))$, and $\|\cdot\|_{H^1(\Omega)} \lesssim \|h_\delta^{-1}\cdot\|_{L_2(\Omega)}$ on $\mathcal{S}^{0,q}_{\mathcal{T}_S^\delta,0}$, imply the uniform boundedness of $\bar{Q}_x^\delta \in \mathcal{L}(H_0^1(\Omega), H_0^1(\Omega))$.

For $p_i \in \mathcal{P}_q(t_i^\delta, t_{i+1}^\delta) \otimes L_2(\Omega)$, $\tilde{p}_i \in L_2(t_i^\delta, t_{i+1}^\delta) \otimes \mathcal{S}^{-1,q-1}_{\mathcal{T}^{\delta_i}}$, and $y \in Y$, we have

$$\begin{aligned}\Big\langle g, (\mathrm{Id} - \bar{Q}^\delta)y\Big\rangle_{L_2(I\times\Omega)} &= \sum_i \sum_{T'\in\mathcal{T}^{\delta_i}} \Big\langle ((\mathrm{Id} - Q_q^i)\otimes \mathrm{Id})(g - p_i), y\Big\rangle_{L_2((t_i^\delta, t_{i+1}^\delta)\times T')} \\ &\quad + \sum_i \sum_{T'\in\mathcal{T}^{\delta_i}} \Big\langle (\mathrm{Id}\otimes(\mathrm{Id} - P_q^\delta)')(g - \tilde{p}_i), Q_q^i \otimes (\mathrm{Id} - \bar{Q}_x^{A,\delta_i})y\Big\rangle_{L_2((t_i^\delta, t_{i+1}^\delta)\times T')},\end{aligned}$$

since $\mathcal{P}_q(t_i^\delta, t_{i+1}^\delta)$ is reproduced by Q_q^i, and $\mathcal{S}^{-1,q-1}_{\mathcal{T}_S^{\delta_i}}$ by $P_q^{\delta\prime}$. The first double sum is bounded by a constant multiple of

$$\sqrt{\textstyle\sum_i \sum_{T'\in\mathcal{T}^{\delta_i}} \|g - p_i\|^2_{L_2((t_i^\delta, t_{i+1}^\delta)\times T')}}\, \|y\|_{L_2(I\times\Omega)}.$$

On account of $\|\hbar_\delta P_q^{\delta\prime}\hbar_\delta^{-1}\|_{\mathcal{L}(L_2(\Omega),L_2(\Omega))} \lesssim 1$ and

$$\|(\mathrm{Id} - \bar{Q}_x^{A,\delta_i})v\|_{L_2(T')} \le (\operatorname{diam} T')|v|_{H^1(\cup_{\{T''\in\mathcal{T}^{\delta_i}\,:\,T''\cap T'\neq\varnothing\}} T'')},$$

one sees that the second double sum is bounded by a constant multiple of $\sqrt{\sum_i \sum_{T'\in\mathcal{T}^{\delta_i}} (\operatorname{diam} T')^2 \|g - \tilde{p}_i\|^2_{L_2((t_i^\delta, t_{i+1}^\delta)\times T')}}\, \|y\|_Y$, showing the result. □

7.6 Numerical experiments

We investigate our formulations for solving the data assimilation problem numerically. As underlying parabolic equation we select a simple heat equation posed on a spatial domain $\Omega \subset \mathbb{R}^d$, and we take $T = 1$, i.e. $I = [0, 1]$.

We use *NGSolve*, [Sch97, Sch14], to assemble the system matrices and for spatial multigrid. We employ a preconditioned conjugate gradient scheme for solving the corresponding Schur complement systems (7.3.13) from §7.3.4 and (7.4.9) from §7.4.2.

7.6.1 Unit interval

We start with the simplest possible situation where $d = 1$, and $\Omega := [0,1]$. We subdivide I and Ω into $1/h_\delta \in \mathbb{N}$ equal subintervals yielding $\mathcal{I}^\delta$ and $\mathcal{T}^\delta$ respectively. We then select our discrete spaces as tensor-product spaces

$$X^\delta := \mathcal{S}^{0,1}_{\mathcal{I}^\delta} \otimes \mathcal{S}^{0,1}_{\mathcal{T}^\delta,0},\ \ Y^\delta_\ell := \bar{Y}^\delta_\ell := \mathcal{S}^{-1,1}_{\mathcal{I}^\delta} \otimes \mathcal{S}^{0,1}_{\mathcal{T}^\delta_\ell,0},\ \ Z^\delta := \mathcal{S}^{-1,0}_{\mathcal{I}^\delta} \otimes \mathcal{S}^{0,1}_{\mathcal{T}^\delta} \quad (7.6.1)$$

with $\mathcal{T}^\delta_\ell$ found from $\mathcal{T}^\delta$ by recursively bisecting every subinterval ℓ times.

As follows from Sect. 7.5, in our current setting, for both second order and FOSLS formulation, for $\ell \geq 2$ uniformly bounded Fortin interpolators exist, i.e., (7.3.8) or (7.4.7) are satisfied, so that the minimizers $u^{\delta,\delta}_\varepsilon \in X^\delta$ and $(\bar{u}^{\delta,\delta}_\varepsilon, \boldsymbol{p}^{\delta,\delta}_\varepsilon) \in X^\delta \times Z^\delta$ of G^δ_ε or H^δ_ε exist uniquely, and satisfy the a priori bounds from Theorem 7.3.9 or Theorem 7.4.6, as well as the a posteriori bounds from Corollary 7.3.13 and Proposition 7.4.7. Moreover, these Fortin interpolators can be selected such that for sufficiently smooth datum g the order of the data-oscillation term $e^\delta_{\mathrm{osc}}(g)$ or $\bar{e}^\delta_{\mathrm{osc}}(g)$, that are present in the a posteriori bounds, exceeds the generally best possible approximation order that can be expected. Consequently, for $\ell = 2$ the expressions $\sqrt{G^\delta_0(u^{\delta,\delta}_\varepsilon)}$ or $\sqrt{H^\delta_0(\bar{u}^{\delta,\delta}_\varepsilon, \boldsymbol{p}^{\delta,\delta}_\varepsilon)}$ are, modulo a constant factor and oscillation terms of higher order, upper bounds for the X_η-norm of $e^\delta_\varepsilon := u_0 - u^{\delta,\delta}_\varepsilon$ or $\bar{e}^\delta_\varepsilon := u_0 - \bar{u}^{\delta,\delta}_\varepsilon$.

We will use this fact to explore in subsequent experiments also whether it would actually be harmful in practice to take $\ell<2$ (resulting in lower computational cost). Note that the choice of the refinement level ℓ in Y^δ_ℓ or $\bar{Y}^\delta_\ell$ affects, on the one hand, the quality of the numerical solution $u^{\delta,\delta}_\varepsilon$ and, on the other hand, the reliability of the a posteriori error bound. We will denote below by ℓ the refinement level used to compute $u^{\delta,\delta}_\varepsilon$, and by L the refinement level in Y^δ_L or $\bar{Y}^\delta_L$ used to compute the a posteriori error bounds. Since these 'reliable' a posteriori error bounds with $L = 2$ apply to any function from X^δ (taking for the second argument of H^δ_ε any argument from Z^δ), we have also used them, in particular, to assess the quality of the numerical approximations based on taking Y^δ_0 or $\bar{Y}^\delta_0$ instead of Y^δ_2 or $\bar{Y}^\delta_2$.

Equipping $\mathcal{S}^{-1,1}_{\mathcal{I}^\delta}$ with basis Φ^δ_t, and $\mathcal{S}^{0,1}_{\mathcal{T}^\delta_\ell,0}$ with $\Phi^\delta_{\boldsymbol{x}}$, the representation of the Riesz isometry $Y^\delta \to Y^{\delta\prime}$ reads as $\boldsymbol{R}^\delta = \langle \Phi^\delta_t, \Phi^\delta_t \rangle_{L_2(I)} \otimes \langle \nabla\Phi^\delta_{\boldsymbol{x}}, \nabla\Phi^\delta_{\boldsymbol{x}} \rangle_{L_2(\Omega)^d}$. Taking Φ^δ_t to be $L_2(I)$-orthogonal, the first factor is diagonal and can be inverted directly. With $\boldsymbol{M}G^\delta_{\boldsymbol{x}} \eqsim \langle \nabla\Phi^\delta_{\boldsymbol{x}}, \nabla\Phi^\delta_{\boldsymbol{x}} \rangle^{-1}_{L_2(\Omega)^d}$ a symmetric spatial multigrid solver, we define $\boldsymbol{K}^\delta_Y := \langle \Phi^\delta_t, \Phi^\delta_t \rangle^{-1}_{L_2(I)} \otimes \boldsymbol{M}G^\delta_{\boldsymbol{x}} \eqsim (\boldsymbol{R}^\delta)^{-1}$, which can be applied at linear cost. As explained in §7.3.4 and §7.4.2, all considerations concerning the discrete approximations $u^{\delta,\delta}_\varepsilon$ or $(\bar{u}^{\delta,\delta}_\varepsilon, \boldsymbol{p}^{\delta,\delta}_\varepsilon)$ remain valid when $\boldsymbol{R}^\delta$ in the matrix vector systems (7.3.12) or (7.4.8), that define these approximations, is replaced by $(\boldsymbol{K}^\delta_Y)^{-1}$, and despite this replacement we continue to denote them by $u^{\delta,\delta}_\varepsilon$ and $(\bar{u}^{\delta,\delta}_\varepsilon, \boldsymbol{p}^{\delta,\delta}_\varepsilon)$.

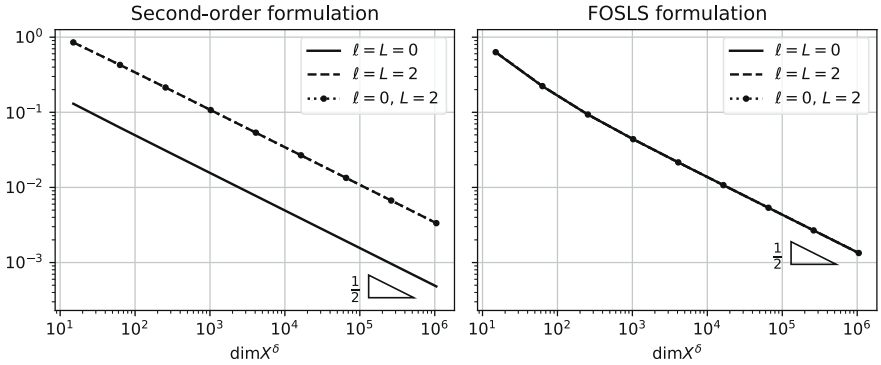


Figure 7.1 Estimated errors for the unit interval problem with consistent data.

Equipping X^δ and Z^δ with similar tensor product bases, for the efficient iterative solution of the Schur complements (7.3.13) or (7.4.9) that define $u_\varepsilon^{\delta,\delta}$ or $(\bar{u}_\varepsilon^{\delta,\delta}, \boldsymbol{p}_\varepsilon^{\delta,\delta})$, for $W \in \{X, Z\}$ we use a preconditioner $\boldsymbol{K}_W^\delta$ that can be applied at linear cost and that is uniformly spectrally equivalent to the inverse of the representation of the Riesz isometry $W^\delta \to W^{\delta\prime}$. The construction of $\boldsymbol{K}_Z^\delta$ does not pose any difficulties, and for the construction of $\boldsymbol{K}_X^\delta$, that builds on a symmetric spatial multigrid solver that is robust for diffusion-reaction problems and the use of a wavelet basis in time that is stable in $L_2(I)$ and $H^1(I)$, we refer to [SvVW21].

Consistent data As a first test we prescribe the solution

$$u(t,x) = (t^3 + 1)\sin(\pi x), \tag{7.6.2}$$

take $\omega = [\frac{1}{4}, \frac{3}{4}]$ and use data (g, f) that are *consistent* with u. We computed $u_\varepsilon^{\delta,\delta}$ and $(\bar{u}_\varepsilon^{\delta,\delta}, \boldsymbol{p}_\varepsilon^{\delta,\delta})$ for $\varepsilon = h_\delta$, being the largest value of ε (up to a constant factor) for which we expect that the regularization doesn't spoil the order of convergence. Indeed, we expect that $e^\delta_{\mathrm{approx}}(u) \eqsim \bar{e}^\delta_{\mathrm{approx}}(u) \eqsim h_\delta$. For both the second order and the FOSLS formulation, Figure 7.1 depicts the a posteriori error estimators $\sqrt{G_0^\delta(u_\varepsilon^{\delta,\delta})}$ and $\sqrt{H_0^\delta(\bar{u}_\varepsilon^{\delta,\delta}, \boldsymbol{p}_\varepsilon^{\delta,\delta})}$ for $(\ell, L) \in \{(2,2),(0,2),(0,0)\}$ as a function of $\dim X^\delta \approx h_\delta^{-2}$. The two formulations show very similar performance. Moreover, the observed convergence rate $1/2$ is the best possible given our discretization of piecewise linears on uniform meshes. Concerning the choices for ℓ and L, the results for $L = 2$, that give reliable a posteriori error bounds, indicate that there is hardly any difference in the numerical approximations for test spaces Y_2^δ or Y_0^δ, respectively, $\bar{Y}_2^\delta$ or $\bar{Y}_0^\delta$, i.e., for $\ell = 2$ or $\ell = 0$, so that we will take $\ell = 0$ in the sequel. For the second order formulation, the value of the a posteriori estimator evaluated for $L = 0$ is significantly smaller than that for $L = 2$, but it shows qualitatively the same behaviour. In view of this observation, we will also use $L = 0$ in what follows.

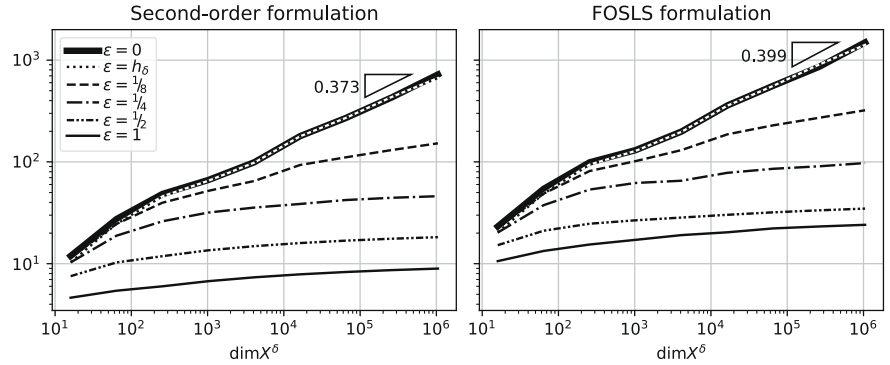


Figure 7.2 Condition numbers of the preconditioned system for a number of regularization parameters.

System conditioning To see how the choice of ε affects the condition number of the preconditioned systems (7.3.13) and (7.4.9), we computed these condition numbers for various ε and decreasing mesh sizes. The results depicted in Figure 7.2 illustrate that for fixed $\varepsilon > 0$, the condition numbers are uniformly bounded. We show the values for $\ell = 2$; for $\ell = 0$, they are very similar. It also reveals that the growth in terms of ε is far more modest than the upper bound $\eqsim \varepsilon^{-2}$ on the condition numbers we found in Sect. 7.3.4 and 7.4.2.

Inconsistent data In case of inconsistent data, there exists no state that exactly explains the data, and $e_{\mathrm{cons}}(u_0) > 0$. In this case, it does not make sense to approximate u_0 within a tolerance that is significantly smaller than $e_{\mathrm{cons}}(u_0)$. Considering for the second order formulation the a priori estimate

$$\|u_0 - u_\varepsilon^{\delta,\delta}\|_{X_\eta} \lesssim e_{\mathrm{cons}}(u_0) + e^\delta_{\mathrm{approx}}(u_0) + \varepsilon\|\gamma_0 u_0\|_{L_2(\Omega)}$$

from Theorem 7.3.9 and taking the fact into account that choosing ε small has an only moderate effect on the conditioning of the preconditioned linear system, in the following we take ε of the order of the best possible approximation error that can be expected, so that $\varepsilon\|\gamma_0 u_0\|_{L_2(\Omega)} \lesssim e^\delta_{\mathrm{approx}}(u_0)$. Then ideally we would like to stop refining our mesh as soon as $e^\delta_{\mathrm{approx}}(u_0) \eqsim e_{\mathrm{cons}}(u_0)$. In order to achieve this we use the a posteriori error estimator. From Corollary 7.3.14 we know that

$$e_{\mathrm{cons}}(u_0) \lesssim \sqrt{G_0^\delta(u_\varepsilon^{\delta,\delta})} + e^\delta_{\mathrm{osc}}(g),$$

where, following the reasoning from the proof of Proposition 7.3.4,

$$\begin{aligned}\sqrt{G_0^\delta(u_\varepsilon^{\delta,\delta})} \le \sqrt{G_\varepsilon^\delta(u_\varepsilon^{\delta,\delta})} \le \sqrt{G_\varepsilon^\delta(P_{X^\delta}u_0)} \le \sqrt{G_\varepsilon(P_{X^\delta}u_0)} \\ \lesssim e^\delta_{\mathrm{approx}}(u_0) + e_{\mathrm{cons}}(u_0) + \varepsilon\|\gamma_0 u\|_{L_2(\Omega)}.\end{aligned}$$

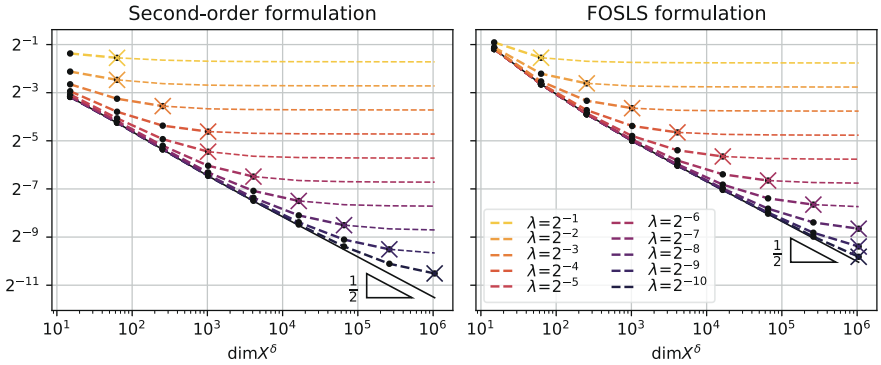


Figure 7.3 Estimated errors for the interval problem for varying inconsistent data.

We selected $(Y^\delta)_{\delta\in\Delta}$ such that, in any case for sufficiently smooth g, the order of $e^\delta_{\mathrm{osc}}(g)$ is equal or higher than the generally best possible order of the approximation error, so that $e^\delta_{\mathrm{osc}}(g) \lesssim e^\delta_{\mathrm{approx}}(u_0)$. In view of our earlier assumption on ε, we conclude that

$$e_{\mathrm{cons}}(u_0) \lesssim \sqrt{G^\delta_0(u^{\delta,\delta}_\varepsilon)} \lesssim e_{\mathrm{cons}}(u_0) + e^\delta_{\mathrm{approx}}(u_0).$$

Exploiting a common uniform or adaptive refinement strategy, it can be expected that $e^\delta_{\mathrm{approx}}(u_0)$ decays with a certain algebraic rate $\rho < 1$. Unless $e_{\mathrm{cons}}(u_0)$ is very large, it can therefore be expected that in the early stage of the iteration the a posteriori error estimator $\sqrt{G^\delta_0(u^{\delta,\delta}_\varepsilon)}$ decays with this rate, whose value therefore can be monitored. By contrast, as soon as $e^\delta_{\mathrm{approx}}(u_0)$ has been reduced to $Ce_{\mathrm{cons}}(u_0)$ for some constant $C > 0$, the reduction of $\sqrt{G^\delta_0(u^{\delta,\delta}_\varepsilon)}$ in the next step cannot be expected to be better than $\frac{1+C\rho}{1+C}$. Taking $C = 1/3$, our strategy will therefore be to stop the iteration as soon as the observed reduction of $\sqrt{G^\delta_0(u^{\delta,\delta}_\varepsilon)}$ is worse than $\frac{1+C\rho}{1+C}$.

We have implemented this strategy, and a similar one for the FOSLS formulation, where we apply the discrete spaces as in (7.6.1), take $\varepsilon = h_\delta$, and again consider the unit interval problem (7.6.2) but now perturb the measured state $f = u|_{I\times\omega}$ by adding $\lambda\mathbb{1}$ to it for various values of λ.

From the results in Figure 7.3 we see that the error estimators decrease at first, but then stagnate in the aforementioned sense, at which point we exit the refinement loop (indicated by a $\times$-sign). Further refinement (indicated by the thin dashed lines) is not very useful, and the error estimators stabilize to a value just below $\lambda|\omega|^{1/2}$, being the $L_2(I\times\Omega)$-norm of the perturbation we added to the consistent f. Knowing that the error estimator converges to $e_{\mathrm{cons}}(u_0)$ (see Remark 7.3.10), we conclude that $(0,\mathbb{1}) \in Y' \times L_2(I\times\omega)$ is close to orthogonal to ran B_ω. We note that $\ell = L = 2$ produces very similar results.

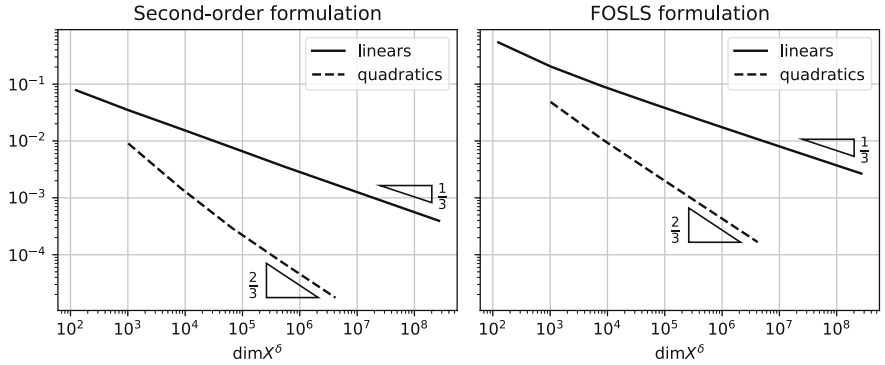


Figure 7.4 Estimated errors for the consistent unit square problem with $\omega := [\frac{1}{4}, \frac{3}{4}]^2$ using piecewise linears and quadratics.

7.6.2 Unit square

We choose $\Omega := (0,1)^2$. We again subdivide I into $1/h_\delta \in \mathbb{N}$ equal subintervals yielding $\mathcal{I}^\delta$, and Ω first into $1/h_\delta \times 1/h_\delta$ squares and then into $2/h_\delta^2$ triangles by connecting the lower left and the upper right corner in each square yielding $\mathcal{T}^\delta$. For a polynomial degree q, we take $\varepsilon = h_\delta^q$. Following the discussion in Sect. 7.6.1, we select $\ell = L = 0$ and take our discrete spaces as

$$X_q^\delta := \mathcal{S}^{0,q}_{\mathcal{I}^\delta} \otimes \mathcal{S}^{0,q}_{\mathcal{T}^\delta,0}, \quad Y_q^\delta := \mathcal{S}^{-1,q}_{\mathcal{I}^\delta} \otimes \mathcal{S}^{0,q}_{\mathcal{T}^\delta,0}, \quad \bar{Y}_q^\delta := \mathcal{S}^{-1,q-1}_{\mathcal{I}^\delta} \otimes \mathcal{S}^{0,q}_{\mathcal{T}^\delta,0},$$

and $Z_q^\delta := \mathcal{S}^{-1,q-1}_{\mathcal{I}^\delta} \otimes \mathcal{Z}^q_{\mathcal{T}^\delta}$, with $\mathcal{Z}^q_{\mathcal{T}^\delta}$ the BDM space of index $\min(1, q-1)$. Note that the degree $q-1$ in the temporal direction of $\bar{Y}_q^\delta$ guarantees an oscillation error of the same order as the approximation error, cf. Footnote 7.

We define the preconditioners $\boldsymbol{K}_Y^\delta$, $\boldsymbol{K}_Z^\delta$, and $\boldsymbol{K}_X^\delta$ similar as in the 1D case.

Consistent data We select $\omega := [\frac{1}{4}, \frac{3}{4}]^2$ with prescribed solution $u(t,x,y) := (t^3+1)\sin(\pi x)\sin(\pi y)$ and consistent data (g,f). Figure 7.4 shows for both formulations and $q \in \{1,2\}$ the error estimators as a function of $\dim X^\delta \eqsim h_\delta^{-3}$. The choice of preconditioners allows us to reach the desired tolerance $\langle \boldsymbol{r}, \boldsymbol{K}_X^\delta \boldsymbol{r}\rangle \leq \varepsilon^2 G_0^\delta(\tilde{u}_\varepsilon^{\delta,\delta}) = 5.937 \cdot 10^{-13}$ for a system with 268 434 945 unknowns in only 96 iterations. The two formulations again exhibit similar performance, and the observed rate $q/3$ is the best possible, in line with Theorem 7.5.5. Moreover, we see that while theory is incomplete for the second order formulation in practice it works well also for piecewise quadratics.

