Multi-level stochastic collocation methods for parabolic and Schrödinger equations

Zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

von der KIT-Fakultät für Mathematik des
Karlsruher Instituts für Technologie (KIT)
genehmigte

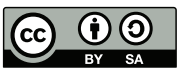
Dissertation

von

Benny Stein

Tag der mündlichen Prüfung: 16.03.2022

1. Referent: Prof. Dr. Tobias Jahnke
2. Referent: Prof. Dr. Christian Wieners



This document is licensed under a Creative Commons
Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0):
https://creativecommons.org/licenses/by-sa/4.0/deed.en

Abstract

In this thesis, we propose, analyse and implement numerical methods for time-dependent non-linear parabolic and Schrödinger-type equations with uncertain parameters. The discretisation of the parameter space which incorporates the uncertainty of the problem is performed via single- and multi-level collocation strategies. To deal with the possibly large dimension of the parameter space, sparse grid collocation techniques are used to alleviate the curse of dimensionality to a certain extent. We prove that the multi-level method is capable of reducing the overall computational costs significantly.

In the parabolic case, the time discretisation is performed via an implicit-explicit splitting strategy of order two which consists shortly speaking of a combination of an implicit trapezoidal rule for the stiff linear part and Heun's method for the non-linear part. In the Schrödinger case, time is discretised via the famous second-order Strang splitting method.

For both problem classes we review known error bounds for both discretizations and prove new error bounds for the time discretisations which take the regularity in the parameter space into account. In the parabolic case, a new error bound for the "implicit-explicit trapezoidal method" (IMEXT) method is proved. To our knowledge, this error bound stating second-order convergence of the IMEXT method closes a current gap in the literature.

Utilising the aforementioned new error bounds for both problem classes, we can rigorously prove convergence of the single- and multi-level methods. Additionally, cost savings of the multi-level methods compared to the single-level approach are predicted and verifed by numerical examples.

The results mentioned above are novel contributions in two areas of mathematics. The first one is (analysis of) numerical methods for uncertainty quantification and the second one is numerical analysis of time-integration schemes for PDEs.

Keywords:
Uncertainty quantification, sparse grids, stochastic collocation method, multi-level method, implicit-explicit methods, splitting methods, Strang splitting, parabolic differential equations, predator-prey equations, Schrödinger equations

Danksagung

Bedanken möchte ich mich zuerst bei Prof. Dr. Tobias Jahnke, der mir nicht nur die (damals für mich unerwartete) Chance gab, eine Promotion anzutreten, sondern der bis zuletzt auch jede noch so kleine Lücke in seinem Terminkalender dafür investierte, dass diese auch zu einem erfolgreichen Schluss kommen konnte. Die Reduktion meines Aufwands in der Lehre und zahllose Verbesserungsvorschläge beim Lesen vieler Entwürfe waren im vergangenen Semester eine große Hilfe beim Fertigstellen meiner Dissertation. Vielen Dank dafür!

Prof. Dr. Christian Wieners gilt mein Dank für das Übernehmen des zweiten Gutachtens und einige erhellende Gespräche zum Thema Uncertainty Quantification. Die Ratschläge zu meiner Dissertation haben mir sehr geholfen. Ihm gilt auch mein Dank für die angenehme kollegiale Atmosphäre in unserer Arbeitsgruppe, um die er sich stets bemüht.

Prof. Dr. Andreas Rieder möchte ich vor allem für spannende Duelle am Tischkicker – dem sozialen Zentrum unserer Arbeitsgruppe in Abwesenheit von Pandemien – danken.

Daniel Weiß habe ich nicht nur eine hervorragende Ausbildung im Hinblick auf meine Lehrtätigkeit zu verdanken, sondern auch an bereits genanntem Tischkicker! Außerdem möchte ich mich für zahllose (fachliche, gesellschafts- und hochschulpolitische, persönliche, ...) Gespräche und sein stets offenes Ohr und Engagement bedanken, ohne die die hervorragende Atmosphäre und Laune in unserer Arbeitsgruppe zweifellos leiden müssten.

Christian habe ich unzählige lustige bis philosophische Pausen zu verdanken, in denen wir "el gordo" um erhebliche Mengen golden brown (nein, nicht *das* golden brown!) erleichtert haben. Außerdem war mir Christian in den letzten Monaten eine große moralische Unterstützung beim Erklimmen des Promotionsgipfels (... und anderer Gipfel).

Der bisher größte Dank geht zweifellos an Felix und Johnny! Ohne euch hätte ich meine Promotion vermutlich weder angetreten noch zu Ende gebracht. So viele Zeilen, wie ich euch widmen müsste, will ich wirklich nicht schreiben – ZwinkerSmiley!

Meinem UQ-Gefährten Niklas gilt mein Dank für das genaue Korrekturlesen der ersten drei Kapitel meiner Dissertation.

Auch den vielen anderen Gefährtinnen und Gefährten in unserer Arbeitsgruppe bin ich sehr dankbar für die gelungene Arbeitsatmosphäre und in den letzten 2 Jahren besonders die Ablenkung vom Corona-Wahnsinn – vielen Dank an (in alphabetischer Reihenfolge) Daniel Z., Daniele, Johannes, Julian B., Julian K., Kevin, Laura, Lukas, Lydia, Marcel, Michael, Nathalie, Patrick, Philip, Ramin, Simon, Sonja und (last but not least) Tine.

Ein besonderer Dank gilt Dr. Zaza Miminoshvili, der immer für den nötigen musikalischen Ausgleich in Studiums- und Promotionszeit sorgte und der mich vom Einklang von Mathematik und Musik überzeugt hat.

Ohne familiären Rückhalt und bedingungslose Unterstützung in allen Bereichen wäre meine persönliche und berufliche Entwicklung nicht so gut gelaufen, wie sie es jetzt ist. Es ist nicht vorstellbar, wo ich ohne sie wäre. Vielen Dank an Helga, Kurt, Nico, Eva – und Leo! ☺

im März 2022

Contents

List of Figures

List of symbols

I	Identity operator or identity matrix
$\mathcal{L}(\mathcal{X}, \mathcal{Y})$	Space of linear bounded operators $\mathcal{X} \to \mathcal{Y}$
$\mathcal{L}(\mathcal{X})$	Space of linear bounded operators $\mathcal{X} \to \mathcal{X}$
$\mathcal{D}(\mathcal{A})$	Domain of the operator $\mathcal{A}$
$\varrho(\mathcal{A})$	Resolvent set of the operator $\mathcal{A}$
$\omega_{\mathcal{A}}$	Growth bound of the operator $\mathcal{A}$
$\overline{A}$	Closure of a set A
$\operatorname{int}(A)$	Interior of a set A
$\mathcal{B}_{\mathcal{X}}(x, r)$	Open ball in $\mathcal{X}$ with center $x \in \mathcal{X}$ and radius $r > 0$
$L^p_{\mathbb{P}}(\Omega)$	Standard L^p-space with respect to the measure $\mathbb{P}$ on Ω
$L^p(\Omega)$	Standard L^p-space with respect to the Lebesgue measure on Ω
$W^{k,p}(D)$	Sobolev space of k times weakly differentiable functions with derivatives in $L^p(D)$
$C^{\boldsymbol{k}}(\Gamma, X)$	Function space of continuously differentiable functions $\Gamma \to \mathcal{X}$ corresponding to multi-index $\boldsymbol{k}$
$\mathbb{P}_k$	Polynomials with real coefficients and degree not larger than k
$\mathbb{E}[Y]$	Expected value of the random variable Y
$\mathbb{V}[Y]$	Variance of the random variable Y
$\operatorname{Cov}[a]$	Covariance function of a random field $a \colon \Omega \times D \to \mathbb{R}$
Γ	Generic parameter set; from Section 2.4 onward $\Gamma = [-1, 1]^d$
$\Sigma(\boldsymbol{\sigma})$	Region bounded by a Bernstein polyellipse, see (2.19)
$\Pi_{K,\infty}$	Set of orthogonal polynomials up to degree K in all variables
$\Pi_{K,1}$	Set of orthogonal polynomials up to total degree K
$\mathcal{U}_n^{p(\ell)}$	One-dimensional interpolation operator with $p(\ell)$ nodes from Section 2.6.1
$\mathcal{I}_L^{p,g}$	Generalised sparse grid interpolation operator
$\mathcal{H}_L^{p,g}$	Underlying sparse grid of $\mathcal{I}_L^{p,g}$

$\eta_L^{p,g}$	Cardinality of the sparse grid $\mathcal{H}_L^{p,g}$
$\Pi_L^{p,g}$	Exact set of the sparse grid $\mathcal{H}_L^{p,g}$
$\mathcal{Q}_n^{p(\ell)}$	One-dimensional quadrature operator with $p(\ell)$ nodes from Section 2.6.3
$\mathcal{Q}_L^{p,g}$	Generalised sparse grid quadrature operator
$\mathcal{R}(\eta, k, d)$	Convergence rate of sparse grid interpolation for functions of finite regularity, see (2.23)
Φ_τ, Φ_{τ,t_n}	Numerical flow of a time integration scheme (different meaning in the individual chapters/sections)
η_ℓ	Number of collocation nodes on level ℓ
$u_{\eta,\tau}^{(\mathrm{SL})}$	Single-level stochastic collocation approximation
$u_J^{(\mathrm{ML})}$	Multi-level stochastic collocation approximation
$\mathcal{I}_\eta$	Generic interpolation operator with η nodes
$C^{(\mathrm{ML})}$	Total computational cost of the multi-level estimator
ν	Outward unit normal
D	Spatial domain in Chapter 4
Σ	Complex parameter set
$A(x, z)$	Elliptic operator defined in (4.4)
$B(x, z)$	First-order boundary operator defined in (4.5)
$W_B^{2,p}(D)$	Sobolev space of twice weakly differentiable functions with first-order boundary restriction, see (4.7)
φ_j	The j-th φ-function, see Definition B.1 in Appendix B
$\mathbb{T} = \mathbb{R}/(2\pi\mathbb{Z})$	One-dimensional torus
$\mathbb{T}_K = \mathbb{R}/(2\pi K\mathbb{Z})$	Scaled one-dimensional torus
$H^s(\mathbb{T}^N)$	Bessel potential space on the N-dimensional torus
$\lVert\cdot\rVert_{\boldsymbol{k},s}$	Norm of the spaces $C^{\boldsymbol{k}}(\Gamma, H^s(\mathbb{T}^N))$, see (5.10)
V	Potential of the NLS, LSE or GPE
$\mathcal{P}^S$	Power set of S
$\mathcal{P}_*^S$	Power set of S without $\emptyset$
$\mathcal{P}_{**}^S$	Power set of S without $\emptyset$ and S
$\Pi(S)$	Set of partitions of S into non-empty subsets

CHAPTER 1

Motivation and Introduction

The laws of physics, chemistry, finance, engineering and biology are often built on partial differential equations (PDEs). Predicting the behaviour of systems from one of the aforementioned sciences by numerical simulations is perhaps the most important task for mathematicians working on numerics and applications. In the last decade the influence of unknown or uncertain parameters to the behaviour and simulation of PDEs has received increasing attention. The main interest in the field of *uncertainty quantification* (UQ) is to understand these influences in any imaginable PDE occuring from real-life phenomena. We give a few examples now. The first three of them are treated in this thesis.

- In biology, *predator-prey models* describe the growth, interaction and possibly eradication of populations of different species in a specified environment. Parameters like food supply for the prey, social friction or environmental and human influences can almost never be quantified satisfactorily and have to be considered with uncertainties to some extent. Clearly, these uncertainties influence the size of the predator and prey populations in a complicated and usually non-linear way.

- In quantum mechanics, central equations of interest are *linear Schrödinger equations* describing wave functions of quantum-mechanical systems. Uncertainties in such equations naturally arise through the potential describing the environment of the system and through the unknown initial state.

- A *non-linear* equation arising in the context of Bose–Einstein condensates is the Gross–Pitaevskii equation which describes the ground state of a system of identical bosons. Here the non-linear behaviour comes from the interaction of the particles. Again, the potential and initial state are unknown.

- In electrodynamics, the *Lugiato-Lefever equation* describes the generation of frequency combs. This equation is a damped non-linear Schrödinger equation with additional driving and detuning terms. The Lugiato-Lefever equation can be used to model the data transmission rate through optical fibers.

Dispersion, detuning and external field parameters heavily influence the shape and frequency range of these combs. Thus, a reliable quantification of the uncertainties introduced by these parameters is crucial for improving data transmission rates in practice.

These four examples (among many others which could have made the list above) demonstrate that a reliable numerical treatment of uncertainties in parameters of PDE systems is both desirable and necessary. Several textbooks offer a good introduction to the topic of uncertainty quantification, e.g. [108, 11, 68]. The treatment of UQ in all of these books is somehow different and each of them presents another perspective on the topic.

The most attractive methods for uncertainty quantification in PDEs are *non-intrusive*. This means that these methods reuse existing (but often very sophisticated) solvers for PDEs with *known* and *deterministic* (in contrast to *unknown* and *uncertain*) parameters. This procedure usually comes at the cost of requiring a huge amount of simulations of deterministic systems. Hence the interesting and non-trivial question is: "How can we use the deterministic solvers in an intelligent way to save a huge amount of computing power?"

Certainly the most popular classes of non-intrusive methods are *stochastic collocation methods* and *Monte Carlo-type methods*. For both of these methods, a large number of PDE solutions for different parameters has to be computed. The difference between the two methods essentially lies in the choice of the parameters for which the PDEs are solved. Loosely speaking, Monte Carlo-type methods always use parameters or samples which are somehow randomised, while stochastic collocation methods rely on parameters with specific approximation properties. For both methods, it is possible to compute important quantities of the solution such as expectations, variances or higher-order moments. For *stochastic collocation methods*, a surrogate for the unknown solution itself can be constructed by interpolation or by using a *generalised polynomial chaos expansion*.

Seminal work on stochastic collocation methods can be attributed to Xiu and Hesthaven [117] and Nobile and Tempone [83, 85, 3]. The literature on stochastic collocation methods will be discussed in detail later in Chapter 3.

Monte Carlo methods, their extensions and numerous applications can be found in [66, 39, 12]. Each of these books features a different point of view on Monte Carlo methods. Specifically for Quasi-Monte Carlo methods, seminal work was done by Niederreiter, Sloan and Woźniakowski [82, 103], see also [102] for further references. It should also be noted that the first successful attemps of using (Quasi-)Monte Carlo methods for differential equations were made in finance, see [13, 94].

One of the most important extensions of the standard Monte Carlo method is the multi-level Monte Carlo method by Heinrich [51] and Giles [36, 37, 38], where the number of samples is more carefully chosen in dependency of the accuracy of the other discretisations involved. This multi-level strategy usually leads to a significant reduction of the computational work. Major contributions on multi-level Monte Carlo methods in the area of UQ can be attributed to Teckentrup, Scheichl and coworkers, see e.g. [19, 15, 27] and [98].

A multi-level approach, however, is also possible for stochastic collocation methods, as introduced and developed in [109, 49, 115]. In these references it was shown that multi-level stochastic collocation (MLSC) methods have a much lower computational cost than standard collocation methods if a high accuracy is required and the regularity of the solution with respect to the uncertain parameters is rather low.

In this thesis, we apply the multi-level stochastic collocation idea to parabolic and Schrödinger equations (linear and non-linear). Here, the second discretisation variable beside the stochastic variable is the *temporal* variable, and not the *spatial* variable. Thus, the multi-level stochastic collocation approach examined here combines the information from approximations computed with different temporal and stochastic refinements via an intelligent strategy. The analysis of the interplay between stochastic and temporal discretisations will require results from numerical analysis of time integration schemes, and in fact a major part of this thesis is to establish such results. In the context of UQ, this interplay between stochastic and temporal discretisation is seldom studied. Usually the focus is on combining different refinement levels of the stochastic and spatial discretisations.

Important note: The spatial discretisation of PDEs is not the focus of this thesis and is only discussed in sections where numerical experiments are shown. The rest of this thesis only deals with temporal discretisations or discretisations of the stochastic/parameter space.

Structure of the thesis. This thesis is structured as follows: In Chapter 2, an introduction to the topic of uncertainty quantification is given, targeted on the methods which are used later on. Chapter 3 discusses the class of (single- and multi-level) stochastic collocation methods. In Chapter 4, these methods are applied to a class of parabolic problems. In Chapter 5, stochastic collocation methods are applied to linear and non-linear Schrödinger equations. The thesis closes with a summary and outlook in Chapter 6.

The links between the chapters are as follows: Chapter 4 and 5 rely on Chapter 2 and 3, although Section 2.2 and 2.5 are not strictly required for Chapter 4 and 5 and are only added for the sake of completeness. Chapter 4 and 5 can be read independently of each other, so Chapter 5 does not require any of the results from Chapter 4 (and vice versa).

Prepublications. Some variants of the results from Chapter 5 in this thesis will appear in a similar form in [61]. Moreover, a few text segments are almost identical to what can be found in the cited work. Other segments in this thesis have already been published (with minor or major changes) in the preprint [62]. We indicate the places where such segments appear.

Computer Architecture. Almost all implementations of the numerical methods in this thesis were realised in Python code. The computations for the numerical experiments in Section 4.6, 4.7 and 5.6 were carried out on an AMD Ryzen Threadripper 2990WX 32-Core Processor machine (unless otherwise stated).

Notation.

- By I, we denote the identity operator or identity matrix (depending on the context).
- The closure and interior of a set A (in an underlying topological space) are denoted by $\overline{A}$ and $\operatorname{int}(A)$, respectively.
- The p-norm in $\mathbb{R}^d$ or $\mathbb{C}^d$ is denoted by $|\cdot|_p$.
- The space of continuous maps $X \to Y$ between topological spaces is denoted by $C(X,Y)$. If $X \neq \emptyset$ is compact and Y a normed space, then it is given the norm

  $$\|u\|_{C(X,Y)} = \sup_{x \in X} \|u(x)\|_Y$$

  for $u \in C(X,Y)$.
- The space of linear and bounded operators $\mathcal{X} \to \mathcal{Y}$ between normed spaces is denoted by $\mathcal{L}(\mathcal{X},\mathcal{Y})$ and carries the norm

  $$\|\mathcal{A}\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})} = \sup_{\|x\|_{\mathcal{X}} \leq 1} \|\mathcal{A}x\|_{\mathcal{Y}}, \qquad \mathcal{A} \in \mathcal{L}(\mathcal{X},\mathcal{Y}).$$

  In the special case $\mathcal{X} = \mathcal{Y}$, we simply write $\mathcal{L}(\mathcal{X})$ instead of $\mathcal{L}(\mathcal{X},\mathcal{X})$.
- For a linear operator $\mathcal{A}\colon \mathcal{X} \supseteq \mathcal{D}(\mathcal{A}) \to \mathcal{Y}$ between Banach spaces $\mathcal{X}$ and $\mathcal{Y}$, the domain $\mathcal{D}(\mathcal{A})$ is usually equipped with the graph norm defined by

  $$\|x\|_{\mathcal{D}(\mathcal{A})} = \|x\|_{\mathcal{X}} + \|\mathcal{A}x\|_{\mathcal{Y}}, \qquad x \in \mathcal{D}(\mathcal{A}).$$

  The operator $\mathcal{A}$ is closed if and only if $\mathcal{D}(\mathcal{A})$ equipped with the graph norm is a Banach space.
- For a closed operator $\mathcal{A}\colon \mathcal{X} \supseteq \mathcal{D}(\mathcal{A}) \to \mathcal{X}$, the resolvent set is defined as

  $$\varrho(\mathcal{A}) = \{\lambda \in \mathbb{C} \mid \lambda I - \mathcal{A}\colon \mathcal{D}(\mathcal{A}) \to \mathcal{X} \text{ is bijective}\}.$$

- If $\mathcal{X}$ and $\mathcal{Y}$ are Banach spaces and $\mathcal{O} \subseteq \mathcal{X}$ is an open set, then the derivative of a map $f\colon \mathcal{O} \to \mathcal{Y}$ is denoted by f' and usually understood in the sense of Fréchet (unless otherwise stated). Thus, f is (Fréchet) differentiable in $x \in \mathcal{O}$ if there exists $A \in \mathcal{L}(\mathcal{X},\mathcal{Y})$ such that

  $$\lim_{\|h\|_{\mathcal{X}} \to 0} \frac{\|f(x+h) - f(x) - Ah\|_{\mathcal{Y}}}{\|h\|_{\mathcal{X}}} = 0.$$

  In this case, $f'(x) = A$ is the Fréchet derivative of f at x. If f is differentiable in every $x \in \mathcal{O}$, then f' defines a map $f'\colon \mathcal{O} \to \mathcal{L}(\mathcal{X},\mathcal{Y})$. If this map f' is continuous, too, then we say that f is continuously (Fréchet) differentiable and write $f \in C^1(\mathcal{O},\mathcal{Y})$.

  Higher Fréchet derivatives are defined iteratively: Let $\mathcal{L}^m(\mathcal{X},\mathcal{Y})$ denote the space of all continuous m-linear maps $\prod_{j=1}^m \mathcal{X} \to \mathcal{Y}$. With the identifications $\mathcal{L}^0(\mathcal{X},\mathcal{Y}) \simeq \mathcal{Y}$ and $\mathcal{L}(\mathcal{X},\mathcal{L}^{m-1}(\mathcal{X},\mathcal{Y})) \simeq \mathcal{L}^m(\mathcal{X},\mathcal{Y})$ for $m \geq 1$, we say that f is m times continuously differentiable if $f', \ldots, f^{(m)}$ exist on $\mathcal{O}$ and

  $$f^{(m)}\colon \mathcal{O} \to \mathcal{L}^m(\mathcal{X},\mathcal{Y})$$

is continuous. If this is the case, then we write $f \in C^m(\mathcal{O}, \mathcal{Y})$. More details on this definition can be found in [1, Chap. VIII, Sec. 5].

Partial Fréchet derivatives occur in some places, too. They are denoted as classical partial derivatives, i.e. the partial derivative of f with respect to the variable u is denoted by $\partial_u f$.

- For abstract probability spaces $(\Omega, \Sigma, \mathbb{P})$, $L^p_{\mathbb{P}}(\Omega)$ denotes the L^p-space with respect to the measure $\mathbb{P}$. If no lower index is given and $\Omega \subseteq \mathbb{R}^N$, then the underlying measure is the Lebesgue measure.
- For open $\Omega \subseteq \mathbb{R}^N$, the space $W^{k,p}(\Omega)$ is the Sobolev space of k times weakly differentiable functions with derivatives in $L^p(\Omega)$.
- For multi-indices $\boldsymbol{j} = (j_1, \dots, j_d)$, $\boldsymbol{k} = (k_1, \dots, k_d) \in \mathbb{N}_0^d$, we write $\boldsymbol{j} \leq \boldsymbol{k}$ if $j_\ell \leq k_\ell$ for $\ell = 1, \dots, d$.
- Let $\emptyset \neq \Gamma \subseteq \mathbb{R}^d$ be compact. For multi-indices $\boldsymbol{k} = (k_1, \dots, k_d) \in \mathbb{N}_0^d$, we define anisotropic spaces of continuously differentiable functions by

$$C^{\boldsymbol{k}}(\Gamma, X) = \left\{ w \colon \Gamma \to X \;\middle|\; \partial_y^{\boldsymbol{j}} w \in C(\Gamma, X),\ \boldsymbol{j} = (j_1, \dots, j_d),\ \mathbf{0} \leq \boldsymbol{j} \leq \boldsymbol{k} \right\}$$

  with corresponding norm

$$\|w\|_{C^{\boldsymbol{k}}(\Gamma,X)} = \max_{\mathbf{0} \leq \boldsymbol{j} \leq \boldsymbol{k}} \|\partial_y^{\boldsymbol{j}} w\|_{C(\Gamma,X)}.$$

- The space of polynomials with real coefficients and degree not larger than k is denoted by $\mathbb{P}_k$.

CHAPTER 2

Uncertainty quantification (UQ)

Assume that we are given a general equation of the form

$$\mathcal{L}(u, \omega) = 0 \tag{2.1}$$

with some operator $\mathcal{L}\colon \mathcal{X} \times \Omega \to \widetilde{\mathcal{X}}$ which acts between Banach spaces $\mathcal{X}$ and $\widetilde{\mathcal{X}}$ and takes an additional argument ω which belongs to a probability space $(\Omega, \Sigma, \mathbb{P})$. This argument ω accounts for the uncertainty in the equation. As we are interested in the situation where $\mathcal{L}$ is a PDE (or a system of PDEs) with spatial and temporal unknowns, let us think of the following situations: The uncertainty in $\mathcal{L}$ enters through uncertain parameters of the equations, uncertain boundary or initial conditions, an unknown spatial domain or some other influences on the equations. The latter might be of interest if it is not clear whether $\mathcal{L}$ accurately models the underlying real-world problem under consideration or not. These uncertainties occur naturally since precise knowledge of all these influences is usually not available (or it is available, but either too complicated to be incorporated in the equation or we are too ignorant to do so). Clearly, a solution u of (2.1) will depend on ω in general, i.e. $u = u(\omega)$. By a solution of (2.1) we mean a function u which satisfies the equation $\mathbb{P}$-almost surely.

In this situation, we may loosely summarise the goal of *uncertainty quantification* as "understanding the solution map $\mathcal{S}\colon \Omega \to \mathcal{X},\ \omega \mapsto u(\omega)$". Of course, different questions may be asked in order to gain this understanding. But even if a specific question can be asked, there are usually numerous ways to approach its answer. If we regard $\mathcal{S}$ as a random variable, one may ask for the probability density or distribution function of $\mathcal{S}$. For almost any problem of interest, these two questions are way too hard. Simpler tasks could be to determine good approximations of stochastic moments of the random variable $\mathcal{S}$, or other statistical quantities. Another simplification of practical interest is not to ask for the solution itself, but rather for a functional of the solution, i.e. the random variable $\omega \mapsto \Upsilon(u(\omega))$ for some explicitly known $\Upsilon\colon \mathcal{X} \to \mathbb{R}$. Such a functional Υ is usually called a *quantity of interest*. Determining this random variable is often a much simpler problem which can sometimes be solved up to an acceptable error. The total error usually contains contributions from several discretisations, since spatial and temporal (or other types of)

discretisations typically have to be combined with some kind of stochastic discretisation. As this work focuses on numerical analysis, we are not only interested in the approximative solution or a functional of it, but in error margins for this total error, too. In general, the relative or absolute error cannot be predicted, but we can sometimes answer how much the error could be reduced by refining the involved discretisations and how much work (meaning computational work, CPU time or simply wall time) is necessary to achieve this.

Let us briefly acknowledge that the previously defined goal of uncertainty quantification is actually only the goal of *forward* uncertainty quantification. Of course, it is also important to consider *inverse problems* under the influence of uncertainty. This equally interesting topic of *backward uncertainty quantification*, however, is not treated in this work.

2.1 From stochastic to parametric PDEs

Although the previous description uses probabilistic notions, a numerical treatment of (2.1) must be in some way deterministic in order to perform meaningful and reproducible computations. The strategy to arrive there is the following: First, we assume that the uncertainty or randomness in the model enters only through a finite number of uncorrelated or even independent random variables. This is the *finite-dimensional noise assumption*. These finitely many, say d, real-valued random variables $Y_1, \dots, Y_d$ can be characterised by a few rather general properties such as the range of values they take, their mean and their variance. We explain the finite-dimensional noise assumption and why it is reasonable in detail later. Among many references which discuss the finite-dimensional noise assumption, we refer to [5, Sec. 2.5], [83, Sec. 1.1] and [3, Sec. 1.1].

Under the finite-dimensional noise assumption, we arrive at the problem of solving an equation[1]

$$\mathcal{L}(\widetilde{u}(\omega), Y_1(\omega), \dots, Y_d(\omega)) = 0$$

for $\widetilde{u}$ and $\mathbb{P}$-almost every $\omega \in \Omega$. For any joint realisation of $Y_1, \dots, Y_d$, say $y_1 = Y_1(\omega), \dots, y_d = Y_d(\omega)$ with $\omega \in \Omega$, the problem of finding $\widetilde{u}(\omega)$ is then a *deterministic* one. Of course not every realisation of this d-dimensional random vector $Y = (Y_1, \dots, Y_d)$ is equally probable, so we have to somehow incorporate the probability of this realisation in some way. We denote the set of realisations by Γ, i.e.

$$\Gamma = Y(\Omega) = \{(Y_1(\omega), \dots, Y_d(\omega)) \colon \omega \in \Omega\}.$$

If we assume that Y admits a bounded and measurable probability density function $\varrho \colon \Gamma \to [0, \infty)$ with respect to the Lebesgue measure on $\mathbb{R}^d$, then we obtain the expected value of Y (if $Y \in L^1_{\mathbb{P}}(\Omega, \Gamma)$) as

$$\mathbb{E}[Y] = \int_\Omega Y(\omega) \mathrm{d}\mathbb{P}(\omega) = \int_\Gamma y \varrho(y) \mathrm{d}y$$

or the probability of a (measurable) event $E \subseteq \Gamma$ as

$$\mathbb{P}(\{\omega \in \Omega \colon Y(\omega) \in E\}) = \int_E \varrho(y) \mathrm{d}y.$$

This procedure now allows us to replace the abstract probability space

$$(\Omega, \Sigma, \mathbb{P}) \qquad \text{by} \qquad (\Gamma, \mathcal{B}(\Gamma), \varrho(y)\mathrm{d}y), \tag{2.2}$$

[1] strictly speaking, the function $\mathcal{L}$ differs from the one in (2.1) in the structure of its input arguments

where $\mathcal{B}(\Gamma)$ denotes the σ-algebra of Borel sets on Γ. By Doob-Dynkin's lemma (see e.g. [89, Lem. 2.1.2]), the "former" solution map $\widetilde{u}\colon \Omega \to \mathcal{X}$ is Y-measurable if and only if the "new" map $u\colon \Gamma \to \mathcal{X}$ with $u(Y(\omega)) = \widetilde{u}(\omega)$ is Borel measurable. Now the initial stochastic problem (2.1) has turned into a deterministic one with an additional d-dimensional parameter. Thus, we search for a function $u = u(y)$ such that

$$\mathcal{L}(u(y_1, \dots, y_d), y_1, \dots, y_d) = 0 \qquad \text{or briefly} \qquad \mathcal{L}(u(y), y) = 0 \tag{2.3}$$

holds for $\varrho(y)\mathrm{d}y$-almost every $y = (y_1, \dots, y_d) \in \Gamma$.

Now we may regard the solution u as a function of $y \in \Gamma$ instead of ω. The random variable $\omega \mapsto \widetilde{u}(\omega)$ is then completely determined (up to null sets) by $y \mapsto u(y)$. Thus, we have replaced the problem of determining the first of these functions $\widetilde{u}$ by determining the second one u instead.

The reader may ask the legitimate question whether the *finite-dimensional noise assumption* is reasonable. To legitimise it, we explain two common approaches to expand uncertain input parameters as a series and how truncating this series leads to a problem which satisfies the finite-dimensional noise assumption. The two approaches are *Karhunen-Loève* and *(generalised) polynomial chaos* expansions. Both of them belong to the class of *spectral expansions*, see e.g. [108, Chap. 11]. The Karhunen-Loève expansion is often used to represent input random variables of a given problem, whereas the polynomial chaos expansion is usually used to represent the random variable/field/process which corresponds to the solution. We give a reason for that in the end of Section 2.2.

If the finite-dimensional noise assumption is dropped and one considers infinitely many random variables instead (i.e. $d = \infty$), then intelligent strategies to deal with infinitely many uncertain inputs have to be derived since for numerical computations, the set of input random variables must always be truncated somehow. Some works treat this case, see e.g. [21, 64, 16]. See also the review article [20] for a detailed discussion on the treatment of general parametric (high-dimensional) PDEs.

We proceed with the explanation of the two aforementioned spectral expansions of random variables.

2.2 Karhunen-Loève expansions

The *Karhunen-Loève expansion* is a very practical representation of random fields as it naturally admits a truncation. The truncated series then leads us to a substitute problem which automatically satisfies the finite-dimensional noise assumption mentioned in the previous section.

For a classical introduction to the Karhunen-Loève expansion, see [69, §37.5], [108, Chap. 11.1], [112] or [35, Chap. 2.3.1]. In the latter reference, a rather detailed derivation with most of the computations is included. Our presentation here is more compact and similar to the ones given in [30, 4, 5] and [83]. Since perfoming a Karhunen-Loève expansion is usually part of the process of replacing an abstract probability space with a more usable one as in (2.2), it is reasonable to start with an abstract probability space here.

Thus, let $(\Omega, \Sigma, \mathbb{P})$ be a probability space and $D \subseteq \mathbb{R}^N$ be a bounded domain. Let $a\colon \Omega \times D \to \mathbb{R}$ be a random field with continuous covariance function $\mathrm{Cov}[a]\colon \overline{D} \times \overline{D} \to \mathbb{R}$ defined by

$$\mathrm{Cov}[a](x_1, x_2) = \mathbb{E}\Big[\big(a(\cdot, x_1) - \mathbb{E}[a(\cdot, x_1)]\big)\big(a(\cdot, x_2) - \mathbb{E}[a(\cdot, x_2)]\big)\Big]$$

for $x_1, x_2 \in D$. Then the compact selfadjoint covariance operator

$$L^2(D) \to C(\overline{D}) \hookrightarrow L^2(D), \qquad g \mapsto \int_D \mathrm{Cov}[a](\cdot, x) g(x) \mathrm{d}x$$

has a sequence of non-negative eigenvalues $(\lambda_k)_{k\in\mathbb{N}}$ and corresponding mutually orthonormal eigenfunctions $(b_k)_{k\in\mathbb{N}}$. The non-negative eigenvalues can be arranged in such a way that

$$\|\operatorname{Cov}[a]\|_{L^2(D\times D,\mathbb{R})} \geq \lambda_1 \geq \lambda_2 \geq \cdots \geq 0$$

and

$$\sum_{k=1}^{\infty} \lambda_k = \int_D \mathbb{V}[a(\cdot,x)](x)\mathrm{d}x.$$

Moreover, there exists a sequence of random variables $\{Y_k\}_{k=1}^{\infty}$ which are mutually uncorrelated, have zero mean and unit variance such that

$$a(\omega,x) = \mathbb{E}[a(\cdot,x)] + \sum_{k=1}^{\infty} \sqrt{\lambda_k} b_k(x) Y_k(\omega),$$

with equality understood in the sense of convergence in $L^2_{\mathbb{P}}(\Omega, L^2(D))$. This is the *Karhunen-Loève expansion* of the random field a. Whenever $\lambda_k > 0$, the random variables Y_k can be written as

$$Y_k(\omega) = \frac{1}{\sqrt{\lambda_k}} \int_D (a(\omega,x) - \mathbb{E}[a(\cdot,x)]) b_k(x)\mathrm{d}x.$$

The *truncated Karhunen-Loève expansion* is now defined as

$$a_K : \Omega \times D \to \mathbb{R}, \qquad a_K(\omega,x) = \mathbb{E}[a(\cdot,x)] + \sum_{k=1}^{K} \sqrt{\lambda_k} b_k(x) Y_k(\omega). \tag{2.4}$$

The question of convergence as $K \to \infty$ in some other norms beside $\|\cdot\|_{L^2_{\mathbb{P}}(\Omega,L^2(D))}$ occurs naturally. The answer is given by Mercer's theorem, see [95, p. 245], [69, p. 145] or [108, Thm. 11.3]. (The formulation of the theorem differs significantly from one source to the next.)

Theorem 2.2.1 (Mercer)**.** *Under the previous assumptions, the following statements hold.*

(a) For $K \to \infty$, we have

$$\sup_{x\in D} \mathbb{E}[(a(\cdot,x) - a_K(\cdot,x))^2] = \sup_{x\in D} (\mathbb{V}[a(\cdot,x)] - \mathbb{V}[a_K(\cdot,x)]) = \sup_{x\in D} \sum_{k=K+1}^{\infty} \lambda_k b_k^2(x) \to 0.$$

(b) Under the additional assumptions that

- *the sets $Y_k(\Omega)$ are uniformly bounded in $k \in \mathbb{N}$,*
- *the covariance function $\operatorname{Cov}[a]$ is smooth (meaning C^∞) on $\overline{D \times D}$ and*
- *for some $s > 1$,*

$$\sqrt{\lambda_k}\|b_k\|_{L^\infty(D)} = \mathcal{O}\left(\frac{1}{1+k^s}\right), \qquad k \to \infty,$$

we additionally obtain

$$\|a - a_K\|_{L^\infty_{\mathbb{P}}(\Omega,L^\infty(D))} \to 0, \qquad K \to \infty.$$

We stress that although the Karhunen-Loève expansion is rather simple to digest theoretically, its practical usage is often limited by either correctly predicting the covariance function of a random field or – if the covariance function is known – determining the eigenpairs of the covariance operator. Practical ways to determine these eigenpairs are given in [30]. In the cited work, several estimates for the decay

of the Karhunen-Loève series in dependency of the regularity of the covariance kernel $\mathrm{Cov}[a]$ are given, too. They can be summarised as follows: The smoother the covariance kernel, the faster the decay of the eigenvalues λ_k as $k \to \infty$. If the covariance kernel is (real) analytic, then the eigenvalues decay exponentially, but if $\mathrm{Cov}[a]$ has only finite Sobolev regularity, then the decay is only algebraic. Other decay estimates for the Karhunen-Loève series are given in [99, 111]. Such results are important since only a rapidly decaying Karhunen-Loève series allows us to choose a small value of K while keeping a good approximation $a_K \approx a$.

In practice, a is not a quantity which is known or given. Instead, it is part of the modelling process to determine a surrogate for the unavailable function a. Often, this is done in a somehow reversed process with the following steps:

- Determine a reasonable covariance operator by measuring e.g. correlation lengths between points (see the example below for the definition of "correlation length").
- Determine the eigenpairs $(\lambda_k, b_k)_{k=1}^{\infty}$ of this covariance operator.
- Define "standard" fluctuations Y_k for $k \in \mathbb{N}$, for example $Y_k \sim \mathcal{U}(0,1)$ (uniform distribution) or $Y_k \sim \mathcal{N}(0,1)$ (normal distribution).
- Guess the expected value $\mathbb{E}[a(\cdot, x)]$ for $x \in D$.
- Define a as the Karhunen-Loève expansion corresponding to $(\lambda_k, b_k, Y_k)_{k=1}^{\infty}$.

This procedure is backed by the preceding paragraph, as basically all sufficiently regular random fields arise as a (in general infinite) Karhunen-Loève series.

Observe that the truncated Karhunen-Loève expansion (2.4) allows us to describe the random variable a through the variables $Y_1, \dots, Y_K$ (at least approximately). Thus, let us briefly return to the context of Section 2.1 and assume that the random variable a from before appears somewhere in $\mathcal{L}$ (for example as a parameter in a PDE). Then it is clear that in the process of replacing a by $Y_1, \dots, Y_K$, a larger value of K corresponds to a larger *dimension* d of the resulting parameter space Γ. It is thus desirable to have a rather small value of K to stay away from high-dimensional parameter spaces and the *curse of dimensionality*.

Let us briefly discuss the most commonly used covariance functions.

Example 2.2.2.

In practice, one often encounters covariance functions of the form

$$\mathrm{Cov}[a](x_1, x_2) = f_a(|x_1 - x_2|_2), \qquad x_1, x_2 \in D,$$

for a function $f_a \colon \mathbb{R}_+ \to \mathbb{R}_+$. A typical choice is the Gaussian kernel given by

$$f_a(z) = \sigma_a^2 \exp\left(-\frac{z^2}{\gamma_a^2 \Lambda_D^2}\right). \tag{2.5}$$

The quantities σ_a, γ_a and Λ_D represent standard deviation and correlation length of a and the diameter of the domain D. As $\mathrm{Cov}[a]$ with f_a from (2.5) admits a holomorphic extension to $\mathbb{C}^d \times \mathbb{C}^d$ (not given by the same formula due to the appearance of $|\cdot|_2$), the eigenvalues of the corresponding Karhunen-Loève expansion decay exponentially, see [99, Prop. 2.19] for a precise statement. Moreover, the larger the

correlation length γ_a is, the smoother is the covariance function and the faster is the convergence of the Karhunen-Loève series.

A slightly more general example is obtained if f_a is replaced by

$$f_{a,p}(z) = \sigma_a^p \exp\left(-\frac{z^p}{\gamma_a^p \Lambda_D^p}\right)$$

for $1 \leq p < \infty$. In the case $p = 1$ and $D = [-b, b]$ with $b > 0$, the eigenpairs of the covariance function can be written down explicitly, see the detailed explanation in [35, p. 26–33]. In this book, some other covariance functions and their eigenpairs are also discussed. A short, but easily accessible explanation on "How to compute eigenpairs of the covariance operator numerically?" for a generic covariance function is given there, too.

For the examples considered in this thesis later on, we usually assume that the uncertain parameters are already given in a simple Karhunen-Loève or polynomial chaos expansion (the latter is discussed in the next section). Thus, we do not examine covariance functions which appear in specific applications, but assume that we are already one step ahead in the modelling process and take some "artificial" uncertain parameters as given.

As we have seen, knowledge of the covariance structure of the random field under consideration is an essential requirement for performing a Karhunen-Loève expansion. Since the covariance structure of *input* random variables of a given problem can often be guessed, these input variables are often represented via Karhunen-Loève expansions. For *solutions* of forward UQ problems, the covariance structure is seldom known and thus such solutions are usually represented via polynomial chaos expansions instead. These are explained in the next section.

2.3 Polynomial chaos expansions

In the previous subsection, we expanded a random function in an affine-linear way in terms of certain "basic" random variables Y_k. Here we derive an expansion of higher polynomial degree with respect to the random variables Y_k. As before, the procedure is well-known and discussed in e.g. [118, 114] and [108, Sec. 11.3]. The *generalised* polynomial chaos expansion we present here was introduced in [118] as an extension to Wiener's classical polynomial chaos expansion from [114].

Here it is assumed that $Y_1, \ldots, Y_d$ are real-valued and stochastically independent random variables, and that each Y_n has a known probability density function $\varrho_n : \Gamma_n \to [0, \infty)$ with respect to the Lebesgue measure. The number of random variables d is denoted by the same letter as the stochastic dimension in Section 2.1 for a good reason (and the index k for Y_k from the previous section is replaced by n here). By stochastically independent, we mean that for all positive measurable functions $f_n \colon \Gamma_n \to \mathbb{R}$, $n = 1, \ldots, d$, we have

$$\mathbb{E}[f_1(Y_1) \cdots f_d(Y_d)] = \mathbb{E}[f_1(Y_1)] \cdots \mathbb{E}[f_d(Y_d)],$$

or – equivalently – the σ-algebras generated by f_n for $n = 1, \ldots, d$ are independent. (See [108, Sec. 2.6] for a brief introduction to stochastic independence.)

The remainder of this section appeared almost literally in [62, Sec. 3.2].

For generalised polynomial chaos expansions, we consider random variables which have finite second moments. Let $n \in \{1, \dots, d\}$ and let $L^2_{\varrho_n}(\Gamma_n)$ be the Hilbert space of measurable, square-integrable functions on Γ_n. Since $Y_1, \dots, Y_d$ are independent by assumption, the probability density function of $Y = (Y_1, \dots, Y_d)$ is given by the product

$$\varrho(y) = (\varrho_1 \otimes \cdots \otimes \varrho_d)(y) = \varrho_1(y_1) \cdots \varrho_d(y_d), \qquad y = (y_1, \dots, y_d) \in \Gamma,$$

where $\Gamma = \Gamma_1 \times \cdots \times \Gamma_d$. The space $L^2_\varrho(\Gamma)$ with norm $\|\cdot\|_\varrho$ induced by the inner product

$$\langle v, w \rangle_\varrho := \int_\Gamma v(y)\overline{w(y)}\varrho(y)\mathrm{d}y \tag{2.6}$$

is again a Hilbert space.

For $n = 1, \dots, d$, let $(\phi_{n,j})_{j \in \mathbb{N}_0}$ be a complete set of real-valued orthonormal polynomials on $L^2_{\varrho_n}(\Gamma_n)$ with the properties $\deg(\phi_{n,j}) = j$ and $\phi_{n,0} \equiv 1$. Such polynomials can be computed efficiently by three-term recursions. Multivariate orthonormal polynomials in $L^2_\varrho(\Gamma)$ can be constructed via tensorisation: If we let

$$\phi_{\boldsymbol{k}}(y) = (\phi_{1,k_1} \otimes \cdots \otimes \phi_{d,k_d})(y) = \prod_{n=1}^{d} \phi_{n,k_n}(y_n), \qquad y = (y_1, \dots, y_d), \tag{2.7}$$

for a multi-index $\boldsymbol{k} = (k_1, \dots, k_d) \in \mathbb{N}_0^d$, then by construction we have $\langle \phi_{\boldsymbol{j}}, \phi_{\boldsymbol{k}} \rangle_\varrho = \delta_{\boldsymbol{jk}}$ for $\boldsymbol{j}, \boldsymbol{k} \in \mathbb{N}_0^d$, where $\delta_{\boldsymbol{jk}}$ is the Kronecker delta. The *generalised polynomial chaos expansion* (gPCE) of $u \in L^2_\varrho(\Gamma)$ is now given by

$$u(y) = \sum_{\boldsymbol{k} \in \mathbb{N}_0^d} u_{\boldsymbol{k}} \phi_{\boldsymbol{k}}(y), \qquad u_{\boldsymbol{k}} = \langle u, \phi_{\boldsymbol{k}} \rangle_\varrho, \tag{2.8}$$

where $\{\phi_{\boldsymbol{k}} \mid \boldsymbol{k} \in \mathbb{N}_0^d\}$ is a complete set of orthogonal polynomials in $L^2_\varrho(\Gamma)$ and equality is understood in the space $L^2_\varrho(\Gamma)$, cf. [118, 114]. The convergence of gPCEs can be established in many cases of interest, since criteria are available which can be verified for most of the usual densities ϱ encountered in practice. It should be noted, however, that for distribution functions that, for example, do not possess finite moments of all orders, the above gPCE may not converge to the correct function. We refer to [29] for a detailed discussion of convergence criteria for gPCEs.

From (2.8), the expectation and the variance of u can easily be derived: Setting $\boldsymbol{0} = (0, \dots, 0) \in \mathbb{N}_0^d$, we get

$$\mathbb{E}[u] = \int_\Gamma u(y)\varrho(y)\mathrm{d}y = \sum_{\boldsymbol{k} \in \mathbb{N}_0^d} u_{\boldsymbol{k}} \int_\Gamma \phi_{\boldsymbol{k}}(y)\varrho(y)\mathrm{d}y = u_{\boldsymbol{0}} \tag{2.9}$$

and

$$\mathbb{V}[u] = \int_\Gamma \Big| \sum_{\boldsymbol{k} \in \mathbb{N}_0^d \setminus \{\boldsymbol{0}\}} u_{\boldsymbol{k}} \phi_{\boldsymbol{k}}(y) \Big|^2 \varrho(y)\mathrm{d}y = \sum_{\boldsymbol{k} \in \mathbb{N}_0^d \setminus \{\boldsymbol{0}\}} |u_{\boldsymbol{k}}|^2 \tag{2.10}$$

due to $u_{\boldsymbol{0}} = \phi_{\boldsymbol{0}}(y)u_{\boldsymbol{0}}$ and $\langle \phi_{\boldsymbol{k}}, \phi_{\boldsymbol{0}} \rangle_\varrho = \delta_{\boldsymbol{k0}}$. From (2.8), one can also derive similar formulas for higher-order moments and other statistical quantities of interest, such as the covariance function and the global sensitivity coefficients, see e.g. [116].

Since only finitely many terms can be computed in practice, we also consider the *truncated polynomial chaos expansion* given by

$$u_\Pi(y) = \sum_{\phi \in \Pi} \langle u, \phi \rangle_\varrho \phi(y) \approx u(y) \tag{2.11}$$

for a set Π which consists of multivariate orthogonal polynomials corresponding to a finite set of multi-indices in $\mathbb{N}_0^d$ which is *downward closed*. This notion is defined as follows.

Definition 2.3.1. A subset $\Xi \subseteq \mathbb{N}_0^d$ is called *downward closed* if it is finite and for each $\boldsymbol{k} \in \Xi$ the set $\{\boldsymbol{k} - \boldsymbol{e}_n \mid n = 1, \dots, d \text{ with } k_n \geq 1\}$ belongs to Ξ. Here, $\boldsymbol{e}_n$ is the n-th canonical unit vector in $\mathbb{R}^d$.

Typical choices for the set Π are:

(i) All polynomials up to degree K in all variables, i.e. $\Pi_{K,\infty} = \{\phi_{\boldsymbol{k}} \colon \ |\boldsymbol{k}|_\infty \leq K\}$.

(ii) All polynomials up to total degree K, i.e. $\Pi_{K,1} = \{\phi_{\boldsymbol{k}} \colon \ |\boldsymbol{k}|_1 \leq K\}$.

The dimensions of the polynomial spaces in these two cases are

$$|\Pi_{K,\infty}| = (K+1)^d \qquad \text{and} \qquad |\Pi_{K,1}| = \binom{K+d}{d} = \binom{K+d}{K}.$$

The first space grows like K^d as $d \to \infty$, and the second one at most as[2] $\exp(K)d^K$. Thus, the second space is much smaller than the first one with increasing dimension. Both examples $\Pi_{K,\infty}$ and $\Pi_{K,1}$ and a third (more generic) choice are depicted in Figure 2.1. The more flexible notation with general Π allows us later to include other polynomial sets which occur as exact sets of sparse grid interpolation operators, such as the one in Figure 2.1(c).

A brief remark on the term "chaos expansion": There is almost nothing "chaotic" in polynomial chaos expansions besides the fact that they are used to represent *random* variables. Although the terminology does not really fit the situation, it is kept for historical reasons and because it is so widely used in the community that it does not make sense to change it.

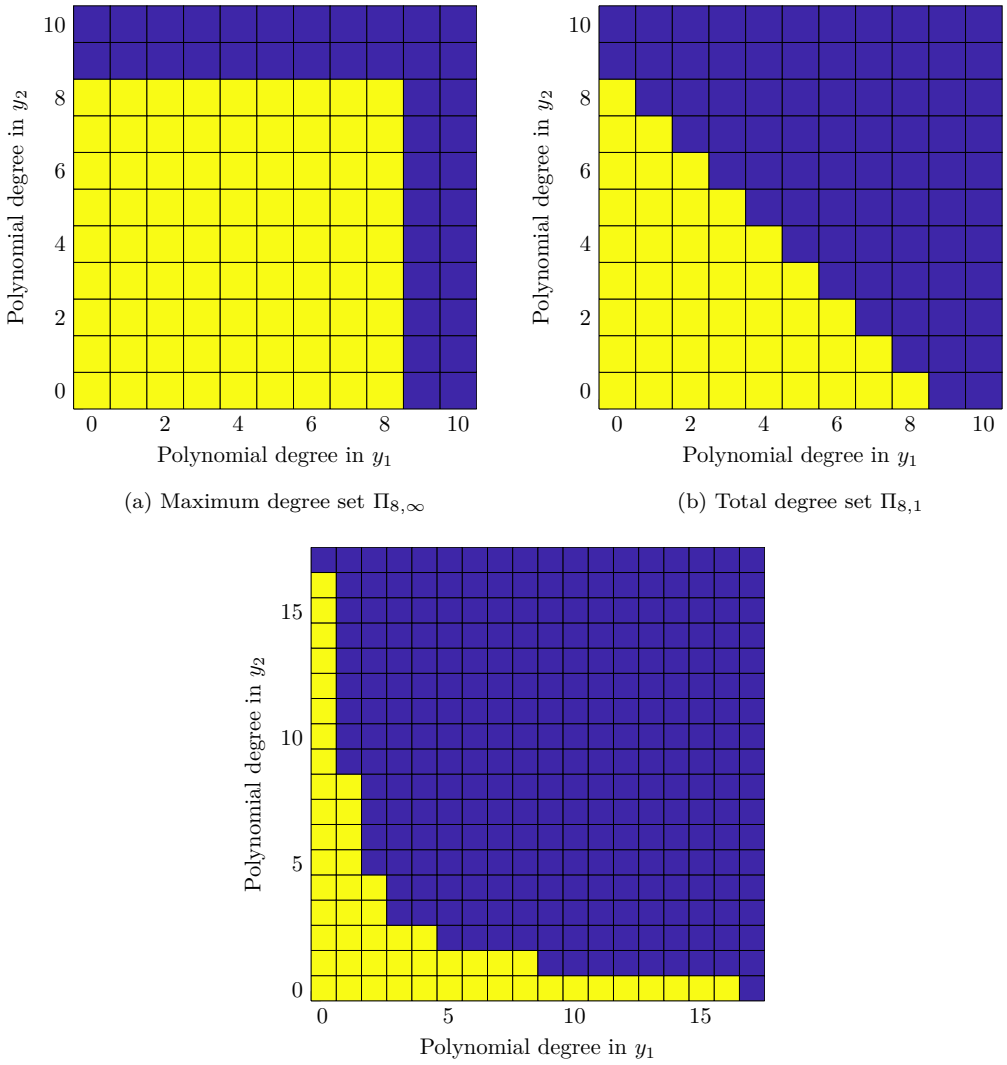
Now that these tools are available, let us briefly connect the previous sections and explain the next step towards *numerical methods* for uncertainty quantification.

2.4 The connection to the following chapters

As we have seen in the previous sections, one of the most important steps before carrying out numerical simulations of stochastic systems such as (2.1) or (2.3), regardless of the form of numerical methods, is to properly identify the basic random variables $Y_1, \dots, Y_d$ so that the relevant input data uncertainty is accurately incorporated in the formulation of the parametric problem. This task is often possible to accomplish when the uncertain inputs are physically meaningful parameters of the system. In this case it is relatively straightforward to identify the independent parameters and to model them as random variables with appropriate distributions based on measurements, experience, or intuition. In Section 2.2, we have seen that the knowledge of the covariance function of the parameters is a powerful tool to achieve this. In this thesis, we will always assume that the random inputs have already been characterised by a set of mutually independent random variables $Y_1, \dots, Y_d$ with good accuracy and focus on the numerical treatment of (2.3) for special classes of operators $\mathcal{L}$.

Before we continue, we stress that by the transformation and simplification from (2.1) to (2.3), we clearly change the solution we initially searched for, too. One may legitimately ask the question why ignoring additional stochastic influences does not fundamentally change the behaviour of the problem

[2] This can be improved significantly, but for the discussion of the limit $d \to \infty$ this crude bound is sufficient.



(a) Maximum degree set $\Pi_{8,\infty}$

(b) Total degree set $\Pi_{8,1}$

(c) Generic (isotropic) set Π

Figure 2.1: Different polynomial spaces span(Π) for the truncation of the gPCE ($d = 2$); pictures (b) and (c) are taken from [62, Fig. 1–2].

or its solution. The answer to this question is somehow unsatisfactory: In almost all scenarios it is too hard to properly quantify the error of replacing the initial stochastic problem with a finite-dimensional parametric one. A noteable exception is the by far most studied example of forward UQ – an elliptic PDE with an uncertain diffusion coefficient. In this setting there are some articles by Schwab and coworkers which examine this type of error, see e.g. [30, 111, 52, 16].

Now we continue with the finite-dimensional parametric problem setting from (2.3). We state the first general assumption which remains throughout this thesis:

Assumption A1. The parameter set $\Gamma = \{(Y_1(\omega), \dots, Y_d(\omega)) \colon \omega \in \Omega\}$ is given by $\Gamma = [-1,1]^d$ and the variable $y \in \Gamma$ corresponds to a realisation of a random variable $Y \sim \mathcal{U}(-1,1)^d$ with uniform probability density $\varrho(y) = 2^{-d}$.

The reader may be surprised about this restrictive assumption, so let us justify this choice. Assumption A1 is mainly made in order to use available estimates for the sparse grid interpolation error in the norm of $L^2_\varrho(\Gamma)$, given later in Section 2.6. For the choices $\Gamma = [-1,1]^d$ and $\varrho(y) = 2^{-d}$ from Assumption A1, such error estimates are well-known. Many of our results could also be adapted to other choices of Γ and ϱ as soon as corresponding estimates are available. Every bounded probability density $\hat{\varrho}$ on the same set Γ could be handled without much effort, since it defines a weaker norm. If $\hat{\varrho}$ is also bounded from below, then the induced norms are even equivalent. For other parameter spaces and probabiliy densities, the estimates are different and our procedure has to be adapted, although it would probably yield rather similar results in the end. In some places we make use of the product structure of $[-1,1]^d$ and thus it is natural to assume that $Y_1, \dots, Y_d$ are independent. (A sketch on the procedure if the random variables are not independent is given in Remark 3.1.2 later.) The requirement that Γ is compact, however, must not be removed. Since practical considerations often make it reasonable to set some (perhaps generous) limits to the random input vector Y, Assumption A1 is not too restrictive from a practical point of view. We stress that this choice of ϱ and Γ guarantees that the generalised polynomial chaos expansion from (2.8) is indeed convergent with the correct limit for any function $u \in L^2_\varrho(\Gamma)$. (One can easily verify the requirements from [29, Thm. 3.6] to see this.)

Now we turn to the discretisation of the finite-dimensional parameter space Γ. More specifically, we continue with the discussion of Monte Carlo methods and sparse grids.

2.5 Monte Carlo and Quasi-Monte Carlo integration

Although Monte Carlo and Quasi-Monte Carlo methods for integration and other purposes are very important classes of methods by themselves, we only describe them very briefly and take one of their disadvantages as a motivation for the construction of sparse grids. It should be noted that Monte Carlo methods are nevertheless very important and they can by no means be entirely replaced by sparse grid techniques. We indicate the limitations of sparse grids and the situations where they are inferior to Monte Carlo methods in the end of this section. Our presentation here follows [108, Sec. 9.5]. Readers familiar with Monte Carlo and Quasi-Monte Carlo methods may skip this section and may wish to continue reading in Section 2.6, since the results below are only presented as a motivation for sparse grids and are not referenced in the remainder of this thesis.

2.5.1 The "vanilla" Monte Carlo method

Suppose $(\Omega, \Sigma, \mathbb{P})$ is a probability space and $Y : \Omega \to \Gamma$ is a μ-distributed random variable. The "vanilla"[3] Monte Carlo method computes approximations to the expected value of a random variable $\omega \mapsto f(Y(\omega))$, where $f : \Gamma \to \mathbb{R}$ is a function which has to be integrated with respect to the density ϱ of Y (e.g. an observable of a solution to a PDE with uncertain parameters). The well-known approximation is

$$E_n(f) = \frac{1}{n}\sum_{i=1}^{n} f(Y^{(i)}) \approx \mathbb{E}[f(Y)] = \int_\Omega f(Y(\omega))\mathrm{d}\mathbb{P}(\omega) = \int_\Gamma f(y)\varrho(y)\mathrm{d}y, \tag{2.12}$$

where $Y^{(i)},\ i = 1, \dots, n$, are independent and identically distributed samples drawn from μ.

Observe that this approximation requires drawing independent samples from μ. This is not always possible and one has to use an intelligent strategy such as a Metropolis–Hastings Markov chain Monte Carlo method to overcome this problem. In such an algorithm, a Markov chain is constructed which has μ as its equilibrium distribution. The state of the chain after sufficiently many steps can be used as a replacement for a sample from μ. Details can be found in, e.g., [108, Sec. 9.5].

The second observation is that the formula for $E_n(f)$ always computes an approximation to some expected value of a random variable, whereas other methods such as the later explained stochastic collocation method compute approximations to the random variable $f(Y)$ itself.

The error of the Monte Carlo integration (2.12) can be quantified as follows. Because of

$$\mathbb{E}[E_n(f)] = \mathbb{E}[f(Y)] \qquad \text{and} \qquad \mathbb{V}[E_n(f)] = \frac{\mathbb{V}[f(Y)]}{n}$$

for any $n \in \mathbb{N}$, we have

$$\mathbb{E}[(E_n(f) - \mathbb{E}[f(Y)])^2] = \mathbb{V}[E_n(f) - \mathbb{E}[f(Y)]] = \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}[f(Y^{(i)})] = \frac{1}{n}\mathbb{V}[f(Y)].$$

The *root mean-square error* is thus given by

$$\sqrt{\mathbb{E}[(E_n(f) - \mathbb{E}[f(Y)])^2]} = \frac{1}{\sqrt{n}}\sqrt{\mathbb{V}[f(Y)]}$$

for $n \in \mathbb{N}$. This means that the convergence rate of the Monte Carlo approximation is proportional to $n^{-1/2}$. This slow convergence of Monte Carlo integration with respect to n is the main disadvantage which sparse grid integration techniques can repair to some extent. We stress that the convergence rate does not depend on the dimension of the domain of f, which is a crucial advantage of the Monte Carlo method no other standard method can offer. This is the reason why Monte Carlo methods are unrivalled for very high-dimensional problems. The variance of $f(Y)$, however, is usually dependent on the dimension.

The convergence rate does not depend on the smoothness of the function f, either. This is often considered a disadvantage if Monte Carlo methods are applied to UQ problems, since many uncertain PDEs (such as the ones considered in this thesis) have solutions which inherently have more smoothness with respect to their uncertain parameters.

Next, we discuss how the convergence rate of Monte Carlo methods can be improved.

[3] the term "vanilla" is taken from [108, Sec. 9.5]

2.5.2 Quasi-Monte Carlo methods

The purpose of *Quasi-Monte Carlo* methods is to improve the convergence rate $n^{-1/2}$ of the vanilla Monte Carlo method. One standard approach to do this is the usage of *low-discrepancy sequences* instead of randomly chosen samples $Y^{(i)}$ as before. Suppose for simplicity that $\varrho \equiv 1$ and $\Gamma = [0,1]^d$.

The remarkable *Koksma-Hlawka theorem* [108, Thm. 9.23] states that for any function $f\colon [0,1]^d \to \mathbb{R}$ with finite Hardy-Krause variation $V^{\mathrm{HK}}(f)$, the inequality

$$\left| \frac{1}{n} \sum_{i=1}^{n} f(y^{(i)}) - \int_{[0,1]^d} f(y) \mathrm{d}y \right| \leq V^{\mathrm{HK}}(f) D^*(y^{(1)}, \dots, y^{(n)})$$

holds for any $y^{(1)}, \dots, y^{(n)} \in [0,1)^d$. The quantity $D^*(y^{(1)}, \dots, y^{(n)})$ is the *star-discrepancy* of the points $y^{(1)}, \dots, y^{(n)}$ defined by

$$D^*(y^{(1)}, \dots, y^{(n)}) = \sup_{R \in \mathcal{R}^*} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_R(y^{(i)}) - \int_{[0,1]^d} \mathbb{1}_R(y) \mathrm{d}y \right|,$$

where $\mathcal{R}^*$ is the set of all rectangles of the form $[0, b_1) \times \cdots \times [0, b_d)$ with $0 \leq b_i < 1$ and $\mathbb{1}_R$ denotes the characteristic function of R. The star discrepancy measures roughly speaking how uniformly the points $y^{(1)}, \dots, y^{(n)}$ are distributed inside the cube $[0,1]^d$. The Hardy-Krause variation is defined in a more complicated way not outlined here[4], but at least we note that it does not require differentiability of f and thus no additional smoothness.

Famous low-discrepancy sequences such as the ones of van der Corput, Halton and Sobol achieve $D^*(y^{(1)}, \dots, y^{(n)}) \leq C(\log(n))^d/n$ and if one uses these sequences instead of random samples, one can improve the error estimate of the Monte Carlo approximation (2.12) to

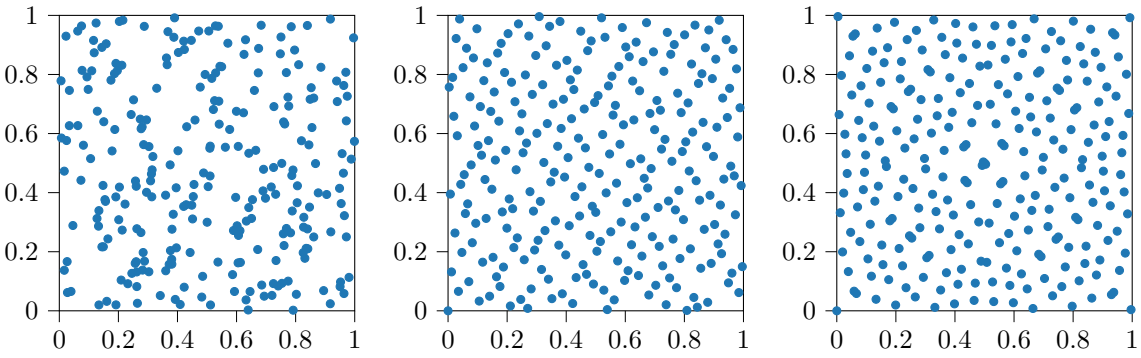
$$\left| \frac{1}{n} \sum_{i=1}^{n} f(y^{(i)}) - \int_{[0,1]^d} f(y) \mathrm{d}y \right| \leq C V^{\mathrm{HK}}(f) \frac{\log(n)^d}{n}.$$

This special choice of nodes $y^{(1)}, \dots, y^{(n)}$ makes it a *Quasi*-Monte Carlo method. In Figure 2.2, we show three sequences of two-dimensional pseudo- or quasi-random vectors. The first one was generated with a standard random number generator (Mersenne Twister), the second one is an extract of a Halton sequence and the third one an extract of the Sobol sequence. From each of the sequences, 256 points are depicted.

Although there is nothing "random" about low-discrepancy sequences, they can loosely be seen as more strategically chosen (pseudo-)random point sets. This can also be verified visually from the pictures in Figure 2.2, since the two low-discrepancy sequences clearly fill the unit cube in a somehow less chaotic way. The reason for this behaviour lies in the number-theoretic properties of these sequences. *Sparse grids* have intrinsically even more structure and thus offer even better approximation properties, but they require more smoothness of the function f. They are discussed in the next section.

Finally we note that the choice of a uniform measure in the above discussion is not a severe restriction since point sets from the cube can often be transformed to the support of other measures (with correct distribution). This is explained in [108, Sec. 9.5].

[4] The Hardy-Krause variation coincides for $d = 1$ with the *total variation* of the function f. For a general definition of the Hardy-Krause variation, see [108, Def. 9.21].



(a) pseudo-random numbers (b) Halton sequence (c) Sobol sequence

Figure 2.2: Two-dimensional pseudo-random and low-discrepancy sequences. The images were generated using SciPy's Quasi-Monte Carlo submodule `scipy.stats.qmc`, see [100].

2.6 Sparse grids

Sparse grids were originally introduced in the context of (spatial) approximation theory on rectangular domains [104, 119], and their usability for UQ was recognized many years later. Nowadays they are well-known and discussed in a plethora of articles, including [34, 113, 86, 87, 88, 14, 107, 6]. In order to have a consistent notation in this thesis, we repeat their construction here. After that, we discuss the approximation properties of sparse grids which are relevant for us and explain how sparse grids can be used for multivariate integration, where they compete with (Quasi-)Monte Carlo methods. Our presentation and notation follow Teckentrup et al. [109, Sec. 5] and Babůška, Nobile and Tempone [3, Sec. 6.1].

2.6.1 Construction of sparse grids

Here we present the construction of sparse grids for the purpose of *interpolation* (in contrast to the previous section, where (Quasi-)Monte Carlo methods were used as a method of *integration* instead). *Integration* with sparse grids is discussed later in Section 2.6.3.

Consider the d-dimensional rectangle $\Gamma = [-1,1]^d$ as in Assumption A1 and one-dimensional interpolation operators for $1 \leq n \leq d$,

$$\mathcal{U}_n^{p(\ell)} \colon C([-1,1]) \to \mathbb{P}_{p(\ell)-1},$$

to be defined shortly. The integer ℓ will control the accuracy of the interpolation operator. The function $p \colon \mathbb{N} \to \mathbb{N}$ is called the *growth rule* and must have the properties

$$p(1) = 1 \qquad \text{and} \qquad p(\ell) < p(\ell+1), \quad \ell = 1, 2, \ldots \tag{2.13}$$

Clearly, p is strictly increasing. The (given) interpolation nodes of $\mathcal{U}_n^{p(\ell)}$ are denoted by

$$\big(y_{n,j}^{(\ell)}\big)_{j=1}^{p(\ell)}.$$

With the Lagrange polynomials $L_{n,j}^{(\ell)}$, $j = 1, \ldots, p(\ell)$, corresponding to exactly these interpolation nodes,

the one-dimensional interpolation operator $\mathcal{U}_n^{p(\ell)}$ is defined as

$$\mathcal{U}_n^{p(\ell)} v(y_n) = \sum_{j=1}^{p(\ell)} v(y_{n,j}^{(\ell)}) L_{n,j}^{(\ell)}(y_n) \tag{2.14}$$

for $v \in C([-1,1])$.