Thanks to $X_\eta \hookrightarrow C([\eta,1], L_2(\Omega))$, the time-slice errors $\|e_\varepsilon^\delta(t)\|_{L_2(\Omega)}$ and $\|\bar{e}_\varepsilon^\delta(t)\|_{L_2(\Omega)}$ are bounded by multiples of $\|e_\varepsilon^\delta\|_{X_\eta}$ or $\|\bar{e}_\varepsilon^\delta\|_{X_\eta}$, respectively. Figure 7.5 shows these time-slice errors for both formulations using piecewise linears, i.e., $q = 1$. We see that for both formulations, the time-slice errors converge with the better rate $2/3$, and that these errors deteriorate for $t \searrow 0$.

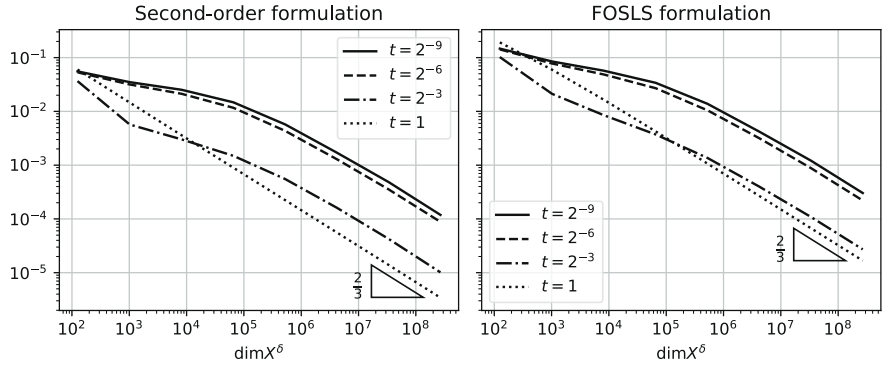


Figure 7.5 Time-slice errors for the consistent unit square problem with $\omega := [\frac{1}{4}, \frac{3}{4}]^2$ using piecewise linears.

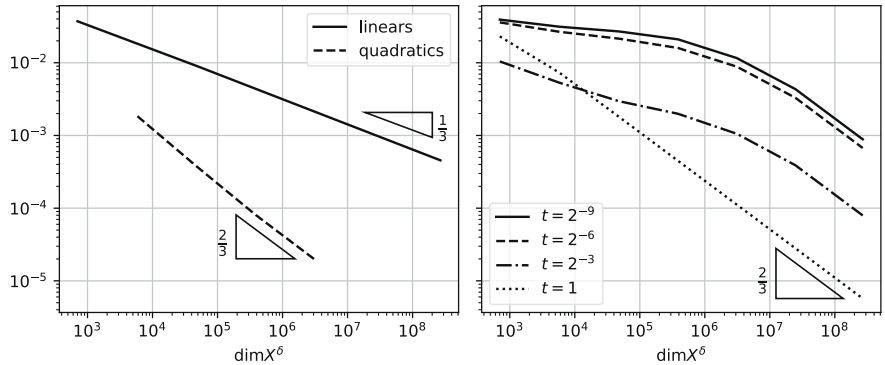


Figure 7.6 Second order formulation for the consistent unit square problem with $\omega := [\frac{7}{16}, \frac{9}{16}]^2$. Left: error estimators; right: time-slice errors.

This deterioration becomes much stronger when $\operatorname{diam}\omega \to 0$: taking for example $\omega := [\frac{7}{16}, \frac{9}{16}]^2$, Figure 7.6 shows that while the error estimators remain nearly unchanged, the time-slice errors fan-out an order of magnitude more than in the case of $\omega := [\frac{1}{4}, \frac{3}{4}]^2$.

Inconsistent data Finally, we return to the case of inconsistent observational data. Again taking $u(t,x,y) := (t^3+1)\sin(\pi x)\sin(\pi y)$ and $\omega := [\frac{1}{4}, \frac{3}{4}]^2$, we select consistent forcing data $g := Bu$ but *perturbed* observational data $f := u|_{I\times\omega} + \lambda\mathbb{1}$. Running the strategy outlined in Sect. 7.6.1 with $C = 1/3$, with uniform refinements and choosing $\varepsilon = h_\delta$, yields the results of Figure 7.7. We see a situation very similar to the unit interval case: the error estimators decrease at first and then stagnate, at which point we exit the refinement loop. Error estimators again stabilize at around $\lambda|\omega|^{1/2}$.

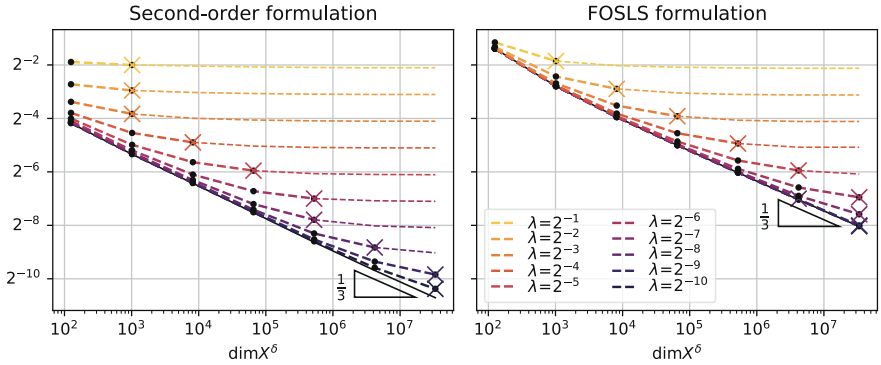


Figure 7.7 Estimated errors for the square problem for varying inconsistent data.

7.7 Concluding remarks

We saw that basing data assimilation for parabolic problems on time-space formulations that are stable at a continuous level, and related regularized least squares functionals, has a number of conceptual advantages: one obtains improved a priori error estimates as well as a posteriori error bounds. Among other things the latter ones are important for determining suitable stopping criteria for iterative solvers. Moreover, the design of corresponding preconditioners is based on the infinite-dimensional variational formulation. We showed that for each fixed regularization parameter ε the preconditioner is optimal relative to the condition of the regularized problem so that the numerical complexity remains under control. Moreover, the regularization parameter is disentangled from the discretizations, so one can optimize its choice.

Furthermore, it will be interesting to relate the present results to the recent state estimation concepts in [BCD$^+$11, CDD$^+$19, MPPY15] providing error bounds in the full energy norm $\|\cdot\|_X$ at the expense of certain stability factors reflecting a geometric relation between X and a certain space of functionals providing the data which, in turn, quantifies the "visibility" of the true states by the sensors. A further important issue is to explore the use of the obtained "static" methods for "dynamic data assimilation". In this context the underlying stable variational formulations are expected to be crucial for the use of certified reduced models.

A price for building on the above "natural" variational formulations—in the sense that no *excess regularity* is implied—is to properly discretize dual norms. As pointed out earlier in Remark 7.4.5, this is avoided in [FK21] by replacing the term $\|C(w,\boldsymbol{q})-\tilde{g}\|^2_{Y'}$ in $H_\varepsilon(w;\boldsymbol{q})$ (for $K=\mathrm{Id}$) by the L_2-residual $\|C(w,\boldsymbol{q})-\tilde{g}\|^2_{L_2(I;L_2(\Omega))}$. Being reduced to using then a somewhat weaker version of the Carleman estimate, we would obtain a statement similar to that in Corollary 7.4.4, but with an approximation error $\bar{e}^\delta_{\mathrm{approx}}(u)$ measured in a

somewhat stronger norm

$$\min_{\substack{\{(w,\boldsymbol{q})\in X^\delta\times Z^\delta:\\ \partial_t w-\mathrm{div}_{\boldsymbol{x}}\,\boldsymbol{q}\in L_2(I;L_2(\Omega))\}}} \|u-w\|_X+\|\nabla_{\boldsymbol{x}}u-\boldsymbol{q}\|_Z+\|\partial_t u-\triangle u-(\partial_t w-\mathrm{div}_{\boldsymbol{x}}\,\boldsymbol{q})\|_{L_2(I;L_2(\Omega))}.$$

Finally, optimal preconditioning in the space $\{(w,\boldsymbol{q})\in X^\delta\times Z^\delta\colon \partial_t w-\mathrm{div}_{\boldsymbol{x}}\,\boldsymbol{q}\in L_2(I;L_2(\Omega))\}$, equipped with the graph norm, is a challenge.

On the other hand, we also have the standard, second order formulation whose implementation is cheaper, and at least in the above experiments performs well also in cases beyond the regime so far covered by theory.

7.A Construction of the biorthogonal projector as in Remark 7.5.3 for $d = q = 2$ and one red-refinement

A basis for $\mathcal{S}^{0,2}_{\mathcal{T}^\delta,0}+\mathcal{S}^{-1,1}_{\mathcal{T}^\delta}$ is given by the sum of the union over $T'\in\mathcal{T}^\delta$ of the usual nodal basis for $P_1(T')$, and the union over the internal edges of $\mathcal{T}^\delta$ of the continuous piecewise quadratic bubble associated to that edge, whose support extends to the two neighbouring triangles in $\mathcal{T}^\delta$. Indeed, one easily verifies that this set of functions is linearly independent, and that each function from either $\mathcal{S}^{0,2}_{\mathcal{T}^\delta,0}$ or $\mathcal{S}^{-1,1}_{\mathcal{T}^\delta}$ is in its span.

We consider the restriction of this basis to one $T'\in\mathcal{T}^\delta$, and subsequently transfer it into a collection of functions on a 'reference triangle' $\hat{T}$ with $|\hat{T}|=1$ by an affine transformation. We denote the resulting functions as indicated in the left of Figure 7.8.

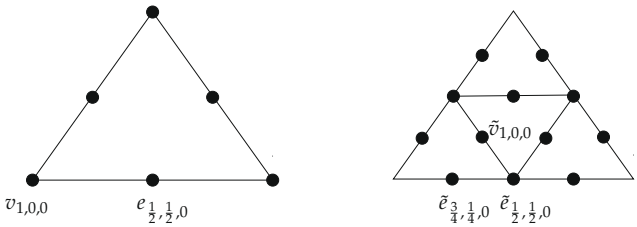


Figure 7.8 Left: Basis functions at the 'primal side' with the indices of the missing basis functions obtained by permuting the barycentric coordinates. Right: Notation at the 'dual side'.

At the 'dual side', we consider the nodal basis of the continuous piecewise quadratics w.r.t. the red-refinement of $\hat{T}$, where we omit the basis functions associated to the vertices of $\hat{T}$. We denote these basis functions as indicated in the right of Figure 7.8.

We now apply the following transformations:

1. On the primal side, we redefine

$$e_{\frac{1}{2},\frac{1}{2},0}\leftarrow e_{\frac{1}{2},\frac{1}{2},0}-\tfrac{7}{10}(v_{1,0,0}+v_{0,1,0})+\tfrac{7}{30}v_{0,0,1},$$

and update $e_{\frac{1}{2},0\frac{1}{2}}$ and $e_{0,\frac{1}{2},\frac{1}{2}}$ analogously. As a consequence, we obtain $\operatorname{span}\{e_{\frac{1}{2},\frac{1}{2},0}, e_{\frac{1}{2},0,\frac{1}{2}}, e_{0,\frac{1}{2},\frac{1}{2}}\} \perp \operatorname{span}\{\tilde{v}_{1,0,0}, \tilde{v}_{0,1,0}, \tilde{v}_{0,0,1}\}$.

2. On the dual side, we redefine

$$\tilde{e}_{\frac{1}{2},\frac{1}{2},0} \leftarrow \tfrac{1}{102}\tilde{e}_{\frac{1}{2},\frac{1}{2},0} - \tfrac{7}{2312}(\tilde{e}_{\frac{3}{4},\frac{1}{4},0} + \tilde{e}_{\frac{1}{4},\frac{3}{4},0}),$$

and update $\tilde{e}_{\frac{1}{2},0\frac{1}{2}}$ and $\tilde{e}_{0,\frac{1}{2},\frac{1}{2}}$ analogously. Then $\{e_{\frac{1}{2},\frac{1}{2},0}, e_{\frac{1}{2},0,\frac{1}{2}}, e_{0,\frac{1}{2},\frac{1}{2}}\}$ and $\{\tilde{e}_{\frac{1}{2},\frac{1}{2},0}, \tilde{e}_{\frac{1}{2},0,\frac{1}{2}}, \tilde{e}_{0,\frac{1}{2},\frac{1}{2}}\}$ became biorthogonal. The functions $e_{\pi(\frac{3}{4},\frac{1}{4},0)}$ for any permutation π, will not play any role anymore, and will be ignored.

3. On the dual side, we redefine

$$\begin{bmatrix} \tilde{v}_{1,0,0} \\ \tilde{v}_{0,1,0} \\ \tilde{v}_{0,0,1} \end{bmatrix} \leftarrow 12 \begin{bmatrix} 3 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 3 \end{bmatrix} \begin{bmatrix} \tilde{v}_{1,0,0} \\ \tilde{v}_{0,1,0} \\ \tilde{v}_{0,0,1} \end{bmatrix}.$$

Now, $\{v_{1,0,0}, v_{0,1,0}, v_{0,0,1}\}$ and $\{\tilde{v}_{1,0,0}, \tilde{v}_{0,1,0}, \tilde{v}_{0,0,1}\}$ are biorthogonal.

After these 3 steps, the 6×6 'local generalized mass matrix' that contains the $L_2(\hat{T})$-inner products between all primal functions, grouped into v- and e-functions, and all (remaining) dual functions, grouped into $\tilde{v}$- and $\tilde{e}$-functions, has the 2×2 block structure $\begin{bmatrix} \mathrm{Id} & \frac{9}{32}\mathrm{Id} - \frac{31}{32}\mathbb{1} \\ 0 & \mathrm{Id} \end{bmatrix}$, with $\mathbb{1}$ the 3×3 all-ones matrix (and with the $\tilde{e}$-functions ordered as the 'opposite' v-functions). The invertibility of this matrix confirms that both collections of 6 primal and 6 dual functions are linearly independent.

We use these primal and dual functions on the reference triangle $\hat{T}$ to construct collections of primal and dual functions on Ω by the usual lifting by means of an affine bijection between $\hat{T}$ and any $T' \in \mathcal{T}^\delta$. When doing so, we connect the functions of e or $\tilde{e}$-type continuously over 'their' edges, and omit them on edges on $\partial\Omega$.

Each function of v or $\tilde{v}$-type is supported on one $T' \in \mathcal{T}^\delta$, and we multiply them by the factor $|T'|^{-\frac{1}{2}}$. The functions of e or $\tilde{e}$-type are supported on two adjacent $T', T'' \in \mathcal{T}^\delta$, and we multiply them by the factor $(|T'| + |T''|)^{-\frac{1}{2}}$.

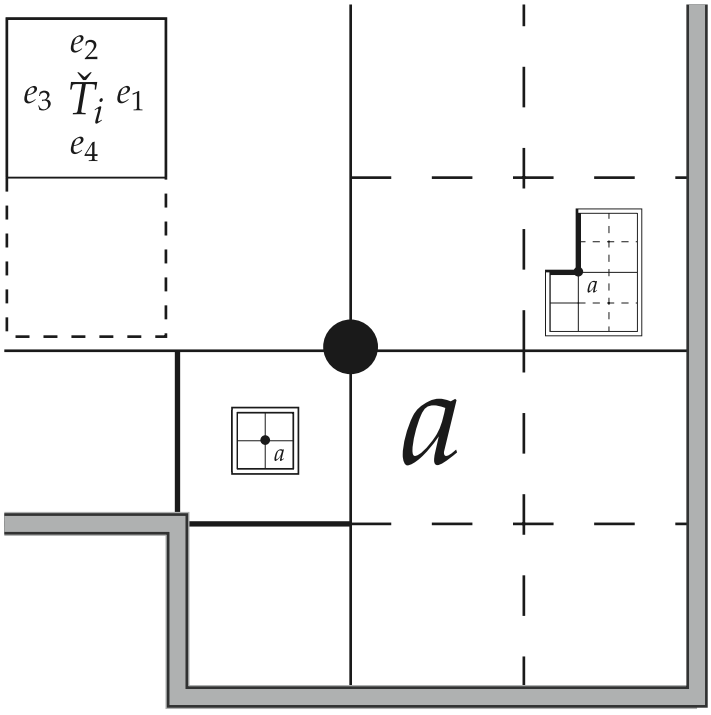
By construction, the resulting primal and dual collections, denoted by Φ^δ and $\tilde{\Phi}^\delta$, are uniformly $L_2(\Omega)$-Riesz systems, with mass matrices whose extremal eigenvalues are inside the interval spanned by the extremal eigenvalues of the corresponding primal or dual mass matrices on the reference triangle.

Furthermore, $\operatorname{span}\Phi^\delta = \mathcal{S}^{0,2}_{\mathcal{T}^\delta,0} + \mathcal{S}^{-1,1}_{\mathcal{T}^\delta}$, and $\operatorname{span}\tilde{\Phi}^\delta \subset \mathcal{S}^{0,2}_{\mathcal{T}^\delta_S,0}$, with $\mathcal{T}^\delta_S$ being constructed from $\mathcal{T}^\delta$ by one uniform red-refinement.

The generalized mass matrix, i.e., the matrix with the $L_2(\Omega)$-inner products between all primal functions, grouped into v- and e-functions, and all

dual functions, grouped into $\tilde{v}$- and $\tilde{e}$-functions, has the 2×2 block structure $\begin{bmatrix} \mathrm{Id} & * \\ 0 & \mathrm{Id} \end{bmatrix}$. The uniform $L_2(\Omega)$-Riesz basis property of both Φ^δ and $\tilde{\Phi}^\delta$ shows that the spectral norm of the non-zero off-diagonal block is uniformly bounded. By now redefining $\Phi^\delta \leftarrow \begin{bmatrix} \mathrm{Id} & -* \\ 0 & \mathrm{Id} \end{bmatrix} \Phi^\delta$, we obtain primal and dual uniformly $L_2(\Omega)$-Riesz systems that are *biorthogonal*, where $\operatorname{span} \Phi^\delta = \mathcal{S}^{0,2}_{\mathcal{T}^\delta,0} + \mathcal{S}^{-1,1}_{\mathcal{T}^\delta}$ and $\operatorname{span} \tilde{\Phi}^\delta \subset \mathcal{S}^{0,2}_{\mathcal{T}^\delta_S,0}$.

In view of the supports of the dual functions, and those of the primal functions before the last transformation, we infer that the support of a function in $\tilde{\Phi}^\delta$ is contained in either one $T' \in \mathcal{T}^\delta$ ($\tilde{v}$-type), or in the union of two triangles from $\mathcal{T}^\delta$ that share an edge ($\tilde{e}$-type), and that the support of a function in Φ^δ is contained in either the union of two triangles from $\mathcal{T}^\delta$ that share an edge (e-type), or in the union of $T' \in \mathcal{T}^\delta$ and those at most three $T'' \in \mathcal{T}^\delta$ that share an edge with T'. We conclude that the biorthogonal projector $P_2^\delta \colon u \mapsto \langle u, \Phi^\delta \rangle_{L_2(\Omega)} \tilde{\Phi}^\delta$ satisfies both conditions (7.5.4) and (7.5.6).


e_2
e_3 $\check{T}_i$ e_1
e_4
a
a
a

8 On p-robust saturation on quadrangulations for elliptic PDEs

Abstract For the Poisson problem in two dimensions, posed on a domain partitioned into axis-aligned rectangles with up to one hanging node per edge, we envision an efficient error reduction step in an instance-optimal hp-adaptive finite element method. Central to this is the problem: Which increase in local polynomial degree ensures p-robust contraction of the error in energy norm? We reduce this problem to a small number of saturation problems on the reference square, and provide strong numerical evidence for their solution.

8.1 Introduction

We consider the Poisson model problem of finding $u : \Omega \to \mathbb{R}$ that satisfies

$$-\Delta u = f \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \partial\Omega, \tag{8.1.1}$$

where $\Omega \subset \mathbb{R}^2$ is a connected union of a finite number of essentially disjoint axis-aligned rectangles, and $f \in L_2(\Omega)$. Given a 1-irregular quadrangulation $\mathcal{T}$ of the domain into essentially disjoint axis-aligned rectangles, let $U_{\mathcal{T}}$ be the space of continuous piecewise polynomials of variable degree w.r.t. $\mathcal{T}$ that vanish on the domain boundary, and let $u_{\mathcal{T}} \in U_{\mathcal{T}}$ be its best approximation of u in energy norm. We are interested in the following *contraction problem*:

Problem Which (hp-)refinement $\overline{\mathcal{T}}$ of $\mathcal{T}$ ensures *contraction* of the energy error, in that

$$\|\nabla u - \nabla u_{\overline{\mathcal{T}}}\|_{L_2(\Omega)} \leq \alpha \|\nabla u - \nabla u_{\mathcal{T}}\|_{L_2(\Omega)}$$

for some fixed $\alpha < 1$ independent of $\mathcal{T}$ and its local polynomial degrees?

It is well known that this problem is equivalent to the *saturation problem* of finding $\overline{\mathcal{T}}$ for which

$$\|\nabla u_{\mathcal{T}} - \nabla u_{\overline{\mathcal{T}}}\|_{L_2(\Omega)} \leq \rho \|\nabla u - \nabla u_{\mathcal{T}}\|_{L_2(\Omega)} \quad \text{for some} \quad \rho > 1;$$

in this work, we will study the saturation problem, posed locally on a patch of rectangles around a given vertex.

This chapter is a minor modification of **On p-Robust Saturation on Quadrangulations**, J. Westerdiep, *Computational Methods in Applied Mathematics*, 20(1):169–186, 2020.

The idea of hp-adaptive finite element methods started gaining momentum in the eighties with the seminal works of Babuška and colleagues [GB86a, GB86b]. They showed that for certain elliptic boundary value problems, careful a priori decisions between h-refinement and p-enrichment can yield a sequence of finite element solutions that exhibit an exponential convergence rate with respect to the number of degrees of freedom (DoFs).

Since then, a lot of research has been done on hp-adaptive refinement driven by a posteriori error estimates, but despite the interest, it was not until 2015 that Canuto, Nochetto, Stevenson and Verani [CNSV17a] proved the instance optimality—and with it, exponential convergence—of one such method. The method alternates between (i) a module that refines the triangulation to reduce the energy error with a sufficiently large fixed factor, and (ii) an hp-coarsening strategy developed by Binev [Bin18] that essentially removes near-redundant degrees of freedom to yield an *instance optimal* triangulation. The sequence of triangulations found after each hp-coarsening step then exhibits the desired exponential decay.

In [CNSV17a], the error reducer of step (i) was a typical h-adaptive loop driven by an element-based Dörfler marking, using the a posteriori error estimator of Melenk and Wohlmuth [MW01]. The efficiency of this error estimator is known to be sensitive to polynomial degrees, which can lead to a runtime that grows exponentially in the number of DoFs.

In [CNSV17b], Canuto *et al.* explore a different error reduction strategy. It is an adaptive p-enrichment loop driven by a vertex-based Dörfler marking using the equilibrated flux estimator, which was shown to be p-robust in [BPS09]. They show that solving a number of *local saturation problems*, posed on patches around a vertex in terms of dual norms of residuals, leads to an efficient error reducer. They were able to reduce the problem, stated on triangulations without hanging nodes, to three problems on a reference triangle, and provided numerical results indicating that uniform saturation holds when increasing the local degree p to $p + \lceil \lambda p \rceil$ for any constant $\lambda > 0$, but that an additive quantity of the form $p + n$ is insufficient.

Finally, in [CNSV19], Canuto *et al.* present a theoretical result solving slightly ill-fitted variant on one of the reference problems. Whereas the former two works discuss partitions of the domain into triangles, the latter proves a result on the reference *square* instead. As a first step towards repairing this inconsistency, the present work considers quadrangulations. Our goal of adaptive approximation requires us to consider partitions with hanging nodes, which introduce complications. A key contribution in this regard has been made by Dolejší, Ern and Vohralík in [DEV16].

Contributions

This work has two related goals. In a larger context, we take a step in the direction of a polynomial-time hp-adaptive FEM with exponential convergence rates. In particular, we are interested in finding an efficient error reducer. To this end, we reduce the *saturation problem* to a small number of problems on the reference square, and provide numerical results suggesting these prob-

lems may be solvable theoretically. We detail the computational aspect as well, so that the numerical results are easily reproducible.

On a lower level, this work aims extends the reduction to reference problems of [CNSV17b] from regular triangulations equipped with polynomials of certain *total degree* to the situation of 1-irregular quadrangulations with polynomials of certain *degree in each variable separately*. Allowing 1-irregularity makes for a rather involved adaptation of the original result, as the refined regular patches are not necessarily composed of elements containing the original vertex.

Organisation

In §2, we will establish our notation. In §3, we show a contractive property within (hp-)adaptive finite element context, under a local patch-based saturation assumption. In §4, we reduce the local saturation assumption to boundedness of a small number of reference saturation coefficients. In §5, we discuss the computation of these coefficients, and in §6 we show numerical results suggesting that these quantities are in fact bounded.

8.2 Notation and setup

In this work, $A \lesssim B$ means that A is bounded by at most a multiple of B, independently of parameters of A and B, and $A \eqsim B$ means $A \lesssim B$ and $B \lesssim A$.

8.2.1 Quadrangulations

We consider partitions $\mathcal{T}$ of the domain into closed axis-aligned rectangles. We impose that $T_1^\circ \cap T_2^\circ = \varnothing$ for $T_1, T_2 \in \mathcal{T}$ distinct, and allow *irregularity* along shared edges, meaning that $T_1 \cap T_2$ may be empty, a shared vertex, or part of a shared edge. Irregularity allows for highly adaptive quadrangulations, but to ensure p-robustness of our main result, we restrict ourselves to *1-irregularity*: every element edge may contain up to one *hanging node*—a vertex in the interior of a neighbor's edge.

To avoid pathological situations, we lastly assume that every $\mathcal{T}$ is found from a regular initial quadrangulation (i.e., without hanging nodes) by means of repeated red-refinement (subdivision into four similar rectangles), thus automatically ensuring uniform shape regularity. We collect the family of such quadrangulations in the set $\mathbb{T}$. See Figure 8.1 for a few examples.

The set of *nonhanging* vertices of a quadrangulation $\mathcal{T} \in \mathbb{T}$ form the set $\mathcal{V}_\mathcal{T}$, and $\mathcal{V}_\mathcal{T}^{\mathrm{ext}}$ (resp. $\mathcal{V}_\mathcal{T}^{\mathrm{int}}$) is its subset of boundary (resp. interior) vertices. The edges of $\mathcal{T}$ form the set $\mathcal{E}_\mathcal{T}$.

8.2.2 Polynomials on quadrangulations

For $T \in \mathcal{T} \in \mathbb{T}$, write $\mathbb{Q}_{p,p'}(T)$ for the space of polynomials on T of degree at most p and p' in the two canonical coordinates. Define $\mathbb{Q}_p(T) := \mathbb{Q}_{p,p}(T)$. Equip each T with a local polynomial degree $p_T = p_{T,\mathcal{T}}$ for which $p_T \geq 1$,

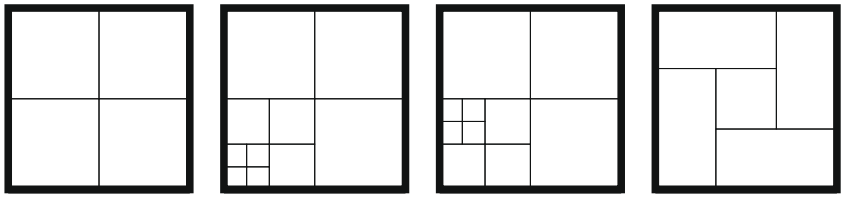


Figure 8.1 (a) Regular initial quadrangulation $\mathcal{T}_a \in \mathbb{T}$ of a square domain; (b) 1-irregular quadrangulation found from $\mathcal{T}_a$ through red-refinement; (c) quadrangulation found from $\mathcal{T}_a$ that is not 1-irregular; (d) typical "pathological" quadrangulation excluded from this chapter.

and write $\boldsymbol{p}_\mathcal{T} := (p_T)_{T\in\mathcal{T}}$ for the collection of these local degrees. Then with $\mathbb{Q}^{-1}_{\boldsymbol{p}_\mathcal{T}}(\mathcal{T}) := \prod_{T\in\mathcal{T}} \mathbb{Q}_{p_T}(T)$ the space of broken piecewise polynomials over $\mathcal{T}$ of degree at most p_T on every element, we introduce the finite-dimensional subspace $U_\mathcal{T}$ of $H^1_0(\Omega)$ as

$$U_\mathcal{T} := H^1_0(\Omega) \cap \mathbb{Q}^{-1}_{\boldsymbol{p}_\mathcal{T}}(\mathcal{T}) \quad (\mathcal{T} \in \mathbb{T}).$$

Denote with $u \in H^1_0(\Omega)$ the weak solution to (8.1.1), and its Galerkin approximation as $u_\mathcal{T} \in U_\mathcal{T}$.

8.2.3 Patches

Let ψ_a be the *hat function* characterized by $\psi_a \in C(\overline{\Omega}) \cap \mathbb{Q}^{-1}_1(\mathcal{T})$ and $\psi_a(b) = \delta_{ab}$ for all $b \in \mathcal{V}_\mathcal{T}$. Let $\omega_a = \omega_{\mathcal{T},a}$ be its support, and denote with $\mathcal{T}_a \subset \mathcal{T}$ the quadrangulation restricted to ω_a; we call this set a *patch*. For each nonhanging vertex $a \in \mathcal{V}_\mathcal{T}$, write

$$\boldsymbol{p}_a := \boldsymbol{p}_{\mathcal{T}_a} = (p_T)_{T\in\mathcal{T}_a}, \quad p_a := \max \boldsymbol{p}_a.$$

We decompose the patch edges $\mathcal{E}_{\mathcal{T}_a} := \{e \in \mathcal{E}_\mathcal{T} : e \subset \omega_a\}$ as

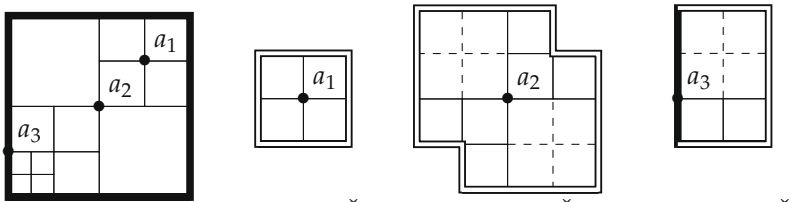
$$\mathcal{E}^{\mathrm{ext}}_a := \left\{e \in \mathcal{E}_{\mathcal{T}_a} : e \subset \partial\omega_a\right\}, \quad \mathcal{E}^{\mathrm{int}}_a := \mathcal{E}_{\mathcal{T}_a} \setminus \mathcal{E}^{\mathrm{ext}}_a.$$

We decompose exterior edges into *Dirichlet* and *Neumann* edges, through

$$\mathcal{E}^{\mathrm{ext},D}_a := \left\{e \in \mathcal{E}^{\mathrm{ext}}_a : a \in e\right\}, \quad \mathcal{E}^{\mathrm{ext},N}_a := \mathcal{E}^{\mathrm{ext}}_a \setminus \mathcal{E}^{\mathrm{ext},D}_a,$$

giving rise to the local spaces

$$H^1_*(\omega_a) := \begin{cases} \left\{v \in H^1(\omega_a) : \langle v, \mathbb{1}\rangle_{\omega_a} = 0\right\} & a \in \mathcal{V}^{\mathrm{int}}_\mathcal{T}, \\ \left\{v \in H^1(\omega_a) : v|_e = 0 \text{ on } e \in \mathcal{E}^{\mathrm{ext},D}_a\right\} & a \in \mathcal{V}^{\mathrm{ext}}_\mathcal{T}. \end{cases}$$



(a) $\mathcal{T}$ (b) $\mathcal{T}_{a_1} = \check{\mathcal{T}}_{a_1}$ (c) $\mathcal{T}_{a_2}$ and $\check{\mathcal{T}}_{a_2}$ (d) $\mathcal{T}_{a_3}$ and $\check{\mathcal{T}}_{a_3}$

Figure 8.2 Example refined patches. (a) Example quadrangulation with three vertices. (b) Regular patch $\mathcal{T}_{a_1}$ of interior vertex $a_1 \in \mathcal{V}_{\mathcal{T}}^{\text{int}}$ that equals its smallest regular refinement $\check{\mathcal{T}}_{a_1}$. (c) 1-irregular patch $\mathcal{T}_{a_2}$ of $a_2 \in \mathcal{V}_{\mathcal{T}}^{\text{int}}$; $\check{\mathcal{T}}_{a_2}$ is a refinement of $\mathcal{T}_{a_2}$ denoted by dashed lines. (d) Patch of boundary vertex $a_3 \in \mathcal{V}_{\mathcal{T}}^{\text{ext}}$. The thick black line indicates edges in $\check{\mathcal{E}}_a^{\text{ext},D}$; the double line edges in $\check{\mathcal{E}}_a^{\text{ext},N}$.

Remark 8.2.1. Our definition of $H^1_*(\omega_a)$ differs from its definition in, e.g., [CNSV17b, DEV16] when $a \in \mathcal{V}_{\mathcal{T}}^{\text{ext}}$. In previous works, functions in $H^1_*(\omega_a)$ vanish on the entire part $\partial\omega_a \cap \partial\Omega$; in our case, they vanish only on those edges $e \subset \partial\omega_a \cap \partial\Omega$ for which $a \in e$. Nonetheless, relevant dual norm properties of the residual in §8.3 carry over to our case. ◊

8.2.4 Refined patches

Given $\mathcal{T}_a$, define the *refined patch* $\check{\mathcal{T}}_a$ as the smallest regular red-refinement of $\mathcal{T}_a$, and let each $\check{T} \in \check{\mathcal{T}}_a$ inherit its local degree $p_{\check{T}}$ from its parent in $\mathcal{T}_a$. The key insight of considering the regular refinement $\check{\mathcal{T}}_a$ instead of $\mathcal{T}_a$ was proposed in [DEV16] and allows us to write the discrete residual below as a sum of inner products with local polynomials.

For the edge sets $\mathcal{E}_a^{\text{int}}, \mathcal{E}_a^{\text{ext}}, \mathcal{E}_a^{\text{ext},N}, \mathcal{E}_a^{\text{ext},D}$, define their $\check{}$-variants as the set of children edges; e.g., $\check{\mathcal{E}}_a^{\text{int}} := \{\check{e} \in \mathcal{E}_{\check{\mathcal{T}}_a} : \exists e \in \mathcal{E}_a^{\text{int}} \text{ s.t. } \check{e} \subset e\}$. See Figure 8.2 for a few examples.

8.3 Reduction to local saturation problems

This section follows the same general structure of [CNSV17b, §3–4]; proofs are omitted for brevity but follow analogously to their counterpart in [CNSV17b].

For ω a proper subset of $\overline{\Omega}$, let $\langle\cdot,\cdot\rangle_\omega$ denote the $L_2(\omega)$- or $[L_2(\omega)]^2$-inner product, and $\|\cdot\|_\omega$ its norm. Unless mentioned otherwise, closed subspaces of $H^1(\omega)$ on which $\|\nabla\cdot\|_\omega$ is equivalent to $\|\cdot\|_{H^1(\omega)}$ are equipped with the $H^1(\omega)$-seminorm $|||\cdot|||_\omega := \|\nabla\cdot\|_\omega$.

8.3.1 Residual

For $\check{e} \in \check{\mathcal{E}}_a^{\mathrm{int}}$, we denote with $[\![\cdot]\!]$ the jump operator and with $\boldsymbol{n}_{\check{e}}$ a unit normal vector of $\check{e}$. We then define the global and localized residuals as

$$r_{\mathcal{T}}(v) := \langle f, v\rangle_\Omega - \langle \nabla u_{\mathcal{T}}, \nabla v\rangle_\Omega, \quad r_a(v) := r_{\mathcal{T}}(\psi_a v) \quad (v \in H^1(\Omega)),$$

and observe that after integration by parts, the localized residual satisfies

$$r_a(v) = \sum_{\check{T}\in\check{\mathcal{T}}_a} \langle \psi_a(f + \Delta u_{\mathcal{T}}), v\rangle_{\check{T}} + \sum_{\check{e}\in\check{\mathcal{E}}_a^{\mathrm{int}}} \langle \psi_a\, [\![\nabla u_{\mathcal{T}} \cdot \boldsymbol{n}_{\check{e}}]\!], v\rangle_{\check{e}}.$$

The following result, first discovered in [CF99] in a slightly different formulation, shows that the residual dual norms $\|r_a\|_{H^1_*(\omega_a)'}$ may be used as a posteriori error indicators.

Proposition (Reliability & Efficiency [CNSV17b, Prop. 3.1])**.** *There is some constant* $C_{\mathrm{eff}} > 0$ *with*

$$|||u - u_{\mathcal{T}}|||_\Omega^2 \le 3 \sum_{a\in\mathcal{V}_{\mathcal{T}}} \|r_a\|^2_{H^1_*(\omega_a)'}, \quad \|r_a\|_{H^1_*(\omega_a)'} \le C_{\mathrm{eff}} |||u - u_{\mathcal{T}}|||_{\omega_a} \ \ (a \in \mathcal{V}_{\mathcal{T}}).$$

8.3.2 Data oscillation and discrete residual

For a rectangle T, define Π_p^T as the $L_2(T)$-orthogonal projection onto $\mathbb{Q}_p(T)$. The approximation $\Pi_{\check{\mathcal{T}}_a} f$ to f is then piecewise defined through $(\Pi_{\check{\mathcal{T}}_a} f)|_{\check{T}} := \Pi^{\check{T}}_{p_{\check{T}}} f|_{\check{T}}$. The difference between f and its approximation is quantified by the *data oscillation*, defined as

$$\mathrm{osc}(f, \mathcal{T})^2 := \sum_{\check{T}\in\check{\mathcal{T}}_a} h^2_{\check{T}} \|f - \Pi^{\check{T}}_{p_{\check{T}}} f\|^2_{\check{T}}.$$

We study the *discrete residual*, computed on discrete data $\Pi_{\check{\mathcal{T}}_a} f$ instead of f:

$$\tilde{r}_a(v) := \sum_{\check{T}\in\check{\mathcal{T}}_a} \langle \phi_{\check{T}}, v\rangle_{\check{T}} + \sum_{\check{e}\in\check{\mathcal{E}}_a^{\mathrm{int}}} \langle \phi_{\check{e}}, v\rangle_{\check{e}} \quad (v \in H^1(\omega_a)) \tag{8.3.1}$$

where

$$\phi_{\check{T}} := \psi_a(\Pi^{\check{T}}_{p_{\check{T}}} f + \Delta u_{\mathcal{T}}) \in \mathbb{Q}_{p_{\check{T}}+1}(\check{T}), \quad \text{and} \quad \phi_{\check{e}} := \psi_a\, [\![\nabla u_{\mathcal{T}} \cdot \boldsymbol{n}_{\check{e}}]\!] \in \mathbb{P}_{p_a+1}(\check{e}).$$

Proposition (Residual Discrepancy [CNSV17b, Corol. 3.4])**.** *There exists a constant* $C_{\mathrm{osc}} > 0$ *with*

$$\left| \sqrt{\sum_{a\in\mathcal{V}_{\mathcal{T}}} \|\tilde{r}_a\|^2_{H^1_*(\omega_a)'}} - \sqrt{\sum_{a\in\mathcal{V}_{\mathcal{T}}} \|r_a\|^2_{H^1_*(\omega_a)'}} \right| \le C_{\mathrm{osc}}\, \mathrm{osc}(f, \mathcal{T}).$$

8.3.3 A theoretical AFEM

We envision an abstract adaptive FEM that loops

$$\text{SOLVE} - \text{ESTIMATE} - \text{MARK} - \text{REFINE},$$

driven by the vertex-based a posteriori error indicators $\|\tilde{r}_a\|_{H^1_*(\omega_a)'}$. The following result provides sufficient conditions for p-robust contraction of the error in energy norm. This AFEM can serve as an efficient error reducer in an instance-optimal hp-AFEM through a coarsening step; cf [Bin18].

Proposition 8.3.1 (Contraction of AFEM [CNSV17b, Prop. 4.1])**.** *Take constants $\theta \in (0,1]$ and $\rho \in [1,\infty)$. Suppose for some $\lambda \in (0, \frac{\theta}{C_{\mathrm{osc}}\rho})$, we have*

(a) small data oscillation:

$$\operatorname{osc}(f,\mathcal{T}) \le \lambda \sqrt{\sum_{a\in\mathcal{V}_\mathcal{T}} \|\tilde{r}_a\|^2_{H^1_*(\omega_a)'}},$$

(b) Dörfler marking: a set $\mathcal{M} \subset \mathcal{V}_\mathcal{T}$ of marked vertices satisfying

$$\sqrt{\sum_{a\in\mathcal{M}} \|\tilde{r}_a\|^2_{H^1_*(\omega_a)'}} \ge \theta \sqrt{\sum_{a\in\mathcal{V}_\mathcal{T}} \|\tilde{r}_a\|^2_{H^1_*(\omega_a)'}},$$

(c) local saturation: a closed subspace $\overline{U} \supset U_\mathcal{T}$ of $H^1_0(\Omega)$ that saturates *each residual dual norm:*

$$\|\tilde{r}_a\|_{H^1_*(\omega_a)'} \le \rho \|\tilde{r}_a\|_{[H^1_*(\omega_a)\cap \overline{U}|_{\omega_a}]'} \quad (a \in \mathcal{M}).$$

Then, with $\overline{u} \in \overline{U}$ the Galerkin approximation of the solution u of (8.1.1)*, we have contraction,*

$$|||u - \overline{u}|||_\Omega \le \alpha |||u - u_\mathcal{T}|||_\Omega, \quad \alpha = \alpha(\theta,\rho,\lambda) := \sqrt{1 - \left(\frac{\theta - C_{\mathrm{osc}}\lambda\rho}{3C_{\mathrm{eff}}(1+C_{\mathrm{osc}}\lambda)\rho}\right)^2},$$

meaning the error is reduced by a factor α, uniformly bounded away from 1.

Remark. Assumption (a) is usually satisfied [CNSV17b, Rem. 4.2], and the Dörfler marking for (b) can be constructed by ordering vertices by $\|\tilde{r}_a\|_{H^1_*(\omega_a)'}$, so we focus on (c). Given a function $q : \mathbb{N} \to \mathbb{N}$ such that

$$\|\tilde{r}_a\|_{H^1_*(\omega_a)'} \le \rho \|\tilde{r}_a\|_{[H^1_*(\omega_a)\cap \mathrm{Q}^{-1}_{q(p_a+1)+1}(\check{\mathcal{T}}_a)]'} \quad (a \in \mathcal{M}),$$

then, (c) is satisfied for any $U_\mathcal{T} \subset \overline{U} \subset H^1_0(\Omega)$ with

$$H^1_*(\omega_a) \cap \mathrm{Q}^{-1}_{q(p_a+1)+1}(\check{\mathcal{T}}_a) \subset H^1_*(\omega_a) \cap \overline{U}|_{\omega_a} \quad (a \in \mathcal{M}).$$

In Theorem 8.4.2 below, we reduce existence of q to a small number of saturation problems on the reference square. Under this assumption, $\overline{U}$ can be constructed as $U_{\check{\mathcal{T}}}$, where $\check{\mathcal{T}}$ is found through the following REFINE step:

(i) for each $a \in \mathcal{M}$, replace $\mathcal{T}_a$ by its smallest regular red-refinement $\check{\mathcal{T}}_a$;

(ii) for each $a \in \mathcal{M}$, for each $\check{T} \in \check{\mathcal{T}}_a$, increase $p_{\check{T}}$ to $q(p_a + 1) + 1$;

(iii) Take $\check{\mathcal{T}}$ the smallest 1-irregular red-refinement of the resulting mesh.

The numerical results of §8.6 suggest that the aforementioned reference problems are solved for $q(p) := p + \lceil \lambda p \rceil$ for any $\lambda > 0$. Each REFINE step multiplies the number of elements by not more than a factor 4, and the local degrees by (up to) a constant factor $1 + \lceil \lambda \rceil$. Therefore, the dimension of the local finite element space is multiplied by not more than a factor $4(1 + \lceil \lambda \rceil)^2$; since the number of REFINE steps necessary for a fixed error reduction factor $\delta \in (0, 1)$ is bounded by $M \leq \lceil \frac{\log \delta}{\log \alpha} \rceil$, leading to an efficient error reducer. ♢

8.3.4 Equivalent computable error quantities

The localized discrete residuals $\tilde{r}_a$ provide, by their dual norms $\|\tilde{r}_a\|_{H^1_*(\omega_a)'}$, reliable and efficient error indicators which can drive an AFEM. These dual norms are, however, not computable.

For $p \geq 0$ and a rectangle T, the *Raviart-Thomas space* of degree p is

$$\mathrm{RT}_p(T) := \mathbb{Q}_{p+1,p}(T) \times \mathbb{Q}_{p,p+1}(T) \subset \boldsymbol{H}(\mathrm{div}; T).$$

The following results underline the importance of this space for p-robustness.

Lemma 8.3.2 (p-Robust Inverse of Divergence [BPS09, Thm. 5])**.** *Let T be a rectangle. For $\varphi \in \mathbb{Q}_p(T)$, there is a $\boldsymbol{\sigma} \in \mathrm{RT}_p(T)$ with*

$$\mathrm{div}\, \boldsymbol{\sigma} = \varphi, \quad \|\boldsymbol{\sigma}\|_T \lesssim \|\langle \varphi, \cdot \rangle_T\|_{H^1_0(T)'}.$$

Lemma 8.3.3 (p-Robust Raviart–Thomas Extension [CM10, Corol. 3.4])**.** *Take T a rectangle with edges $\{e_1, e_2, e_3, e_4\}$, and take γ the union of one or more edges. Suppose we have a $\varphi \in L_2(\gamma)$ such that $\varphi|_e \in \mathbb{P}_p(e)$ for all edges $e \subset \gamma$, and when $\gamma = \partial T$, also $\langle \varphi, \mathbb{1} \rangle_\gamma = 0$. Then there is a $\boldsymbol{\sigma} \in \mathrm{RT}_p(T)$ with*

$$\mathrm{div}\, \boldsymbol{\sigma} = 0, \quad (\boldsymbol{\sigma} \cdot \boldsymbol{n}_T)|_\gamma = \varphi, \quad \|\boldsymbol{\sigma}\|_T \lesssim \inf_{\{\boldsymbol{\tau} \in \boldsymbol{H}(\mathrm{div};T) : \mathrm{div}\, \boldsymbol{\tau} = 0, (\boldsymbol{\tau} \cdot \boldsymbol{n}_T)|_\gamma = \varphi\}} \|\boldsymbol{\tau}\|_T.$$

Remark. Lemma 8.3.3 follows from a careful reading of [CDD08, §3.3–3.4], where we sum over only those polynomial lifts U_j that correspond with an edge in γ. Their result is on tangential derivatives and $\boldsymbol{H}(\mathrm{curl}; T)$, but rotating over 90° recovers our result for the normal derivatives and $\boldsymbol{H}(\mathrm{div}; T)$. ♢

In [DEV16], Dolejší *et al.* use these two lemmas (stated on triangles) to find a Raviart–Thomas flux $\boldsymbol{\sigma}_a \in \prod_{\check{T} \in \check{\mathcal{T}}_a} \mathrm{RT}_{p_a}(\check{T})$ with p-robust norm equivalence $\|\boldsymbol{\sigma}_a\|_{\omega_a} \eqsim \|\tilde{r}_a\|_{H^1_*(\omega_a)'}$, and present an efficient algorithm for its construction. The error indicators $\|\boldsymbol{\sigma}_a\|_{\omega_a}$ *can* be computed, and so can drive an AFEM.

8.4 Reduction to reference problems

In this section, we prove the main theorem of this work, reducing the local p-robust saturation problem to a small number of saturation problems on the reference square.

8.4.1 Saturation coefficients

Let $\hat{T} := [-1,1]^2$ be the reference square. For a closed subspace $\hat{\mathscr{H}} \subset H^1(\hat{T})$ on which the $H^1(\hat{T})$-seminorm is a norm, a finite-dimensional subspace $\hat{\mathscr{V}} \subset \hat{\mathscr{H}}$, and a set of functionals $\hat{\mathscr{F}} \subset \hat{\mathscr{H}}'$, define the *saturation coefficient*

$$S(\hat{\mathscr{H}}, \hat{\mathscr{V}}, \hat{\mathscr{F}}) := \sup_{\hat{F} \in \hat{\mathscr{F}}} \frac{\|\hat{F}\|_{\hat{\mathscr{H}}'}}{\|\hat{F}\|_{\hat{\mathscr{V}}'}}$$

which, if bounded, shows that $\hat{\mathscr{V}}$ is large enough to *saturate* $\hat{\mathscr{H}}$ over $\hat{\mathscr{F}}$.

Lemma 8.4.1 (Saturation extends to rectangles)**.** *For any $T \in \mathcal{T} \in \mathbb{T}$,*

$$\sup_{F \in \mathscr{F}} \frac{\|F\|_{\mathscr{H}'}}{\|F\|_{\mathscr{V}'}} \lesssim \kappa_2(\boldsymbol{B}) S(\hat{\mathscr{H}}, \hat{\mathscr{V}}, \hat{\mathscr{F}})$$

where $F_T(\boldsymbol{x}) := \boldsymbol{B}\boldsymbol{x} + \boldsymbol{b}$ is an affine mapping from T to $\hat{T}$, and $\mathscr{H}, \mathscr{V}, \mathscr{F}$ are determined by the pull-back, pull-back, resp. push-forward; cf. [BS08, p. 82]. In words, saturation on the reference square extends to uniformly shape regular rectangles.