Now consider the multivariate case. Smolyak's idea [104] was to combine certain full tensor grids with few nodes to obtain a sparse grid whose number of points does not grow too fast with the dimension d. For ease of notation, we set $\mathcal{U}_n^{p(0)} = 0$ and define the difference operator $\Delta_n^{p(\ell)}$ as

$$\Delta_n^{p(\ell)} = \mathcal{U}_n^{p(\ell)} - \mathcal{U}_n^{p(\ell-1)}.$$

Let $\boldsymbol{\ell} = (\ell_1, \dots, \ell_d) \in \mathbb{N}^d$ and $L \in \mathbb{N}$. The quantity L will be called the *depth*[5] of the sparse grid and determines the accuracy. The *generalised sparse grid interpolant* is defined as

$$\mathcal{I}_L^{p,g} v = \sum_{g(\boldsymbol{\ell}) \le L} (\Delta_1^{p(\ell_1)} \otimes \cdots \otimes \Delta_d^{p(\ell_d)}) v \qquad \text{for} \qquad v \in C(\Gamma), \tag{2.15}$$

where the function $g \colon \mathbb{N}^d \to \mathbb{N}$ is yet to be specified. So far, we only assume that g is strictly increasing. The *sparse grid* which corresponds to this definition is given by the set

$$\mathcal{H}_L^{p,g} = \bigcup_{\boldsymbol{\ell} \in \mathcal{J}_L} \prod_{n=1}^{d} \left\{ y_{n,j}^{(\ell_n)} \right\}_{j=1}^{p(\ell_n)} \qquad \text{with} \qquad \mathcal{J}_L = \{\boldsymbol{\ell} \in \mathbb{N}^d \colon L - d + 1 \le g(\boldsymbol{\ell}) \le L\}. \tag{2.16}$$

The cardinality of $\mathcal{H}_L^{p,g}$ is denoted by $\eta_L^{p,g}$.

We briefly state the four most common choices for the functions p and g in Table 2.1, although we will use only the "classical" Smolyak choice in this thesis. (The meaning of the last column and $\mathcal{E}_L^{p,g}$ will be explained soon.)

Approximation type	$p(\ell)$	$g(\boldsymbol{\ell})$	$y_1^{k_1} \cdots y_d^{k_d} \in \mathcal{E}_L^{p,g}$
Tensor product	ℓ	$\max_{n=1}^d (\ell_n - 1)$	$\max_{n=1}^d k_n \le L$
Total degree	ℓ	$\sum_{n=1}^d (\ell_n - 1)$	$\sum_{n=1}^d k_n \le L$
Hyperbolic cross	ℓ	$\left(\prod_{n=1}^d \ell_n\right) - 1$	$\prod_{n=1}^d (k_n + 1) \le L + 1$
Smolyak[6] [7]	$2^{\ell-1} + 1$	$\sum_{n=1}^d (\ell_n - 1)$	$\sum_{n=1}^d f(k_n) \le f(L)$

Table 2.1: Common choices of p and g for $\mathcal{I}_L^{p,g}$. The table is taken from [6, Table 1].

We stress that these spaces and hence the resulting grids are all isotropic in the sense that all d dimensions are equally enriched. Only the refinement along the "diagonals" is different, see Figure 2.3. One may define anisotropic sparse grids by using functions $g = g_{\boldsymbol{\alpha}}$ which treat the input variables differently, for example

$$g_{\boldsymbol{\alpha}}(\boldsymbol{\ell}) = \sum_{n=1}^{d} \frac{\alpha_n}{\min_{m=1}^d \alpha_m} (\ell_n - 1), \qquad \boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_d) \in (0, \infty)^d,$$

[5] Perhaps the term "level" would be more appropriate than "depth", but we do not wish to confuse the reader later when another notion of level (in the context of single-level and multi-level collocation methods) appears. Thus, we use "depth" for the parameter which determines the accuracy of a sparse grid.

[6] The given formula for $p(\ell)$ is to be understood as stated except for $\ell = 1$, where we set $p(1) = 1$ as required.

[7] Here, f is defined by $f(0) = 0$, $f(1) = 1$ and $f(p) = \lceil \log_2(p) \rceil$ for $p \ge 2$.

which corresponds to g for the total degree space if $\boldsymbol{\alpha} = (1, \dots, 1)$. See [6, p. 49] for anisotropic versions of the other functions g from Table 2.1. More details on anisotropic sparse grids are given in [3, Sec. 6.1], [84] and [109, Rem. 5.7].

Observe that the sparse grid $\mathcal{H}_L^{p,g}$ from (2.16) is a union of different tensorised one-dimensional grids. An analogous statement holds for the multivariate polynomial sets $\mathcal{E}_L^{p,g}$ on which the corresponding interpolation operators are exact (meaning $\mathcal{I}_L^{p,g} f = f$), see Figure 2.4 for a visual representation of these facts and [8, Lem. 2 and Prop. 3] for precise statements. Now the meaning of the last column in Table 2.1 should be clear, too.

A word of warning about calling $\mathcal{I}_L^{p,g}$ an *interpolant*: It is actually not clear from definition (2.15) whether $\mathcal{I}_L^{p,g}$ is interpolatory in the nodes $\mathcal{H}_L^{p,g}$ or not. In [8, Prop. 6], it was shown that $\mathcal{I}_L^{p,g}$ is interpolatory whenever the one-dimensional grids are *nested*. Now we give an example for this situation.

By far the most common choice of interpolation nodes is $y_{n,1}^{(1)} = 0$ and

$$y_{n,j}^{(\ell)} = -\cos\left(\frac{\pi(j-1)}{p(\ell)-1}\right), \qquad j = 1, \dots, p(\ell), \tag{2.17}$$

for $\ell \geq 2$. These are the extrema of the Chebyshev polynomials (including the endpoints) and often called *Clenshaw-Curtis abscissas* since they are also the nodes of the famous Clenshaw-Curtis quadrature rule [18]. Combined with the "Smolyak" choice from Table 2.1, the abscissas become nested, i.e.

$$\left\{y_{n,j}^{(\ell)}\right\}_{j=1}^{p(\ell)} \subsetneq \left\{y_{n,j}^{(\ell+1)}\right\}_{j=1}^{p(\ell+1)},$$

and thus also the corresponding sparse grids,

$$\mathcal{H}_L^{p,g} \subseteq \mathcal{H}_{L+1}^{p,g}.$$

For sake of completeness, we note that if the function g is chosen as in the "Smolyak" case (but not necessarily the function p, too), then the general formula for the sparse-grid interpolant becomes

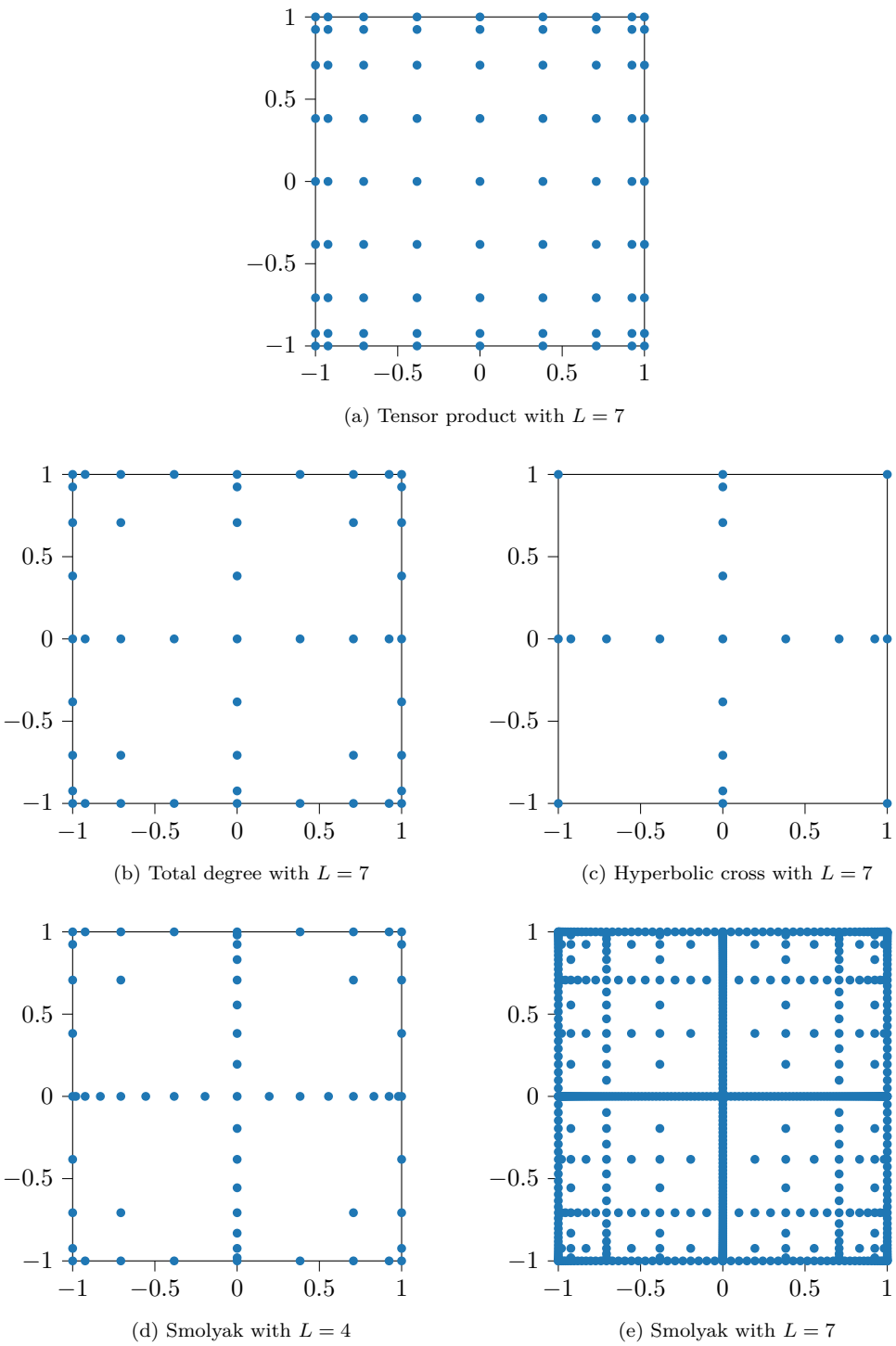
$$\mathcal{I}_L^{p,g} v = \sum_{\boldsymbol{\ell} \in \mathcal{J}_L} c(\boldsymbol{\ell}) (\mathcal{U}_1^{p(\ell_1)} \otimes \cdots \otimes \mathcal{U}_d^{p(\ell_d)})(v), \tag{2.18}$$

where

$$c(\boldsymbol{\ell}) = (-1)^{L+d-|\boldsymbol{\ell}|} \binom{d-1}{L+d-|\boldsymbol{\ell}|} \qquad \text{and} \qquad \mathcal{J}_L = \{\boldsymbol{\ell} \in \mathbb{N}^d \mid L+1 \leq |\boldsymbol{\ell}| \leq L+d\}.$$

This formula is also common and appears in many classical articles on sparse grids, among them [8, 23, 34, 86, 87, 88, 113], and it is in fact more suitable for implementation than the (error-prone) difference representation (2.15). For general functions g, similar formulas can be derived from which it is clear that $\mathcal{I}_L^{p,g}$ is just a linear combination of tensor product interpolation operators (which do not contain the difference operators $\Delta_n^{p(\ell_n)}$ anymore). We do not state these formulas here since we do not use them throughout our work. The interested reader is referred to [3, Eq. (6.2)].

Remark 2.6.1 (Dimension- and depth-dependence of sparse grids)**.** Although sparse grids were invented to *reduce* the amount of nodes required to approximate functions on d-dimensional rectangles, they still require too many nodes with increasing depth and/or dimension (although less than tensor grids, of course). To give the reader an impression of the growth of $\eta_L^{p,g}$, Table 2.2 shows the number of nodes in



(a) Tensor product with $L = 7$

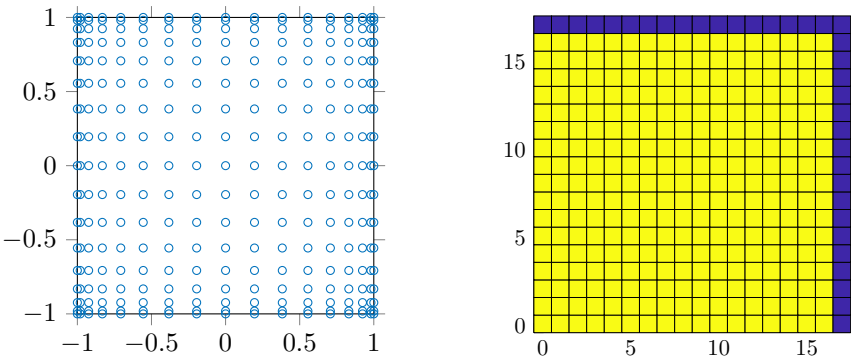
(b) Total degree with $L = 7$

(c) Hyperbolic cross with $L = 7$

(d) Smolyak with $L = 4$

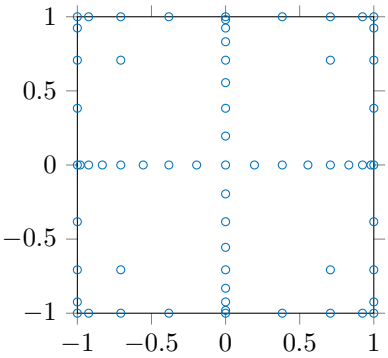
(e) Smolyak with $L = 7$

Figure 2.3: Different types of two-dimensional sparse grids based on Clenshaw-Curtis abscissas. The images were generated using the Tasmanian libraries, see [106, 105].

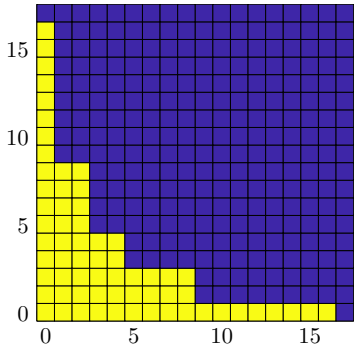


(a) Nodes of full grid

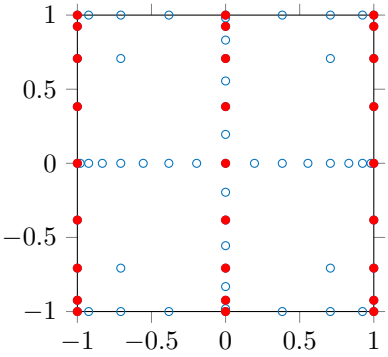
(b) Exact set (yellow)



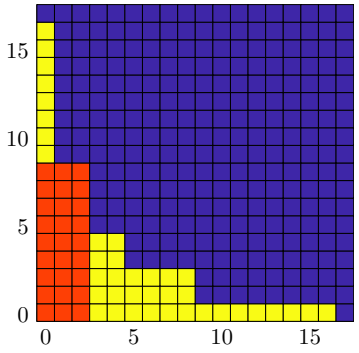
(c) Nodes of sparse grid



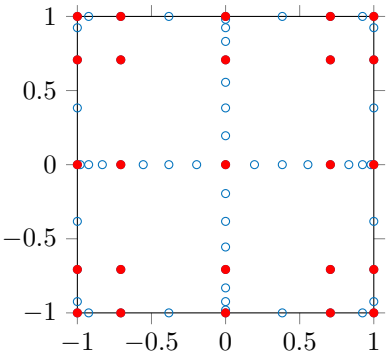
(d) Exact set (yellow)



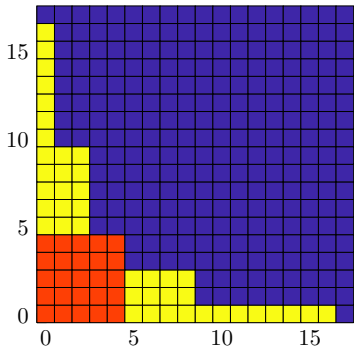
(e) Contained full grid in red



(f) Exact set (yellow + red)



(g) Contained full grid in red



(h) Exact set (yellow + red)

Figure 2.4: Clenshaw-Curtis tensor grid and Smolyak sparse grid of depth $L = 4$ ($d = 2$)

	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$	$L=7$	$L=8$
$d=2$	5	13	29	65	145	321	705	1537
$d=5$	11	61	241	801	2433	6993	19313	51713
$d=10$	21	221	1581	8801	41265	171425	652065	2320385
$d=20$	41	841	11561	120401	1018129	7314609	46106289	261163009

Table 2.2: Growth of $\eta_L^{p,g}$ with d and L

isotropic sparse grids with Smolyak polynomial space and Clenshaw-Curtis abscissas in dependency of d and L.

Especially for dimension $d = 20$, the usage of grids of depths ≥ 4 seems to be rather unpractical, since more than 100.000 stochastic degrees of freedom are usually not feasible in costly PDE simulations. This clearly indicates that *isotropic* sparse grids still suffer from a mitigated curse of dimensionality. ◇

2.6.2 Approximation properties of sparse grids

Infinite regularity. Here we examine the approximation properties of generalised sparse grids for functions with *infinite* regularity, meaning analytic functions on a complex polyellipse.

We remind the reader of the fact from complex analysis that for functions $\Sigma \to \mathcal{X}$ with open $\Sigma \subseteq \mathbb{C}^d$ and a *complex* Banach space $\mathcal{X}$, the terms "analytic" and "holomorphic" are equivalent. So if such a function has one complex derivative, then it is analytic.

Consider for $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_d) \in (1, \infty)^d$ the set

$$\Sigma(\boldsymbol{\sigma}) = \prod_{n=1}^{d} \Sigma(\sigma_n) \subseteq \mathbb{C}^d, \tag{2.19}$$

where $\Sigma(\sigma_n)$ denotes the closed region bounded by the Bernstein ellipse

$$\partial\Sigma(\sigma_n) = \left\{ \frac{1}{2}(z + z^{-1}) \colon z \in \mathbb{C},\ |z| = \sigma_n \right\},$$

see Figure 2.5 for a visualisation.

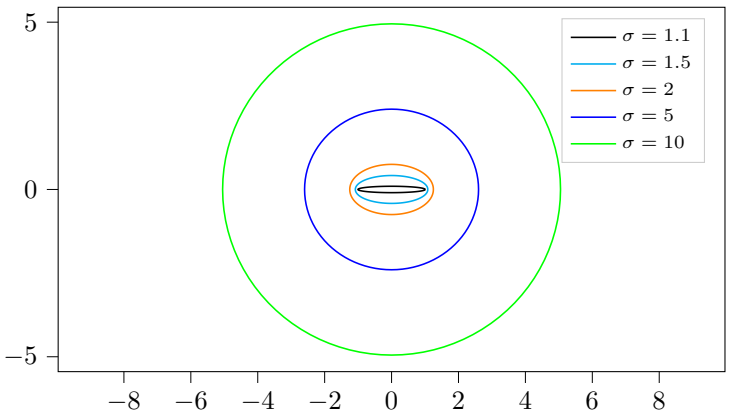


Figure 2.5: Bernstein ellipses $\partial\Sigma(\sigma)$ for different values of $\sigma > 1$

The set $\Sigma(\boldsymbol{\sigma})$ is the cartesian product of ellipses ("polyellipse") in the complex plane with foci ± 1. An analytic extension to such a polyellipse is advantageous for global polynomial interpolation. In the context of sparse grid interpolation, the following theorem from [109, Thm. 5.5] holds for Clenshaw-Curtis abscissas, see also [83, Thm. 3.10 and 3.11] for a more detailed statement with proof, explicit constant and rate. Recall the definition of $\mathcal{I}_L^{p,g}$ from (2.15). Here we choose the functions p and g from the "Smolyak" case in Table 2.1.

Theorem 2.6.2. *Let $\mathcal{X}$ be a Banach space and let $v\colon \Sigma(\boldsymbol{\sigma}) \to \mathcal{X}$ be analytic. Setting $\sigma_{\min} = \min_{n=1}^{d} \sigma_n$, there exist constants $C(\sigma_{\min}, d)$ and $\mu(\sigma_{\min}, d)$ such that*

$$\|v - \mathcal{I}_L^{p,g} v\|_{L^2_\varrho(\Gamma,\mathcal{X})} \leq C(\sigma_{\min}, d)\eta_L^{-\mu(\sigma_{\min},d)} \max_{z\in\Sigma(\boldsymbol{\sigma})} \|v(z)\|_{\mathcal{X}},$$

where p and g are given by the "Smolyak" case from Table 2.1 with Clenshaw-Curtis abscissas (2.17) and where $\eta_L = \eta_L^{p,g}$ is the number of points used by $\mathcal{I}_L^{p,g}$, i.e. $\eta_L = |\mathcal{H}_L^{p,g}|$. The rate $\mu(\sigma_{\min}, d)$ is given by

$$\mu(\sigma_{\min}, d) = \frac{\sigma^*}{1 + \log(2d)} \qquad \text{with} \qquad \sigma^* = \frac{1}{2}\log\left(\sigma_{\min} + \sqrt{1 + \sigma_{\min}^2}\right).$$

Remark 2.6.3 (Dimension-dependence of the convergence rate)**.** From the formula for $\mu = \mu(\sigma_{\min}, d)$, we see that the convergence rate μ deteriorates with increasing dimension d. But in the tensor product case, we would obtain $\mu(\sigma_{\min}, d) \sim \sigma^*/d$, which is clearly much worse than $\mu(\sigma_{\min}, d) \sim \sigma^*/\log(2d)$ in the Smolyak case.

For anisotropic Smolyak sparse grid interpolants, a result similar to Theorem 2.6.2 holds, see [84]. In the anisotropic case, one can sometimes get rid of the dimension-dependence of the convergence rate μ and the constant C. This remarkable fact is explained in [84, Rem. 3.11] and [3, Thm. 6.3] (for Gaussian abscissas) and is one of the rare occasions where the curse of dimensionality can be completely overcome for sparse grid interpolation. ◇

Remark 2.6.4 (Comparison with Monte Carlo sampling)**.** It is also worth comparing Theorem 2.6.2 to the convergence rate of Monte Carlo sampling. Since

$$\|\mathbb{E}[v - \mathcal{I}_L^{p,g} v]\|_{\mathcal{X}} \leq \mathbb{E}[\|v - \mathcal{I}_L^{p,g} v\|_{\mathcal{X}}] \leq \|v - \mathcal{I}_L^{p,g} v\|_{L^2_\varrho(\Gamma,\mathcal{X})} \leq C\eta_L^{-\mu},$$

sparse grids are not really relevant in cases where $\mu \leq \frac{1}{2}$, since Monte Carlo sampling would then give a faster convergence rate for the term on the left (if $\mathbb{E}[\mathcal{I}_L^{p,g} v]$ is replaced by the Monte Carlo estimator). In the light of the previous remark, this is (for fixed σ^*) a limitation on the dimension d for which sparse grids should be used. Roughly speaking, sparse grids are only more suitable than Monte Carlo sampling if

$$\frac{1}{2} \lesssim \frac{\sigma^*}{\log(2d)} \qquad \text{or, equivalently,} \qquad d \lesssim \frac{1}{2}\exp(2\sigma^*)$$

is satisfied. ◇

Remark 2.6.5 (Subexponential convergence)**.** For depths $L > d/\log(2)$, a much better result is available where *subexponential* convergence can be shown instead of *algebraic* convergence as in Theorem 2.6.2, see [83, Thm. 3.11]. For $d = 2$, the condition $L > 2/\log(2) \approx 2.89$ implies that subexponential convergence is achieved for all depths except $L = 1$ and $L = 2$. For more realistic applications with e.g. $d > 5$, the condition $L > d/\log(2)$ is almost never satisfied in practice (since the grids contain already too many

points for these values of L) and one only gets the algebraic convergence discussed earlier. Thus, we are mainly interested in the situations where algebraic convergence occurs. Nevertheless, we will observe the faster convergence rate in one of our numerical experiments in Example 4.6.4 later on and thus this fact cannot be ignored. ◇

Next we examine the approximation properties of generalised sparse grids for functions with *finite* regularity. This will be used in the context of Schrödinger equations later on.

Finite regularity. Again, we restrict ourselves to the "Smolyak" case from Table 2.1 with Clenshaw-Curtis abscissas (2.17), because we are not aware of a more general statement in the literature. The function p is thus given by

$$p(1) = 1, \qquad p(\ell) = 2^{\ell-1} + 1, \qquad \ell \geq 2. \tag{2.20}$$

The correct function spaces for sparse grid interpolation with finite regularity are spaces of *dominating mixed smoothness*. For a multi-index $\boldsymbol{k} = (k_1, \dots, k_d) \in \mathbb{N}_0^d$, we consider

$$C^{\boldsymbol{k}}(\Gamma, \mathcal{X}) = \left\{ w \colon \Gamma \to \mathcal{X} \,\middle|\, \partial_y^{\boldsymbol{j}} w \in C(\Gamma, \mathcal{X}),\ \boldsymbol{j} = (j_1, \dots, j_d),\ \mathbf{0} \leq \boldsymbol{j} \leq \boldsymbol{k} \right\}$$

with the norm

$$\|w\|_{C^{\boldsymbol{k}}(\Gamma,\mathcal{X})} = \max_{\mathbf{0} \leq \boldsymbol{j} \leq \boldsymbol{k}} \|\partial_y^{\boldsymbol{j}} w\|_{C(\Gamma,\mathcal{X})}, \qquad \|w\|_{C(\Gamma,\mathcal{X})} = \sup_{y \in \Gamma} \|w(y)\|_{\mathcal{X}}.$$

For example, all functions $w \in C^{(1,1)}([-1,1]^2, \mathcal{X})$ have derivatives

$$\frac{\partial w}{\partial y_1},\ \frac{\partial w}{\partial y_2},\ \frac{\partial^2 w}{\partial y_1 \partial y_2} \in C([-1,1]^2, \mathcal{X}).$$

In particular the latter "mixed" derivative is characteristic for such a space. The reader should be reminded of the definition of the classical Sobolev space $H^1((-1,1)^2)$, where functions $w \in H^1((-1,1)^2)$ have

$$\frac{\partial w}{\partial y_1},\ \frac{\partial w}{\partial y_2} \in L^2((-1,1)^2), \qquad \text{but in general} \qquad \frac{\partial^2 w}{\partial y_1 \partial y_2} \notin L^2((-1,1)^2).$$

The error bound for sparse grid interpolation in spaces of dominating mixed smoothness is as follows. For $v \in C^{\boldsymbol{k}}(\Gamma, \mathcal{X})$ with $\boldsymbol{k} = (k, \dots, k) \in \mathbb{N}_0^d$, we have

$$\|v - \mathcal{I}_L^{p,g} v\|_{C(\Gamma,\mathcal{X})} \leq C(k,d)(L+1)^{2d} 2^{-kL} \|v\|_{C^{\boldsymbol{k}}(\Gamma,\mathcal{X})} \tag{2.21}$$

by [83, Eq. (3.28)] for a constant

$$C(k,d) = \frac{c(c(1+2^k))^d}{|c(1+2^k) - 1|},$$

where c does not depend on d or L. The number of nodes in the sparse grid is again denoted by $\eta_L^{p,g}$. Combining the previous discussion with a counting lemma for $\eta_L^{p,g}$, one obtains a version of (2.21) where the L-dependency is replaced by an $\eta_L^{p,g}$-dependency. The following statement was given in [83, Sec. 3.1.1].

Theorem 2.6.6. *Let $\mathcal{X}$ be a Banach space and $v \in C^{\boldsymbol{k}}(\Gamma, \mathcal{X})$ with $\boldsymbol{k} = (k, \dots, k) \in \mathbb{N}_0^d$. The sparse grid interpolation error can be estimated by*

$$\|v - \mathcal{I}_L^{p,g} v\|_{C(\Gamma,\mathcal{X})} \leq C(k,d)\mathcal{R}(\eta_L, k, d)\|v\|_{C^{\boldsymbol{k}}(\Gamma,\mathcal{X})}, \tag{2.22}$$

where p and g are given by the "Smolyak" case from Table 2.1 with Clenshaw-Curtis abscissas (2.17), $\eta_L = \eta_L^{p,g}$ is the number of points used by $\mathcal{I}_L^{p,g}$ and

$$C(k,d) = \frac{(c(1+2^k))^d}{|1+2^k-1/c|},$$
$$\mathcal{R}(\eta,k,d) = \left(1+\log_2\left(1+\frac{\eta}{d}\right)\right)^{2d} \min\left\{2^k \eta^{-\frac{k\log(2)}{1+\log(2d)}}, \eta^{-k}\left(1+\log_2\left(1+\frac{\eta}{d}\right)\right)^{dk}\right\}. \tag{2.23}$$

In particular, the above estimate implies

$$\|v - \mathcal{I}_L^{p,g} v\|_{C(\Gamma,\mathcal{X})} \leq C(k,d)\eta_L^{-k}\left(1+\log_2\left(1+\frac{\eta_L}{d}\right)\right)^{(k+2)d} \|v\|_{C^{\boldsymbol{k}}(\Gamma,\mathcal{X})}.$$

In the earlier paper [8], the estimate

$$\|v - \mathcal{I}_L^{p,g} v\|_{C(\Gamma,\mathcal{X})} \leq C(k,d)\eta_L^{-k}(\log(\eta_L))^{(k+2)(d-1)+1} \|v\|_{C^{\boldsymbol{k}}(\Gamma,\mathcal{X})} \tag{2.24}$$

for $\eta_L > 1$ was given, which has the charm that it is less intricate.

In the next section, we discuss how sparse grids can be used in the context of high-dimensional integration. This procedure is almost identical to the one for interpolation.

2.6.3 Sparse grid integration

We start again by introducing some notation. Let $n \in \{1,\dots,d\}$ and let p be a growth rule, i.e. a function satisfying the conditions (2.13). For $\ell \in \mathbb{N}$, we define a $p(\ell)$-point quadrature formula for one-dimensional integrals over $[-1,1]$ with weight function ϱ_n by

$$\mathcal{Q}_n^{p(\ell)}(f) = \sum_{j=1}^{p(\ell)} \omega_{n,j}^{(\ell)} f(y_{n,j}^{(\ell)}) \approx \int_{\Gamma_n} f(y_n)\varrho_n(y_n)\mathrm{d}y_n,$$

where $f\colon [-1,1] \to \mathbb{C}$. Here, $y_{n,j}^{(\ell)}$ are the nodes and $\omega_{n,j}^{(\ell)}$ are the weights for $j = 1,\dots,p(\ell)$. The formula $\mathcal{Q}_n^{p(\ell)}$ has *degree of exactness*[8] $q_n(\ell) \in \mathbb{N}$ if $\mathcal{Q}_n^{p(\ell)}$ is exact for all polynomials whose degree in y_n is not larger than $q_n(\ell)$. We assume that the degree of exactness of $\mathcal{Q}_n^{p(\ell)}$ increases with ℓ. (This fits to the previous requirement that the growth rule p is increasing, too.)

Clearly, multi-dimensional quadrature formulas for integrals over $\Gamma = [-1,1]^d$ could in principle be constructed via tensorisation: For $f\colon \Gamma \to \mathbb{C}$ and $\boldsymbol{\ell} = (\ell_1,\dots,\ell_d) \in \mathbb{N}^d$ let

$$\mathcal{Q}_{\boldsymbol{\ell}} f = (\mathcal{Q}_1^{p(\ell_1)} \otimes \cdots \otimes \mathcal{Q}_d^{p(\ell_d)})(f) = \sum_{j_1=1}^{p(\ell_1)} \cdots \sum_{j_d=1}^{p(\ell_d)} \prod_{n=1}^{d} \omega_{n,j_n}^{p(\ell_n)} f(y_{1,j_1}^{p(\ell_1)},\dots,y_{d,j_d}^{p(\ell_d)}) \approx \int_\Gamma f(y)\varrho(y)\mathrm{d}y, \tag{2.25}$$

where $\varrho = \varrho_1 \otimes \cdots \otimes \varrho_d$. We call such a quadrature rule a *tensor product quadrature rule* and the corresponding grid

$$\left\{\left(y_{1,j_1}^{p(\ell_1)},\dots,y_{d,j_d}^{p(\ell_d)}\right) : j_n = 1,\dots,p(\ell_n) \text{ for } n = 1,\dots,d\right\}$$

a *full grid* or *tensor grid*. Note that this approximation requires a total number of $p(\ell_1)\cdots p(\ell_d)$ function evaluations.

[8] The degree of exactness is not to be confused with the *order* of a quadrature rule. By definition, a one-dimensional quadrature rule with degree of exactness q has the order $q+1$.

Remark 2.6.7. Now let us briefly return to the gPCE from (2.8). One important application of quadrature formulas for integrals over Γ with weight ϱ is the approximation of the polynomial chaos coefficients $\langle f, \phi \rangle_\varrho$ for the polynomials ϕ in the truncation set Π from (2.11). Our quadrature formula should be able to do this at least if f belongs to the set Π, too. So $\mathcal{Q}_{\boldsymbol{\ell}}$ should be chosen in such a way that

$$\langle \phi, \psi \rangle_\varrho = \int_\Gamma \phi(y)\psi(y)\varrho(y)\mathrm{d}y = \mathcal{Q}_{\boldsymbol{\ell}}(\phi\psi) \qquad \text{for all} \qquad \phi, \psi \in \Pi. \tag{2.26}$$

Since the product $\phi\psi$ is again a polynomial, we define the *half-exact set*

$$\Pi'_{\boldsymbol{\ell}} = \{\phi_{\boldsymbol{k}} \colon \ \boldsymbol{k} \in \mathbb{N}_0, \ k_n \le \lfloor \tfrac{q_n(\ell_n)}{2} \rfloor \ \text{ for } \ n = 1, \dots, d\}, \tag{2.27}$$

with $\phi_{\boldsymbol{k}}$ defined by (2.7). Here $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x. If we choose $\Pi = \Pi'_{\boldsymbol{\ell}}$, then (2.26) holds. Since we have already decided to avoid tensor product quadrature rules, we have to see how Π can be chosen for a quadrature rule based on a sparse grid. ◇

Now we set $\mathcal{Q}_n^{p(0)} = 0$ and define the difference operator[9] $\Delta_n^{p(\ell)}$ as $\Delta_n^{p(\ell)} = \mathcal{Q}_n^{p(\ell)} - \mathcal{Q}_n^{p(\ell-1)}$. The sparse grid quadrature formula is given by

$$\mathcal{Q}_L^{p,g} v = \sum_{g(\boldsymbol{\ell}) \le L} (\Delta_1^{p(\ell_1)} \otimes \cdots \otimes \Delta_d^{p(\ell_d)}) v \qquad \text{for} \qquad v \in C(\Gamma). \tag{2.28}$$

For the "Smolyak" choice of g from Table 2.1, we can simplify the above formula to

$$\mathcal{Q}_L^{p,g}(f) = \sum_{\boldsymbol{\ell} \in \mathcal{J}_L} c(\boldsymbol{\ell}) \mathcal{Q}_{\boldsymbol{\ell}}, \tag{2.29}$$

where

$$c(\boldsymbol{\ell}) = (-1)^{L+d-|\boldsymbol{\ell}|} \binom{d-1}{L+d-|\boldsymbol{\ell}|} \qquad \text{and} \qquad \mathcal{J}_L = \{\boldsymbol{\ell} \in \mathbb{N}^d \colon L+1 \le |\boldsymbol{\ell}| \le L+d\}. \tag{2.30}$$

Remark 2.6.8. Here it is worth to spend some time on investigating the exactness properties of this quadrature rule. We define

$$\Pi_L \coloneqq \bigcup_{\boldsymbol{\ell} \in \mathcal{J}_L} \Pi'_{\boldsymbol{\ell}} \tag{2.31}$$

with $\Pi'_{\boldsymbol{\ell}}$ from (2.27). In [22, Cor. 3.3], it was shown that

$$\Pi_L \subseteq \{\phi_{\boldsymbol{k}} \mid \mathcal{Q}_L^{p,g}(\phi_{\boldsymbol{k}}\phi_{\boldsymbol{k}}) = \langle \phi_{\boldsymbol{k}}, \phi_{\boldsymbol{k}} \rangle_\varrho = 1\}.$$

This set Π_L can be used as a basis for the truncated gPCE, i.e. we can choose $\Pi = \Pi_L$. We stress, however, that (2.26) is usually *not* true for $\Pi = \Pi_L$ and $\mathcal{Q}_L^{p,g}$ instead of $\mathcal{Q}_{\boldsymbol{\ell}}$, i.e. there are $\phi_{\boldsymbol{k}}$, $\phi_{\boldsymbol{j}} \in \Pi_L$ such that $\mathcal{Q}_L^{p,g}(\phi_{\boldsymbol{k}}\phi_{\boldsymbol{j}}) \neq \delta_{\boldsymbol{k}\boldsymbol{j}}$. This leads to the phenomenon that

$$\phi_{\boldsymbol{k}} \neq \sum_{\phi_{\boldsymbol{j}} \in \Pi_L} \mathcal{Q}_L^{p,g}(\phi_{\boldsymbol{k}}\phi_{\boldsymbol{j}})\phi_{\boldsymbol{j}} \qquad \text{instead of} \qquad \phi_{\boldsymbol{k}} = \sum_{\phi_{\boldsymbol{j}} \in \Pi_L} \langle \phi_{\boldsymbol{k}}, \phi_{\boldsymbol{j}} \rangle_\varrho \, \phi_{\boldsymbol{j}}. \tag{2.32}$$

Hence, the usage of such a quadrature rule for gPCEs has to be modified; this will be explained in Section 3.2 later. It should be noted that there is a notable exception to (2.32), namely the case $\boldsymbol{k} = \mathbf{0}$ corresponding to $\phi_{\mathbf{0}} \equiv 1$. Here,

$$1 = \phi_{\mathbf{0}} = \sum_{\phi_{\boldsymbol{j}} \in \Pi_L} \mathcal{Q}_L^{p,g}(\phi_{\mathbf{0}}\phi_{\boldsymbol{j}})\phi_{\boldsymbol{j}} = \sum_{\phi_{\boldsymbol{j}} \in \Pi_L} \mathcal{Q}_L^{p,g}(\phi_{\boldsymbol{j}})\phi_{\boldsymbol{j}} \tag{2.33}$$

[9] We use the same notation $\Delta_n^{p(\ell)}$ as in the earlier section, although the letter $\mathcal{U}$ is now replaced by $\mathcal{Q}$ in the definition, thus slightly overusing notation. No confusion should arise from that.

is indeed true, which is quite useful for computing expected values. We omit a proof of (2.33) since it is a simple consequence of a result we discuss later. ◇

We summarise the choices we have to make in order to construct different sparse grid quadrature formulas: In each dimension, we select a sequence of one-dimensional quadrature formulas $(\mathcal{Q}_n^{p(\ell)})_{\ell\in\mathbb{N}}$. Moreover, we choose a growth rule p and the function g from (2.28). The choice of p should depend on the family of quadrature formulas. As explained before, an exponential growth rule is beneficial for for Clenshaw-Curtis nodes to obtain a nested family of sparse grid rules. The choice of g depends on the available "mixed" regularity of the functions we wish to interpolate or integrate. The depth L is chosen depending on the desired accuracy. These choices together determine the sparse grid, the corresponding weights and the appropriate polynomial basis Π_L which can be used for gPCEs.

One can derive error estimates for sparse grid quadrature formulas in a similar fashion as for sparse grid interpolation. Such results can be found in the work of Novak and Ritter [86, 87, 88] and Wasilkowski and Woźniakowski [113], just to name some of the most important articles (among many others) on this topic. We do not state such results here and we do not use them explicitly in this thesis.

Now that sparse grids are available, we are ready to discuss the class of single- and multi-level stochastic collocation methods in the next chapter.

CHAPTER 3

Stochastic collocation (SC) methods

Standard discretisations of the parameter space Γ in the context of UQ are Monte Carlo or Quasi-Monte Carlo discretisations or variants thereof. These methods are *sampling-based* and thus *non-intrusive*, which means that existing solvers for the corresponding deterministic problems can be used without changes. More sophisticated non-intrusive methods are well-known nowadays, the most important ones being Gaussian process regression and stochastic collocation methods. We consider the latter with the special choice of a sparse grid collocation strategy. It is arguably the most prominent method for moderately high-dimensional parameter spaces. Pioneering work on stochastic collocation was done by Xiu and Hesthaven in [117], and by Nobile, Tempone and Webster in [83], with important follow-up articles [3, 84] for elliptic, [85] for parabolic and [80, 81] for hyperbolic equations. Another early work on stochastic collocation methods in UQ for computational fluid dynamics is [77].

Later in this chapter we explain how stochastic collocation methods can be adjusted to a second discretisation – in our case a temporal discretisation. This leads to the *multi-level stochastic collocation* (MLSC) method which is the main method under consideration for discretising the parameter space in this thesis.

3.1 The method

Consider again the problem from (2.3) of finding u such that

$$\mathcal{L}(u(y_1, \dots, y_d), y_1, \dots, y_d) = 0 \tag{3.1}$$

for $\varrho(y)\mathrm{d}y$-almost every $y = (y_1, \dots, y_d) \in \Gamma$. Suppose that we have a numerical method which computes reliable approximations of the solution of (3.1) for any given $y = (y_1, \dots, y_d) \in \Gamma = [-1, 1]^d$. The application of such a method is often very expensive by itself (think of $\mathcal{L}$ as describing a large PDE system), so we try to avoid random sampling from $\varrho(y)\mathrm{d}y$ and thus omit the computation of a huge number of solutions for different samples $y^{(1)}, \dots, y^{(\eta)} \in \Gamma$. A better option is a non-random choice of the

points in the parameter space Γ. Readers familiar with *Monte Carlo* methods may think of Quasi-Monte Carlo methods, where one replaces (pseudo-)random numbers by low-discrepancy sequences, which are also less arbitrarily chosen in a rather vague sense, as discussed in Section 2.5. For stochastic collocation methods, one chooses a set of points $y^{(1)}, \dots, y^{(\eta)} \in \Gamma$ which allow the construction of an interpolant

$$\mathcal{I}_\eta u(y) = \sum_{j=1}^{\eta} \widetilde{u}_j \ell_j(y) \approx u(y), \tag{3.2}$$

where $\widetilde{u}_j$ is the approximation of the solution to (3.1) for $(y_1, \dots, y_d) = y^{(j)} = (y_1^{(j)}, \dots, y_d^{(j)})$, and where ℓ_j are suitably chosen basis functions. The importance of (3.2) lies in the fact that although only the computation of η approximations $\widetilde{u}_1, \dots, \widetilde{u}_\eta$ is required, we should obtain a good approximation $u(y) \approx \mathcal{I}_\eta u(y)$ for *all* $y \in \Gamma$ if u is sufficiently smooth with respect to y. For sparse grid interpolants, this has already been discussed in Section 2.6.2. The interpolant reveals the fundamental difference between Monte Carlo (including its many variants) and collocation methods: While Monte Carlo methods compute approximations of deterministic quantities such as expectations of (functionals of) the solution, collocation methods compute an approximation or surrogate $\mathcal{I}_\eta u$ of the solution u *itself*.

Of course, the expectation and variance of a quantity of interest Υ of u can be approximated for collocation methods, too, e.g. via

$$\mathbb{E}[\Upsilon(u)] \approx \mathbb{E}[\Upsilon(\mathcal{I}_\eta u)] \approx \sum_{j=1}^{\eta} \Upsilon(\widetilde{u}_j)\mathbb{E}[\ell_j].$$

This approximation suggests that $y^{(j)}$ and $\mathbb{E}[\ell_j]$ should be choosen as nodes and weights of a sufficiently accurate quadrature formula.

One may now define a specific collocation method by simply choosing an interpolant, or equivalently by choosing nodes $y^{(1)}, \dots, y^{(\eta)}$ and corresponding basis functions $\ell_1, \dots, \ell_\eta$. In principle, global or piecewise polynomial or spline interpolation could be used. In the one-dimensional case, the interpolant from (2.14) is one option. This is a global polynomial interpolant since it is based on (global) Lagrange polynomials. In higher dimensions, it is crucial that the number of collocation points does not grow too fast. In regard of the previous section, our collocation methods are always based on sparse grids. Thus, our interpolant $\mathcal{I}_\eta$ is (unsurprisingly) exactly $\mathcal{I}_L^{p,g}$ from (2.15) and $\eta = \eta_L^{p,g} = |\mathcal{H}_L^{p,g}|$.

From a practical perspective it should be noted that the costly part of stochastic collocation methods is usually the computation of the sample solutions $\widetilde{u}_j$, which typically requires a costly PDE simulation for each $j = 1, \dots, \eta$. The construction of the interpolant afterwards is comparatively cheap.

Remark 3.1.1. Students are taught in undergraduate courses on numerical mathematics that polynomial interpolation has several issues and is often inferior to other interpolation strategies, for example (cubic) spline interpolation. The reason for us to stick to global polynomial interpolation is the fact that we use interpolation in the parameter space Γ and the solution u we aim to approximate is often very smooth with respect to the variable $y \in \Gamma$, even analytic in some situations. In these cases global polynomial interpolation is attractive again.

But nevertheless we have to make sure that the interpolation is not based on equidistant nodes, but rather on abscissas with a Lebesgue constant which is not growing too fast. Thus, the following choices are reasonable: Either choose the roots of the Chebyshev polynomials or their extrema. In both cases, the Lebesgue constant only grows logarithmically, see e.g. [58]. Since the roots of the Chebyshev polynomials

do not give a *nested* family of nodes, choosing the extrema of the Chebyshev polynomials (= nodes of the Clenshaw-Curtis quadrature rule) is much better for us. ◇

It is clear that the error of a stochastic collocation method is in fact an interpolation error. Thus, known interpolation error bounds can be used to quantify the error of collocation methods. These error bounds usually have a different form than the error bounds for (Quasi-)Monte Carlo methods, since the latter are formulated using probabilistic notions. The downside of interpolation error bounds is that the appearing optimal constants are seldom known and it is usually not possible to quantify a confidence interval, as is the case for Monte Carlo methods. The good news is that interpolation error bounds are often formulated using norms of standard function spaces and utilise the regularity of the function which is interpolated. Thus, stochastic collocation methods can be examined in a straightforward way from the perspective of numerical analysis.

Stochastic collocation methods are usually limited to moderately large dimensions (up to 10 or perhaps 20 depending on context) since sparse grids contain too many points in even higher dimensions and thus are too costly, as discussed in Remark 2.6.1 before. If the problem has even more dimensions in the parameter space, then either anisotropic sparse grids have to be used for the collocation method or one should fall back to Monte Carlo methods again.

Remark 3.1.2 (Dependent random variables)**.** Here we give a remark similar to [3, Sec. 2] which explains how stochastic collocation methods can be used if the density of the random input vector $Y\colon \Omega \to \Gamma$ from Assumption A1 in Section 2.4 does not factorise, i.e. $\varrho(y) \neq \varrho_1(y_1)\cdots\varrho_d(y_d)$. This means that the entries of Y are not assumed to be stochastically independent. In this case, one can introduce an auxiliary probability density function $\hat{\varrho}\colon \Gamma = [-1,1]^d \to [0,\infty)$ which factorises as

$$\hat{\varrho}(y_1,\dots,y_d) = \prod_{j=n}^{d} \hat{\varrho}_n(y_n), \qquad \text{and additionally satisfies} \qquad \left\| \frac{\varrho}{\hat{\varrho}} \right\|_{L^\infty(\Gamma)} < \infty.$$

It is preferable that the last norm is as small as possible, since it enters the error estimates for collocation methods with general ϱ, see [83, Sec. 3.1.2]. The cited article by Nobile, Tempone and Webster treats this more general case of a density ϱ which does not factorize. The reader who is interested in such a setting is referred to the aforementioned article and the work [3].

As a reasonable set of collocation points for the density ϱ might not be available, one can use quadrature rules for the auxiliary density $\hat{\varrho}$ to compute ϱ-weighted integrals such as

$$\mathbb{E}[v] \approx \bar{v}_L = \mathcal{Q}_{L,\hat{\varrho}}\left(\frac{\varrho}{\hat{\varrho}} v\right) \qquad \text{and} \qquad \mathbb{V}[v] \approx \mathcal{Q}_{L,\hat{\varrho}}\left(\frac{\varrho}{\hat{\varrho}} (v-\bar{v}_L)^2\right),$$

where $\mathcal{Q}_{L,\hat{\varrho}}$ is the sparse grid quadrature operator of depth L from (2.28), but now defined for the separable density $\hat{\varrho}$ instead of ϱ. ◇

A final remark on the terminology: Actually there is nothing "stochastic" about stochastic collocation methods. The term "stochastic" only accounts for the fact that the collocation points are chosen in the stochastic/parameter space.

Now we point out the connection between polynomial chaos expansions and stochastic collocation methods.

3.2 Interlude: Reconstruction of the gPCE

This section appeared in a similar form in [62].

In the previous section we explained that stochastic collocation methods compute an interpolant of the unavailable solution. In fact, once an interpolant $\mathcal{I}_\eta u$ of the solution u as in (3.2) has been computed, it is possible to reconstruct an approximation of the polynomial chaos expansion of u, which was introduced in Section 2.3. However, it is not straightforward to do this for *sparse grid* interpolants, which is why we devote this whole section to the procedure.

Let $f \in L^2_\varrho(\Gamma)$ be given by

$$f(y) = \sum_{\boldsymbol{k} \in \mathbb{N}_0^d} f_{\phi_{\boldsymbol{k}}} \phi_{\boldsymbol{k}}(y), \qquad y \in \Gamma,$$

where the equality comes from expanding f into a polynomial chaos expansion as in (2.8). Suppose that we seek an approximation $\widetilde{f}$ of f of the form

$$\widetilde{f}(y) = \sum_{\phi \in \Pi} \widetilde{f}_\phi \phi(y) \approx f(y), \qquad y \in \Gamma,$$

where Π is an orthonormal multivariate polynomial basis as in (2.11). One should think of $f(y) = \mathcal{I}_\eta u(y)$ with the interpolant $\mathcal{I}_\eta$ from (3.2). We further suppose that the true coefficients $f_{\phi_{\boldsymbol{k}}} = \langle f, \phi_{\boldsymbol{k}} \rangle_\varrho$ are not known and thus have to be approximated. If f is at least continuous, a standard approach is to replace the integral in

$$f_{\phi_{\boldsymbol{k}}} = \langle f, \phi_{\boldsymbol{k}} \rangle_\varrho = \int_\Gamma f(y) \phi_{\boldsymbol{k}}(y) \varrho(y) \mathrm{d}y \tag{3.3}$$

by a quadrature formula, say $\mathcal{Q}_L(f\phi_{\boldsymbol{k}})$, where $\mathcal{Q}_L$ is the sparse grid quadrature operator from (2.29) (and the upper indices p and g are omitted). Unfortunately, this naive approach produces wildly inaccurate approximations for $f(y)$ in most cases. The reason is that the two types of errors involved so far, coming from truncation $\sum_{\boldsymbol{k} \in \mathbb{N}_0^d} \rightsquigarrow \sum_{\phi \in \Pi}$ and quadrature, influence each other. This leads to a phenomenon called *internal aliasing* which is responsible for the poor accuracy. Another consequence of internal aliasing is (2.32). This problem was examined in [23], where also the correct modification was given. See also [22] for a related discussion. We explain this correct modification now.

Let $y^{(1)}, \dots, y^{(\eta)}$ denote the quadrature nodes of $\mathcal{Q}_L$. Let $\boldsymbol{\ell} \in \mathcal{J}_L$ be the multi-index of one fixed full grid contained in the sparse grid (see formula (2.29)), and let $\omega_{\boldsymbol{\ell}}(y^{(j)})$ denote the weight from quadrature rule $\mathcal{Q}_{\boldsymbol{\ell}}$ (see (2.25)) at the node $y^{(j)}$ for $j \in \{1, \dots, \eta\}$. (If $y^{(j)}$ is not a node for $\boldsymbol{\ell}$, then $\omega_{\boldsymbol{\ell}}(y^{(j)})$ is zero.) Then $\mathcal{Q}_L(f\phi)$ can be expressed equivalently as

$$\mathcal{Q}_L(f\phi) = \sum_{\boldsymbol{\ell} \in \mathcal{J}_L} c(\boldsymbol{\ell}) (\mathcal{Q}^{(1)}_{m_1} \otimes \cdots \otimes \mathcal{Q}^{(d)}_{m_d})(f\phi) = \sum_{j=1}^{\eta} \sum_{\boldsymbol{\ell} \in \mathcal{J}_L} c(\boldsymbol{\ell}) \omega_{\boldsymbol{\ell}}(y^{(j)}) \phi(y^{(j)}) f(y^{(j)}),$$

where $c(\boldsymbol{\ell})$ was defined in (2.30). For the set Π, we choose $\Pi = \Pi_L$ defined as in (2.31). According to [23] the correct way to compute the approximations $\widetilde{f}_\phi \approx f_\phi$ for $\phi \in \Pi$ is given by

$$\widetilde{f}_\phi = \sum_{j=1}^{\eta} \sum_{\substack{\boldsymbol{\ell} \in \mathcal{J}_L \\ \phi \in \Pi'_{\boldsymbol{\ell}}}} c(\boldsymbol{\ell}) \omega_{\boldsymbol{\ell}}(y^{(j)}) \phi(y^{(j)}) f(y^{(j)}), \tag{3.4}$$

where the crucial part is that the sum only runs over those $\boldsymbol{\ell}$ where $\phi \in \Pi'_{\boldsymbol{\ell}}$. (The half-exact set $\Pi'_{\boldsymbol{\ell}}$ was defined in (2.27).) This method of computing the polynomial chaos coefficients is called the *sparse pseudospectral approximation method*, see [23]. Since only evaluations of f in $y^{(j)}$ must be available for this computation, we conclude that choosing the interpolation points from $\mathcal{I}_\eta$ as $(y^{(j)})_{j=1}^{\eta}$ gives indeed $\widetilde{\mathcal{I}_\eta u} = \widetilde{u}$. Note that we also have

$$\widetilde{f}(y) = f(y), \qquad f \in \Pi, \tag{3.5}$$

which would not be true for the naive approach due to (2.32). A proof of (3.5) is given in [23, Thm. 2].

It should be noted that the expected value (which, by (2.9), is the first coefficient of the gPCE) can be computed "naively". This was also noticed in [23, Cor. 1] and can be seen as follows. Suppose that $\phi \equiv 1$. Then $\phi \in \Pi'_{\boldsymbol{\ell}}$ for any $\boldsymbol{\ell} \in \mathcal{J}_L$ and thus, by (3.4),

$$\widetilde{f}_\phi = \sum_{j=1}^{\eta} \sum_{\boldsymbol{\ell} \in \mathcal{J}_L} c(\boldsymbol{\ell}) \omega_{\boldsymbol{\ell}}(y^{(j)}) f(y^{(j)}) = \mathcal{Q}_L(f).$$

Together with (3.5), this proves (2.33) from the previous section, too.

The aforementioned procedure of "stochastic collocation + computation of the polynomial chaos coefficients" is sometimes called *non-intrusive spectral projection* (NISP) and some authors (see e.g. [108, Sec. 13.1]) distinguish this method from stochastic collocation methods. The difference is mainly whether one accepts the interpolant as an approximation or computes the polynomial chaos expansion afterwards. We do not care whether it is just a post-processed stochastic collocation method or a method in its own right, but the connections between the two approaches are apparent. The main benefit of having an approximation to the PCE coefficients is that expectations, variances and other statistical quantities can be computed immediately. (One should perhaps recall the formulas (2.9) and (2.10) at this point.)

3.3 The single-level stochastic collocation method (SLSC)

While the first section of this chapter discussed stochastic collocation methods in a general manner, the focus of this section is on the combination of the collocation method in the parameter space and the temporal discretisation of a time-dependent PDE. Parts of this and the subsequent sections are taken from [61].

Consider a time-dependent PDE of the form

$$\partial_t u(t,y) + \mathcal{L}(u(t,y), y) = 0, \qquad t \geq 0,\ y \in \Gamma, \tag{3.6a}$$

$$u(0,y) = u_0(y), \qquad y \in \Gamma \tag{3.6b}$$

on a Banach space $\mathcal{X}$ for given $u_0 \colon \Gamma \to \mathcal{X}$ and $\Gamma = [-1,1]^d$ which admits a solution u (in some sense). For the time being, we do not specify the operator $\mathcal{L}$ as this section only explains the general procedure of single-level stochastic collocation methods.

Furthermore, we assume that we already have a reasonable time-stepping method which successively computes approximations $u_n(y) \approx u(t_n, y)$ to the solution u of (3.6) at times $t_n = n\tau$ for any given $y \in \Gamma$, where $\tau > 0$ is the step-size of the time-stepping method.

The strategy to define a single-level collocation method is rather simple: Choose collocation points from a sparse grid $\mathcal{H}_L^{p,g} = \{y_j\}_{j=1}^{\eta}$ and compute approximations $u_n(y_j) \approx u(t_n, y_j)$ at times $t_n = n\tau$ for

each $j = 1, \dots, \eta$. If we compute the interpolant from the values $u_n(y_1), \dots, u_n(y_\eta)$ for each n, then we obtain an approximation of $u(t_n, y)$ for any $y \in \Gamma$ and $n \in \mathbb{N}$. Afterwards, one may wish to compute a gPCE as explained in Section 3.2.

The step-size τ determines the accuracy of the temporal approximations and the accuracy of the interpolation in y is determined by selecting a coarser/finer grid of interpolation points, which corresponds to a smaller/larger value of η.

Let us be more specific about this approach. We denote the numerical flow of the time-integrator with step-size $\tau > 0$ by Φ_τ, so we set

$$\Phi_\tau^n(u_0, y) = \Phi_\tau(\Phi_\tau^{n-1}(u_0, y), y) = u_n(y), \qquad n \in \mathbb{N}, \qquad \Phi_\tau^0(u_0, y) = u_0(y)$$

for $u_0 \colon \Gamma \to \mathcal{X}$ and $y \in \Gamma$. The second argument $y \in \Gamma$ accounts for the fact that the flow usually depends on the parameter $y \in \Gamma$ since $\mathcal{L}$ from (3.6a) depends on y, too. Moreover, let $\mathcal{I}_L$ denote the depth-L-sparse grid interpolant from (2.15), but with the upper indices p and g omitted in the notation.

In a formula, the stochastic collocation approximation of depth L at time $t_n = n\tau$ is given by

$$u_{L,n}(y) = \mathcal{I}_L \Phi_\tau^n(u_0, y) \approx u(t_n, y), \qquad n \in \mathbb{N}_0,\ y \in \Gamma. \tag{3.7}$$

We distinguish between:

$$\begin{aligned} u(t,\cdot) &\longleftrightarrow \text{solution of (3.6) at time } t \\ \mathcal{I}_L u(t,\cdot) &\longleftrightarrow \text{interpolation of } u(t) \text{ on the depth-}L\text{-grid} \\ u_n = \Phi_\tau^n(u_0,\cdot) &\longleftrightarrow \text{time-discrete approximation after } n \text{ time-steps} \\ u_{L,n} = \mathcal{I}_L \Phi_\tau^n(u_0,\cdot) &\longleftrightarrow \text{approximation after } n \text{ time-steps, interpolated on the depth-}L\text{-grid} \end{aligned}$$

From now on, the variable y indicated with the dot $\cdot$ above will usually be hidden in the notation when it is convenient.

Let us now discuss the *accuracy* of this method by examining the different contributions to the total error. We discuss this for the norm of the space $L^q_\varrho(\Gamma, \mathcal{X})$, $q \in [1, \infty)$, since convergence in this norm corresponds (in probabilistic notion) to *convergence in the q-th mean* due to

$$\|v\|^q_{L^q_\varrho(\Gamma,\mathcal{X})} = \int_\Gamma \|v(y)\|^q_{\mathcal{X}} \varrho(y) \mathrm{d}y = \mathbb{E}[y \mapsto \|v(y)\|^q_{\mathcal{X}}], \qquad v \in L^q_\varrho(\Gamma, \mathcal{X}).$$

Note that Markov's inequality shows that convergence in the q-th mean implies convergence in probability, too. The error after n steps with step-size $\tau > 0$ can be split into

$$\|u(t_n) - u_{L,n}\|_{L^q_\varrho(\Gamma,\mathcal{X})} \le \|u(t_n) - \Phi_\tau^n(u_0)\|_{L^q_\varrho(\Gamma,\mathcal{X})} + \|\Phi_\tau^n(u_0) - \mathcal{I}_L\Phi_\tau^n(u_0)\|_{L^q_\varrho(\Gamma,\mathcal{X})} =: \texttt{(I)} + \texttt{(II)}. \tag{3.8}$$

The first part of the error, `(I)`, corresponds to the time-discretisation error, whereas the second part `(II)` corresponds to the stochastic collocation error, applied to time-discrete approximation $u_n = \Phi_\tau^n(u_0)$. Both error contributions have to be examined separately for each problem under consideration.

The term `(I)` can often be estimated by standard tools from numerical analysis. Examples are given later in this thesis for specific problem classes.

The term `(II)` corresponds to the interpolation error for the sparse grid of depth L, but applied to the *time-discrete* approximation. Error bounds for sparse grid interpolation were given in Section 2.6.2

before. To apply these results, one has to show that the time-discrete solution has enough y-regularity and that the norm of $\Phi_\tau^n(u_0)$ in spaces such as $\{v\colon \Sigma(\boldsymbol{\sigma}) \to \mathcal{X} \mid v \text{ is analytic}\}$ or $C^{\boldsymbol{k}}(\Gamma, \mathcal{X})$ can be somehow related to the corresponding norm of the solution itself, which is hard and surprisingly technical for, e.g., splitting methods. Two of the main results in this thesis (stated in Section 4.5.2 and Section 5.3.2 later) address exactly this issue.

Suppose now that we are able to quantify the error $\|e_{L,n}\|_{L^q_\varrho(\Gamma,\mathcal{X})}$ from (3.8), where $e_{L,n} \coloneqq u(t_n) - u_{L,n}$. Then we can use the relations between different L^q-spaces to obtain some other error estimates. Using $\|\mathbb{E}[e_{L,n}]\|_{\mathcal{X}} \leq \mathbb{E}[\|e_{L,n}\|_{\mathcal{X}}] \leq \|e_{L,n}\|_{L^q_\varrho(\Gamma,\mathcal{X})}$, we get an estimate for the error of the expected value in $\mathcal{X}$. We can also treat a smooth functional $\Upsilon\colon \mathcal{X} \to \mathbb{R}$ of u by the estimate

$$|\mathbb{E}[\Upsilon(u(t_n)) - \Upsilon(u_{L,n})]| \leq \left(\int_0^1 \|\Upsilon'(u_{L,n} + \theta e_{L,n})\|_{L^{q^*}_\varrho(\Gamma,\mathcal{X}^*)} \mathrm{d}\theta \right) \|e_{L,n}\|_{L^q_\varrho(\Gamma,\mathcal{X})},$$

where Υ' denotes the Fréchet derivative of Υ, q^* denotes the Hölder conjugate of q and $\mathcal{X}^*$ the dual space of $\mathcal{X}$. Such an estimate is important if one tries to quantify the error in the computation of $\mathbb{E}[\Upsilon(u)]$ instead of u itself.

Now we are finally ready to discuss the multi-level stochastic collocation method from the thesis title.

3.4 The multi-level stochastic collocation method (MLSC)

In situations where a very accurate approximation of the solution is sought-after or where the regularity of the solution in the parameter space is comparatively low (such that a very fine sparse grid is required), the efficiency of single-level stochastic collocation methods can be improved considerably by a multi-level strategy – at least if certain conditions are met. Such methods have been proposed and analyzed in [109, 115]. For more recent works containing remarkable extensions of the approach we refer to [45, 44, 65]. We will briefly discuss these extensions in Section 3.8 later.

Our presentation of the multi-level strategy follows very closely the work by Teckentrup et al. [109]. In this reference an elliptic problem is considered, and different spatial and stochastic discretisations are combined with a multi-level strategy. In contrast to that, we combine different temporal and stochastic discretisations. Most of the material in this and the next section will appear in a similar form in [61].

Throughout this section, let $T > 0$ be fixed. We choose $N_0 \in \mathbb{N}$, set $\tau_0 = \frac{T}{N_0}$ and define the decreasing sequence of step-sizes $(\tau_j)_{j\in\mathbb{N}_0}$ via $\tau_j = 2^{-j}\tau_0$ for $j \in \mathbb{N}_0$. To each of these step-sizes corresponds a numerical flow Φ_{τ_j} and a number of time-steps $N_j = 2^j N_0$ to reach the given final time T, so $T = \tau_j N_j$ for all $j \in \mathbb{N}_0$. For simplicity, the notation

$$u_{\tau_j} = \Phi_{\tau_j}^{N_j}(u_0)$$

for $j \in \mathbb{N}_0$ is used henceforth. (We will not use the previous notation u_n for $n \in \mathbb{N}$ in the remainder of this chapter.)

The first requirement for the construction of a multi-level method is a convergence result for the temporal discretisation. We assume that convergence is obtained with respect to the $L^2_\varrho(\Gamma,\mathcal{X})$-norm, although one could use $L^q_\varrho(\Gamma,\mathcal{X})$ for any $q \in [1,\infty]$ instead.

Assumption B1. Suppose that there exist constants $\alpha, C_T > 0$ such that

$$\|u(T) - u_{\tau_j}\|_{L^2_\varrho(\Gamma,\mathcal{X})} \leq C_T \tau_j^\alpha \tag{3.9}$$

for all $j \in \mathbb{N}_0$.

Usually, α is the classical order of the time integration scheme. In this thesis, we consider second-order methods and thus we typically have $\alpha = 2$, at least when $\mathcal{X}$ is a spatial L^2-space and the requirements for convergence are met. The value α could be smaller than the classical order if either the solution is not smooth enough or the method suffers from order reduction due to, e.g., stiffness of the problem. The constant C_T will usually depend on some norm of the exact solution $u(t)$ (in general on the whole time interval $[0, T]$) and other problem-dependent quantities, too. The precise dependencies are not crucial because in practice, one determines C_T numerically by regression.

Remark 3.4.1. Convergence results for the temporal discretisation often have a step-size restriction, so the error bound only holds for all step-sizes $0 < \tau \leq \tau_{\max}$ for some $\tau_{\max} > 0$. In particular for non-linear problems this can usually not be avoided. If the restriction is strong and hence $\tau_{\max}$ is already small, then one has to make sure that $\tau_0 \leq \tau_{\max}$ is satisfied. This might prohibit the use of the multi-level approach for some problems with severe step-size restrictions.

Another type of step-size restrictions are CFL conditions arising from the interplay between spatial and temporal discretisation for certain problem classes. Often the maximal temporal step-size is dictated by the smallest mesh width in the spatial mesh (for example if wave equations are solved with explicit time-stepping methods), and thus the straightforward multi-level procedure might not be usable in combination with fine or locally refined meshes. Resolving this issue is a topic of current research, see e.g. the recent work [42].

For the problems considered in this thesis, we only observe rather mild step-size restrictions or none at all as we are working with relatively coarse and regular spatial meshes. ◇

Notation. In the previous sections, the sparse grid interpolation operator was denoted by $\mathcal{I}_L^{p,g}$ or $\mathcal{I}_L$, where the index L is the depth of the sparse grid. In this section, we change the lower index to η, the number of nodes in the corresponding sparse grid and write $\mathcal{I}_\eta$ instead, as the depth parameter is not as relevant as in the previous sections. This number η corresponds to the degrees of freedom of the interpolation operator $\mathcal{I}_\eta$ and it is also the quantity which appears in the interpolation estimates (2.22) and (2.24). This is the reason why η plays a more important role in the following. One should be aware of the fact that for sparse grids, it is not possible (at least not with the procedure described in Section 2.6) to define a sparse grid interpolation operator for *arbitrary* $\eta \in \mathbb{N}$. This issue will be addressed after Example 3.5.3.