8.4.2 Enumerating the interior edges of a refined patch

Refined patches will play an integral role in the proof of the forthcoming Theorem. Take $a \in \mathcal{V}_{\mathcal{T}}$, and let $\check{\mathcal{T}}_a$ be its refined patch. We will construct an enumeration of the interior edges $\check{\mathcal{E}}_a^{\mathrm{int}}$ of $\check{\mathcal{T}}_a$ as $(\check{e}_i)_{i=1}^{n_a}$, where $n_a := \#\check{\mathcal{E}}_a^{\mathrm{int}}$, and for each interior edge, choose a specific square $\check{T}_i \in \check{\mathcal{T}}_a$ adjacent to $\check{e}_i$.

Because every patch $\mathcal{T}_a$ is a 1-irregular collection of axis-aligned rectangles, there is only a finite number of different refined patch types. In fact, it can be shown that up to rotation/flipping of $\check{\mathcal{T}}_a$, all patches fall in one of the thirteen types on the right of Figure 8.3.

Overlay the vertex a with the $\hat{a}$ in the 4×4 grid to the left of Figure 8.3. Then every $\check{e} \in \check{\mathcal{E}}_a^{\mathrm{int}}$ inherits a number $1 \le k(i) \le 24$ from the grid. We then enumerate $(\check{e}_i)_{i=1}^{n_a}$ in increasing order of the values $k(i)$, and we choose $\check{T}_i$ as the square above or to the left of $\check{e}_i$ (whichever is applicable).

8.4.3 Main theorem

Let T be a rectangle. When $\gamma \subset \partial T$ with $\mathrm{meas}(\gamma) > 0$, the space $H^1_{0,\gamma}(T)$ denotes the closure in $H^1(T)$ of the smooth functions on $\overline{T}$ that vanish on γ. By abuse of notation, when $\mathcal{E} = \{\gamma\}$ is a collection of such parts of the boundary, $H^1_{0,\mathcal{E}}(T)$ will denote the closure of smooth functions that vanish on every γ separately.

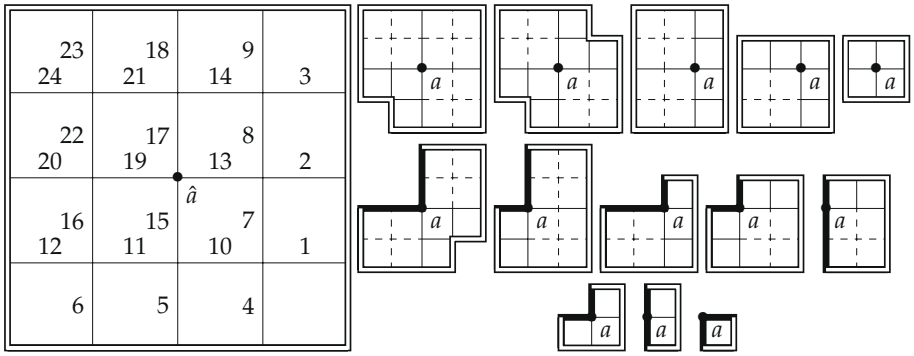


Figure 8.3 Left: a 4×4 grid with vertex $\hat{a}$, and enumeration of its interior edges. Right: the thirteen fundamentally different refined patch types, with the double line indicating Neumann edges $\check{\mathcal{E}}_a^{\mathrm{ext},N}$ of the patch boundary, and the thick black line Dirichlet edges $\check{\mathcal{E}}_a^{\mathrm{ext},D}$. We enumerate interior edges of a patch by overlaying its vertex a with $\hat{a}$ in the left grid, and numbering them in increasing order.

For brevity, write the restriction of $H^1_*(\omega_a)$ to piecewise polynomials as

$$H^1_{*,p}(\check{\mathcal{T}}_a) := H^1_*(\omega_a) \cap \mathbb{Q}_p^{-1}(\check{\mathcal{T}}_a) \quad (p \in \mathbb{N},\ a \in \mathcal{V}_{\mathcal{T}}).$$

We enumerate the edges of the reference square $\hat{T}$ as $\mathcal{E}_{\hat{T}} = (\hat{e}_1, \hat{e}_2, \hat{e}_3, \hat{e}_4)$, in counterclockwise fashion, starting from the rightmost edge.

Theorem 8.4.2 (Reduction of p-robust saturation)**.** *Given the sets*

$$\mathbb{E}^{(A)} := \left\{\mathcal{E} \subset \mathcal{E}_{\hat{T}} : \mathcal{E} \neq \varnothing\right\}, \quad \mathbb{E}^{(B)} := \left\{\left\{\hat{e}_2\right\}, \left\{\hat{e}_3\right\}, \left\{\hat{e}_2, \hat{e}_3\right\}, \left\{\hat{e}_2, \hat{e}_3, \hat{e}_4\right\}\right\}, \tag{8.4.1}$$

of subsets of $\mathcal{E}_{\hat{T}}$, define the reference saturation coefficients

$$S^{(A)}_{\mathcal{E},p,q} := S\Big(H^1_{0,\mathcal{E}}(\hat{T}), H^1_{0,\mathcal{E}}(\hat{T}) \cap \mathbb{Q}_q(\hat{T}), \{h \mapsto \langle \phi, h\rangle_{\hat{T}} : \phi \in \mathbb{Q}_p(\hat{T})\}\Big) \ (\mathcal{E} \in \mathbb{E}^{(A)}),$$
$$S^{(B)}_{\mathcal{E},p,q} := S\Big(H^1_{0,\mathcal{E}}(\hat{T}), H^1_{0,\mathcal{E}}(\hat{T}) \cap \mathbb{Q}_q(\hat{T}), \{h \mapsto \langle \phi, h\rangle_{\hat{e}_1} : \phi \in \mathbb{P}_p(\hat{e}_1)\}\Big) \ (\mathcal{E} \in \mathbb{E}^{(B)}),$$
$$S^{(C)}_{p,q} := S\Big(H^1(\hat{T})/\mathbb{R}, \mathbb{Q}_q(\hat{T})/\mathbb{R}, \{h \mapsto \langle \phi, h\rangle_{\hat{e}_1} : \phi \in \mathbb{P}_p(\hat{e}_1)/\mathbb{R}\}\Big).$$

If for some function $q : \mathbb{N} \to \mathbb{N}$, it holds that

$$\hat{S} := \sup_p \max\left\{S^{(A)}_{\mathcal{E},p,q(p)} : \mathcal{E} \in \mathbb{E}^{(A)}\right\} \cup \left\{S^{(B)}_{\mathcal{E},p,q(p)} : \mathcal{E} \in \mathbb{E}^{(B)}\right\} \cup \left\{S^{(C)}_{p,q(p)}\right\} < \infty,$$

then we have p-robust saturation, in that

$$\|\tilde{r}_a\|_{H^1_*(\omega_a)'} \lesssim \|\tilde{r}_a\|_{H^1_{*,q(p_a+1)+1}(\check{\mathcal{T}}_a)'} \tag{8.4.2}$$

dependent on $\hat{S}$, but independent of the quadrangulation $\mathcal{T}$ and its local degrees.

Outline of proof

Our proof is similar in taste to [BPS09, Thm. 7] and [CNSV17b, Thm. 7.1], with some details requiring a more involved approach. We will perform three steps. Write, as in (8.3.1),

$$\tilde{r}_a(v) = \sum_{\check{T}\in\check{\mathcal{T}}_a} \langle \phi_{\check{T}}, v\rangle_{\check{T}} + \sum_{\check{e}\in\check{\mathcal{E}}_a^{\mathrm{int}}} \langle \phi_{\check{e}}, v\rangle_{\check{e}} \quad (v \in H^1(\omega_a))$$

for some $\phi_{\check{T}} \in \mathbb{Q}_{p_{\check{T}}+1}(\check{T})$ and $\phi_{\check{e}} \in \mathbb{P}_{p_a+1}(\check{e})$. In Step (A) below, we bound the dual norm of the set of element terms; in Steps (B) and (C), we do the same for the set of edge terms. Throughout the proof, we will use the assumption $p_{\check{T}} \geq 1$ to find that, for interior vertices $a \in \mathcal{V}_{\mathcal{T}}^{\mathrm{int}}$, the residual vanishes on constants ($\psi_a \in U_{\mathcal{T}}$ so $\tilde{r}_a(\mathbb{1}) = \tilde{r}(\psi_a \mathbb{1}) = \tilde{r}(\psi_a) = 0$).

In Step (A), we employ $\sup_p \max_{\mathcal{E}\in\mathbb{E}^{(A)}} S^{(A)}_{\mathcal{E},p,q(p)} < \infty$ to find, on every rectangle $\check{T} \in \check{\mathcal{T}}_a$, a functional $\tilde{r}_{\check{T}} \in H^1_*(\omega_a)'$ with

$$\|\tilde{r}_{\check{T}}\|_{H^1_*(\omega_a)'} \lesssim \|\tilde{r}_a\|_{H^1_{*,q(p_a+1)}(\check{\mathcal{T}}_a)'} \quad \text{and} \quad \tilde{r}_{\check{T}}(\mathbb{1}) = 0, \tag{8.4.3}$$

that removes the $\check{T}$-contribution from $\tilde{r}_a$, in the sense that the residual $\tilde{r}_a^{(0)} := \sum_{\check{T}\in\check{\mathcal{T}}_a} \tilde{r}_{\check{T}}$ satisfies, for $v \in H^1_*(\omega_a)$,

$$\tilde{r}_a(v) - \tilde{r}_a^{(0)}(v) = \sum_{\check{e}\in\check{\mathcal{E}}_a^{\mathrm{int}}} \langle \phi_{\check{e}}^{(0)}, v\rangle_{\check{e}} \quad \text{for some} \quad \phi_{\check{e}}^{(0)} \in \mathbb{P}_{p_a+1}(\check{e}). \tag{8.4.4}$$

In Step (B), we use the enumeration $(\check{e}_j)_{j=1}^{n_a}$ of the interior edges $\check{\mathcal{E}}_a^{\mathrm{int}}$ where $n_a := \#\check{\mathcal{E}}_a^{\mathrm{int}}$. At step $i \in \{1,\dots,n_a-1\}$, we use $\sup_p \max_{\mathcal{E}\in\mathbb{E}^{(B)}} S^{(B)}_{\mathcal{E},p,q(p)} < \infty$ and Lemma 8.4.3 below to find a functional $\tilde{r}_{\check{e}_i} = \tilde{r}_{\check{T}_i,\check{e}_i} \in H^1_*(\omega_a)'$ with

$$\|\tilde{r}_{\check{e}_i}\|_{H^1_*(\omega_a)'} \lesssim \|\tilde{r}_a\|_{H^1_{*,q(p_a+1)+1}(\check{\mathcal{T}}_a)'} \quad \text{and} \quad \tilde{r}_{\check{e}_i}(\mathbb{1}) = 0, \tag{8.4.5}$$

that removes the $\check{e}_i$-contribution from $\tilde{r}_a^{(i-1)}$ while not re-introducing contributions on edges $\check{e}_j$ for $j < i$, in the sense that the residual $\tilde{r}_a^{(i)} := \tilde{r}_a^{(i-1)} + \tilde{r}_{\check{e}_i}$ satisfies, for $v \in H^1_*(\omega_a)$,

$$\tilde{r}_a(v) - \tilde{r}_a^{(i)}(v) = \sum_{j\geq i+1} \langle \phi_{\check{e}_j}^{(i)}, v\rangle_{\check{e}_j} \quad \text{for some} \quad \phi_{\check{e}_j}^{(i)} \in \mathbb{P}_{p_a+1}(\check{e}_j). \tag{8.4.6}$$

Lastly, in Step (C), the final iteration $i = n_a$, we make a distinction. When $a \in \mathcal{V}_{\mathcal{T}}^{\mathrm{ext}}$ is a boundary vertex, we construct a $\tilde{r}_{\check{e}_{n_a}} \in H^1_*(\omega_a)'$ for which (8.4.5) and (8.4.6) hold once more. Then through the triangle inequality, $\#\check{\mathcal{T}}_a \leq 16$, and $\#\check{\mathcal{E}}_a^{\mathrm{int}} \leq 24$ we find

$$\|\tilde{r}_a\|_{H^1_*(\omega_a)'} \leq \sum_{\check{T}\in\check{\mathcal{T}}_a} \|\tilde{r}_{\check{T}}\|_{H^1_*(\omega_a)'} + \sum_{j=1}^{n_a} \|\tilde{r}_{\check{e}_j}\|_{H^1_*(\omega_a)'} \lesssim \|\tilde{r}_a\|_{H^1_{*,q(p_a+1)+1}(\check{\mathcal{T}}_a)'}.$$

When $a \in \mathcal{V}_{\mathcal{T}}^{\mathrm{int}}$ is an interior vertex, we use $\sup_p S^{(C)}_{p,q(p)} < \infty$ to bound

$$\|\tilde{r}_a - \tilde{r}_a^{(n_a-1)}\|_{H^1_*(\omega_a)'} \lesssim \|\tilde{r}_a\|_{H^1_{*,q(p_a+1)}(\check{\mathcal{T}}_a)'} \tag{8.4.7}$$

which implies that

$$\begin{aligned}\|\tilde{r}_a\|_{H^1_*(\omega_a)'} &\leq \|\tilde{r}_a - \tilde{r}_a^{(n_a-1)}\|_{H^1_*(\omega_a)'} + \sum_{\check{T}\in\check{\mathcal{T}}_a} \|\tilde{r}_{\check{T}}\|_{H^1_*(\omega_a)'} + \sum_{j=1}^{n_a-1} \|\tilde{r}_{\check{e}_j}\|_{H^1_*(\omega_a)'} \\ &\lesssim \|\tilde{r}_a\|_{H^1_{*,q(p_a+1)+1}(\check{\mathcal{T}}_a)'}.\end{aligned}$$

In either case, we conclude that (8.4.2) must hold.

Extension lemma

Proving, in particular, inequality (8.4.5) requires some creativity. Assume for now that a is a boundary vertex (the other case is handled in the main proof). We will require the intermediate result that for some specific finite-dimensional subspace of polynomials $\mathscr{V}_i \subset H^1(\check{T}_i)$, there is, for each $v \in \mathscr{V}_i$, a piecewise polynomial $Ev \in H^1_*(\omega_a)$ with

$$|||Ev|||_{\omega_a} \lesssim |||v|||_{\check{T}_i}, \quad \text{and} \quad \langle \phi^{(i-1)}_{\check{e}_i}, v\rangle_{\check{e}_i} = \tilde{r}_a(Ev) - \tilde{r}_a^{(i-1)}(Ev).$$

Our approach is the following. Note that $\langle \phi^{(i-1)}_{\check{e}_i}, v\rangle_{\check{e}_i}$ is an inner product over a single edge, whereas $\tilde{r}_a(Ev) - \tilde{r}_a^{(i-1)}(Ev)$ is a sum of inner products $\langle \phi^{(i-1)}_{\check{e}_j}, Ev\rangle_{\check{e}_j}$ on interior edges $\check{e}_j$ with $j \geq i$ (see (8.4.6)). The desired equality holds for all $v \in \mathscr{V}_i$ surely when Ev extends v (in that $Ev|_{\check{T}_i} = v$), and $Ev|_{\check{e}_j} = 0$ for every $j \geq i+1$. Moreover, $a \in \mathcal{V}_{\mathcal{T}}^{\mathrm{ext}}$, so $Ev \in H^1_*(\omega_a)$ should vanish on all edges in $\check{\mathcal{E}}_a^{\mathrm{ext},D}$. This gives rise to the *set of patch (resp. local) Dirichlet edges*,

$$\check{\mathcal{E}}^D_{a,i} := \check{\mathcal{E}}_a^{\mathrm{ext},D} \cup \left\{\check{e}_j \in \check{\mathcal{E}}_a^{\mathrm{int}} : j \geq i+1\right\}, \quad \check{\mathcal{E}}^{\mathrm{loc},D}_{a,i} := \check{\mathcal{E}}^D_{a,i} \cap \check{\mathcal{E}}_{\check{T}_i} \quad (i = 1, \dots, n_a),$$

and for v that vanishes on all local Dirichlet edges, Ev then vanishes on all *patch* Dirichlet edges $\check{e} \in \check{\mathcal{E}}^D_{a,i}$. Existence of this Ev depends on the enumeration $(\check{e}_i)_{i=1}^{n_a}$ of interior edges. The following lemma shows that with our particular construction, we can build a suitable E.

Lemma 8.4.3 (Bounded polynomial extension)**.** *Let $\check{\mathcal{T}}_a$ be one of the thirteen refined patch types of Figure 8.3. Let n_a, $(\check{e}_i)_{i=1}^{n_a}$, and $(\check{T}_i)_{i=1}^{n_a}$ be as defined in §8.4.2. For each square $\check{T}_i$, we enumerate its edges as (e_1, e_2, e_3, e_4), in counterclockwise fashion, starting from the rightmost edge.*

For $1 \leq i \leq n_a - 1$, and $i = n_a$ when a is an external vertex, we have:

1. *The set $\check{\mathcal{E}}^{\mathrm{loc},D}_{a,i}$ is nonempty. In fact, one of five situations occurs:*

   (a) $\check{\mathcal{E}}^{\mathrm{loc},D}_{a,i} = \{e_1, e_2, e_3\}$, (b) $\check{\mathcal{E}}^{\mathrm{loc},D}_{a,i} = \{e_2, e_3, e_4\}$, (c) $\check{\mathcal{E}}^{\mathrm{loc},D}_{a,i} = \{e_2, e_3\}$,
   (d) $\check{\mathcal{E}}^{\mathrm{loc},D}_{a,i} = \{e_2\}$ *and* $e_3 \in \check{\mathcal{E}}_a^{\mathrm{ext},N}$, (e) $\check{\mathcal{E}}^{\mathrm{loc},D}_{a,i} = \{e_3\}$ *and* $e_2 \in \check{\mathcal{E}}_a^{\mathrm{ext},N}$.

2. *There is a bounded linear map*

$$E : H^1_{0,\check{\mathcal{E}}^{\mathrm{loc},D}_{a,i}}(\check{T}_i) \cap \mathbb{Q}_{q(p_a+1)}(\check{T}_i) \to H^1(\omega_a) \cap \mathbb{Q}^{-1}_{q(p_a+1)+1}(\check{\mathcal{T}}_a)$$

so that for all v, its extension Ev vanishes on patch Dirichlet edges; specifically,

$$Ev|_{\check{T}_i} = v, \quad \|\!|Ev|\!\|_{\omega_a} \lesssim \|\!|v|\!\|_{\check{T}_i}, \quad Ev|_{\check{e}} = 0 \ \ (\check{e} \in \check{\mathcal{E}}^D_{a,i}).$$

Proof. A careful visual inspection of the enumeration for each of the thirteen patch types of Figure 8.3 shows that condition (1) holds: by enumerating the edges right-to-left, bottom-to-top, we ensure e_2 and e_3 are (situations (a–c)) both in $\check{\mathcal{E}}^{\mathrm{ext},D}_a$ or equal to some $\check{e}_j$ for $j > i$, or (situations (d–e)) when $\check{T}_i$ is in the topmost row or leftmost column, either e_2 or e_3 is in $\check{\mathcal{E}}^{\mathrm{ext},N}_a$, but never both.

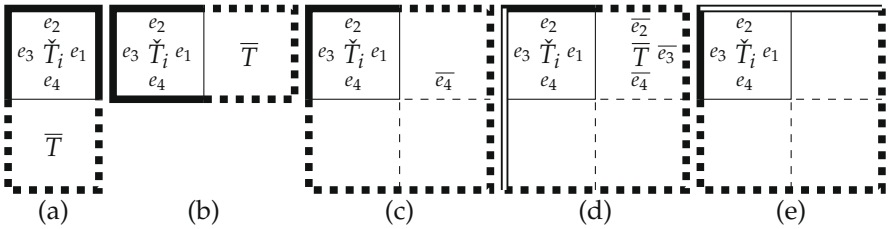


Figure 8.4 The different extensions of Lemma 8.4.3: (a) $\check{\mathcal{E}}^{\mathrm{loc},D}_{a,i} = \{e_1, e_2, e_3\}$, (b) $\check{\mathcal{E}}^{\mathrm{loc},D}_{a,i} = \{e_2, e_3, e_4\}$, (c) $\check{\mathcal{E}}^{\mathrm{loc},D}_{a,i} = \{e_2, e_3\}$, (d) $\check{\mathcal{E}}^{\mathrm{loc},D}_{a,i} = \{e_2\}, e_3 \in \check{\mathcal{E}}^{\mathrm{ext},N}_a$, (e) $\check{\mathcal{E}}^{\mathrm{loc},D}_{a,i} = \{e_3\}, e_2 \in \check{\mathcal{E}}^{\mathrm{ext},N}_a$. The full thick line on $\partial\check{T}_i$ denotes its local Dirichlet boundary $\check{\mathcal{E}}^{\mathrm{loc},D}_{a,i}$, and the dashed thick line shows the Dirichlet boundary of the extension; double lines indicate edges in $\check{\mathcal{E}}^{\mathrm{ext},N}_a$.

By the first result of this Lemma, there are essentially five cases to look at. See Figure 8.4. Denote with $\mathscr{T}$ the union of squares in the appropriate case. Let $v \in H^1_{0,\check{\mathcal{E}}^{\mathrm{loc},D}_{a,i}}(\check{T}_i) \cap \mathbb{Q}_{q(p_a+1)}(\check{T}_i)$. We will use multiple reflections of v to find a piecewise polynomial $\underline{v} \in H^1(\mathscr{T})$ (of degree $q(p_a+1)+1$) that vanishes on the part of $\partial\mathscr{T}$ denoted by the thick line. Restricting $\underline{v}$ to $\mathscr{T} \cap \omega_a$ (because $\mathscr{T}$ may contain squares outside $\check{\mathcal{T}}_a$) yields a function that vanishes on the edges $\check{e} \in \check{\mathcal{E}}^{\mathrm{int}}_a$ with $\check{e} \subset \partial\mathscr{T}$, so that we can easily zero-extend $\underline{v}|_{\mathscr{T}\cap\omega_a}$ to $Ev \in H^1(\omega_a) \cap \mathbb{Q}^{-1}_{q(p_a+1)+1}(\check{\mathcal{T}}_a)$.

The choice of $\check{T}_i$ ensures that $\check{e}_i$ is its right or bottom edge. Moreover, the enumeration is bottom-right to top-left, so that every patch Dirichlet edge is positioned either above or to the left of $\check{T}_i$. On the other hand, the support of our extension Ev is—as we will shortly see—to the right or bottom of $\check{T}_i$. Therefore, Ev necessarily vanishes on all of $\check{\mathcal{E}}^D_{a,i}$.

It remains to construct $\underline{v}$ with the properties above, for each situation.

(a) Denote with $\overline{v}, \overline{T}$ the reflections of v and $\check{T}_i$ across e_4. Then $\overline{v}|_{e_4} = v|_{e_4}$ and $\|\!|\overline{v}|\!\|_{\overline{T}} = \|\!|v|\!\|_{\check{T}}$, so the extension $\underline{v}$ defined by $\underline{v}|_{\check{T}} := v$ and $\underline{v}|_{\overline{T}} := \overline{v}$

vanishes on all of $\partial(\check{T}_i \cup \overline{T})$, is continuous globally, and polynomial on both squares separately.

(b) The proof of this case is analogous to that of (a).

(c) Denote with $\overline{e_4}$ the reflection of e_4 across e_1. Denote with $\overline{\overline{v}}$ the extension of v on $\check{T}_i \cup \overline{T}$. The extension $\underline{v}$ of $\overline{\overline{v}}$ across $e_4 \cup \overline{e_4}$ is the desired function.

(d) Denote with $\overline{v}, \overline{T}, \overline{e_2}, \overline{e_3}, \overline{e_4}$ the reflections of v, T, e_2, e_3, e_4 across e_1, respectively. Let $\overline{\phi} \in \mathbb{Q}_1(\overline{T})$ be a decay function defined by $\overline{\phi}|_{e_1} = 1$ and $\overline{\phi}|_{\overline{e_3}} = 0$. Then

$$(\overline{v}\overline{\phi})|_{e_1} = v|_{e_1}, \quad \overline{v}\overline{\phi} \in H^1_{0,\overline{e_2}\cup\overline{e_3}}(\overline{T}) \cap \mathbb{Q}_{q(p_a+1)+1}(\overline{T}),$$

and we thus see that the function $\overline{\overline{v}}$ defined by $\overline{\overline{v}}|_{\check{T}_i} := v$, $\overline{\overline{v}}|_{\overline{T}} := \overline{v}\overline{\phi}$ is a continuous polynomial extension of v that moreover vanishes on $\overline{e_3}$. Its norm satisfies $|||\overline{\overline{v}}|||_{\check{T}_i\cup\overline{T}} \lesssim |||v|||_{\check{T}_i}$ (proof is analogous to [EV15, (3.29)]). The desired function $\underline{v}$ is found as the extension of $\overline{\overline{v}}$ across $e_4 \cup \overline{e_4}$.

(e) The proof of this case is analogous to that of (d). □

Proof of Theorem 8.4.2

We proceed in several steps.

Step (A0) For every $\check{T} \in \check{\mathcal{T}}_a$, we will find our functional $\tilde{r}_{\check{T}} \in H^1_*(\omega_a)'$ by constructing a Raviart–Thomas flux $\boldsymbol{\sigma}_{\check{T}} \in \mathrm{RT}_{p_a+1}(\check{T})$, and write $\tilde{r}_{\check{T}}(v) = \langle \boldsymbol{\sigma}_{\check{T}}, \nabla v\rangle_{\check{T}}$. Let $\check{T} \in \check{\mathcal{T}}_a$.

Step (A1) We construct $\tilde{r}_{\check{T}}$. By Lemma 8.3.2, there is a $\boldsymbol{\sigma}^{(1)}_{\check{T}} \in \mathrm{RT}_{p_a+1}(\check{T})$ with

$$\operatorname{div} \boldsymbol{\sigma}^{(1)}_{\check{T}} = \phi_{\check{T}} \quad \text{and} \quad \|\boldsymbol{\sigma}^{(1)}_{\check{T}}\|_{\check{T}} \lesssim \|\langle \phi_{\check{T}}, \cdot\rangle_{\check{T}}\|_{H^1_0(\check{T})'}. \tag{8.4.8}$$

By definition, $\tilde{r}_a$ has no contributions on the exterior edges of $\check{\mathcal{T}}_a$. However, $\boldsymbol{\sigma}^{(1)}_{\check{T}}$ *can have* a nonzero normal component on edges in $\check{\mathcal{E}}^{\mathrm{ext}}_{\check{T}}$. Let's resolve this.

Without loss of generality we assume $\check{\mathcal{E}}^{\mathrm{int}}_{\check{T}} \neq \varnothing$,[1] so the Galerkin problem

$$\langle \nabla w_{\check{T}}, \nabla v\rangle_{\check{T}} = F_{\check{T}}(v) := \sum_{\check{e}\in\check{\mathcal{E}}^{\mathrm{ext}}_{\check{T}}} \langle \boldsymbol{\sigma}^{(1)}_{\check{T}} \cdot \boldsymbol{n}_{\check{T}}, v\rangle_{\check{e}} \ (v \in \mathscr{H}_{\check{T}}) \text{ where } \mathscr{H}_{\check{T}} := H^1_{0,\check{\mathcal{E}}^{\mathrm{int}}_{\check{T}}}(\check{T}),$$

has a unique solution $w_{\check{T}} \in \mathscr{H}_{\check{T}}$ for which it follows directly that

$$\operatorname{div} \nabla w_{\check{T}} = 0, \quad |||w_{\check{T}}|||_{\check{T}} \le \|F_{\check{T}}\|_{\mathscr{H}'_{\check{T}}}, \quad \nabla w_{\check{T}} \cdot \boldsymbol{n}_{\check{T}} = -\boldsymbol{\sigma}^{(1)}_{\check{T}} \cdot \boldsymbol{n}_{\check{T}} \text{ for } \check{e} \in \check{\mathcal{E}}^{\mathrm{ext}}_{\check{T}}.$$

[1] When $\check{\mathcal{E}}^{\mathrm{int}}_{\check{T}} = \varnothing$, then $\check{\mathcal{T}}_a$ consists of a single element $\check{T}$, in which case $H^1_*(\omega_a) = H^1_{0,\check{\mathcal{E}}^{\mathrm{ext}}_a}(\check{T})$ so that we may invoke the assumption $\sup_p S^{(A)}_{\check{\mathcal{E}}^{\mathrm{ext}}_a,p,q(p)} < \infty$ directly to find the saturation result (8.4.2).

Now, integration by parts tells us that

$$F_{\check{T}}(v) = \langle \operatorname{div} \boldsymbol{\sigma}_{\check{T}}^{(1)}, v\rangle_{\check{T}} + \langle \boldsymbol{\sigma}_{\check{T}}^{(1)}, \nabla v\rangle_{\check{T}} \quad \big(v \in \mathscr{H}_{\check{T}}\big)\,.$$

Then, through (8.4.8) and (for the final inequality) $H_0^1(\check{T}) \subset \mathscr{H}_{\check{T}}$, we have

$$\begin{aligned}\|F_{\check{T}}\|_{\mathscr{H}_{\check{T}}'} &\le \|\langle \operatorname{div} \boldsymbol{\sigma}_{\check{T}}^{(1)}, \cdot\rangle_{\check{T}}\|_{\mathscr{H}_{\check{T}}'} + \|\boldsymbol{\sigma}_{\check{T}}^{(1)}\|_{\check{T}} \lesssim \|\langle \boldsymbol{\phi}_{\check{T}}, \cdot\rangle_{\check{T}}\|_{\mathscr{H}_{\check{T}}'} + \|\langle \boldsymbol{\phi}_{\check{T}}, \cdot\rangle_{\check{T}}\|_{H_0^1(\check{T})'} \\ &\le 2\|\langle \boldsymbol{\phi}_{\check{T}}, \cdot\rangle_{\check{T}}\|_{\mathscr{H}_{\check{T}}'},\end{aligned}$$

so that $\|\!|w_{\check{T}}|\!\|_{\check{T}} \lesssim \|\langle \boldsymbol{\phi}_{\check{T}}, \cdot\rangle_{\check{T}}\|_{\mathscr{H}_{\check{T}}'}$. We invoke Lemma 8.3.3 with $\gamma := \cup_{\check{e}\in\check{\mathcal{E}}_{\check{T}}^{\mathrm{ext}}}\check{e}$, $\varphi \in L_2(\gamma)$ piecewise defined as $\varphi|_{\check{e}} := -\boldsymbol{\sigma}_{\check{T}}^{(1)} \cdot \boldsymbol{n}_{\check{T}}$, and take $\boldsymbol{\tau} := \nabla w_{\check{T}}$, yielding a $\boldsymbol{\sigma}_{\check{T}}^{(2)} \in \mathrm{RT}_{p_a+1}(\check{T})$ for which

$$\operatorname{div} \boldsymbol{\sigma}_{\check{T}}^{(2)} = 0, \quad \boldsymbol{\sigma}_{\check{T}}^{(2)} \cdot \boldsymbol{n}_{\check{T}} = -\boldsymbol{\sigma}_{\check{T}}^{(1)} \cdot \boldsymbol{n}_{\check{T}} \text{ for } \check{e} \in \check{\mathcal{E}}_{\check{T}}^{\mathrm{ext}}, \quad \|\boldsymbol{\sigma}_{\check{T}}^{(2)}\|_{\check{T}} \lesssim \|\langle \boldsymbol{\phi}_{\check{T}}, \cdot\rangle_{\check{T}}\|_{\mathscr{H}_{\check{T}}'}, \tag{8.4.9}$$

so that $\boldsymbol{\sigma}_{\check{T}} := \boldsymbol{\sigma}_{\check{T}}^{(1)} + \boldsymbol{\sigma}_{\check{T}}^{(2)}$ has bounded norm, with normal components vanishing on $\check{\mathcal{E}}_{\check{T}}^{\mathrm{ext}}$. We then define $\tilde{r}_{\check{T}} \in H_*^1(\omega_a)'$ and $\tilde{r}_a^{(0)} \in H_*^1(\omega_a)'$ as

$$\tilde{r}_{\check{T}}(v) := \langle \boldsymbol{\sigma}_{\check{T}}, \nabla v\rangle_{\check{T}} \quad \text{and} \quad \tilde{r}_a^{(0)} := \sum_{\check{T}\in\check{\mathcal{T}}_a} \tilde{r}_{\check{T}}.$$

Step (A2) We verify (8.4.4). Partial integration yields that for $v \in H_*^1(\omega_a)$,

$$\tilde{r}_{\check{T}}(v) = -\langle \boldsymbol{\phi}_{\check{T}}, v\rangle_{\check{T}} + \sum_{\check{e}\in\check{\mathcal{E}}_{\check{T}}^{\mathrm{int}}} \langle \boldsymbol{\phi}_{\check{T},\check{e}}^{(0)}, v\rangle_{\check{e}}, \quad \text{where} \quad \boldsymbol{\phi}_{\check{T},\check{e}}^{(0)} := \boldsymbol{\sigma}_{\check{T}} \cdot \boldsymbol{n}_{\check{T}} \in \mathbb{P}_{p_a+1}(\check{e}).$$

Therefore, $\tilde{r}_a^{(0)}$ removes all element contributions from $\tilde{r}_a$; it follows that indeed, $\tilde{r}_a - \tilde{r}_a^{(0)}$ is a sum of contributions over (interior) edges:

$$\tilde{r}_a(v) - \tilde{r}_a^{(0)}(v) = \sum_{\check{e}\in\check{\mathcal{E}}_a^{\mathrm{int}}} \langle \boldsymbol{\phi}_{\check{e}}, v\rangle_{\check{e}} - \sum_{(\check{T},\check{e})\in\check{\mathcal{T}}_a\times\check{\mathcal{E}}_{\check{T}}^{\mathrm{int}}} \langle \boldsymbol{\phi}_{\check{T},\check{e}}^{(0)}, v\rangle_{\check{e}} =: \sum_{\check{e}\in\check{\mathcal{E}}_a^{\mathrm{int}}} \langle \boldsymbol{\phi}_{\check{e}}^{(0)}, v\rangle_{\check{e}} \ \ (v \in H_*^1(\omega_a)).$$

Now, every $\boldsymbol{\phi}_{\check{e}}^{(0)}$ is a sum of polynomials $\boldsymbol{\phi}_{\check{T},\check{e}}^{(0)}$, so indeed $\boldsymbol{\phi}_{\check{e}}^{(0)} \in \mathbb{P}_{p_a+1}(\check{e})$.

Step (A3) We verify (8.4.3). By definition, $\tilde{r}_{\check{T}}(\mathbb{1}) = 0$. Cauchy–Schwarz, (8.4.8), and (8.4.9) imply

$$\|\tilde{r}_{\check{T}}\|_{H_*^1(\omega_a)'} \le \|\boldsymbol{\sigma}_{\check{T}}\|_{\check{T}} = \|\boldsymbol{\sigma}_{\check{T}}^{(1)} + \boldsymbol{\sigma}_{\check{T}}^{(2)}\|_{\check{T}} \lesssim \|\langle \boldsymbol{\phi}_{\check{T}}, \cdot\rangle_{\check{T}}\|_{\mathscr{H}_{\check{T}}'}.$$

Moreover, $\check{\mathcal{E}}_{\check{T}}^{\mathrm{int}} \neq \varnothing$, hence it can be identified with a set $\mathcal{E} \in \mathbb{E}^{(A)}$ from (8.4.1). By assumption, $\sup_p S_{\mathcal{E},p+1,q(p+1)}^{(A)} \le \hat{S}$, so that through Lemma 8.4.1, we find

$$\|\langle \boldsymbol{\phi}_{\check{T}}, \cdot\rangle_{\check{T}}\|_{\mathscr{H}_{\check{T}}'} \lesssim \|\langle \boldsymbol{\phi}_{\check{T}}, \cdot\rangle_{\check{T}}\|_{\mathscr{V}_{\check{T}}'}, \quad \text{where} \quad \mathscr{V}_{\check{T}} := \mathscr{H}_{\check{T}} \cap \mathbb{Q}_{q(p_a+1)}(\check{T}).$$

Every $v \in \mathscr{V}_{\check{T}}$ vanishes on interior edges; write its zero-extension to ω_a as $\overline{v} \in H^1(\omega_a)$. Then

$$H^1_{*,q(p_a+1)}(\check{\mathcal{T}}_a) \ni \overline{\overline{v}} := \begin{cases} \overline{v} - \langle \overline{v}, \mathbb{1} \rangle_{\omega_a} & a \in \mathcal{V}_{\mathcal{T}}^{\mathrm{int}}, \\ \overline{v}, & a \in \mathcal{V}_{\mathcal{T}}^{\mathrm{ext}}. \end{cases}$$

By $\tilde{r}_a(\mathbb{1}) = 0$ for $a \in \mathcal{V}_{\mathcal{T}}^{\mathrm{int}}$, we have $\langle \phi_{\check{T}}, v \rangle_{\check{T}} = \tilde{r}_a(\overline{v}) = \tilde{r}_a(\overline{\overline{v}})$; moreover, $|||v|||_{\check{T}} = |||\overline{\overline{v}}|||_{\omega_a}$, so

$$\begin{aligned} \|\langle \phi_{\check{T}}, \cdot \rangle_{\check{T}}\|_{\mathscr{V}'_{\check{T}}} &:= \sup_{0 \neq v \in \mathscr{V}_{\check{T}}} \frac{|\langle \phi_{\check{T}}, v \rangle_{\check{T}}|}{|||v|||_{\check{T}}} = \sup_{0 \neq v \in \mathscr{V}_{\check{T}}} \frac{\left|\tilde{r}_a(\overline{\overline{v}})\right|}{|||\overline{\overline{v}}|||_{\omega_a}} \\ &\leq \sup_{0 \neq w \in H^1_{*,q(p_a+1)}(\check{\mathcal{T}}_a)} \frac{\left|\tilde{r}_a(w)\right|}{|||w|||_{\omega_a}} =: \|\tilde{r}_a\|_{H^1_{*,q(p_a+1)}(\check{\mathcal{T}}_a)'}. \end{aligned}$$

Chaining the dual norm inequalities in this step yields (8.4.3).

Step (B0) We traverse the interior edges $\check{\mathcal{E}}_a^{\mathrm{int}}$ in the order $(\check{e}_j)_{j=1}^{n_a}$ constructed in §8.4.2. Let $(\check{T}_j)_{j=1}^{n_a}$ be the sequence of squares for each $\check{e}_j$.

At each iteration of the traversal, we use result (1) of Lemma 8.4.3 to remove the $\check{e}_i$-contribution from the previous residual by—in a fashion similar to Step (A1)—solving a local Galerkin problem and constructing a Raviart–Thomas flux σ_i with specific properties. The resulting functional $\tilde{r}_{\check{e}_i} \in H^1_*(\omega_a)'$ will be found as $\langle \sigma_i, \nabla v \rangle_{\check{T}_i}$. We then use result (2) of the Lemma to establish the dual norm bound of (8.4.5), similar to Step (A3).

We continue by induction. Let $i = 1$.

Step (B1) We construct $\tilde{r}_{\check{e}_i}$. By result (1) of Lemma 8.4.3, we have $\check{\mathcal{E}}^{\mathrm{loc},D}_{a,i} \neq \varnothing$, so the problem

$$\langle \nabla w^{(i)}, \nabla v \rangle_{\check{T}_i} = \langle \phi^{(i-1)}_{\check{e}_i}, v \rangle_{\check{e}_i} \quad (v \in \mathscr{H}_i) \quad \text{where} \quad \mathscr{H}_i := H^1_{0,\check{\mathcal{E}}^{\mathrm{loc},D}_{a,i}}(\check{T}_i) \tag{8.4.10}$$

has a unique solution $w^{(i)} \in \mathscr{H}_i$ for which it holds that $\operatorname{div} \nabla w^{(i)} = 0$, and

$$|||w^{(i)}|||_{\check{T}_i} \leq \|\langle \phi^{(i-1)}_{\check{e}_i}, \cdot \rangle_{\check{e}_i}\|_{\mathscr{H}'_i}, \quad \begin{cases} \nabla w^{(i)} \cdot \boldsymbol{n}_{\check{T}_i} = -\phi^{(i-1)}_{\check{e}_i} \text{ on } \check{e}_i, \\ \nabla w^{(i)} \cdot \boldsymbol{n}_{\check{T}_i} = 0 \text{ on } \left\{ \check{e}_j \in \check{\mathcal{E}}^{\mathrm{int}}_{\check{T}_i} : j < i \right\}, \\ \nabla w^{(i)} \cdot \boldsymbol{n}_{\check{T}_i} = 0 \text{ on } \check{\mathcal{E}}^{\mathrm{ext},N}_{\check{T}_i}. \end{cases}$$

By Lemma 8.3.3, there is a $\sigma_i \in \mathrm{RT}_{p_a+1}(\check{T}_i)$ with the same normal components on $\check{\mathcal{E}}_{\check{T}_i} \setminus \check{\mathcal{E}}^{\mathrm{loc},D}_{a,i}$ as $\nabla w^{(i)}$, such that

$$\operatorname{div} \sigma_i = 0, \quad \|\sigma_i\|_{\check{T}_i} \lesssim \|\langle \phi^{(i-1)}_{\check{e}_i}, \cdot \rangle_{\check{e}_i}\|_{\mathscr{H}'_i}. \tag{8.4.11}$$

We then define $\tilde{r}_{\check{e}_i} \in H^1_*(\omega_a)'$ and $\tilde{r}_a^{(i)} \in H^1_*(\omega_a)'$ as

$$\tilde{r}_{\check{e}_i}(v) := \langle \boldsymbol{\sigma}_i, \nabla v\rangle_{\check{T}_i} \quad \text{and} \quad \tilde{r}_a^{(i)}(v) := \tilde{r}_a^{(i-1)}(v) + \tilde{r}_{\check{e}_i}(v) \quad (v \in H^1_*(\omega_a)).$$

Step (B2) Let us look at (8.4.6). In light of (8.4.4) when $i = 1$, or (8.4.6) for $i \geq 2$, suppose we have

$$\tilde{r}_a(v) - \tilde{r}_a^{(i-1)}(v) = \sum_{j\geq i} \langle \phi_{\check{e}_j}^{(i-1)}, v\rangle_{\check{e}_j} \text{ for some } \phi_{\check{e}_j}^{(i-1)} \in \mathbb{P}_{p_a+1}(\check{e}_j) \quad (v \in H^1_*(\omega_a)). \tag{8.4.12}$$

Using that $v \in H^1_*(\omega_a)$ vanishes along edges in $\check{\mathcal{E}}^{\mathrm{ext},D}_{\check{T}_i}$, and considering the normal components of $\boldsymbol{\sigma}_i$, integration by parts yields (8.4.6):

$$\begin{aligned}
\tilde{r}_a(v) - \tilde{r}_a^{(i)}(v) &= \Big[\tilde{r}_a(v) - \tilde{r}_a^{(i-1)}(v)\Big] - \tilde{r}_{\check{e}_i}(v) = \sum_{j\geq i} \langle \phi_{\check{e}_j}^{(i-1)}, v\rangle_{\check{e}_j} - \langle \boldsymbol{\sigma}_i \cdot \boldsymbol{n}_{\check{T}_i}, v\rangle_{\partial \check{T}_i} \\
&= \sum_{j\geq i} \langle \phi_{\check{e}_j}^{(i-1)}, v\rangle_{\check{e}_j} - \langle \boldsymbol{\sigma}_i \cdot \boldsymbol{n}_{\check{T}_i}, v\rangle_{\check{e}_i} - \sum_{\check{e}_j \in \check{\mathcal{E}}^{\mathrm{int}}_{\check{T}_i} : j>i} \langle \boldsymbol{\sigma}_i \cdot \boldsymbol{n}_{\check{T}_i}, v\rangle_{\check{e}_j} \\
&= \sum_{j\geq i+1} \langle \phi_{\check{e}_j}^{(i)}, v\rangle_{\check{e}_j} - \sum_{\check{e}_j \in \check{\mathcal{E}}^{\mathrm{int}}_{\check{T}_i} : j>i} \langle \boldsymbol{\sigma}_i \cdot \boldsymbol{n}_{\check{T}_i}, v\rangle_{\check{e}_j} \\
&=: \sum_{j\geq i+1} \langle \phi_{\check{e}_j}^{(i)}, v\rangle_{\check{e}_j} \quad \text{for some} \quad \phi_{\check{e}_j}^{(i)} \in \mathbb{P}_{p_a+1}(\check{e}_j).
\end{aligned}$$

Step (B3) We verify (8.4.5). By definition, $\tilde{r}_{\check{e}_i}(\mathbb{1}) = 0$. Moreover, by result (1) of Lemma 8.4.3, $\check{\mathcal{E}}^{\mathrm{loc},D}_{a,i}$ corresponds with an $\mathcal{E} \in \mathbb{E}^{(B)}$ from (8.4.1).[2] By assumption, $S^{(B)}_{\mathcal{E},p_a+1,q(p_a+1)} \leq \hat{S}$, so (8.4.11) and Lemma 8.4.1 yield

$$\|\tilde{r}_{\check{e}_i}\|_{H^1_*(\omega_a)'} \leq \|\boldsymbol{\sigma}_i\|_{\check{T}_i} \lesssim \|\langle \phi_{\check{e}_i}^{(i-1)}, \cdot\rangle_{\check{e}_i}\|_{\mathscr{H}_i'} \lesssim \|\langle \phi_{\check{e}_i}^{(i-1)}, \cdot\rangle_{\check{e}_i}\|_{\mathscr{V}_i'},$$

where $\mathscr{V}_i := \mathscr{H}_i \cap \mathbb{Q}_{q(p_a+1)}(\check{T}_i)$. To establish (8.4.5), it suffices to show

$$\|\langle \phi_{\check{e}_i}^{(i-1)}, \cdot\rangle_{\check{e}_i}\|_{\mathscr{V}_i'} \lesssim \|\tilde{r}_a - \tilde{r}_a^{(i-1)}\|_{H^1_{*,q(p_a+1)+1}(\check{\mathcal{T}}_a)'}, \tag{8.4.13}$$

as by $\tilde{r}_a - \tilde{r}_a^{(i-1)} = \tilde{r}_a - \tilde{r}_a^{(0)} + \sum_{j=1}^{i-1} \tilde{r}_{\check{e}_j}$, a triangle inequality, (8.4.5), and $\#\check{\mathcal{T}}_a \leq 16$,

$$\|\tilde{r}_a - \tilde{r}_a^{(i-1)}\|_{H^1_{*,q(p_a+1)+1}(\check{\mathcal{T}}_a)'} \lesssim \|\tilde{r}_a - \tilde{r}_a^{(0)}\|_{H^1_{*,q(p_a+1)+1}(\check{\mathcal{T}}_a)'} \lesssim \|\tilde{r}_a\|_{H^1_{*,q(p_a+1)+1}(\check{\mathcal{T}}_a)'}.$$

We proceed as in Step (A3). Take $v \in \mathscr{V}_i$. Result (2) of Lemma 8.4.3 guarantees a bounded extension from v to a $Ev \in H^1(\omega_a) \cap \mathbb{Q}^{-1}_{q(p_a+1)+1}(\check{\mathcal{T}}_a)$ that vanishes on interior edges $\check{e}_j$ with $j > i$. Moreover, Ev is zero on edges $\check{e} \in \check{\mathcal{E}}^{\mathrm{ext},D}_a$

[2] The set $\check{\mathcal{E}}^{\mathrm{loc},D}_{a,i}$ is in one of five states, whereas $\mathbb{E}^{(B)}$ has four; situation (a) and (b) of Lemma 8.4.3 correspond with the same $\mathcal{E} \in \mathbb{E}^{(B)}$.

whenever $a \in \mathcal{V}_{\mathcal{T}}^{\mathrm{ext}}$, so that in fact

$$H^1_{*,q(p_a+1)+1}(\check{\mathcal{T}}_a) \ni \overline{\overline{v}} := \begin{cases} Ev - \langle Ev, \mathbb{1}\rangle_{\omega_a} & a \in \mathcal{V}_{\mathcal{T}}^{\mathrm{int}}, \\ Ev & a \in \mathcal{V}_{\mathcal{T}}^{\mathrm{ext}}, \end{cases}$$

with $\|\!|\overline{\overline{v}}|\!\|_{\omega_a} = \|\!|Ev|\!\|_{\omega_a} \lesssim \|\!|v|\!\|_{\check{T}}$.

Now, $\tilde{r}_a(\mathbb{1}) - \tilde{r}_a^{(i-1)}(\mathbb{1}) = 0$ when $a \in \mathcal{V}_{\mathcal{T}}^{\mathrm{int}}$, so $\tilde{r}_a(\overline{\overline{v}}) - \tilde{r}_a^{(i-1)}(\overline{\overline{v}}) = \tilde{r}_a(Ev) - \tilde{r}_a^{(i-1)}(Ev)$. Moreover, $Ev|_{\check{e}_i} = v|_{\check{e}_i}$ and $Ev|_{\check{e}_j} = 0$ for $j > i$, so almost all terms of (8.4.12) vanish when we plug in Ev, which yields $\tilde{r}_a(Ev) - \tilde{r}_a^{(i-1)}(Ev) = \langle \phi_{\check{e}_i}^{(i-1)}, Ev\rangle_{\check{e}_i} = \langle \phi_{\check{e}_i}^{(i-1)}, v\rangle_{\check{e}_i}$. Then, (8.4.13) follows by

$$\begin{aligned} \|\langle \phi_{\check{e}_i}^{(i-1)}, \cdot\rangle_{\check{e}_i}\|_{\mathcal{V}_i'} &= \sup_{0\neq v\in\mathcal{V}_i} \frac{\left|\langle \phi_{\check{e}_i}^{(i-1)}, v\rangle_{\check{e}_i}\right|}{\|\!|v|\!\|_{\check{T}_i}} \lesssim \sup_{0\neq v\in\mathcal{V}_i} \frac{\left|\tilde{r}_a(\overline{\overline{v}}) - \tilde{r}_a^{(i-1)}(\overline{\overline{v}})\right|}{\|\!|\overline{\overline{v}}|\!\|_{\omega_a}} \\ &\leq \|\tilde{r}_a - \tilde{r}_a^{(i-1)}\|_{H^1_{*,q(p_a+1)+1}(\mathcal{T}_a)'}. \end{aligned}$$

Step (B4) We repeat Steps (B1)–(B3) for $i \in \{2, \dots, n_a - 1\}$, at each step finding functionals $\tilde{r}_{\check{e}_i}$ and $\tilde{r}_a^{(i)}$ for which (8.4.5) and (8.4.6) hold.