Let $(\eta_\ell)_{\ell \in \mathbb{N}_0}$ be increasing (not necessarily strictly increasing) and consider an abstract sequence $(\mathcal{I}_{\eta_\ell})_{\ell \in \mathbb{N}_0}$ of interpolation operators $\mathcal{I}_{\eta_\ell} \colon C(\Gamma, \mathcal{X}) \to L^2_\varrho(\Gamma, \mathcal{X})$. We think of η_ℓ as being the number of interpolation points for $\mathcal{I}_{\eta_\ell}$. Since $(\eta_\ell)_{\ell \in \mathbb{N}_0}$ is increasing, it is reasonable that $\mathcal{I}_{\eta_{\ell+1}}$ is more accurate than $\mathcal{I}_{\eta_\ell}$ if $\eta_{\ell+1} > \eta_\ell$ for $\ell \in \mathbb{N}_0$.

We make the following assumption for the interpolation error.

Assumption B2. There exist constants $C_I, C_\zeta, \beta > 0$ and a Banach space $\Lambda(\Gamma, \mathcal{X}) \hookrightarrow L^2_\varrho(\Gamma, \mathcal{X})$ such that $\{u_{\tau_j} = \Phi_{\tau_j}^{N_j}(u_0) \colon j \in \mathbb{N}_0\}$ is contained in $\Lambda(\Gamma, \mathcal{X})$ and that for all $v \in \Lambda(\Gamma, \mathcal{X})$, it holds

$$\|v - \mathcal{I}_{\eta_\ell} v\|_{L^2_\varrho(\Gamma, \mathcal{X})} \leq C_I \kappa_\ell \zeta(v) \tag{3.10}$$

for some decreasing sequence $(\kappa_\ell)_{\ell\in\mathbb{N}_0}$ and a function $\zeta\colon \Lambda(\Gamma,\mathcal{X})\to\mathbb{R}$ that satisfies

$$\zeta(u_{\tau_j})\le C_\zeta\tau_0^\beta \qquad \text{and} \qquad \zeta(u_{\tau_{j+1}}-u_{\tau_j})\le C_\zeta\tau_{j+1}^\beta \tag{3.11}$$

for all $j\in\mathbb{N}_0$.

It is allowed that $\eta_\ell=\eta_{\ell+1}$ and $\mathcal{I}_{\eta_\ell}=\mathcal{I}_{\eta_{\ell+1}}$ for some $\ell\in\mathbb{N}_0$ and hence also $\kappa_\ell=\kappa_{\ell+1}$. In light of the Theorems 2.6.2 and 2.6.6, the space $\Lambda(\Gamma,\mathcal{X})$ will be either

- a space of analytic functions $\Sigma(\boldsymbol{\sigma})\to\mathcal{X}$ on a polyellipse $\Sigma(\boldsymbol{\sigma})$ as in (2.19) or
- the space $C^{\boldsymbol{k}}(\Gamma,\mathcal{X})$ with $\boldsymbol{k}=(k,\dots,k)\in\mathbb{N}_0^d$.

Therefore the functions u_{τ_j} and $u_{\tau_{j+1}}-u_{\tau_j}$ need a certain amount of regularity in the stochastic variable in order to fulfill Assumption B2. The sequence $(\kappa_\ell)_{\ell\in\mathbb{N}_0}$ is determined by the convergence rate of the interpolation error, so $\kappa_\ell=\eta_\ell^{-\mu}$ for some $\mu>0$ in the analytic case by Theorem 2.6.2. The function ζ is usually some norm which appears on the "right-hand side" of an interpolation error bound. In the analytic case, $\zeta(v)=\max_{z\in\Sigma(\boldsymbol{\sigma})}\|v(z)\|_{\mathcal{X}}$.

For many time integration schemes, it is not easy to verify Assumption B2. This is especially true for splitting methods (as we will see in Chapter 4 and Chapter 5).

Remark 3.4.2. Remark 3.4.1 concerning step-size restrictions also applies to the estimate

$$\zeta(u_{\tau_{j+1}}-u_{\tau_j})\le C_\zeta\tau_{j+1}^\beta,$$

so again one has to make sure that τ_0 is sufficiently small. ◇

After these preparations we are in the position to formulate the *multi-level stochastic collocation* (MLSC) method. We choose $L\in\mathbb{N}_0$, set $u_{\tau_{-1}}=0$ and start with the telescoping sum

$$u_{\tau_J}=\sum_{j=0}^{J}(u_{\tau_j}-u_{\tau_{j-1}}), \tag{3.12}$$

where we have used the notation

$$u_{\tau_j}=\Phi_{\tau_j}^{N_j}(u_0),$$

such that all u_{τ_j}, $j\in\mathbb{N}_0$, are approximations of the solution u at the same time $T=N_j\tau_j$.

Only an interpolation of u_{τ_J} can be computed in practice, and in principle, one could simply interpolate every difference under the sum with the same interpolation operator. In order to reach a given accuracy, however, it is much more efficient to balance the two errors caused by time integration and interpolation in a near-optimal way. If j increases, then Assumption B2 implies that $\zeta(u_{\tau_j}-u_{\tau_{j-1}})$ decreases and $u_{\tau_j}-u_{\tau_{j-1}}$ can thus be interpolated with a coarser (but cheaper) interpolation operator. Conversely, a more accurate interpolation can be used for the summands with small j, for which the time integration is less costly.

Thus we define the multi-level approximation $u_J^{(\mathrm{ML})}$ by

$$u_J^{(\mathrm{ML})}=\sum_{j=0}^{J}\mathcal{I}_{\eta_{J-j}}[u_{\tau_j}-u_{\tau_{j-1}}]. \tag{3.13}$$

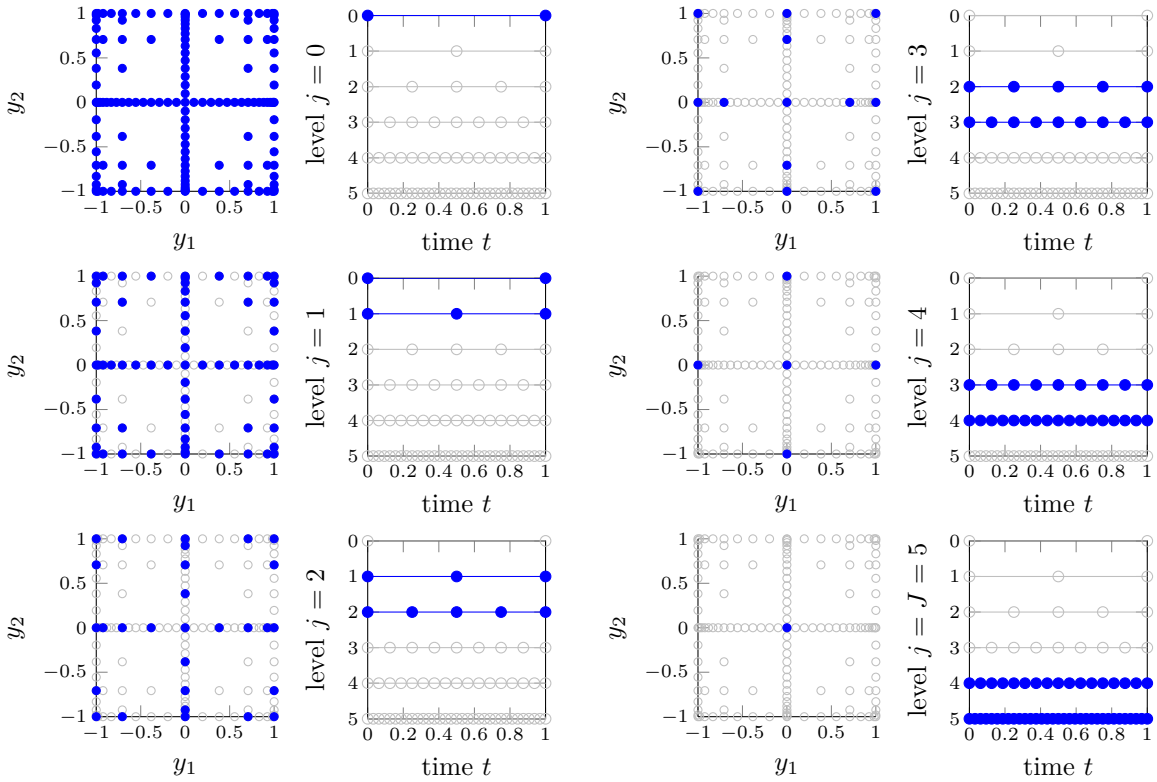


Figure 3.1: Combination of stochastic and temporal discretisation in the multi-level estimator

The most accurate interpolation operator $\mathcal{I}_{\eta_J}$ is used for the coarsest temporal approximation u_{τ_0} and the least accurate interpolation operator $\mathcal{I}_{\eta_0}$ is used for the difference of the two finest temporal approximations $u_{\tau_J} - u_{\tau_{J-1}}$, see Figure 3.1.

Note that for this combination of interpolation operators and temporal approximations, the two indices j and ℓ from before are reduced to only one index j, to which the other index ℓ will be tuned. The correct tuning of coarse interpolation / fine temporal resolution (and vice versa) will be achieved by selecting a special sequence $(\eta_j)_{j\in\mathbb{N}_0}$ later on. The refinement of the temporal discretisations has already been fixed by the choice $\tau_j = 2^{-j}\tau_0$ for $j \in \mathbb{N}_0$.

Let us now examine the convergence of the multi-level approximation $u_J^{(\mathrm{ML})}$ in the norm of $L^2_\varrho(\Gamma, \mathcal{X})$. We start with a triangle inequality similar to (3.8), i.e.

$$\|u(T) - u_J^{(\mathrm{ML})}\|_{L^2_\varrho(\Gamma,\mathcal{X})} \le \|u(T) - u_{\tau_J}\|_{L^2_\varrho(\Gamma,\mathcal{X})} + \|u_{\tau_J} - u_J^{(\mathrm{ML})}\|_{L^2_\varrho(\Gamma,\mathcal{X})} =: (\mathtt{I}) + (\mathtt{II}). \tag{3.14}$$

We show that for a suitable choice of $(\eta_j)_{j\in\mathbb{N}_0}$, the error components (I) and (II) converge at the same rate. Assumption B1 implies that there exist $\alpha, C_T > 0$ such that

$$(\mathtt{I}) \le C_T \tau_J^\alpha.$$

The second term (II) can be estimated by the triangle inequality as

$$(\mathtt{II}) \le \sum_{j=0}^{J} \left\| (u_{\tau_j} - u_{\tau_{j-1}}) - \mathcal{I}_{\eta_{J-j}}(u_{\tau_j} - u_{\tau_{j-1}}) \right\|_{L^2_\varrho(\Gamma,\mathcal{X})} \le \sum_{j=0}^{J} C_I C_\zeta \kappa_{J-j} \tau_j^\beta$$

due to (3.12), (3.13) and Assumption B2. Choosing a sequence $(\eta_j)_{j\in\mathbb{N}_0}$ such that the corresponding values κ_{J-j} in (3.10) satisfy

$$\kappa_{J-j} \leq C_T\big((J+1)C_I C_\zeta\big)^{-1}\tau_J^\alpha \tau_j^{-\beta} \tag{3.15}$$

yields

$$(\mathtt{II}) \leq \sum_{j=0}^{J} C_T\big((J+1)C_I C_\zeta\big)^{-1}\tau_J^\alpha \tau_j^{-\beta} C_I C_\zeta \tau_j^\beta = C_T \tau_J^\alpha,$$

such that the error contribution from (II) and (I) is almost the same. With this choice of $(\eta_j)_{j=0}^J$, we obtain

$$\|u(T) - u_J^{(\mathrm{ML})}\|_{L^2_\varrho(\Gamma,\mathcal{X})} \leq 2C_T\tau_J^\alpha, \tag{3.16}$$

which means that the multi-level approximation converges as $J \to \infty$.

Remark 3.4.3. In the original work [109], the Assumptions B1 and B2 were formulated with u_h instead of u_τ, where u_h is the finite element approximation of a solution to an elliptic PDE on a mesh with width h. In contrast to our goal of approximating $u(T)$ for a fixed time T, their goal was the computation of an approximation on the whole spatial domain. ◇

Next we examine the computational cost of $u_J^{(\mathrm{ML})}$ in detail. Clearly, we want to keep the cost as low as possible while still satisfying the inequality (3.15) which ensures convergence of $u_J^{(\mathrm{ML})}$ as $J \to \infty$.

3.5 Cost analysis of the multi-level method

Here we analyse the ε-cost of the MLSC estimator from (3.13), which will be defined as the computational cost required to achieve a desired accuracy ε. The analysis relies on the convergence rates from Assumptions B1 and B2. From now on, we add the following restriction to Assumption B2: The value κ_ℓ is related to η_ℓ by $\kappa_\ell = \eta_\ell^{-\mu}$ for some $\mu > 0$.

Remark 3.5.1. The relation $\kappa_\ell = \eta_\ell^{-\mu}$ corresponds to the interpolation error for *analytic* functions, as stated in Theorem 2.6.2.

For functions of finite regularity, the error estimates usually contain a logarithmic factor $\log(\eta_\ell)^E$, where the exponent E depends on the regularity of the function and the dimension d, see e.g. (2.22) or (2.24). To fit this case into our current setting, note that $\log(\eta_\ell)^E \leq C\eta_\ell$ for a constant C which again depends on the regularity and on d, but is independent of η_ℓ. The constant C is given explicitly in Lemma A.1. If we use this for $E = (k+2)(d-1)+1$, which corresponds to the interpolation error bound (2.24) for functions belonging to $C^{\boldsymbol{k}}(\Gamma,\mathcal{X})$ with $\boldsymbol{k} = (k,\dots,k) \in \mathbb{N}_0^d$, we get

$$\kappa_\ell = \eta_\ell^{-k}(\log(\eta_\ell))^{(k+2)(d-1)+1} \leq C\eta_\ell^{-(k-1)}$$

and conclude that the choice $\kappa_\ell = \eta_\ell^{-\widetilde{\mu}}$ with $\widetilde{\mu} = k-1$ can be used in this case, too.

The question occurs whether it would be possible to incorporate the logarithmic factor into the construction of the multi-level method. To the best of our knowledge, this is not possible to an extent where it would be of any practical value. More specifically, an explicit formula for η_j such as (3.18) given later could not be derived by the same strategy if the logarithmic factor was present. ◇

Notation. We write $a \lesssim b$ if and only if $a \leq Cb$ for some constant C which is independent of the step-sizes $(\tau_j)_{j\in\mathbb{N}_0}$, the numbers of interpolation points $(\eta_\ell)_{\ell\in\mathbb{N}_0}$ and the accuracy ε. Similarly, we write $a \eqsim b$ if and only if $a = Cb$ for some constant C with the same properties.

We denote the cost of "evaluating" $u_{\tau_j} - u_{\tau_{j-1}}$ at a sample y by C_j and assume the following.

Assumption B3. The cost C_j satisfies $C_j \lesssim \tau_j^{-1}$ for all $j \in \mathbb{N}_0$.

For time-stepping schemes, the number of time-steps is usually proportional to the inverse of the step-size. Thus, the above assumption says that the cost is proportional to the number of time-steps. This is often reasonable, but it neglects computational work in the setup. For the parabolic problems consider later in Chapter 4, matrix assembly is part of the setup process and thus not neglectable from the total work if only few time-steps are computed. In such cases it requires an intelligent strategy of reusing already assembled matrices for collocation points contained in sparse grids of lower depth to avoid computational overhead for the multi-level estimator.

Now we define the *total computational cost* of the MLSC approximation (3.13) as

$$C^{(\mathrm{ML})} = \sum_{j=0}^{J} \eta_{J-j} C_j.$$

This definition is reasonable if the interpolation operator $\mathcal{I}_{\eta_{J-j}}$ is based on η_{J-j} points. The following result is the main result of this subsection. It quantifies the cost which is needed to achieve an accuracy of ε with the MLSC approximation.

Theorem 3.5.2 (Multi-level ε-cost theorem)**.**
Suppose that Assumptions B1 – B3 hold with $\kappa_\ell = \eta_\ell^{-\mu}$ for some $\mu > 0$ and assume that $\alpha \geq \min\{\beta, \mu\}$. Then, for given $\varepsilon < \mathrm{e}^{-1}$, there exists $J \in \mathbb{N}_0$ and a sequence $(\eta_j)_{j=0}^J$ of real numbers such that

$$\|u(T) - u_J^{(\mathrm{ML})}\|_{L_\varrho^2(\Gamma,\mathcal{X})} \leq \varepsilon \tag{3.17}$$

and simultaneously

$$C^{(\mathrm{ML})} \lesssim \begin{cases} \varepsilon^{-\frac{1}{\mu}}, & \beta > \mu, \\ \varepsilon^{-\frac{1}{\mu}} |\log(\varepsilon)|^{1+\frac{1}{\mu}}, & \beta = \mu, \\ \varepsilon^{-\frac{1}{\mu} - \frac{\mu-\beta}{\alpha\mu}}, & \beta < \mu. \end{cases}$$

The sequence $(\eta_j)_{j=0}^J$ is given by

$$\eta_{J-j} = (2C_I C_\zeta \tau_0^\beta S_J)^{1/\mu} \varepsilon^{-1/\mu} 2^{-\frac{j(\beta+1)}{\mu+1}}, \qquad j = 0, \dots, J, \tag{3.18}$$

where

$$S_J = \sum_{j=0}^{J} 2^{-j\frac{\beta-\mu}{\mu+1}}.$$

Proof. Although a complete proof of this result was given in [109, Thm. 4.2], we repeat it here since the result is crucial to this thesis.

Consider the error from (3.14). To achieve $(\mathtt{I}) \leq \frac{\varepsilon}{2}$, we need

$$\tau_J \leq \left(\frac{\varepsilon}{2C_T}\right)^{1/\alpha}$$

by Assumption B1 and thus

$$J = \left\lceil \frac{1}{\alpha} \log_2 \left(\frac{2C_T}{\varepsilon} \right) + \log_2(\tau_0) \right\rceil \tag{3.19}$$

is sufficient. Our goal is to minimise the computational cost $C^{(\mathrm{ML})}$ subject to the requirement $(\mathtt{II}) \leq \frac{\varepsilon}{2}$. Thus we define the Lagrange function $\mathcal{L}$ as

$$\mathcal{L}(\eta_0, \dots, \eta_J, \lambda) = \tau_0^{-1} \sum_{j=0}^{J} \eta_{J-j} 2^j + \lambda \left(\sum_{j=0}^{J} C_I C_\zeta \eta_{J-j}^{-\mu} 2^{-j\beta} \tau_0^\beta - \frac{\varepsilon}{2} \right).$$

The Lagrange multiplier method suggests to look for critical points of $\mathcal{L}$. Therefore, we try to solve the equations

$$\frac{\partial \mathcal{L}}{\partial \eta_{J-j}} = 2^j \tau_0^{-1} - \lambda C_I C_\zeta \mu \eta_{J-j}^{-(\mu+1)} 2^{-j\beta} \tau_0^\beta = 0, \qquad j = 0, \dots, J, \tag{3.20a}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = \sum_{j=0}^{J} C_I C_\zeta \eta_{J-j}^{-\mu} 2^{-j\beta} \tau_0^\beta - \frac{\varepsilon}{2} = 0. \tag{3.20b}$$

The first $J+1$ equations (3.20a) can be rearranged as

$$\eta_{J-j} = (C_I C_\zeta \mu \lambda \tau_0^{\beta+1})^{1/(\mu+1)} 2^{-\frac{j(\beta+1)}{\mu+1}}, \qquad j = 0, \dots, J, \tag{3.21}$$

which in turn implies

$$\lambda^{\frac{1}{\mu+1}} = 2^{1/\mu} (C_I C_\zeta)^{\frac{1}{(\mu+1)\mu}} \tau_0^{\frac{\beta-\mu}{\mu(\mu+1)}} \mu^{-\frac{1}{\mu+1}} \varepsilon^{-1/\mu} S_J^{1/\mu}$$

via (3.20b). Inserting this into (3.21), we arrive at

$$\eta_{J-j} = (2 C_I C_\zeta \tau_0^\beta S_J)^{1/\mu} \varepsilon^{-1/\mu} 2^{-\frac{j(\beta+1)}{\mu+1}}.$$

So far, we only know how to choose J and η_{J-j} for $j = 0, \dots, J$. We still have to verify that the cost $C^{(\mathrm{ML})}$ scales as claimed. This is the next step. We have

$$\begin{aligned} C^{(\mathrm{ML})} &= \sum_{j=0}^{J} \eta_{J-j} C_j \eqsim \sum_{j=0}^{J} \eta_{J-j} 2^j \lesssim \sum_{j=0}^{J} \varepsilon^{-1/\mu} S_J^{1/\mu} 2^{-j\frac{\beta-\mu}{\mu+1}} \\ &\eqsim \varepsilon^{-1/\mu} S_J^{1+1/\mu}. \end{aligned}$$

Now we distinguish three cases.

The case $\beta > \mu$. Here, S_J converges as $J \to \infty$ and thus $C^{(\mathrm{ML})} \lesssim \varepsilon^{-1/\mu}$.

The case $\beta = \mu$. Here we have $S_J = J + 1$ and thus

$$C^{(\mathrm{ML})} \lesssim \varepsilon^{-1/\mu} (J+1)^{1+1/\mu}.$$

Using $\varepsilon^{-1/\alpha} \leq \varepsilon^{-1/\mu}$ for $\varepsilon \leq 1$ and

$$(J+1)^{1+1/\mu} < \left(\frac{1}{\alpha} \log_2(2C_T/\varepsilon) + \log_2(\tau_0) + 2 \right)^{1+1/\mu} \lesssim \left(-\frac{1}{\alpha} \log_2(\varepsilon) \right)^{1+1/\mu} = \left(\frac{1}{\alpha} |\log_2(\varepsilon)| \right)^{1+1/\mu}$$

for $\varepsilon < 1/\mathrm{e}$, we arrive at $C^{(\mathrm{ML})} \lesssim \varepsilon^{-1/\mu} |\log_2(\varepsilon)|^{1+1/\mu}$ as claimed.

The case $\beta < \mu$. Here we have

$$S_J = \sum_{j=0}^{J} 2^{(j-J)\frac{\beta-\mu}{\mu+1}} = 2^{-J\frac{\beta-\mu}{\mu+1}} \sum_{j=0}^{J} 2^{-j\frac{\mu-\beta}{\mu+1}} \lesssim 2^{J\frac{\mu-\beta}{\mu+1}}.$$

Inserting the value of J from (3.19) gives

$$S_J^{1+1/\mu} \lesssim \varepsilon^{-\frac{\mu-\beta}{\alpha\mu}}$$

and we finally obtain

$$C^{(\mathrm{ML})} \lesssim \varepsilon^{-\frac{1}{\mu}-\frac{\mu-\beta}{\alpha\mu}}.$$

This concludes the proof of the theorem. □

Example 3.5.3.

For $\alpha = \beta$ (which will be the situation in our model problems later on), the requirement $\alpha \geq \min\{\beta, \mu\}$ is clearly satisfied. Hence, Theorem 3.5.2 implies that $\|u(T) - u_J^{(\mathrm{ML})}\|_{L^2_\varrho(\Gamma,\mathcal{X})} \leq \varepsilon$ can be achieved with

$$C^{(\mathrm{ML})} \lesssim \begin{cases} \varepsilon^{-\frac{1}{\mu}}, & \mu < 2, \\ \varepsilon^{-\frac{1}{\mu}} |\log(\varepsilon)|^{1+\frac{1}{\mu}}, & \mu = 2, \\ \varepsilon^{-\frac{1}{\alpha}}, & \mu > 2. \end{cases}$$

The optimal choice for η_{J-j} specified by (3.18) gives in general not an integer. In practice, however, the interpolation operators $\mathcal{I}_m$ are only available for certain integer depths $L \in \mathbb{N}$ corresponding to $m = m_L$, the number of points in a sparse grid. To determine a practicable family $(\widetilde{\eta}_j)_{j=0}^J$ as a replacement for $(\eta_j)_{j=0}^J$, one has to choose a number of the form $m_{\ell(j)}$, $\ell(j) \in \mathbb{N}$, for which an interpolation operator (and hence an associated sparse grid) is available. An obvious choice is

$$\widetilde{\eta}_j = \eta_j^{\mathrm{up}} = \min\{m_\ell \colon \ell \in \mathbb{N},\ \eta_j \leq m_\ell\}, \qquad j = 0, \dots, J. \tag{3.22}$$

We indicate this *rounding strategy* by "up" since η_j is rounded up to the next admissible integer. One should be aware of the fact that such a replacement sequence may not lead precisely to the cost estimate from Theorem 3.5.2.

Remark 3.5.4. The sequence $(m_\ell)_{\ell\in\mathbb{N}}$ usually grows exponentially in case of nested point sequences. Hence, η_j^{up} might be up to twice as large as η_j in some cases, which could be crucial in large stochastic dimensions d or in case that very accurate solutions (and hence large values of η_J) are required. This could heavily influence the cost scaling of the MLSC method. This is the main reason why other rounding strategies are necessary, as explained in the following. ◇

Similarly, we define the rounding strategy "down" by always choosing the next sparse grid with less points than η_j,

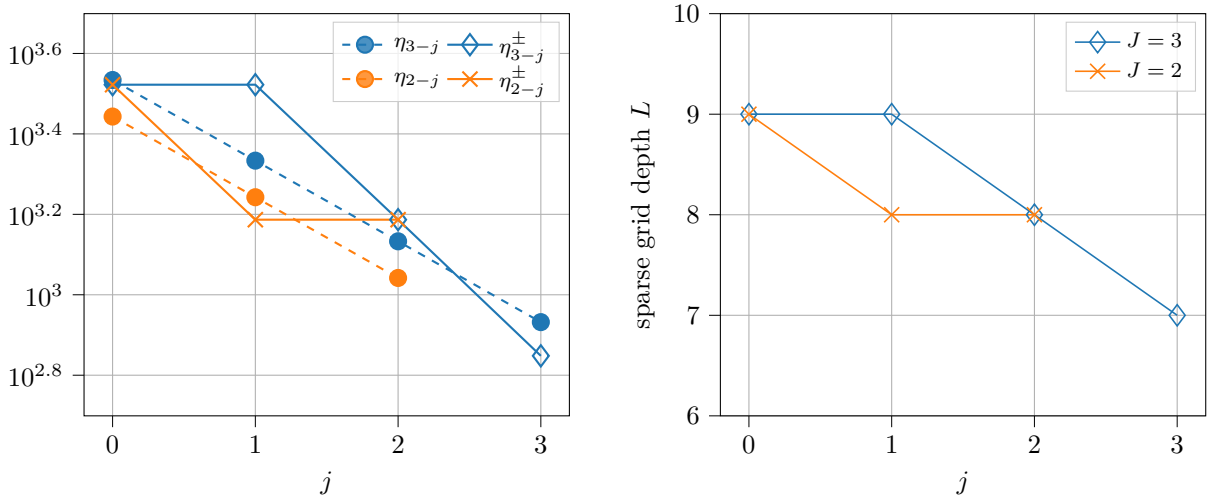
$$\eta_j^{\mathrm{down}} = \max\{1, \max\{m_\ell \colon \ell \in \mathbb{N},\ m_\ell \leq \eta_j\}\}, \qquad j = 0, \dots, J.$$

This gives a cheaper estimator, but it is more likely that an error below ε is not achieved.

Another rounding strategy $(\eta_j^{\pm})_{j=0}^J$ for $(\eta_j)_{j=0}^J$ named "up/down" was discussed in [109, Rem. 6.1 and 6.3]. This strategy can be described as follows: At first, all η_j are rounded to the nearest number

for which a sparse grid is available. If more η_j are rounded down than up in this procedure, we choose the number which was rounded down by the largest amount and instead round it up. This is iterated until equally many η_j are rounded up and down (or one more number is rounded up than down). If more η_j are rounded up than down in the beginning, then the procedure is analogous. This way, we arrive at the sequence $(\eta_j^{\pm})_{j=0}^{J}$. A visualisation of the "up/down" rounding strategy (in a scenario with fictional constants, rates and ε) is given in Figure 3.2a for different values of J. The corresponding sparse grid depths are depicted in Figure 3.2b.

Although the rounding strategy "up/down" is cheaper than the strategy "up", it can still suffer from the issue described in Remark 3.5.4, but it is less pronounced for "up/down".



(a) η_{J-j} and $\eta_{J-j}^{\pm}$ for $J = 2$ (orange) and $J = 3$ (blue)

(b) Sparse grid levels corresponding to $\eta_{J-j}^{\pm}$

Figure 3.2: "Up/down" rounding strategy

3.5.1 Comparison with single-level collocation methods

Given a set of collocation points $y_1, \dots, y_\eta \in \Gamma = [-1, 1]^d$ from a sparse grid and a step-size $\tau = \frac{T}{N}$ for $N \in \mathbb{N}$, we denote the single-level approximation from (3.7) by

$$u_{\eta,\tau}^{(\mathrm{SL})} := \mathcal{I}_\eta \Phi_\tau^N(u_0),$$

where $\mathcal{I}_\eta$ is the interpolation operator corresponding to $\{y_j\}_{j=1}^{\eta}$. (As the final time T is fixed as in the previous section, the step-size determines the integer $N \in \mathbb{N}$.) The upper index "(SL)" stands for single-level, referring to the fact that only a single point set $\{y_j\}_{j=1}^{\eta}$ is used to compute $u_{\eta,\tau}^{(\mathrm{SL})}$.

Under the same assumptions as in Theorem 3.5.2, the error of the *single-level* collocation method with $\eta = \eta_\ell \in \mathbb{N}$ and $\tau = \tau_j$ can be bounded by

$$\begin{aligned}
\|u(T) - u_{\eta,\tau}^{(\mathrm{SL})}\|_{L_\varrho^2(\Gamma,\mathcal{X})} &\leq \|u(T) - u_\tau\|_{L_\varrho^2(\Gamma,\mathcal{X})} + \|u_\tau - \mathcal{I}_\eta u_\tau\|_{L_\varrho^2(\Gamma,\mathcal{X})} \\
&\leq C_T \tau^\alpha + C_I \zeta(u_\tau) \eta^{-\mu} \\
&\leq C_T \tau^\alpha + C_I C_\zeta \tau_0^\beta \eta^{-\mu}.
\end{aligned}$$

To make both contributions equal to $\varepsilon/2$ (or ε, since we ignore constants anyway), choose η and τ such that $\eta \eqsim \varepsilon^{-\frac{1}{\mu}}$ and $\tau \eqsim \varepsilon^{\frac{1}{\alpha}}$. The computational cost for the single-level method to achieve a total error less or equal than ε is thus given by

$$C_\varepsilon^{(\mathrm{SL})} \eqsim \frac{\eta}{\tau} \eqsim \varepsilon^{-\frac{1}{\mu}-\frac{1}{\alpha}}. \tag{3.23}$$

Now we compare this with the result for $C_\varepsilon^{(\mathrm{ML})}$ from Theorem 3.5.2. To this end, we define the *cost reduction* of the multi-level approach compared to the single-level approach as $C_\varepsilon^{(\mathrm{ML})}/C_\varepsilon^{(\mathrm{SL})}$. The lower the quotient $C_\varepsilon^{(\mathrm{ML})}/C_\varepsilon^{(\mathrm{SL})}$ is, the better is the performance of the multi-level method. (The inverse of the cost reduction thus corresponds to the *speed-up* of the multi-level approach.) By Theorem 3.5.2 and (3.23), we have

$$\frac{C_\varepsilon^{(\mathrm{ML})}}{C_\varepsilon^{(\mathrm{SL})}} \eqsim \begin{cases} \varepsilon^{\frac{1}{\alpha}}, & \beta > \mu, \\ \varepsilon^{\frac{1}{\alpha}} |\log(\varepsilon)|^{1+\frac{1}{\mu}}, & \beta = \mu, \\ \varepsilon^{\frac{\beta}{\alpha\mu}}, & \beta < \mu. \end{cases}$$

Note that the cost savings discussed above ignore constants which appear in $C_\varepsilon^{(\mathrm{ML})}$ and $C_\varepsilon^{(\mathrm{SL})}$, so only the decay rate in ε and μ is meaningful. We observe that the cost reduction tends to zero for $\varepsilon \to 0$ for every constellation of β and μ, so the multi-level approach is beneficial whenever the tolerance ε is sufficiently small. For large tolerances, the benefits of using various levels disappear. As the constant in the cost reduction is not available, it is not immediately clear for which particular tolerances one of the methods is better than the other. This has to be examined in practical situations.

Example 3.5.5.

We return to the case $\alpha = \beta$ from Example 3.5.3. Here, the cost reductions are $\varepsilon^{\frac{1}{\alpha}} = \varepsilon^{\frac{1}{2}}$ for $\mu < 2$ ("low regularity"), $\varepsilon^{\frac{1}{2}} |\log(\varepsilon)|^{\frac{3}{2}}$ for $\mu = 2$ and $\varepsilon^{\frac{\beta}{\alpha\mu}} = \varepsilon^{\frac{1}{\mu}}$ for $\mu > 2$ ("high regularity"). Clearly, the savings are most noticeable if either the rate μ or the tolerance ε is small. In both cases, more levels are required in total for the multi-level estimator and thus the profit of using various levels increases. Figure 3.3 below gives a picture of this situation (dark is best, bright means "no savings").

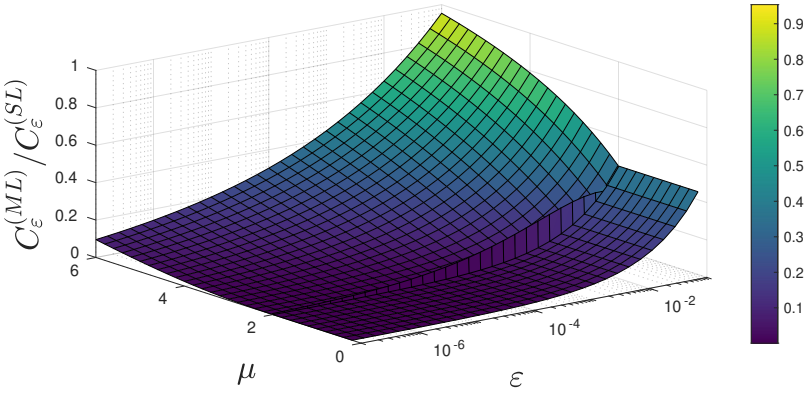


Figure 3.3: Savings of the multi-level approach in dependency of ε and μ

3.5.2 Practical considerations

It should be noted that Theorem 3.5.2 itself is not directly helpful in practice yet. The open questions which have to addressed for an implementation of the method are the following.

(1) How to estimate the constant C_T and the rate α from Assumption B1?

(2) How to estimate $C_I C_\zeta$, β and μ in $\kappa_\ell = \eta_\ell^{-\mu}$ from Assumption B2?

(3) How to find the correct value of J from Theorem 3.5.2?

All questions have already been answered satisfactorily in [109, Sec. 6.3], but we repeat the most important steps here.

Ad (1). If a sufficiently accurate reference solution $u_{\mathrm{ref}}(T) \approx u(T)$ is available, we can use it as a replacement for $u(T)$ in (3.9) and use least-squares fitting for the errors obtained with approximations u_{τ_1} and u_{τ_2} for two different step-sizes τ_1 and τ_2. This would give values for C_T and α. If such a reference solution is not available (which is the practically relevant situation), then we can use yet another approximation u_{τ_3} with $\tau_3 < \min\{\tau_1, \tau_2\}$ as a replacement for $u(T)$ in (3.9). For $j = 1, 2$, we have

$$\begin{aligned} \|u_{\tau_3} - u_{\tau_j}\|_{L^2_\varrho(\Gamma,\mathcal{X})} &\le \|u_{\tau_3} - u(T)\|_{L^2_\varrho(\Gamma,\mathcal{X})} + \|u(T) - u_{\tau_j}\|_{L^2_\varrho(\Gamma,\mathcal{X})} \\ &\le C_T((\tau_3/\tau_j)^\alpha + 1)\tau_j^\alpha \qquad (3.24) \\ &\le 2C_T\tau_j^\alpha, \qquad (3.25) \end{aligned}$$

from which we can reconstruct the values of C_T and α very easily if we assume that equality holds here.

More precisely, we fit the data

$$(\tau_1, \|u_{\tau_3} - u_{\tau_1}\|_{L^2_\varrho(\Gamma,\mathcal{X})}), \quad (\tau_2, \|u_{\tau_3} - u_{\tau_2}\|_{L^2_\varrho(\Gamma,\mathcal{X})})$$

to the model function

$$f(x) = \hat{C}x^{\hat{\alpha}} \qquad \text{or} \qquad \log(f(x)) = \log(\hat{C}) + \hat{\alpha}\log(x)$$

to determine $\hat{C}$ and $\hat{\alpha}$. Then we take $\hat{C}$ and $\hat{\alpha}$ as replacements for C_T and α. This is a linear regression for the function $\log(f(x))$ with respect to $\log(x)$.

In practice, we can only compute $\|\mathcal{I}_\eta u_{\tau_3} - \mathcal{I}_\eta u_{\tau_j}\|_{L^2_\varrho(\Gamma,\mathcal{X})}$ for an interpolant $\mathcal{I}_\eta$. The choice of η is usually not crucial and yields almost identical values for the constant C_T and rate α as long as step-sizes τ_j, $j = 1, 2, 3$, are used for which the asymptotic behaviour already shows up.

Ad (2). Here the procedure is similar. Additionally to the approximations with different step-sizes, we also need different interpolation levels η_0, η_1 and η_2 to estimate $C_I C_\zeta$, β and μ from Assumption B2 (with $\kappa_\ell = \eta_\ell^{-\mu}$). If we assume that Assumption B2 is satisfied with $\kappa_\ell = \eta_\ell^{-\mu}$ and the value of β is known (which is the case in all of our applications later), then we can proceed as follows to estimate the product $C = C_I C_\zeta$ and μ. By Assumption B2, we have

$$\|(u_{\tau_{k+1}} - u_{\tau_k}) - \mathcal{I}_{\eta_\ell}(u_{\tau_{k+1}} - u_{\tau_k})\|_{L^2_\varrho(\Gamma,\mathcal{X})} \le C_I C_\zeta \eta_\ell^{-\mu}\tau_{k+1}^\beta$$

for $k = -1, \dots, j$ (with $u_{\tau_{-1}} = 0$). This implies

$$\begin{aligned}
\|u_{\tau_{j+1}} - \mathcal{I}_{\eta_\ell} u_{\tau_{j+1}}\|_{L^2_\varrho(\Gamma,\mathcal{X})} &\le \left\| \sum_{k=-1}^{j} \big(u_{\tau_{k+1}} - u_{\tau_k}\big) - \sum_{k=-1}^{\ell} \mathcal{I}_{\eta_\ell}(u_{\tau_{k+1}} - u_{\tau_k}) \right\|_{L^2_\varrho(\Gamma,\mathcal{X})} \\
&\le \sum_{k=-1}^{j} \big\|(u_{\tau_{k+1}} - u_{\tau_k}) - \mathcal{I}_{\eta_\ell}(u_{\tau_{k+1}} - u_{\tau_k})\big\|_{L^2_\varrho(\Gamma,\mathcal{X})} \\
&\le \sum_{k=-1}^{j} C_I C_\zeta \eta_\ell^{-\mu} \tau_{k+1}^{\beta} = C_I C_\zeta \eta_\ell^{-\mu} \tau_0^{\beta} \sum_{k=0}^{j+1} 2^{-k\beta} \\
&= C_I C_\zeta \eta_\ell^{-\mu} \tau_0^{\beta} \frac{1 - 2^{-(j+2)\beta}}{1 - 2^{-\beta}} \le C_I C_\zeta \eta_\ell^{-\mu} \tau_0^{\beta} \frac{1}{1 - 2^{-\beta}}.
\end{aligned} \tag{3.26}$$

Now take $\eta_0 < \eta_1 < \eta_2$, step-sizes τ_1, τ_2 and determine $\hat{C}$ and μ such that

$$\|\mathcal{I}_{\eta_2} u_{\tau_{j+1}} - \mathcal{I}_{\eta_\ell} u_{\tau_{j+1}}\|_{L^2_\varrho(\Gamma,\mathcal{X})} \le \hat{C} \eta_\ell^{-\mu} \tau_0^{\beta}$$

is satisfied for $j = 0, 1$ and $\ell = 0, 1$. This can be done by a fitting similar to the previous paragraph. Then we take $\hat{C}$ as a replacement for $C_I C_\zeta$. As the factor $1/(1 - 2^{-\beta})$ in (3.26) is larger than 1, the determined constant will match the first inequality in (3.11), too.

Ad (3). In the proof of Theorem 3.5.2, the total number of levels J was fixed by formula (3.19). Another strategy of computing J is the following, which was used in [109, Sec. 6.3] and is quite similar to what was described for multi-level *Monte Carlo* methods in [38, Sec. 4.2]. It has the benefit that an error estimator is available during the computation.

First of all we assume that equality holds in (3.9). Then we can estimate

$$\begin{aligned}
\|u_{\tau_j} - u_{\tau_{j-1}}\|_{L^2_\varrho(\Gamma,\mathcal{X})} &\le \|u_{\tau_j} - u(T)\|_{L^2_\varrho(\Gamma,\mathcal{X})} + \|u(T) - u_{\tau_{j-1}}\|_{L^2_\varrho(\Gamma,\mathcal{X})} \\
&\le C_T(1 + 2^\alpha)\tau_j^\alpha, \\
&= (1 + 2^\alpha)\|u_{\tau_j} - u(T)\|_{L^2_\varrho(\Gamma,\mathcal{X})},
\end{aligned} \tag{3.27}$$

and thus it is reasonable to check the necessary condition

$$\|u_{\tau_j} - u_{\tau_{j-1}}\|_{L^2_\varrho(\Gamma,\mathcal{X})} \le (1 + 2^\alpha)\varepsilon/2 \tag{3.28}$$

instead of the (uncomputable) $\|u_{\tau_j} - u(T)\|_{L^2_\varrho(\Gamma,\mathcal{X})} \le \varepsilon/2$. A stronger condition which is even sufficient can be derived as follows. The triangle inequality yields

$$\begin{aligned}
\|u_{\tau_j} - u_{\tau_{j-1}}\|_{L^2_\varrho(\Gamma,\mathcal{X})} &\ge \|u_{\tau_{j-1}} - u(T)\|_{L^2_\varrho(\Gamma,\mathcal{X})} - \|u_{\tau_j} - u(T)\|_{L^2_\varrho(\Gamma,\mathcal{X})} \\
&= C_T(2^\alpha - 1)\tau_j^\alpha = (2^\alpha - 1)\|u_{\tau_j} - u(T)\|_{L^2_\varrho(\Gamma,\mathcal{X})}
\end{aligned}$$

and thus it is reasonable to check the condition

$$\|u_{\tau_j} - u_{\tau_{j-1}}\|_{L^2_\varrho(\Gamma,\mathcal{X})} \le (2^\alpha - 1)\varepsilon/2 \tag{3.29}$$

instead of $\|u_{\tau_j} - u(T)\|_{L^2_\varrho(\Gamma,\mathcal{X})} \le \varepsilon/2$.

This explains the following procedure to find J and compute the corresponding MLSC approximation:

1. Set $J = 1$.
2. Calculate η_j for $j = 0, \dots, J$ according to the formula (3.18) and use a rounding strategy $\widetilde{\eta}_j \approx \eta_j$ to determine the corresponding sparse grid depths.
3. Compute $u^{(\mathrm{SL})}_{\widetilde{\eta}_{J-j},\tau_j} - u^{(\mathrm{SL})}_{\widetilde{\eta}_{J-j},\tau_{j-1}}$ for $j = 0, \dots, J$.
4. Test for convergence by checking (3.29) for $j = J$ with $u^{(\mathrm{SL})}_{\widetilde{\eta}_{J-j},\tau_j} - u^{(\mathrm{SL})}_{\widetilde{\eta}_{J-j},\tau_{j-1}}$ instead of $u_{\tau_j} - u_{\tau_{j-1}}$.
5. If (3.29) is not satisfied, increase J by 1 and return to step 2. Otherwise stop.

We now present a variant of the multi-level stochastic collocation method which specifically targets the approximation of quantities of interest of the solution.

3.6 MLSC for quantities of interest

So far, the goal of the multi-level approach was the computation of an approximation to the solution with tolerance ε. In applications, one is often not interested in the solution $u(T)$ itself, but rather a quantity of interest $\Upsilon(u(T))$ of the solution. In such a case it is reasonable to aim for

$$\begin{aligned} &\big|\mathbb{E}\big[\Upsilon(u(T)) - \Upsilon(u_J^{(\mathrm{ML})})\big]\big| \le \varepsilon \\ \text{instead of} \qquad &\|u(T) - u_J^{(\mathrm{ML})}\|_{L^2_\varrho(\Gamma,\mathcal{X})} \le \varepsilon. \end{aligned}$$

This is usually a much simpler goal, at least if the functional Υ is relatively smooth. We explain below how the multi-level approach can be adapted to this new objective. Throughout this section, we use the notation from the previous sections. The content presented here is essentially contained in [109, Sec. 4.3].

Let $\Upsilon\colon W \to \mathbb{C}$ be a functional and suppose that the subset $W \subseteq \mathcal{X}$ is large enough such that $u(T,y) \in W$ and $u_{\tau_j}(y) \in W$ for almost every $y \in \Gamma$ and all $j \in \mathbb{N}_0$. (One should think of spaces such as $L^2(D)$ or $H^1(D)$ for a spatial domain $D \subseteq \mathbb{R}^N$. The latter space is useful in cases where spatial derivatives occur in the functional Υ.) Without loss of generality, we additionally assume that $0 \in W$ and $\Upsilon(0) = 0$.

A single-level estimator for $\Upsilon(u)$ can be defined as

$$\Upsilon^{(\mathrm{SL})}_{\eta_\ell,\tau_j}[u] = \mathcal{I}_{\eta_\ell}(\Upsilon(u_{\tau_j})),$$

whereas the multi-level estimator for $\Upsilon(u)$ is defined as

$$\Upsilon^{(\mathrm{ML})}_J[u] = \sum_{j=0}^{J} \mathcal{I}_{\eta_{J-j}}\big(\Upsilon(u_{\tau_j}) - \Upsilon(u_{\tau_{j-1}})\big) = \sum_{j=0}^{J} \Upsilon(u^{(\mathrm{SL})}_{\eta_{J-j},\tau_j}) - \Upsilon(u^{(\mathrm{SL})}_{\eta_{J-j},\tau_{j-1}}),$$

where we set $u_{\tau_{-1}} = 0$. In the special case of a linear functional Υ, we have

$$\Upsilon^{(\mathrm{SL})}_{\eta_j,\tau}[u] = \Upsilon(u^{(\mathrm{SL})}_{\eta_j,\tau}) \qquad \text{and} \qquad \Upsilon^{(\mathrm{ML})}_J[u] = \Upsilon(u^{(\mathrm{ML})}_J).$$

We denote the cost of evaluating $\Upsilon(u_{\tau_j}) - \Upsilon(u_{\tau_{j-1}})$ at a sample y by C_j. Note that one usually has to compute both u_{τ_j} and $u_{\tau_{j-1}}$ first, and thus the constant C_j will often be at least as large as the earlier defined constant C_j in Assumption B3.

The following theorem is a variant of the ε-cost theorem for the approximation of functionals from [109, Prop. 4.5].

Theorem 3.6.1 (Multi-level ε-cost theorem for QoIs)**.**
Suppose that there exist constants $\alpha, \beta, \mu, C_T, C_I, C_\zeta > 0$ with $\alpha \geq \min\{\beta, \mu\}$, a Banach space $\Lambda(\Gamma) \hookrightarrow L^2_\varrho(\Gamma, \mathbb{C})$ with $\{\Upsilon(u_{\tau_j})\colon j \in \mathbb{N}_0\} \subseteq \Lambda(\Gamma)$ and an operator $\zeta\colon \Lambda(\Gamma) \to \mathbb{R}$ such that the estimates

$$\begin{aligned}
\left|\mathbb{E}\left[\Upsilon(u(T)) - \Upsilon(u_{\tau_j})\right]\right| &\leq C_T \tau_j^\alpha, \\
\left|\mathbb{E}\left[(\Upsilon(u_{\tau_{j+1}}) - \Upsilon(u_{\tau_j})) - \mathcal{I}_{\eta\ell}(\Upsilon(u_{\tau_{j+1}}) - \Upsilon(u_{\tau_j}))\right]\right| &\leq C_I \eta_\ell^{-\mu} \zeta(\Upsilon(u_{\tau_{j+1}}) - \Upsilon(u_{\tau_j})), \\
\zeta(\Upsilon(u_{\tau_j})) &\leq C_\zeta \tau_0^\beta, \\
\zeta(\Upsilon(u_{\tau_{j+1}}) - \Upsilon(u_{\tau_j})) &\leq C_\zeta \tau_{j+1}^\beta, \\
C_j &\lesssim \tau_j^{-1}
\end{aligned}$$

hold for all $j, \ell \in \mathbb{N}_0$.

Then, for any $\varepsilon < \mathrm{e}^{-1}$, there exists $J \in \mathbb{N}_0$ and a sequence $(\eta_j)_{j=0}^J$ such that

$$\left|\mathbb{E}\left[\Upsilon(u(T)) - \Upsilon_J^{(\mathrm{ML})}[u]\right]\right| \leq \varepsilon$$

and simultaneously

$$C^{(\mathrm{ML})} \lesssim \begin{cases} \varepsilon^{-\frac{1}{\mu}}, & \beta > \mu, \\ \varepsilon^{-\frac{1}{\mu}} |\log(\varepsilon)|^{1+\frac{1}{\mu}}, & \beta = \mu, \\ \varepsilon^{-\frac{1}{\mu} - \frac{\mu-\beta}{\alpha\mu}}, & \beta < \mu, \end{cases}$$

where $C^{(\mathrm{ML})}$ is now the computational cost for $\Upsilon_J^{(\mathrm{ML})}[u]$. The sequence $(\eta_j)_{j=0}^J$ is given by (3.18) again.

Of course, the assumptions of the above theorem are modifications of the Assumptions B1 – B3. Although the constants and rates above are denoted by the same symbols, they may take different values than the earlier defined constants and rates. We always indicate in our numerical experiments later if we apply the multi-level approach to the solution or a quantity of interest of it and the constants and rates are then understood as the correct ones for this approach.

It should also be noted that for some functionals Υ, the assumptions from Theorem 3.6.1 can be verified directly from Assumption B1 – B3. An example for this are bounded linear functionals Υ.

3.7 MLSC or MLMC?

Readers familiar with multi-level Monte Carlo (MLMC) methods might be interested in the performance of MLSC compared to MLMC. In the article from Teckentrup et al. [109, Sec. 6], it was shown that for an elliptic problem on a unit interval/square with Karhunen-Loève expanded (logarithmic) diffusion coefficient and either $(N, d) = (2, 10)$ or $(N, d) = (1, 20)$ (spatial dimension N, stochastic dimension d), MLSC requires much less computational work to achieve the same accuracy as MLMC. Here, "much less" usually means several orders of magnitude, but the overall cost reduction is dependent on the desired accuracy. Thus, for problems with sufficent regularity in y (and moderately large stochastic dimension) the usage of MLSC is usually preferable, as is the case for stochastic collocation and Monte Carlo methods without multi-level approach. The situation might be different if unusual basic random variables appear and no reasonable set of collocation points is available for the discretisation of the parameter space, or when the dimension of the parameter space is too large to use collocation at all. In such cases Monte Carlo

methods are attractive again. It is yet to be seen how collocation and Monte Carlo methods compare with each other when applied to more complicated problems. Although this is certainly an interesting question for future research, it is not the topic of this thesis to compare MLSC and MLMC.

We conclude this chapter with a brief summary of known extensions of MLSC.

3.8 Multi-index stochastic collocation and other extensions

In many problems, the spatial discretisation is typically "more costly" than the temporal discretisation as the temporal dimension is always 1. Hence, it is often more advantageous to balance the cost of spatial and stochastic discretisations, and not temporal and stochastic discretisations, as is the case in this thesis. Especially for this reason, the multi-index stochastic collocation (MISC) approach from [45, 44] is perhaps the most important extension of MLSC.

This approach computes an estimator based on mixed difference operators in all individual spatiotemporal and stochastic dimensions. This is different for the MLSC method described here, where the refinement in the stochastic (and temporal) dimensions is determined by a single parameter (and where the spatial discretisation is not discussed at all). Based on profits computed from a priori work and error bounds, one arrives at a knapsack problem for the optimal index set of the difference operator. A quasi-optimal multi-index set is then selected by solving a slightly simplified knapsack problem.

The MISC method is not only capable of balancing spatial, temporal and stochastic components of the error, but the knapsack problem approach avoids the usage of rounding strategies, which is an attractive feature in comparison with the MLSC method.

Another remarkable extension of MLSC was presented by Lang, Scheichl and Silvester in [65], where the approach from Teckentrup et al. [109] was extended in such a way that the adaptive (spatial) mesh refinement is allowed to vary with the samples. This allows for an optimisation of the computational work in each stochastic collocation point and was shown to be superior to strategies which are only adaptive in the spatial or stochastic discretisation, but ignore properties of individual samples.

CHAPTER 4

Multi-level stochastic collocation for parabolic equations

4.1 Motivation

Among the most famous PDE systems appearing in biology are *predator-prey* systems. The equations for the prey and predator densities $u = u(t,x)$ and $v = v(t,x)$ usually take the form

$$\partial_t u = \delta_1 \Delta u + R_1(u,v), \qquad \text{in } [0,T] \times D, \tag{4.1a}$$

$$\partial_t v = \delta_2 \Delta v + R_2(u,v), \qquad \text{in } [0,T] \times D, \tag{4.1b}$$

$$u(0,x) = u_0(x), \qquad \text{for } x \in D, \tag{4.1c}$$

$$v(0,x) = v_0(x), \qquad \text{for } x \in D, \tag{4.1d}$$

$$\frac{\partial u}{\partial \nu} = \frac{\partial v}{\partial \nu} = 0, \qquad \text{on } [0,T] \times \partial D, \tag{4.1e}$$

where ν denotes the outward unit normal to the boundary ∂D of the spatial domain $D \subseteq \mathbb{R}^N$, $N \in \{1,2,3\}$, and $\delta_1, \delta_2 > 0$ are diffusion constants. The functions R_1, R_2 account for the reaction dynamics between predators and prey. The behaviour of u and v depends primarily on R_1 and R_2. Although solving this quite general system under the influence of uncertainty will be one of our main objectives, we now give a more concrete setting described in the literature.

In [32], the system

$$\partial_t u = \Delta u + u(1-u) - vh(au), \qquad \text{in } [0,T] \times D, \tag{4.2a}$$

$$\partial_t v = \delta \Delta v + bvh(au) - cv, \qquad \text{in } [0,T] \times D, \tag{4.2b}$$

$$u(0,x) = u_0(x), \qquad \text{for } x \in D, \tag{4.2c}$$

$$v(0,x) = v_0(x), \qquad \text{for } x \in D, \tag{4.2d}$$

$$\frac{\partial u}{\partial \nu} = \frac{\partial v}{\partial \nu} = 0, \qquad \text{on } [0,T] \times \partial D, \tag{4.2e}$$

with parameters a, b, $c > 0$ was discussed. Here, the function h represents the instantaneous feeding

rate as a function of prey abundance. A typical choice for h is

$$h_{\mathrm{Hol}}(w) = \frac{w}{1+w},$$

or its second-order Taylor polynomial at $w = 0$ given by $h(w) = w(1-w)$. This choice is originally due to Holling [54] and h_{Hol} is thus called the "Holling type II functional response". See [25] for more details and other functional responses. The dynamics of this system, saddle and stationary points are discussed in [32].

Let us briefly describe the meaning of the individual terms in the system (4.2). The terms Δu and $\delta\Delta v$ represent (spatial) dispersion of prey and predators, $u(1-u)$ describes birth, death and social friction of the prey in absence of predators and $-vh(au)$ the effect of predation on the prey. The influence of predation on the predators is incorporated via the term $bvh(au)$ and the death of the predators is described by $-cv$. Observe that the term $u(1-u)$ implies logistic growth of the prey in the absence of predators. The appearing constants δ, a, b and c are usually not known and have to be guessed from observations. Hence it is reasonable to model them as uncertain parameters and study the behaviour of the system under the influence of these uncertainties.

An introduction to predator-prey systems can be found in [55]. In this work, the PDE setting and thus spatial *hetero*geneity is covered in detail (in contrast to many other articles on population dynamics). Considering spatial heterogeneity is important since the famous ODE setting known as the *Lotka-Volterra equations* corresponds to spatial *homo*geneity and is not capable of predicting the behaviour of species near boundaries of a habitat, their spatial distribution and behaviour in scenarios such as invasion or exposure to a new environment.

In [32, 31], a treatment of the system (4.2) via finite element methods for the spatial discretisation is explained. Both theoretical and practical aspects are discussed there.

4.2 A model problem

Although the application we have in mind is a coupled system of two PDEs with uncertain parameters, we restrict ourselves to a single PDE here, supplied with initial and boundary conditions. In the context of predator-prey equations from the previous section, we may interpret the PDE stated below as a model for the prey dynamics in absence of predators. Relevant analytical and numerical aspects are discussed for this PDE and we only return to the coupled system at the end of the chapter. The setting here is derived from [75, Sec. 3.1].

We consider a parametric PDE of the form[1]

$$\partial_t u(t,x,z) = A(x,z)u(t,x,z) + R(u(t,x,z),z) + S(t,x,z), \qquad t \geq 0,\ x \in D,\ z \in \Sigma, \tag{4.3a}$$

$$B(x,z)u(t,x,z) = 0, \qquad t \geq 0,\ x \in \partial D,\ z \in \Sigma, \tag{4.3b}$$

$$u(0,x,z) = u_0(x,z), \qquad x \in D,\ z \in \Sigma. \tag{4.3c}$$

Here, the spatial domain is an open bounded subset $D \subsetneq \mathbb{R}^N$ with C^2-boundary ∂D and exterior unit normal vector $\nu(x)$ at $x \in \partial D$. The parameter set Σ is some bounded domain $\Sigma \subseteq \mathbb{C}^d$ which contains $\Gamma = [-1,1]^d$. Although we are primarily interested in $z \in [-1,1]^d$ (in which case we called this parameter

[1] The notation in (4.3a) can be explained as follows: R for "reaction" and S for "source".

y instead of z in previous chapters), we need complex parameters later on and thus allow them from the beginning. Hence, the solution u is *complex-valued*, too. Moreover, $R\colon \mathbb{C}\times\Sigma\to\mathbb{C}$, $S\colon [0,\infty)\times D\times\Sigma\to\mathbb{C}$ and $u_0\colon D\times\Sigma\to\mathbb{C}$ are given functions and[2]

$$A(x,z)=\sum_{i,j=1}^{N} a_{ij}(x,z)\partial_{ij}+\sum_{i=1}^{N} a_i(x,z)\partial_i+a_0(x,z)I, \tag{4.4}$$

where $a_{ij},\ a_i$ are uniformly continuous coefficient functions on $\overline{D}\times\overline{\Sigma}$. The complex matrix $(a_{ij}(x,z))_{i,j=1}^{N}$ is assumed to be symmetric for every $(x,z)\in D\times\Sigma$.

We assume *strong ellipticity* for the diffusion part in the sense that

$$\operatorname{Re}\left(\sum_{i,j=1}^{N} a_{ij}(x,z)\xi_i\xi_j\right)\geq a_{\min}|\xi|^2,\qquad x\in\overline{D},\ z\in\overline{\Sigma},\ \xi\in\mathbb{R}^N$$

for some constant $a_{\min}>0$. Note that this is just a requirement on the principal part of the differential operator. For the boundary operator B in (4.3b), we assume

$$B(x,z)u(x)=\sum_{i=1}^{N}\beta_i(x,z)\partial_i u(x)+\alpha(x,z)u(x) \tag{4.5}$$

with functions $\beta_i,\ \alpha\in C^1(\overline{D}\times\overline{\Sigma})$ satisfying the non-tangentiality condition

$$\inf_{x\in\partial D}\left|\sum_{i=1}^{N}\beta_i(x,z)\nu_i(x)\right|>0.$$

A typical example for this situation is $\beta_i(x,z)=\sum_{j=1}^{N}a_{ij}(x,z)\nu_j(x)$ and $\alpha(x,z)=0$, such that

$$B(x,z)u(x)=\sum_{i,j=1}^{N}a_{ij}(x,z)\nu_j(x)\partial_i u(x)$$

is the conormal derivative corresponding to a Neumann boundary condition. The regularity of R,S and u_0 will be specified later on.

Often, we hide the spatial variable x in our exposition and hence abbreviate system (4.3) as

$$\begin{aligned}
\partial_t u(t,z)&=A(z)u(t,z)+R(u(t,z),z)+S(t,z), && t\geq 0,\ z\in\Sigma,\\
B(z)u(t,z)|_{\partial D}&=0, && t\geq 0,\ z\in\Sigma,\\
u(0,z)&=u_0(z), && z\in\Sigma.
\end{aligned}$$

Let us now discuss some more advanced properties of the linear operator $A(z)$.

4.2.1 Properties of the linear part

Consider the Banach space $X=L^p(D)$ with $1<p<\infty$ and the domain of the operators $A(z)$ defined in (4.4), given by

$$\mathcal{D}(A(z))=W^{2,p}_{B(z)}(D)=\{v\in W^{2,p}(D)\colon B(z)v|_{\partial D}=0\}, \tag{4.7}$$

where the restriction $|_{\partial D}$ is understood as an application of the trace operator. We now discuss in detail that the operator $A(z)$ with domain $\mathcal{D}(A(z))$ has the important property that it is *sectorial* for any given $z\in\Sigma$. This notion is defined as follows.

[2]The partial derivatives ∂_{ij} and ∂_i act on the spatial variable x, and not on z.

Definition 4.2.1. A linear operator $\mathcal{A}\colon \mathcal{X} \supseteq \mathcal{D}(\mathcal{A}) \to \mathcal{X}$ on a Banach space $\mathcal{X}$ is called *sectorial* if there exist $\vartheta \in \left(\frac{\pi}{2}, \pi\right)$, $\omega \in \mathbb{R}$ and $M > 0$ such that the sector

$$S_{\vartheta,\omega} = \{\lambda \in \mathbb{C}\colon\ \lambda \neq \omega, |\arg(\lambda - \omega)| < \vartheta\}$$

is contained in the resolvent set $\varrho(\mathcal{A})$ and that the corresponding resolvents are bounded by

$$\|(\lambda I - \mathcal{A})^{-1}\|_{\mathcal{L}(\mathcal{X})} \leq \frac{M}{|\lambda - \omega|} \qquad \text{for all} \quad \lambda \in S_{\vartheta,\omega}.$$

(Occasionally, we will write $\omega_{\mathcal{A}}$ instead of ω if we want to emphasise the corresponding operator.)

Notation. To distinguish better between abstract definitions and results and statements about the specific model problem, "abstract" quantities, operators and spaces are denoted by calligraphic symbols such as $\mathcal{A}$ and $\mathcal{X}$, whereas the specific second-order elliptic operator from (4.4) is denoted by the standard capital letter A (and similar for the corresponding Banach space $X = L^p(D)$).

It is well-known that sectorial operators generate analytic semigroups in the sense of part (b) of the following lemma, shown in [75, Prop. 2.1.1].

Lemma 4.2.2. *Let $\mathcal{A}\colon \mathcal{X} \supseteq \mathcal{D}(\mathcal{A}) \to \mathcal{X}$ be sectorial with corresponding quantities ϑ, ω and M.*

(a) There are constants M_k, $k \in \mathbb{N}_0$, such that the strongly continuous semigroup $(\mathrm{e}^{t\mathcal{A}})_{t\geq 0}$ generated by $\mathcal{A}$ satisfies the inequalities

$$\|\mathrm{e}^{t\mathcal{A}}\|_{\mathcal{L}(\mathcal{X})} \leq M_0 \mathrm{e}^{\omega t}, \qquad \|t^k(\mathcal{A} - \omega I)^k \mathrm{e}^{t\mathcal{A}}\|_{\mathcal{L}(\mathcal{X})} \leq M_k \mathrm{e}^{\omega t}$$

for $k \in \mathbb{N}$ and all $t > 0$. More precisely, the constant M_0 only depends on M and some angle $\varphi \in (\frac{\pi}{2}, \vartheta)$, whereas M_1 is explicitly given by $M_1 = \frac{M}{\pi}\left(\frac{1}{|\cos(\vartheta)|} + \mathrm{e}\vartheta\right)$.

(b) The map $(0,\infty) \to \mathcal{L}(\mathcal{X})$, $t \mapsto \mathrm{e}^{t\mathcal{A}}$ belongs to $C^\infty((0,\infty), \mathcal{L}(\mathcal{X}))$ and has an analytic extension in the sector

$$S_{\vartheta - \frac{\pi}{2}, 0} = \{\lambda \in \mathbb{C}\colon \lambda \neq 0,\ |\arg(\lambda)| < \vartheta - \tfrac{\pi}{2}\}.$$

Note that the sector appearing in part (b) is not the same sector as in the definition of "sectorial", but – metaphorically speaking – the sector reduced by "half of a cake" and shifted by ω.

We introduce a bit of notation for the next theorem. Let

$$\|\mathscr{D}u\|_p = \left(\sum_{i=1}^N \|\partial_i u\|_{L^p(D)}^p\right)^{1/p}, \qquad \|\mathscr{D}^2 u\|_p = \left(\sum_{i,j=1}^N \|\partial_{ij} u\|_{L^p(D)}^p\right)^{1/p}$$

and

$$M^* = \max\{\|a_{ij}\|_\infty, \|a_i\|_\infty\} < \infty,$$

where the supremum norm $\|\cdot\|_\infty$ is taken with respect to $(x,z) \in \overline{D} \times \overline{\Sigma}$. Now we present the main result of this section, which is proved in [75, Thm. 3.1.3].

Theorem 4.2.3. *Let $1 < p < \infty$ and $z \in \Sigma$. The operators $A(z) = A(\cdot, z)$ and $B(z) = B(\cdot, z)$ are defined as in (4.4) and (4.5) with the assumptions stated in Section 4.2.*

We have the following results:

(a) *There exists* $\omega_0 \in \mathbb{R}$ *such that if* $\operatorname{Re}(\lambda) \geq \omega_0$, *then for every* $v \in L^p(D)$ *and* $w \in W^{1,p}(D)$, *the problem*

$$(\lambda I - A(z))u = v \quad \text{in } D, \qquad B(z)u = w|_{\partial D} \quad \text{on } \partial D$$

has a unique solution $u \in W^{2,p}(D)$ *which depends continuously on* v *and* w.

Taking $w = 0$, *it follows that* $\{\lambda \in \mathbb{C} \colon \operatorname{Re}(\lambda) \geq \omega_0\} \subseteq \varrho(A(z))$.

(b) *There exist* $\omega \geq \omega_0$ *and* $\widetilde{M} > 0$ *(with both quantities depending only on* $N, p, a_{\min}, M^*$ *and* D*) such that if* $\operatorname{Re}(\lambda) \geq \omega$, *then for every* $u \in W^{2,p}(D)$ *we have, setting* $w = B(z)u|_{\partial D}$,

$$|\lambda|\|u\|_{L^p(D)} + |\lambda|^{\frac{1}{2}}\|\mathscr{D}u\|_p + \|\mathscr{D}^2 u\|_p \leq \widetilde{M}\left(\|\lambda u - A(z)u\|_{L^p(D)} + |\lambda|^{\frac{1}{2}}\|w_1\|_{L^p(D)} + \|\mathscr{D}w_1\|_p\right),$$

where w_1 *is any extension of* w *belonging to* $W^{1,p}(D)$.

Taking $w = 0$, *we obtain*

$$|\lambda|\|u\|_{L^p(D)} \leq \widetilde{M}\|\lambda u - A(z)u\|_{L^p(D)}$$

for any $u \in \mathcal{D}(A(z))$ *or, equivalently,*

$$\|\lambda(\lambda I - A(z))^{-1}v\|_{L^p(D)} \leq \widetilde{M}\|v\|_{L^p(D)}$$

for any $v \in L^p(D)$.

The careful reader might notice that the angle $\vartheta > \frac{\pi}{2}$ from the definition of sectoriality is missing yet. So far, the resolvent set only contains a half-plane where the resolvent estimates are available. But a standard result says that this is in fact enough to ensure sectoriality. (Roughly speaking, this follows from the fact that the resolvent set $\varrho(\mathcal{A})$ is open.) A proof of the result below is given in [75, Prop. 2.1.11].