Step (C) When $a \in \mathcal{V}_{\mathcal{T}}^{\mathrm{ext}}$, the results of Lemma 8.4.3 are satisfied once more for $i = n_a$. This allows us to repeat Steps (B1)–(B3) for a $\tilde{r}_{\check{e}_{n_a}} \in H^1_*(\omega_a)'$ satisfying (8.4.5) and (8.4.6).

When $a \in \mathcal{V}_{\mathcal{T}}^{\mathrm{int}}$, we cannot continue the iteration; the set $\mathcal{E}_{n_a}^{\mathrm{loc},D}$ is empty so we cannot solve (8.4.10). However, from (8.4.6), we find for $v \in H^1_*(\omega_a)$,

$$\tilde{r}_a(v) - \tilde{r}_a^{(n_a-1)}(v) = \langle \phi_{\check{e}_{n_a}}^{(n_a-1)}, v\rangle_{\check{e}_{n_a}} \quad \text{for some} \quad \phi_{\check{e}_{n_a}}^{(n_a-1)} \in \mathbb{P}_{p_a+1}(\check{e}_j).$$

With $\|\!|v - \langle v, \mathbb{1}\rangle_{\check{T}_{n_a}}|\!\|_{\check{T}_{n_a}} \leq \|\!|v|\!\|_{\omega_a}$ and $\langle \phi_{\check{e}_{n_a}}^{(n_a-1)}, \mathbb{1}\rangle_{\check{e}_{n_a}} = 0$ (by $\tilde{r}_a(\mathbb{1}) = 0 = \tilde{r}_a^{(n_a-1)}(\mathbb{1})$),

$$\|\langle \phi_{\check{e}_{n_a}}^{(n_a-1)}, \cdot\rangle_{\check{e}_{n_a}}\|_{H^1_*(\omega_a)'} \leq \|\langle \phi_{\check{e}_{n_a}}^{(n_a-1)}, \cdot\rangle_{\check{e}_{n_a}}\|_{\mathscr{H}_{n_a}'} \quad \text{where} \quad \mathscr{H}_{n_a} := H^1(\check{T}_{n_a})/\mathbb{R}.$$

The fact $\langle \phi_{\check{e}_{n_a}}^{(n_a-1)}, \mathbb{1}\rangle_{\check{e}_{n_a}} = 0$ also allows us to use $S^{(C)}_{p_a,q(p_a+1)} \leq \hat{S}$ and invoke Lemma 8.4.1 yielding, with $\mathscr{V}_{n_a} := \mathscr{H}_{n_a} \cap \mathbb{Q}_{q(p_a+1)}(\check{T}_{n_a})$,

$$\|\langle \phi_{\check{e}_{n_a}}^{(n_a-1)}, \cdot\rangle_{\check{e}_{n_a}}\|_{\mathscr{H}_{n_a}'} \lesssim \|\langle \phi_{\check{e}_{n_a}}^{(n_a-1)}, \cdot\rangle_{\check{e}_{n_a}}\|_{\mathscr{V}_{n_a}'}.$$

Any $v \in \mathscr{V}_{n_a}$ has zero mean, so reflecting across every row and column of $\check{\mathcal{T}}_a$ yields a mean-zero extension $\overline{v} \in H^1_{*,q(p_a+1)}(\check{\mathcal{T}}_a)$ with, by $\#\check{\mathcal{T}}_a \leq 16$, norm $\|\!|\overline{v}|\!\|_{\omega_a} = \sqrt{\#\check{\mathcal{T}}_a}\|\!|v|\!\|_{\check{T}_{n_a}} \lesssim \|\!|v|\!\|_{\check{T}_{n_a}}$.

Finally, by $\langle \phi^{(n_a-1)}_{\check{e}_{n_a}}, \overline{v}\rangle_{\check{e}_{n_a}} = \langle \phi^{(n_a-1)}_{\check{e}_{n_a}}, v\rangle_{\check{e}_{n_a}}$, we have the dual norm bound

$$\begin{aligned}\|\langle \phi^{(n_a-1)}_{\check{e}_{n_a}}, \cdot\rangle_{\check{e}_{n_a}}\|_{\mathscr{V}'_{n_a}} &= \sup_{0\neq v\in\mathscr{V}_{n_a}} \frac{\left|\langle \phi^{(n_a-1)}_{\check{e}_{n_a}}, \overline{v}\rangle_{\check{e}_{n_a}}\right|}{|||v|||_{\check{T}_{n_a}}} \lesssim \sup_{0\neq v\in\mathscr{V}_{n_a}} \frac{\left|\langle \phi^{(n_a-1)}_{\check{e}_{n_a}}, \overline{v}\rangle_{\check{e}_{n_a}}\right|}{|||\overline{v}|||_{\omega_a}} \\ &\leq \|\langle \phi^{(n_a-1)}_{\check{e}_{n_a}}, \cdot\rangle_{\check{e}_{n_a}}\|_{H^1_{*,\mathcal{A}(p_a+1)}(\check{T}_a)'};\end{aligned}$$

then, through the triangle inequality, we find

$$\begin{aligned}\|\langle \phi^{(n_a-1)}_{\check{e}_{n_a}}, \cdot\rangle_{\check{e}_{n_a}}\|_{H^1_{*,\mathcal{A}(p_a+1)}(\check{T}_a)'} &\leq \|\tilde{r}_a\|_{H^1_{*,\mathcal{A}(p_a+1)}(\check{T}_a)'} + \|\tilde{r}_a^{(n_a-1)}\|_{H^1_{*,\mathcal{A}(p_a+1)}(\check{T}_a)'} \\ &\lesssim \|\tilde{r}_a\|_{H^1_{*,\mathcal{A}(p_a+1)}(\check{T}_a)'}.\end{aligned}$$

Chaining the dual norm inequalities in Step (C) then yields (8.4.7). □

8.5 Computation of reference saturation coefficients

In this section, we detail on the numerical computation of the saturation coefficient $S(\hat{\mathscr{H}}, \hat{\mathscr{V}}, \hat{\mathscr{F}})$. To allow computation of this coefficient, we first write it as the solution to a generalized Eigenvalue problem. We then discuss the construction of bases for the spaces $\hat{\mathscr{H}}$, $\hat{\mathscr{V}}$, and $\hat{\mathscr{F}}$ involved in the specific saturation coefficients of Theorem 8.4.2.

8.5.1 An equivalent problem

In our applications, $\hat{\mathscr{F}}$ is a *finite-dimensional subspace* of $\hat{\mathscr{H}}'$ rather than just a subset, allowing us to write $S(\hat{\mathscr{H}}, \hat{\mathscr{V}}, \hat{\mathscr{F}})$ as $\sup_{\left\{\hat{F}\in\hat{\mathscr{F}}:\|\hat{F}\|_{\hat{\mathscr{V}}'}=1\right\}} \|\hat{F}\|_{\hat{\mathscr{H}}'}$. Since $\hat{\mathscr{F}}$ is finite-dimensional, this supremum is attained, so we may equivalently solve the problem of finding the largest $0 < \mu = S(\hat{\mathscr{H}}, \hat{\mathscr{V}}, \hat{\mathscr{F}})$ such that

$$\text{for some } 0 \neq \hat{F} \in \hat{\mathscr{F}}, \quad \|\hat{F}\|^2_{\hat{\mathscr{H}}'} = \mu^2 \|\hat{F}\|^2_{\hat{\mathscr{V}}'}. \tag{8.5.1}$$

Proposition 8.5.1 (Equivalent Eigenvalue problem)**.** *Let $\Xi_{\hat{\mathscr{H}}}$, $\Xi_{\hat{\mathscr{V}}}$, and $\Sigma_{\hat{\mathscr{F}}}$ be bases for the three spaces. For $\hat{U} \in \{\hat{\mathscr{H}}, \hat{\mathscr{V}}\}$, define stiffness- and load matrices*

$$\boldsymbol{A}_{\hat{U}} := \langle \nabla \Xi_{\hat{U}}, \nabla \Xi_{\hat{U}}\rangle_{\hat{T}} = \left[\langle \nabla \xi_{i,\hat{U}}, \nabla \xi_{j,\hat{U}}\rangle_{\hat{T}}\right]_{i,j=1}^{\#\Xi_{\hat{U}}}, \boldsymbol{L}_{\hat{U}} := \Sigma_{\hat{\mathscr{F}}}(\Xi_{\hat{U}}) := \left[\sigma_i(\xi_{j,\hat{U}})\right]_{i,j}$$

Then problem (8.5.1) *is equivalent to finding the largest generalized Eigenvalue μ^2 of*

$$\boldsymbol{R}_{\hat{\mathscr{H}}}\boldsymbol{F} = \mu^2 \boldsymbol{R}_{\hat{\mathscr{V}}}\boldsymbol{F}, \quad \textit{where} \quad \boldsymbol{R}_{\hat{U}} := \boldsymbol{L}_{\hat{U}} \boldsymbol{A}_{\hat{U}}^{-\top} \boldsymbol{L}_{\hat{U}}^{\top}. \tag{8.5.2}$$

Proof. For $\hat{U} \in \{\hat{\mathscr{H}}, \hat{\mathscr{V}}\}$ and $\hat{F} \in \hat{\mathscr{F}}$, take $u_{\hat{U}} = u_{\hat{U}}(\hat{F})$ the unique solution to

$$\langle \nabla u_{\hat{U}}, \nabla v_{\hat{U}}\rangle_{\hat{T}} = \hat{F}(v_{\hat{U}}) \quad (v_{\hat{U}} \in \hat{U}).$$

Recalling that we equip $\hat{U}$ with $|||\cdot|||_{\hat{T}} := \|\nabla\cdot\|_{\hat{T}}$, and so $\hat{U}'$ with its dual norm, $\|\hat{F}\|_{\hat{U}'} = |||u_{\hat{U}}|||_{\hat{T}}$. Write $\hat{F} = \boldsymbol{F}^\top \Sigma_{\mathscr{F}}$ and $u_{\hat{U}} = \boldsymbol{u}_{\hat{U}}^\top \Xi_{\hat{U}}$; then $\boldsymbol{u}_{\hat{U}}^\top \boldsymbol{A}_{\hat{U}} = \langle \nabla u_{\hat{U}}, \nabla \Xi_{\hat{U}} \rangle_{\hat{T}} = \hat{F}(\Xi_{\hat{U}}) = \boldsymbol{F}^\top \boldsymbol{L}_{\hat{U}}$, or $\boldsymbol{A}_{\hat{U}}^\top \boldsymbol{u}_{\hat{U}} = \boldsymbol{L}_{\hat{U}}^\top \boldsymbol{F}$. Now, $\boldsymbol{A}_{\hat{U}}$ is invertible, so $\boldsymbol{u}_{\hat{U}} = \boldsymbol{A}_{\hat{U}}^{-\top} \boldsymbol{L}_{\hat{U}}^\top \boldsymbol{F}$. We see $\|\hat{F}\|_{\hat{U}'}^2 = |||u_{\hat{U}}|||_{\hat{T}}^2 = \boldsymbol{u}_{\hat{U}}^\top \boldsymbol{A}_{\hat{U}} \boldsymbol{u}_{\hat{U}} = \boldsymbol{F}^\top \boldsymbol{L}_{\hat{U}} \boldsymbol{A}_{\hat{U}}^{-\top} \boldsymbol{L}_{\hat{U}}^\top \boldsymbol{F} = \boldsymbol{F}^\top \boldsymbol{R}_{\hat{U}} \boldsymbol{F}$, reducing problem (8.5.1) to finding the largest $\mu > 0$ s.t.

$$\text{for some } \boldsymbol{F} \neq \boldsymbol{0}, \quad \boldsymbol{F}^\top \boldsymbol{R}_{\mathscr{H}} \boldsymbol{F} = \mu^2 \boldsymbol{F}^\top \boldsymbol{R}_{\mathscr{V}} \boldsymbol{F} \iff \mu^2 = \frac{\boldsymbol{F}^\top \boldsymbol{R}_{\mathscr{H}} \boldsymbol{F}}{\boldsymbol{F}^\top \boldsymbol{R}_{\mathscr{V}} \boldsymbol{F}},$$

which, as both $\boldsymbol{R}_{\mathscr{H}}$ and $\boldsymbol{R}_{\mathscr{V}}$ are symmetric positive-definite, is a Rayleigh quotient for the generalized Eigenvalue problem of (8.5.2) (cf. [Li15]). □

8.5.2 Discrete saturation coefficients

In all of the cases of Theorem 8.4.2, the space $\mathscr{H}$ is an infinite-dimensional closed subspace of $H^1(\hat{T})$, so computing $S(\mathscr{H}, \mathscr{V}, \mathscr{F})$ by means of (8.5.2) will likely not be possible. However, the following result shows that we may restrict ourselves to a finite-dimensional subspace that is large enough.

Lemma 8.5.2. *Since $\mathscr{F}$ is a finite-dimensional subspace of $\mathscr{H}'$, a compactness argument shows that the* discrete saturation coefficient $S(\mathscr{H} \cap \mathbb{Q}_r(\hat{T}), \mathscr{V}, \mathscr{F})$ *tends to $S(\mathscr{H}, \mathscr{V}, \mathscr{F})$ for $r \to \infty$.*

8.5.3 Bases for the subspaces

Solving (8.5.2) hinges on computing the stiffness matrix $\boldsymbol{A}_{\mathscr{H}}$ and load matrix $\boldsymbol{L}_{\mathscr{H}}$, which depend on the choice of basis. In practice, we are able to choose these bases as tensor products. For instance, when $\Xi_{\mathscr{H}} =: \Xi^1 \otimes \Xi^2$, we see

$$\boldsymbol{A}_{\mathscr{H}} = \langle \tfrac{\mathrm{d}}{\mathrm{d}x}\Xi^1, \tfrac{\mathrm{d}}{\mathrm{d}x}\Xi^1 \rangle_{\hat{I}} \otimes \langle \Xi^2, \Xi^2 \rangle_{\hat{I}} + \langle \Xi^1, \Xi^1 \rangle_{\hat{I}} \otimes \langle \tfrac{\mathrm{d}}{\mathrm{d}x}\Xi^2, \tfrac{\mathrm{d}}{\mathrm{d}x}\Xi^2 \rangle_{\hat{I}},$$

where $\hat{I} := [-1, 1]$, so that $\hat{T} = \hat{I} \times \hat{I}$, and with $\otimes$ denoting the Kronecker product. Essentially, computation of the saturation coefficient boils down to computing a number of inner products.

Define L_k as the kth Legendre polynomial, with $\deg L_k = k$ and $L_k(1) = 1$. The functions

$$\varphi_k(x) := \sqrt{k + \tfrac{1}{2}} L_k(x), \quad (k \geq 0)$$

then constitute an $L_2(\hat{I})$-orthonormal basis called the *Legendre* basis. Moreover, the functions

$$\xi_k(x) := \sqrt{k - \tfrac{1}{2}} \int_x^1 L_{k-1}(s)\,\mathrm{d}s = \frac{1}{\sqrt{4k-2}}(L_{k-2}(x) - L_k(x)), \quad (k \geq 2)$$

constitute an orthonormal basis with respect to the $H^1(\hat{I})$-seminorm which we call the *Babuška–Shen* basis. With respect to the $L_2(\hat{I})$-inner product, this basis is quasi-orthogonal in that

$$\langle \xi_k, \xi_m \rangle_{\hat{I}} = 0 \iff k - m \notin \{-2, 0, 2\}, \text{ and } \langle \varphi_k, \xi_m \rangle_{\hat{I}} = 0 \iff k - m \notin \{0, 2\}.$$

We can supplement the Babuška–Shen basis with $\xi_1(x) = \frac{1}{2}\sqrt{2}(1-x)$ to find an orthonormal basis for $H^1_{0,\{1\}}(\hat{I})$, and with $\tilde{\xi}_1(x) := \xi_1(-x)$ for a basis for $H^1_{0,\{-1\}}(\hat{I})$. The supplemented functions are $L_2(\hat{I})$-orthogonal to ξ_m for $m \geq 4$.

Recall the saturation coefficients from Theorem 8.4.2. The space $\hat{\mathscr{V}}$ is in every case just $\hat{\mathscr{H}}$ restricted to polynomials of lower degree, so we will focus on building bases for $\hat{\mathscr{H}}$ and $\hat{\mathscr{F}}$.

First discrete coefficient $S^{(A)}_{\mathcal{E},p,q,r}$ Denote

$$\hat{\mathscr{H}} := H^1_{0,\mathcal{E}}(\hat{T}) \cap \mathbb{Q}_r(\hat{T}), \quad \hat{\mathscr{F}} := \{h \mapsto \langle \phi, h\rangle_{\hat{T}} : \phi \in \mathbb{Q}_p(\hat{T})\} \subset \hat{\mathscr{H}}'.$$

A tensorized basis $\Xi_{\hat{\mathscr{H}}} = \Xi^1 \otimes \Xi^2$ for $\hat{\mathscr{H}}$ is readily constructed through the Babuška–Shen basis, supplemented to account for boundary conditions, up to degree r in each coordinate.

Choosing $\Phi := \Phi^1 \otimes \Phi^2$ with $\Phi^1 := \Phi^2$ the Legendre basis up to degree p, we set $\Sigma_{\hat{\mathscr{F}}} := \langle \Phi, \cdot \rangle_{\hat{T}}$. Then, the load matrix can be computed from $\boldsymbol{L}_{\hat{\mathscr{H}}} = \langle \Phi, \Xi_{\hat{\mathscr{H}}} \rangle_{\hat{T}} = \langle \Phi^1, \Xi^1 \rangle_{\hat{I}} \otimes \langle \Phi^2, \Xi^2 \rangle_{\hat{I}}$.

Second discrete coefficient $S^{(B)}_{\mathcal{E},p,q,r}$ The space $\hat{\mathscr{H}}$ is the same as in $S^{(A)}_{\mathcal{E},p,q,r}$ so its basis $\Xi_{\hat{\mathscr{H}}} = \Xi^1 \otimes \Xi^2$ is readily constructed. For $\hat{\mathscr{F}} := \left\{h \mapsto \langle \phi, h\rangle_{\hat{e}_1} : \phi \in \mathbb{P}_p(\hat{e}_1)\right\}$, we choose the basis $\Sigma_{\hat{\mathscr{F}}} := \langle \Phi, \cdot \rangle_{\hat{e}_1}$ with Φ the Legendre basis for $\mathbb{P}_p(\hat{e}_1)$. Then

$$\boldsymbol{L}_{\hat{\mathscr{H}}} = \langle \Phi, \Xi_{\hat{\mathscr{H}}} \rangle_{\hat{e}_1} = \Xi^1(1) \otimes \langle \Phi, \Xi^2 \rangle_{\hat{I}} \quad \text{where} \quad \Xi^1(x) := \big(\xi(x)\big)_{\xi \in \Xi^1}. \tag{8.5.3}$$

The polynomials $\xi \in \Xi^1$ with $\deg \xi \geq 2$ have $\xi(1) = 0$, so $\boldsymbol{L}_{\hat{\mathscr{H}}}$ is sparse with entire zero rows.

Third discrete coefficient $S^{(C)}_{p,q,r}$ To create a basis for $\hat{\mathscr{H}} := \mathbb{Q}_r(\hat{T})/\mathbb{R}$, we build

$$X := \left\{\chi_0, \xi_1 - \langle \xi_1, \mathbb{1}\rangle_{\hat{I}}, \xi_2 - \langle \xi_2, \mathbb{1}\rangle_{\hat{I}}, \dots, \xi_r - \langle \xi_r, \mathbb{1}\rangle_{\hat{I}}\right\}, \quad \text{where} \quad \chi_0 := \mathbb{1}/\sqrt{2}$$

which is a basis for $\mathbb{Q}_r(\hat{I})$, (almost) orthogonal w.r.t. the $H^1(\hat{I})$-seminorm, with every element except χ_0 having zero mean. The set $\Xi_{\hat{\mathscr{H}}} := X \times X \setminus \{\chi_0 \otimes \chi_0\}$ then consists of linearly independent polynomials with zero mean, and is of correct cardinality, hence a basis for $\hat{\mathscr{H}}$. Legendre polynomials ϕ_k of degree $k \geq 1$ have mean zero, so $\Phi_* := \left\{\phi_k : 1 \leq k \leq p\right\}$ is a basis for $\mathbb{P}_p(\hat{e}_1)/\mathbb{R}$, and $\Sigma_{\hat{\mathscr{F}}} := \langle \Phi_*, \cdot \rangle_{\hat{e}_1}$ a basis for $\hat{\mathscr{F}} := \left\{h \mapsto \langle \phi, h\rangle_{\hat{e}_1} : \phi \in \mathbb{P}_p(\hat{e}_1)/\mathbb{R}\right\}$. Its load matrix is formed analogously to (8.5.3).

8.6 Numerical verification

In Theorem 8.4.2, we showed that patch-based p-robust saturation holds, under the assumption that a number of quantities on the reference square are

finite. More specifically, we are interested in finding a function $q : \mathbb{N} \to \mathbb{N}$ such that the *saturation coefficients*

$$S^{(A)}_{\mathcal{E},p,q(p)}, \quad S^{(B)}_{\mathcal{F},p,q(p)}, \quad S^{(C)}_{p,q(p)} \quad (\mathcal{E} \in \mathbb{E}^{(A)}, \ \mathcal{F} \in \mathbb{E}^{(B)})$$

(cf. Thm. 8.4.2) are uniformly bounded in p. Unable to compute the limit $p \to \infty$, we resort to computing them for a number of large but finite values of p, and extrapolate from this progression. Moreover, with our current approach, we are unable to compute the above quantities, so we instead compute the *discrete saturation coefficients*

$$S^{(A)}_{\mathcal{E},p,q,r}, \quad S^{(B)}_{\mathcal{F},p,q,r}, \quad S^{(C)}_{p,q,r} \quad (\mathcal{E} \in \mathbb{E}^{(A)}, \ \mathcal{F} \in \mathbb{E}^{(B)})$$

(cf. Lemma 8.5.2) for values of r that are large relative to p and q, and expect to see *r-stabilization* of the discrete coefficient to its "continuous" counterpart.

In [CNSV17b], it was shown that (for a slightly different setting), a strategy of the form $q(p) = p + n$ with $n \in \mathbb{N}$ is insufficient, whereas for *any* $\lambda > 0$, the choice $q(p) = p + \lceil \lambda p \rceil$ exhibits saturation. This motivates our choice to run numerical computations for

$$q(p) = p + 4, \quad q(p) = p + \lceil p/7 \rceil, \quad q(p) = 2p.$$

By symmetry, we have five different configurations of sets of edges in $\mathbb{E}^{(A)}$:

$$\mathcal{E}_1 := \{\hat{e}_1\}, \quad \mathcal{E}_2 := \{\hat{e}_1, \hat{e}_2\}, \quad \mathcal{E}_3 := \{\hat{e}_1, \hat{e}_3\},$$
$$\mathcal{E}_4 := \{\hat{e}_1, \hat{e}_2, \hat{e}_3\}, \quad \mathcal{E}_5 := \{\hat{e}_1, \hat{e}_2, \hat{e}_3, \hat{e}_4\}.$$

Moreover, enumerating the elements of $\mathbb{E}^{(B)}$ as

$$\mathcal{F}_1 := \{\hat{e}_2\}, \ \mathcal{F}_2 := \{\hat{e}_3\}, \quad \mathcal{F}_3 := \{\hat{e}_2, \hat{e}_3\}, \quad \mathcal{F}_4 := \{\hat{e}_2, \hat{e}_3, \hat{e}_4\},$$

we conclude that there are $5 + 4 + 1 = 10$ reference problems to investigate.

Results were gathered using the sparse matrix library `scipy.sparse` with `float64` matrices, using `linalg.spsolve` and `linalg.eigsh`, with default settings. Sparsity of the matrices ensures highly accurate results.

See Table 8.6 for the computed results. First we study the r-stabilization by means of the three 'hardest' problems (ordered by saturation coefficient for $p = 4$, $q = 8$, $r = 16$). There is little difference between $r = 2q$, $r = 4q$, and $r = 8q$, indicating that $r = 2q$ is sufficient.

Choosing $q = p + 4$ is insufficient for p-robust saturation: for every problem, the discrete saturation coefficients increase as a function of p. For the two strategies $q = p + \lceil \frac{p}{7} \rceil$ and $q = 2p$, we see that these coefficients *decrease* as a function of p, strongly suggesting p-robust saturation for $p \to \infty$. For $q = 2p$, these values even tend to 1, indicating full saturation.