Lemma 4.2.4. *Let* $\mathcal{A} \colon \mathcal{X} \supseteq \mathcal{D}(\mathcal{A}) \to \mathcal{X}$ *be a linear operator such that* $\varrho(\mathcal{A})$ *contains a half-plane* $\{\lambda \in \mathbb{C} \colon \operatorname{Re}(\lambda) \geq \omega\}$ *and*

$$\|\lambda(\lambda I - \mathcal{A})^{-1}\|_{\mathcal{L}(\mathcal{X})} \leq \widetilde{M}, \qquad \operatorname{Re}(\lambda) \geq \omega,$$

with $\omega \in \mathbb{R}$, $\widetilde{M} > 0$. *Then* $\mathcal{A}$ *is sectorial with* $\vartheta = \pi - \arctan(\widetilde{M})$ *and a constant* M *which depends only on* $\widetilde{M}$.

Thus, Theorem 4.2.3 and Lemma 4.2.4 imply the following.

Corollary 4.2.5. *Under the assumptions stated in Section 4.2, the operators* $A(z) \colon X \supseteq \mathcal{D}(A(z)) \to X$ *defined by* (4.4) *with* $\mathcal{D}(A(z)) = W^{2,p}_{B(z)}(D)$ *and* $X = L^p(D)$ *are sectorial for any* $z \in \Sigma$.

In addition to the sectoriality, the dependencies stated in the results Theorem 4.2.3 and Lemma 4.2.4 show that the characteristic quantities ω, ϑ and M can be chosen uniformly in $z \in \Sigma$. We observe that the constants M_0 and M_1 in Lemma 4.2.2 (applied to $A(z)$) can be chosen uniformly in $z \in \Sigma$, too.

Remark 4.2.6 (Dirichlet boundary conditions)**.** In the previous discussion, the Neumann boundary condition $B(z)u|_{\partial D} = 0$ and the domain $\mathcal{D}(A(z)) = W^{2,p}_{B(z)}(D)$ can be replaced by a Dirichlet boundary condition $u|_{\partial D} = 0$ and corresponding domain $\mathcal{D}(A(z)) = W^{2,p}(D) \cap W^{1,p}_0(D)$. The statements in Theorem 4.2.3 and Corollary 4.2.5 remain true up to some obvious changes. They are stated in detail in [75, Sec. 3.1.1] and are thus omitted here. This remark is included since we discuss two examples later in which we consider Dirichlet or mixed Dirichlet-Neumann boundary conditions. ◇

Remark 4.2.7 (Exponential stability)**.** Later we will assume for $A(z)$ that the constant ω from the definiton of "sectorial" satisfies $\omega < 0$. This corresponds to the definition of an *exponentially stable sectorial* operator. This is not a severe restriction, since one can achieve this by a suitable scaling of the operator $A(z)$, i.e. replacing $A(z)$ by $A(z) - \omega I$. (More details are given shortly.) As a consequence, we will often assume

$$0 \in \varrho(A(z)) \qquad \text{and} \qquad A(z)^{-1} \in \mathcal{L}(X, \mathcal{D}(A)).$$

Both properties will be used in some places later on.

Let us now sketch the scaling procedure which allows us to replace the sectorial operator $A(z)$ by an exponentially stable sectorial operator without significantly changing the problem under consideration. As this discussion is not affected by the variable z, we omit it in the presentation. Suppose we have a solution to the initial value problem

$$\begin{aligned} u'(t) &= Au(t) + R(u(t)) + S(t), \qquad t \geq 0, \\ u(0) &= u_0, \end{aligned}$$

then $v(t) = \mathrm{e}^{-\omega' t} u(t)$ is a solution to

$$\begin{aligned} v'(t) &= (A - \omega' I)v(t) + \widetilde{R}(t, v(t)) + \widetilde{S}(t), \qquad t \geq 0, \\ v(0) &= u_0 \end{aligned}$$

with $\widetilde{R}(t, v(t)) = \mathrm{e}^{-\omega' t} R(\mathrm{e}^{\omega' t} v(t))$ and $\widetilde{S}(t) = \mathrm{e}^{-\omega' t} S(t)$. This follows from

$$\begin{aligned} v'(t) &= -\omega' \mathrm{e}^{-\omega' t} u(t) + \mathrm{e}^{-\omega' t} u'(t) \\ &= -\omega' v(t) + A\mathrm{e}^{-\omega' t} u(t) + \mathrm{e}^{-\omega' t} R(u(t)) + \mathrm{e}^{-\omega' t} S(t) \\ &= (A - \omega' I)v(t) + \mathrm{e}^{-\omega' t} R(\mathrm{e}^{\omega' t} v(t)) + \mathrm{e}^{-\omega' t} S(t). \end{aligned}$$

If we set $\omega' = \omega + \varepsilon$ for some $\varepsilon > 0$ (where ω is the quantity from Definition 4.2.1), then $A - \omega' I$ has a growth bound less than zero as desired. It should be noted, however, that there is a small structural difference between the two initial value problems for u and v: The function $\widetilde{R}$ depends explicitly on t, which is not the case for R. This is not crucial in our exposition. ◇

As a final remark, we stress that the operators $\mathrm{e}^{tA(z)}$ are *not contractive* in general (unless $p = 2$). A criterion for the contractivity of semigroups generated by strongly elliptic second-order differential operators with complex coefficients was given in [17] in case of Dirichlet boundary conditions.

4.3 Wellposedness and regularity

Here we consider (non-parametric) evolution equations of the form

$$u'(t) = \mathcal{A}u(t) + \mathcal{F}(t, u(t)), \qquad t \geq 0, \qquad u(0) = u_0, \tag{4.8}$$

where $\mathcal{A}\colon \mathcal{X} \supseteq \mathcal{D}(\mathcal{A}) \to \mathcal{X}$ is sectorial with dense domain $\mathcal{D}(\mathcal{A})$ and $\mathcal{F}\colon [0, T] \times \mathcal{O} \to \mathcal{X}$ is continuous, where $\mathcal{O} \subseteq \mathcal{X}_\gamma$ is an open set. Here, $0 < \gamma < 1$ and $\mathcal{X}_\gamma$ is either the domain $\mathcal{D}((-\mathcal{A})^\gamma)$ of a fractional power of $-\mathcal{A}$, a continuous real interpolation space $\mathcal{D}_{\mathcal{A}}(\gamma)$ or a complex interpolation space[3] $[\mathcal{D}(\mathcal{A}), \mathcal{X}]_\gamma$.

[3] See the beginning of [75, Chap. 7] for a more general description of $\mathcal{X}_\gamma$ which includes the three above-mentioned special cases.

For explanations of these Banach space constructions, we refer again to [75]. However, we do not need the definitions of these spaces and instead refer to some of their properties whenever we make use of them. Boundary conditions of an underlying initial boundary value problem for (4.8) are assumed to be incorporated into the domain $\mathcal{D}(\mathcal{A})$. We further assume that $\mathcal{F}$ is locally Lipschitz continuous in the second variable, which means that for every $v_0 \in \mathcal{O}$ there are $r,\ L > 0$ such that

$$\|\mathcal{F}(t, v_2) - \mathcal{F}(t, v_1)\|_{\mathcal{X}} \leq L\|v_2 - v_1\|_{\mathcal{X}_\gamma} \quad \text{for all} \quad t \in [0, T],\ v_1, v_2 \in B_{\mathcal{X}_\gamma}(v_0, r).$$

For the initial value, we assume that $u_0 \in \mathcal{O}$.

We stress that the abstract setting here of course allows that both $\mathcal{A}$ and $\mathcal{F}$ depend on the spatial variable x of a specific problem such as (4.3). This dependency does not appear explicitly in (4.8) since it is hidden inside the Banach space $\mathcal{X}$, which is typically some function space such as, e.g., $L^p(D)$.

We consider the following solution concepts for (4.8).

Definition 4.3.1 (Solution concepts for (4.8))**.** Let $b > 0$.

- A function $u \in C^1((0, b], \mathcal{X}) \cap C([0, b], \mathcal{X}) \cap C((0, b], \mathcal{D}(\mathcal{A}))$ such that $u(t) \in \mathcal{O}$ for every $t \in [0, b]$ is said to be a *classical solution* of (4.8) in the interval $[0, b]$ if $u'(t) = \mathcal{A}u(t) + \mathcal{F}(t, u(t))$ for each $t \in (0, b]$ and $u(0) = u_0$.
- A function $u \in C^1([0, b], \mathcal{X}) \cap C([0, b], \mathcal{D}(\mathcal{A}))$ such that $u(t) \in \mathcal{O}$ for every $t \in [0, b]$ is said to be a *strict solution* of (4.8) in the interval $[0, b]$ if $u'(t) = \mathcal{A}u(t) + \mathcal{F}(t, u(t))$ for each $t \in [0, b]$ and $u(0) = u_0$.

Clearly, strict solutions are classical solutions, but the reverse is not true. There are many theorems on wellposedness and regularity for the problem considered here. We state one of them which assumes Hölder-continuity in time for $\mathcal{F}$ (which is trivially satisfied if $\mathcal{F}$ is autonomous). The result is taken from [75, Thm. 7.1.10], where a proof is given, too.

Theorem 4.3.2. *Assume that there exists $\theta \in (0, 1)$ such that for each $u_0 \in \mathcal{O}$ there are $r,\ K$ such that*

$$\|\mathcal{F}(t_2, v) - \mathcal{F}(t_1, v)\|_{\mathcal{X}} \leq K(t_2 - t_1)^\theta, \quad 0 \leq t_1 \leq t_2 \leq T, \quad \|v - u_0\|_{\mathcal{X}_\gamma} \leq r.$$

Then the following statements are true:

(a) There exists a classical solution u of (4.8) on a (possibly half-open or unbounded) maximal interval of existence denoted by $I(u_0)$.

Now fix any compact interval $[0, b] \subseteq I(u_0)$.

(b) If also $u_0 \in \mathcal{D}(\mathcal{A})$ holds, then u is bounded as a map $[0, b] \to \mathcal{D}(\mathcal{A})$ and Lipschitz continuous as a map $[0, b] \to \mathcal{X}$ for any compact interval $[0, b] \subseteq I(u_0)$.

(c) If also $\mathcal{A}u_0 + \mathcal{F}(0, u_0) \in \mathcal{D}(\mathcal{A})$, then $u \in C^1([0, b], \mathcal{X}) \cap C([0, b], \mathcal{D}(\mathcal{A}))$ and it is a strict solution.

Let us apply this theorem to the model problem with a fixed parameter $z \in \Sigma$.

Example 4.3.3 (Application to (4.3) – low regularity)**.**

Consider the model problem (4.3) for fixed $z \in \Sigma$, where $A(z) = A(\cdot, z)$ is given by (4.4) with domain $\mathcal{D}(A(z)) = W^{2,p}_{B(z)}(D)$ and $X = L^p(D)$ for $1 < p < \infty$. Assume that there exist $\theta \in (0,1)$ and K_z such that

$$\|S(t_2, \cdot, z) - S(t_1, \cdot, z)\|_X \leq K_z (t_2 - t_1)^\theta, \qquad 0 \leq t_1 \leq t_2 \leq T.$$

We further assume that R is locally Lipschitz continuous, i.e. for every $v_0 \in X_\gamma$ there exist $r > 0$ and $L_{r,z} > 0$ such that

$$\|R(v_2, z) - R(v_1, z)\|_X \leq L_{r,z} \|v_2 - v_1\|_{X_\gamma} \quad \text{for all} \quad v_1, v_2 \in \mathcal{B}_{X_\gamma}(v_0, r).$$

If

$$u_0(\cdot, z) \in \mathcal{D}(A(z)) \qquad \text{and} \qquad A(z)u_0(\cdot, z) + R(u_0(\cdot, z), z) + S(0, \cdot, z) \in \mathcal{D}(A(z)),$$

then, by sectoriality of $A(z)$ and Theorem 4.3.2(c), there exists a strict solution

$$u(\cdot, \cdot, z) \in C^1([0, b], X) \cap C([0, b], \mathcal{D}(A(z)))$$

to (4.3).

Now we discuss an example with higher regularity. In the following, the domains of powers of a linear operator $\mathcal{A} \colon \mathcal{X} \supseteq \mathcal{D}(\mathcal{A}) \to \mathcal{X}$ are defined iteratively by

$$\mathcal{D}(\mathcal{A}^{k+1}) = \{v \in \mathcal{D}(\mathcal{A}^k) \colon \mathcal{A}v \in \mathcal{D}(\mathcal{A}^k)\}, \qquad k \in \mathbb{N}.$$

These domains are equipped with the graph norm of $\mathcal{A}^{k+1}$.

Example 4.3.4 (Application to (4.3) – high regularity)**.**

Consider again (4.3) for fixed $z \in \Sigma$. Assume that there exist $\theta \in (0,1)$ and K_z such that

$$\|S(t_2, \cdot, z) - S(t_1, \cdot, z)\|_{\mathcal{D}(A(z)^2)} \leq K_z (t_2 - t_1)^\theta, \qquad 0 \leq t_1 \leq t_2 \leq T.$$

Further assume that R is locally Lipschitz continuous in $\mathcal{D}(A(z)^2)_\gamma$ (where $\mathcal{D}(A(z)^2)_\gamma$ lies between $\mathcal{D}(A(z)^2)$ and $\mathcal{D}(A(z)^3)$) for each $z \in \Sigma$. Then the following statements follow from the sectoriality of $A(z)$ and Theorem 4.3.2.

(a) If $u_0(\cdot, z) \in \mathcal{D}(A(z)^2)_\gamma$, there exists a classical solution u with regularity

$$u(\cdot, \cdot, z) \in C^1((0, b], \mathcal{D}(A(z)^2)) \cap C([0, b], \mathcal{D}(A(z)^2)) \cap C((0, b], \mathcal{D}(A(z)^3))$$

to (4.3) for each $[0, b] \subseteq I(u_0)$.

(b) If

$$u_0(\cdot, z) \in \mathcal{D}(A(z)^3) \qquad \text{and} \qquad A(z)u_0(\cdot, z) + R(u_0(\cdot, z), z) + S(0, \cdot, z) \in \mathcal{D}(A(z)^3),$$

then there exists a strict solution

$$u(\cdot, \cdot, z) \in C^1([0, b], \mathcal{D}(A(z)^2)) \cap C([0, b], \mathcal{D}(A(z)^3))$$

of (4.3).

Let us further assume more temporal regularity: Suppose that $R(\cdot,z) \in C^1(\mathcal{D}(A(z))_\gamma, \mathcal{D}(A(z))_\gamma)$ (where $\mathcal{D}(A(z))_\gamma$ lies between $\mathcal{D}(A(z))$ and $\mathcal{D}(A(z)^2)$), $S(\cdot,\cdot,z) \in C^1([0,T],\mathcal{D}(A(z)))$ and that the Fréchet derivatives $\partial_u R$ and $\partial_t S$ are θ-Hölder continuous in $\mathcal{D}(A(z))$. Then we may differentiate (4.3) with respect to t, resulting in the problem

$$\partial_t v(t,\cdot,z) = A(z)v(t,\cdot,z) + \partial_u R(u(t,\cdot,z),z)v(t,\cdot,z) + \partial_t S(t,\cdot,z), \qquad t \geq 0,\ z \in \Sigma, \tag{4.9a}$$

$$B(\cdot,z)v(t,\cdot,z)|_{\partial D} = 0, \qquad t \geq 0,\ z \in \Sigma, \tag{4.9b}$$

$$v(0,\cdot,z) = v_0(\cdot,z), \qquad z \in \Sigma, \tag{4.9c}$$

where we have set $v_0(\cdot,z) = \partial_t u(0,\cdot,z)$. The following statements hold:

(c) If $v_0(\cdot,z) \in \mathcal{D}(A(z))_\gamma$, there exists a classical solution v to (4.9) with regularity

$$v(\cdot,\cdot,z) \in C^1((0,b],\mathcal{D}(A(z))) \cap C([0,b],\mathcal{D}(A(z))) \cap C((0,b],\mathcal{D}(A(z)^2))$$

for each $[0,b] \subseteq I(v_0(\cdot,z))$.

(d) If

$$v_0(\cdot,z) \in \mathcal{D}(A(z)^2) \tag{4.10a}$$

$$\text{and} \qquad A(z)v_0(\cdot,z) + \partial_u R(u_0(\cdot,z),z)v_0(\cdot,z) + \partial_t S(0,z) \in \mathcal{D}(A(z)^2), \tag{4.10b}$$

then, by Theorem 4.3.2 and the sectoriality of $A(z)$, there exists a strict solution

$$v(\cdot,\cdot,z) \in C^1([0,b],\mathcal{D}(A(z))) \cap C([0,b],\mathcal{D}(A(z)^2))$$

to (4.9). This implies that

$$u(\cdot,\cdot,z) \in C^2([0,b],\mathcal{D}(A(z))) \cap C^1([0,b],\mathcal{D}(A(z)^2))$$

for $[0,b] \subseteq I(v_0(\cdot,z)) \cap I(u_0(\cdot,z))$.

Note that we can get rid of v_0 in (4.10) to obtain the (equivalent) conditions

$$A(z)u_0(\cdot,z) + R(u_0(\cdot,z),z) + S(0,\cdot,z) \in \mathcal{D}(A(z)^2)$$

$$\text{and} \qquad (A(z) + \partial_u R(u_0(\cdot,z),z))[A(z)u_0(\cdot,z) + R(u_0(\cdot,z),z) + S(0,\cdot,z)] + \partial_t S(0,\cdot,z) \in \mathcal{D}(A(z)^2),$$

in which u_0 appears instead of v_0.

Further regularity properties and dependence on the data are examined in [75, Sec. 8.3]. One particular result concerning analytic dependence on the data will be briefly discussed later in Section 4.6 and is especially relevant for stochastic collocation methods.

Now we briefly explain the idea behind trapezoidal splitting methods. These methods can be used for the time discretisation of certain (ordinary or partial) differential equations and are a common choice for advection-diffusion-reaction problems, see [57, Sec. IV.2.3].

4.4 Interlude: Trapezoidal splitting method

In general, the trapezoidal splitting method for an ordinary initial value problem

$$\begin{aligned} w'(t) &= \mathcal{F}_1(t, w(t)) + \cdots + \mathcal{F}_s(t, w(t)), \qquad t \geq 0, \\ w(t) &= w_0 \end{aligned}$$

computes approximations $w_n \approx w(t_n)$ at times $t_n = n\tau$ for $n \in \mathbb{N}_0$ and a given step-size $\tau > 0$ via the recursion

$$\begin{aligned} v_0 &= w_n, \\ v_i &= v_{i-1} + \frac{\tau}{2}\mathcal{F}_i(t_n, v_{i-1}), && i = 1, \dots, s, \\ v_{s+i} &= v_{s+i-1} + \frac{\tau}{2}\mathcal{F}_{s+1-i}(t_{n+1}, v_{s+i}), && i = 1, \dots, s, \\ w_{n+1} &= v_{2s}. \end{aligned}$$

Loosely speaking, this can be interpreted as the following procedure to compute w_{n+1} from w_n:

1. Successively compute explicit Euler steps for $\mathcal{F}_1, \dots, \mathcal{F}_{s-1}$ over the interval $\frac{\tau}{2}$.
2. Compute an explicit Euler step for $\mathcal{F}_s$ over the interval $\frac{\tau}{2}$.
3. Compute an implicit Euler step for $\mathcal{F}_s$ over the interval $\frac{\tau}{2}$.
4. Successively compute implicit Euler steps for $\mathcal{F}_{s-1}, \dots, \mathcal{F}_1$ over the interval $\frac{\tau}{2}$.

Of course, steps 2. and 3. together are equivalent to one step of the (implicit) trapezoidal rule for $\mathcal{F}_s$ over the interval τ. Replacing this trapezoidal rule by Heun's method (which is essentially an explicit version of the trapezoidal rule) is exactly what we will do in the next section to derive the *implicit-explicit trapezoidal* (IMEXT) method. More specifically, the equations for v_s and v_{s+1} above will be replaced by

$$\begin{aligned} v_s &= v_{s-1} + \tau\mathcal{F}_s(t_n, v_{s-1}), \\ v_{s+1} &= v_{s-1} + \frac{\tau}{2}\left(\mathcal{F}_s(t_n, v_{s-1}) + \mathcal{F}_s(t_{n+1}, v_s)\right). \end{aligned}$$

The advantage is that the function $\mathcal{F}_s$ does not have to be treated implicitly anymore, but explicitly, which makes it (besides some other benefits which are explained later) computationally attractive. Note that only the steps involving $\mathcal{F}_s$ are altered, whereas the other steps of the trapezoidal splitting method remain unchanged for IMEXT.

The way the solutions of the individual subproblems were combined above is motivated by the construction of (second-order) splitting methods, see e.g. [43, Sec. II.5] or [78] for an introduction. For a thorough discussion of trapezoidal splitting methods in particular, we refer to [57, Sec. IV.2.3].

Remark 4.4.1. It was explained in [56] why the usage of a "midpoint splitting method" (defined analogously, but with the implicit and explicit Euler steps interchanged) is usually a bad idea. The reason is (roughly speaking) that more explicit than implicit steps are applied to some of the internal stages and thus the resulting method is much more prone to stability issues. See also [57, Rem. IV.2.3] for a more compact discussion of this issue. ◇

We continue with the presentation and analysis of the IMEXT method in the PDE setting, which is then applied to equations of the form (4.3) or (4.1) afterwards.

It should be noted that the time integrator presented in the next section is not a crucial choice for the application of a multi-level stochastic collocation method. Other time integrators may be used in a multi-level approach, too, as long as they satisfy the abstract requirements presented in Section 3.4 earlier. As usual, the choice of integrator should be dictated by the problem under consideration.

4.5 The implicit-explicit trapezoidal method (IMEXT)

We now describe the numerical method we employ to solve our model problem (4.3) in practice. To simplify the presentation, we look at an abstract version of problem (4.3) and forget about the specific value of $z \in \Sigma$, too. Thus, we consider

$$\partial_t u(t) = \mathcal{A}u(t) + f(u(t)) + g(t), \qquad t \geq 0, \tag{4.11a}$$

$$u(0) = u_0 \in \mathcal{D}(\mathcal{A}) \tag{4.11b}$$

and assume that $\mathcal{A}$ is a linear but, in general, unbounded operator. Note that (4.3) without the variable z can be cast into the form (4.11), at least if the homogeneous boundary conditions are incorporated in the domain $\mathcal{D}(\mathcal{A})$ of the operator $\mathcal{A}$ (which is the setting we have in mind). Since the purpose of this section is to describe the basic idea of the method, we do not rigorously keep track of the requirements for the method to be well-defined. This will be done later, where more assumptions on $\mathcal{A}, f, g$ and u_0 will be given.

In many situations it is advantageous to split an equation like (4.11) into several parts and solve these parts in an alternating fashion. If we regard the term "$\mathcal{A}u(t)$" as first and the non-linearity and inhomogeneity "$f(u(t)) + g(t)$" as second part, we arrive at the two – usually simpler – problems

$$\partial_t v(t) = \mathcal{A}v(t), \qquad t \geq 0, \qquad\qquad v(0) = v_0$$

and

$$\partial_t w(t) = f(w(t)) + g(t), \qquad t \geq 0, \qquad\qquad w(0) = w_0.$$

In the notation of the previous section, these subproblems are related to $\mathcal{F}_1$ and $\mathcal{F}_2$ via $\mathcal{F}_1(t, v(t)) = \mathcal{A}v(t)$ and $\mathcal{F}_2(t, w(t)) = f(w(t)) + g(t)$. If we solve these two subproblems one after another with the IMEXT approach outlined in Section 4.4 (with $s = 2$), we obtain an approximation to the solution of the "full" problem (4.11). We arrive at the implicit-explicit trapezoidal (IMEXT) method for (4.11),

$$\begin{aligned} u_n^+ &= \left(I + \frac{\tau}{2}\mathcal{A}\right) u_n, \\ u_{n+1} &= \left(I - \frac{\tau}{2}\mathcal{A}\right)^{-1} \left[u_n^+ + \frac{\tau}{2}\left(f(u_n^+) + f(u_n^+ + \tau f(u_n^+)) + g(t_n) + g(t_{n+1})\right)\right], \end{aligned}$$

which successively computes approximations $u_n \approx u(t_n)$ at times $t_n = n\tau$ with $n \in \mathbb{N}_0$ for given u_0 and a temporal step-size $\tau > 0$.

The linear part is treated implicitly since we think of $\mathcal{A}$ as being a differential operator (or its spatial discretisation) and thus intrinsically stiff. Since the f- and g-parts are treated explicitly, the method

is not unconditionally stable and thus suffers from a step-size restriction. A rather crude bound on the usable step-sizes for this method is given later in the proof of Theorem 4.5.6.

The "classical" order of this method is two, but a convergence analysis in our PDE setting is not obvious and requires a careful inspection of the problem and the method. We provide a proof for second-order convergence in the following subsections.

For later reference, we write down a version of the IMEXT method where we keep track of $z \in \Sigma$ if $\mathcal{A}$, f, g and u_0 depend on the parameter z, too. In this case,

$$u_n^+(z) = \left(I + \frac{\tau}{2}\mathcal{A}(z)\right) u_n(z), \tag{4.12a}$$

$$u_{n+1}(z) = \left(I - \frac{\tau}{2}\mathcal{A}(z)\right)^{-1} \left[u_n^+(z) + \frac{\tau}{2}\left(f(u_n^+(z), z) + f(u_n^+(z) + \tau f(u_n^+(z), z), z)\right) + \frac{\tau}{2}\left(g(t_n, z) + g(t_{n+1}, z)\right) \right]. \tag{4.12b}$$

Before we go into details of the error analysis, let us discuss the potential benefits of the IMEXT method in some applications, starting with the predator-prey system from the introduction.

Remark 4.5.1. The purpose of this remark is to explain why using a *splitting method* is important for problems such as (4.1). Typically, the reaction terms of predator-prey systems such as (4.1) are local in the sense that the dynamics at each spatial point x are decoupled if the diffusion part is absent. An example for this was already given in the introduction of this chapter, see (4.2). Thus, solving the non-linear part implicitly only requires the solution of a non-linear system in which the size of the system corresponds to the number of *species* (which is two if one predator and one prey species are considered) at each spatial grid point. This is relatively cheap and thus it is advantageous to employ a method with separates the diffusion and reaction parts. However, if we use a splitting method, it is not really required for these PDE systems that the non-linear part is treated explicitly so far – at least not to reduce the computational cost. ◇

Now that it is clear why we use a splitting method, let us explain why we treat the non-linear part explicitly. Often, the most important argument for using explicit methods is to avoid the solution of a large non-linear system in each time-step. As we pointed out in the previous example, this is not the important reason for us since the splitting approach allows us to decouple the large non-linear system into tiny non-linear systems in each spatial grid point.

For the model problems we have in mind, the non-linear part is treated explicitly for two reasons: The first one is ease of implementation and the second one is that it is not stiff compared to the linear part. Thus, an implicit treatment of the non-linear part should not be necessary.

For more general problems, the explicit treatment of the non-linear part might possibly be a good idea in one of the following situations:

- The derivative of the non-linear part is not availabile or too complicated to compute (analytically or numerically). For an implicit method, the derivative is usually required in order to perform some kind of Newton iteration.
- The problem is complicated by itself and a rather simple and less error-prone implementation is requested. Although this is not a crucial point from a theoretical point of view, it certainly is in complicated applications.

- A very accurate solution is required and thus tiny step-sizes have to be used. Then the step-size restriction coming from stability is usually not a problem anymore, so the necessity for solving the non-linear part implicitly disappears.

Now we begin with the analysis of the IMEXT method, starting with some preliminaries for later reference. The reader who is only interested in the assumptions and results of the error analysis may skip the following subsection and is directly referred to Section 4.5.2.

4.5.1 Preliminaries for the error analysis

We start with a discrete Gronwall lemma.

Lemma 4.5.2 (Discrete Gronwall lemma)**.** *Let $(e_n)_{n\in\mathbb{N}_0}$ and $(g_n)_{n\in\mathbb{N}_0}$ denote non-negative sequences and $a \geq 0$. If*

$$e_n \leq a + \sum_{k=0}^{n-1} g_k e_k \qquad \text{for all} \qquad n \in \mathbb{N}_0,$$

then also

$$e_n \leq a \prod_{k=0}^{n-1} (1 + g_k) \leq a \exp\left(\sum_{k=0}^{n-1} g_k\right).$$

In the special case where $g_k = b$ for all $k \in \mathbb{N}_0$, we obtain

$$e_n \leq a(1+b)^n \leq a \exp(bn), \qquad n \in \mathbb{N}_0.$$

As a service to the reader, we sketch a short proof of this result.

Proof. The case $n = 0$ is clear. The key observation in the induction step $n \rightsquigarrow n+1$ is that by successively extracting factors $1 + g_0,\ 1 + g_1,\ \ldots,\ 1 + g_{n-1}$ as follows

$$1 + \sum_{k=0}^{n} g_k \prod_{j=0}^{k-1} (1 + g_j) = (1 + g_0) \left[1 + \sum_{k=1}^{n} g_k \prod_{j=1}^{k-1} (1 + g_j) \right] = (1 + g_0)(1 + g_1) \left[1 + \sum_{k=2}^{n} g_k \prod_{j=2}^{k-1} (1 + g_j) \right],$$

we arrive at

$$1 + \sum_{k=0}^{n} g_k \prod_{j=0}^{k-1} (1 + g_j) = \prod_{j=0}^{n} (1 + g_j).$$

With this equation, the reasoning in the induction step is straightforward. □

We also need some other standard tools for the analysis of time integration schemes. These are the φ-functions defined in Appendix B. The most important properties are given there, too. They are very useful in the derivation of an error formula for the IMEXT method. Readers unfamiliar with φ-functions might want to have a look at Appendix B now, although we clearly indicate the usage of any of the properties from the Appendix where necessary.

For later reference, we collect some lesser known formulas for φ-functions. Again, the reader might skip this part and return to it later when the formulas are needed.

Lemma 4.5.3 (Midpoint approximation of the linear part)**.**
Let $\mathcal{A}$ be a sectorial operator with $\omega_{\mathcal{A}} \leq 0$. For $\tau > 0$, $a = \frac{\tau}{2}\mathcal{A}$ and $\alpha = (I-a)^{-1}$, we have

$$\left(I - \frac{\tau}{2}\mathcal{A}\right)^{-1}\left(I + \frac{\tau}{2}\mathcal{A}\right) - \mathrm{e}^{\tau\mathcal{A}} = \alpha(I+a) - \varphi_0(\tau\mathcal{A}) = \tau(\alpha - \varphi_1(\tau\mathcal{A}))\mathcal{A} \tag{4.13}$$

and

$$\alpha - \varphi_1(\tau\mathcal{A}) = \tau\alpha\mathcal{A}\left(\tfrac{1}{2}\varphi_0(\tau\mathcal{A}) - \varphi_1(\tau\mathcal{A})\right) + \tau\mathcal{A}(\varphi_1(\tau\mathcal{A}) - \varphi_2(\tau\mathcal{A})) \tag{4.14}$$

as operators defined on $\mathcal{D}(\mathcal{A})$.

Proof. The formula $\varphi_0(\tau\mathcal{A}) = I + \tau\mathcal{A}\varphi_1(\tau\mathcal{A})$ from (B.1) implies

$$\begin{aligned}\alpha(I+a) - \varphi_0(\tau\mathcal{A}) &= \alpha(I+a) - I - \tau\mathcal{A}\varphi_1(\tau\mathcal{A}) \\ &= \alpha[(I+a) - (I-a)] - \tau\varphi_1(\tau\mathcal{A})\mathcal{A} \\ &= \tau(\alpha - \varphi_1(\tau\mathcal{A}))\mathcal{A}.\end{aligned}$$

Note that we are allowed to interchange the order of $\mathcal{A}$ and $\varphi_1(\tau\mathcal{A})$ if we restrict ourselves to the domain $\mathcal{D}(\mathcal{A})$. This establishes (4.13).

From Lemma B.3 in Appendix B, we also obtain

$$\begin{aligned}\alpha - \varphi_1(\tau\mathcal{A}) &= \alpha - \varphi_0(\tau\mathcal{A}) + \tau\mathcal{A}(\varphi_1(\tau\mathcal{A}) - \varphi_2(\tau\mathcal{A})) \\ &= \alpha(I - \varphi_0(\tau\mathcal{A}) + a\varphi_0(\tau\mathcal{A})) + \tau\mathcal{A}(\varphi_1(\tau\mathcal{A}) - \varphi_2(\tau\mathcal{A})).\end{aligned}$$

Together with the recursion formula for φ-functions (B.1) in the form $I - \varphi_0(\tau\mathcal{A}) = -\tau\mathcal{A}\varphi_1(\tau\mathcal{A})$ we arrive at

$$\alpha - \varphi_1(\tau\mathcal{A}) = \tau\alpha\mathcal{A}\left(\tfrac{1}{2}\varphi_0(\tau\mathcal{A}) - \varphi_1(\tau\mathcal{A})\right) + \tau\mathcal{A}(\varphi_1(\tau\mathcal{A}) - \varphi_2(\tau\mathcal{A}))$$

and thus (4.14) holds. □

The following result was stated and proved in [40, Lem. 5.1].

Lemma 4.5.4 (Resolvent smoothing)**.** *Let $\mathcal{A}$ be sectorial with $\omega_{\mathcal{A}} < 0$ and $0 \leq \gamma \leq 1$. Then there exist constants $\omega_{\mathcal{A}} < \omega_1 < 0$ and $C > 0$ such that*

$$\|(-\mathcal{A})^{\gamma}(I - \tau\mathcal{A})^{-n}\|_{\mathcal{L}(\mathcal{X})} \leq C\frac{\mathrm{e}^{\omega_1 t_n}}{t_n^{\gamma}} \leq \frac{C}{t_n^{\gamma}}$$

holds for $t_n = n\tau$ whenever $\tau > 0$ and $n \in \mathbb{N}$. Taking $n = 1$ and $\gamma = 1$, we obtain

$$\|\mathcal{A}(I - \tau\mathcal{A})^{-1}\|_{\mathcal{L}(\mathcal{X})} \leq C\frac{\mathrm{e}^{\omega_1\tau}}{\tau} \leq \frac{C}{\tau}.$$

After these preliminaries, we are now ready for the error analysis of the IMEXT method.

4.5.2 Error analysis

Consider again the abstract Cauchy problem from (4.11). The IMEXT method that we analyse here is given by

$$u_n^{+} = \left(I + \frac{\tau}{2}\mathcal{A}\right)u_n, \tag{4.15a}$$

$$u_{n+1} = \left(I - \frac{\tau}{2}\mathcal{A}\right)^{-1}\left[u_n^{+} + \frac{\tau}{2}\left(f(u_n^{+}) + f(u_n^{+} + \tau f(u_n^{+})) + g(t_n) + g(t_{n+1})\right)\right] \tag{4.15b}$$

for $n = 0, 1, \ldots$ and a step-size $\tau > 0$. We make the following assumptions.

Assumption C. (1) The operator $\mathcal{A}$ is sectorial with $\omega_{\mathcal{A}} < 0$.

(2) $f \in C^2(\mathcal{D}(\mathcal{A}), \mathcal{D}(\mathcal{A}))$ (This implies that f and f' are locally Lipschitz continuous.)

(3) The function f satisfies a "mixed" local Lipschitz property in $\mathcal{X}$: For any $r > 0$, there is a constant L_r^{mix} such that

$$\|f(v_2) - f(v_1)\|_{\mathcal{X}} \leq L_r^{\mathrm{mix}} \|v_2 - v_1\|_{\mathcal{X}} \tag{4.16}$$

for all $v_1, v_2 \in \mathcal{B}_{\mathcal{D}(\mathcal{A})}(0, r)$.

(4) $g \in C^2([0,T], \mathcal{X}) \cap C^1([0,T], \mathcal{D}(\mathcal{A}))$

(5) There is a strict solution u of (4.11) which has the regularity

$$u \in C^2([0,T], \mathcal{D}(\mathcal{A})) \cap C^1([0,T], \mathcal{D}(\mathcal{A}^2)).$$

In particular, we have $u_0 \in \mathcal{D}(\mathcal{A}^2)$.

(6) For u from (5), $f(u(\cdot)) + g(\cdot) \in C^1([0,T], \mathcal{D}(\mathcal{A}^2))$.

Observe that the above assumptions imply that the method (4.15) is indeed well-defined.

Example 4.5.5.

For the model problem (4.3), we discussed in Section 4.2.1 that (1) is satisfied, probably after some rescaling explained in Remark 4.2.7. Assumptions (2) and (3) are often fulfilled for polynomials f. This follows from the product rule for Sobolev functions and the Sobolev embedding theorem for domains. The only problem is the homogeneous boundary condition dictated by the operator B from (4.5). If the function α in (4.5) is equally zero, then all polynomials f satisfy (2) and (3). For $\alpha \neq 0$, the situation is more complicated. The regularity of the solution u in (5) can be replaced by requirements on f, g and u_0 as follows. If f and g are sufficiently smooth, then the compatibility conditions

$$u_0 \in \mathcal{D}(\mathcal{A}^3), \quad \mathcal{A}u_0 + f(u_0) + g(0) \in \mathcal{D}(\mathcal{A}^3), \quad (\mathcal{A} + f'(u_0))(\mathcal{A}u_0 + f(u_0) + g(0)) + g'(0) \in \mathcal{D}(\mathcal{A}^2)$$

are sufficient to obtain

$$u \in \bigcap_{j=0}^{2} C^j([0,T], \mathcal{D}(\mathcal{A}^{3-j})),$$

which is clearly more than (5) demands, see Example 4.3.4 from before for details. (Example 4.3.4 treated the model problem (4.3), but an analogous statement holds for (4.11), too.) The remaining two points (4) and (6) are essentially compatibility conditions for the inhomogeneity g.

We stress here that the semigroup generated by $\mathcal{A}$ is not assumed to be contractive or quasi-contractive in Assumption C since this requirement would be too strict for our model problem (4.3) in the case $z \in \Sigma \subseteq \mathbb{C}^d$, as already mentioned in the comment after Remark 4.2.7.

The only goal of this section is to show the following theorem. To the best of our knowledge, this theorem has not been stated or proved in the literature yet.

Theorem 4.5.6 (Error bound for IMEXT)**.** *Under Assumption C, there exists $\tau_0 > 0$ such that the error of the* IMEXT *method (4.15) after n steps with initial value $u(0) = u_0$ and step-size $\tau \in (0, \tau_0]$ is bounded by*

$$\|u_n - u(t_n)\|_{\mathcal{X}} \leq C\tau^2 \tag{4.17}$$

for $0 \leq t_n = n\tau \leq T$ and a constant $C > 0$ which depends on the solution u on $[0, T]$, but is independent of τ and n.

The proof of this result and the formulation of a parametric version will occupy the remainder of this section and requires several preparations. If the reader is not primarily interested in the proof of this convergence result, we recommend to skip the very lengthy remainder of this section. Our advise is to continue reading in Section 4.5.3 where the theorem above is verified numerically.

Remark 4.5.7. For parabolic problems, many results which discuss convergence or stability of specific time integrators can be found in the literature, see e.g. [2, 91] for implicit-explicit methods, [70] for linearly-implicit methods, [53] for exponential integrators, [41, 71, 92] for general Runge-Kutta and/or multi-step methods, and [26] for the Peaceman-Rachford scheme, just to name a few. An error analysis for the IMEXT scheme discussed here seems to be missing so far. ◇

In the following, we replace the graph norms of the spaces $\mathcal{D}(\mathcal{A})$ and $\mathcal{D}(\mathcal{A}^2)$ by the equivalent norms

$$\begin{aligned} \|v\|_{\mathcal{D}(\mathcal{A})} &= \|\mathcal{A}v\|_{\mathcal{X}}, & v &\in \mathcal{D}(\mathcal{A}), \\ \|v\|_{\mathcal{D}(\mathcal{A}^2)} &= \|\mathcal{A}^2 v\|_{\mathcal{X}}, & v &\in \mathcal{D}(\mathcal{A}^2). \end{aligned}$$

These two definitions give indeed norms since $0 \in \varrho(\mathcal{A})$ by Assumption C(1) and thus $\mathcal{A}\colon \mathcal{D}(\mathcal{A}) \to \mathcal{X}$ is invertible. We indicate the norm in the big O notation as an index, so $\mathcal{O}_Y(\tau^j)$ denotes a function which is $\mathcal{O}(\tau^j)$ as $\tau \to 0$ in the norm of the space Y.

We start the error analysis with the local error, which is the error after performing just one step of the IMEXT method (with exact starting value). To this end, we introduce the numerical flow Φ_{τ,t_n} defined by

$$\Phi_{\tau,t_n}(u_n) = u_{n+1}, \qquad n \in \mathbb{N}_0,$$

where u_n and u_{n+1} are related via (4.15). For later reference, we also define powers of the numerical flow as

$$\Phi^m_{\tau,t_n}(v) = \Phi_{\tau,t_{n+m-1}}(\Phi^{m-1}_{\tau,t_n}(v)), \qquad n, m = 1, 2, \ldots, \qquad \Phi^0_{\tau,t_n}(v) = v,$$

with the special case

$$\Phi^m_{\tau}(v) = \Phi^m_{\tau,0}(v), \qquad m = 1, 2, \ldots \tag{4.18}$$

For the local error of the IMEXT method in the norm of $\mathcal{X}$, we have the following result.

Proposition 4.5.8 (Local error in $\mathcal{X}$)**.** *Under Assumption C, the error of the* IMEXT *method (4.15) after one step with step-size $\tau > 0$ is bounded by*

$$\|\Phi_{\tau,t_n}(u(t_n)) - u(t_{n+1})\|_{\mathcal{X}} \leq C\tau^3 \tag{4.19}$$

for a constant $C > 0$ which is independent of τ and n.

Before the proof, we set up some notation similar to [91]. In this work, a similar error bound for a dimension splitting method with two linear operators A and B was established. Our setting here roughly corresponds to $B = 0$ in their work. Another important difference is that the (typically non-linear) function f from our setting is absent in their work, which makes our situation in some sense more complicated. We will come back to [91] now and then in our proof. Let us abbreviate

$$a = \frac{\tau}{2}\mathcal{A}, \qquad \alpha = (I - a)^{-1},$$

as in Lemma 4.5.3. Moreover, we set

$$u_n = u(t_n) \qquad \text{and} \qquad u_n^+ = (I + a)u_n \tag{4.20}$$

for convenience, the latter being the result after one half-step with the explicit Euler method starting with u_n. We also use the abbreviation

$$\mathcal{F}(t, u) = f(u) + g(t).$$

With this notation, (4.15) can be written as a single formula

$$u_{n+1} = \alpha\left(u_n^+ + \frac{\tau}{2}\left(\mathcal{F}(t_n, u_n^+) + \mathcal{F}(t_{n+1}, u_n^+ + \tau\mathcal{F}(t_n, u_n^+))\right)\right).$$

Proof of Proposition 4.5.8. Using the function $\widetilde{\mathcal{F}}(h) = \mathcal{F}(t_n + h, u_n^+ + h\mathcal{F}(t_n, u_n^+))$, we may expand

$$\begin{aligned}\mathcal{F}(t_{n+1}, u_n^+ + \tau\mathcal{F}(t_n, u_n^+)) &= \mathcal{F}(t_n, u_n^+) + \tau\partial_u\mathcal{F}(t_n, u_n^+)\mathcal{F}(t_n, u_n^+) + \tau\partial_t\mathcal{F}(t_n, u_n^+) + r_n^{(1)}(\tau)\\ &= f(u_n^+) + g(t_n) + \tau f'(u_n^+)[f(u_n^+) + g(t_n)] + \tau g'(t_n) + r_n^{(1)}(\tau),\end{aligned}$$

where

$$r_n^{(1)}(\tau) = \int_0^\tau (\tau - \sigma)\widetilde{\mathcal{F}}''(\sigma)\mathrm{d}\sigma.$$

Setting $R_n^{(1)} = \frac{\tau}{2}\alpha r_n^{(1)}(\tau)$, we get

$$u_{n+1} = \alpha\left(u_n^+ + \tau[f(u_n^+) + g(t_n)] + \frac{\tau^2}{2}[f'(u_n^+)[f(u_n^+) + g(t_n)] + g'(t_n)]\right) + R_n^{(1)}.$$

Observe that $R_n^{(1)} = \mathcal{O}_{\mathcal{X}}(\tau^3)$. For the solution itself, the variation-of-constants formula yields

$$u(t_{n+1}) = \mathrm{e}^{\tau\mathcal{A}}u(t_n) + \int_0^\tau \mathrm{e}^{(\tau-s)\mathcal{A}}[f(u(t_n + s)) + g(t_n + s)]\mathrm{d}s.$$

Taylor expansion of the function

$$h_n(s) = f(u(t_n + s)) + g(t_n + s) \tag{4.21}$$

shows that

$$u(t_{n+1}) = \mathrm{e}^{\tau\mathcal{A}}u(t_n) + \int_0^\tau \mathrm{e}^{(\tau-s)\mathcal{A}}\left[f(u(t_n)) + g(t_n) + s[f'(u(t_n))u'(t_n) + g'(t_n)] + \int_0^s (s - \sigma)h_n''(\sigma)\mathrm{d}\sigma\right]\mathrm{d}s.$$

Using the definition of the φ-functions $\varphi_j = \varphi_j(\tau\mathcal{A})$ (see Definition B.1), we arrive at

$$u(t_{n+1}) = \varphi_0 u(t_n) + \tau\varphi_1[f(u(t_n)) + g(t_n)] + \tau^2\varphi_2[f'(u(t_n))u'(t_n) + g'(t_n)] + R_n^{(2)}$$

with

$$R_n^{(2)} = \int_0^\tau \mathrm{e}^{(\tau-s)\mathcal{A}} \left(\int_0^s (s-\sigma) h_n''(\sigma) \mathrm{d}\sigma \right) \mathrm{d}s.$$

The error $e_{n+1} = u_{n+1} - u(t_{n+1})$ is now given by

$$\begin{aligned}
e_{n+1} &= \alpha(I+a)e_n + [\alpha(I+a) - \varphi_0]\, u(t_n) + \tau \left[\alpha f(u_n^+) - \varphi_1 f(u(t_n))\right] \\
&\quad + \tau^2 \left[\frac{1}{2}\alpha f'(u_n^+)[f(u_n^+) + g(t_n)] - \varphi_2 f'(u(t_n))u'(t_n)\right] \\
&\quad + \tau(\alpha - \varphi_1) g(t_n) + \tau^2 \left[\frac{1}{2}\alpha - \varphi_2\right] g'(t_n) + R_n^{(1)} - R_n^{(2)} \\
&= \alpha(I+a)e_n + [\alpha(I+a) - \varphi_0]\, u(t_n) + \tau\alpha \left[f(u_n^+) - f(u(t_n))\right] + \tau[\alpha - \varphi_1] f(u(t_n)) \\
&\quad + \frac{\tau^2}{2}\alpha \left[f'(u_n^+)[f(u_n^+) + g(t_n)] - f'(u(t_n))u'(t_n)\right] + \tau^2 \left[\frac{1}{2}\alpha - \varphi_2\right] f'(u(t_n))u'(t_n) \\
&\quad + \tau(\alpha - \varphi_1) g(t_n) + \tau^2 \left[\frac{1}{2}\alpha - \varphi_2\right] g'(t_n) + R_n^{(1)} - R_n^{(2)}.
\end{aligned}$$

By (4.20), $u(t_n) = u_n$ and thus $e_n = 0$. The terms with differences of f or f' with different arguments are new and do not appear in [91]. They deserve a special treatment later on. The other terms are almost the same as in [91, Eq. (31) ff.] (although with a different notation). Following the strategy in this article, we can simplify the term

$$\begin{aligned}
E_n(\tau) &\coloneqq [\alpha(I+a) - \varphi_0]\, u(t_n) + \tau\, [\alpha - \varphi_1]\, f(u(t_n)) + \tau^2 \left[\frac{1}{2}\alpha - \varphi_2\right] f'(u(t_n))u'(t_n) \\
&\quad + \tau(\alpha - \varphi_1) g(t_n) + \tau^2 \left[\frac{1}{2}\alpha - \varphi_2\right] g'(t_n) \\
&= [\alpha(I+a) - \varphi_0]\, u(t_n) + \tau\, [\alpha - \varphi_1]\, h_n(0) + \tau^2 \left[\frac{1}{2}\alpha - \varphi_2\right] h_n'(0)
\end{aligned}$$

with h_n from (4.21) as follows: First, we use $I = \alpha - \alpha a$ and get

$$\begin{aligned}
E_n(\tau) &= [\alpha(I+a) - (\alpha - \alpha a)\varphi_0]\, u(t_n) + \tau\, [\alpha - (\alpha - \alpha a)\varphi_1]\, h_n(0) + \tau^2 \left[\frac{1}{2}\alpha - (\alpha - \alpha a)\varphi_2\right] h_n'(0) \\
&= \alpha\, [I + a - \varphi_0 + a\varphi_0]\, u(t_n) + \tau\alpha\, [I - \varphi_1 + a\varphi_1]\, h_n(0) + \tau^2\alpha \left[\frac{1}{2} I - \varphi_2 + a\varphi_2\right] h_n'(0).
\end{aligned}$$

With $\frac{1}{2}I - \varphi_2 = -\tau\mathcal{A}\varphi_3$ and $\varphi_0 = I + \tau\mathcal{A}\varphi_1$ on $\mathcal{X}$ (see (B.1) for $j = 3$ and $j = 1$), we obtain

$$\begin{aligned}
\frac{1}{2}I - \varphi_2 + a\varphi_2 &= 2a \left(\frac{1}{2}\varphi_2 - \varphi_3\right), \\
I + a - \varphi_0 &= 2a \left(\frac{1}{2} I - \varphi_1\right),
\end{aligned}$$

which then yields

$$\begin{aligned}
E_n(\tau) &= \tau\alpha\mathcal{A} \left[\frac{1}{2}I - \varphi_1 + \frac{1}{2}\varphi_0\right] u(t_n) + \tau\alpha\, [I - \varphi_1 + a\varphi_1]\, h_n(0) + \tau^2 \alpha 2a \left[\frac{1}{2}\varphi_2 - \varphi_3\right] h_n'(0) \\
&= \tau\alpha \left[\frac{1}{2} I - \varphi_1\right] (\mathcal{A}u(t_n) + h_n(0)) + \frac{\tau}{2}\alpha\, [(I + \tau\mathcal{A}\varphi_1)\mathcal{A}u(t_n) + (I + \tau\mathcal{A}\varphi_1)h_n(0)] \\
&\quad + \tau^2 \alpha 2a \left[\frac{1}{2}\varphi_2 - \varphi_3\right] h_n'(0) \\
&= \tau\alpha \left[I - \varphi_1 + \frac{\tau}{2}\mathcal{A}\varphi_1\right] u'(t_n) + \tau^2 \alpha 2a \left[\frac{1}{2}\varphi_2 - \varphi_3\right] h_n'(0)
\end{aligned}$$

because u solves (4.11). Due to $I - \varphi_1 = -\tau\mathcal{A}\varphi_2$ (see (B.2) for $j = 2$), we get

$$I - \varphi_1 + \frac{\tau}{2}\mathcal{A}\varphi_1 = \tau\mathcal{A}\left[\frac{1}{2}\varphi_1 - \varphi_2\right] = \tau\mathcal{A}\left[\frac{1}{2}(\varphi_1 - I) + \frac{1}{2}I - \varphi_2\right] = (\tau\mathcal{A})^2\left[\frac{1}{2}\varphi_2 - \varphi_3\right]$$

and

$$u''(t_n) = \mathcal{A}u'(t_n) + h_n'(0) = \mathcal{A}^2 u(t_n) + \mathcal{A}h_n(0) + h_n'(0).$$

With the last two formulas, we arrive at

$$\begin{aligned} E_n(\tau) &= \tau^3\alpha\left[\frac{1}{2}\varphi_2 - \varphi_3\right]\mathcal{A}^2 u'(t_n) + \tau^2\alpha 2a\left[\frac{1}{2}\varphi_2 - \varphi_3\right]h_n'(0) \\ &= \tau^3\alpha\left[\frac{1}{2}\varphi_2 - \varphi_3\right]\mathcal{A}\left(\mathcal{A}^2 u(t_n) + \mathcal{A}h_n(0) + h_n'(0)\right) \\ &= \tau^3\alpha\left[\frac{1}{2}\varphi_2 - \varphi_3\right]\mathcal{A}u''(t_n), \end{aligned}$$

which is the final formula for $E_n(\tau)$. For e_{n+1}, we infer

$$\begin{aligned} e_{n+1} = \tau^3\alpha\left[\frac{1}{2}\varphi_2 - \varphi_3\right]\mathcal{A}u''(t_n) + R_n^{(1)} - R_n^{(2)} \\ + \tau\alpha\left[f(u_n^+) - f(u(t_n))\right] + \frac{\tau^2}{2}\alpha\left[f'(u_n^+)[f(u_n^+) + g(t_n)] - f'(u(t_n))u'(t_n)\right]. \end{aligned} \tag{4.22}$$

The difference $f(u_n^+) - f(u(t_n))$ has to be treated by another Taylor expansion for the function

$$\widetilde{f}(\tau) = f(u_n + \tfrac{\tau}{2}\mathcal{A}u_n),$$

which reads

$$\widetilde{f}(\tau) = \widetilde{f}(0) + \tau\widetilde{f}'(0) + r_n^{(3)}(\tau), \qquad r_n^{(3)}(\tau) = \int_0^\tau (\tau - \sigma)\widetilde{f}''(\sigma)\mathrm{d}\sigma,$$

or equivalently

$$f(u_n^+) = f(u_n) + \frac{\tau}{2}f'(u_n)\mathcal{A}u_n + r_n^{(3)}(\tau)$$

with $r_n^{(3)}(\tau) = \mathcal{O}_{\mathcal{X}}(\tau^2)$. In (4.22), we have the term $\frac{\tau^2}{2}\alpha f'(u_n^+)[f(u_n^+) + g(t_n)]$, too, which is basically $\frac{\tau^2}{2}\alpha f'(u_n)[f(u_n) + g(t_n)] + \mathcal{O}_{\mathcal{X}}(\tau^3)$. Using this and $u(t_n) = u_n$, we get

$$\begin{aligned} &\tau\alpha\left[f(u_n^+) - f(u(t_n))\right] + \frac{\tau^2}{2}\alpha\left[f'(u_n^+)[f(u_n^+) + g(t_n)] - f'(u(t_n))u'(t_n)\right] \\ &= \frac{\tau^2}{2}\alpha f'(u_n)\left[\mathcal{A}u_n + f(u_n) + g(t_n)\right] - \frac{\tau^2}{2}\alpha f'(u(t_n))u'(t_n) + \mathcal{O}_{\mathcal{X}}(\tau^3). \end{aligned}$$

The $\mathcal{O}_{\mathcal{X}}(\tau^2)$ terms on the right-hand side cancel due to $u_n = u(t_n)$ and $u'(t_n) = \mathcal{A}u(t_n) + f(u(t_n)) + g(t_n)$. If we now use $R_n^{(j)} = \mathcal{O}_{\mathcal{X}}(\tau^3)$ for $j = 1, 2$, we can identify the local error (4.22) as $\mathcal{O}_{\mathcal{X}}(\tau^3)$. □

Now that we have a local error bound in $\mathcal{X}$, we continue by proving an error bound in the stronger norm of $\mathcal{D}(\mathcal{A})$. As usual, we loose one power of τ here.

Proposition 4.5.9 (Local error in $\mathcal{D}(\mathcal{A})$)**.** *Under Assumption C, the error of the* IMEXT *method* (4.15) *after one step with step-size $\tau > 0$ is bounded by*

$$\|\Phi_{\tau,t_n}(u(t_n)) - u(t_{n+1})\|_{\mathcal{D}(\mathcal{A})} \leq C\tau^2$$

for a constant $C > 0$ which is independent of τ and n.

Proof. Using again the function $\widetilde{\mathcal{F}}(h) = \mathcal{F}(t_n + h, u_n^+ + h\mathcal{F}(t_n, u_n^+))$, we may expand

$$\mathcal{F}(t_{n+1}, u_n^+ + \tau\mathcal{F}(t_n, u_n^+)) = \mathcal{F}(t_n, u_n^+) + r_n^{(1)}(\tau) = f(u_n^+) + g(t_n) + r_n^{(1)}(\tau),$$

where

$$r_n^{(1)}(\tau) = \int_0^\tau \widetilde{\mathcal{F}}'(\sigma)\mathrm{d}\sigma.$$

The derivative of $\widetilde{\mathcal{F}}$ is

$$\begin{aligned}\widetilde{\mathcal{F}}'(\sigma) &= \partial_t\mathcal{F}(t_n + \sigma, u_n^+ + \sigma\mathcal{F}(t_n, u_n^+)) + \partial_u\mathcal{F}(t_n + \sigma, u_n^+ + \sigma\mathcal{F}(t_n, u_n^+))\mathcal{F}(t_n, u_n^+)\\ &= g'(t_n + \sigma) + f'(u_n^+ + \sigma\mathcal{F}(t_n, u_n^+))(f(u_n^+) + g(t_n))\end{aligned}$$

and bounded in $\mathcal{D}(\mathcal{A})$ for $g \in C^1([0,T], \mathcal{D}(\mathcal{A}))$ and $u \in C([0,T], \mathcal{D}(\mathcal{A}^2))$. Thus,

$$u_{n+1} = \alpha\left(u_n^+ + \tau f(u_n^+) + \tau g(t_n)\right) + R_n^{(1)} \qquad \text{with} \qquad R_n^{(1)} = \tfrac{\tau}{2}\alpha r_n^{(1)}(\tau) = \mathcal{O}_{\mathcal{D}(\mathcal{A})}(\tau^2). \tag{4.23}$$

The solution u is expanded as

$$u(t_{n+1}) = \varphi_0 u(t_n) + \tau\varphi_1\left[f(u(t_n)) + g(t_n)\right] + R_n^{(2)},$$

where

$$R_n^{(2)} = \int_0^\tau \mathrm{e}^{(\tau-s)\mathcal{A}} \int_0^s f'(u(t_n + \sigma))u'(t_n + \sigma) + g'(t_n + \sigma)\mathrm{d}\sigma\mathrm{d}s, \tag{4.24}$$

which is $\mathcal{O}_{\mathcal{D}(\mathcal{A})}(\tau^2)$ for $g \in C^1([0,T], \mathcal{D}(\mathcal{A}))$ and $u \in C^1([0,T], \mathcal{D}(\mathcal{A}))$. Let $e_{n+1} = u_{n+1} - u(t_{n+1})$. We obtain

$$\begin{aligned}e_{n+1} = \alpha(I + a)e_n &+ (\alpha(I + a) - \varphi_0)u(t_n) + \tau\alpha\left[f(u_n^+) - f(u(t_n))\right] + \tau[\alpha - \varphi_1]f(u(t_n))\\ &+ \tau[\alpha - \varphi_1]g(t_n) + R_n^{(1)} - R_n^{(2)}\end{aligned}$$

with $e_n = 0$ according to (4.20). By (4.13), we have

$$(\alpha(I + a) - \varphi_0)u(t_n) = \tau(\alpha - \varphi_1)\mathcal{A}u(t_n)$$

and hence the error formula becomes

$$e_{n+1} = \tau[\alpha - \varphi_1]u'(t_n) + \tau\alpha\left[f(u_n^+) - f(u(t_n))\right] + \mathcal{O}_{\mathcal{D}(\mathcal{A})}(\tau^2). \tag{4.25}$$

Now we need a Taylor expansion for the function $\widetilde{f}(\tau) = f(u_n + \frac{\tau}{2}\mathcal{A}u_n)$, which reads

$$\widetilde{f}(\tau) = \widetilde{f}(0) + r_n^{(3)}(\tau), \qquad r_n^{(3)}(\tau) = \int_0^\tau \widetilde{f}'(\sigma)\mathrm{d}\sigma$$

with

$$\widetilde{f}'(\sigma) = \frac{1}{2}f'(u_n + \tfrac{\sigma}{2}\mathcal{A}u_n)\mathcal{A}u_n,$$

or equivalently

$$f(u_n^+) = f(u_n) + r_n^{(3)}(\tau)$$

with $r_n^{(3)}(\tau) = \mathcal{O}_{\mathcal{D}(\mathcal{A})}(\tau)$.

The other term in (4.25) can be treated by (4.14), which gives

$$(\alpha - \varphi_1)u'(t_n) = \tau\alpha\mathcal{A}\left(\tfrac{1}{2}\varphi_0 - \varphi_1\right)u'(t_n) + \tau\mathcal{A}(\varphi_1 - \varphi_2)u'(t_n).$$

This term is $\mathcal{O}_{\mathcal{D}(\mathcal{A})}(\tau)$, since $u \in C^1([0,T], \mathcal{D}(\mathcal{A}^2))$.

From (4.25), we now conclude $e_{n+1} = \mathcal{O}_{\mathcal{D}(\mathcal{A})}(\tau^2)$. □

To establish a bound on the norm of the numerical solution in $\mathcal{D}(\mathcal{A})$ – at least under a step-size restriction – we use the local error of order 2 in the norm of $\mathcal{D}(\mathcal{A})$ shown above, but also a local error of order 1 in the norm of $\mathcal{D}(\mathcal{A}^2)$. This result is established in the next proposition.

Proposition 4.5.10 (Local error in $\mathcal{D}(\mathcal{A}^2)$)**.** *Under Assumption C, the error of the* IMEXT *method* (4.15) *after one step with step-size $\tau > 0$ is bounded by*

$$\|\Phi_{\tau,t_n}(u(t_n)) - u(t_{n+1})\|_{\mathcal{D}(\mathcal{A}^2)} \leq C\tau$$

for a constant $C > 0$ which is independent of τ and n.

Proof. The proof follows along the same lines as the proof for the local error in $\mathcal{D}(\mathcal{A})$, but with minor changes. For $R_n^{(1)} = \frac{\tau}{2}\alpha r_n^{(1)}(\tau)$ from (4.23), we get $R_n^{(1)} = \mathcal{O}_{\mathcal{D}(\mathcal{A}^2)}(\tau)$ by using Lemma 4.5.4:

$$\|R_n^{(1)}\|_{\mathcal{D}(\mathcal{A}^2)} = \|\mathcal{A}R_n^{(1)}\|_{\mathcal{D}(\mathcal{A})} = \frac{\tau}{2}\|\mathcal{A}\alpha r_n^{(1)}(\tau)\|_{\mathcal{D}(\mathcal{A})} \leq C\|r_n^{(1)}(\tau)\|_{\mathcal{D}(\mathcal{A})} = \mathcal{O}(\tau)$$

For the expansion of the solution u in $\mathcal{D}(\mathcal{A}^2)$, we need that u, $f(u(\cdot)) + g(\cdot) \in C([0,T], \mathcal{D}(\mathcal{A}^2))$ and that the (time) derivative of $f(u(\cdot)) + g(\cdot)$ is in $L^1([0,T], \mathcal{D}(\mathcal{A}^2))$. Then we get $R_n^{(2)} = \mathcal{O}_{\mathcal{D}(\mathcal{A}^2)}(\tau)$ with $R_n^{(2)}$ from (4.24) as follows:

$$\begin{aligned}\|R_n^{(2)}\|_{\mathcal{D}(\mathcal{A}^2)} &\leq C\int_0^\tau \int_0^s \|f'(u(t_n+r))u'(t_n+r) + g'(t_n+r)\|_{\mathcal{D}(\mathcal{A}^2)}\mathrm{d}r\mathrm{d}s \\ &\leq C\tau^2\|f(u(\cdot)) + g(\cdot)\|_{C^1([0,T],\mathcal{D}(\mathcal{A}^2))}\end{aligned}$$

Formula (4.25), Lemma 4.5.4 and the local Lipschitz continuity of f in $\mathcal{D}(\mathcal{A})$ thus imply

$$\begin{aligned}\|e_{n+1}\|_{\mathcal{D}(\mathcal{A}^2)} &\leq \|\tau[\alpha - \varphi_1]u'(t_n)\|_{\mathcal{D}(\mathcal{A}^2)} + \tau\|\alpha[f(u_n^+) - f(u(t_n))]\|_{\mathcal{D}(\mathcal{A}^2)} + \mathcal{O}(\tau) \\ &\leq C\tau\|u\|_{C^1([0,T],\mathcal{D}(\mathcal{A}^2))} + C\|f(u_n^+) - f(u(t_n))\|_{\mathcal{D}(\mathcal{A})} + \mathcal{O}(\tau) \\ &\leq C\left\|u_n + \tfrac{\tau}{2}\mathcal{A}u_n - u(t_n)\right\|_{\mathcal{D}(\mathcal{A})} + \mathcal{O}(\tau) \\ &= \mathcal{O}(\tau),\end{aligned}$$

where we have used $u(t_n) = u_n$ and $u \in C([0,T], \mathcal{D}(\mathcal{A}^2))$ in the last step. □

It can be seen from (4.15) for $f \equiv 0$ and $g \equiv 0$ that $[(I+a)\alpha]^n$ has to be uniformly bounded in $n \in \mathbb{N}$ to obtain a *stable* IMEXT method, again with the abbreviations

$$a = \frac{\tau}{2}\mathcal{A}, \qquad \alpha = (I-a)^{-1}.$$

Several stability results from the literature apply to our setting, such as [93, Thm. 1] and [74, Thm. 3.5]. Thus, we have indeed

$$C = \sup_{n\in\mathbb{N}_0} \|[(I+a)\alpha]^n\|_{\mathcal{L}(\mathcal{X})} < \infty. \tag{4.26}$$

It can be seen from [93, Thm. 1] that the constant C above depends only on the quantities M and ϑ of $\mathcal{A}$ in the definition of "sectorial".

It turns out that as long as the intermediate steps u_n^+ from (4.15a) remain bounded in $\mathcal{D}(\mathcal{A})$, the IMEXT method is stable. The precise statement is the following.

Proposition 4.5.11 (Stability). *Let*

$$\max_{j=0}^{n-1}\{\|u_j^+\|_{\mathcal{D}(\mathcal{A})}, \|v_j^+\|_{\mathcal{D}(\mathcal{A})}\} \le M_n,$$

where v_j^+ is defined as u_j^+ in (4.15a), but with initial value $v_0 \in \mathcal{D}(\mathcal{A}^2)$ instead of u_0. Moreover, let L_n and $\widetilde{L}_n$ denote the Lipschitz constants of f for the balls $\mathcal{B}_{\mathcal{D}(\mathcal{A})}(0, 2M_n)$ and $\mathcal{B}_{\mathcal{D}(\mathcal{A})}(0, 2M_n(1+\tau L_n))$. Similarly, let L_n^{mix} and $\widetilde{L}_n^{\text{mix}}$ denote the Lipschitz constants of f for the same $\mathcal{D}(\mathcal{A})$-balls, but where both norms are $\|\cdot\|_{\mathcal{X}}$ (see part (3) of Assumption C). Then we have the following inequalities

$$\|u_n^+ - v_n^+\|_{\mathcal{D}(\mathcal{A})} \le C \exp\left(C\widetilde{L}_n(2+\tau\widetilde{L}_n)n\tau/2\right)\|u_0^+ - v_0^+\|_{\mathcal{D}(\mathcal{A})}, \tag{4.27a}$$

$$\|u_n - v_n\|_{\mathcal{X}} \le \widetilde{C} \exp\left(\widetilde{C}\widetilde{L}_n^{\text{mix}}(2+\tau\widetilde{L}_n^{\text{mix}})n\tau/2\right)\|u_0^+ - v_0^+\|_{\mathcal{X}}, \tag{4.27b}$$

with constants $C, \widetilde{C} > 0$ independent of n.

Proof. The recursion formula (4.15) yields

$$\begin{aligned}
u_n^+ - v_n^+ &= (I+a)\alpha(u_{n-1}^+ - v_{n-1}^+) + \frac{\tau}{2}(I+a)\alpha(f(u_{n-1}^+) - f(v_{n-1}^+)) \\
&\quad + \frac{\tau}{2}(I+a)\alpha\left(f(u_{n-1}^+ + \tau f(u_{n-1}^+)) - f(v_{n-1}^+ + \tau f(v_{n-1}^+))\right) \\
&= [(I+a)\alpha]^n(u_0^+ - v_0^+) \\
&\quad + \frac{\tau}{2}\sum_{j=0}^{n-1}[(I+a)\alpha]^{n-j}\left(f(u_j^+) - f(v_j^+) + f(u_j^+ + \tau f(u_j^+)) - f(v_j^+ + \tau f(v_j^+))\right).
\end{aligned}$$

Utilising the bound

$$C = \sup_{n\in\mathbb{N}_0}\|[(I+a)\alpha]^n\|_{\mathcal{L}(\mathcal{X})} < \infty$$

from (4.26) (which holds in the norm of $\mathcal{L}(\mathcal{D}(\mathcal{A}))$, too), we obtain

$$\begin{aligned}
\|u_n^+ - v_n^+\|_{\mathcal{D}(\mathcal{A})} &\le C\|u_0^+ - v_0^+\|_{\mathcal{D}(\mathcal{A})} \\
&\quad + \frac{\tau}{2}C\sum_{j=0}^{n-1}\left\|f(u_j^+) - f(v_j^+)\|_{\mathcal{D}(\mathcal{A})} + \|f(u_j^+ + \tau f(u_j^+)) - f(v_j^+ + \tau f(v_j^+))\right\|_{\mathcal{D}(\mathcal{A})}.
\end{aligned}$$

With the definitions of L_n and $\widetilde{L}_n$, we arrive at

$$\|u_n^+ - v_n^+\|_{\mathcal{D}(\mathcal{A})} \le C\|u_0^+ - v_0^+\|_{\mathcal{D}(\mathcal{A})} + \frac{\tau}{2}C\sum_{j=0}^{n-1}\widetilde{L}_n(2+\tau\widetilde{L}_n)\|u_j^+ - v_j^+\|_{\mathcal{D}(\mathcal{A})}.$$

Now the discrete Gronwall lemma from Lemma 4.5.2 shows the first inequality.

For the proof of the second inequality (4.27b), consider

$$\begin{aligned}
\alpha^{-1}(u_n - v_n) &= (u_{n-1}^+ - v_{n-1}^+) + \frac{\tau}{2}(f(u_{n-1}^+) - f(v_{n-1}^+)) \\
&\quad + \frac{\tau}{2}(f(u_{n-1}^+ + \tau f(u_{n-1}^+)) - f(v_{n-1}^+ + \tau f(v_{n-1}^+))) \\
&= [(I+a)\alpha]^{n-1}(u_0^+ - v_0^+) \\
&\quad + \frac{\tau}{2}\sum_{j=0}^{n-1}[(I+a)\alpha]^{n-1-j}(f(u_j^+) - f(v_j^+) + f(u_j^+ + \tau f(u_j^+)) - f(v_j^+ + \tau f(v_j^+))),
\end{aligned}$$

which can be estimated using the mixed local Lipschitz property (4.16) to obtain

$$\|\alpha^{-1}(u_n - v_n)\|_{\mathcal{X}} \le C\|u_0^+ - v_0^+\|_{\mathcal{X}} + \frac{\tau}{2} C \sum_{j=0}^{n-1} \widetilde{L}_n^{\mathrm{mix}}(2 + \tau \widetilde{L}_n^{\mathrm{mix}})\|u_j^+ - v_j^+\|_{\mathcal{X}}.$$

The estimate

$$\|u_j^+ - v_j^+\|_{\mathcal{X}} = \|(I + a)\alpha\alpha^{-1}(u_j - v_j)\|_{\mathcal{X}} \le C\|\alpha^{-1}(u_j - v_j)\|_{\mathcal{X}}$$

with C from (4.26) now allows us to apply Lemma 4.5.2. It yields

$$\|\alpha^{-1}(u_n - v_n)\|_{\mathcal{X}} \le C \exp(C^2 \widetilde{L}_n^{\mathrm{mix}}(2 + \tau \widetilde{L}_n^{\mathrm{mix}}) n\tau/2)\|u_0^+ - v_0^+\|_{\mathcal{X}}. \tag{4.28}$$

By Lemma 4.5.4 for $\gamma = 0$ and $n = 1$, we have

$$\|u_n - v_n\|_{\mathcal{X}} \le \|\alpha\|_{\mathcal{L}(\mathcal{X})}\|\alpha^{-1}(u_n - v_n)\|_{\mathcal{X}} \le \widetilde{C}\|\alpha^{-1}(u_n - v_n)\|.$$

The second inequality (4.27b) now follows from (4.28). □

Now we can finally prove Theorem 4.5.6.

Proof of Theorem 4.5.6

In the following, an upper index "+" corresponds as before to a half-step with the explicit Euler method for $\mathcal{A}$, so $v^+ = (I + \frac{\tau}{2}\mathcal{A})v$. Moreover, we omit the time t_j from the notation Φ_{τ,t_j} and simply write Φ_τ. It is always clear from the context which time t_j is used.

Our first goal is to establish a bound on the approximations $\Phi_\tau^n(u(t_\ell))^+$ in the norm of $\mathcal{D}(\mathcal{A})$, uniformly in n and ℓ. This is the key ingredient which allows us to combine the local error and stability bounds later to arrive at a global error bound.

More precisely, we show by induction on n that there exists a step-size $\tau_0 > 0$ such that for all $n \in \mathbb{N}_0$ and $\ell \in \mathbb{N}_0$ with $(\ell + n)\tau \le T$, it holds that

$$\|\Phi_\tau^n(u(t_\ell))^+\|_{\mathcal{D}(\mathcal{A})} \le 2M^{(1)}, \qquad M^{(1)} = \max_{t\in[0,T]}(\|u(t)\|_{\mathcal{D}(\mathcal{A})} + \|u(t)\|_{\mathcal{D}(\mathcal{A}^2)}) \tag{4.29}$$

for all $\tau \le \tau_0$. This is immediately clear for $n = 0$ if we choose $\tau_0 \le 2$. Now assume that

$$\|\Phi_\tau^k(u(t_\ell))^+\|_{\mathcal{D}(\mathcal{A})} \le 2M^{(1)} \quad \text{for} \quad k = 0, \dots, n-1, \ \ell \in \mathbb{N}_0 \quad \text{with} \quad (\ell + k)\tau \le T.$$

Consider the telescoping sum

$$\Phi_\tau^n(u(t_\ell))^+ = u(t_{\ell+n})^+ + \sum_{j=0}^{n-1} \big(\Phi_\tau^{n-j}(u(t_{\ell+j}))^+ - \Phi_\tau^{n-j-1}(u(t_{\ell+j+1}))^+\big). \tag{4.30}$$

Combining the stability estimate (4.27a) with the induction hypothesis, we get

$$\|\Phi_\tau^{n-j}(u(t_{\ell+j}))^+ - \Phi_\tau^{n-j-1}(u(t_{\ell+j+1}))^+\|_{\mathcal{D}(\mathcal{A})} \le C \exp\Big(C\widetilde{L}(2 + \tau\widetilde{L})(n - j - 1)\tau/2\Big)\|\Phi_\tau(u(t_{\ell+j}))^+ - u(t_{\ell+j+1})^+\|_{\mathcal{D}(\mathcal{A})}$$

for $0 \le j \le n-2$, where L and $\widetilde{L}$ are the Lipschitz constants corresponding to the balls of radius $4M^{(1)}$ and $4M^{(1)}(1+\tau L)$. Note that

$$\begin{aligned}\|\Phi_\tau(u(t_{\ell+j}))^+ - u(t_{\ell+j+1})^+\|_{\mathcal{D}(\mathcal{A})} \\ \le \|\Phi_\tau(u(t_{\ell+j})) - u(t_{\ell+j+1})\|_{\mathcal{D}(\mathcal{A})} + \frac{\tau}{2}\|\Phi_\tau(u(t_{\ell+j})) - u(t_{\ell+j+1})\|_{\mathcal{D}(\mathcal{A}^2)}.\end{aligned} \tag{4.31}$$

The second-order local error bound in $\mathcal{D}(\mathcal{A})$ from Proposition 4.5.9 and the first-order local error in $\mathcal{D}(\mathcal{A}^2)$ from Proposition 4.5.10 now show that (4.31) is $\mathcal{O}(\tau^2)$. Now we can estimate (4.30) as

$$\begin{aligned}\|\Phi_\tau^n(u(t_\ell))^+\|_{\mathcal{D}(\mathcal{A})} &\le \|u(t_{\ell+n})^+\|_{\mathcal{D}(\mathcal{A})} + \sum_{j=0}^{n-1} \left\|\Phi_\tau^{n-j}(u(t_{\ell+j}))^+ - \Phi_\tau^{n-j-1}(u(t_{\ell+j+1}))^+\right\|_{\mathcal{D}(\mathcal{A})} \\ &\le \|u(t_{\ell+n})\|_{\mathcal{D}(\mathcal{A})} + \frac{\tau}{2}\|u(t_{\ell+n})\|_{\mathcal{D}(\mathcal{A}^2)} + T\widetilde{C}\exp\left(C\widetilde{L}(2+\tau\widetilde{L})T/2\right)\tau.\end{aligned} \tag{4.32}$$

We have to show that this term is bounded by $2M^{(1)}$ for sufficiently small τ. Let us choose $\tau \le 2$ to get rid of the τ in the exponential in (4.32) and in the definition of $\widetilde{L}$. By definition of $M^{(1)}$, we also have $\|u(t_{\ell+n})\|_{\mathcal{D}(\mathcal{A})} + \frac{\tau}{2}\|u(t_{\ell+n})\|_{\mathcal{D}(\mathcal{A}^2)} \le M^{(1)}$. Thus, we achieve our goal for all

$$\tau \le \tau_0 := \min\left\{2, \frac{M^{(1)}}{T\widetilde{C}\exp\left(C\widetilde{L}(1+\widetilde{L})T\right)}\right\}.$$

This finishes the induction step and establishes (4.29).

The global error can be estimated by the triangle inequality as

$$\|\Phi_\tau^n(u_0) - u(t_n)\|_{\mathcal{X}} \le \sum_{j=0}^{n-1} \|\Phi_\tau^j(\Phi_\tau(u(t_{n-j-1})) - \Phi_\tau^j(u(t_{n-j}))\|_{\mathcal{X}}.$$

Now we apply the stability result (4.27b) to each of the j summands. Note that with the result from (4.29) at hand, we may always choose $M_n = 2M^{(1)}$ in Proposition 4.5.11. Thus, $\widetilde{L}_n$ and $\widetilde{L}_n^{\mathrm{mix}}$ from Proposition 4.5.11 can be chosen independently of n, too, say $\hat{L}$. Together with the local error bounds from Proposition 4.5.8 and Proposition 4.5.9, the global error is now estimated by

$$\begin{aligned}\|\Phi_\tau^n(u_0) - u(t_n)\|_{\mathcal{X}} &\le \sum_{j=0}^{n-1} \widetilde{C}\exp\left(\widetilde{C}\hat{L}(2+\tau\hat{L})j\tau/2\right)\|\Phi_\tau(u(t_{n-j-1}))^+ - u(t_{n-j})^+\|_{\mathcal{X}} \\ &\le \widetilde{C}\sum_{j=0}^{n-1}\exp\left(\widetilde{C}\hat{L}(2+\tau\hat{L})j\tau/2\right)\Big(\|\Phi_\tau(u(t_{n-j-1})) - u(t_{n-j})\|_{\mathcal{X}} \\ &\qquad + \tau\|\Phi_\tau(u(t_{n-j-1})) - u(t_{n-j})\|_{\mathcal{D}(\mathcal{A})}\Big) \\ &\le \widetilde{C}\frac{\exp\left(\widetilde{C}\hat{L}(2+\tau\hat{L})n\tau/2\right) - 1}{\exp\left(\widetilde{C}\hat{L}(2+\tau\hat{L})\tau/2\right) - 1}C_{\mathrm{loc}}\tau^3 \\ &\le \frac{\exp\left(\widetilde{C}\hat{L}(2+\tau_0\hat{L})T/2\right) - 1}{\hat{L}}C_{\mathrm{loc}}\tau^2.\end{aligned}$$

In the last step, we used $1 + x \le \mathrm{e}^x$ for $x \ge 0$ and $t_n \le T$.

Thus, we have finally completed the proof of Theorem 4.5.6. □

Now we return to the parametric setting and explain how the dependency on z can be incorporated. Consider the initial value problem

$$\partial_t u(t,z) = \mathcal{A}(z)u(t,z) + f(u(t,z),z) + g(t,z), \qquad t \geq 0,\ z \in \Sigma, \tag{4.33a}$$

$$u(0,z) = u_0(z) \in \mathcal{D}(\mathcal{A}(z)), \qquad z \in \Sigma, \tag{4.33b}$$

which differs from (4.11) only in the appearance of $z \in \Sigma$, where $\Sigma \subseteq \mathbb{C}^d$ is again a bounded domain as in the model problem from Section 4.2. In the next remark, we state how the assumptions and the result of this section can be extended to the parametric case.

Remark 4.5.12 (Uniform convergence of IMEXT)**.** We make the following assumptions, which are "parametric" versions of the ones stated in Assumption C.

(1) The operator $\mathcal{A}(z)$ is sectorial for all $z \in \Sigma$, $\sup_{z\in\Sigma} \omega_{\mathcal{A}(z)} < 0$ and the domain $\mathcal{D}(\mathcal{A}) = \mathcal{D}(\mathcal{A}(z))$ is independent of $z \in \Sigma$ with equivalent graph norms.

(2) $f \in C^2(\mathcal{D}(\mathcal{A}) \times \Sigma, \mathcal{D}(\mathcal{A}))$

(3) The function f satisfies a "mixed" local Lipschitz property in $\mathcal{X}$: For any $r > 0$, there is a constant L_r^{mix} such that

$$\sup_{z\in\Sigma} \|f(v_2,z) - f(v_1,z)\|_{\mathcal{X}} \leq L_r^{\mathrm{mix}} \|v_2 - v_1\|_{\mathcal{X}}$$

for all $v_1, v_2 \in \mathcal{B}_{\mathcal{D}(\mathcal{A})}(0,r)$.

(4) $g \in C^2([0,T] \times \Sigma, \mathcal{X}) \cap C^1([0,T] \times \Sigma, \mathcal{D}(\mathcal{A}))$

(5) There is a strict solution u of (4.33) which has the regularity

$$u(\cdot,z) \in C^2([0,T], \mathcal{D}(\mathcal{A})) \cap C^1([0,T], \mathcal{D}(\mathcal{A}^2)).$$

and this solution depends continuously on $z \in \Sigma$.

(6) $f(u(\cdot,\cdot),\cdot) + g(\cdot,\cdot) \in C^1([0,T] \times \Sigma, \mathcal{D}(\mathcal{A}^2))$

In particular Assumption C is satisfied if we fix $z \in \Sigma$. The following statement holds:

For any compact $\Sigma' \subseteq \Sigma$, there exists $\tau_0 > 0$ such that the global error of the IMEXT method (4.12) after n steps with initial value $u(0) = u_0$ and step-size $\tau \in (0, \tau_0]$ is bounded by

$$\max_{z\in\Sigma'} \|u_n(z) - u(t_n,z)\|_{\mathcal{X}} \leq C\tau^2 \tag{4.34}$$

for $0 \leq t_n = n\tau \leq T$ and a constant $C > 0$ which depends on the solution u on $[0,T] \times \Sigma'$, but is independent of τ, n and z.

The proof is omitted since it is completely analogous to the proof of Theorem 4.5.6 before. In fact, one only has to check in the proof of Theorem 4.5.6 that the estimates are uniform in $z \in \Sigma'$ such that the constant C is also independent of z. ◇

For the model problem (4.3), the result can be stated as follows. By Corollary 4.2.5, the operators $A(z)$ are sectorial.

Example 4.5.13 (Uniform convergence of IMEXT for the model problem (4.3))**.**

Consider the setting of Section 4.2. Assume that all of the following hold:

(1) Assume that $\sup_{z\in\Sigma}\omega_{A(z)} < 0$ and the domain $\mathcal{D}(A(z)) = W^{2,p}_{B(z)}(D)$ is independent of $z \in \Sigma$ with equivalent graph norms.

(2) $f \in C^2(W^{2,p}_B(D)\times\Sigma, W^{2,p}_B(D))$

(3) For any $r > 0$, there is a constant L^{mix}_r such that

$$\sup_{z\in\Sigma}\|f(v_2,z)-f(v_1,z)\|_{L^p(D)} \le L^{\mathrm{mix}}_r\|v_2-v_1\|_{L^p(D)}$$

for all $v_1, v_2 \in W^{2,p}_B(D)$ with $\|v_j\|_{W^{2,p}(D)} \le r,\ j = 1, 2.$

(4) $g \in C^2([0,T]\times\Sigma, L^p(D)) \cap C^1([0,T]\times\Sigma, W^{2,p}_B(D))$

(5) There is a strict solution u of (4.3) which has the regularity

$$u(\cdot,z) \in C^2([0,T], W^{2,p}_B(D)) \cap C^1([0,T], \widetilde{W}^{4,p}_{B,z}(D)) \tag{4.35}$$

with

$$\widetilde{W}^{4,p}_{B,z}(D) = \{u \in W^{4,p}(D)\colon u,\ A(z)u \in W^{2,p}_B(D)\} \tag{4.36}$$

and the norm of $u(\cdot,z)$ in the space on the right-hand side of (4.35) is uniformly bounded in $z \in \Sigma$.

(6) Assume that $f(u(\cdot,z),z) + g(\cdot,z) \in C^1([0,T], \widetilde{W}^{4,p}_{B,z}(D))$ and

$$\|f(u(\cdot,z),z)+g(\cdot,z)\|_{C^1([0,T],\widetilde{W}^{4,p}_{B,z}(D))}$$

is uniformly bounded in $z \in \Sigma$.

Then the following statement holds: For any compact $\Sigma' \subseteq \Sigma$, there exists $\tau_0 > 0$ such that the error of the IMEXT method after n steps with initial value $u(0) = u_0$ and step-size $\tau \in (0, \tau_0]$ is bounded by

$$\max_{z\in\Sigma'}\|u_n(z)-u(t_n,z)\|_{L^p(D)} \le C\tau^2$$

for $0 \le t_n = n\tau \le T$ and a constant $C > 0$ which depends on the solution u on $[0,T]\times\Sigma'$, but is independent of τ, n and z.

Remark 4.5.14. Consider again (4.3) with the assumptions from Section 4.2. In the previous example, the requirement (1) that the domain $\mathcal{D}(A(z)) = W^{2,p}_{B(z)}(D)$ is independent of $z \in \Sigma$ is automatically satisfied for a homogeneous Neumann boundary condition in the case of isotropic diffusion, i.e. the diffusion coefficients satisfy

$$a_{ij}(x,z) = \delta_{ij}a_{11}(x,z), \qquad i,j = 1,\dots,N.$$

In this case, the boundary operator is

$$B(x,z)u(x) = \sum_{i,j=1}^{N}\delta_{ij}a_{11}(x,z)\nu_j(x)\partial_i u(x) = a_{11}(x,z)\sum_{i=1}^{N}\nu_i(x)\partial_i u(x) = a_{11}(x,z)\nu(x)^\top\nabla u(x)$$

and since $a_{11}(x,z) > 0$ by ellipticity, it follows that $B(x,z^*)u(x) = 0$ holds for $z^* \in \Sigma$ if and only if $B(x,z)u(x) = 0$ for *all* $z \in \Sigma$. ◇

Now we examine the convergence of the IMEXT method in practice.

4.5.3 Numerical verification

Let us verify the second-order convergence of the IMEXT scheme from Theorem 4.5.6 for a simple one-dimensional diffusion equation with a finite difference discretisation of the spatial variable. Additionally, we compare it to some other time integrators.

Example 4.5.15.

Consider the problem

$$\begin{aligned} \partial_t u(t,x) &= \partial_{xx}^2 u(t,x) - u(t,x)^2, && t \geq 0,\ x \in [0,1], \\ u(0,x) &= u_0(x), && x \in [0,1], \\ u(t,0) &= u(t,1) = 0, && t \geq 0, \end{aligned}$$

with

$$u_0(x) = \begin{cases} \exp\left(8 - \frac{1}{x(1-x)}\right), & x \in (0,1), \\ 0, & x = 0 \text{ or } x = 1. \end{cases}$$

In the previous notation, we have

$$\mathcal{A}u = Au \coloneqq \partial_{xx}^2 u, \qquad f(u) = -u^2, \qquad g(t) = 0$$

and $\mathcal{D}(A) = W^{2,p}([0,1]) \cap W_0^{1,p}([0,1])$.

We use the standard central finite difference approximation of the second derivative to discretise the spatial variable x. Thus, the matrix

$$\widetilde{A} = \frac{1}{h^2} \begin{pmatrix} -2 & 1 & 0 & \dots & 0 \\ 1 & -2 & 1 & \ddots & \vdots \\ 0 & 1 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 1 \\ 0 & \dots & 0 & 1 & -2 \end{pmatrix} \tag{4.37}$$

with mesh width $h > 0$ becomes the discrete analogue of the second derivative operator ∂_{xx}^2. For this test problem, the computation of the matrix exponential $\mathrm{e}^{t\widetilde{A}}$ is feasible and relatively cheap (note that $\widetilde{A}$ is sparse, even tridiagonal). We compare the following methods:

EXPLIE .... a first-order splitting method where the linear part is propagated by $\mathrm{e}^{t\widetilde{A}}$ and the non-linear part by an implicit Euler method

RESLIE .... a first-order splitting method where both parts are propagated successively by the implicit Euler method

EXPSTH ... a second-order splitting method where the linear part is propagated by $\mathrm{e}^{t\widetilde{A}}$ and the non-linear part by Heun's method

EXPSTT ... a second-order splitting method where the linear part is propagated by $\mathrm{e}^{t\widetilde{A}}$ and the non-linear part by an implicit trapezoidal rule

TRAP ..... the "classical" trapezoidal splitting method introduced in the beginning of Section 4.4

IMEXT ..... the implicit-explicit trapezoidal method introduced in Section 4.4 and discussed throughout this section

PEACE ..... the Peaceman-Rachford method, see e.g. [26, 48]

Note that all of these methods rely on splitting the equation into the same linear and non-linear part. Methods which do not utilise such a splitting are often not competitive, since in each step solving a potentially large non-linear system would be required. An exception to this are exponential integrators [53], but they also require the computation of an exponential of a matrix, which is not desirable in many applications. Thus they are not examined here.

The linear systems corresponding to the linear subproblem which have to be solved in some of the methods are symmetric, positive definite and tridiagonal and can be solved effortless via a Cholesky decomposition.

Of course, we do not expect the first-order methods to be competitive, but we decided to keep them in the list for comparison.

Let us fix the spatial discretisation with 1000 equidistant subintervals in $[0, 1]$ (mesh width $h = 10^{-3}$) and vary the temporal step-size τ. As a reference solution u_{ref}, we use an approximation with the classical trapezoidal splitting method, but with a much smaller time-step $\tau = 2^{-15}$.

The temporal convergence of these methods is shown in Figure 4.1. The error is measured in the discrete L^2-norm at the final time $T = 1$. A work-precision diagram is depicted in Figure 4.2.

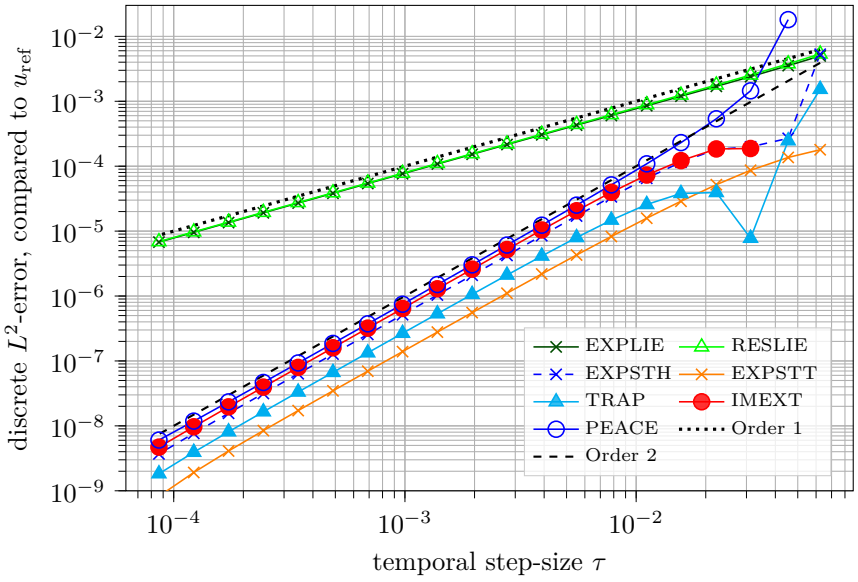


Figure 4.1: Convergence test series for all time integrators from the list

As expected, the first-order methods are not competitive. The performance of IMEXT and PEACE is quite similar, which is not surprising since these methods are constructed in a similar way (although f is treated implicitly in PEACE). All methods using the matrix exponential (EXP`XXX`) are more costly than their competitors, but sometimes yield smaller errors when the same step-size is used. Finally, we compare TRAP and IMEXT. Recall that the only difference is that a full-step with the trapezoidal rule for f in TRAP is replaced by a full-step with Heun's method in IMEXT. Although their computational

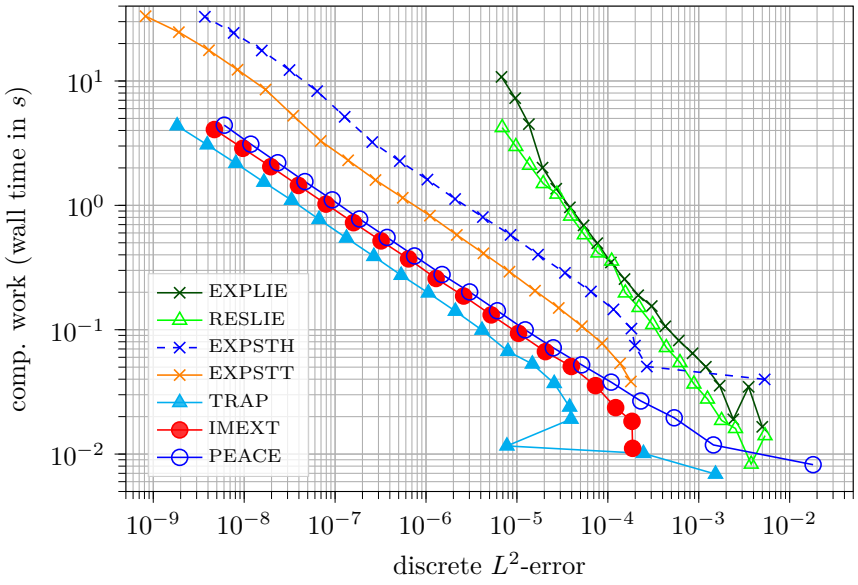


Figure 4.2: Work-precision diagrams for all time integrators from the list

cost is almost the same, TRAP yields smaller errors (2–3 times smaller) and thus performs slightly better (as the reader may have guessed). But the non-linearity in this toy problem is very simple and solving it implicitly is not much harder then solving it explicitly. This will not be the case anymore for more interesting problems. We draw the conclusion that IMEXT is indeed a competitive method and the theoretically predicted order 2 is confirmed. Thus, it may possibly be an attractive choice for the temporal discretisation of several problems of interest.

The next example shows that the convergence order of IMEXT is reduced if less regularity is available.

Example 4.5.16.

One may ask the question whether the regularity of the solution

$$u \in C^1([0,T], \mathcal{D}(\mathcal{A}^2)) \cap C^2([0,T], \mathcal{D}(\mathcal{A}))$$

from Assumption C(5) in Section 4.5.2 can be reduced while keeping second-order convergence. We now show numerically that in case of initial data $u_0 \in \mathcal{D}(\mathcal{A}) \setminus \mathcal{D}(\mathcal{A}^2)$, the convergence order is strictly less than two. We change the setting from the previous example as follows: We remove the non-linearity (i.e. $f(u) = 0$) and change the initial value to

$$u_0(x) = \left|x - \frac{1}{2}\right| \left(x - \frac{1}{2}\right)^2 - \frac{1}{8} \qquad \text{with derivative} \qquad u_0'(x) = 3\left|x - \frac{1}{2}\right| \left(x - \frac{1}{2}\right)$$

such that $u_0 \in W_0^{2,p}([0,1]) \setminus W^{4,p}([0,1])$. Thus, we have indeed $u_0 \in \mathcal{D}(A) \setminus \mathcal{D}(A^2)$. The IMEXT method reduces to the trapezoidal/midpoint method in this case. (So in fact we use the famous *Crank-Nicolson* method here.) The interval $[0,1]$ is again discretised with 1000 equidistant subintervals. To obtain a reference solution, we use the numerically computed matrix exponential $\mathrm{e}^{t\widetilde{A}}$ with $\widetilde{A}$ from (4.37). The result of a convergence test series is given in Figure 4.3, with the error measured at the final time $T = 1$. We obtain an experimental order of convergence (EOC) of approximately 1.44.

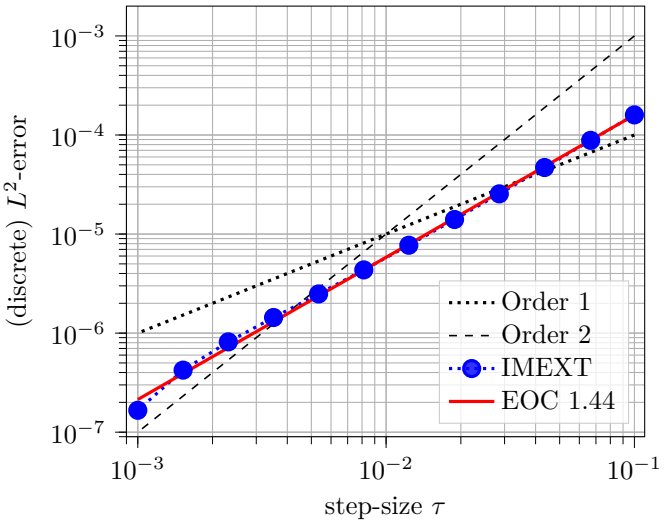


Figure 4.3: Order reduction for initial data in $\mathcal{D}(A) \setminus \mathcal{D}(A^2)$

So we verified numerically that for initial data not belonging to $\mathcal{D}(A^2)$, we do not obtain second-order convergence, i.e.

$$\|\mathrm{e}^{\tau n A} v - \left((I - \tfrac{\tau}{2}A)^{-1}(I + \tfrac{\tau}{2}A)\right)^n v\|_X \neq \mathcal{O}(\tau^2) \qquad \text{for} \qquad v \in \mathcal{D}(A) \setminus \mathcal{D}(A^2).$$

Of course this observation is not new. It is stated for example in [47], where such a behaviour is not only discussed for the trapezoidal method, but for other single-step methods, too. If one assumes $v \in \mathcal{D}(A^2)$, then $\mathcal{O}(\tau^2)$ is obtained as expected for this problem, see [67, Thm. 4.2].

Now we return to parametric parabolic equations and discuss stochastic collocation methods.

4.6 Single-level stochastic collocation

We consider the problem (4.33) and assume that the assumptions for uniform convergence of the IMEXT method in $z \in \Sigma$ from Remark 4.5.12 hold. Suppose additionally that for $\Sigma \supseteq \Sigma(\boldsymbol{\sigma}) \supseteq \Gamma = [-1,1]^d$ with $\boldsymbol{\sigma} \in (1,\infty)^d$, the maps

$$u \colon \Sigma(\boldsymbol{\sigma}) \to C^1([0,T],\mathcal{X}) \qquad \text{and} \qquad u_n \colon \Sigma(\boldsymbol{\sigma}) \to \mathcal{X}$$

are analytic for all $n \in \mathbb{N}_0$, where u is the solution of (4.33) and u_n is the IMEXT approximation after n steps with step-size $\tau > 0$, as defined in (4.12).

Remark 4.6.1. This assumption is not as strict as it may seem and such an assumption can often be verified for elliptic and parabolic problems, see e.g. [21] in case of an elliptic and [85, Sec. 3] in case of a parabolic problem. In the latter article, an equation is treated which is a special case of (4.3), albeit linear.

For non-linear parabolic problems, an abstract result is given in [75, Thm. 8.4.4] which states that the solution depends analytically on the parameters if these parameters enter the equation in an analytic

way. The precise statement uses a decent amount of Banach space interpolation theory and it is still under research under which precise assumptions this result can be applied to our model problem.

For the time-discrete approximation u_n, analyticity can be deduced from the explicit formula (4.12) under suitable requirements on $\mathcal{A}, f, g$ and u_0. ◇

The single-level approximation of the solution u of (4.33) at time $t_n = n\tau$ is given by

$$u_{L,n} = \mathcal{I}_L^{p,g} u_n,$$

where $\mathcal{I}_L^{p,g}$ is the sparse grid interpolant from (2.15). Thus, we compute the IMEXT approximations at time t_n for η_L values of z, namely the nodes of the sparse grid $\mathcal{H}_L^{p,g}$ from (2.16). This definition is of course in agreement with the general definition (3.7) from Section 3.3.

We stress that although the parameter vector $z \in \Sigma$ was assumed to be *complex* before, we are in the end only interested in solving the problem (4.33) for *real* vectors $z \in [-1,1]^d$. The analytic extension to a complex polyellipse is just a useful tool which allows us to apply the sparse grid interpolation error estimate from Theorem 2.6.2.

Notation. Whenever we consider *real* parameters belonging to $\Gamma = [-1,1]^d$, we denote them by y. In contrast to that, parameters from the *complex* set $\Sigma \supsetneq \Gamma$ are denoted by the letter z for distinction.

The analyticity of u allows us to use Theorem 2.6.2, which implies that the sparse grid interpolation error is bounded by

$$\|u - \mathcal{I}_L^{p,g} u\|_{L^2_\varrho(\Gamma, C^1([0,T],\mathcal{X}))} \leq C \eta_L^{-\mu} \max_{z \in \Sigma(\boldsymbol{\sigma})} \|u(\cdot, z)\|_{C^1([0,T],\mathcal{X})},$$

for some constants C and μ if we use Clenshaw-Curtis abscissas and p and g as in Theorem 2.6.2.

Now consider the time-discrete approximation u_n for $n \in \mathbb{N}_0$. By the triangle inequality and (4.34), its norm is bounded by

$$\max_{z \in \Sigma(\boldsymbol{\sigma})} \|u_n(z)\|_{\mathcal{X}} \leq \max_{z \in \Sigma(\boldsymbol{\sigma})} \|u_n(z) - u(t_n, z)\|_{\mathcal{X}} + \|u(t_n, z)\|_{\mathcal{X}} \leq C\tau_0^2 + \max_{z \in \Sigma(\boldsymbol{\sigma})} \max_{t \in [0,T]} \|u(t, z)\|_{\mathcal{X}}$$

for some $\tau_0 > 0$, in particular independently of n and τ. Again, Theorem 2.6.2 implies

$$\|u_n - \mathcal{I}_L^{p,g} u_n\|_{L^2_\varrho(\Gamma,\mathcal{X})} \leq C \eta_L^{-\mu} \max_{z \in \Sigma(\boldsymbol{\sigma})} \|u_n(z)\|_{\mathcal{X}}. \tag{4.38}$$

Splitting the total error of the single-level stochastic collocation method as

$$\|u(t_n) - u_{L,n}\|_{L^2_\varrho(\Gamma,\mathcal{X})} \leq \|u(t_n) - u_n\|_{L^2_\varrho(\Gamma,\mathcal{X})} + \|u_n - \mathcal{I}_L^{p,g} u_n\|_{L^2_\varrho(\Gamma,\mathcal{X})}$$

and using (4.34) and (4.38), we arrive at the following theorem.

Theorem 4.6.2 (Error of single-level collocation)**.** *Under the previous assumptions, there is a step-size $\tau_0 > 0$ such that for all temporal step-sizes $\tau \in (0, \tau_0]$, the single-level stochastic collocation error is bounded by*

$$\|u(t_n) - u_{L,n}\|_{L^2_\varrho(\Gamma,\mathcal{X})} \leq C(\tau^2 + \eta_L^{-\mu})$$

as long as $0 \leq t_n = n\tau \leq T$ for $n \in \mathbb{N}_0$. The constant C depends on the solution u of (4.33) on $[0,T] \times \Sigma(\boldsymbol{\sigma})$, the dimension d, $\boldsymbol{\sigma}$, but is independent of η_L, τ and n.

Thus, we obtain algebraic convergence with respect to the temporal step-size τ and the number of collocation points η_L. Before we are able to verify this theorem numerically for the predator-prey system (4.1) from the introductory section of this chapter, we briefly have to discuss the spatial discretisation which we use.

Remark 4.6.3 (Finite element discretisation)**.** Consider the system

$$\partial_t u_1 = \operatorname{div}_x(\delta_1(x,y)\nabla_x u_1) + R_1(u_1,u_2,y), \qquad \text{in } [0,T]\times D\times\Gamma, \tag{4.39a}$$
$$\partial_t u_2 = \operatorname{div}_x(\delta_2(x,y)\nabla_x u_2) + R_2(u_1,u_2,y), \qquad \text{in } [0,T]\times D\times\Gamma, \tag{4.39b}$$
$$u_1(0,x,y) = u_{1,0}(x,y), \qquad \text{for } (x,y)\in D\times\Gamma, \tag{4.39c}$$
$$u_2(0,x,y) = u_{2,0}(x,y), \qquad \text{for } (x,y)\in D\times\Gamma, \tag{4.39d}$$
$$\frac{\partial u_1}{\partial \nu} = \frac{\partial u_2}{\partial \nu} = 0, \qquad \text{on } [0,T]\times\partial D\times\Gamma, \tag{4.39e}$$

with $u_j = u_j(t,x,y)$ for $j=1,2$. By div_x, we denote the divergence operator with respect to the spatial variable x. This system is a parametric version of (4.1) from the introduction of this chapter which additionally contains heterogeneous diffusion coefficients δ_1 and δ_2.

A standard weak formulation of this problem for fixed value of $y\in\Gamma$ is given as follows: Determine

$$u_j(\cdot,\cdot,y)\in L^2((0,T),H^2(D)) \qquad \text{with} \qquad \partial_t u_j(\cdot,\cdot,y)\in L^2((0,T),L^2(D))$$

for $j=1,2$ such that for all $(v_1,v_2)\in H^2(D)\times H^2(D)$,

$$\int_D \partial_t u_1(t,x,y)v_1(x)\mathrm{d}x + \int_D \delta_1(x,y)\nabla u_1(t,x,y)\nabla v_1(x)\mathrm{d}x = \int_D R_1(u_1(t,x,y),u_2(t,x,y))v_1(x)\mathrm{d}x,$$
$$\int_D \partial_t u_2(t,x,y)v_2(x)\mathrm{d}x + \int_D \delta_2(x,y)\nabla u_2(t,x,y)\nabla v_2(x)\mathrm{d}x = \int_D R_2(u_1(t,x,y),u_2(t,x,y))v_2(x)\mathrm{d}x,$$
$$u_1(0,x,y) = u_{1,0}(x,y), \qquad x\in D,$$
$$u_2(0,x,y) = u_{2,0}(x,y), \qquad x\in D.$$

From this weak formulation, a space-discrete system for a finite element space $V_h\subseteq C(\overline{D})$ with mesh width $h>0$ can be derived. We choose a triangular mesh (the same mesh for all $y\in\Gamma$) and define V_h as the space of continuous functions on $\overline{D}$ which are linear on each triangle ($\mathbb{P}_1$ elements).

As the time discretisation is done by splitting the equation into two parts prior to the spatial discretisation, we actually arrive at two spatially discrete systems which are solved in an alternating fashion dictated by the IMEXT method. With mass and stiffness matrices M_h and $A_{1,h}(y),\ A_{2,h}(y)\in\mathbb{R}^{\mathcal{N}_h\times\mathcal{N}_h}$, where $\mathcal{N}_h=\dim(V_h)$, the first system is of the form

$$M_h\partial_t u_{1,h}^{(1)}(t,y) + A_{1,h}(y)u_{1,h}^{(1)}(t,y) = 0, \qquad t\geq 0, \tag{4.40a}$$
$$M_h\partial_t u_{2,h}^{(1)}(t,y) + A_{2,h}(y)u_{2,h}^{(1)}(t,y) = 0, \qquad t\geq 0, \tag{4.40b}$$

supplied with initial values for $u_{1,h}^{(1)}(0,y)$ and $u_{2,h}^{(1)}(0,y)$ and has to be solved for $u_{j,h}^{(1)}(t,y)\in\mathbb{R}^{\mathcal{N}_h}$, $j=1,2$. As these two equations (4.40a) and (4.40b) are decoupled ($u_{1,h}^{(1)}$ does not appear in the second equation and $u_{2,h}^{(1)}$ does not appear in the first equation), they can be solved independently of each other.

The system corresponding to the second subproblem is of the form

$$\partial_t u_{1,h}^{(2)}(t,y) = R_1(u_{1,h}^{(2)}(t,y),u_{2,h}^{(2)}(t,y),y), \qquad t\geq 0, \tag{4.41a}$$
$$\partial_t u_{2,h}^{(2)}(t,y) = R_2(u_{1,h}^{(2)}(t,y),u_{2,h}^{(2)}(t,y),y), \qquad t\geq 0, \tag{4.41b}$$

which is again supplied with initial values and has to be solved for $u_{j,h}^{(2)}(t,y) \in \mathbb{R}^{\mathcal{N}_h}$, $j = 1,2$. The reaction terms R_1 and R_2 in this example do not depend explicitly on x, and thus (4.41a) and (4.41b) give two-dimensional non-linear ODE systems in each point x of the spatial mesh. The IMEXT method now consists of the following three steps in each time-step: First, (4.40) is solved by an explicit Euler method with step-size $\tau/2$, then (4.41) is solved by Heun's method with step-size τ, and finally (4.40) is solved by an implicit Euler method with step-size $\tau/2$. The initial values are always taken as the results of the previous step. ◇

Now we are ready to verify Theorem 4.6.2 numerically for this predator-prey system.

Example 4.6.4.

Consider the predator-prey system (4.39) with polynomial reaction terms[4]

$$\begin{aligned} R_1(u_1,u_2,y) &= (c_1 - c_2u_1 - c_3 - c_4u_2)u_1, \\ R_2(u_1,u_2,y) &= (c_5u_1 - c_6)u_2 \end{aligned}$$

for given parameters $c_j = c_j(y)$, $j = 1,\dots,6$. The parameters c_j represent the following quantities:

- c_1: "natural" birth rate prey
- c_2: social friction prey
- c_3: "natural" death rate prey
- c_4: death rate prey due to predation
- c_5: birth rate predators
- c_6: death rate predators

Clearly, if c_4 or c_5 is not equal to zero, then these reaction terms couple the equations (4.39a) and (4.39b) for u_1 and u_2. Here, the reaction constants $c_1,\dots,c_6$ are partly uncertain and given by

$$\begin{aligned} c_1(y) &= 1 + y_1, & c_2(y) &= 0, & c_3(y) &= \frac{1}{2}, \\ c_4(y) &= \frac{1}{2}(1+y_2), & c_5(y) &= \frac{3}{2}(1+y_2), & c_6(y) &= \frac{1}{10} \end{aligned}$$

for $y = (y_1,y_2) \in \Gamma = [-1,1]^2$ (and thus the stochastic dimension is $d = 2$). The diffusion is uncertain, too, and given by

$$\begin{aligned} \delta_1(x,y) &= 0.1 + 0.00625y_1\left(x_1^2 + x_2^2\right), \\ \delta_2(x,y) &= 0.3 + 0.15y_2\exp\left(-x_1^2 - x_2^2\right), \end{aligned}$$

where $x = (x_1,x_2) \in D$ and $y = (y_1,y_2) \in \Gamma$. It should be noted that it is unusual that the variables y_1 and y_2 appear in the diffusion *and* reaction terms since it is unlikely that these quantities depend on the same randomness in applications. But this example is mainly for demonstrational purposes and we want to keep the stochastic dimension rather low. This allows us to do convergence tests in a wider range of sparse grid depths and temporal step-sizes than for more complicated high-dimensional problems. In a later example we increase the stochastic dimension d to cover a more realistic scenario.

The circular domain D is discretised with $\mathbb{P}_1$ elements on the triangulation shown in Figure 4.4. The spatial and temporal discretisation are realised as described in Remark 4.6.3 before.

[4] This choice of reaction terms is used in several books and articles which treat the ODE system, see e.g. [46, Eq. (60.7)].

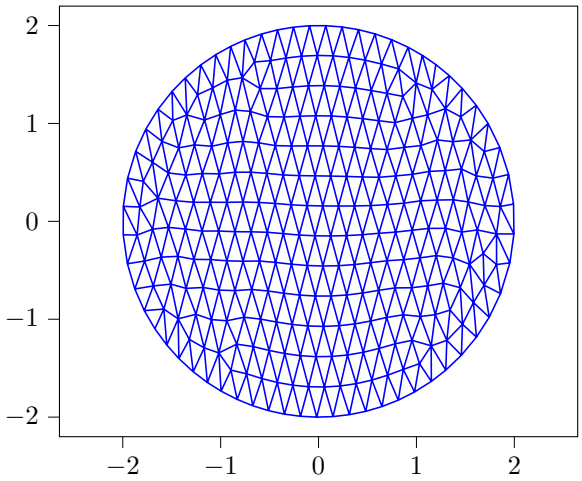


Figure 4.4: Spatial domain and its triangulation; mesh width $h = 0.28$, $\dim(V_h) = 726$

As all computations in this example are carried out on the same coarse mesh shown in Figure 4.4, the error contribution from the space discretisation will usually be the dominant one compared to the stochastic and temporal error. So our results are at first sight only meaningful for a semi-discrete problem. The computed errors in the experiments should not be understood as errors compared to the solution of the continuous problem under consideration, but compared to the spatially discrete problem.

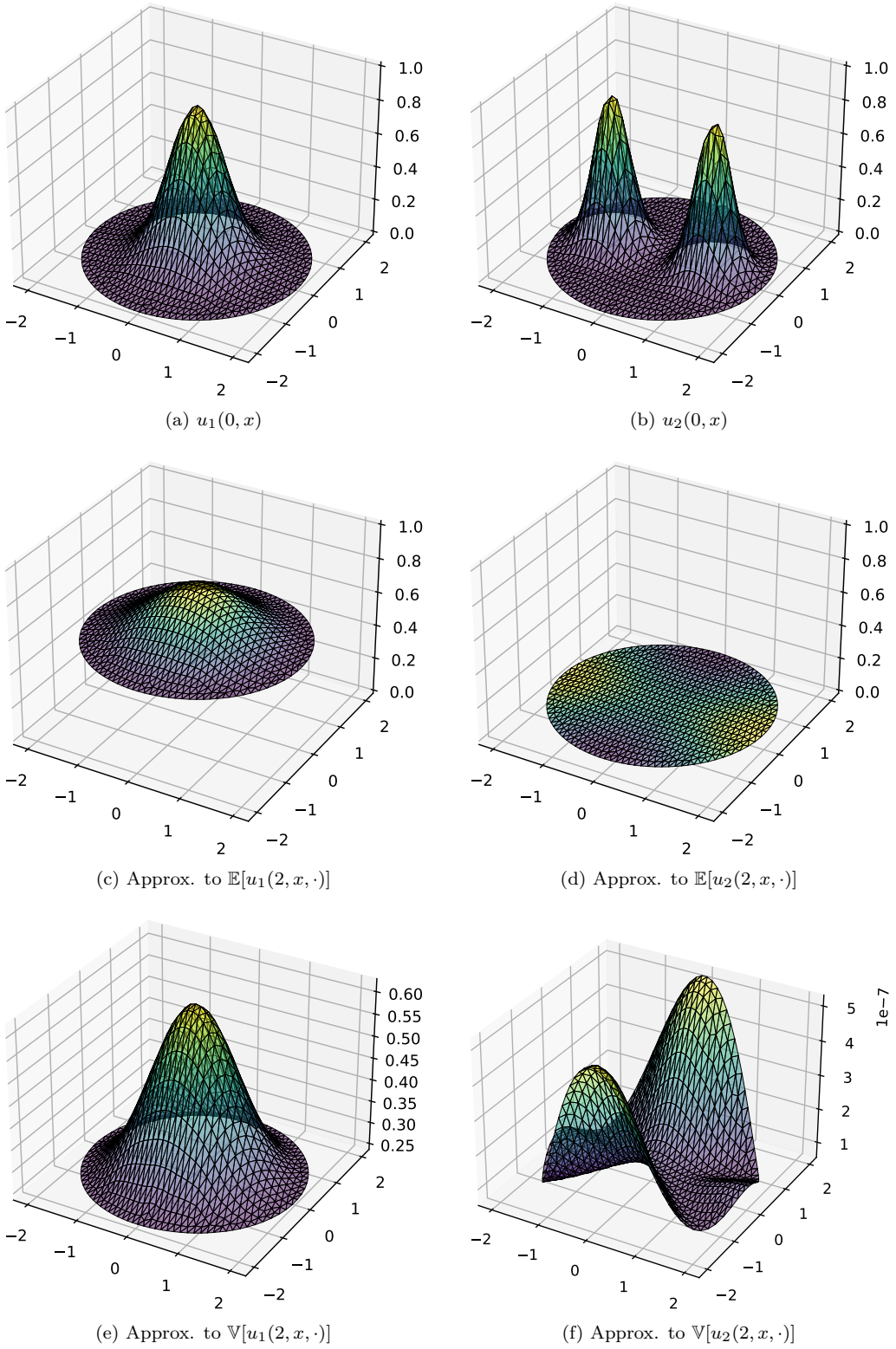
The initial configurations $u_{1,0}$ and $u_{2,0}$ of the prey and predator densities are independent of y in this example and depicted in Figure 4.5a and Figure 4.5b.

Let us describe the influence of the uncertainty in this system from the application side. Since $c_2 \equiv 0$, there is no social friction between the prey individuals, and there is no uncertainty in the death rates of predators and prey, too. But there is uncertainty in the two interaction constants c_4 and c_5, in the diffusion parameters and most importantly in the reproduction rate of the prey, c_1. Observe that for $y_1 \to -1$, no prey individuals are born and the prey will almost be eradicated after a short time. For $y_1 \to 1$, the prey will not be eradicated and the system will approach a (spatially constant) equilibrium between prey and predators. This prediction should be confirmed by our computations.

We use a temporal step-size $\tau = 0.01$ and a sparse grid of depth $L = 4$ with $\eta_L = 65$ nodes. The collocation strategy in the parameter space is based on the Smolyak polynomial space and Clenshaw-Curtis abscissas. (This will always be the case throughout our numerical examples.) At time $t = 2$, we compute the expected values and variances of both u_1 and u_2 in each grid point. The results are shown in the remaining images (c) – (f) in Figure 4.5.

Observe that the prey has not yet reached the spatially constant state, but the predators have (almost). Notice that $\mathbb{E}[u_2(T, x, \cdot)]$ is not zero, but takes values ≈ 0.074 throughout the domain. The vastly different behaviour we predicted for the prey is captured in the large variance of u_1. (The variance becomes even larger if one considers a larger time interval.) However, the variance of the predators is very small and thus we infer that the uncertain birth rate of the prey does not substantially change the resulting predator population. This example also shows that by incorporating *uncertainty* into the computations, we may sometimes arrive at a very *certain* conclusion, in this case the predator population in the equilibrium.

Let us now examine the convergence with respect to τ and L. By doing this, we aim to confirm



(a) $u_1(0,x)$

(b) $u_2(0,x)$

(c) Approx. to $\mathbb{E}[u_1(2,x,\cdot)]$

(d) Approx. to $\mathbb{E}[u_2(2,x,\cdot)]$

(e) Approx. to $\mathbb{V}[u_1(2,x,\cdot)]$

(f) Approx. to $\mathbb{V}[u_2(2,x,\cdot)]$

Figure 4.5: SLSC + IMEXT with $\tau = 0.01$, $\eta_L = 65$, $h = 0.28$, $\dim(V_h) = 726$

the statement of Theorem 4.6.2. It is reasonable that the theorem applies to the current PDE system, as it consists of two PDEs which are essentially of the form (4.3), but coupled. A formal proof that Theorem 4.6.2 applies to this PDE system is not given here.

Convergence test. We consider the time interval $[0,T] = [0,1]$ and use a single-level collocation approximation as reference solution $u_{\mathrm{ref}} = (u_{L_{\mathrm{ref}},\tau_{\mathrm{ref}},1}, u_{L_{\mathrm{ref}},\tau_{\mathrm{ref}},2}) \approx (u_1(T), u_2(T))$ computed with a sparse grid of depth $L_{\mathrm{ref}} = 9$ ($\eta_{L_{\mathrm{ref}}} = 3329$) and temporal step-size $\tau_{\mathrm{ref}} = 5 \cdot 10^{-6}$. The norm in which we measure the error is

$$\|(e_1, e_2)\|_* = \sqrt{\|e_1\|^2_{L^2_{\hat{\varrho}}(\Gamma, L^2(D))} + \|e_2\|^2_{L^2_{\hat{\varrho}}(\Gamma, L^2(D))}} \tag{4.42}$$

(more strictly speaking its discrete analogue), computed at the final time T. In the following pictures, the quantity

$$\|(e_1, e_2)\|_* \qquad \text{with} \qquad e_k = u_{L_{\mathrm{ref}},\tau_{\mathrm{ref}},k} - u_{L,\tau,k}, \quad k = 1, 2,$$

will be labelled by "error in $\|\cdot\|_*$". Results are depicted in Figure 4.6.

The convergence with respect to τ is almost as expected, although the full order 2 does not show up over the entire range of step-sizes τ. This is not necessarily a systematic order reduction but could also be explained by an insufficiently fine reference solution. In particular, order reduction due to stiffness would typically arise for large step-sizes, and not for small step-sizes as in our situation here.

From Theorem 2.6.2, we expect that the convergence with respect to η_L is $\mathcal{O}(\eta_L^{-\mu})$ for some $\mu > 0$, which should give us a line with slope $-\mu$ in the logarithmic plot. At first sight, the convergence with respect to η_L seems to be much better than expected, as the error does not decrease linearly in the picture, but faster. This is not surprising in light of Remark 2.6.5, where we adressed the subexponential convergence for depths $L > d/\log(2)$. For $d = 2$, the improved convergence rate occurs for all levels $L \geq 3$ and thus it can be observed here, too. We included a reference line ⋯ with subexponential convergence in the picture which resembles the behaviour of the error much better.

For both convergence test series, the error does not improve anymore if the error contribution of the other discretisation type becomes the dominant term. This can be verified in the pictures, too.

A work-precision diagram is also shown in Figure 4.6(c). In the next section, we will examine how the multi-level method compares to this picture.

Now we turn to the multi-level setting.

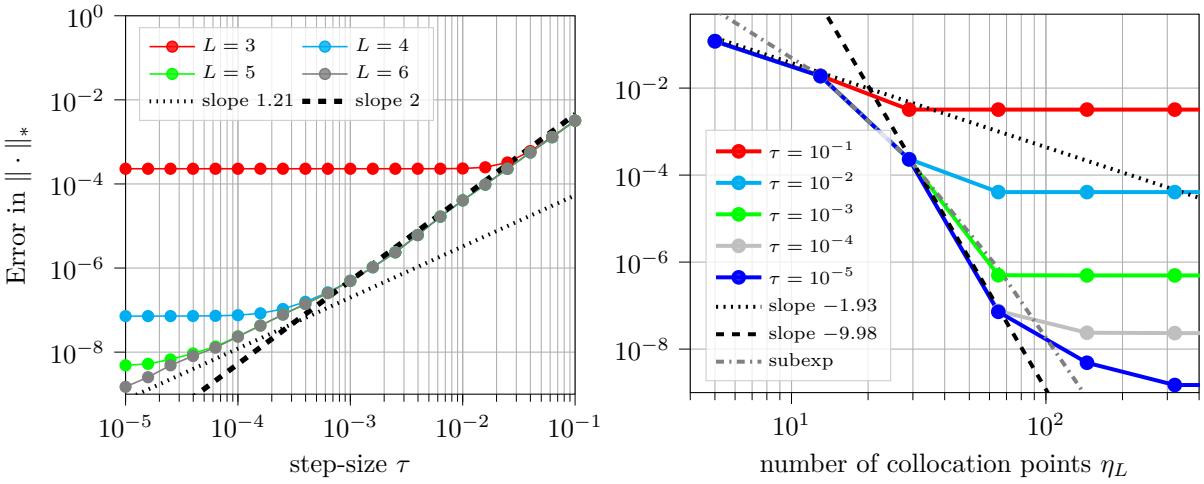
4.7 Multi-level stochastic collocation

Here we explain that the general Assumptions B1 – B3 for the multi-level method from Section 3.4 and Section 3.5 are satisfied for problem (4.33) if we make the same assumptions as in the previous section.

Let $T > 0$. We use the notation introduced in Section 3.4,

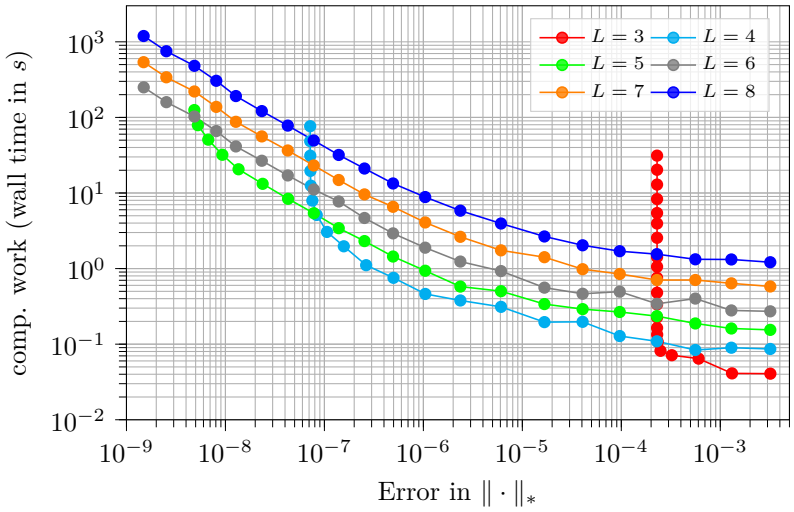
$$u_{\tau_j} = \Phi^{N_j}_{\tau_j}(u_0) \qquad \text{and} \qquad u_J^{(\mathrm{ML})}$$

[5] Lines for $L = 7$ and $L = 8$ are not included in the picture as they coincide visually with the line for $L = 6$.



(a) Conv. of temporal discretisation[5]

(b) Conv. of stochastic discretisation



(c) Work-precision diagram

Figure 4.6: Convergence test series and cost scaling of SLSC + IMEXT; $T = 1$, $h = 0.28$, $\dim(V_h) = 726$

with the numerical flow of the IMEXT scheme Φ_τ^m from (4.18), $\tau_j = 2^{-j}\tau_0$ and $T = \tau_j N_j$ for $j \in \mathbb{N}_0$. The largest step-size τ_0 with $N_0 = T/\tau_0 \in \mathbb{N}$ is chosen either as the (in practice unknown) value τ_0 from Remark 4.5.12 or smaller, but not larger.

To show Assumption B1, we apply (4.34) and obtain

$$\|u(T) - \Phi_{\tau_j}^{N_j}(u_0)\|_{L^2_\varrho(\Gamma,\mathcal{X})} \le \|u(T) - \Phi_{\tau_j}^{N_j}(u_0)\|_{L^\infty(\Gamma,\mathcal{X})} \le \max_{z\in\Sigma'} \|u(T,z) - \Phi_{\tau_j}^{N_j}(u_0,z)\|_{\mathcal{X}} \le C\tau_j^2$$

for $j \in \mathbb{N}_0$. Thus, Assumption B1 is satisfied with $\alpha = 2$.

Next we show that Assumption B2 is satisfied for $\zeta\colon \Lambda(\Gamma,\mathcal{X}) \to \mathbb{R}$ with[6]

$$\Lambda(\Gamma,\mathcal{X}) = \{v\colon \Gamma \to \mathcal{X} \text{ analytic} \mid v \text{ has an analytic extension to an open set containing } \Sigma(\boldsymbol{\sigma})\},$$
$$\zeta(v) = \max_{z\in\Sigma(\boldsymbol{\sigma})} \|v(z)\|_{\mathcal{X}},$$

for some $\boldsymbol{\sigma} \in (1,\infty)^d$ and $\beta = 2$. The interpolation error estimate in Assumption B2, (3.10), follows directly from Theorem 2.6.2 with $\kappa_\ell = \eta_\ell^{-\mu}$ for some $\mu > 0$. We stress that we have already *assumed* the analyticity of u and u_{τ_j} in the previous section.

For the remaining estimates, we apply the triangle inequality and (4.34) to arrive at

$$\zeta(u_{\tau_{j+1}} - u_{\tau_j}) \le \max_{z\in\Sigma(\boldsymbol{\sigma})} \|u_{\tau_{j+1}}(z) - u(T,z)\|_{\mathcal{X}} + \max_{z\in\Sigma(\boldsymbol{\sigma})} \|u(T,z) - u_{\tau_j}(z)\|_{\mathcal{X}} \le C(\tau_{j+1}^2 + \tau_j^2) \le 5C\tau_{j+1}^2,$$
$$\zeta(u_{\tau_j}) \le \max_{z\in\Sigma(\boldsymbol{\sigma})} \|u_{\tau_j}(z) - u(T,z)\|_{\mathcal{X}} + \max_{z\in\Sigma(\boldsymbol{\sigma})} \|u(T,z)\|_{\mathcal{X}} \le \left(C + \tau_0^{-2} \max_{z\in\Sigma(\boldsymbol{\sigma})} \|u(T,z)\|_{\mathcal{X}}\right) \tau_0^2,$$

and thus Assumption B2 is satisfied with $\beta = 2$, too. Assumption B3 basically says that the cost of evaluating $u_{\tau_j} - u_{\tau_{j-1}}$ is proportional to the number of time-steps, which is reasonable in practice. Thus, we can apply Theorem 3.5.2 and obtain the ε-cost bound

$$C^{(\mathrm{ML})} \lesssim \begin{cases} \varepsilon^{-\frac{1}{\mu}}, & \mu < 2, \\ \varepsilon^{-\frac{1}{\mu}} |\log(\varepsilon)|^{1+\frac{1}{\mu}}, & \mu = 2, \\ \varepsilon^{-\frac{1}{2}}, & \mu > 2, \end{cases}$$

for a multi-level approximation $u_J^{(\mathrm{ML})}$ satisfying $\|u(T) - u_J^{(\mathrm{ML})}\|_{L^2_\varrho(\Gamma,\mathcal{X})} \le \varepsilon$.

Now we use the multi-level method in the setting from Example 4.6.4.

Example 4.7.1.

The purpose of this example is threefold: We show that

- the error of the multi-level estimator in the norm $\|\cdot\|_*$ from (4.42) stays below (or almost below) the given tolerance ε,
- the computational cost scales as predicted by Theorem 3.5.2,
- the benefits of the multi-level approach in low dimensions (such as $d = 2$) are limited to settings where the regularity in the parameter space is rather low and the tolerance small.

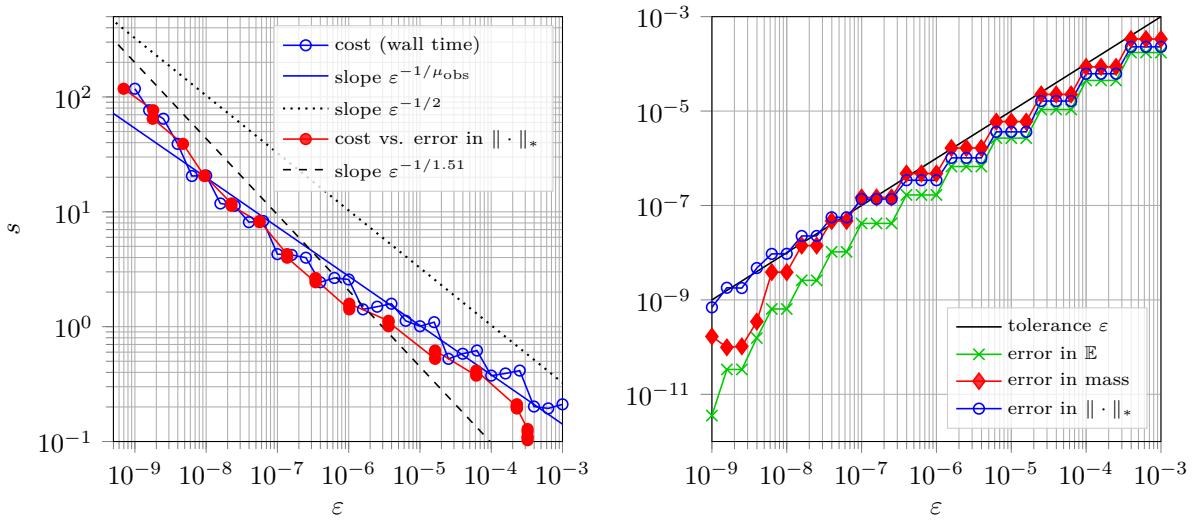
[6] Note that $\Lambda(\Gamma,\mathcal{X})$ is indeed a Banach space with norm ζ. This can be proved using Morera's theorem and Cauchy's integral theorem, see e.g. [96, Thm. 10.28]. The analytic extension of a function $u \in \Lambda(\Gamma,\mathcal{X})$ to $\Sigma(\boldsymbol{\sigma})$ is unique by the identity theorem for analytic functions (as $\Gamma = [-1,1]^d$ clearly has an accumulation point).

The third point is explained in detail later.

We return to the setting for the convergence tests in Example 4.6.4, but now we apply the multi-level method with final time $T = 1$. For the constants $\mu = 9.35$, $C_I C_\zeta = 2.80 \cdot 10^{11}$, $C_T = 0.25$ and $\alpha = 1.90$, we could verify the Assumptions B1 and B2 numerically on the first few levels[7], and we have set $\beta = \alpha$. We refer to the explanations in Section 3.5.2 how this is done. We use the rounding strategy "up/down" explained in the end of Section 3.5 and the maximal step-size $\tau_0 = 0.1$.

To address the questions above, we apply the multi-level method for different values of ε. For each of these values, we compute the error compared to the reference solution u_{ref} with $L_{\text{ref}} = 9$ ($\eta_{L_{\text{ref}}} = 3329$) and $\tau_{\text{ref}} = 5 \cdot 10^{-6}$ (the same reference solution as for the convergence tests for the single-level method in Example 4.6.4) and keep track of the time required to compute the multi-level approximation. By "time", we mean the wall time to perform the time integration itself and ignore matrix assembly and other pre- or postprocessing computations (which are heavily dependent on the specific implementation anyway). The time for the single-level method in Figure 4.6(c) was measured in the same way, such that the comparison is somehow fair.

The results are shown in Figure 4.7.



(a) Cost vs. given ε

(b) Error (compared to u_{ref}) vs. given ε

Figure 4.7: MLSC for the predator-prey system (4.39): $d = 2$, $T = 1$, $\mu = 9.35$, $\mu_{\text{obs}} = 2.33$, $h = 0.28$

Since $2 = \beta < \mu = 9.35$, we expect from Example 3.5.3 that the computational cost scales as $\varepsilon^{-1/2}$ for $\varepsilon \to 0$. We *observe* the overall cost scaling $\varepsilon^{-1/\mu_{\text{obs}}}$ for $\mu_{\text{obs}} = 2.33$ (——), which is better than the predicted scaling $\varepsilon^{-1/2}$. The slope of - - - is fitted to the six smallest tolerances ε and suggests that the cost scales rather like $\varepsilon^{-1/1.51}$ in this area, which is worse than expected. Two possible reasons for that could be the following.

In the iterative process of finding the correct value of J from Theorem 3.5.2 described in the end of Section 3.5.2, some single-level collocation approximations $u^{(\text{SL})}_{\eta_{j-j},\tau_j} - u^{(\text{SL})}_{\eta_{j-j},\tau_{j-1}}$ for $j = 0, \dots, \hat{J}$ and $\hat{J} = 0, \dots, J-1$ have to be computed, which may or may not enter the multi-level estimator in the end.

[7] The huge value of $C_I C_\zeta$ is not implausible, because it is accompanied by the large rate μ.

Their computation is included in the time depicted in Figure 5.2(a), but is not included in the theoretical cost from Theorem 3.5.2. This could explain why slightly more effort than expected is necessary for smaller tolerances ε. On the other hand, one can reuse most of these approximations for the multi-level approximation $u_J^{(\mathrm{ML})}$, so the overhead is not too large. Another effect which contributes to the slightly worse cost behaviour which we observe is the overestimation of the theoretical value of η_{J-j} explained in Remark 3.5.4. The overestimation can be somehow controlled by the choice of rounding strategy.