		$q(p) = p + 4$				$q(p) = p + \lceil p/7 \rceil$			$q(p) = 2p$			
		$p = 4$	12	28	60	$p = 14$	28	56	$p = 4$	8	16	32
	r	$q = 8$	16	32	64	$q = 16$	32	64	$q = 8$	16	32	64
$S^{(A)}_{\mathcal{E}_1,p,q,r}$	$2q$	1.0017	1.0344	1.1562	1.4015	1.1905	1.1562	1.1431	1.0017	1.0005	1.0003	1.0002
$S^{(A)}_{\mathcal{E}_2,p,q,r}$	$2q$	1.0120	1.1076	1.3505	1.7715	1.3970	1.3505	1.3334	1.0120	1.0060	1.0042	1.0035
$S^{(A)}_{\mathcal{E}_3,p,q,r}$	$2q$	1.0017	1.0350	1.1580	1.4039	1.1945	1.1580	1.1440	1.0017	1.0006	1.0003	1.0002
$S^{(A)}_{\mathcal{E}_4,p,q,r}$	$2q$	1.0138	1.1112	1.3541	1.7744	1.4038	1.3541	1.3352	1.0138	1.0065	1.0044	1.0036
$S^{(A)}_{\mathcal{E}_5,p,q,r}$	$2q$	1.0150	1.1143	1.3575	1.7772	1.4101	1.3575	1.3370	1.0150	1.0070	1.0046	1.0037
$S^{(B)}_{\mathcal{F}_1,p,q,r}$	$2q$	1.0295	1.1012	1.2092	1.3429	1.2380	1.2092	1.1952	1.0295	1.0204	1.0165	1.0147
	$4q$	1.0317	1.1075	1.2196	1.3570	1.2502	1.2196	1.2048	1.0317	1.0218	1.0176	1.0156
	$8q$	1.0318	1.1079	1.2203	—	1.2511	1.2203	—	1.0318	1.0219	1.0176	—
$S^{(B)}_{\mathcal{F}_2,p,q,r}$	$2q$	1.0013	1.0106	1.0314	1.0634	1.0385	1.0314	1.0277	1.0013	1.0006	1.0004	1.0003
$S^{(B)}_{\mathcal{F}_3,p,q,r}$	$2q$	1.0295	1.1012	1.2092	1.3429	1.2380	1.2092	1.1952	1.0295	1.0204	1.0165	1.0147
	$4q$	1.0317	1.1075	1.2196	1.3570	1.2502	1.2196	1.2048	1.0317	1.0218	1.0176	1.0156
	$8q$	1.0318	1.1079	1.2203	—	1.2511	1.2203	—	1.0318	1.0219	1.0176	—
$S^{(B)}_{\mathcal{F}_4,p,q,r}$	$2q$	1.0346	1.1055	1.2118	1.3443	1.2430	1.2118	1.1965	1.0346	1.0226	1.0175	1.0152
	$4q$	1.0374	1.1123	1.2227	1.3587	1.2563	1.2227	1.2064	1.0374	1.0242	1.0186	1.0161
	$8q$	1.0376	1.1128	1.2234	—	1.2572	1.2234	—	1.0376	1.0243	1.0187	—
$S^{(C)}_{p,q,r}$	$2q$	1.0013	1.0106	1.0313	1.0634	1.0384	1.0313	1.0277	1.0013	1.0006	1.0004	1.0003

Table 8.1 Discrete saturation coefficients for different p, q, and r. We discern three 'bands' of columns, one for each function q, and within each band, different values of p, one per column. We moreover see a number of different 'bands' of rows, one for each reference problem; within each band, a number of different discrete saturation coefficients are shown, one for each (p, q, r)-tuple.

Bibliography

[AC75] S. M. Allen and J. W. Cahn. Coherent and incoherent equilibria in iron-rich iron-aluminum alloys. *Acta Metall.*, 23(9):1017–1026, 1975.

[AK01] O. Axelsson and I. Kaporin. Error norm estimation and stopping criteria in preconditioned conjugate gradient iterations. *Numer. Linear Algebr. with Appl.*, 8(4):265–286, 2001.

[And12] R. Andreev. *Stability of space-time Petrov–Galerkin discretizations for parabolic evolution equations*. PhD thesis, ETH Zürich, 2012.

[And13] R. Andreev. Stability of sparse space-time finite element discretizations of linear parabolic evolution equations. *IMA J. Numer. Anal.*, 33(1):242–260, 2013.

[And16] R. Andreev. Wavelet-in-time multigrid-in-space preconditioning of parabolic evolution equations. *SIAM J. Sci. Comput.*, 38(1):A216–A242, 2016.

[Ash15] A. Ashton. Elliptic PDEs with constant coefficients on convex polyhedra via the unified method. *J. Math. Anal. Appl.*, 425(1):160–177, 2015.

[AT15] R. Andreev and C. Tobler. Multilevel preconditioning and low-rank tensor iteration for space-time simultaneous discretizations of parabolic PDEs. *Numer. Linear Algebr. with Appl.*, 22(2):317–337, 2015.

[BBFD15] E. Bécache, L. Bourgeois, L. Franceschini, and J. Dardé. Application of mixed formulations of quasi-reversibility to solve ill-posed problems for heat and wave equations: The 1D case. *Inverse Probl. Imaging*, 9(4):971–1002, 2015.

[BBFV20] M. Boulakia, E. Burman, M. A. Fernández, and C. Voisembert. Data assimilation finite element method for the linearized Navier–Stokes equations in the low Reynolds regime. *Inverse Probl.*, 36(8):085003, 2020.

[BCD+11] P. Binev, A. Cohen, W. Dahmen, R. DeVore, G. Petrova, and P. Wojtaszczyk. Convergence rates for greedy algorithms in reduced basis methods. *SIAM J. Math. Anal.*, 43(3):1457–1472, 2011.

[BE76] H. Brézis and I. Ekeland. Un principe variationnel associé à certaines équations paraboliques. Le cas dépendant du temps. *C. R. Acad. Sci. Paris Sér. A-B*, 282(20):Ai, A1197–A1198, 1976.

[BEEN19] T. Boiveau, V. Ehrlacher, A. Ern, and A. Nouy. Low-rank approximation of linear parabolic equations by space-time tensor Galerkin methods. *ESAIM Math. Model. Numer. Anal.*, 53(2):635–658, 2019.

[Ben99] S. J. Benbow. Solving generalized least-squares problems with LSQR. *SIAM J. Matrix Anal. Appl.*, 21(1):166–177, 1999.

[BFB+18] C. P. Blanken, E. S. Farag, S. M. Boekholdt, T. Leiner, J. Kluin, A. J. Nederveen, P. Ooij, and R. N. Planken. Advanced cardiac MRI techniques for evaluation of left-sided valvular heart disease. *J. Magn. Reson. Imaging*, 48(2):318–329, 2018.

[BFMO21] E. Burman, A. Feizmohammadi, A. Münch, and L. Oksanen. Space time stabilized finite element methods for a unique continuation problem subject to the wave equation. *ESAIM Math. Model. Numer. Anal.*, 55:S969–S991, 2021.

[BG09] P. B. Bochev and M. D. Gunzburger. *Least-squares finite element methods*, volume 166 of *Applied Mathematical Sciences*. Springer, New York, 2009.

[BGIP21] R. Becker, G. Gantner, M. Innerberger, and D. Praetorius. Goal-oriented adaptive finite element methods with optimal computational complexity. 2021, arXiv:2101.11407.

[BIHO18] E. Burman, J. Ish-Horowicz, and L. Oksanen. Fully discrete finite element data assimilation method for the heat equation. *ESAIM Math. Model. Numer. Anal.*, 52(5):2065–2082, 2018.

[Bin18] P. Binev. Tree approximation for hp-adaptivity. *SIAM J. Numer. Anal.*, 56(6):3346–3357, 2018.

[BJ89] I. Babuška and T. Janik. The h-p version of the finite element method for parabolic equations. Part I. The p-version in time. *Numer. Methods Partial Differ. Equ.*, 5(4):363–399, 1989.

[BJ90] I. Babuška and T. Janik. The h-p version of the finite element method for parabolic equations. II. The h-p version in time. *Numer. Methods Partial Differ. Equ.*, 6(4):343–369, 1990.

[Bla19] T. Blanken. *Model-based estimation and control of the particle distribution and discharge supervision in nuclear fusion reactors*. PhD thesis, Eindhoven University of Technology, 2019.

[BLO18] E. Burman, M. G. Larson, and L. Oksanen. Primal-dual mixed finite element methods for the elliptic cauchy problem. *SIAM J. Numer. Anal.*, 56(6):3480–3509, 2018.

[BO18] E. Burman and L. Oksanen. Data assimilation for the heat equation using stabilized finite element methods. *Numer. Math.*, 139(3):505–528, 2018.

[BP88] J. H. Bramble and J. E. Pasciak. A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems. *Math. Comput.*, 50(181):1, 1988.

[BPS09] D. Braess, V. Pillwein, and J. Schöberl. Equilibrated residual error estimates are p-robust. *Comp. Meth. Appl. Mech. Eng.*, 198(13-14):1189–1197, 2009.

[BR18] L. Bourgeois and A. Recoquillay. A mixed formulation of the Tikhonov regularization and its application to inverse PDE problems. *ESAIM Math. Model. Numer. Anal.*, 52(1):123–145, 2018.

[Bra81] A. Brandt. Multigrid solvers on parallel computers. In *Elliptic Probl. Solvers*, pages 39–83. Elsevier, 1981.

[Bra07] D. Braess. *Finite Elements*. Cambridge University Press, Cambridge, 2007.

[BRU20] N. Beranek, M. A. Reinhold, and K. Urban. A space-time variational method for optimal control problems. 2020, arXiv:2010.00345.

[BS08] S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer New York, New York, NY, 2008.

[BS14] D. Broersen and R. Stevenson. A robust Petrov–Galerkin discretisation of convection-diffusion equations. *Comput. Math. with Appl.*, 68(11):1605–1618, 2014.

[BVVW20] R. Brokkelkamp, R. van Venetië, M. de Vries, and J. Westerdiep. PACE Solver Description: tdULL. In Y. Cao and M. Pilipczuk, editors, *15th Int. Symp. Parameterized Exact Comput. (IPEC 2020)*, pages 29:1–29:4, Dagstuhl, Germany, 2020. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

[BY14] R. E. Bank and H. Yserentant. On the H^1-stability of the L_2-projection onto finite element spaces. *Numer. Math.*, 126(2):361–381, 2014.

[BZ96] R. Balder and C. Zenger. The solution of multidimensional real Helmholtz equations on sparse grids. *SIAM J. Sci. Comput.*, 17(3):631–646, 1996.

[Car01] C. Carstensen. Merging the Bramble–Pasciak–Steinbach and the Crouzeix–Thomée criterion for H^1-stability of the L^2-projection onto finite element spaces. *Math. Comput.*, 71(237):157–164, 2001.

[CDD08] M. Costabel, M. Dauge, and L. Demkowicz. Polynomial extension operators for H^1, $\boldsymbol{H}(\mathbf{curl})$ and $\boldsymbol{H}(\mathbf{div})$-spaces on a cube. *Math. Comput.*, 77(264):1967–1999, 2008.

[CDD+19] A. Cohen, W. Dahmen, R. DeVore, J. Fadili, O. Mula, and J. Nichols. Optimal reduced model algorithms for data-based state estimation, 2019.

[CDG14] C. Carstensen, L. Demkowicz, and J. Gopalakrishnan. A posteriori error control for DPG methods. *SIAM J. Numer. Anal.*, 52(3):1335–1353, 2014.

[CDW12] A. Cohen, W. Dahmen, and G. Welper. Adaptivity and variational stabilization for convection-diffusion equations. *ESAIM Math. Model. Numer. Anal.*, 46:1247–1273, 2012.

[CEQ14] J. Chan, J. A. Evans, and W. Qiu. A dual Petrov–Galerkin finite element method for the convection-diffusion equation. *Comput. Math. with Appl.*, 68(11):1513–1529, 2014.

[CF99] C. Carstensen and S. A. Funken. Fully reliable localized error control in the FEM. *SIAM J. Sci. Comput.*, 21(4):1465–1484, 1999.

[CFLQ14] H. Chen, G. Fu, J. Li, and W. Qiu. First order least squares method with weakly imposed boundary condition for convection dominated diffusion problems. *Comput. Math. with Appl.*, 68(12):1635–1652, 2014.

[Che14] L. Chen. A simple construction of a Fortin operator for the two dimensional Taylor–Hood element. *Comput. Math. Appl.*, 68(10):1368–1373, 2014.

[CM10] M. Costabel and A. McIntosh. On Bogovskiĭ and regularized Poincaré integral operators for de Rham complexes on Lipschitz domains. *Math. Zeitschrift*, 265(2):297–320, 2010.

[CNSV17a] C. Canuto, R. H. Nochetto, R. Stevenson, and M. Verani. Convergence and optimality of hp-AFEM. *Numer. Math.*, 135(4):1073–1119, 2017.

[CNSV17b] C. Canuto, R. H. Nochetto, R. Stevenson, and M. Verani. On p-robust saturation for hp-AFEM. *Comput. Math. with Appl.*, 73(9):2004–2022, 2017.

[CNSV19] C. Canuto, R. H. Nochetto, R. Stevenson, and M. Verani. A saturation property for the spectral-Galerkin approximation of a Dirichlet problem in a square. *ESAIM Math. Model. Numer. Anal.*, 53(4):987–1003, 2019.

[Dal94] R. Daley. *Atmospheric data analysis*. Cambridge Atmospheric and Space Science Series. Cambridge University Press, Cambridge, UK, 1994.

[DC07] P. Díez and G. Calderón. Goal-oriented error estimation for transient parabolic problems. *Comput. Mech.*, 39(5):631–646, 2007.

[DEV16] V. Dolejší, A. Ern, and M. Vohralík. hp-adaptation driven by polynomial-degree-robust a posteriori error estimates for elliptic problems. *SIAM J. Sci. Comput.*, 38(5):A3220–A3246, 2016.

[Dev20] D. Devaud. Petrov–Galerkin space-time hp-approximation of parabolic equations in H1/2. *IMA J. Numer. Anal.*, 40(4):2717–2745, 2020.

[DG11] L. Demkowicz and J. Gopalakrishnan. A class of discontinuous Petrov–Galerkin methods. II. Optimal test functions. *Numer. Methods Partial Differ. Equations*, 27(1):70–105, 2011.

[DG14] L. F. Demkowicz and J. Gopalakrishnan. An overview of the discontinuous Petrov Galerkin method. In *Recent Dev. discontinuous Galerkin finite Elem. methods Partial Differ. equations*, volume 157 of *IMA Vol. Math. Appl.*, pages 149–180. Springer, Cham, 2014.

[DHH13] J. Dardé, A. Hannukainen, and N. Hyvönen. An H_{div}-based mixed quasi-reversibility method for solving elliptic cauchy problems. *SIAM J. Numer. Anal.*, 51(4):2123–2148, 2013.

[Dij09] T. Dijkema. *Adaptive tensor product wavelet methods for the solution of PDEs*. PhD thesis, Utrecht University, 2009.

[DKS16] L. Diening, C. Kreuzer, and R. Stevenson. Instance optimality of the adaptive maximum strategy. *Found. Comput. Math.*, 16(1):33–68, 2016.

[DL92] R. Dautray and J.-L. Lions. *Mathematical analysis and numerical methods for science and technology: Volume 5*. Springer-Verlag, Berlin, 1992.

[DL00] R. Dautray and J.-L. Lions. *Mathematical analysis and numerical methods for science and technology: Volume 6*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2000.

[DPS05] L. Dalcín, R. Paz, and M. Storti. MPI for Python. *J. Parallel Distrib. Comput.*, 65(9):1108–1115, 2005.

[DS18] D. Devaud and C. Schwab. Space-time hp-approximation of parabolic equations. *Calcolo*, 55(3):35, 2018.

[DS20] L. Diening and J. Storn. A space-time DPG method for the heat equation. 2020, arXiv:2012.13229.

[DST20] L. Diening, J. Storn, and T. Tscherpel. On the Sobolev and L^p-Stability of the L^2-projection. 2020, arXiv:2008.01801.

[DSW21] W. Dahmen, R. Stevenson, and J. Westerdiep. Accuracy controlled data assimilation for parabolic problems. *Math. Comput.*, 2021, arXiv:2105.05836.

[Dup82] T. Dupont. Mesh modification for evolution equations. *Math. Comput.*, 39(159):85, 1982.

[DV13] M. D'Elia and A. Veneziani. Uncertainty quantification for data assimilation in a steady incompressible Navier–Stokes problem. *ESAIM Math. Model. Numer. Anal.*, 47(4):1037–1057, 2013.

[ESV17] A. Ern, I. Smears, and M. Vohralík. Guaranteed, locally space-time efficient, and polynomial-degree robust a posteriori error estimates for high-order discretizations of parabolic problems. *SIAM J. Numer. Anal.*, 55(6):2811–2834, 2017.

[EV15] A. Ern and M. Vohralík. Polynomial-degree-robust a posteriori estimates in a unified setting for conforming, nonconforming, discontinuous Galerkin, and mixed discretizations. *SIAM J. Numer. Anal.*, 53(2):1058–1081, 2015.

[Eva10] L. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, Rhode Island, 2010.

[FFK+14] R. D. Falgout, S. Friedhoff, T. V. Kolev, S. P. MacLachlan, and J. B. Schroder. Parallel time integration with multigrid. *SIAM J. Sci. Comput.*, 36(6):C635–C661, 2014.

[FK21] T. Führer and M. Karkulik. Space-time least-squares finite elements for parabolic equations. *Comput. Math. with Appl.*, 92:27–36, 2021.

[For77] M. Fortin. An analysis of the convergence of mixed finite element methods. *RAIRO. Anal. numérique*, 11(4):341–354, 1977.

[Gan15] M. J. Gander. 50 Years of time parallel time integration. In *Mult. Shoot. Time Domain Decompos. Methods*, chapter 3, pages 69–113. Springer, Cham, 2015.

[GB86a] B. Guo and I. Babuška. The h-p version of the finite element method – Part 1: The basic approximation results. *Comput. Mech.*, 1(1):21–41, 1986.

[GB86b] B. Guo and I. Babuška. The h-p version of the finite element method – Part 2: General results and applications. *Comput. Mech.*, 1(3):203–220, 1986.

[GHPS18] G. Gantner, A. Haberl, D. Praetorius, and B. Stiftner. Rate optimal adaptive FEM with inexact solver for nonlinear operators. *IMA J. Numer. Anal.*, 38(4):1797–1831, 2018.

[GHS16] F. D. Gaspoz, C.-J. Heine, and K. G. Siebert. Optimal grading of the newest vertex bisection and H^1-stability of the L_2-projection. *IMA J. Numer. Anal.*, 36(3):1217–1241, 2016.

[GK11] M. D. Gunzburger and A. Kunoth. Space-time adaptive wavelet methods for optimal control problems constrained by parabolic evolution equations. *SIAM J. Control Optim.*, 49(3):1150–1170, 2011.

[GK12] M. J. Gander and F. Kwok. Chladni figures and the Tacoma bridge: motivating PDE eigenvalue problems via vibrating plates. *SIAM Rev.*, 54(3):573–596, 2012.

[GL01] V. Girault and J.-L. Lions. Two-grid finite-element schemes for the transient Navier-Stokes problem. *ESAIM Math. Model. Numer. Anal.*, 35(5):945–980, 2001.

[GM97] G. H. Golub and G. Meurant. Matrices, moments and quadrature II: How to compute the norm of the error in iterative methods. *BIT Numer. Math.*, 37(3):687–705, 1997.

[GN16] M. J. Gander and M. Neumüller. Analysis of a new space-time parallel multigrid algorithm for parabolic problems. *SIAM J. Sci. Comput.*, 38(4):A2173–A2208, 2016.

[GS84] P. Gray and S. Scott. Autocatalytic reactions in the isothermal, continuous stirred tank reactor. *Chem. Eng. Sci.*, 39(6):1087–1097, 1984.

[GS19] H. Gimperlein and J. Stocek. Space-time adaptive finite elements for nonlocal parabolic variational inequalities. *Comput. Methods Appl. Mech. Engrg.*, 352:137–171, 2019.

[GS21] G. Gantner and R. Stevenson. Further results on a space-time FOSLS formulation of parabolic PDEs. *ESAIM Math. Model. Numer. Anal.*, 55(1):283–299, 2021.

[GW12] M. J. Gander and G. Wanner. From Euler, Ritz, and Galerkin to modern computing. *SIAM Rev.*, 54(4):627–666, 2012.

[Hac85] W. Hackbusch. *Multi-grid methods and applications*, volume 4 of *Springer Series in Computational Mathematics*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1985.

[Hac12] W. Hackbusch. *Tensor spaces and numerical tensor calculus*, volume 42 of *Springer Series in Computational Mathematics*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[Hip06] R. Hiptmair. Operator preconditioning. *Comput. Math. with Appl.*, 52(5):699–706, 2006.

[HPSV21] A. Haberl, D. Praetorius, S. Schimanko, and M. Vohralík. Convergence and quasi-optimal cost of adaptive algorithms for nonlinear operators including iterative linearization and algebraic solver. *Numer. Math.*, 147(3):679–725, 2021.

[HS52] M. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *J. Res. Natl. Bur. Stand. (1934).*, 49(6):409, 1952.

[HVW95] G. Horton, S. Vandewalle, and P. Worley. An algorithm with polylog parallel complexity for solving parabolic partial differential equations. *SIAM J. Sci. Comput.*, 16(3):531–541, 1995.

[Kat60] T. Kato. Estimation of iterated Matrices, with application to the von Neumann condition. *Numer. Math.*, 2(1):22–29, 1960.

[KKL+21] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang. Physics-informed machine learning. *Nat. Rev. Phys.*, 2021.

[KS14] S. Kestler and R. Stevenson. Fast evaluation of system matrices w.r.t. multi-tree collections of tensor product refinable basis functions. *J. Comput. Appl. Math.*, 260:103–116, 2014.

[KSU15] S. Kestler, K. Steih, and K. Urban. An efficient space-time adaptive wavelet Galerkin method for time-periodic parabolic partial differential equations. *Math. Comput.*, 85(299):1309–1333, 2015.

[Li15] R.-C. Li. Rayleigh Quotient Based Optimization Methods for Eigenvalue Problems. In *Matrix Funct. Matrix Equations*, number 2, pages 76–108. nov 2015.

[Lio71] J.-L. Lions. *Optimal control of systems governed by partial differential equations*. Springer-Verlag Berlin Heidelberg, 1 edition, 1971.

[LLD06] J. M. Lewis, S. Lakshmivarahan, and S. Dhall. *Dynamic data assimilation*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, Cambridge, 2006.

[LMN16] U. Langer, S. E. Moore, and M. Neumüller. Space-time isogeometric analysis of parabolic evolution problems. *Comp. Meth. Appl. Mech. Eng.*, 306:342–363, 2016.

[LMT01] J.-L. Lions, Y. Maday, and G. Turinici. Résolution d'EDP par un schéma en temps «pararéel». *Comptes Rendus l'Académie des Sci. - Ser. I - Math.*, 332(7):661–668, 2001.

[LSTY21] U. Langer, O. Steinbach, F. Tröltzsch, and H. Yang. Space-time finite element discretization of parabolic optimal control problems with energy regularization. *SIAM J. Numer. Anal.*, 59(2):675–695, 2021.

[Maj16] A. J. Majda. *Introduction to turbulent dynamical systems in complex systems*, volume 5 of *Frontiers in Applied Dynamical Systems: Reviews and Tutorials*. Springer International Publishing, Cham, 2016.

[MFL+91] O. A. McBryan, P. O. Frederickson, J. Lindenand, A. Schüller, K. Solchenbach, K. Stüben, C.-A. Thole, and U. Trottenberg. Multigrid methods on parallel computers–A survey of recent developments. *IMPACT Comput. Sci. Eng.*, 3(1):1–75, 1991.

[MMCP+19] J. Muñoz-Matute, V. M. Calo, D. Pardo, E. Alberdi, and K. G. van der Zee. Explicit-in-time goal-oriented adaptivity. *Comp. Meth. Appl. Mech. Eng.*, 347:176–200, 2019.

[MPPY15] Y. Maday, A. T. Patera, J. D. Penn, and M. Yano. A parameterized-background data-weak approach to variational data assimilation: formulation, analysis, and application to acoustics. *Int. J. Numer. Methods Eng.*, 102(5):933–965, 2015.

[MPT21] G. Meurant, J. Papež, and P. Tichý. Accurate error estimation in CG. 2021, arXiv:2101.03931.

[MRS90] P. A. Markowich, C. A. Ringhofer, and C. Schmeiser. *Semiconductor equations*. Springer Vienna, Vienna, 1990.

[MS09] M. S. Mommer and R. Stevenson. A goal-oriented adaptive finite element method with convergence rates. *SIAM J. Numer. Anal.*, 47(2):861–886, 2009.

[MS17] A. Münch and D. A. Souza. Inverse problems for linear parabolic equations using mixed formulations – Part 1: Theoretical analysis. *J. Inverse Ill-posed Probl.*, 25(4), 2017.

[MT13] G. Meurant and P. Tichý. On computing quadrature-based bounds for the A-norm of the error in conjugate gradients. *Numer. Algorithms*, 62(2):163–191, 2013.