The right picture shows that the multi-level estimator indeed achieves an error almost equal to ε in the norm of interest $\|\cdot\|_*$ from (4.42). Note that we choose the weaker stopping criterion (3.28) instead of the stronger one from (3.29) and thus it is not guaranteed that the error is less than ε. But the result is convincing nevertheless. The other lines and are included only to show the error in some other quantities one might be interested in:

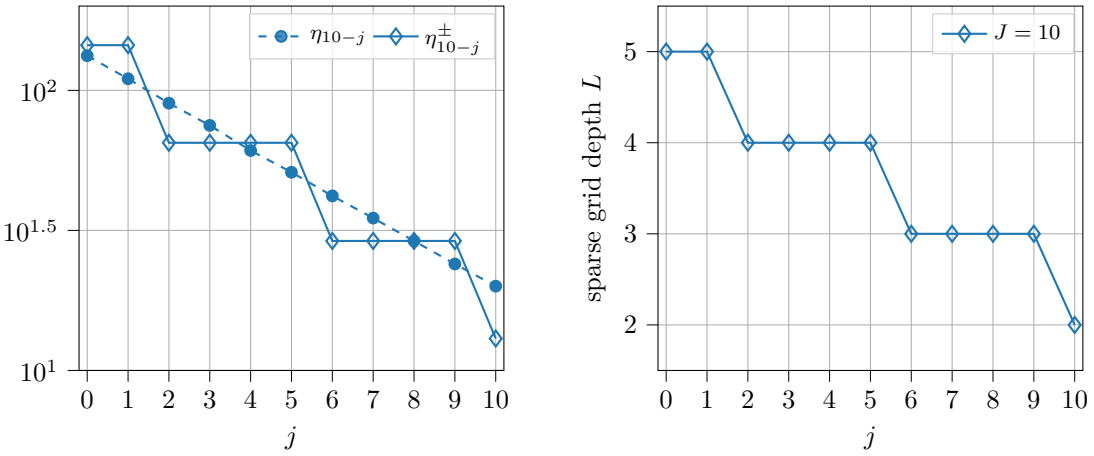
- By "error in $\mathbb{E}$", we mean the spatially discrete analogue of

$$\left(\sum_{k=1}^{2} \| \mathbb{E}[u_{J,k}^{(\mathrm{ML})} - u_{L_{\mathrm{ref}},\tau_{\mathrm{ref}},k}(T)] \|_{L^2(D)}^2 \right)^{1/2}$$

- By "error in mass", we mean the spatially discrete analogue of the quantity

$$\sum_{k=1}^{2} \left| \mathbb{E}\left[\|u_{J,k}^{(\mathrm{ML})}\|_{L^1(D)} - \|u_{L_{\mathrm{ref}},\tau_{\mathrm{ref}},k}(T)\|_{L^1(D)} \right] \right|.$$

Figure 4.8 shows for the particular choice of $\varepsilon = 10^{-8}$ how many levels are required, how many collocation points belong to each level and which sparse grid depths correspond to these levels.



(a) η_{J-j} and $\eta^{\pm}_{J-j}$

(b) Sparse grid depths corresponding to $\eta^{\pm}_{J-j}$

Figure 4.8: $\eta^{\pm}_{J-j}$ and corresponding sparse grid depths for $\varepsilon = 10^{-8}$ and $J = 10$

Now we compare the work-precision-diagrams from the single-level and multi-level experiments. To this end, we pick a few tolerances and compare the wall times which are required for both approaches to achieve these tolerances. Since we have a setting with very large $\mu = 9.35$, we expect from Example 3.5.5 that the benefits of the multi-level approach will only show up for very small tolerances. This is confirmed

Tolerance	wall time SL	wall time ML
10^{-6}	0.46s	1.42s
10^{-7}	3s	4.2s
10^{-8}	32s	20.7s
$2 \cdot 10^{-9}$	250s	76.5s
10^{-9}	400s (estimated)	117.9s

Table 4.1: Cost comparison for the single- and multi-level methods

by Table 4.1. One can see this as a limitation of the multi-level approach: For problems with very high regularity in the parameter space and large tolerances, the benefit of using more levels disappears. Intuitively, this is clear: If μ is large, then the stochastic discretisation is cheap and a single-level method computed with a rather coarse grid is sufficient. We stress, however, that the large value of μ in this example only comes from the phenomenon of subexponential convergence, which is absent in higher dimensions, as explained in Remark 2.6.5. It should also be noted that it is not clear for the single-level method which combination of depth L and time-step τ is best to achieve a given tolerance, and thus the "best" combination of L and τ to achieve a tolerance ε is not known a priori.

Before we come to the next (and final) example of this chapter, let us briefly discuss the practical meaning of boundary conditions for predator-prey problems. The most relevant boundary conditions in this context are

- homogeneous Neumann boundary conditions ("zero flux"), which are used whenever one is incapable of determining the complete extent of a habitat or the habitat is bounded by a fence, and

- homogeneous Dirichlet boundary conditions, which correspond to "lethal borders", where the individuals are either killed or leave the domain without return,

see [31, Sec. 2.1] for a more detailed discussion. Now we examine a problem with mixed boundary conditions in the sense that one part of the boundary is supplied with Neumann boundary conditions, and the other with Dirichlet conditions.

Example 4.7.2.

Consider a predator-prey system on a domain with a hole, where the outer boundary ∂D_N is supplied with homogeneous Neumann boundary conditions and the inner boundary ∂D_D is supplied with homogeneous Dirichlet boundary conditions. The domain D and its spatial discretisation with linear triangular elements is shown in Figure 4.9.

In this example we assume that the uncertainty of the system is described completely by $d = 10$

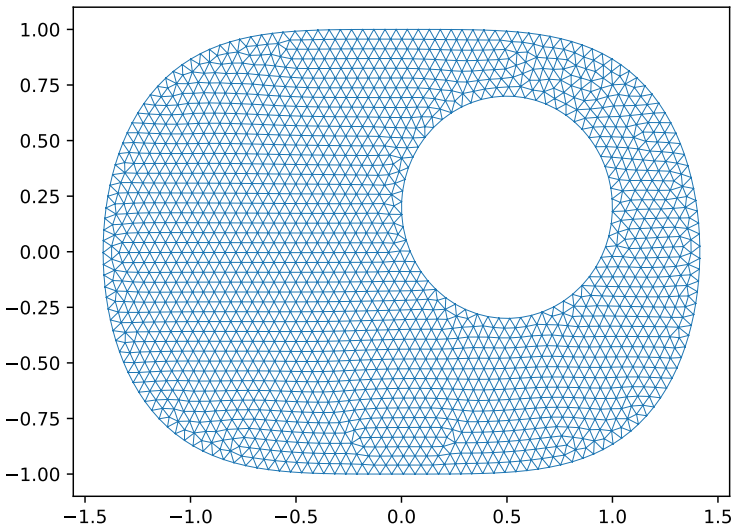


Figure 4.9: Spatial domain with hole and its triangulation; mesh width $h = 0.067$, $\dim(V_h) = 1860$

unknown parameters, so $y = (y_1, \dots, y_{10}) \in [-1, 1]^{10}$. The PDE system is given by

$$\partial_t u_1 = \delta_1(y)\Delta u_1 + R_1(u_1, u_2, y), \qquad \text{in } [0,T] \times D \times \Gamma, \tag{4.43a}$$

$$\partial_t u_2 = \delta_2(y)\Delta u_2 + R_2(u_1, u_2, y), \qquad \text{in } [0,T] \times D \times \Gamma, \tag{4.43b}$$

$$u_1(0, x, y) = u_{1,0}(x, y), \qquad \text{for } (x, y) \in D \times \Gamma, \tag{4.43c}$$

$$u_2(0, x, y) = u_{2,0}(x, y), \qquad \text{for } (x, y) \in D \times \Gamma, \tag{4.43d}$$

$$\frac{\partial u_1}{\partial \nu} = \frac{\partial u_2}{\partial \nu} = 0, \qquad \text{on } [0,T] \times \partial D_N \times \Gamma, \tag{4.43e}$$

$$u_1 = u_2 = 0, \qquad \text{on } [0,T] \times \partial D_D \times \Gamma, \tag{4.43f}$$

with polynomial reaction terms

$$R_1(u_1, u_2, y) = u_1(1 - u_1) - u_2 h(a(y)u_1),$$
$$R_2(u_1, u_2, y) = b(y)u_2 h(a(y)u_1) - c(y)u_2$$

as in (4.2) from the introductory section of this chapter, where

$$h(w) = h_{\text{Hol}}(w) = \frac{w}{1 + w}.$$

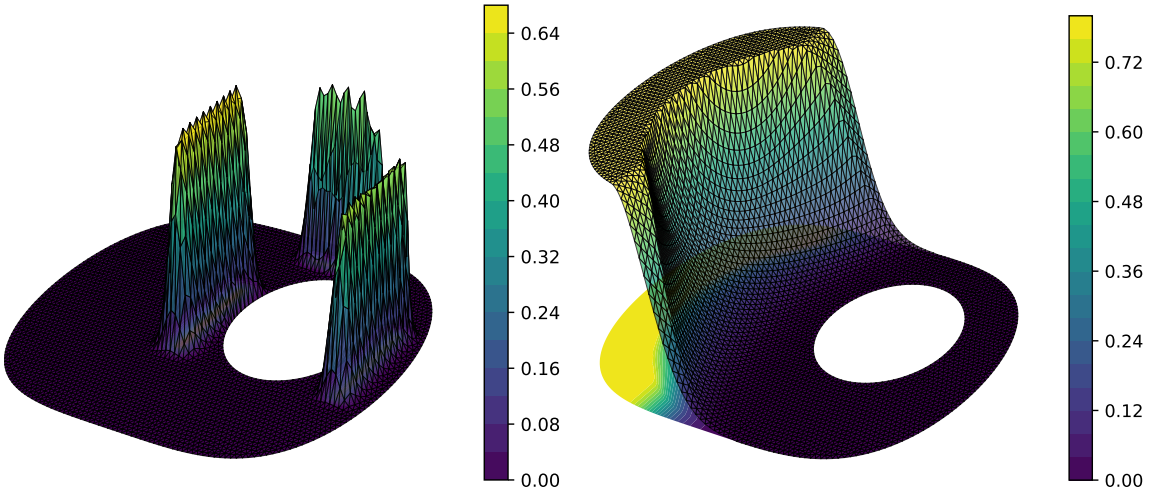
The uncertain parameters $\delta_1, \delta_2, a, b, c$ are given by

$$\begin{aligned} \delta_1(y) &= 0.2 + 0.025y_1, \\ \delta_2(y) &= 0.3 + 0.015y_2, \\ a(y) &= 1 + 0.5y_3 + 0.25y_4, \\ b(y) &= 2 + y_5 + 0.5y_6, \\ c(y) &= 0.1 + 0.05y_7 + 0.025y_8. \end{aligned}$$

The parameters y_9 and y_{10} come from the initial distributions of prey and predators given by

$$u_{1,0}(x, y) = (1 + 0.5y_9)\,\bar{u}_1(x) \qquad \text{and} \qquad u_{2,0}(x, y) = (1 + 0.5y_{10})\,\bar{u}_2(x)$$

for $\bar{u}_1$ and $\bar{u}_2$ as depicted in Figure 4.10. These functions are in fact smooth, but have large gradients in some areas.



(a) Initial prey configuration $\bar{u}_1$ (b) Initial predator configuration $\bar{u}_2$

Figure 4.10: Initial configurations of prey and predators

A weak formulation of this problem can be derived almost as in Remark 4.6.3. The main difference here is that a Dirichlet boundary appears in addition to the Neumann boundary. The Dirichlet boundary condition can be incorporated into the weak formulation by replacing the space $H^1(D)$ with $H^1_{\partial D_D}(D) = \{v \in H^1(D) \colon v|_{\partial D_D} = 0\}$. A similar change affects the spatially discrete system.

We continue with the description of the setting from an application point of view. The initial configuration depicted in Figure 4.10 could describe a scenario where predators invade a habitat around a "lethal area" (e.g. an area surrounded by an electric fence) from the west, and three groups of prey are located around this area. The outer Neumann boundary says that equally many individuals (of both species) enter and leave the habitat such that the total flux over this boundary is zero, and the inner Dirichlet boundary ensures that no individuals enter or cross the lethal area. A reasonable question for a ranger of this habitat could be "how many of the prey individuals will survive the invasion?" and a quantity of interest is $\mathbb{E}[\Upsilon]$ for

$$\Upsilon(y) = \int_D u_1(T, x, y) \mathrm{d}x \tag{4.44}$$

at a time T. The larger the quantity Υ, the more individuals of the prey survived the invasion. The reaction and diffusion constants in mean might be available from past observations, and a guess on their variability, too.

Let us proceed with the simulation of the system. Since this problem is high-dimensional in the parameter space, we aim for a simpler goal than approximating the solutions u_1 and u_2 completely, since the spatial distribution of both species might not be important. So instead of approximating u_1 and u_2 in the whole domain with a small error, we consider the quantity Υ from (4.44) instead, which we want to approximate with a small error (in expectation). Thus, we use the multi-level approach for QoIs outlined in Section 3.6.

The constants were computed as $\mu = 1.41$, $C = C_I C_\zeta = 2312$, $C_T = 40.87$ and $\beta = \alpha = 2$. The

maximal step-size is chosen as $\tau_0 = 0.05$. Results of the multi-level approach for $T = 1$ are shown in Figure 4.11 with three different rounding strategies.

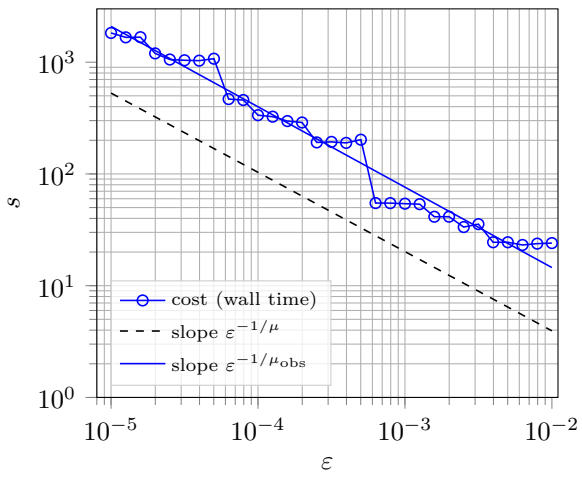
We see from Figure 4.11 that the theoretical scaling of the MLSC method is confirmed, even slightly improved in case of the "down" rounding strategy as $1.50 = \mu_{\text{obs}} > \mu = 1.41$. The scaling in case of the "up/down" rounding strategy is noticeably smaller.

The reference value Υ_{ref} for Υ is computed by a single-level method with $L_{\text{ref}} = 5$ ($\eta_{L_{\text{ref}}} = 41265$) and $\tau_{\text{ref}} = 10^{-5}$. We have $\mathbb{E}[\Upsilon_{\text{ref}}] = 0.063$. The error behaviour shown in Figure 4.11(b) is also in agreement with the theory, as $|\mathbb{E}[\Upsilon_J^{(\text{ML})} - \Upsilon_{\text{ref}}]|$ (denoted by "error in Υ" in the picture) is almost always below the given tolerance ε. For large tolerances, we even observe that the error is much smaller than expected. The errors in $\|\cdot\|_*$ and $\mathbb{E}$ are included for completeness, but we stress that the multi-level approach was used to approximate Υ, and thus no prediction about these errors can be made.

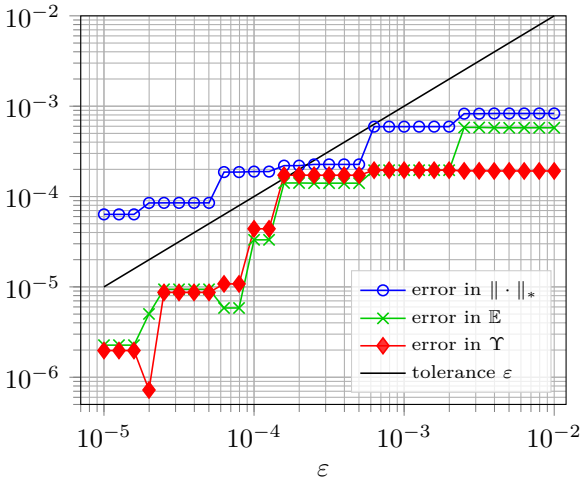
The approximation for $\varepsilon = 10^{-4}$ at times $t_* \in \{0.25, 0.5, 1\}$ is depicted in Figure 4.12 – Figure 4.14 on the following pages. (Note the different scalings of the colorbars in these images.) These snapshots of the multi-level approximation are available as all of these times t_* are multiples of the maximal step-size $\tau_0 = 0.05$.

A test run for a longer time interval with final time $T = 3$ is shown in Figure 4.15. Here, the constants are computed as $\mu = 0.98$, $C = 273$, $C_T = 37.7$ and $\beta = \alpha = 2$. The approximation for $\varepsilon = 10^{-4}$ at the final time $T = 3$ is shown in Figure 4.16.

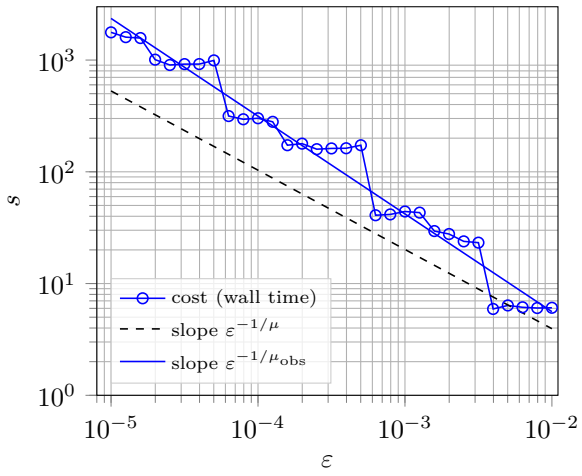
We conclude this example with a description of the behaviour of the solution from the application side. In the beginning of the time interval, the dynamics are only driven by diffusion since mathematically speaking, the supports of predator and prey populations are disjoint and there is no interaction between the species. As the prey is out of reach for the predators, the total amount of predators decreases in the beginning. Both species spread over the whole domain and their spatial distributions become smoother as expected. At time $t = 1$, the spatial distribution of u_1 and u_2 is almost as homogeneous as possible for these boundary conditions. From now on the dynamics are dictated by the reaction between the species. As the prey population is very small compared to the amount of predators, the predator density decreases further. This is clearly visible from the images at time $t = 3$ in Figure 4.16. The prey population recovers slightly between $t = 1$ and $t = 3$. Note that the variance of u_2 becomes smaller and smaller over time, so we conclude that the state of u_2 is very certain despite the input uncertainty in the system.



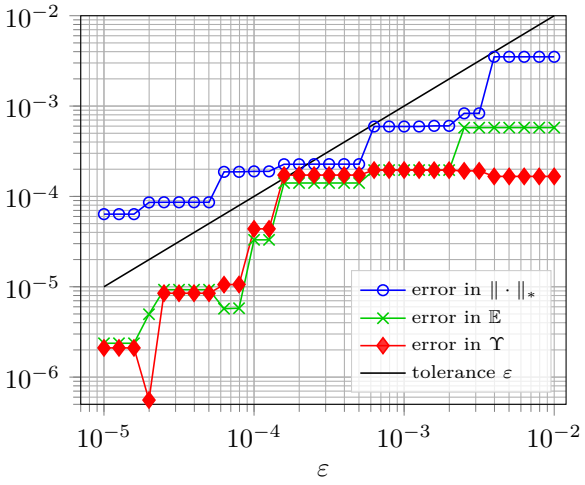
(a) Cost vs. given ε, rounding "up", $\mu_{\text{obs}} = 1.39$



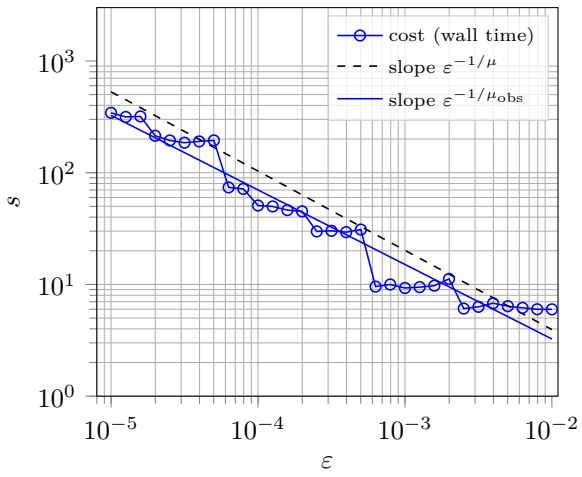
(b) Error vs. given ε, rounding "up"



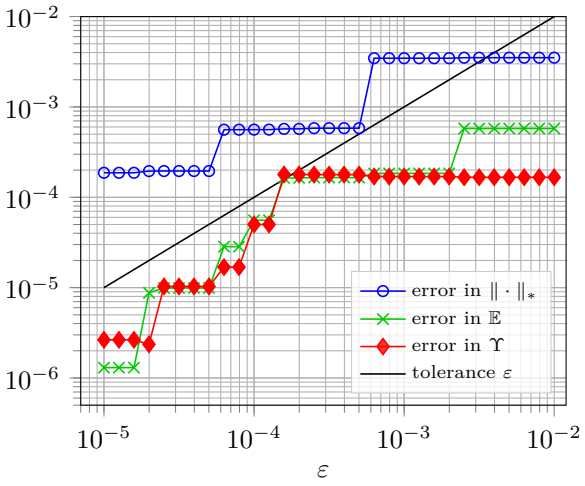
(c) Cost vs. given ε, rounding "up/down", $\mu_{\text{obs}} = 1.14$



(d) Error vs. given ε, rounding "up/down"

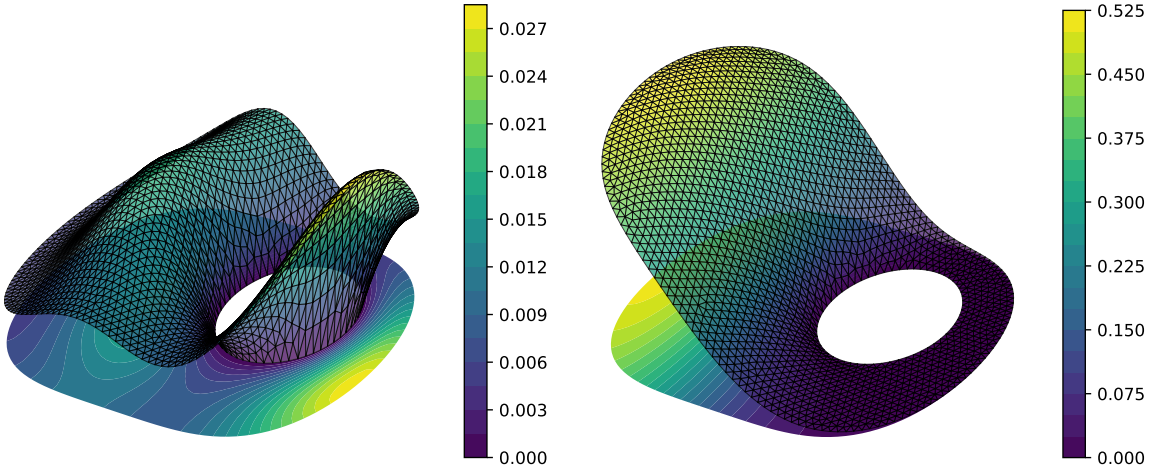


(e) Cost vs. given ε, rounding "down", $\mu_{\text{obs}} = 1.50$



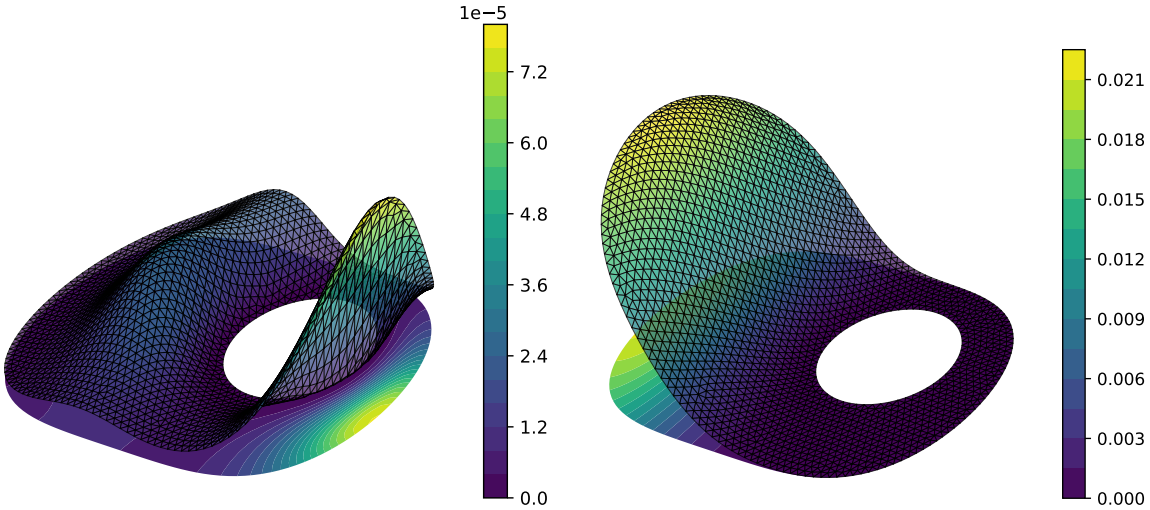
(f) Error vs. given ε, rounding "down"

Figure 4.11: MLSC for the predator-prey system (4.43): $d = 10$, $T = 1$, $\mu = 1.41$, $h = 0.067$



(a) Approx. to $\mathbb{E}[u_1(0.25, x, \cdot)]$

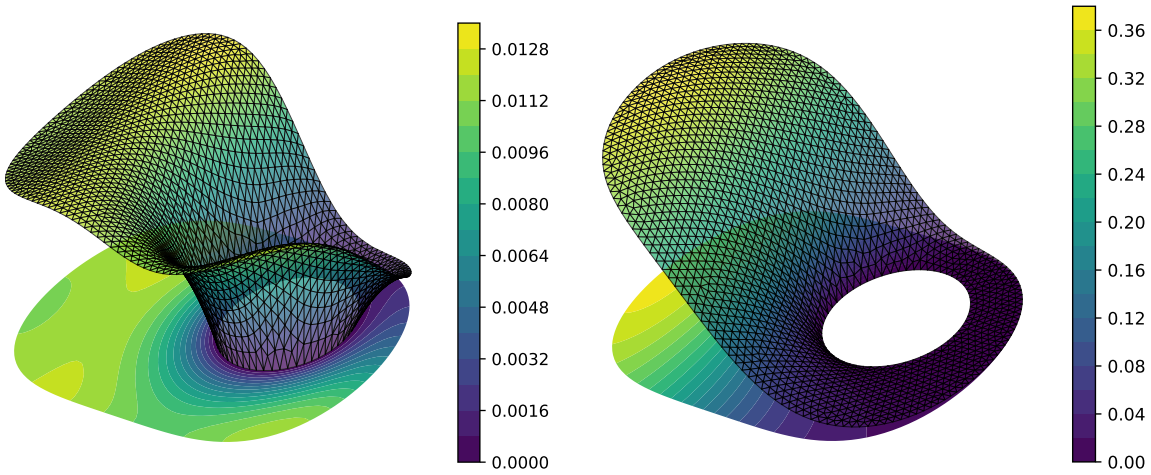
(b) Approx. to $\mathbb{E}[u_2(0.25, x, \cdot)]$



(c) Approx. to $\mathbb{V}[u_1(0.25, x, \cdot)]$

(d) Approx. to $\mathbb{V}[u_2(0.25, x, \cdot)]$

Figure 4.12: Approximations at time $t = 0.25$ computed with MLSC + IMEXT



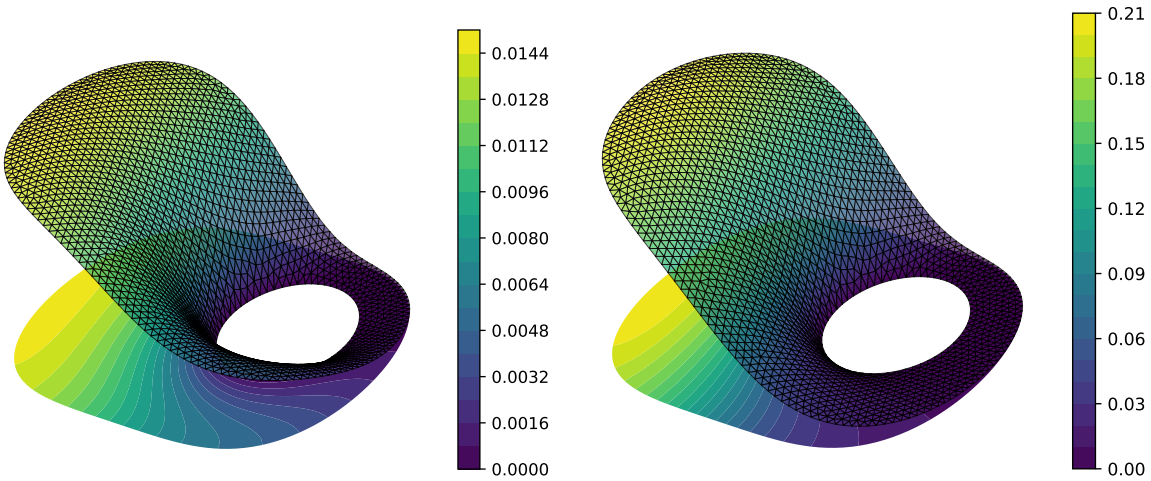
(a) Approx. to $\mathbb{E}[u_1(0.5, x, \cdot)]$

(b) Approx. to $\mathbb{E}[u_2(0.5, x, \cdot)]$

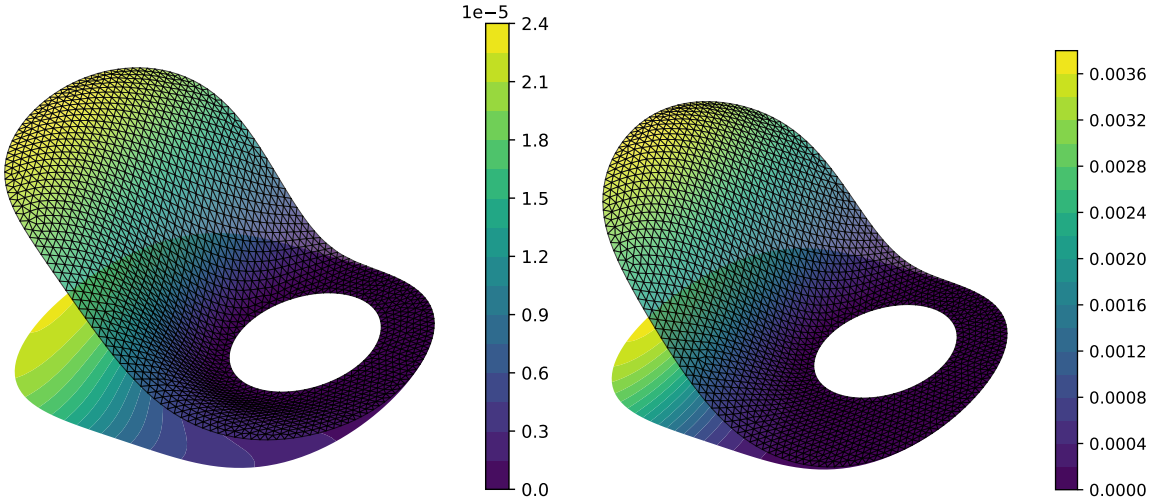
(c) Approx. to $\mathbb{V}[u_1(0.5, x, \cdot)]$

(d) Approx. to $\mathbb{V}[u_2(0.5, x, \cdot)]$

Figure 4.13: Approximations at time $t = 0.5$ computed with MLSC + IMEXT

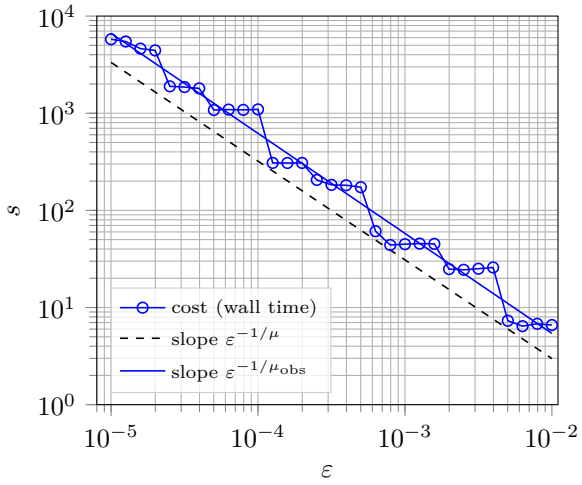


(a) Approx. to $\mathbb{E}[u_1(1,x,\cdot)]$ (b) Approx. to $\mathbb{E}[u_2(1,x,\cdot)]$

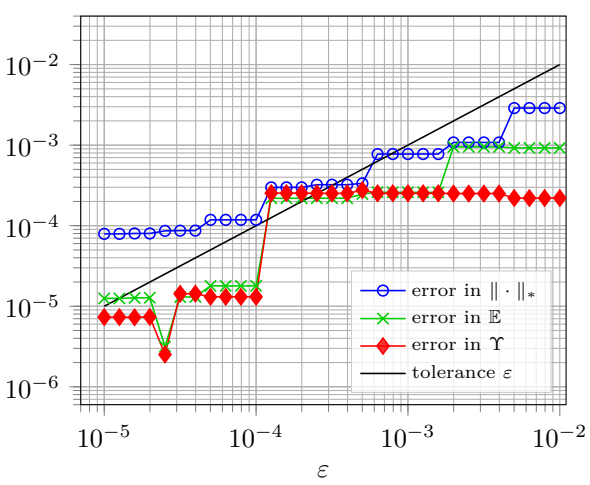


(c) Approx. to $\mathbb{V}[u_1(1,x,\cdot)]$ (d) Approx. to $\mathbb{V}[u_2(1,x,\cdot)]$

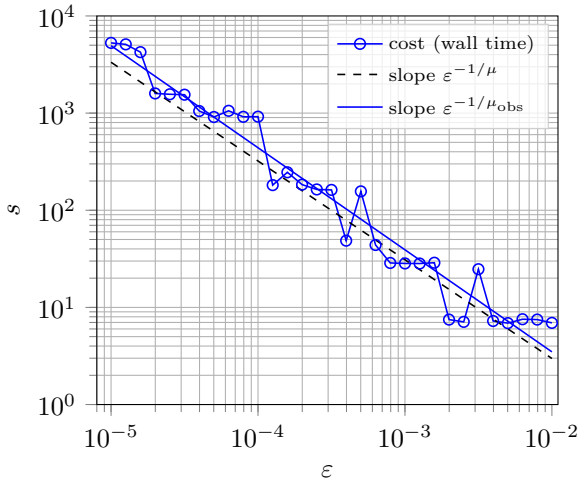
Figure 4.14: Approximations at time $t = 1$ computed with MLSC + IMEXT



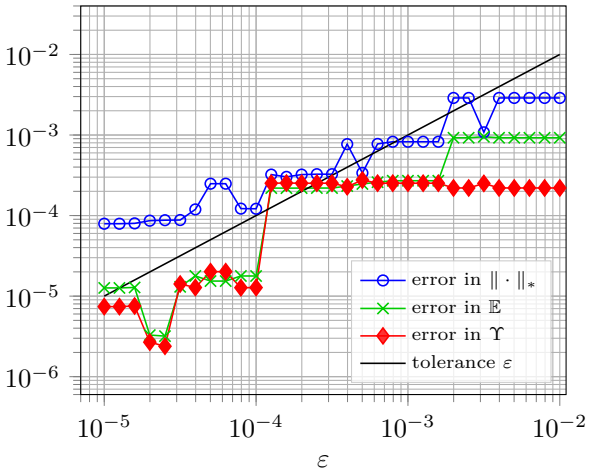
(a) Cost vs. given ε, rounding "up", $\mu_{\text{obs}} = 0.97$



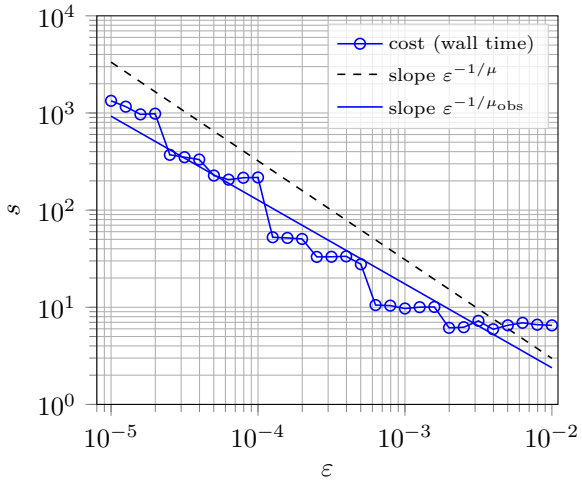
(b) Error vs. given ε, rounding "up"



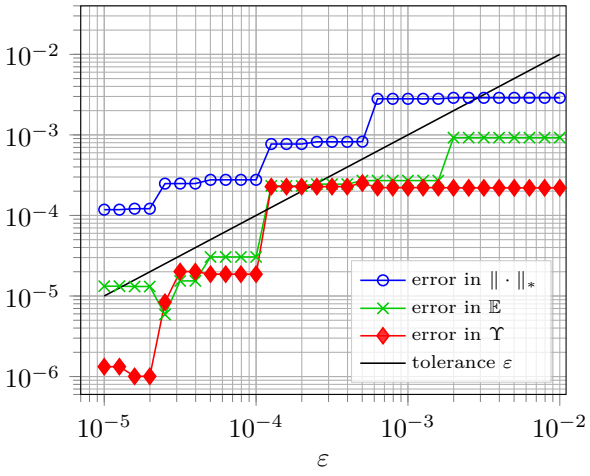
(c) Cost vs. given ε, rounding "up/down", $\mu_{\text{obs}} = 0.95$



(d) Error vs. given ε, rounding "up/down"

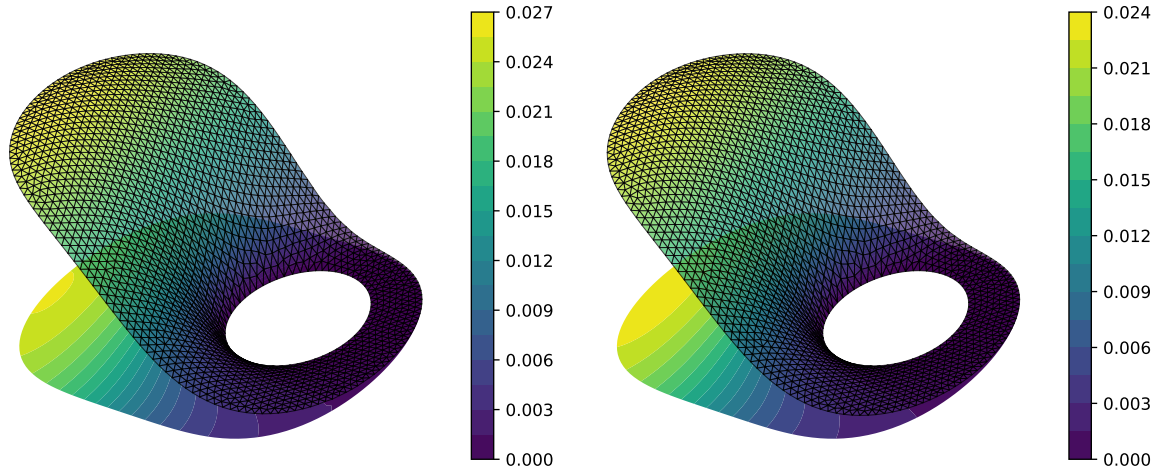


(e) Cost vs. given ε, rounding "down", $\mu_{\text{obs}} = 1.16$



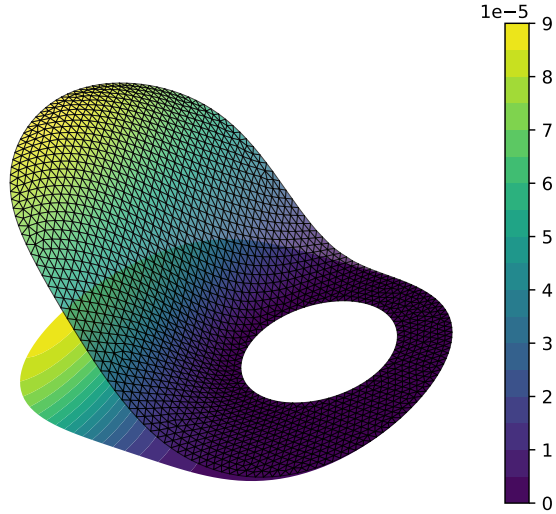
(f) Error vs. given ε, rounding "down"

Figure 4.15: MLSC for the predator-prey system (4.43): $d = 10$, $T = 3$, $\mu = 0.98$, $h = 0.067$

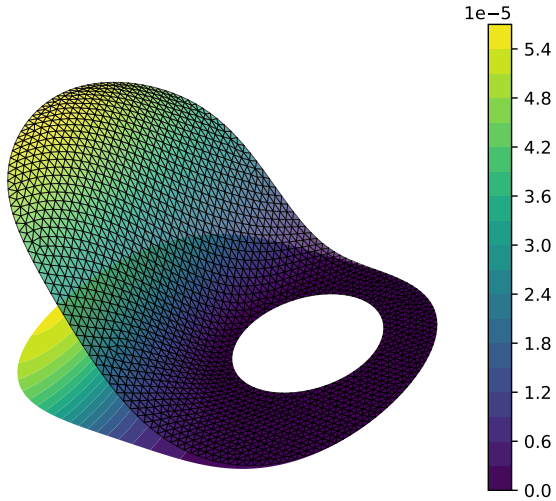


(a) Approx. to $\mathbb{E}[u_1(3,x,\cdot)]$

(b) Approx. to $\mathbb{E}[u_2(3,x,\cdot)]$



(c) Approx. to $\mathbb{V}[u_1(3,x,\cdot)]$



(d) Approx. to $\mathbb{V}[u_2(3,x,\cdot)]$

Figure 4.16: Approximations at time $t = 3$ computed with MLSC + IMEXT

CHAPTER 5

Multi-level stochastic collocation for Schrödinger equations

5.1 Motivation

The *non-linear Schrödinger equation* (NLS) is an important partial differential equation in non-linear optics and condensed matter physics. It appears in the context of *Bose-Einstein condensates*[1], for example, where it is usually called *Gross-Pitaevskii equation* (GPE).

Let us briefly describe this specific field of research to explain the role of the NLS there. A Bose-Einstein condensate (BEC) is a gas of bosons which are in a certain identical quantum state. The specific quantum state of a BEC is achieved by cooling a gas of very low density to temperatures close to absolute zero. Bosons are particles that follow Bose-Einstein statistics and thus have integer spin. Among them are fundamental particles such as photons and the famous Higgs boson, but also composite particles such as deuterium, helium-4 and many alkaline isotopes. Nowadays, laboratories all over the world are able to routinely produce BECs despite the fact that they are unstable. BECs have been able to answer a variety of questions in fundamental physics and are still intensively under study today. For an easily accessible introduction to the topic, see [101, 63] (German). For a more recent book in English, see e.g. [79].

Let us now discuss why the NLS is central to this research area and how it is derived. As each of the bosons in a BEC is in the same quantum state (ground state), they have the same wave function, which we denote by ψ. In practice, this only happens if the temperature is basically absolute zero. A free quantum particle is described by a single-particle Schrödinger equation, but modelling a system of $N_{\mathrm{bos}} \in \mathbb{N}$ bosons should take interactions between the particles into account. The Hartree-Fock approximation uses the product ansatz

$$\mathbf{\Psi}(t, x^{(1)}, \dots, x^{(N_{\mathrm{bos}})}) = \psi(t, x^{(1)}) \cdots \psi(t, x^{(N_{\mathrm{bos}})})$$

for the (time-dependent) wave-function $\mathbf{\Psi}$ of the complete system, where $x^{(j)} \in \mathbb{R}^3$ are the positions of the bosons. For low densities in which the distance between the particles is comparatively larger than their scattering length (a gas with this property is said to be *dilute*), the interactions can be

[1] The term is dedicated to Satyendra Nath Bose and Albert Einstein, both of whom were pioneers in this field.

approximated via a pseudopotential. For extremely low temperatures, this pseudopotential approach yields the (distributional) Hamiltonian

$$H = \sum_{j=1}^{N_{\text{bos}}} \left(-\frac{\hbar^2}{2m} \Delta_{x^{(j)}} + V(x^{(j)}) \right) + \sum_{k=j+1}^{N_{\text{bos}}} \frac{4\pi\hbar^2 a_s}{m} \delta(x^{(j)} - x^{(k)}),$$

where $\hbar$ is the reduced Planck constant, m the (identical) mass of the bosons, V the external potential, δ the Dirac delta and a_s the boson-boson s-wave scattering length. The latter quantity models the scattering process in the low-temperature setting. Using Heisenberg's equation of motion, one arrives at the (mean-field) equation

$$\mathrm{i}\hbar\partial_t \psi(t,x) = \left(-\frac{\hbar^2}{2m}\Delta + V(x) + \chi|\psi(t,x)|^2 \right) \psi(t,x) \tag{5.1}$$

with m and V from before and a quantity $\chi = 4\pi\hbar^2 a_s/m$ which represents particle interactions. This is the way the Gross-Pitaevskii equation is usually formulated. The normalisation implied here is

$$\int_{\mathbb{R}^3} |\psi(t,x)|^2 \mathrm{d}x = N_{\text{bos}}.$$

Although equation (5.1) above is formulated in three-dimensional space, its two- and one-dimensional variants (with $x \in \mathbb{R}^2$ or $x \in \mathbb{R}$) represent important special cases and are thus also studied in the physical and mathematical literature. We refer to [7, Sec. 2.3] for an explanation of the dimension reduction and in which physically meaningful scenarios it is valid. Simply put, the two- and one-dimensional variants correspond to condensates which are "disc-shaped" or "cigar-shaped" ([7, p. 324, 326]).

We stress that the applicability of the NLS in BECs is strictly limited to temperatures very close to absolute zero since the temperature dependence is not taken into account in its derivation. The NLS does not describe condensates of photons and some other particles which are in ground state at room temperature.

In applications, the external potential V is often quadratic, modelling a harmonic trap. Thus, the potential energy of a particle increases quadratically with its distance from the centre.

Another external potential of interest is a helical one modeling a rotating optical dipole trap, described in detail in [90]. It should be noted that such a potential is time-dependent and hence it does not strictly fall into the setting we discuss in this chapter (although it could be incorporated with some changes in the procedure).

The NLS, its numerical solution and physical importance are discussed in [24, 7] and references therein. Observe that if we set $\chi = 0$ in (5.1), we obtain a *linear Schrödinger equation* (LSE). It is an important equation on its own and fundamental to quantum mechanics. However, we regard it as a special case of the non-linear Schrödinger equation.

Let us now discuss how uncertainties naturally occur in this equation. The quantities $\hbar$ and m are known and do not introduce uncertainty. The external potential V is the main source of uncertainty, but also the quantity χ is unknown since it depends on the scattering length a_s. The initial state of the gas $\psi(0,\cdot)$ is typically unknown, too. As a consequence, we will introduce once again a parameter y to this equation which accounts for the uncertainty in the potential V, the scattering length[2] a_s and the initial state $\psi(0,\cdot)$.

Let us now describe our problem setting in detail.

[2] more precisely: in χ, which is proportional to a_s

5.2 Problem setting

The content of the following sections will appear in a similar form in [61] for a *linear* Schrödinger equation. Here we consider a parametric non-linear Schrödinger equation for $u\colon \mathbb{R}_+ \times \mathbb{T}^N \times \Gamma \to \mathbb{C}$ given by

$$\partial_t u(t,x,y) = \mathrm{i}\Delta u(t,x,y) + \mathrm{i}V(x,y)u(t,x,y) + \mathrm{i}\chi|u(t,x,y)|^2 u(t,x,y), \quad t \geq 0,\ x \in \mathbb{T}^N,\ y \in \Gamma, \tag{5.2a}$$

$$u(0,x,y) = u_0(x,y), \qquad x \in \mathbb{T}^N,\ y \in \Gamma \tag{5.2b}$$

with spatial dimension $N \in \mathbb{N}$, $\mathbb{T} = \mathbb{R}/(2\pi\mathbb{Z})$, $\Gamma = [-1,1]^d$, $V\colon \mathbb{T}^N \times \Gamma \to \mathbb{R}$, $\chi \in \mathbb{R}$ and initial value $u_0\colon \mathbb{T}^N \times \Gamma \to \mathbb{C}$. Despite having changed the notation $\psi \rightsquigarrow u$ and applied some harmless transformations, the reader should observe the similarity to the GPE (5.1) from the previous section. Here, t is the temporal variable, x the spatial variable and y is some parameter which accounts for the uncertainty in the potential V and the initial data u_0. The variable y is new compared to the previous section. The choice of the spatial domain $\mathbb{T}^N$ corresponds to imposing periodic boundary conditions. These boundary conditions are a standard choice in the literature as they admit an efficient spatial discretisation via Fourier collocation methods. The spatial dimension N is not to be confused with the number of bosons N_{bos} from the introductory section. In the context of BECs, we have $N \in \{1,2,3\}$, but N_{bos} is typically a large integer of approximate size between 10^2 and 10^7 according to [7, Eq. (2.4)].

We note that the results in this chapter may be extended to the case where the parameter χ in the non-linear term is a function which depends on y.

Most of the time, the spatial variable x will be hidden in our exposition since we regard u as a function in the two variables t and y which takes values in a Banach space X. This Banach space will often be $L^2(\mathbb{T}^N)$, but occasionally also some higher-order Sobolev spaces $H^s(\mathbb{T}^N)$ with $s \in \mathbb{N}$ appear.

With this convention, (5.2) can be formulated as a parameter-dependent Cauchy problem

$$\partial_t u(t,y) = \mathrm{i}\Delta u(t,y) + \mathrm{i}V(y)u(t,y) + \mathrm{i}\chi|u(t,y)|^2 u(t,y), \qquad t \geq 0,\ y \in \Gamma, \tag{5.3a}$$

$$u(0,y) = u_0(y), \qquad y \in \Gamma, \tag{5.3b}$$

where the function $u\colon \mathbb{R}_+ \times \Gamma \to L^2(\mathbb{T}^N)$ is sought-after.

Now we discuss the time integration of (5.3) via the Strang splitting method.

5.3 Strang splitting

Despite the fact that some solution formulas for the NLS may be derived for specific initial values u_0 and potentials V, the rather complicated structure of the NLS demands a reliable numerical solver in general. The most prominent methods for this purpose are splitting methods such as the second-order Strang splitting[3]. Combined with a space discretisation via trigonometric polynomials (for periodic boundary conditions as in our case) or Hermite polynomials (in case of the full space $\mathbb{R}^N$ without boundary conditions), this method is second-order convergent and stable under reasonable assumptions. We refer to [33] for an in-depth analysis of such a splitting method in case of the full space $\mathbb{R}^N$ with a generic smooth potential. More comments on the usage of Strang splitting in the literature will be given at the end of Section 5.3.1.

[3] Higher-order splitting methods are also possible for the NLS, see [110], but are used less frequently.

Now we describe the Strang splitting method in detail and then study how the additional parameter y affects this method. Later in this section, our focus will be on the numerical analysis of the error of this method in the norm of the space $C^{\boldsymbol{k}}(\Gamma, L^2(\mathbb{T}^N))$.

5.3.1 Description of the method

The splitting method we use for (5.3) is derived by dividing the problem into the two subproblems

$$\partial_t v(t,y) = \mathrm{i}(V(y) + \chi|v(t,y)|^2)v(t,y), \qquad t \geq 0,\ y \in \Gamma, \tag{5.4a}$$
$$v(0,y) = v_0(y), \qquad y \in \Gamma. \tag{5.4b}$$

and

$$\partial_t w(t,y) = \mathrm{i}\Delta w(t,y), \qquad t \geq 0,\ y \in \Gamma, \tag{5.5a}$$
$$w(0,y) = w_0(y), \qquad y \in \Gamma. \tag{5.5b}$$

Both of these problems are simpler than the "full" problem in the following sense: Equation (5.4) does not contain spatial derivatives anymore and reduces to an ODE in each (spatial) grid point after performing a space discretisation, whereas (5.5) is a linear equation. We introduce the abbreviation

$$B[v,y] = V(y) + \chi|v(y)|^2. \tag{5.6}$$

The solution of (5.4) is given by

$$v(t,y) = \mathrm{e}^{\mathrm{i}tB[v_0,y]}v_0(y) \tag{5.7}$$

for any $t \geq 0$ and $y \in \Gamma$. This is not obvious, so a proof is provided in Lemma A.2 in Appendix A.

The solution of (5.5) is given by

$$w(t,y) = \mathrm{e}^{\mathrm{i}t\Delta}w_0(y),$$

where $(\mathrm{e}^{\mathrm{i}t\Delta})_{t\in\mathbb{R}}$ is the strongly continuous group generated by $\mathrm{i}\Delta$ on $L^2(\mathbb{T}^N)$. In fact, this group is unitary on each of the spaces $H^s(\mathbb{T}^N)$ for $s \in \mathbb{N}_0$ by Stone's theorem [28, Thm. II.3.24].

Now that the solutions of both subproblems are known, we can combine them to obtain an approximation with step-size $\tau > 0$ to the full problem (5.3) as follows:

1. Solve (5.4) over a time-step $\tau/2$ with the result from the previous time-step as starting value.
2. Solve (5.5) over a time-step τ with the result from 1. as starting value.
3. Solve (5.4) over a time-step $\tau/2$ with the result from 2. as starting value.
4. Repeat steps 1.–3. until a final time T is reached.

In a formula, the corresponding numerical flow can be written as

$$\Phi_\tau(v,y) = \mathrm{e}^{\frac{\tau}{2}\mathrm{i}B[u^+,y]}u^+(y), \qquad \text{where} \qquad u^+(y) = \mathrm{e}^{\tau\mathrm{i}\Delta}\mathrm{e}^{\frac{\tau}{2}\mathrm{i}B[v,y]}v(y).$$

Thus, we successively compute approximations $u_n(y) \approx u(t_n, y)$ at times $t_n = n\tau$ via

$$u_n(y) \coloneqq \Phi_\tau(u_{n-1}, y), \qquad n = 1, 2, \ldots, \tag{5.8}$$

where $\tau > 0$ is a given step-size and $y \in \Gamma$. If we set

$$\Phi_\tau^n(u_0, y) \coloneqq \Phi_\tau(\Phi_\tau^{n-1}(u_0, y), y), \qquad n \in \mathbb{N}, \qquad \Phi_\tau^0(u_0, y) = u_0(y), \tag{5.9}$$

then (5.8) can be cast into the form

$$u_n(y) = \Phi_\tau^n(u_0, y), \qquad n \in \mathbb{N}_0,\ y \in \Gamma.$$

Observe that the flow $\Phi_\tau(\cdot, y)$ preserves the $L^2(\mathbb{T}^N)$-norm for any $y \in \Gamma$. In the next section, we show that the method is convergent of order 2 in the norm of $L^2(\mathbb{T}^N)$ under reasonable assumptions. Thus, we will be able to efficiently compute good approximations for the solution of (5.2) for any given value of $y \in \Gamma$.

As we can only compute approximations for finitely many values of y in practice, we have to characterise the regularity of the solution and its approximations with respect to the variable y in order to successfully apply a stochastic collocation method. A result which is given in the next section implies that the splitting approximation has the regularity $C^{\boldsymbol{k}}$ with respect to the variable y if the solution u and the potential V already belong to the class $C^{\boldsymbol{k}}$. This will be crucial to obtain good convergence rates for the approximations in the parameter space later on.

Before we continue with the error analysis, let us briefly comment on the usage of this method in the literature. Strang splitting has been applied to numerous problems by different authors, see e.g. [59, 73, 60, 110, 7, 33] and references therein for applications to different Schrödinger equations. Nowadays the Strang splitting scheme seems to be the most attractive method for the numerical simulation of the NLS in the literature. Besides its efficiency, the preservation of the $L^2(\mathbb{T}^N)$-norm is an important feature in applications. In the context of BECs presented in the introductory section of this chapter, this preservation corresponds to a constant amount of bosons. Similar interpretations remain true in other physical applications. It should be noted, however, that finite difference and other spectral schemes are also present in the literature [76], but usually not preferred in light of the discussion at the end of [7, Sec. 3.1].

Let us now discuss the main results of our error analysis for the Strang splitting scheme and the assumptions under which they are valid.

5.3.2 Error analysis: The results

In the following, we consider the Banach spaces $C^{\boldsymbol{k}}(\Gamma, H^s(\mathbb{T}^N))$ for $\boldsymbol{k} \in \mathbb{N}_0^d$ and $s \in \mathbb{N}_0$ and abbreviate

$$\|w\|_{\boldsymbol{k},s} \coloneqq \|w\|_{C^{\boldsymbol{k}}(\Gamma, H^s(\mathbb{T}^N))} = \max_{\boldsymbol{0} \le \boldsymbol{j} \le \boldsymbol{k}} \|\partial_y^{\boldsymbol{j}} w\|_{C(\Gamma, H^s(\mathbb{T}^N))} \tag{5.10}$$

for $w \in C^{\boldsymbol{k}}(\Gamma, H^s(\mathbb{T}^N))$. Note that the case $s = 0$ corresponds to the space $L^2(\mathbb{T}^N)$, so $\|\cdot\|_{\boldsymbol{0},0}$ is the usual norm of the space $C(\Gamma, L^2(\mathbb{T}^N))$.

We make the following assumptions.

Assumption E1. Let $s > \max\left\{\frac{N}{2}, 1\right\}$ be an integer and $\boldsymbol{k} \in \mathbb{N}_0^d$. Assume that there is a function

$$u \in C([0,T], C^{\boldsymbol{k}}(\Gamma, H^{s+2}(\mathbb{T}^N))) \cap C^1([0,T], C^{\boldsymbol{k}}(\Gamma, H^s(\mathbb{T}^N))) \cap C^2([0,T], C^{\boldsymbol{k}}(\Gamma, L^2(\mathbb{T}^N))) \tag{5.11}$$

that solves (5.2), in particular $u(0, \cdot, \cdot) = u_0$.

This strong assumption cannot be easily verified. In the *linear* case corresponding to $\chi = 0$, however, it is possible to prove similar regularity under assumptions on u_0 and V. This will be discussed later in Section 5.3.6, in particular in Theorem 5.3.10.

The requirement $s > \frac{N}{2}$ in Assumption E1 implies that $H^s(\mathbb{T}^N)$ is an algebra. This property plays a central role throughout the error analysis and allows us to incorporate the non-linearity correctly. We also need the following assumption on the potential.

Assumption E2. Let $V \in C^{\boldsymbol{k}}(\Gamma, H^{s+2}(\mathbb{T}^N))$ for $\boldsymbol{k} = (k_1, \dots, k_d) \in \mathbb{N}_0^d$, so in particular

$$\|\partial_{x_1}^{r_1} \cdots \partial_{x_N}^{r_N} \partial_{y_1}^{m_1} \cdots \partial_{y_d}^{m_d} V\|_{\mathbf{0},0} < \infty$$

for $|\boldsymbol{r}|_1 \leq s+2$ and $\boldsymbol{m} = (m_1, \dots, m_d) \leq \boldsymbol{k}$.

Notation. As $u(t, \cdot, \cdot)$ belongs to $C^{\boldsymbol{k}}(\Gamma, H^{s+2}(\mathbb{T}^N))$ for each $t \in [0, T]$ by Assumption E1, it is reasonable to write $u(t)$ instead of $u(t, \cdot, \cdot)$ and omit both variables x and y in the notation. Other functions will be treated similarly. Thus, we use the following convention from now on:

The variables x and y are omitted in the notation if possible.

This allows us to write expressions like $\|u(t)\|_{\boldsymbol{k},s}$ for any $t \in [0, T]$ and reduces the number of function arguments in lengthy formulas which will appear later on.

Now we present the main results of our error analysis.

Theorem 5.3.1 (Local error)**.** *Assume that Assumptions E1 and E2 hold with the same values of s and $\boldsymbol{k}$. The local error of the Strang splitting method with initial value $v = u(0)$ and step-size $0 < \tau \leq T$ is bounded by*

$$\|\Phi_\tau(v) - u(\tau)\|_{\boldsymbol{k},s} \leq C\tau^2, \tag{5.12}$$

$$\|\Phi_\tau(v) - u(\tau)\|_{\boldsymbol{k},0} \leq \widetilde{C}\tau^3, \tag{5.13}$$

where C and $\widetilde{C}$ depend on

$$\max_{t\in[0,\tau]} \|u(t)\|_{\boldsymbol{k},s+2} \qquad \text{and} \qquad \|V\|_{C^{\boldsymbol{k}}(\Gamma, H^{s+2}(\mathbb{T}^N))}.$$

From this *local* error bound, the following *global* result will be deduced.

Theorem 5.3.2 (Global error)**.** *Assume that Assumptions E1 and E2 hold with the same values of s and $\boldsymbol{k}$. Then there exists a step-size $\tau_0 > 0$ such that the global error of the Strang splitting method after n steps with step-size $\tau \in (0, \tau_0]$ is bounded by*

$$\|\Phi_\tau^n(u_0) - u(t_n)\|_{\boldsymbol{k},s} \leq C\tau, \tag{5.14}$$

$$\|\Phi_\tau^n(u_0) - u(t_n)\|_{\boldsymbol{k},0} \leq \widetilde{C}\tau^2, \tag{5.15}$$

as long as $t_n = n\tau \leq T$. The constants C and $\widetilde{C}$ depend on t_n,

$$M_{\boldsymbol{k}}^{(s+2)} := \max_{t\in[0,t_n]} \|u(t)\|_{\boldsymbol{k},s+2} \qquad \text{and} \qquad \|V\|_{C^{\boldsymbol{k}}(\Gamma, H^{s+2}(\mathbb{T}^N))}.$$

If the reader is not primarily interested in the proof of this convergence result, we recommend to skip the next three subsections completely. Our advise is then to continue reading either in Section 5.3.6, where simplifications for the linear Schrödinger equation are presented, or in Section 5.4, where stochastic collocation methods are discussed for the NLS.

Brave readers who are interested in the proofs are now confronted with some tools to compute multivariate derivatives of u and its numerical approximation.

5.3.3 Interlude: Multivariate differentiation formulas

In this technical section we state multivariate product and chain rules for certain differential operators which we need in our error analysis later on. The idea of this framework is taken from [50], although our notation is slightly different. We discovered recently that a similar framework was also used in an UQ context in [9, App. A].

We introduce some notation. Let

$$\mathscr{D}_{\boldsymbol{\eta}} \coloneqq \frac{\partial^m}{\partial y_{\eta_1} \cdots \partial y_{\eta_m}} \qquad \text{for} \qquad \boldsymbol{\eta} = (\eta_1, \dots, \eta_m) \in \{1, \dots, d\}^m.$$

The vector $\boldsymbol{\eta}$ contains the (perhaps multiply occuring) directions of the partial derivatives. Its length m is exactly total order of the differential operator $\mathscr{D}_{\boldsymbol{\eta}}$. Let $M = \{1, \dots, m\}$. We can associate a multi-index to $\boldsymbol{\eta}$ which contains the number of derivatives in each direction,

$$\boldsymbol{k}(\boldsymbol{\eta}) = (|\{s \in M \colon \eta_s = 1\}|, \dots, |\{s \in M \colon \eta_s = d\}|) \in \mathbb{N}_0^d.$$

Note that $|\boldsymbol{k}(\boldsymbol{\eta})|_1 = m$. Since we may exchange the order of the derivatives if the corresponding functions are continuously differentiable up to the required order, it holds

$$\mathscr{D}_{\boldsymbol{\eta}} f = \frac{\partial^m f}{\partial y_{\eta_1} \cdots \partial y_{\eta_m}} = \frac{\partial^m f}{\partial y^{\boldsymbol{k}(\boldsymbol{\eta})}}$$

for such functions f. If one tries to write down a "product rule" for such a differential operator $\mathscr{D}_{\boldsymbol{\eta}}$, i.e. a formula for $\mathscr{D}_{\boldsymbol{\eta}}(fg)$, one observes that all differential operators of lower order than $\mathscr{D}_{\boldsymbol{\eta}}$ (i.e. operators with some of the ∂y_{η_j} omitted) are necessary to write it down. To make this more precise, we now associate differential operators to *subsets* of $M = \{1, \dots, m\}$. This allows us later to state product and chain rules which look familiar.

We fix $\boldsymbol{\eta}$ and the corresponding value of m. (The following definitions only make sense for given $\boldsymbol{\eta}$ and m, but we do not indicate this in the notation.) For a set $S \subseteq M = \{1, \dots, m\}$, we define

$$\frac{\partial^{|S|}}{\partial y^S} \coloneqq \frac{\partial^{|S|}}{\prod_{j \in S} \partial y_{\eta_j}}.$$

As a special case, we have

$$\frac{\partial^m}{\partial y^M} = \frac{\partial^m}{\partial y_{\eta_1} \cdots \partial y_{\eta_m}} = \mathscr{D}_{\boldsymbol{\eta}}.$$

The power set of S is denoted by $\mathcal{P}^S$. We further define

$$\mathcal{P}_*^S = \mathcal{P}^S \setminus \{\emptyset\} \qquad \text{and} \qquad \mathcal{P}_{**}^S = \mathcal{P}_*^S \setminus \{S\}.$$

The set of partitions of S into non-empty subsets ("blocks") is denoted by $\Pi(S)$. For a set $S \subseteq M$, the complement S^c is always understood as the complement in M, i.e. $S^c = M \setminus S$.

Example 5.3.3.

Let $M = \{1, 2, 3\}$. Then the five elements of $\Pi(M)$ are the following.

$$\begin{aligned}
&\text{Partitions with 1 block: } \{\{1,2,3\}\}\\
&\text{Partitions with 2 blocks: } \{\{1\},\{2,3\}\}, \quad \{\{2\},\{1,3\}\}, \quad \text{and} \quad \{\{3\},\{1,2\}\}\\
&\text{Partitions with 3 blocks: } \{\{1\},\{2\},\{3\}\}
\end{aligned}$$

We have

$$\begin{aligned}
\mathcal{P}_*^M &= \big\{\{1\},\{2\},\{3\},\{1,2\},\{2,3\},\{1,3\},M\big\},\\
\mathcal{P}_{**}^M &= \big\{\{1\},\{2\},\{3\},\{1,2\},\{2,3\},\{1,3\}\big\}.
\end{aligned}$$

The empty set $\emptyset$ has exactly one partition, namely $\emptyset$ itself.

Now that a suitable notation is available, the multivariate chain rule (also known as *Faà di Bruno's formula*) may be stated in the form

$$\frac{\partial^{|S|}}{\partial y^S} f(g(y)) = \sum_{\pi \in \Pi(S)} f^{|\pi|}(g(y)) \prod_{B \in \pi} \frac{\partial^{|B|} g(y)}{\partial y^B} \tag{5.16}$$

for a set $S \subseteq M$, where $|\pi|$ is the number of blocks in the partition π and f^j denotes the j-th derivative of f (and not the j-th power). The multivariate product rule may be stated in the form

$$\frac{\partial^{|S|}}{\partial y^S}(fg) = \sum_{T \in \mathcal{P}^S} \frac{\partial^{|T|} f}{\partial y^T} \frac{\partial^{|S\setminus T|} g}{\partial y^{S\setminus T}}. \tag{5.17}$$

We refer to [50, Prop. 1 and 5] for proofs of these equations and instead give an example for both formulas.

Example 5.3.4.

The following examples are taken from Ex. 1 and 2 and the beginning of Sec. 6 in [50].

- For $\boldsymbol{\eta} = (1, 2, 3)$ and hence $m = 3$, (5.16) yields the formula

$$\begin{aligned}
\frac{\partial^3}{\partial y_1\,\partial y_2\,\partial y_3} f(g(y)) &= f'(g(y)) \frac{\partial^3 g(y)}{\partial y_1\,\partial y_2\,\partial y_3}\\
&\quad + f''(g(y)) \left(\frac{\partial g(y)}{\partial y_1} \cdot \frac{\partial^2 g(y)}{\partial y_2\,\partial y_3} + \frac{\partial g(y)}{\partial y_2} \cdot \frac{\partial^2 g(y)}{\partial y_1\,\partial y_3} + \frac{\partial g(y)}{\partial y_3} \cdot \frac{\partial^2 g(y)}{\partial y_1\,\partial y_2} \right)\\
&\quad + f'''(g(y)) \frac{\partial g(y)}{\partial y_1} \cdot \frac{\partial g(y)}{\partial y_2} \cdot \frac{\partial g(y)}{\partial y_3}.
\end{aligned}$$

  By (5.17), we have

$$\begin{aligned}
\frac{\partial^3}{\partial y_1\,\partial y_2\,\partial y_3}(fg) &= f \cdot \frac{\partial^3 g}{\partial y_1\,\partial y_2\,\partial y_3} + \frac{\partial f}{\partial y_1} \cdot \frac{\partial^2 g}{\partial y_2\,\partial y_3} + \frac{\partial f}{\partial y_2} \cdot \frac{\partial^2 g}{\partial y_1\,\partial y_3} + \frac{\partial f}{\partial y_3} \cdot \frac{\partial^2 g}{\partial y_1\,\partial y_2}\\
&\quad + \frac{\partial^2 f}{\partial y_1\,\partial y_2} \cdot \frac{\partial g}{\partial y_3} + \frac{\partial^2 f}{\partial y_1\,\partial y_3} \cdot \frac{\partial g}{\partial y_2} + \frac{\partial^2 f}{\partial y_2\,\partial y_3} \cdot \frac{\partial g}{\partial y_1} + \frac{\partial^3 f}{\partial y_1\,\partial y_2\,\partial y_3} \cdot g.
\end{aligned}$$

- For $\boldsymbol{\eta} = (1,2,2)$ and $\boldsymbol{k}(\boldsymbol{\eta}) = (1,2)$, (5.16) yields

$$\frac{\partial^3}{\partial y_1 \partial y_2^2} f(g(y)) = f'(g(y)) \frac{\partial^3 g(y)}{\partial y_1 \partial y_2^2} + f'''(g(y)) \frac{\partial g(y)}{\partial y_1} \cdot \left(\frac{\partial g(y)}{\partial y_2} \right)^2 + f''(g(y)) \left(\frac{\partial g(y)}{\partial y_1} \cdot \frac{\partial^2 g(y)}{\partial y_2^2} + \frac{\partial g(y)}{\partial y_2} \cdot \frac{\partial^2 g(y)}{\partial y_1 \partial y_2} + \frac{\partial g(y)}{\partial y_2} \cdot \frac{\partial^2 g(y)}{\partial y_1 \partial y_2} \right).$$

In fact, we use a slightly more general product rule than the one stated in (5.17). To explain it, let X_1, X_2 and Z be general Banach spaces. We say that a function $P\colon X_1 \times X_2 \to Z$ is a *product* if P is bilinear and continuous in the sense that there is $C > 0$ such that $\|P(x_1, x_2)\|_Z \leq C \|x_1\|_{X_1} \|x_2\|_{X_2}$ holds for all $x_1 \in X_1$ and $x_2 \in X_2$. Since there is a product rule for functions of the form $y \mapsto P(f(y), g(y))$ (see [97, Kap. 2, Satz 2.7]) and the combinatorics of higher derivatives remain the same as in formula (5.17), we get

$$\frac{\partial^{|S|}}{\partial y^S} P(f,g) = \sum_{T \in \mathcal{P}^S} P\left(\frac{\partial^{|T|} f}{\partial y^T}, \frac{\partial^{|S \setminus T|} g}{\partial y^{S \setminus T}} \right).$$

We use this formula later for

$$P\colon \mathcal{L}(X_2, Z) \times X_2 \to Z, \qquad P(A,x) = Ax$$

and

$$P\colon H^s(\mathbb{T}^N) \times H^s(\mathbb{T}^N) \to H^s(\mathbb{T}^N), \qquad P(u,v) = u \mathrm{e}^{\mathrm{i}\tau\Delta} v$$

for $s > \frac{N}{2}$, which are both products.

Now we state several other results which will be needed later on.

5.3.4 Preliminaries for the error analysis

The following lemma is crucial for the treatment of the non-linearity in (5.2). It is an extension of the well-known product estimate

$$\|vw\|_r \leq C \|v\|_{r^*} \|w\|_r \qquad \text{with} \qquad r^* = \max\left\{ r, \lfloor \tfrac{N}{2} \rfloor + 1 \right\} \tag{5.18}$$

for $v \in H^{r^*}(\mathbb{T}^N)$ and $w \in H^r(\mathbb{T}^N)$ which is a consequence of the Sobolev embedding theorem, see [10, Sec. 2.1]. Here and in the following, we always set

$$r^* = \max\left\{ r, \lfloor \tfrac{N}{2} \rfloor + 1 \right\}.$$

It is easy to see that $\lfloor x \rfloor + 1$ is the smallest integer which is strictly larger than x.

Lemma 5.3.5. *Let $\boldsymbol{k} \in \mathbb{N}_0$ and $r \in \mathbb{N}_0$. If $v \in C^{\boldsymbol{k}}(\Gamma, H^{r^*}(\mathbb{T}^N))$ and $w \in C^{\boldsymbol{k}}(\Gamma, H^r(\mathbb{T}^N))$, then also $vw \in C^{\boldsymbol{k}}(\Gamma, H^r(\mathbb{T}^N))$ with*

$$\|vw\|_{\boldsymbol{k},r} \leq C \|v\|_{\boldsymbol{k},r^*} \|w\|_{\boldsymbol{k},r}$$

for a constant $C > 0$ independent of v and w.

For the proof, we set $m = |\boldsymbol{k}|_1$, $M = \{1, \dots, m\}$ and

$$\boldsymbol{\eta} = (\underbrace{1, \dots, 1}_{k_1}, \underbrace{2, \dots, 2}_{k_2}, \dots, \underbrace{d, \dots, d}_{k_d}) \in \{1, \dots, d\}^m.$$

The notation introduced in Section 5.3.3 will be used throughout the error analysis.

Proof. The product rule (5.17) implies that

$$\frac{\partial^{|S|}(vw)}{\partial y^S} = \sum_{T\in\mathcal{P}^S} \frac{\partial^{|T|}v}{\partial y^T}\frac{\partial^{|S\setminus T|}w}{\partial y^{S\setminus T}}$$

for $S \subseteq M$. By applying the norm of $C(\Gamma, H^r(\mathbb{T}^N))$, the triangle inequality and (5.18), we get

$$\begin{aligned}
\left\|\frac{\partial^{|S|}(vw)}{\partial y^S}\right\|_{\mathbf{0},r} &\le C \sum_{T\in\mathcal{P}^S} \left\|\frac{\partial^{|T|}v}{\partial y^T}\right\|_{\mathbf{0},r^*} \left\|\frac{\partial^{|S\setminus T|}w}{\partial y^{S\setminus T}}\right\|_{\mathbf{0},r} \\
&\le C \max_{T\in\mathcal{P}^S} \left\|\frac{\partial^{|T|}v}{\partial y^T}\right\|_{\mathbf{0},r^*} \sum_{T\in\mathcal{P}^S} \left\|\frac{\partial^{|S\setminus T|}w}{\partial y^{S\setminus T}}\right\|_{\mathbf{0},r} \\
&\le C \max_{T\in\mathcal{P}^S} \left\|\frac{\partial^{|T|}v}{\partial y^T}\right\|_{\mathbf{0},r^*} \max_{T\in\mathcal{P}^S} \left\|\frac{\partial^{|S\setminus T|}v}{\partial y^{S\setminus T}}\right\|_{\mathbf{0},r} \\
&\le C\|v\|_{\boldsymbol{k},r^*}\|w\|_{\boldsymbol{k},r}.
\end{aligned}$$

As $S \subseteq M$ was arbitrary, the statement follows. □

Another simple consequence of Lemma 5.3.5 and the definition of B in (5.6) is the following result.

Corollary 5.3.6. *If $V \in C^{\boldsymbol{k}}(\Gamma, H^r(\mathbb{T}^N))$ and $w \in C^{\boldsymbol{k}}(\Gamma, H^{r^*}(\mathbb{T}^N))$, then also $B[w] \in C^{\boldsymbol{k}}(\Gamma, H^r(\mathbb{T}^N))$ and*

$$\|B[w]\|_{\boldsymbol{k},r} \le \|V\|_{\boldsymbol{k},r} + |\chi|\|w\|_{\boldsymbol{k},r^*}\|w\|_{\boldsymbol{k},r}.$$

The above result will be frequently used throughout the next section and will not be mentioned explicitly.

The next lemma is a variant of [60, Lem. 3] with y-dependency and is used later in the final step of the proofs of the local error bounds. Here and in the entire next section, we abbreviate

$$B_*(w) = \frac{\mathrm{i}}{2}B[w]. \tag{5.19}$$

Lemma 5.3.7. *Assume that Assumption E1 holds with $u(0) = v$. Let $u^+(\tau) = \mathrm{e}^{i\Delta\tau}\mathrm{e}^{B_*(v)\tau}v$ and set*

$$b(\tau) = \frac{\partial^{|S|}}{\partial y^S}\left[B[u(\tau)] - B[u^+(\tau)]\right] = \chi\frac{\partial^{|S|}}{\partial y^S}\left[|u(\tau)|^2 - |u^+(\tau)|^2\right].$$

Then we have

$$b(\tau) = \int_0^\tau \partial_\tau b(r)\mathrm{d}r = \int_0^\tau\int_0^r \partial_\tau^2 b(\widetilde{r})\mathrm{d}\widetilde{r}\mathrm{d}r$$

with

$$\begin{aligned}
\partial_\tau b(\tau) &= 2\chi\frac{\partial^{|S|}}{\partial y^S}\Big(\operatorname{Re}(\overline{u(\tau)}\partial_\tau u(\tau)) - 2\operatorname{Re}(\overline{u^+(\tau)}\partial_\tau u^+(\tau))\Big), \\
\partial_\tau^2 b(\tau) &= 2\chi\frac{\partial^{|S|}}{\partial y^S}\operatorname{Re}\left(\overline{u(\tau)}\partial_\tau^2 u(\tau) + \overline{\partial_\tau u(\tau)}\partial_\tau u(\tau) - \overline{u^+(\tau)}\partial_\tau^2 u^+(\tau) - \overline{\partial_\tau u^+(\tau)}\partial_\tau u^+(\tau)\right),
\end{aligned}$$

where $\partial_\tau b(\tau) \in C(\Gamma, H^s(\mathbb{T}^N))$ and $\partial_\tau^2 b(\tau) \in C(\Gamma, L^2(\mathbb{T}^N))$ are uniformly bounded in $\tau \in [0, T]$.

Proof. By linearity of differentiation in y, it is enough to show the statement for $S = \emptyset$. By the fundamental theorem of calculus, we have

$$b(\tau) = b(0) + \int_0^\tau \partial_\tau b(r)\mathrm{d}r = b(0) + \tau\partial_\tau b(0) + \int_0^\tau\int_0^r \partial_\tau^2 b(\widetilde{r})\mathrm{d}\widetilde{r}\mathrm{d}r.$$

Since $u(0) = v$, it is clear that $b(0) = 0$. Now we show $\partial_\tau b(0) = 0$. We have

$$\partial_\tau b(\tau) = 2\chi \operatorname{Re}(\overline{u(\tau)}\partial_\tau u(\tau)) - 2\chi \operatorname{Re}(\overline{u^+(\tau)}\partial_\tau u^+(\tau)) \tag{5.20}$$

by the product rule. Since $v = u(0)$ by Assumption E1, it holds

$$\begin{aligned}\partial_\tau u|_{\tau=0} &= \mathrm{i}\Delta v + \mathrm{i}B[v]v,\\ \partial_\tau u^+|_{\tau=0} &= \mathrm{i}\Delta v + B_*(v)v,\end{aligned}$$

such that

$$\partial_\tau b(0) = 2\chi \operatorname{Re}(\overline{v}\partial_\tau u(0)) - 2\chi \operatorname{Re}(\overline{v}\partial_\tau u^+(0)) = 2\chi \operatorname{Re}(\overline{v}\mathrm{i}B[v]v) - 2\chi \operatorname{Re}(\overline{v}B_*(v)v) = 0,$$

because both functions inside Re are purely imaginary. The second derivative can be obtained by differentiating (5.20). To see the boundedness, observe that

$$\begin{aligned}\partial_\tau u^+(\tau) &= \mathrm{i}\Delta \mathrm{e}^{\tau \mathrm{i}\Delta}\mathrm{e}^{\tau B_*(v)}v + \mathrm{e}^{\tau \mathrm{i}\Delta}\mathrm{e}^{\tau B_*(v)}B_*(v)v,\\ \partial_\tau^2 u^+(\tau) &= (\mathrm{i}\Delta)^2\mathrm{e}^{\tau \mathrm{i}\Delta}\mathrm{e}^{\tau B_*(v)}v + 2\mathrm{i}\Delta \mathrm{e}^{\tau \mathrm{i}\Delta}\mathrm{e}^{\tau B_*(v)}B_*(v)v + \mathrm{e}^{\tau \mathrm{i}\Delta}\mathrm{e}^{\tau B_*(v)}B_*^2(v)v.\end{aligned}$$

Now we apply $\|\cdot\|_{\mathbf{0},s}$ to the first equation and $\|\cdot\|_{\mathbf{0},0}$ to the second one. By Assumption E1, we get that these expressions are uniformly bounded in $\tau \in [0, T]$. □

5.3.5 Error analysis: The proofs

We start the proof of the two main theorems with the proof of the local error bound in $H^s(\mathbb{T}^N)$. This choice is made for two reasons: First, it is the simpler and shorter proof of the two local error bounds and second, the *global* error bound in $L^2(\mathbb{T}^N)$ requires the *local* error bound in $H^s(\mathbb{T}^N)$ such that the $L^2(\mathbb{T}^N)$-results somehow depend on the $H^s(\mathbb{T}^N)$-results. Thus, this subsection is structured as follows.

1. Proof of the local error bound (5.12) in $H^s(\mathbb{T}^N)$
2. Proof of the local error bound (5.13) in $L^2(\mathbb{T}^N)$
3. Proof of two stability results
4. Proof of the global error bound (5.15) in $L^2(\mathbb{T}^N)$
5. Proof of the global error bound (5.14) in $H^s(\mathbb{T}^N)$ (sketched)

In the proofs, we use the notation

$$\mathcal{O}_{\boldsymbol{k},s}(\tau^j)$$

for $j \in \mathbb{N}$ to indicate that a function is $\mathcal{O}(\tau^j)$ as $\tau \to 0$ with respect to the norm $\|\cdot\|_{\boldsymbol{k},s}$.

Proof of (5.12)

As a preparation, we set $m = |\boldsymbol{k}|_1$, $M = \{1, \dots, m\}$,

$$\boldsymbol{\eta} = (\underbrace{1, \dots, 1}_{k_1}, \underbrace{2, \dots, 2}_{k_2}, \dots, \underbrace{d, \dots, d}_{k_d}) \in \{1, \dots, d\}^m$$

and abbreviate

$$\mathscr{D} = \mathscr{D}_{\boldsymbol{\eta}} = \frac{\partial^m}{\partial y^M} = \frac{\partial^{|\boldsymbol{k}|_1}}{\partial y^{\boldsymbol{k}}} \tag{5.21}$$

with the notation from Section 5.3.3. Recall the Cauchy problem for the NLS

$$\partial_t u(t) = \mathrm{i}\Delta u(t) + \mathrm{i}B[u(t)]u(t), \qquad t \geq 0, \tag{5.22a}$$

$$u(0) = v \tag{5.22b}$$

from (5.3). Note that we have set $v = u_0$ in order to distinguish more clearly between the solution u and the initial value v. We only prove that

$$\|\mathscr{D}u(\tau) - \mathscr{D}(\Phi_\tau(v))\|_{\mathbf{0},s} \leq C\tau^2$$

for $\mathscr{D}$ from (5.21), since the procedure for differential operators with lower order than $\mathscr{D}$ is completely analogous.

The proof is divided into five steps.

Step 1: Representation of the solution. For the solution u of (5.22), we have

$$\partial_t \mathscr{D}u(t) = \mathrm{i}\Delta \mathscr{D}u(t) + \mathrm{i}\sum_{S\in\mathcal{P}^M} \frac{\partial^{|S|}B[u(t)]}{\partial y^S}\frac{\partial^{|S^c|}u(t)}{\partial y^{S^c}}$$

and the variation-of-constants formula yields

$$\mathscr{D}u(\tau) = \mathrm{e}^{\mathrm{i}\tau\Delta}\mathscr{D}v + \sum_{S\in\mathcal{P}^M} I_1(S) \tag{5.23}$$

with

$$I_1(S) = \mathrm{i}\int_0^\tau \mathrm{e}^{\mathrm{i}(\tau-r)\Delta}\left(\frac{\partial^{|S|}B[u(r)]}{\partial y^S}\frac{\partial^{|S^c|}u(r)}{\partial y^{S^c}}\right)\mathrm{d}r. \tag{5.24}$$

For the integrand in (5.24), we apply the variation-of-constants formula once again to obtain

$$I_1(S) = \mathrm{i}\int_0^\tau \mathrm{e}^{\mathrm{i}(\tau-r)\Delta}\left(\frac{\partial^{|S|}B[u(r)]}{\partial y^S}\mathrm{e}^{\mathrm{i}r\Delta}\frac{\partial^{|S^c|}v}{\partial y^{S^c}}\right)\mathrm{d}r + \mathcal{O}_{\mathbf{0},s}(\tau^2). \tag{5.25}$$

Step 2: Representation of the numerical approximation. We use the notation

$$\Phi_\tau[\widetilde{v}] = \mathrm{e}^{\frac{\tau}{2}\mathrm{i}B[u^+]}\mathrm{e}^{\mathrm{i}\tau\Delta}\mathrm{e}^{\frac{\tau}{2}\mathrm{i}B[v]}\widetilde{v} \qquad \text{with} \qquad u^+ = \mathrm{e}^{\mathrm{i}\tau\Delta}\mathrm{e}^{\frac{\tau}{2}\mathrm{i}B[v]}v$$

(observe the distinction between v and $\widetilde{v}$ and note that $\Phi_\tau[v] = \Phi_\tau(v)$).

By the product rule (5.17), we have

$$\mathscr{D}(\Phi_\tau[v]) = \Phi_\tau[\mathscr{D}v] + \sum_{S\in\mathcal{P}^M_*} \frac{\partial^{|S|}\Phi_\tau}{\partial y^S}\frac{\partial^{|S^c|}v}{\partial y^{S^c}}. \tag{5.26}$$

Using again the product rule (5.17), $B_*(w) = \frac{\mathrm{i}}{2}B[w]$ and Faà di Bruno's formula (5.16), we compute (formally)

$$\begin{aligned}\frac{\partial^{|S|}\Phi_\tau}{\partial y^S} &= \sum_{T\in\mathcal{P}^S}\frac{\partial^{|T|}\mathrm{e}^{\tau B_*(u^+)}}{\partial y^T}\mathrm{e}^{\mathrm{i}\tau\Delta}\frac{\partial^{|S\setminus T|}\mathrm{e}^{\tau B_*(v)}}{\partial y^{S\setminus T}} \\ &= \sum_{T\in\mathcal{P}^S}\sum_{\pi\in\Pi(T)}\sum_{\sigma\in\Pi(S\setminus T)}\tau^{|\pi|+|\sigma|}\prod_{C_1\in\pi}\frac{\partial^{|C_1|}B_*(u^+)}{\partial y^{C_1}}\Phi_\tau\prod_{C_2\in\sigma}\frac{\partial^{|C_2|}B_*(v)}{\partial y^{C_2}}.\end{aligned} \tag{5.27}$$

If we separate the terms with $T = \emptyset$ and $T = S$, we get

$$\frac{\partial^{|S|}\Phi_\tau}{\partial y^S} = \tau f_1(S) + \mathcal{O}_{\mathbf{0},s}(\tau^2)$$

with

$$f_1(S) = \Phi_\tau \frac{\partial^{|S|} B_*(v)}{\partial y^S} + \frac{\partial^{|S|} B_*(u^+)}{\partial y^S}\Phi_\tau. \tag{5.28}$$

Thus, we arrive at

$$\mathscr{D}(\Phi_\tau[v]) = \Phi_\tau[\mathscr{D}v] + \sum_{S\in\mathcal{P}_*^M} \tau f_1(S)\frac{\partial^{|S^c|} v}{\partial y^{S^c}} + \mathcal{O}_{\mathbf{0},s}(\tau^2), \tag{5.29}$$

where expressions like $f_1(S)w$ have to be understood as

$$f_1(S)w = \Phi_\tau\left[\frac{\partial^{|S|} B_*(v)}{\partial y^S} w\right] + \frac{\partial^{|S|} B_*(u^+)}{\partial y^S}\cdot \Phi_\tau[w].$$

Eq. (5.29) still contains the numerical flow Φ_τ in the first term on the right-hand side and also in the definition of f_1, too. The next step is to get rid of it.

Step 3: Eliminate numerical flows Φ_τ. Using the "Taylor expansion" for φ-functions (B.2) in Appendix B, we obtain

$$\mathrm{e}^{\tau B_*(w)} v = \sum_{j=0}^{J-1} \frac{\tau^j}{j!} B_*^j(w)v + \tau^J B_*^J(w)\varphi_J(\tau B_*(w))v, \qquad J\in\mathbb{N}. \tag{5.30}$$

We use this formula to expand the terms containing Φ_τ in (5.29). For any $\widetilde{v}\in L^2(\mathbb{T}^N)$, it holds

$$\begin{aligned}\Phi_\tau[\widetilde{v}] &= \mathrm{e}^{\tau B_*(u^+)}\mathrm{e}^{\mathrm{i}\tau\Delta}\mathrm{e}^{\tau B_*(v)}\widetilde{v}\\ &= \mathrm{e}^{\mathrm{i}\tau\Delta}\mathrm{e}^{\tau B_*(v)}\widetilde{v} + \tau B_*(u^+)\mathrm{e}^{\mathrm{i}\tau\Delta}\mathrm{e}^{\tau B_*(v)}\widetilde{v} + \mathcal{O}_{\mathbf{0},s}(\tau^2)\\ &= \mathrm{e}^{\mathrm{i}\tau\Delta}\widetilde{v} + \tau\left(\mathrm{e}^{\mathrm{i}\tau\Delta}B_*(v) + B_*(u^+)\mathrm{e}^{\mathrm{i}\tau\Delta}\right)\widetilde{v} + \mathcal{O}_{\mathbf{0},s}(\tau^2).\end{aligned} \tag{5.31}$$

Using this for the first and second summand in (5.29), we get

$$\begin{aligned}\mathscr{D}(\Phi_\tau[v]) = \mathrm{e}^{\mathrm{i}\tau\Delta}\mathscr{D}v &+ \tau\left(\mathrm{e}^{\mathrm{i}\tau\Delta}B_*(v) + B_*(u^+)\mathrm{e}^{\mathrm{i}\tau\Delta}\right)\mathscr{D}v\\ &+ \sum_{S\in\mathcal{P}_*^M} \tau\left(\mathrm{e}^{\mathrm{i}\tau\Delta}\frac{\partial^{|S|}B_*(v)}{\partial y^S} + \frac{\partial^{|S|}B_*(u^+)}{\partial y^S}\mathrm{e}^{\mathrm{i}\tau\Delta}\right)\frac{\partial^{|S^c|}v}{\partial y^{S^c}} + \mathcal{O}_{\mathbf{0},s}(\tau^2).\end{aligned} \tag{5.32}$$

In total, we have the representation

$$\mathscr{D}(\Phi_\tau[v]) = \mathrm{e}^{\mathrm{i}\tau\Delta}\mathscr{D}v + \sum_{S\in\mathcal{P}^M} \tau\widetilde{f}_1(S)\frac{\partial^{|S^c|}v}{\partial y^{S^c}} + \mathcal{O}_{\mathbf{0},s}(\tau^2), \tag{5.33}$$

where $\widetilde{f}_1$ is defined as f_1 in (5.28), but all numerical flows Φ_τ are replaced by $\mathrm{e}^{\mathrm{i}\tau\Delta}$. The expression (5.33) is now free from numerical flows Φ_τ and can be compared to the expansion of the solution (5.23) – at least after some quadrature approximations. The next step is to derive these quadrature expressions.

Step 4: Quadrature approximation. Let

$$h(r) := \mathrm{e}^{(\tau-r)\mathrm{i}\Delta}\left(\frac{\partial^{|S|}B[u(r)]}{\partial y^S}\mathrm{e}^{r\mathrm{i}\Delta}\frac{\partial^{|S^c|}v}{\partial y^{S^c}}\right).$$

We approximate $I_1(S)$ from (5.25) by

$$I_1(S) + \mathcal{O}_{\mathbf{0},s}(\tau^2) = \mathrm{i}\int_0^\tau h(r)\mathrm{d}r \approx \frac{\mathrm{i}\tau}{2}\Big(h(0)+h(\tau)\Big) =: I_1^{\square}(S)$$

and use the first-order Peano form of the error of the trapezoidal rule

$$E = I_1^{\square}(S) - \mathrm{i}\int_0^\tau h(r)\mathrm{d}r = \frac{\mathrm{i}\tau}{2}\Big(h(0)+h(\tau)\Big) - \mathrm{i}\int_0^\tau h(r)\mathrm{d}r = -\mathrm{i}\tau^2\int_0^1 (\tfrac{1}{2}-\theta)h'(\theta\tau)\mathrm{d}\theta.$$

The last term can be estimated in a standard way: We have

$$\begin{aligned} h'(r) = &-\mathrm{i}\Delta\mathrm{e}^{(\tau-r)\mathrm{i}\Delta}\left(\frac{\partial^{|S|}B[u(r)]}{\partial y^S}\mathrm{e}^{r\mathrm{i}\Delta}\frac{\partial^{|S^c|}v}{\partial y^{S^c}}\right)\\ &+ 2\chi\mathrm{e}^{(\tau-r)\mathrm{i}\Delta}\left(\frac{\partial^{|S|}(\mathrm{Re}(u(r)\overline{\partial_r u(r)}))}{\partial y^S}\mathrm{e}^{r\mathrm{i}\Delta}\frac{\partial^{|S^c|}v}{\partial y^{S^c}}\right)\\ &+ \mathrm{e}^{(\tau-r)\mathrm{i}\Delta}\left(\frac{\partial^{|S|}B[u(r)]}{\partial y^S}\mathrm{i}\Delta\mathrm{e}^{r\mathrm{i}\Delta}\frac{\partial^{|S^c|}v}{\partial y^{S^c}}\right)\end{aligned}$$

and observe that h' is bounded in $\|\cdot\|_{\mathbf{0},s}$ since $u(r), v \in C^{\boldsymbol{k}}(\Gamma, H^s(\mathbb{T}^N))$ by Assumption E1. For (5.23), we obtain the updated expansion

$$\mathscr{D}u(\tau) = \mathrm{e}^{\tau\mathrm{i}\Delta}\mathscr{D}v + \sum_{S\in\mathcal{P}^M} I_1^{\square}(S) + \mathcal{O}_{\mathbf{0},s}(\tau^2). \tag{5.34}$$

Step 5: Error expansion. Subtracting (5.33) from (5.34) yields

$$\begin{aligned}\mathscr{D}u(\tau) - \mathscr{D}(\Phi_\tau[v]) &= \sum_{S\in\mathcal{P}^M}\left(I_1^{\square}(S) - \tau\widetilde{f}_1(S)\frac{\partial^{|S^c|}v}{\partial y^{S^c}}\right) + \mathcal{O}_{\mathbf{0},s}(\tau^2)\\ &= \sum_{S\in\mathcal{P}^M} E_1(S) + \mathcal{O}_{\mathbf{0},s}(\tau^2),\end{aligned} \tag{5.35}$$

and with B_* from (5.19), we get

$$\begin{aligned}E_1(S) &= \tau\left(\mathrm{e}^{\mathrm{i}\tau\Delta}\frac{\partial^{|S|}B_*(v)}{\partial y^S} + \frac{\partial^{|S|}B_*(u(\tau))}{\partial y^S}\mathrm{e}^{\mathrm{i}\tau\Delta} - \mathrm{e}^{\mathrm{i}\tau\Delta}\frac{\partial^{|S|}B_*(v)}{\partial y^S} - \frac{\partial^{|S|}B_*(u^+)}{\partial y^S}\mathrm{e}^{\mathrm{i}\tau\Delta}\right)\frac{\partial^{|S^c|}v}{\partial y^{S^c}}\\ &= \tau\left(\frac{\partial^{|S|}B_*(u(\tau))}{\partial y^S} - \frac{\partial^{|S|}B_*(u^+)}{\partial y^S}\right)\mathrm{e}^{\mathrm{i}\tau\Delta}\frac{\partial^{|S^c|}v}{\partial y^{S^c}}.\end{aligned}$$

Lemma 5.3.7 implies that $E_1(S)$ is $\mathcal{O}_{\mathbf{0},s}(\tau^2)$. Plugging this into (5.35), we arrive at

$$\mathscr{D}u(\tau) - \mathscr{D}(\Phi_\tau[v]) = \mathcal{O}_{\mathbf{0},s}(\tau^2),$$

which finishes the proof. □

Now we can turn to the proof of the error bound in $L^2(\mathbb{T}^N)$.

Proof of (5.13)

This proof is divided into the same five steps as the previous proof. Again, we only show that

$$\|\mathscr{D}u(\tau) - \mathscr{D}(\Phi_\tau(v))\|_{\mathbf{0},0} \leq C\tau^3,$$

since the procedure for differential operators with lower order than $\mathscr{D}$ is completely analogous. The main differences compared to the previous proof of (5.12) are that the norm $\|\cdot\|_{\mathbf{0},s}$ is replaced by $\|\cdot\|_{\mathbf{0},0}$ and that we need an additional power of τ in the remainder terms.

Step 1: Representation of the solution. Expanding (5.24) once more via the variation-of-constants formula, we obtain

$$\mathscr{D}u(\tau) = \mathrm{e}^{\tau \mathrm{i}\Delta}\mathscr{D}v + \sum_{S\in\mathcal{P}^M}\left(I_1(S) + \sum_{T\in\mathcal{P}^{S^c}} I_2(S,T)\right) \tag{5.36}$$

with

$$I_1(S) = \mathrm{i}\int_0^\tau \mathrm{e}^{\mathrm{i}(\tau-r)\Delta}\frac{\partial^{|S|}B[u(r)]}{\partial y^S}\mathrm{e}^{r\mathrm{i}\Delta}\frac{\partial^{|S^c|}v}{\partial y^{S^c}}\mathrm{d}r, \tag{5.37}$$

$$I_2(S,T) = \mathrm{i}^2\int_0^\tau\int_0^r \mathrm{e}^{\mathrm{i}(\tau-r)\Delta}\frac{\partial^{|S|}B[u(r)]}{\partial y^S}\mathrm{e}^{(r-\widetilde{r})\mathrm{i}\Delta}\frac{\partial^{|T|}B[u(\widetilde{r})]}{\partial y^T}\frac{\partial^{|S^c\setminus T|}u(\widetilde{r})}{\partial y^{S^c\setminus T}}\mathrm{d}\widetilde{r}\mathrm{d}r. \tag{5.38}$$

Step 2: Representation of the numerical approximation. We start with a technical computation. Equation (5.27) yields

$$\frac{\partial^{|S|}\Phi_\tau}{\partial y^S} = \sum_{T\in\mathcal{P}^S}\sum_{\pi\in\Pi(T)}\sum_{\sigma\in\Pi(S\setminus T)}\tau^{|\pi|+|\sigma|}\prod_{C_1\in\pi}\frac{\partial^{|C_1|}B_*(u^+)}{\partial y^{C_1}}\Phi_\tau\prod_{C_2\in\sigma}\frac{\partial^{|C_2|}B_*(v)}{\partial y^{C_2}},$$

again with $B_*(w) = \frac{\mathrm{i}}{2}B[w]$. If we separate the terms with $T=\emptyset$ and $T=S$, we get

$$\begin{aligned}\frac{\partial^{|S|}\Phi_\tau}{\partial y^S} = {}& \sum_{T\in\mathcal{P}^S_{**}}\sum_{\pi\in\Pi(T)}\sum_{\sigma\in\Pi(S\setminus T)}\tau^{|\pi|+|\sigma|}\prod_{C_1\in\pi}\frac{\partial^{|C_1|}B_*(u^+)}{\partial y^{C_1}}\Phi_\tau\prod_{C_2\in\sigma}\frac{\partial^{|C_2|}B_*(v)}{\partial y^{C_2}}\\ &+\sum_{\sigma\in\Pi(S)}\tau^{|\sigma|}\Phi_\tau\prod_{C\in\sigma}\frac{\partial^{|C|}B_*(v)}{\partial y^C} + \sum_{\pi\in\Pi(S)}\tau^{|\pi|}\prod_{C\in\pi}\frac{\partial^{|C|}B_*(u^+)}{\partial y^C}\Phi_\tau.\end{aligned} \tag{5.39}$$

Let us reduce the horror of this equation a bit: First, we only have to keep an eye on the terms of order 1 and 2 in τ, so everything else can be hidden in $\mathcal{O}_{\mathbf{0},0}(\tau^3)$. Second, $T\in\mathcal{P}^S_{**}$ implies than we can only achieve $|\pi|+|\sigma|=2$ in the first line for $|\pi|=|\sigma|=1$, so $\pi=\{T\}$ and $\sigma=\{S\setminus T\}$. Hence,

$$\begin{aligned}\frac{\partial^{|S|}\Phi_\tau}{\partial y^S} = {}& \sum_{T\in\mathcal{P}^S_{**}}\tau^2\frac{\partial^{|T|}B_*(u^+)}{\partial y^T}\Phi_\tau\frac{\partial^{|S\setminus T|}B_*(v)}{\partial y^{S\setminus T}}\\ &+\sum_{\substack{\sigma\in\Pi(S)\\|\sigma|\leq 2}}\tau^{|\sigma|}\left[\Phi_\tau\prod_{C\in\sigma}\frac{\partial^{|C|}B_*(v)}{\partial y^C} + \prod_{C\in\sigma}\frac{\partial^{|C|}B_*(u^+)}{\partial y^C}\Phi_\tau\right] + \mathcal{O}_{\mathbf{0},0}(\tau^3)\\ = {}& \sum_{T\in\mathcal{P}^S_{**}}\tau^2 f_2(S,T) + \tau f_1(S) + \mathcal{O}_{\mathbf{0},0}(\tau^3)\end{aligned} \tag{5.40}$$

with

$$f_1(S) = \Phi_\tau \frac{\partial^{|S|} B_*(v)}{\partial y^S} + \frac{\partial^{|S|} B_*(u^+)}{\partial y^S} \Phi_\tau, \tag{5.41}$$

$$f_2(S,T) = \frac{\partial^{|T|} B_*(u^+)}{\partial y^T} \Phi_\tau \frac{\partial^{|S\setminus T|} B_*(v)}{\partial y^{S\setminus T}} + \frac{1}{2} \Phi_\tau \frac{\partial^{|T|} B_*(v)}{\partial y^T} \frac{\partial^{|S\setminus T|} B_*(v)}{\partial y^{S\setminus T}} + \frac{1}{2} \frac{\partial^{|T|} B_*(u^+)}{\partial y^T} \frac{\partial^{|S\setminus T|} B_*(u^+)}{\partial y^{S\setminus T}} \Phi_\tau. \tag{5.42}$$

The last equality in (5.40) follows from the fact that every partition $\sigma \in \Pi(S)$ with $|\sigma| = 2$ consists of an arbitrary subset $\emptyset \subsetneq T \subsetneq S$ and its complement in S. If we go through all such subsets T and notice that T is the complement of $S \setminus T$, we have counted each partition $\sigma \in \Pi(S)$ with $|\sigma| = 2$ twice. Hence the factor $1/2$ appears in the second and third term in the definition of $f_2(S,T)$.

Now we have to deal with some set-theoretic considerations. In fact, $S \in \mathcal{P}_*^M$ and $T \in \mathcal{P}_{**}^S$ is equivalent to saying that $T \in \mathcal{P}_*^M$ and $M \supseteq S \supsetneq T$. A set $S \supsetneq T$ can be written in a unique way as $S = S' \cup T$ with $S' \in \mathcal{P}_*^{T^c}$. Hence, for any function f, we have the identity

$$\sum_{S\in\mathcal{P}_*^M} \sum_{T\in\mathcal{P}_{**}^S} f(S,T) = \sum_{T\in\mathcal{P}_*^M} \sum_{S'\in\mathcal{P}_*^{T^c}} f(S'\cup T, T) = \sum_{S\in\mathcal{P}_*^M} \sum_{T\in\mathcal{P}_*^{S^c}} f(T\cup S, S).$$

The last step is changing the names of T and S' to S and T. We will apply this formula to

$$f(S,T) = f_2(S,T) \frac{\partial^{|S^c|} v}{\partial y^{S^c}},$$

and thus we compute

$$f_2(T\cup S, S) = \frac{\partial^{|S|} B_*(u^+)}{\partial y^S} \Phi_\tau \frac{\partial^{|T|} B_*(v)}{\partial y^T} + \frac{1}{2} \Phi_\tau \frac{\partial^{|S|} B_*(v)}{\partial y^S} \frac{\partial^{|T|} B_*(v)}{\partial y^T} + \frac{1}{2} \frac{\partial^{|S|} B_*(u^+)}{\partial y^S} \frac{\partial^{|T|} B_*(u^+)}{\partial y^T} \Phi_\tau, \tag{5.43}$$

$$\frac{\partial^{|(T\cup S)^c|} v}{\partial y^{(T\cup S)^c}} = \frac{\partial^{|S^c|} v}{\partial y^{S^c}}$$

for $T \in \mathcal{P}_*^{S^c}$. (For the second equality, use $(T\cup S)^c = M \setminus (T\cup S) = S^c \setminus T$.) Plugging (5.40) into (5.26) yields

$$\begin{aligned} \mathscr{D}(\Phi_\tau[v]) &= \Phi_\tau[\mathscr{D}v] + \sum_{S\in\mathcal{P}_*^M} \frac{\partial^{|S|}\Phi_\tau}{\partial y^S} \frac{\partial^{|S^c|} v}{\partial y^{S^c}} \\ &= \Phi_\tau[\mathscr{D}v] + \sum_{S\in\mathcal{P}_*^M} \tau f_1(S) \frac{\partial^{|S^c|} v}{\partial y^{S^c}} + \sum_{S\in\mathcal{P}_*^M} \sum_{T\in\mathcal{P}_*^{S^c}} \tau^2 f_2(T\cup S, S) \frac{\partial^{|S^c\setminus T|} v}{\partial y^{S^c\setminus T}} + \mathcal{O}_{\mathbf{0},0}(\tau^3). \end{aligned} \tag{5.44}$$

This expression still contains the numerical flow Φ_τ in the first term and in the definitions of f_1 and f_2. In the next step, we will eliminate it from this expression.

Step 3: Eliminate numerical flows Φ_τ. We use (5.30) to expand the terms containing Φ_τ in (5.44). For $\widetilde{v} \in L^2(\mathbb{T}^N)$, we have

$$\begin{aligned} \Phi_\tau[\widetilde{v}] &= \mathrm{e}^{\tau B_*(u^+)} \mathrm{e}^{\mathrm{i}\tau\Delta} \mathrm{e}^{\tau B_*(v)} \widetilde{v} \\ &= \mathrm{e}^{\mathrm{i}\tau\Delta} \mathrm{e}^{\tau B_*(v)} \widetilde{v} + \tau B_*(u^+) \mathrm{e}^{\mathrm{i}\tau\Delta} \mathrm{e}^{\tau B_*(v)} \widetilde{v} + \frac{\tau^2}{2} B_*^2(u^+) \mathrm{e}^{\mathrm{i}\tau\Delta} \mathrm{e}^{\tau B_*(v)} \widetilde{v} + \mathcal{O}_{\mathbf{0},0}(\tau^3) \\ &= \mathrm{e}^{\mathrm{i}\tau\Delta} \widetilde{v} + \tau \left(\mathrm{e}^{\mathrm{i}\tau\Delta} B_*(v) + B_*(u^+) \mathrm{e}^{\mathrm{i}\tau\Delta} \right) \widetilde{v} \\ &\qquad + \frac{\tau^2}{2} \left(B_*^2(u^+) \mathrm{e}^{\mathrm{i}\tau\Delta} + 2 B_*(u^+) \mathrm{e}^{\mathrm{i}\tau\Delta} B_*(v) + \mathrm{e}^{\mathrm{i}\tau\Delta} B_*^2(v) \right) \widetilde{v} + \mathcal{O}_{\mathbf{0},0}(\tau^3). \end{aligned} \tag{5.45}$$

Using this for the first and second summand in (5.44), we get

$$\begin{aligned}\mathscr{D}(\Phi_\tau[v]) = {}& \mathrm{e}^{\mathrm{i}\tau\Delta}\mathscr{D}v + \tau\left(\mathrm{e}^{\mathrm{i}\tau\Delta}B_*(v) + B_*(u^+)\mathrm{e}^{\mathrm{i}\tau\Delta}\right)\mathscr{D}v \\ &+ \sum_{S\in\mathcal{P}_*^M} \tau\left(\mathrm{e}^{\mathrm{i}\tau\Delta}\frac{\partial^{|S|}B_*(v)}{\partial y^S} + \frac{\partial^{|S|}B_*(u^+)}{\partial y^S}\mathrm{e}^{\mathrm{i}\tau\Delta}\right)\frac{\partial^{|S^c|}v}{\partial y^{S^c}} \\ &+ \sum_{S\in\mathcal{P}_*^M} \tau^2\Big[\mathrm{e}^{\mathrm{i}\tau\Delta}B_*(v)\frac{\partial^{|S|}B_*(v)}{\partial y^S} + B_*(u^+)\mathrm{e}^{\mathrm{i}\tau\Delta}\frac{\partial^{|S|}B_*(v)}{\partial y^S} \\ &\qquad + \frac{\partial^{|S|}B_*(u^+)}{\partial y^S}\mathrm{e}^{\mathrm{i}\tau\Delta}B_*(v) + \frac{\partial^{|S|}B_*(u^+)}{\partial y^S}B_*(u^+)\mathrm{e}^{\mathrm{i}\tau\Delta}\Big]\frac{\partial^{|S^c|}v}{\partial y^{S^c}} \\ &+ \tau^2\left(\tfrac{1}{2}B_*^2(u^+)\mathrm{e}^{\mathrm{i}\tau\Delta} + B_*(u^+)\mathrm{e}^{\mathrm{i}\tau\Delta}B_*(v) + \tfrac{1}{2}\mathrm{e}^{\mathrm{i}\tau\Delta}B_*^2(v)\right)\mathscr{D}v \\ &+ \sum_{S\in\mathcal{P}_*^M}\sum_{T\in\mathcal{P}_*^{S^c}} \tau^2 f_2(T\cup S, S)\frac{\partial^{|S^c\setminus T|}v}{\partial y^{S^c\setminus T}} + \mathcal{O}_{\mathbf{0},0}(\tau^3). \end{aligned} \tag{5.46}$$

Since B commutes with all of its derivatives, we may expand the big [ ]-term to

$$\begin{aligned}&\mathrm{e}^{\mathrm{i}\tau\Delta}B_*(v)\frac{\partial^{|S|}B_*(v)}{\partial y^S} + B_*(u^+)\mathrm{e}^{\mathrm{i}\tau\Delta}\frac{\partial^{|S|}B_*(v)}{\partial y^S} + \frac{\partial^{|S|}B_*(u^+)}{\partial y^S}\mathrm{e}^{\mathrm{i}\tau\Delta}B_*(v) + \frac{\partial^{|S|}B_*(u^+)}{\partial y^S}B_*(u^+)\mathrm{e}^{\mathrm{i}\tau\Delta} \\ &= \tfrac{1}{2}\mathrm{e}^{\mathrm{i}\tau\Delta}B_*(v)\frac{\partial^{|S|}B_*(v)}{\partial y^S} + B_*(u^+)\mathrm{e}^{\mathrm{i}\tau\Delta}\frac{\partial^{|S|}B_*(v)}{\partial y^S} + \tfrac{1}{2}B_*(u^+)\frac{\partial^{|S|}B_*(u^+)}{\partial y^S}\mathrm{e}^{\mathrm{i}\tau\Delta} \\ &\quad + \tfrac{1}{2}\mathrm{e}^{\mathrm{i}\tau\Delta}\frac{\partial^{|S|}B_*(v)}{\partial y^S}B_*(v) + \frac{\partial^{|S|}B_*(u^+)}{\partial y^S}\mathrm{e}^{\mathrm{i}\tau\Delta}B_*(v) + \tfrac{1}{2}\frac{\partial^{|S|}B_*(u^+)}{\partial y^S}B_*(u^+)\mathrm{e}^{\mathrm{i}\tau\Delta} \\ &= f_2(S,\emptyset) + f_2(S,S) + \mathcal{O}_{\mathbf{0},0}(\tau).\end{aligned}$$

In total, we get the rather compact representation

$$\mathscr{D}(\Phi_\tau[v]) = \mathrm{e}^{\mathrm{i}\tau\Delta}\mathscr{D}v + \sum_{S\in\mathcal{P}^M}\tau\widetilde{f}_1(S)\frac{\partial^{|S^c|}v}{\partial y^{S^c}} + \sum_{S\in\mathcal{P}^M}\sum_{T\in\mathcal{P}^{S^c}}\tau^2\widetilde{f}_2(T\cup S,S)\frac{\partial^{|S^c\setminus T|}v}{\partial y^{S^c\setminus T}} + \mathcal{O}_{\mathbf{0},0}(\tau^3), \tag{5.47}$$

where $\widetilde{f}_1$, $\widetilde{f}_2$ are defined as f_1 and f_2 in (5.41) and (5.42), but all numerical flows Φ_τ are replaced by $\mathrm{e}^{\mathrm{i}\tau\Delta}$. The expression (5.47) is now free from numerical flows Φ_τ and can be compared to the expansion of the solution u after some quadrature approximations.

Step 4: Quadrature approximation. Let

$$h(r) \coloneqq \mathrm{e}^{\mathrm{i}(\tau-r)\Delta}\left(\frac{\partial^{|S|}B[u(r)]}{\partial y^S}\mathrm{e}^{r\mathrm{i}\Delta}\frac{\partial^{|S^c|}v}{\partial y^{S^c}}\right).$$

We approximate $I_1(S)$ from (5.37) by

$$I_1(S) = \mathrm{i}\int_0^\tau h(r)\mathrm{d}r \approx \frac{\mathrm{i}\tau}{2}\Big(h(0)+h(\tau)\Big) =: I_1^{\square}(S)$$

and use the second-order Peano form for the quadrature error of the trapezoidal rule, i.e.

$$E = I_1^{\square}(S) - I_1(S) = \frac{\mathrm{i}\tau}{2}\Big(h(0)+h(\tau)\Big) - \mathrm{i}\int_0^\tau h(r)\mathrm{d}r = -\frac{\mathrm{i}\tau^3}{2}\int_0^1\theta(1-\theta)h''(\theta\tau)\mathrm{d}\theta.$$

Again, this error can be estimated in a standard way. The second derivative of h looks rather complicated, but is not hard to calculate. We observe that h'' is bounded in $\|\cdot\|_{\mathbf{0},0}$.

In a similar fashion, we proceed for $I_2(S,T)$ from (5.38). Let

$$g(r,\widetilde{r}) = \mathrm{e}^{\mathrm{i}(\tau-r)\Delta}\frac{\partial^{|S|}B[u(r)]}{\partial y^S}\mathrm{e}^{\mathrm{i}(r-\widetilde{r})\Delta}\frac{\partial^{|T|}B[u(\widetilde{r})]}{\partial y^T}\frac{\partial^{|S^c\setminus T|}v}{\partial y^{S^c\setminus T}}.$$

We approximate

$$I_2(S,T)+\mathcal{O}_{\mathbf{0},0}(\tau^3)=\mathrm{i}^2\int_0^\tau\int_0^r g(r,\widetilde{r})\mathrm{d}\widetilde{r}\mathrm{d}r\approx\frac{1}{2}\left(\frac{\mathrm{i}\tau}{2}\right)^2[g(0,0)+2g(\tau,0)+g(\tau,\tau)]=:I_2^{\square}(S,T)\quad(5.48)$$

via a standard triangle rule. To do that with $\mathcal{O}_{\mathbf{0},0}(\tau^3)$ for the approximation indicated by $\approx$ in (5.48), we need both partial derivatives of g with respect to r and $\widetilde{r}$ to be bounded in $\|\cdot\|_{\mathbf{0},0}$. They are given by

$$\begin{aligned}\partial_r g(r,\widetilde{r})&=\mathrm{e}^{\mathrm{i}(\tau-r)\Delta}\left[-\mathrm{i}\Delta\frac{\partial^{|S|}B[u(r)]}{\partial y^S}+2\chi\frac{\partial^{|S|}\big(\mathrm{Re}(u(r)\overline{\partial_r u(r)})\big)}{\partial y^S}+\frac{\partial^{|S|}B[u(r)]}{\partial y^S}\mathrm{i}\Delta\right]\\&\quad\times\mathrm{e}^{\mathrm{i}(r-\widetilde{r})\Delta}\frac{\partial^{|T|}B[u(\widetilde{r})]}{\partial y^T}\frac{\partial^{|S^c\setminus T|}v}{\partial y^{S^c\setminus T}},\\\partial_{\widetilde{r}} g(r,\widetilde{r})&=\mathrm{e}^{\mathrm{i}(\tau-r)\Delta}\frac{\partial^{|S|}B[u(r)]}{\partial y^S}\mathrm{e}^{\mathrm{i}(r-\widetilde{r})\Delta}\left[-\mathrm{i}\Delta\frac{\partial^{|T|}B[u(\widetilde{r})]}{\partial y^T}\frac{\partial^{|S^c\setminus T|}v}{\partial y^{S^c\setminus T}}\right.\\&\quad\left.+2\chi\frac{\partial^{|T|}\big(\mathrm{Re}(u(\widetilde{r})\overline{\partial_{\widetilde{r}}u(\widetilde{r})}\big)}{\partial y^T}\frac{\partial^{|S^c\setminus T|}v}{\partial y^{S^c\setminus T}}\right],\end{aligned}$$

and thus indeed bounded in $\|\cdot\|_{\mathbf{0},0}$. For (5.36), we get the updated expansion

$$\mathscr{D}u(\tau)=\mathrm{e}^{\tau\mathrm{i}\Delta}\mathscr{D}v+\sum_{S\in\mathcal{P}^M}\left(I_1^{\square}(S)+\sum_{T\in\mathcal{P}^{S^c}}I_2^{\square}(S,T)\right)+\mathcal{O}_{\mathbf{0},0}(\tau^3).\quad(5.49)$$

Step 5: Error expansion. Subtracting (5.47) from (5.49), we arrive at

$$\begin{aligned}\mathscr{D}u(\tau)-\mathscr{D}(\Phi_\tau[v])&=\sum_{S\in\mathcal{P}^M}\left(I_1^{\square}(S)-\tau\widetilde{f}_1(S)\frac{\partial^{|S^c|}v}{\partial y^{S^c}}\right)\\&\quad+\sum_{S\in\mathcal{P}^M}\sum_{T\in\mathcal{P}^{S^c}}\left(I_2^{\square}(S,T)-\tau^2\widetilde{f}_2(T\cup S,S)\frac{\partial^{|S^c\setminus T|}v}{\partial y^{S^c\setminus T}}\right)+\mathcal{O}_{\mathbf{0},0}(\tau^3)\\&=\sum_{S\in\mathcal{P}^M}E_1(S)+\sum_{S\in\mathcal{P}^M}\sum_{T\in\mathcal{P}^{S^c}}E_2(S,T)+\mathcal{O}_{\mathbf{0},0}(\tau^3),\end{aligned}\quad(5.50)$$

where

$$\begin{aligned}E_1(S)&=\tau\left(\mathrm{e}^{\mathrm{i}\tau\Delta}\frac{\partial^{|S|}B_*(v)}{\partial y^S}+\frac{\partial^{|S|}B_*(u(\tau))}{\partial y^S}\mathrm{e}^{\mathrm{i}\tau\Delta}-\mathrm{e}^{\mathrm{i}\tau\Delta}\frac{\partial^{|S|}B_*(v)}{\partial y^S}-\frac{\partial^{|S|}B_*(u^+)}{\partial y^S}\mathrm{e}^{\mathrm{i}\tau\Delta}\right)\frac{\partial^{|S^c|}v}{\partial y^{S^c}}\\&=\tau\left(\frac{\partial^{|S|}B_*(u(\tau))}{\partial y^S}-\frac{\partial^{|S|}B_*(u^+)}{\partial y^S}\right)\mathrm{e}^{\mathrm{i}\tau\Delta}\frac{\partial^{|S^c|}v}{\partial y^{S^c}}\end{aligned}$$

and

$$\begin{aligned}E_2(S,T)&=\tau^2\left[\tfrac{1}{2}\mathrm{e}^{\mathrm{i}\tau\Delta}\tfrac{\partial^{|S|}B_*(v)}{\partial y^S}\tfrac{\partial^{|T|}B_*(v)}{\partial y^T}+\tfrac{\partial^{|S|}B_*(u(\tau))}{\partial y^S}\mathrm{e}^{\mathrm{i}\tau\Delta}\tfrac{\partial^{|T|}B_*(v)}{\partial y^T}+\tfrac{1}{2}\tfrac{\partial^{|S|}B_*(u(\tau))}{\partial y^S}\tfrac{\partial^{|T|}B_*(u(\tau))}{\partial y^T}\mathrm{e}^{\mathrm{i}\tau\Delta}\right.\\&\quad-\tfrac{\partial^{|S|}B_*(u^+)}{\partial y^S}\mathrm{e}^{\mathrm{i}\tau\Delta}\tfrac{\partial^{|T|}B_*(v)}{\partial y^T}-\tfrac{1}{2}\mathrm{e}^{\mathrm{i}\tau\Delta}\tfrac{\partial^{|S|}B_*(v)}{\partial y^S}\tfrac{\partial^{|T|}B_*(v)}{\partial y^T}-\tfrac{1}{2}\tfrac{\partial^{|S|}B_*(u^+)}{\partial y^S}\tfrac{\partial^{|T|}B_*(u^+)}{\partial y^T}\mathrm{e}^{\mathrm{i}\tau\Delta}\Big]\tfrac{\partial^{|S^c\setminus T|}v}{\partial y^{S^c\setminus T}}\\&=\tau^2\left[\tfrac{\partial^{|S|}B_*(u(\tau))}{\partial y^S}\mathrm{e}^{\mathrm{i}\tau\Delta}\tfrac{\partial^{|T|}B_*(v)}{\partial y^T}+\tfrac{1}{2}\tfrac{\partial^{|S|}B_*(u(\tau))}{\partial y^S}\tfrac{\partial^{|T|}B_*(u(\tau))}{\partial y^T}\mathrm{e}^{\mathrm{i}\tau\Delta}\right.\\&\quad-\tfrac{\partial^{|S|}B_*(u^+)}{\partial y^S}\mathrm{e}^{\mathrm{i}\tau\Delta}\tfrac{\partial^{|T|}B_*(v)}{\partial y^T}-\tfrac{1}{2}\tfrac{\partial^{|S|}B_*(u^+)}{\partial y^S}\tfrac{\partial^{|T|}B_*(u^+)}{\partial y^T}\mathrm{e}^{\mathrm{i}\tau\Delta}\Big]\tfrac{\partial^{|S^c\setminus T|}v}{\partial y^{S^c\setminus T}}\\&=\tau^2\left[\tfrac{\partial^{|S|}B_*(u(\tau))}{\partial y^S}-\tfrac{\partial^{|S|}B_*(u^+)}{\partial y^S}\right]\mathrm{e}^{\mathrm{i}\tau\Delta}\tfrac{\partial^{|T|}B_*(v)}{\partial y^T}\\&\quad+\tfrac{\tau^2}{2}\left[\tfrac{\partial^{|S|}B_*(u(\tau))}{\partial y^S}-\tfrac{\partial^{|S|}B_*(u^+)}{\partial y^S}\right]\tfrac{\partial^{|T|}B_*(u^+)}{\partial y^T}\mathrm{e}^{\mathrm{i}\tau\Delta}+\mathcal{O}_{\mathbf{0},0}(\tau^3).\end{aligned}$$

By Lemma 5.3.7, we obtain that

$$\frac{\partial^{|S|}B_*(u(\tau))}{\partial y^S} - \frac{\partial^{|S|}B_*(u^+)}{\partial y^S} = \mathcal{O}_{\mathbf{0},0}(\tau^2)$$

and thus both $E_1(S)$ and $E_2(S,T)$ are $\mathcal{O}_{\mathbf{0},0}(\tau^3)$. Plugging this into (5.50), we arrive at

$$\mathscr{D}u(\tau) - \mathscr{D}(\Phi_\tau(v)) = \mathcal{O}_{\mathbf{0},0}(\tau^3),$$

which finishes the proof. □

Now that the local error bounds are established, we may turn to the global error bounds. As usual, the transition from local to global error bounds requires some stability results. In our case, we have the following two lemmas. The first one deals with the flow of the non-linear subproblem.

Lemma 5.3.8. *Let $\boldsymbol{k} \in \mathbb{N}_0^d$, $r \in \mathbb{N}_0$ and $r^* = \max\{r, \lfloor \frac{N}{2} \rfloor + 1\}$. Moreover, assume that $V \in C^{\boldsymbol{k}}(\Gamma, H^{r^*}(\mathbb{T}^N))$ and $v_0, w_0 \in C^{\boldsymbol{k}}(\Gamma, H^{r^*}(\mathbb{T}^N))$ with*

$$\|v_0\|_{\boldsymbol{k},r^*} \leq R \qquad \text{and} \qquad \|w_0\|_{\boldsymbol{k},r^*} \leq R.$$

Then we have

$$\|\mathrm{e}^{\mathrm{i}tB[v_0]}v_0 - \mathrm{e}^{\mathrm{i}tB[w_0]}w_0\|_{\boldsymbol{k},r} \leq \mathrm{e}^{C(\|V\|_{\boldsymbol{k},r^*}+|\chi|R^2)t}\|v_0 - w_0\|_{\boldsymbol{k},r},$$
$$\|\mathrm{e}^{\mathrm{i}tB[v_0]}v_0\|_{\boldsymbol{k},r} \leq R\mathrm{e}^{C(\|V\|_{\boldsymbol{k},r^*}+|\chi|R^2)t},$$

both for $t \geq 0$ and the constant $C > 0$ does not depend on V, v_0 and w_0.

Proof. First observe that the functions $v(t) = \mathrm{e}^{\mathrm{i}tB[v_0]}v_0$ and $w(t) = \mathrm{e}^{\mathrm{i}tB[w_0]}w_0$ solve the initial value problems

$$v'(t) = \mathrm{i}B[v_0]v(t), \qquad v(0) = v_0,$$
$$w'(t) = \mathrm{i}B[w_0]w(t), \qquad w(0) = w_0.$$

We start with the proof of the second inequality and then use it to show the first one.

Second inequality. Starting with

$$v(t) = v_0 + \int_0^t \mathrm{i}B[v_0]v(s)\mathrm{d}s, \tag{5.51}$$

Lemma 5.3.5 and Corollary 5.3.6 imply

$$\|v(t)\|_{\boldsymbol{k},r} \leq \|v_0\|_{\boldsymbol{k},r} + \int_0^t \|B[v_0]v(s)\|_{\boldsymbol{k},r}\mathrm{d}s \leq \|v_0\|_{\boldsymbol{k},r} + C\int_0^t \left(\|V\|_{\boldsymbol{k},r^*} + |\chi|\|v_0\|_{\boldsymbol{k},r^*}^2\right)\|v(s)\|_{\boldsymbol{k},r}\mathrm{d}s,$$

and thus, by Gronwall's lemma,

$$\|v(t)\|_{\boldsymbol{k},r} \leq \mathrm{e}^{C(\|V\|_{\boldsymbol{k},r^*}+|\chi|R^2)t}\|v_0\|_{\boldsymbol{k},r}. \tag{5.52}$$

This shows the second inequality. Note that the same inequality also holds with r replaced by r^*.

First inequality. We have

$$\begin{aligned} B[v_0]v(t) - B[w_0]w(t) &= (V + \chi|v_0|^2)v(t) - (V + \chi|w_0|^2)w(t) \\ &= V(v(t) - w(t)) + \chi(|v_0|^2v(t) - |w_0|^2w(t)) \\ &= V(v(t) - w(t)) + \chi\big[(v_0 - w_0)\overline{v_0}v(t) + w_0(\overline{v_0} - \overline{w_0})v(t) + |w_0|^2(v(t) - w(t))\big] \end{aligned}$$

and hence, using (5.52) with r^* instead of r,

$$\begin{aligned}\|B[v_0]v(t) - B[w_0]w(t)\|_{\boldsymbol{k},r} \leq C\Big[&\|V\|_{\boldsymbol{k},r^*}\|v(t)-w(t)\|_{\boldsymbol{k},r} + |\chi|\|v_0\|_{\boldsymbol{k},r^*}\|v(t)\|_{\boldsymbol{k},r^*}\|v_0-w_0\|_{\boldsymbol{k},r}\\ &+ |\chi|\|w_0\|_{\boldsymbol{k},r^*}\|v(t)\|_{\boldsymbol{k},r^*}\|v_0-w_0\|_{\boldsymbol{k},r} + |\chi|\|w_0\|^2_{\boldsymbol{k},r^*}\|v(t)-w(t)\|_{\boldsymbol{k},r}\Big]\\ \leq C\Big(&[\|V\|_{\boldsymbol{k},r^*} + |\chi|R^2]\|v(t)-w(t)\|_{\boldsymbol{k},r}\\ &+ 2|\chi|R^2 \mathrm{e}^{C(\|V\|_{\boldsymbol{k},r^*}+|\chi|R^2)t}\|v_0-w_0\|_{\boldsymbol{k},r}\Big).\end{aligned}$$

Using (5.51) (and its analogue for $w(t)$), we obtain

$$\begin{aligned}\|v(t)-w(t)\|_{\boldsymbol{k},r} &\leq \|v_0-w_0\|_{\boldsymbol{k},r} + \int_0^t \|B[v_0]v(s) - B[w_0]w(s)\|_{\boldsymbol{k},r}\mathrm{d}s\\ &\leq \left(1 + 2C|\chi|R^2\int_0^t \mathrm{e}^{C(\|V\|_{\boldsymbol{k},r^*}+|\chi|R^2)s}\mathrm{d}s\right)\|v_0-w_0\|_{\boldsymbol{k},r}\\ &\quad + C(\|V\|_{\boldsymbol{k},r^*}+|\chi|R^2)\int_0^t \|v(s)-w(s)\|_{\boldsymbol{k},r}\mathrm{d}s.\end{aligned}$$

Using that the integral is monotonic, we estimate

$$1 + 2C|\chi|R^2\int_0^t \mathrm{e}^{C(\|V\|_{\boldsymbol{k},r^*}+|\chi|R^2)s}\mathrm{d}s \leq \mathrm{e}^{2C(\|V\|_{\boldsymbol{k},r^*}+|\chi|R^2)t}.$$

By Gronwall's lemma, we arrive at

$$\begin{aligned}\|v(t)-w(t)\|_{\boldsymbol{k},r} &\leq \mathrm{e}^{2C(\|V\|_{\boldsymbol{k},r^*}+|\chi|R^2)t}\|v_0-w_0\|_{\boldsymbol{k},r}\mathrm{e}^{tC(\|V\|_{\boldsymbol{k},r^*}+|\chi|R^2)}\\ &= \mathrm{e}^{3C(\|V\|_{\boldsymbol{k},r^*}+|\chi|R^2)t}\|v_0-w_0\|_{\boldsymbol{k},r}.\end{aligned}$$

□

The previous result only deals with the non-linear part of the problem. Now we extend the result to treat the "full" numerical flow.

Lemma 5.3.9. *If $V \in C^{\boldsymbol{k}}(\Gamma, H^{r^*}(\mathbb{T}^N))$, $v_0, w_0 \in C^{\boldsymbol{k}}(\Gamma, H^{r^*}(\mathbb{T}^N))$ with $\|v_0\|_{\boldsymbol{k},r^*} \leq R$ and $\|w_0\|_{\boldsymbol{k},r^*} \leq R$ for some $r \in \mathbb{N}_0$ and $\boldsymbol{k} \in \mathbb{N}_0^d$, then there exists a constant $C > 0$ such that*

$$\|\Phi_\tau(v_0) - \Phi_\tau(w_0)\|_{\boldsymbol{k},r} \leq \mathrm{e}^{C(\|V\|_{\boldsymbol{k},r^*}+|\chi|\widetilde{R}^2)\tau}\|v_0-w_0\|_{\boldsymbol{k},r}$$

for all $0 < \tau \leq 2$, where $\widetilde{R} = \mathrm{e}^{C(\|V\|_{\boldsymbol{k},r^*}+|\chi|R^2)}R$. The constant $C > 0$ does not depend on V, v_0 and w_0.

Proof. We set $v^+ = \mathrm{e}^{\mathrm{i}\Delta\tau}\mathrm{e}^{\tau B_*(v_0)}v_0$ and $w^+ = \mathrm{e}^{\mathrm{i}\Delta\tau}\mathrm{e}^{\tau B_*(w_0)}w_0$. Applying Lemma 5.3.8 twice, we obtain

$$\begin{aligned}\|\Phi_\tau(v_0) - \Phi_\tau(w_0)\|_{\boldsymbol{k},r} &= \|\mathrm{e}^{\tau B_*(v^+)}v^+ - \mathrm{e}^{\tau B_*(w^+)}w^+\|_{\boldsymbol{k},r}\\ &\leq \mathrm{e}^{C(\|V\|_{\boldsymbol{k},r^*}+|\chi|\widetilde{R}^2)\tau/2}\|v^+-w^+\|_{\boldsymbol{k},r}\\ &= \mathrm{e}^{C(\|V\|_{\boldsymbol{k},r^*}+|\chi|\widetilde{R}^2)\tau/2}\|\mathrm{e}^{\tau B_*(v_0)}v_0 - \mathrm{e}^{\tau B_*(w_0)}w_0\|_{\boldsymbol{k},r}\\ &\leq \mathrm{e}^{C(\|V\|_{\boldsymbol{k},r^*}+|\chi|\widetilde{R}^2)\tau}\|v_0-w_0\|_{\boldsymbol{k},r}.\end{aligned}$$

□

We are now in the position to prove (5.15), the global error bound in $L^2(\mathbb{T}^N)$.

Proof of Theorem 5.3.2 / (5.15)

In this proof, we combine the local error bounds (5.12) and (5.13) with the $H^s(\mathbb{T}^N)$-conditional stability estimate from Lemma 5.3.9 (with $r = 0$) to derive the global error bound. This is a typical "Lady Windermere's fan" argument.