[MW01] J. M. Melenk and B. I. Wohlmuth. On residual-based a posteriori error estimation in hp-FEM. *Adv. Comput. Math.*, 15(1-4):311–331, 2001.

[Nay76] B. Nayroles. Deux théorèmes de minimum pour certains systèmes dissipatifs. *C. R. Acad. Sci. Paris Sér. A-B*, 282(17):Aiv, A1035—-A1038, 1976.

[Nie64] J. Nievergelt. Parallel methods for integrating ordinary differential equations. *Commun. ACM*, 7(12):731–733, 1964.

[NS19] M. Neumüller and I. Smears. Time-parallel iterative solvers for parabolic evolution equations. *SIAM J. Sci. Comput.*, 41(1):C28–C51, 2019.

[OR00] M. A. Olshanskii and A. Reusken. On the convergence of a multigrid method for linear reaction-diffusion problems. *Computing*, 65(3):193–202, 2000.

[Pab15] R. Pabel. *Adaptive wavelet methods for variational formulations of nonlinear elliptic PDEs on tensor-product domains*. PhD thesis, Universität zu Köln, 2015.

[Pea93] J. E. Pearson. Complex patterns in a simple system. *Science (80-.).*, 261(5118):189–192, 1993.

[Pfl10] D. Pflüger. *Spatially adaptive sparse grids for high-dimensional problems*. PhD thesis, Institut für Informatik, Technische Universität München, 2010.

[PP20] C.-M. Pfeiler and D. Praetorius. Dörfler marking with minimal cardinality is a linear complexity problem. *Math. Comput.*, 89(326):2735–2752, 2020.

[PS75] C. C. Paige and M. A. Saunders. Solution of sparse indefinite systems of linear equations. *SIAM J. Numer. Anal.*, 12(4):617–629, 1975.

[PS82] C. C. Paige and M. A. Saunders. LSQR: an algorithm for sparse linear equations and sparse least squares. *ACM Trans. Math. Softw.*, 8(1):43–71, 1982.

[Rek18] N. Rekatsinas. *Optimal adaptive wavelet methods for solving first order system least squares*. PhD thesis, University of Amsterdam, 2018.

[RS18a] N. Rekatsinas and R. Stevenson. A quadratic finite element wavelet Riesz basis. *Int. J. Wavelets, Multiresolution Inf. Process.*, 16(04):1850033, 2018.

[RS18b] N. Rekatsinas and R. Stevenson. An optimal adaptive wavelet method for first order system least squares. *Numer. Math.*, 140(1):191–237, 2018.

[RS19] N. Rekatsinas and R. Stevenson. An optimal adaptive tensor product wavelet solver of a space-time FOSLS formulation of parabolic evolution problems. *Adv. Comput. Math.*, 45(2):1031–1066, 2019.

[Sch97] J. Schöberl. NETGEN An advancing front 2D/3D-mesh generator based on abstract rules. *Comput. Vis. Sci.*, 1(1):41–52, 1997.

[Sch14] J. Schöberl. C++11 implementation of finite elements in NGSolve. Technical report, Institute for Analysis and Scientific Computing, Vienna University of Technology, 2014.

[SD18] A. Schwarz and R. P. Dwight. Data assimilation for Navier–Stokes using the least-squares finite-element method. *Int. J. Uncertain. Quantif.*, 8(5):383–403, 2018.

[SS09] C. Schwab and R. Stevenson. Space-time adaptive wavelet methods for parabolic evolution problems. *Math. Comput.*, 78(267):1293–1318, 2009.

[SS17] C. Schwab and R. Stevenson. Fractional space-time variational formulations of (Navier–) Stokes equations. *SIAM J. Math. Anal.*, 49(4):2442–2467, 2017.

[Ste98] R. Stevenson. Stable three-point wavelet bases on general meshes. *Numer. Math.*, 80(1):131–158, 1998.

[Ste03] R. Stevenson. Locally supported, piecewise polynomial biorthogonal wavelets on nonuniform meshes. *Constr. Approx.*, 19(4):477–508, 2003.

[Ste07] R. Stevenson. Optimality of a standard adaptive finite element method. *Found. Comput. Math.*, 7(2):245–269, 2007.

[Ste08] R. Stevenson. The completion of locally refined simplicial partitions created by bisection. *Math. Comput.*, 77(261):227–241, 2008.

[Ste14] R. Stevenson. Adaptive wavelet methods for linear and nonlinear least-squares problems. *Found. Comput. Math.*, 14(2):237–283, 2014.

[Ste15] O. Steinbach. Space-time finite element methods for parabolic problems. *Comput. Methods Appl. Math.*, 15(4):551–566, 2015.

[SvVW21] R. Stevenson, R. van Venetië, and J. Westerdiep. A wavelet-in-time, finite element-in-space adaptive method for parabolic evolution equations. 2021, arXiv:2101.03956.

[SW21a] R. Stevenson and J. Westerdiep. Stability of Galerkin discretizations of a mixed space-time variational formulation of parabolic evolution equations. *IMA J. Numer. Anal.*, 41(1):28–47, 2021.

[SW21b] R. Stevenson and J. Westerdiep. Minimal residual space-time discretizations of parabolic equations: Asymmetric spatial operators. 2021, arXiv:2106.01090.

[SY18] O. Steinbach and H. Yang. Comparison of algebraic multigrid methods for an adaptive space-time finite-element discretization of the heat equation in 3D and 4D. *Numer. Linear Algebr. with Appl.*, 25(3):e2143, 2018.

[SY19] O. Steinbach and H. Yang. 7. Space-time finite element methods for parabolic evolution equations: discretization, a posteriori error estimation, adaptivity and solution. In *Space-Time Methods*, pages 207–248. De Gruyter, Berlin, 2019.

[SZ90] L. R. Scott and S. Zhang. Finite element interpolation of nonsmooth functions satisfying boundary conditions. *Math. Comput.*, 54(190):483, 1990.

[SZ20] O. Steinbach and M. Zank. Coercive space-time finite element methods for initial boundary value problems. *ETNA - Electron. Trans. Numer. Anal.*, 52:154–194, 2020.

[Tho06] V. Thomée. *Galerkin finite element methods for parabolic problems*, volume 25 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, second edition, 2006.

[Tur52] A. Turing. The chemical basis of morphogenesis. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, 237(641):37–72, 1952.

[TV16] F. Tantardini and A. Veeser. The L^2-Projection and quasi-optimality of Galerkin methods for parabolic equations. *SIAM J. Numer. Anal.*, 54(1):317–340, 2016.

[TZA+20] J. Töger, M. J. Zahr, N. Aristokleous, K. Markenroth Bloch, M. Carlsson, and P. O. Persson. Blood flow imaging by optimal matching of computational fluid dynamics to 4D-flow data. *Magn. Reson. Med.*, 84(4), 2020.

[UP13] K. Urban and A. T. Patera. An improved error bound for reduced basis approximation of linear parabolic problems. *Math. Comput.*, 83(288):1599–1615, 2013.

[Vir20] P. Virtanen. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods*, 17(3):261–272, 2020.

[VR18] I. Voulis and A. Reusken. A time dependent Stokes interface problem: well-posedness and space-time finite element discretization. *ESAIM Math. Model. Numer. Anal.*, 52(6):2187–2213, 2018.

[vVW20a] R. van Venetië and J. Westerdiep. A parallel algorithm for solving linear parabolic evolution equations. In *9th Parallel-in-Time Work.*, 2020.

[vVW20b] R. van Venetië and J. Westerdiep. Implementation of: A parallel algorithm for solving linear parabolic evolution equations, 2020, doi:10.5281/zenodo.4475959.

[vVW21a] R. van Venetië and J. Westerdiep. Efficient space-time adaptivity for parabolic evolution equations using wavelets in time and finite elements in space. 2021, arXiv:2104.08143.

[vVW21b] R. van Venetië and J. Westerdiep. Implementation of: Efficient space-time adaptivity for parabolic evolution equations using wavelets in time and finite elements in space, 2021, doi:10.5281/zenodo.4697250.

[Wes20] J. Westerdiep. On p-Robust Saturation on Quadrangulations. *Comput. Methods Appl. Math.*, 20(1):169–186, 2020.

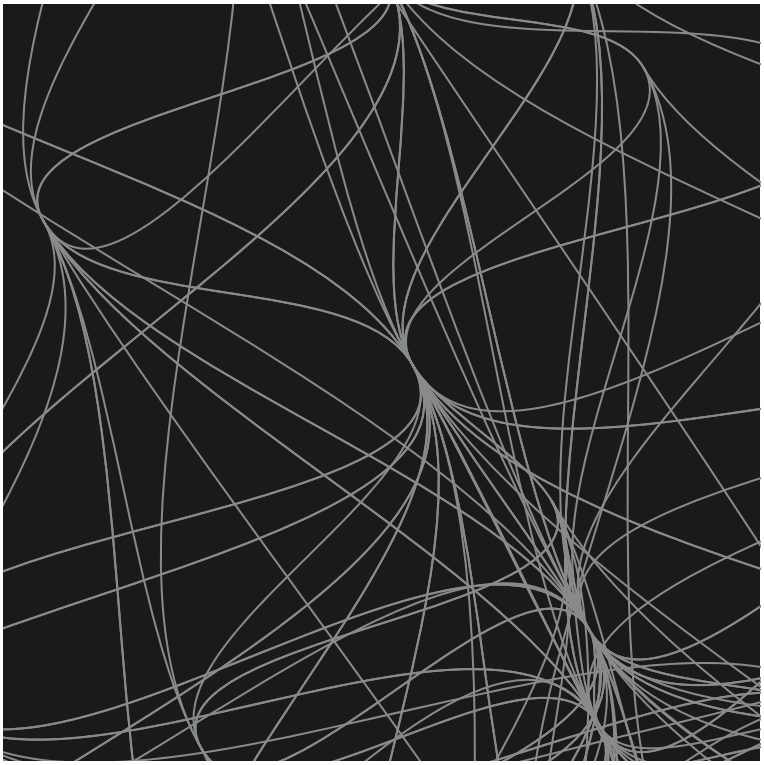
[Wlo82] J. Wloka. *Partielle Differentialgleichungen: Sobolevräume und Randwertaufgaben*. Vieweg+ Teubner Verlag, Stuttgart, 1982.

[Wor91] P. H. Worley. Limits on parallelism in the numerical solution of linear partial differential equations. *SIAM J. Sci. Stat. Comput.*, 12(1):1–35, 1991.

[WZ17] J. Wu and H. Zheng. Uniform convergence of multigrid methods for adaptive meshes. *Appl. Numer. Math.*, 113:109–123, 2017.

[XZ03] J. Xu and L. Zikatanov. Some observations on Babuška and Brezzi theories. *Numer. Math.*, 94(1):195–202, 2003.

[ZBF+20] Z. Zainib, F. Ballarin, S. Fremes, P. Triverio, L. Jiménez-Juan, and G. Rozza. Reduced order methods for parametric optimal flow control in coronary bypass grafts, towards patient-specific data assimilation. *Int. J. Numer. Method. Biomed. Eng.*, 2020.



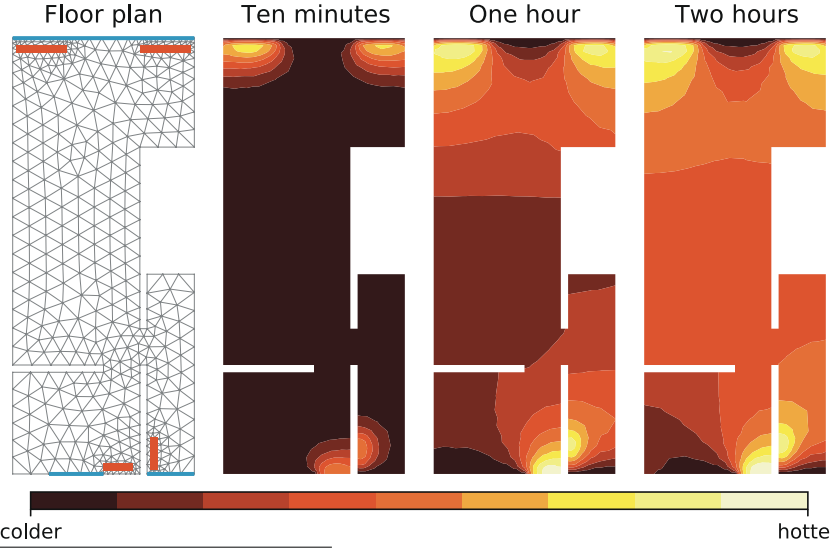
Summary

My partner and I live in an apartment in Amsterdam. See its floor plan below. It's a regular 1900s apartment: the walls are well-insulated (the windows not so much), and there are heaters in the kitchen, living room, and bedroom.

Imagine we just came back from a winter holiday and turn on the heaters. The apartment starts warming up—more so near heaters, less so near windows. How does its temperature evolve as a result?

The temperature of the apartment is governed by a *parabolic partial differential equation*. These equations model time-dependent phenomena like heat conduction, chemical concentration, and fluid flow. *Parabolic evolution equations* describe how a function evolves from a given initial state as governed by the PDE. It is generally impossible to solve such problems using pen and paper, so to understand their solution, we turn to *numerical approximation*.

Initially, the apartment has the ambient temperature. We turn our heaters to some constant setting, and model the temperature of the apartment using a *heat equation*[1]; its *space-time minimal residual* approximation is shown in the figure. The apartment is relatively well-heated, except for the bedroom.[2]



[1]This illustration is heavily inspired by [GW12, §8]. I apologize to my physicist friends, as modeling the temperature of our apartment using the heat equation likely oversimplifies matters.

[2]We know this all too well, and in fact, are working on getting double glazing.

Historically, parabolic evolution equations are solved numerically through *time-stepping*. Here, one first discretizes the problem in space to obtain a coupled system of *ordinary* differential equations in time, which is then solved using an ODE-solver. A drawback of these methods is that they are unable to efficiently resolve details of the solution localized in space and time, such as the high temperature gradients near the heaters just after turning them on.

We take a different approach and consider the equation in space and time simultaneously. *Space-time methods can* resolve these local details, and even produce approximations that are, up to some constant, the best possible. In this thesis, we focus on the *minimal residual (MR) method* introduced in [And12].

Chapter 2 We discuss the preliminaries from an abstract viewpoint. Starting from some linear operator equation posed on Hilbert spaces, we introduce equivalent conditions for *well-posedness* of the problem, discuss its *MR approximation* and derive conditions for *uniform quasi-optimality*.

Chapter 3 Next, for parabolic equations with a symmetric spatial partial differential operator (like our temperature model), we show that uniform quasi-optimality reduces to *uniform stability of trial- and test spaces*. We verify this condition for discretizations of the space-time cylinder into *time slabs*.

Chapter 4 We then take on evolution problems where the spatial partial differential operator is not necessarily symmetric. We find that also for these problems, MR solutions are quasi-optimal. As an application, we consider a *convection-diffusion-reaction equation*.

Chapter 5 We now consider adaptive refinement locally in space and time, and aim for *optimal convergence* at *optimal linear cost*. We use a matrix-free iterative solver to produce approximate MR solutions that are still quasi-optimal, and see that this translates to a highly efficient algorithm in practice.

Chapter 6 We then explore the MR method in parallel computation. We show that our method has a polylogarithmic parallel complexity, which is on par with the best-known algorithms for *elliptic* problems. The result is a highly scalable algorithm producing a solution for the heat equation with over four billion unknowns on over two thousand parallel cores in under two minutes.

Chapter 7 We turn to the ill-posed *data assimilation* problem of recovering some unknown state from (noisy) observational data and a known underlying parabolic evolution equation; somewhat like inferring the temperature in the kitchen just from looking at the thermostat in the living room.

Chapter 8 We finish with a chapter on *p-robust saturation for the Poisson equation*, discussing which increase in local polynomial degrees ensures error reduction. We derive sufficient conditions, reduce them to conditions on a *reference square*, and provide numerical evidence that these hold.

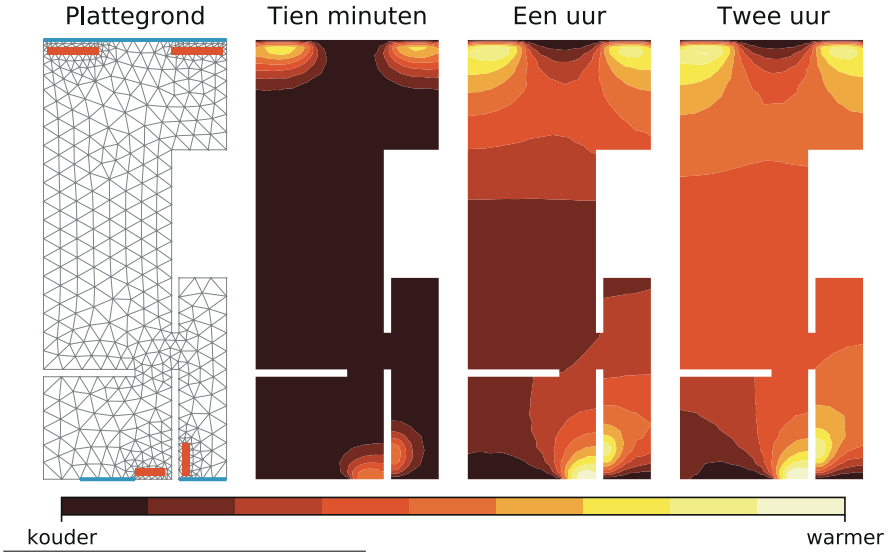
Samenvatting

Carmen en ik wonen in een appartement in Amsterdam. Zie de plattegrond hieronder. Het is typisch jaren-1900: de muren zijn goed geïsoleerd (de ramen helaas niet), en we hebben cv's in de keuken, woonkamer, en slaapkamer.

Stel je voor: we komen net terug van vakantie. We hebben het koud, en doen de verwarming aan. Het huis warmt op—wat sneller in de buurt van de cv, wat minder snel bij ramen. Maar hoe verspreidt de warmte zich *precies*?

De temperatuur van ons huis wordt bepaald door een *parabolische partiële differentiaalvergelijking*. Deze vergelijkingen modelleren tijdsafhankelijke processen zoals warmtegeleiding en vloeistofstroming. Hun *evolutievergelijkingen* beschrijven hoe een functie zich ontwikkelt vanuit e.o.a. beginstaat. Het is typisch onmogelijk om deze problemen met pen en papier op te lossen. Om te snappen hoe zo'n oplossing werkt, maken we een *numerieke benadering*.

In eerste instantie is het huis zo warm als de buitenlucht. We zetten de verwarming aan, en modelleren het temperatuurverloop van het huis met de *warmtevergelijking*[1]; de figuur laat z'n *ruimte-tijd kleinste residu*-benadering zien. Ons huis is relatief goed verwarmd, behalve de slaapkamer.[2]



[1] Deze illustratie is flink geïnspireerd door [GW12, §8]. Mijn excuses tegen m'n natuurkundevrienden: de warmtevergelijking is hoogstwaarschijnlijk een veel te simplistisch model.

[2] Dit weten we maar al te goed. Sterker nog, we zijn bezig om dubbel glas te laten zetten.

Historisch lost men parabolische evolutievergelijkingen op met *tijdstappers*. Je discretiseert het probleem eerst in ruimte, en krijgt dan een gekoppeld systeem van *gewone* differentiaalvergelijkingen in tijd. Dit systeem los je dan op met een ODE-solver. Een belangrijk nadeel is dat het niet lukt om efficiënt details in de oplossing weer te geven die lokaal in ruimte en tijd zitten, zoals 't grote temperatuurverschil in de buurt van de cv wanneer die net aanslaat.

Wij nemen een andere route, en beschouwen de vergelijking tegelijk in ruimte en tijd. *Ruimte-tijdmethodes* kunnen deze lokale details *wel* weergeven, en vinden zelfs oplossingen die *quasi-best* (op een constante na, zo goed als maar kan) zijn. In dit proefschrift bekijken we de methode van Andreev, en noemen die de *kleinste residu*-methode (MR-methode, van *minimal residual*).

Hoofdstuk 2 We beginnen abstract, bij een *lineaire operatorvergelijking* tussen Hilbertruimtes. We bepalen equivalente eisen voor *goedgesteldheid* van het probleem, bespreken wat een MR-benadering precies is, en bepalen wanneer deze *uniform quasi-best* is.

Hoofdstuk 3 Voor parabolische evolutievergelijkingen waarbij de ruimtelijke differentiaaloperator symmetrisch is (zoals in ons warmtemodel) zien we dat MR-benaderingen *uniform quasi-best* zijn voor *uniform stabiele zoek- en testruimtes*. We bewijzen dit voor opdelingen van de ruimte-tijdcylinder in *tijdsplakjes*.

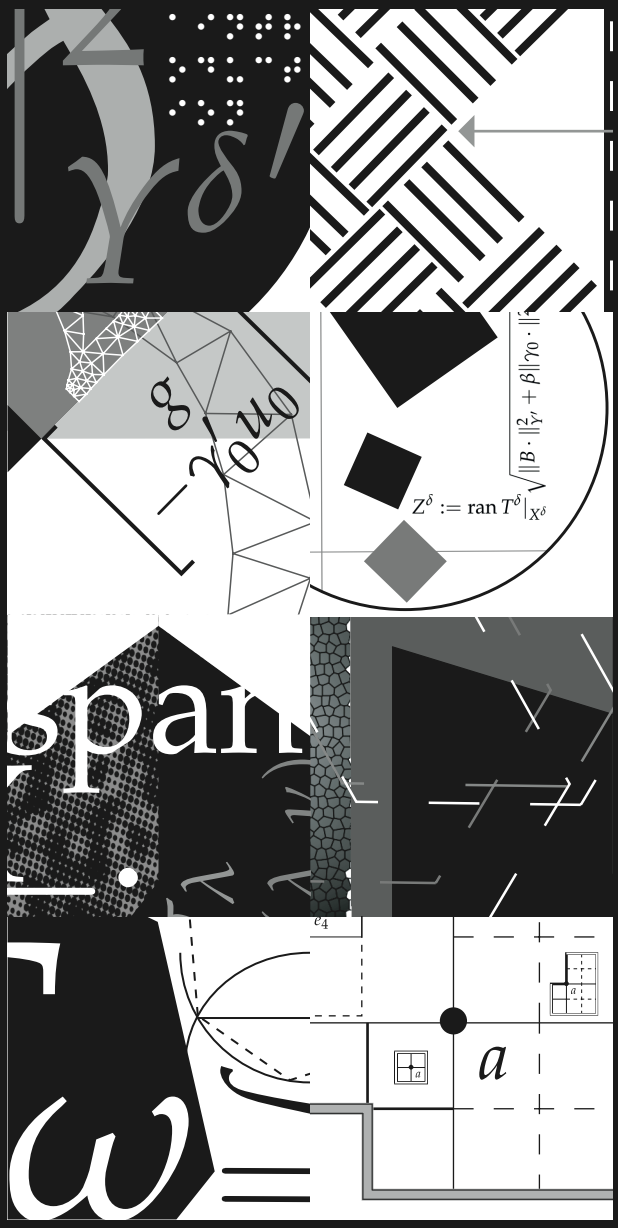
Hoofdstuk 4 Nu bekijken we evolutievergelijkingen waar de ruimtelijke operator *niet* symmetrisch is. We vinden dat MR-benaderingen voor dit soort vergelijkingen *ook* uniform quasi-best zijn. Als toepassing bekijken we een *convectie-diffusie-reactievergelijking*.

Hoofdstuk 5 Dan bekijken we *verfijning adaptief in ruimte-tijd*, en mikken op *optimale convergentie* in *optimale lineaire tijd*. We gebruiken een iteratief proces zonder matrices om benaderingen van MR-benaderingen te vinden die *ook* quasi-best zijn, en zien dat dit vertaalt naar een enorm efficiënt algoritme.

Hoofdstuk 6 We verkennen de MR-methode op een parallele computer. We zien dat onze methode een polylogaritmische parallele complexiteit heeft, even snel als de beste resultaten voor *elliptische* problemen. Het resultaat is een algoritme dat meer dan 2000 processoren tegelijk gebruikt om een probleem met meer dan 4 miljard onbekenden op te lossen in minder dan 2 minuten.

Hoofdstuk 7 We bekijken nu het slechtgestelde *data assimilatie*-probleem. We willen een onbekende functie herleiden vanuit (ruizige) metingen en een bekende parabolische evolutievergelijking. Een beetje zoals bepalen hoe warm het is in de keuken door enkel te kijken naar een thermostaat in de woonkamer.

Hoofdstuk 8 We eindigen met een hoofdstuk over *p-robuuste verzadiging voor de Poissonvergelijking*, dus hoe we lokale polynomiale graden moeten verhogen om verzekerd te zijn van foutverlaging. We zien dat deze vraag reduceert tot eisen op een referentievierkant, en geven numeriek bewijs dat deze kloppen.


$\gamma_{\delta'}$
$\sqrt{\|B\cdot\|_{\gamma'}^2+\beta\|\gamma_0\cdot\|^2}$
$Z^\delta := \operatorname{ran} T^\delta|_{X^\delta}$
span
ω
e_4
a