As a preparation, we prove by induction on n that there exists $\tau_0 > 0$ such that for all $n \in \mathbb{N}_0$ and $\ell \in \mathbb{N}_0$ with $t_{\ell+n} = (\ell+n)\tau \le T$, it holds that

$$\|\Phi_\tau^n(u(t_\ell))\|_{\boldsymbol{k},s} \le 2M_{\boldsymbol{k}}^{(s)} \tag{5.53}$$

for all $0 < \tau \le \tau_0$, where

$$M_{\boldsymbol{k}}^{(s)} = \max_{t\in[0,T]} \|u(t)\|_{\boldsymbol{k},s}.$$

This is clear for $n = 0$. Now assume that

$$\|\Phi_\tau^k(u(t_\ell))\|_{\boldsymbol{k},s} \le 2M_{\boldsymbol{k}}^{(s)} \quad \text{for} \quad k = 0,\dots,n-1,\ \ell \in \mathbb{N}_0 \quad \text{with} \quad t_{\ell+k} \le T.$$

We estimate the norm of $\Phi_\tau^n(u(t_\ell))$ using the telescoping sum

$$\Phi_\tau^n(u(t_\ell)) = u(t_{\ell+n}) + \sum_{j=0}^{n-1} \left(\Phi_\tau^{n-j}(u(t_{\ell+j})) - \Phi_\tau^{n-j-1}(u(t_{\ell+j+1}))\right) \tag{5.54}$$

and

$$\Phi_\tau^{n-j}(u(t_{\ell+j})) - \Phi_\tau^{n-j-1}(u(t_{\ell+j+1})) = \Phi_\tau(\Phi_\tau^{n-j-1}(u(t_{\ell+j}))) - \Phi_\tau(\Phi_\tau^{n-j-2}(u(t_{\ell+j+1})))$$

whenever $0 \le j \le n-2$. Combining the stability result from Lemma 5.3.9 with the induction hypothesis, we get

$$\begin{aligned}
&\|\Phi_\tau^{n-j}(u(t_{\ell+j})) - \Phi_\tau^{n-j-1}(u(t_{\ell+j+1}))\|_{\boldsymbol{k},s} \\
&\qquad \le \mathrm{e}^{C(\|V\|_{\boldsymbol{k},s}+|\chi|\widetilde{R}^2)\tau}\|\Phi_\tau^{n-j-1}(u(t_{\ell+j})) - \Phi_\tau^{n-j-2}(u(t_{\ell+j+1}))\|_{\boldsymbol{k},s} \\
&\qquad \le \mathrm{e}^{C(\|V\|_{\boldsymbol{k},s}+|\chi|\widetilde{R}^2)\tau(n-j-1)}\|\Phi_\tau(u(t_{\ell+j})) - u(t_{\ell+j+1})\|_{\boldsymbol{k},s} \\
&\qquad \le \mathrm{e}^{C(\|V\|_{\boldsymbol{k},s}+|\chi|\widetilde{R}^2)\tau(n-j-1)}C_{\mathrm{loc},s}\tau^2
\end{aligned} \tag{5.55}$$

for all $0 < \tau \le 2$, where $\widetilde{R} = \mathrm{e}^{C(\|V\|_{\boldsymbol{k},s}+|\chi|R^2)}R$ and $R = 2M_{\boldsymbol{k}}^{(s)}$. The constant $C_{\mathrm{loc},s}$ comes from (5.12), the local error bound in $H^s(\mathbb{T}^N)$. Note that (5.55) also holds for $j = n-1$ (no stability result is needed then, just the local error). Using (5.55) for (5.54) yields

$$\begin{aligned}
\|\Phi_\tau^n(u(t_\ell))\|_{\boldsymbol{k},s} &\le \|u(t_{\ell+n})\|_{\boldsymbol{k},s} + \sum_{j=0}^{n-1} \mathrm{e}^{C(\|V\|_{\boldsymbol{k},s}+|\chi|\widetilde{R}^2)\tau j}C_{\mathrm{loc},s}\tau^2 \\
&\le M_{\boldsymbol{k}}^{(s)} + T\mathrm{e}^{C(\|V\|_{\boldsymbol{k},s}+|\chi|\widetilde{R}^2)T}C_{\mathrm{loc},s}\tau
\end{aligned}$$

for $(\ell+n)\tau \le T$, which can be bounded by $2M_{\boldsymbol{k}}^{(s)}$ for any

$$\tau \le \tau_0 := \min\left\{\frac{M_{\boldsymbol{k}}^{(s)}}{T\mathrm{e}^{C(\|V\|_{\boldsymbol{k},s}+|\chi|\widetilde{R}^2)T}C_{\mathrm{loc},s}}, 2\right\}. \tag{5.56}$$

This finishes the induction step and establishes (5.53).

Now we can tackle the proof of the theorem itself. By Lemma 5.3.9 for $r = 0$ and the local error bound (5.13), we have

$$\begin{aligned}
\|\Phi_\tau^n(u_0) - u(t_n)\|_{\boldsymbol{k},0} &\leq \sum_{j=0}^{n-1} \|\Phi_\tau^j(\Phi_\tau(u(t_{n-j-1})) - \Phi_\tau^j(u(t_{n-j}))\|_{\boldsymbol{k},0} \\
&\leq \sum_{j=0}^{n-1} \mathrm{e}^{C(\|V\|_{\boldsymbol{k},s}+|\chi|\widetilde{R}^2)\tau j} \|\Phi_\tau(u(t_{n-j-1})) - u(t_{n-j})\|_{\boldsymbol{k},0} \\
&\leq \sum_{j=0}^{n-1} \mathrm{e}^{C(\|V\|_{\boldsymbol{k},s}+|\chi|\widetilde{R}^2)\tau j} C_{\mathrm{loc},0}\tau^3 \\
&\leq \frac{\mathrm{e}^{C(\|V\|_{\boldsymbol{k},s}+|\chi|\widetilde{R}^2)\tau n} - 1}{\mathrm{e}^{C(|\chi|\widetilde{R}^2+\|V\|_{\boldsymbol{k},s})\tau} - 1} C_{\mathrm{loc},0}\tau^3 \\
&\leq \frac{\mathrm{e}^{C(\|V\|_{\boldsymbol{k},s}+|\chi|\widetilde{R}^2)t_n} - 1}{C(|\chi|\widetilde{R}^2 + \|V\|_{\boldsymbol{k},s})} C_{\mathrm{loc},0}\tau^2.
\end{aligned}$$

In the last step, we used $1 + x \leq \mathrm{e}^x$ for $x \geq 0$. Note that $C_{\mathrm{loc},0}$ is exactly the constant from the local error bound (5.13) and thus depends on $M_{\boldsymbol{k}}^{(s+2)}$. □

Proof of Theorem 5.3.2 / (5.14)

Here we have to combine the local error bound (5.12) with the $H^s(\mathbb{T}^N)$-conditional stability estimate from Lemma 5.3.9 for $r = s = r^*$. The procedure is completely analogous to the previous proof and thus omitted here. □

Now that we have shown both error bounds which were stated in Theorem 5.3.2, the proof is completed.

We do not verify Theorem 5.3.2 numerically, since it has to be seen merely as a tool for the analysis of the multi-level method. The error in the norms $\|\cdot\|_{\boldsymbol{k},0}$ for $\boldsymbol{k} \neq \boldsymbol{0}$ is usually not a quantity one is interested in. Another reason for omitting a numerical verification of the theorem is the difficulty of computing the norms $\|\cdot\|_{\boldsymbol{k},0}$ for $\boldsymbol{k} \neq \boldsymbol{0}$ in practice – the main reason for that is the presence of the (in practice unavailable) y-derivatives of the solution. That being said, the case $\boldsymbol{k} = \boldsymbol{0}$ is an exception – but in this case the statement of the theorem is not new.

5.3.6 Error analysis for the linear Schrödinger equation

All results from this chapter so far hold for the special case $\chi = 0$, too. However, some of the results can be simplified. We present these simplifications now. The content presented here will appear in a more detailed version in [61].

Throughout this section we consider the linear Schrödinger equation (LSE) for $u\colon \mathbb{R}_+ \times \mathbb{T}^N \times \Gamma \to \mathbb{C}$ given by

$$\partial_t u(t,x,y) = \mathrm{i}\Delta u(t,x,y) + \mathrm{i}V(x,y)u(t,x,y), \qquad t \geq 0,\ x \in \mathbb{T}^N,\ y \in \Gamma, \tag{5.57a}$$

$$u(0,x,y) = u_0(x,y), \qquad x \in \mathbb{T}^N,\ y \in \Gamma, \tag{5.57b}$$

where $N \in \mathbb{N}$, $\mathbb{T} = \mathbb{R}/(2\pi\mathbb{Z})$, $\Gamma = [-1,1]^d$, $V\colon \mathbb{T}^N \times \Gamma \to \mathbb{R}$ and $u_0\colon \mathbb{T}^N \times \Gamma \to \mathbb{C}$. This equation is equal to (5.2) for $\chi = 0$.

Since this PDE is linear in u, it is not required anymore to use the algebra structure of $H^r(\mathbb{T}^N)$ for $r > \frac{N}{2}$ in the analysis and we may easily generalise our results from the previous sections for Sobolev spaces of lower order. The precise assumptions and results are stated below.

The first assumption deals with the initial value.

Assumption E3. Let $u_0 \in C^{\boldsymbol{k}}(\Gamma, H^2(\mathbb{T}^N))$ for a multi-index $\boldsymbol{k} = (k_1, \dots, k_d) \in \mathbb{N}_0^d$.

For the potential, we assume the following.

Assumption E4a. Let $V \in C^{\boldsymbol{k}}(\Gamma, W^{2,\infty}(\mathbb{T}^N))$ for a multi-index $\boldsymbol{k} = (k_1, \dots, k_d) \in \mathbb{N}_0^d$.

We also consider the following stronger version of E4a.

Assumption E4b. Let $V \in C^{\boldsymbol{k}}(\Gamma, W^{4,\infty}(\mathbb{T}^N))$ for a multi-index $\boldsymbol{k} = (k_1, \dots, k_d) \in \mathbb{N}_0^d$.

Assumption E3, E4a, and E4b can be checked rather easily in practice since they are all requirements on the given data of the problem, and not on the solution. This is different from the non-linear case before, as Assumption E1 was an assumption on the (unknown) solution.

For the linear equation considered here, the above assumptions can be used to show the existence of a classical solution, as stated in the next theorem.

Theorem 5.3.10. *Suppose that E3 and E4a hold for the same multi-index $\boldsymbol{k} \in \mathbb{N}_0^d$. Then, a classical solution of the initial value problem (5.57) exists and has the regularity*

$$u \in C^1\big([0,T], C^{\boldsymbol{k}}(\Gamma, L^2(\mathbb{T}^N))\big) \cap C\big([0,T], C^{\boldsymbol{k}}(\Gamma, H^2(\mathbb{T}^N))\big).$$

A proof of Theorem 5.3.10 will appear in the supplementary material of [61]. Theorem 5.3.10 also implies that the quantity

$$M_{\boldsymbol{k}}^{(s)} = \max_{t\in[0,T]} \|u(t)\|_{C^{\boldsymbol{k}}(\Gamma, H^s(\mathbb{T}^N))} \tag{5.58}$$

is finite for $s \in \{1, 2\}$.

The following theorem is the main convergence result for the linear Schrödinger equation.

Theorem 5.3.11. *Let $0 < \tau \le 1$ and set $t_n = n\tau$ for $n \in \mathbb{N}_0$. Let $H(y) = \Delta + V(y)$ such that the solution of (5.57) is $u(t, x, y) = \mathrm{e}^{itH(y)} u_0(x, y)$.*

(a) If E3 and E4a hold for the same $\boldsymbol{k} \in \mathbb{N}_0^d$, then $\Phi_\tau^n u_0 \in C^{\boldsymbol{k}}(\Gamma, H^2(\mathbb{T}^N))$ for all $n \in \mathbb{N}_0$. Moreover, there is a constant C such that

$$\|u(t_n) - \Phi_\tau^n u_0\|_{C^{\boldsymbol{k}}(\Gamma, L^2(\mathbb{T}^N))} \le C M_{\boldsymbol{k}}^{(1)} \tau$$

as long as $0 \le t_n \le T$, with $M_{\boldsymbol{k}}^{(1)}$ from (5.58). The constant C depends on T and on the norm $\|V\|_{C^{\boldsymbol{k}}(\Gamma, W^{2,\infty}(\mathbb{T}^N))}$.

(b) If, in addition, E4b holds (again with the same $\boldsymbol{k} \in \mathbb{N}_0^d$), then there is a constant C such that

$$\|u(t_n, \cdot) - \Phi_\tau^n u_0\|_{C^{\boldsymbol{k}}(\Gamma, L^2(\mathbb{T}^N))} \le C M_{\boldsymbol{k}}^{(2)} \tau^2$$

as long as $0 \le t_n \le T$, with $M_{\boldsymbol{k}}^{(2)}$ from (5.58). The constant C depends on T and on the norm $\|V\|_{C^{\boldsymbol{k}}(\Gamma, W^{4,\infty}(\mathbb{T}^N))}$

The proof of Theorem 5.3.11 is not given here, but is similar to the proof of Theorem 5.3.2 and uses the same techniques. It will appear in [61].

Let us now return to the NLS with uncertain parameters and discuss the application of single-level collocation methods.

5.4 Single-level stochastic collocation

The single-level approximation of the solution u of the NLS (5.2) at time $t_n = n\tau$ is given by

$$u_{L,n} = \mathcal{I}_L^{p,g} u_n, \tag{5.59}$$

where $\mathcal{I}_L^{p,g}$ is the sparse grid interpolant from (2.15) and u_n is the Strang splitting approximation after n steps with step-size $\tau > 0$, as defined in (5.8). The sparse grid interpolant $\mathcal{I}_L^{p,g}$ is based on Clenshaw-Curtis abscissas and p and g are given by the "Smolyak" case from Table 2.1.

Unfortunately, the error analysis for this method cannot be performed with the same tools as in the parabolic case in Section 4.6. The deduction of an error bound has to be adapted to the fact that the solution of the NLS does usually not have an analytic extension to a complex polyellipse in the parameter space, but instead belongs to $C^{\boldsymbol{k}}(\Gamma, X)$ with $X = L^2(\mathbb{T}^N)$. Let us explain this briefly, since it is the main reason why the approach from the parabolic case is not suitable here.

Recall the fact from complex analysis that the map $\mathbb{C} \to \mathbb{C},\ z \mapsto |z|^2$ is not holomorphic (= complex analytic). We observe that the right-hand side in the NLS (5.2a) contains the term $|u|^2 u$ and thus a classical solution u of (5.2) cannot have an analytic derivative $\partial_t u$ in a region in $\mathbb{C}^d$. So it is very unlikely that u itself is analytic.

Remark 5.4.1. Solutions of wave equations are typically not analytic with respect to their parameters either. This was examined in detail in [80] and [81] for linear wave equations. It should be noted, however, that a noticeable exception was also given there: If very smooth quantities of interest of the solution are considered, then analyticity arguments can be used again. ◇

Now we explain how the error of the stochastic collocation method for the NLS can be analysed instead. Suppose that the Assumptions E1 and E2 hold for $\boldsymbol{k} = (k, \dots, k) \in \mathbb{N}_0^d$. Then Theorem 2.6.6 implies that the sparse grid interpolation error at time $t \in [0, T]$ is bounded by

$$\|u(t) - \mathcal{I}_L^{p,g} u(t)\|_{L_\varrho^2(\Gamma,X)} \leq C\mathcal{R}(\eta_L, k, d)\|u(t)\|_{C^{\boldsymbol{k}}(\Gamma,X)}$$

with $\mathcal{R}(\eta_L, k, d)$ from (2.23) and $X = L^2(\mathbb{T}^N)$.

Now consider the splitting approximation u_n for $n \in \mathbb{N}_0$. By Theorem 5.3.2, $u_n \in C^{\boldsymbol{k}}(\Gamma, X)$ for every $n \in \mathbb{N}_0$ such that $0 \leq n\tau \leq T$. Moreover, the triangle inequality and (5.15) imply that its norm is bounded by

$$\|u_n\|_{C^{\boldsymbol{k}}(\Gamma,X)} \leq \|u_n - u(t_n)\|_{C^{\boldsymbol{k}}(\Gamma,X)} + \|u(t_n)\|_{C^{\boldsymbol{k}}(\Gamma,X)} \leq \widetilde{C}\tau_0^2 + \max_{t\in[0,T]} \|u(t)\|_{C^{\boldsymbol{k}}(\Gamma,X)},$$

in particular independently of n and τ. The quantities $\widetilde{C}$ and τ_0 are the ones from Theorem 5.3.2. Again, Theorem 2.6.6 implies

$$\|u_n - \mathcal{I}_L^{p,g} u_n\|_{L_\varrho^2(\Gamma,X)} \leq C\mathcal{R}(\eta_L, k, d)\|u_n\|_{C^{\boldsymbol{k}}(\Gamma,X)}.$$

The total error of the single-level stochastic collocation method after n steps with step-size $\tau > 0$ can be split into

$$\|u(t_n) - u_{L,n}\|_{L^2_\varrho(\Gamma,X)} \le \|u(t_n) - u_n\|_{L^2_\varrho(\Gamma,X)} + \|u_n - \mathcal{I}_L^{p,g} u_n\|_{L^2_\varrho(\Gamma,X)}.$$

The first term can be treated by Theorem 5.3.2 for $\boldsymbol{k} = \mathbf{0}$, and the second term was estimated immediately before. Thus we have shown the following theorem.

Theorem 5.4.2 (Error of single-level collocation)**.** *Suppose that the Assumptions E1 and E2 hold for* $s > \max\{\frac{N}{2}, 1\}$ *and* $\boldsymbol{k} = (k, \dots, k) \in \mathbb{N}_0^d$*. Then there exists* $\tau_0 > 0$ *such that for all step-sizes* $\tau \in (0, \tau_0]$*, the stochastic collocation error is bounded by*

$$\|u(t_n) - u_{L,n}\|_{L^2_\varrho(\Gamma,L^2(\mathbb{T}^N))} \le C(\tau^2 + \mathcal{R}(\eta_L, k, d))$$

as long as $0 \le n\tau \le T$*, where* $\mathcal{R}(\eta_L, k, d)$ *is given by* (2.23)*. The constant* C *depends on*

$$\max_{t\in[0,T]} \|u(t)\|_{C^{\boldsymbol{k}}(\Gamma,H^{s+2}(\mathbb{T}^N))},$$

k *and* d*, but is independent of* η_L*,* τ *and* n*.*

We proceed with the multi-level method.

5.5 Multi-level stochastic collocation

Here we verify the general Assumptions B1 and B2 for the multi-level method from Section 3.4. In particular Assumption B2 cannot be verified as in the parabolic case due to the fact that an analytic extension of the solution u to a complex polyellipse is not available here, as discussed in the previous section.

We suppose as in the previous section that the Assumptions E1 and E2 hold for $\boldsymbol{k} = (k, \dots, k) \in \mathbb{N}_0^d$. Let $T > 0$. Recall the notation introduced in Section 3.4,

$$u_{\tau_j} = \Phi_{\tau_j}^{N_j}(u_0)$$

with the numerical flow of the Strang splitting scheme Φ_τ from (5.9). Let $\tau_j = 2^{-j}\tau_0$ and $T = \tau_j N_j$ for $j \in \mathbb{N}_0$, where the largest step-size τ_0 with $N_0 = T/\tau_0 \in \mathbb{N}$ is chosen either as the (in practice unknown) value τ_0 from Theorem 5.3.2 or smaller, but not larger.

To show Assumption B1, we apply Theorem 5.3.2 for $\boldsymbol{k} = \mathbf{0}$ and obtain

$$\|u(T) - u_{\tau_j}\|_{L^2_\varrho(\Gamma,X)} \le \|u(T) - u_{\tau_j}\|_{C^{\mathbf{0}}(\Gamma,X)} \le C\tau_j^\alpha$$

for $j \in \mathbb{N}_0$ with $\alpha = 1$ for $X = H^s(\mathbb{T}^N)$ or $\alpha = 2$ for $X = L^2(\mathbb{T}^N)$. This verifies Assumption B1.

The main challenge is to prove that Assumption B2 is true. We choose $\zeta\colon \Lambda(\Gamma, \mathcal{X}) \to \mathbb{R}$ with

$$\Lambda(\Gamma, \mathcal{X}) = C^{\boldsymbol{k}}(\Gamma, X) \qquad \text{and} \qquad \zeta(v) = \|v\|_{C^{\boldsymbol{k}}(\Gamma,X)}.$$

The interpolation error estimate in Assumption B2, (3.10), follows directly from Theorem 2.6.6 with

$$\kappa_\ell = \mathcal{R}(\eta_\ell, k, d) \tag{5.60}$$

from (2.23). Alternatively, it holds with

$$\kappa_\ell = \eta_\ell^{-k}(\log(\eta_\ell))^{(k+2)(d-1)+1}$$

by (2.24). To show the remaining estimates in (3.11), we apply the triangle inequality and Theorem 5.3.2 to arrive at

$$\begin{aligned}\zeta(u_{\tau_{j+1}} - u_{\tau_j}) &\leq \|u_{\tau_{j+1}} - u(T)\|_{C^k(\Gamma,X)} + \|u(T) - u_{\tau_j}\|_{C^k(\Gamma,X)} \leq C(1+2^\beta)\tau_{j+1}^\beta,\\ \zeta(u_{\tau_j}) &\leq \|u_{\tau_j} - u(T)\|_{C^k(\Gamma,X)} + \|u(T)\|_{C^k(\Gamma,X)} \leq \big(C + \tau_0^{-\beta}\|u(T)\|_{C^k(\Gamma,X)}\big)\tau_0^\beta,\end{aligned}$$

where $\beta = 1$ for $X = H^s(\mathbb{T}^N)$ and $\beta = 2$ for $X = L^2(\mathbb{T}^N)$. Thus, Assumption B2 is satisfied.

So far, the sequence $(\kappa_\ell)_{\ell\in\mathbb{N}_0}$ from (5.60) was different from $\kappa_\ell = \eta_\ell^{-\mu}$ (as required by the ε-cost theorem). By Remark 3.5.1, however, we may indeed choose $\kappa_\ell = \eta_\ell^{-(k-1)}$ and thus $\mu = k-1$.

Altogether, we have verified the assumptions from the ε-cost theorem Theorem 3.5.2. Thus, we obtain the ε-cost scalings for the multi-level estimator which were stated in Example 3.5.3.

Let us put that theory into practice in the next section.

5.6 Numerical experiments

Here we present some numerical experiments for linear and non-linear Schrödinger equations with uncertain parameters which confirm the theoretical predictions from the previous section. We start with the simpler one, the linear Schrödinger equation.

5.6.1 Application to the linear Schrödinger equation

This section will appear in a similar form in [61].

We consider the one-dimensional parametric linear Schrödinger equation

$$\partial_t u(t,x,y) = \frac{\mathrm{i}}{2}\partial_x^2 u(t,x,y) + \mathrm{i}V(x,y)u(t,x,y), \qquad t\in[0,T],\ x\in\mathbb{T}_K,\ y\in\Gamma, \tag{5.61a}$$

$$u(0,x,y) = u_0(x,y), \qquad x\in\mathbb{T}_K,\ y\in\Gamma \tag{5.61b}$$

on a scaled torus $\mathbb{T}_K = \mathbb{R}/(2\pi K\mathbb{Z})$ with potential $V\colon \mathbb{T}_K\times\Gamma\to\mathbb{R}$ and initial data $u_0\colon \mathbb{T}_K\times\Gamma\to\mathbb{C}$. Typical quantities of interest are the *position* of the particle

$$P\colon L^2(\mathbb{T}_K)\to\mathbb{R}, \qquad u\mapsto \int_{\mathbb{T}_K} x|u(x)|^2\mathrm{d}x, \tag{5.62}$$

and the probability that the particle is located in a set $S\subseteq\mathbb{T}_K$,

$$M_S\colon L^2(\mathbb{T}_K)\to\mathbb{R}, \qquad u\mapsto \int_S |u(x)|^2\mathrm{d}x. \tag{5.63}$$

Besides approximating the solution of (5.61) itself, we are also interested in computing these two quantities.

In order to study the convergence of the MLSC method, we compare the multi-level approximation $u_J^{(\mathrm{ML})} = u_J^{(\mathrm{ML})}(T)$ from equation (3.13) computed at time T to a reference $u_{\mathrm{ref}}(T)$. Now we explain how such a reference can be obtained without relying on a very fine collocation approximation and without

using a splitting method for the time integration. The reference solution is thus independent of the methods we try to verify here.

In order to simplify the formulas, the factor $1/2$ in front of the second derivative was introduced in (5.61a). This factor was missing in (5.2) but does not affect the preceding analysis substantially.

Reference solutions in spatial dimension $N = 1$. If we replace the scaled torus $\mathbb{T}_K$ by $\mathbb{R}$ and assume that the potential is a quadratic polynomial in x of the form

$$V(x, y) = -\nu(y)(x - \kappa(y))^2 - \gamma(y), \tag{5.64}$$

then a family of solutions to the linear Schrödinger equation (5.61) is given by

$$u(t, x, y) = \exp(w(t, x)) \tag{5.65a}$$

$$\text{with} \qquad w(t, x, y) = \frac{\mathrm{i}}{2} C(t, y)(x - q(t, y))^2 + \mathrm{i}p(t, y)(x - q(t, y)) + \mathrm{i}\xi(t, y), \tag{5.65b}$$

see [72, Sec. II.4.1]. The quantities $p(t, y), q(t, y) \in \mathbb{R}$ and $C(t, y), \xi(t, y) \in \mathbb{C}$ are related by the (ordinary) differential equations

$$\partial_t q(t, y) = p(t, y), \tag{5.66a}$$

$$\partial_t p(t, y) = -2\nu(y)(q(t, y) - \kappa(y)), \tag{5.66b}$$

$$\partial_t \xi(t, y) = \frac{\mathrm{i}C(t, y)}{2} + \frac{1}{2} p(t, y)^2 - \nu(y)(q(t, y) - \kappa(y))^2 - \gamma(y), \tag{5.66c}$$

$$\partial_t C(t, y) = -C(t, y)^2 - 2\nu(y), \tag{5.66d}$$

supplied with initial values. If the imaginary part of $C(t, y)$ is strictly positive for $t = 0$, then this is the case for all $t \geq 0$ and $|u(t, \cdot, y)|$ is a real-valued Gaussian. However, neither the potential (5.64) nor the solution (5.65) are periodic in space, and thus this construction does not seem to be compatible with the periodic boundary conditions in (5.61). But as the absolute value of the Gaussian (5.65a) decays exponentially, the error caused by imposing periodic boundary conditions at $\pm\pi K$ is negligible if K is chosen sufficiently large. If this is the case, then (5.65) provides indeed sufficiently accurate solutions to the Schrödinger equation (5.61) on $\mathbb{T}_K = \mathbb{R}/(2\pi K\mathbb{Z})$.

To obtain a reference solution for (5.61) with potential (5.64) and initial value given by (5.65) for $t = 0$, $\eta_{\mathrm{ref}} = 100.000$ vectors $y^1, \dots, y^{\eta_{\mathrm{ref}}} \in \Gamma$ from a Halton sequence in $\Gamma = [-1, 1]^d$ were used. For each y^j, the ODE system (5.66) was solved with a Dormand-Prince method with relative error tolerance set to 10^{-9}. Since we focus on the error induced by discretising the stochastic and temporal variables, we use the same space discretisation for the reference solution and the approximations coming from the MLSC method, namely a Fourier collocation method with $M = 2^{10}$ grid points.

In the following examples, computations are made on the time interval $[0, 1]$ and the spatial domain $[-3\pi, 3\pi]$ with periodic boundary conditions, so $K = 3$. All errors are computed at the endpoint of the time interval $T = 1$.

Example 5.6.1 (Two-dimensional example)**.**

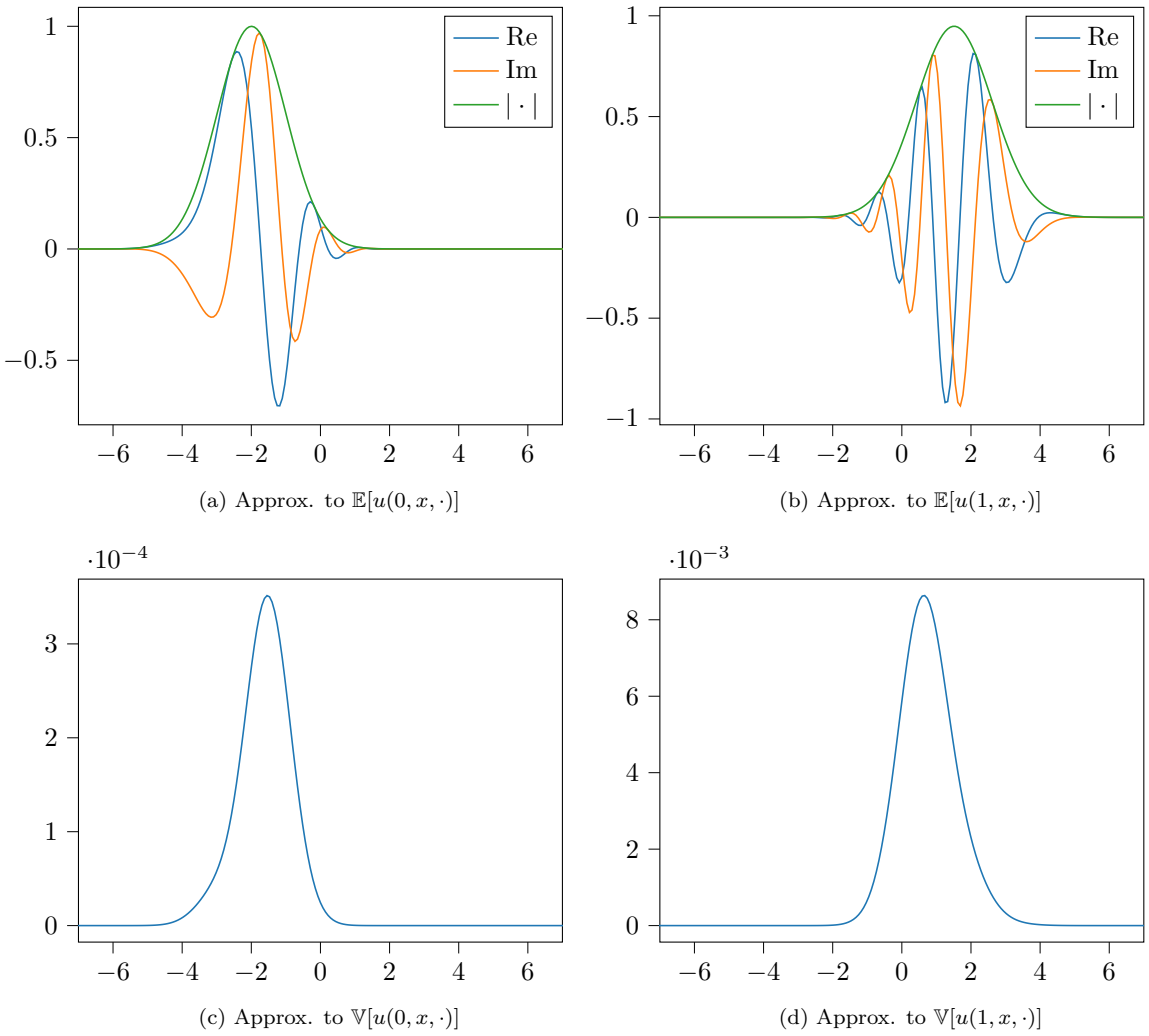
As a first test, we consider a toy problem with two-dimensional parameter space, so $\Gamma = [-1, 1]^2$. For $y = (y_1, y_2) \in \Gamma$, the potential is given by (5.64) with

$$\nu(y) = 1 + \frac{\delta}{3}(y_1 + 2y_2), \qquad \kappa(y) = \frac{1}{2}\left(1 + \frac{\delta}{2}(y_1 + y_2)\right), \qquad \gamma(y) = 1 + \frac{\delta}{3}(y_1 + y_2^2)$$

and $\delta = \frac{1}{20}$. The initial values for (5.66) at time $t = 0$ are set to

$$\Big(C(0,y), q(0,y), p(0,y), \xi(0,y)\Big) = \left(1 + \frac{\delta}{4} y_2^2 + \mathrm{i}, -2 + \delta y_1^2 y_2^2, 2, 1\right),$$

which defines $u(0,x,y)$ via (5.65). The initial value $u(0,x,y)$ and the corresponding reference solution u_{ref} at time $T = 1$ is shown in Figure 5.1. The reference solution u_{ref} was computed as explained in the paragraph before this example ($\eta_{\mathrm{ref}} = 100.000$, $M = 2^{10}$).



(a) Approx. to $\mathbb{E}[u(0,x,\cdot)]$

(b) Approx. to $\mathbb{E}[u(1,x,\cdot)]$

(c) Approx. to $\mathbb{V}[u(0,x,\cdot)]$

(d) Approx. to $\mathbb{V}[u(1,x,\cdot)]$

Figure 5.1: Gaussian solution to the LSE at times $t = 0$ and $t = 1$

The error in the norm of the space $L^2_\varrho(\Gamma, X)$ with $X = L^2(\mathbb{T}_K)$ is computed numerically by

$$\Big(\frac{2\pi K}{N_{\mathrm{ref}} M} \sum_{j=1}^{N_{\mathrm{ref}}} \sum_{k=1}^{M} |u_J^{(\mathrm{ML})}(x_k, y^j) - u_{\mathrm{ref}}(T, x_k, y^j)|^2\Big)^{1/2} \approx \|u_J^{(\mathrm{ML})} - u_{\mathrm{ref}}(T)\|_{L^2_\varrho(\Gamma,X)},$$

where $x_k \in \mathbb{T}_K$, $k = 1, \dots, M = 2^{10}$, are the Fourier collocation points. This error is labelled by "error in $L^2_\varrho(\Gamma, X)$" in diagrams. We investigate the following two other types of error:

- The error in the quantity P from (5.62),

$$\left|\mathbb{E}\left[P\left(u_J^{(\mathrm{ML})}(T)\right) - P(u_{\mathrm{ref}}(T))\right]\right|. \tag{5.67}$$

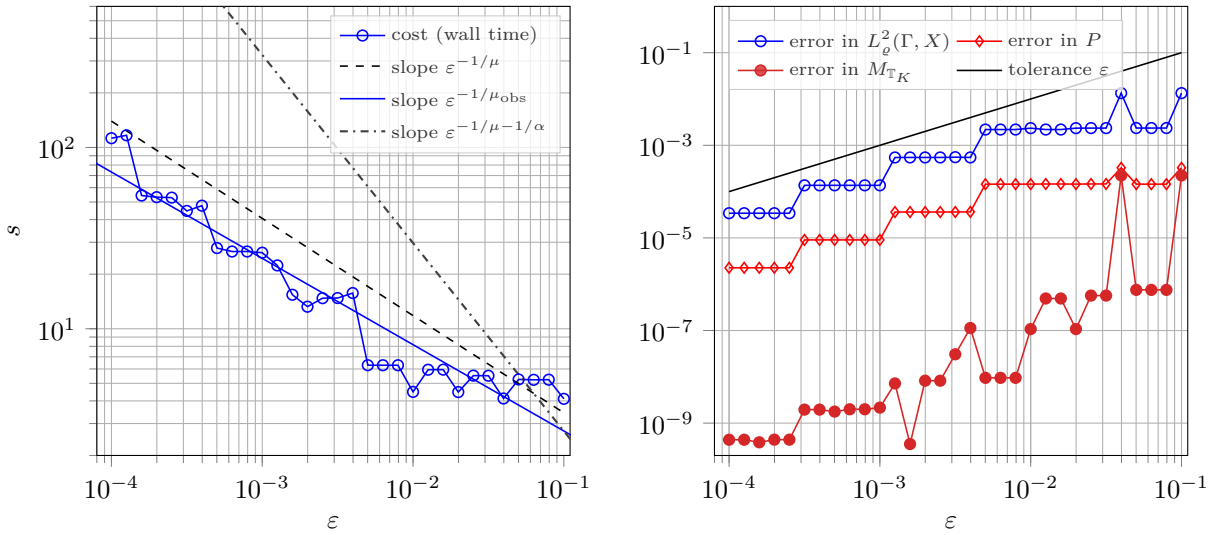
- The error in $M_{\mathbb{T}_K}$ defined in (5.63), which is

$$\left|\mathbb{E}\left[M_{\mathbb{T}_K}\left(u_J^{(\mathrm{ML})}(T)\right) - M_{\mathbb{T}_K}\left(u_{\mathrm{ref}}(T)\right)\right]\right|. \tag{5.68}$$

These two errors are denoted by "error in P" and "error in $M_{\mathbb{T}_K}$" in diagrams. Of course, (5.67) and (5.68) are computed by suitable discrete versions. The quantity $M_{\mathbb{T}_K}$ is a conserved quantity of the NLS and its splitting approximation. It is nevertheless worth examining the error in this quantity, since the stochastic discretisation contributes to the error, but not the temporal discretisation.

As the solution is smooth enough, the requirements B1 and B2 for the multi-level method are satisfied with $\alpha = \beta = 2$ for $\mathcal{X} = L^2(\mathbb{T}_K)$. Using the maximal step-size $\tau_0 = 0.1$, we could confirm Assumption B1 numerically with $C_T = 0.89$, $\alpha = 1.99$ and, after setting $\beta = \alpha$, Assumption B2 with $\mu = 1.86$ and $C = C_I C_\star = 1291$. This was done as explained in Section 3.5.2. Note that the value $\alpha = 1.99$ agrees very well with the order 2 expected from Theorem 5.3.11 for $\boldsymbol{k} = \mathbf{0}$.

In this example we use the "up/down" rounding strategy. The result of a convergence test series is shown in Figure 5.2. In contrast to all other numerical examples, the computations for this small toy problem were carried out on a laptop with Intel(R) Core(TM) i7-7500U CPU, so comparing the wall times here to the ones in other experiments later does not make any sense.



(a) Cost vs. given ε, rounding "up/down"

(b) Error vs. given ε, rounding "up/down"

Figure 5.2: MLSC for the LSE: $d = 2$, $T = 1$, $\mu = 1.86$, $\mu_{\mathrm{obs}} = 2.05$, $M = 2^{10}$

Figure 5.2(b) confirms that the error in $L^2_\varrho(\Gamma, X)$ is indeed smaller than the given tolerance ε. The same is true for the other two types of error. Since $2 = \beta > \mu = 1.86$, we expect from the ε-cost theorem (Theorem 3.5.2) that the computational cost scales as $\varepsilon^{-1/\mu}$. Figure 5.2(a) shows, however, that the wall time of the method —o— scales rather as $\varepsilon^{-1/\mu_{\mathrm{obs}}}$ (——) with the slightly larger value $\mu_{\mathrm{obs}} = 2.05 > \mu$.

We have also included a line -·-·- with slope $\varepsilon^{-1/\mu-1/\alpha}$ which corresponds to the theoretical scaling of the single-level collocation method in light of (3.23).

In the next example, we examine more uncertainty in the problem in the sense of a larger stochastic dimension.

Example 5.6.2 (Ten-dimensional example)**.**

Here we consider a ten-dimensional parameter space $\Gamma = [-1,1]^{10}$ and the quadratic potential (5.64) with

$$\nu(y) = 1 + \frac{\delta}{3}(y_1 + 2y_2), \qquad \kappa(y) = \frac{1}{2}\left(1 + \frac{\delta}{2}(y_3 + y_4)\right), \qquad \gamma(y) = 1 + \frac{\delta}{3}(y_5 + y_6^2)$$

and $\delta = \frac{1}{20}$ for $y = (y_1, \dots, y_{10}) \in \Gamma$. The initial values at time $t = 0$ are given by

$$\Big(C(0,y), q(0,y), p(0,y), \xi(0,y)\Big) = \left(1 + \frac{\delta}{4}y_7^2 + \mathrm{i}, -2 + \delta y_8^2 y_9^2, 2 + \delta y_{10}, 1\right).$$

The remaining parameters of the equation are the same as in the two-dimensional example before.

Here we apply the multi-level approach for QoIs from Section 3.6 to approximate the functional P from (5.62). Thus, our is goal is to achieve

$$|\mathbb{E}[P(u(T)) - P(u_J^{(\mathrm{ML})}(T))]| \leq \varepsilon$$

for given $\varepsilon > 0$, where $T = 1$. In general, approximating the functional P should be easier than approximating the solution itself. The challenge in this example is certainly the large dimension of the parameter set Γ.

The assumptions for the ε-cost theorem for QoIs Theorem 3.6.1 with $\Upsilon = P$ were confirmed numerically with constants and parameters $\mu = 1.11$, $C = C_I C_\star = 265$, $C_T = 0.077$ and $\beta = \alpha = 2.10$.

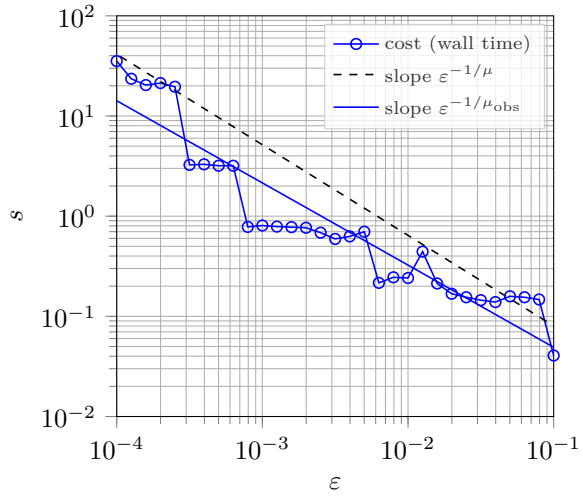
The reference solution u_{ref} was computed as in the previous example with $\eta_{\mathrm{ref}} = 100.000$ and $M = 2^{10}$.

Results are shown in Figure 5.3 for maximal step-size $\tau_0 = 0.1$ and two different choices of rounding strategy, "up/down" and "down". The second one is considered because we expect that the overhead of rounding up in this dimension could be too large and cause a deterioration of the computational cost.

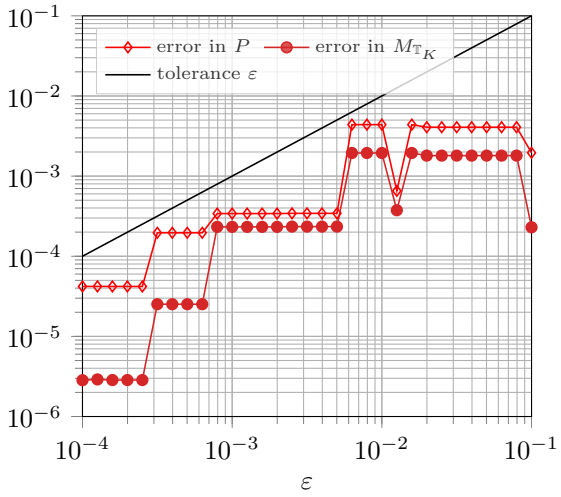
Figure 5.3(a) shows that the computational cost —o— scales as $\varepsilon^{-1/\mu_{\mathrm{obs}}}$ (——) with $\mu_{\mathrm{obs}} = 1.21$. This is significantly better than expected, because Theorem 3.5.2 states that the computational cost grows proportional to $\varepsilon^{-1/\mu}$ (- - -) with $\mu = 1.11$ when $\varepsilon \to 0$. We explain this as follows: In dimensions where generally less levels are feasible, it is possible to reuse already computed solutions more often. This is plausible since rounding to the nearest sparse grid depth often requires a solution which has already been computed. So whenever a single-level collocation approximation with depth L and τ is required in the multi-level estimator, we first check if such a solution has already been computed. If this is the case, then there is no necessity to compute it again. This explains that sometimes improved cost scalings can be observed.

Figure 5.3(b) shows that the error in the quantity P stays below the tolerance for all ε. Thus the results agree with the theoretical statement from Theorem 3.5.2.

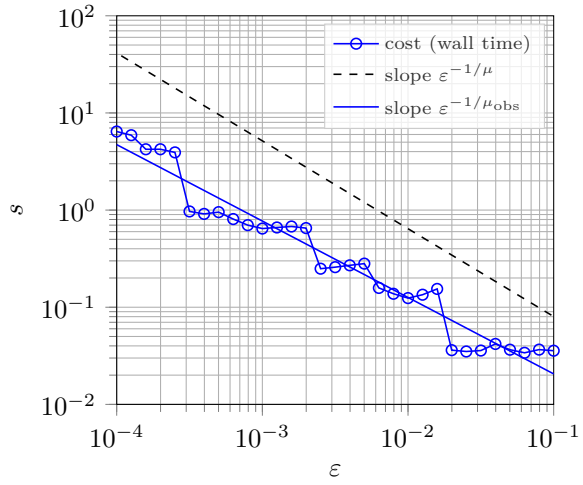
In the next section, we finally treat the non-linear Schrödinger equation.



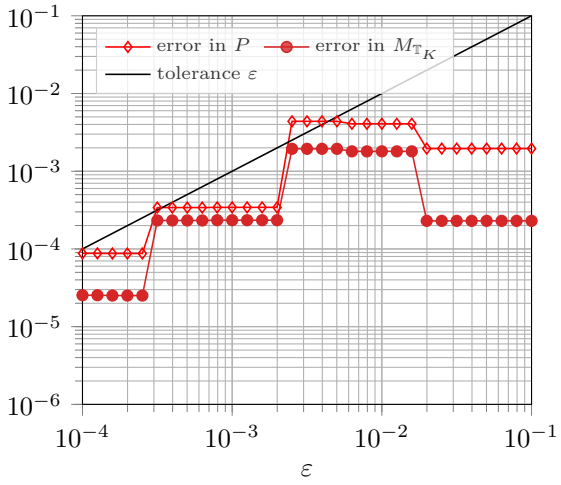
(a) Cost vs. given ε, "up/down", $\mu_{\text{obs}} = 1.21$



(b) Error vs. given ε, rounding "up/down"



(c) Cost vs. given ε, "down", $\mu_{\text{obs}} = 1.27$



(d) Error vs. given ε, rounding "down"

Figure 5.3: MLSC for the LSE: $d = 10$, $T = 1$, $\mu = 1.11$, $M = 2^{10}$

5.6.2 Application to the non-linear Schrödinger equation

We examine a problem with spatial dimension $N = 1$ and a five-dimensional parameter space, so $d = 5$. Thus, we consider

$$\begin{aligned} \partial_t u(t,x,y) &= \mathrm{i}\partial_x^2 u(t,x,y) + \mathrm{i}V(x,y)u(t,x,y) + \mathrm{i}|u(t,x,y)|^2 u(t,x,y), && t \in [0,T],\ x \in \mathbb{T}_K,\ y \in \Gamma, \\ u(0,x,y) &= u_0(x,y), && x \in \mathbb{T}_K,\ y \in \Gamma \end{aligned}$$

with $\Gamma = [-1,1]^5$ and $K = 2$. The initial data is given by

$$u_0(x,y) = \sin(0.5x)^5 \cdot (1 + 0.1y_1 + 0.01y_2).$$

The potential is almost quadratic, but smoothened towards the boundary such that its derivatives also fulfill the periodic boundary condition. We choose

$$V(x,y) = -\kappa(x)\frac{x^2}{2} \cdot (1 + 0.1y_3 + 0.05y_4 + 0.025y_5)$$

with the smoothing function

$$\kappa(x) = \begin{cases} \exp(-1.2 \cdot (\pi - x)^2) + \exp(-1.2 \cdot (-\pi - x)^2), & |x| > \pi, \\ 1, & \text{else.} \end{cases}$$

Important quantities of interest for the solution of this equation in the context of BECs are the energy inside the set $S \subseteq \mathbb{T}_K$ given by

$$E_S(u(t,\cdot,y),y) = \int_S |\nabla u(t,x,y)|^2 - \left(V(x,y) + \frac{1}{2}|u(t,x,y)|^2\right)|u(t,x,y)|^2 \mathrm{d}x, \tag{5.70}$$

or the expected amount of bosons located in S,

$$M_S(u(t,\cdot,y)) = \int_S |u(t,x,y)|^2 \mathrm{d}x.$$

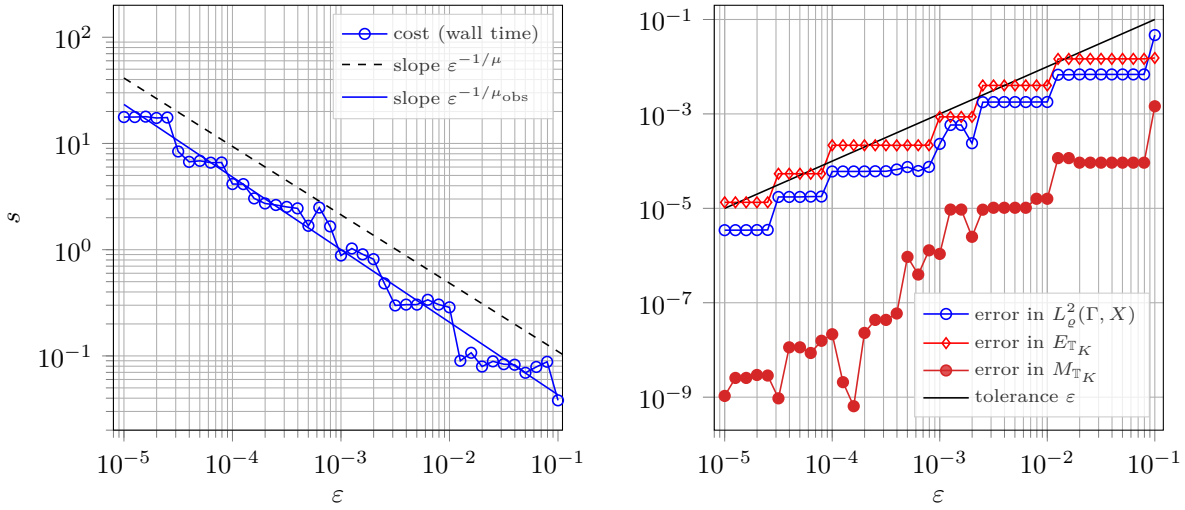
Thus, these two quantities are also examined in this example. As for the LSE, $M_{\mathbb{T}_K}$ is a conserved quantity of the solution and the splitting approximation, so

$$M_{\mathbb{T}_K}(u(t,\cdot,y)) = M_{\mathbb{T}_K}(u_0(\cdot,y)) \quad \text{for all} \quad t \in [0,T].$$

To obtain a reference solution for this problem, we take a fine single-level collocation approximation resolved with a sparse grid of depth $L_{\text{ref}} = 8$ (with $\eta_{L_{\text{ref}}} = 51713$ nodes) and a temporal step-size $\tau_{\text{ref}} = 10^{-5}$. The spatial discretisation for the reference solution and the approximations is done via Fourier collocation, as in case of the linear Schrödinger equation before. We use $M = 2^{10}$ spatial grid points.

The constants for the multi-level approximation of the solution itself at final time $T = 1$ are computed as $\mu = 1.55$, $C = 2770$, $C_T = 3.49$ and $\beta = \alpha = 2$. The result with rounding strategy "down" and $\tau_0 = 0.1$ is depicted in Figure 5.4. The error in the energy $E_{\mathbb{T}_K}$ from (5.70) is displayed, too, and denoted by "error in $E_{\mathbb{T}_K}$".

Although the spatial dimension in all of the examples in this section was $N = 1$, it is of course possible to consider $N = 2$ or $N = 3$, too. We treated the one-dimensional case mainly in order to spent the computing power in the resolution of the parameter space, and not the spatial domain. Of course, the simulation of the higher-dimensional equations is certainly more interesting. Nevertheless, the above examples can be seen as a proof of concept for the MLSC method. The simulation of the two- and three-dimensional non-linear Schrödinger equations with uncertain parameters is left as a goal for future research.



(a) Cost vs. given ε, rounding "down", $\mu_{\text{obs}} = 1.46$

(b) Error vs. given ε, rounding "down"

Figure 5.4: MLSC for the NLS: $d = 5$, $T = 1$, $\mu = 1.55$, $\mu_{\text{obs}} = 1.46$, $M = 2^{10}$

CHAPTER 6

Summary and outlook

Summary

In this thesis, we examined two different types of time-dependent partial differential equations, namely non-linear parabolic and non-linear Schrödinger equations. For both of them, we assumed that some of their parameters are only given with an amount of uncertainty which is modelled by an additional parameter. This parameter, the *stochastic variable*, is thought of being a realisation of a uniformly distributed d-dimensional random variable and is discretised via a collocation strategy based on sparse grids. The choice of sparse grids takes into account that d is typically larger than the spatiotemporal dimensions of the equations under consideration and the parameter space is therefore affected by the *curse of dimensionality*.

As the PDEs are time-dependent, computing approximations to the solutions of these equations requires discretisations for the stochastic *and* the temporal variable (not to mention the third – spatial – discretisation, which was not studied in detail here). These two discretisations were combined via single-level and multi-level strategies, and the goal of using various levels is the reduction of the total computational cost. A multi-level strategy with respect to these two variables is seldom studied in the UQ literature and our work thus presents a relevant contribution to this research area. We stated assumptions under which the multi-level approximation is both convergent and computable with a more preferable scaling of the computational cost compared to a naive single-level approach. The verification of these assumptions, however, is not straightforward but possible, as we demonstrate in two specific situations:

- In the *parabolic* case, the uncertainty investigated here enters in diffusion, convection and reaction terms and the initial data in a smooth, more specifically *analytic* way. The regularity of the solution in terms of the stochastic variable is then usually analytic, too. For the temporal discretisation, we choose an implicit-explicit trapezoidal splitting (IMEXT) method which treats the potentially stiff linear terms implicitly and the reaction terms explicitly. We prove that this method is second-order convergent under certain regularity assumptions. Both the result and some of the techniques to

prove it seem to be new and are not only interesting in the context of uncertainty quantification, but also for the time integration community. To our knowledge, the error bound stating second-order convergence of the IMEXT method closes a current gap in the literature. Additionally, the convergence result is uniform in the parameter space, which is used in the convergence analysis of the single-level method and for the verification of the assumptions for the multi-level approach.

- In the *Schrödinger* case, the uncertainty enters in the potential and the initial data of the problem, but not in the dispersive part. Here, the regularity in terms of the stochastic variable y is not analytic anymore, but finite in the sense of C^k with respect to the individual dimensions of the parameter space. If the solution has this y-regularity, then the approximations of the Strang splitting method used for the temporal discretisation have the same regularity, too. In the corresponding (multivariate) C^k-norm, we prove that the method is second-order convergent. This new result is an extension of the convergence results for the "deterministic" non-linear Schrödinger equation (meaning the NLS without uncertain parameters).

For both problem classes, convergence results for the single- and multi-level stochastic collcation methods are stated and proved. They are accompanied by statements about their theoretical cost.

These theoretical results are supported by several numerical experiments. They confirm the cost and error analysis for the multi-level method. Additionally, we indicate some practical limitations of the multi-level stochastic collocation approach – both in comparison with the single-level method, and regarding the usability for very high stochastic dimensions, problems with low regularity in the stochastic variable, or in connection with time integration methods with severe step-size restrictions.

The results developed in this thesis are novel contributions in two areas of mathematics: The first one is (analysis and implementation of) numerical methods for uncertainty quantification and the second one is numerical analysis of time integration schemes for partial differential equations. The in-depth analysis of the interplay between stochastic and temporal discretisations in this thesis is a relevant contribution to the current state of the literature.

Outlook

There are many ways to proceed from here: Further theoretical study of multi-level methods, extension of the problem classes for which they can be used, extension of the class of numerical methods with which they can be combined, or improvement of their implementation. More specifically, we propose further research on the following points:

- Extension of the error analysis for multi-level methods and verification of the conditions for convergence for other time integrators or in a more general framework
- Incorporate the spatial discretisation as a third discretisation type into the multi-level approach; for example by using the multi-index stochastic collocation method or one of its variants
- Extension of the problem class to hyperbolic problems
- Extension of the problem class to problems where the time integration itself is very demanding, such as highly oscillatory problems

- Anisotropy in the stochastic variables, potentially extending the stochastic dimension even further in applications
- Implementations of multi-level methods which are not only parallel with respect to the different stochastic collocation points, but also parallel in space and/or in time. This could potentially increase their usability on large supercomputers to tackle more interesting real-world problems.
- Gain further insights concerning the limitations of stochastic collocation methods by comparing them with (multi-level) Monte Carlo methods and Gaussian regression in situations where they might be competitive.

This list almost fills itself as multi-level stochastic collocation methods rely on many techniques from different areas and most of them have some extensions which might be worth incorporating. Nevertheless, we believe that the above-mentioned points are certainly the most interesting ones and some of them could be tackled with the current state of knowledge, but we do not claim that this list is anywhere near complete.

APPENDIX A

Miscellaneous results

Lemma A.1 (Growth of log-powers)**.** *For $E \in \mathbb{N}$, we have the estimate*

$$\log(\eta)^E \leq \left(\frac{E}{\mathrm{e}}\right)^E \eta$$

for all $\eta \geq 1$.

Proof. Consider the function $f\colon [1,\infty) \to \mathbb{R},\ f(x) = x^{-1}\left(\log(x)\right)^E$. The derivative of f is

$$f'(x) = \frac{E\log(x)^{E-1} - \log(x)^E}{x^2} = \log(x)^{E-1}\frac{E - \log(x)}{x^2},$$

and thus $f'(x) = 0$ holds if and only if

$$x = x_* \coloneqq \mathrm{e}^E \qquad \text{or} \qquad x = x_{**} \coloneqq 1.$$

As $f(x_{**}) = 0$, the maximum of f is to be found presumably at x_*. The second derivative of f is

$$\begin{aligned} f''(x) &= (E-1)\log(x)^{E-2}\frac{E-\log(x)}{x^3} - \log(x)^{E-1}\frac{1+(E-\log(x))2}{x^3} \\ &= \frac{\log(x)^{E-2}}{x^3}\left[(E-1)E - 3E\log(x) + 2\log(x)^2\right] \end{aligned}$$

and it holds

$$f''(x_*) = -\frac{E^{E-1}}{\mathrm{e}^{3E}} < 0.$$

Clearly, $f(1) = 0$ and $\lim_{x\to\infty} f(x) = 0$, thus we obtain $f(x) \leq f(x_*)$ for all $x \in [1,\infty)$ and hence

$$\log(x)^E \leq f(x_*)x = \left(\frac{E}{\mathrm{e}}\right)^E x$$

as claimed. □

Lemma A.2. *For $V, \chi \in \mathbb{R}$ and $w \in \mathbb{C}$, let*

$$B[w] = V + \chi|w|^2.$$

The solution of the initial value problem

$$v'(t) = iB[v(t)]v(t), \qquad t \geq 0, \tag{A.1a}$$

$$v(0) = v_0 \tag{A.1b}$$

for given $v_0 \in \mathbb{C}$ is given by

$$v(t) = \mathrm{e}^{itB[v_0]}v_0 \tag{A.2}$$

for $t \geq 0$.

Proof. First, observe that

$$|\mathrm{e}^{\mathrm{i}tB[v_0]}v_0| = |v_0|$$

for $t \geq 0$ since $B[v_0] \in \mathbb{R}$. Now we compute

$$\frac{\mathrm{d}}{\mathrm{d}t}(\mathrm{e}^{\mathrm{i}tB[v_0]}v_0) = \mathrm{i}B[v_0]\mathrm{e}^{\mathrm{i}tB[v_0]}v_0 = \mathrm{i}B\left[\mathrm{e}^{\mathrm{i}tB[v_0]}v_0\right]\mathrm{e}^{\mathrm{i}tB[v_0]}v_0$$

and obtain that v from (A.2) solves (A.1). □

APPENDIX B

On φ-functions

Here we introduce the so-called φ-functions which often appear in the context of inhomogeneous ODEs or PDEs and in the construction of exponential integrators, see e.g. [53, 60].

For us, they are quite useful in the derivation of error formulas for the IMEXT method in Section 4.5.2 and the splitting method in Section 5.3.5.

Definition B.1 (φ-functions)**.** Let $(\mathcal{A}, \mathcal{D}(\mathcal{A}))$ be a generator of a strongly continuous semigroup on $\mathcal{X}$. For $t \geq 0$, we define $\varphi_j(t\mathcal{A})\colon \mathcal{X} \to \mathcal{X}$ as the linear operator on $\mathcal{X}$ with

$$\begin{aligned} \varphi_0(t\mathcal{A})v &= \mathrm{e}^{t\mathcal{A}}v, \\ \varphi_j(t\mathcal{A})v &= \int_0^1 \frac{\vartheta^{j-1}}{(j-1)!}\mathrm{e}^{(1-\vartheta)t\mathcal{A}}v\,\mathrm{d}\vartheta, \qquad j \in \mathbb{N}, \end{aligned}$$

for $v \in \mathcal{X}$.

The most important properties of φ-functions are summarised in the following lemma.

Lemma B.2 (Properties of φ-functions)**.** *Let $(\mathcal{A}, \mathcal{D}(\mathcal{A}))$ be a generator of a strongly continuous semigroup on $\mathcal{X}$. The following statements hold.*

(a) For every $j, k \in \mathbb{N}_0$ and $t \geq 0$, the operator

$$\varphi_j(t\mathcal{A})\colon \mathcal{D}(\mathcal{A}^k) \to \mathcal{D}(\mathcal{A}^k)$$

is bounded. If $(\mathcal{A}, \mathcal{D}(\mathcal{A}))$ generates a semigroup of contractions, then $\|\varphi_j(t\mathcal{A})\|_{\mathcal{L}(\mathcal{D}(\mathcal{A}^k))} \leq \frac{1}{j!}$.

(b) If $v \in \mathcal{D}(\mathcal{A}^k)$ and $t > 0$, then $\varphi_j(t\mathcal{A})v \in \mathcal{D}(\mathcal{A}^{k+1})$ for all $j \in \mathbb{N}$ and the equation

$$t\mathcal{A}\varphi_j(t\mathcal{A})v = \varphi_{j-1}(t\mathcal{A})v - \frac{1}{(j-1)!}v, \qquad j \in \mathbb{N}, \tag{B.1}$$

holds in $\mathcal{D}(\mathcal{A}^k)$.

(c) For every $m \in \mathbb{N}$, recursion (B.1) implies the "Taylor expansion"

$$\mathrm{e}^{t\mathcal{A}}v = \sum_{j=0}^{m-1} \frac{t^j}{j!}\mathcal{A}^j v + (t\mathcal{A})^m \varphi_m(t\mathcal{A})v \tag{B.2}$$

for any $v \in \mathcal{D}(\mathcal{A}^{m-1})$.

The following lemma is useful to extract additional t-powers from differences of successive φ-functions.

Lemma B.3. *The φ-functions satisfy the recursion formula*

$$\varphi_{j-1}(t\mathcal{A})v - \varphi_j(t\mathcal{A})v = t\mathcal{A}(\varphi_j(t\mathcal{A}) - \varphi_{j+1}(t\mathcal{A}))v + \frac{j-1}{j!}v$$

for $v \in \mathcal{X}$, $j \in \mathbb{N}$ and $t > 0$.

Observe that for $j \geq 2$, the left-hand side is contained in $\mathcal{D}(\mathcal{A})$, whereas the individual summands on the right-hand side only belong to $\mathcal{X}$ in general.

Proof. By (B.1), we have

$$\begin{aligned} t\mathcal{A}(\varphi_j(t\mathcal{A}) - \varphi_{j+1}(t\mathcal{A})) &= t\mathcal{A}\varphi_j(t\mathcal{A}) - \varphi_j(t\mathcal{A}) + \frac{1}{j!} \\ &= \varphi_{j-1}(t\mathcal{A}) - \frac{1}{(j-1)!} - \varphi_j(t\mathcal{A}) + \frac{1}{j!} \\ &= \varphi_{j-1}(t\mathcal{A}) - \varphi_j(t\mathcal{A}) + \frac{1}{j!} - \frac{1}{(j-1)!}. \end{aligned}$$

This shows the statement. □

Bibliography

[1] H. Amann and J. Escher. *Analysis II*. Translated from the 1999 German original by Silvio Levy and Matthew Cargo. Birkhäuser Verlag, Basel, 2008. ISBN: 978-3-7643-7472-3.

[2] U. M. Ascher, S. J. Ruuth, and R. J. Spiteri. 'Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations'. In: *Applied Numerical Mathematics* 25.2 (1997). Special Issue on Time Integration, pp. 151–167. ISSN: 0168-9274. DOI: 10.1016/S0168-9274(97)00056-1.

[3] I. Babuška, F. Nobile, and R. Tempone. 'A stochastic collocation method for elliptic partial differential equations with random input data'. In: *SIAM Review* 52.2 (2010), pp. 317–355. ISSN: 0036-1445. DOI: 10.1137/100786356.

[4] I. Babuška, R. Tempone, and G. E. Zouraris. 'Galerkin finite element approximations of stochastic elliptic partial differential equations'. In: *SIAM Journal on Numerical Analysis* 42.2 (2004), pp. 800–825. ISSN: 0036-1429. DOI: 10.1137/S0036142902418680.

[5] I. Babuška, R. Tempone, and G. E. Zouraris. 'Solving elliptic boundary value problems with uncertain coefficients by the finite element method: the stochastic formulation'. In: *Computer Methods in Applied Mechanics and Engineering* 194.12-16 (2005), pp. 1251–1294. ISSN: 0045-7825. DOI: 10.1016/j.cma.2004.02.026.

[6] J. Bäck, F. Nobile, L. Tamellini, and R. Tempone. 'Stochastic spectral Galerkin and collocation methods for PDEs with random coefficients: a numerical comparison'. In: *Spectral and high order methods for partial differential equations*. Vol. 76. Lect. Notes Comput. Sci. Eng. Springer, Heidelberg, 2011, pp. 43–62. DOI: 10.1007/978-3-642-15337-2_3.

[7] W. Bao, D. Jaksch, and P. A. Markowich. 'Numerical solution of the Gross–Pitaevskii equation for Bose–Einstein condensation'. In: *Journal of Computational Physics* 187.1 (2003), pp. 318–342. ISSN: 0021-9991. DOI: 10.1016/S0021-9991(03)00102-5.

[8] V. Barthelmann, E. Novak, and K. Ritter. 'High dimensional polynomial interpolation on sparse grids'. In: *Advances in Computational Mathematics* 12.4 (2000), pp. 273–288. DOI: 10.1023/A:1018977404843.

[9] J. Beck, R. Tempone, F. Nobile, and L. Tamellini. 'On the optimal polynomial approximation of stochastic PDEs by Galerkin and collocation methods'. In: *Mathematical Models and Methods in Applied Sciences* 22.9 (2012), pp. 1250023, 33. ISSN: 0218-2025. DOI: 10.1142/S0218202512500236.

[10] Á. Bényi and T. Oh. 'The Sobolev inequality on the torus revisited'. In: *Publicationes Mathematicae Debrecen* 83.3 (2013), pp. 359–374. ISSN: 0033-3883. DOI: 10.5486/PMD.2013.5529.

[11] H. Bijl, D. Lucor, S. Mishra, and C. Schwab. *Uncertainty Quantification in Computational Fluid Dynamics*. Springer International Publishing, 2013. ISBN: 978-3-319-00885-1. DOI: 10.1007/978-3-319-00885-1.

[12] K. Binder and D. W. Heermann. *Monte Carlo Simulation in Statistical Physics: An Introduction*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. ISBN: 978-3-642-03163-2. DOI: 10.1007/978-3-642-03163-2.

[13] P. P. Boyle. 'Options: A Monte Carlo approach'. In: *Journal of Financial Economics* 4.3 (1977), pp. 323–338. ISSN: 0304-405X. DOI: https://doi.org/10.1016/0304-405X(77)90005-8.

[14] H.-J. Bungartz and M. Griebel. 'Sparse grids'. In: *Acta Numerica* 13 (2004), pp. 147–269. ISSN: 0962-4929. DOI: 10.1017/S0962492904000182.

[15] J. Charrier, R. Scheichl, and A. L. Teckentrup. 'Finite element error analysis of elliptic PDEs with random coefficients and its application to multilevel Monte Carlo methods'. In: *SIAM Journal on Numerical Analysis* 51.1 (2013), pp. 322–352. ISSN: 0036-1429. DOI: 10.1137/110853054.

[16] A. Chkifa, A. Cohen, and C. Schwab. 'Breaking the curse of dimensionality in sparse polynomial approximation of parametric PDEs'. In: *Journal de Mathématiques Pures et Appliquées. Neuvième Série* 103.2 (2015), pp. 400–428. ISSN: 0021-7824. DOI: 10.1016/j.matpur.2014.04.009.

[17] A. Cialdea and V. Maz'ya. 'Criterion for the L^p-dissipativity of second order differential operators with complex coefficients'. In: *Journal de Mathématiques Pures et Appliquées. Neuvième Série* 84.8 (2005), pp. 1067–1100. ISSN: 0021-7824. DOI: 10.1016/j.matpur.2005.02.003.

[18] C. W. Clenshaw and A. R. Curtis. 'A method for numerical integration on an automatic computer'. In: *Numerische Mathematik* 2 (1960), pp. 197–205. ISSN: 0029-599X. DOI: 10.1007/BF01386223.

[19] K. A. Cliffe, M. B. Giles, R. Scheichl, and A. L. Teckentrup. 'Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients'. In: *Computing and Visualization in Science* 14.1 (2011), pp. 3–15. ISSN: 1432-9360. DOI: 10.1007/s00791-011-0160-x.

[20] A. Cohen and R. DeVore. 'Approximation of high-dimensional parametric PDEs'. In: *Acta Numerica* 24 (2015), pp. 1–159. ISSN: 0962-4929. DOI: 10.1017/S0962492915000033.

[21] A. Cohen, R. DeVore, and C. Schwab. 'Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDE's'. In: *Analysis and Applications* 9.1 (2011), pp. 11–47. ISSN: 0219-5305. DOI: 10.1142/S0219530511001728.

[22] P. R. Conrad and Y. M. Marzouk. 'Adaptive Smolyak pseudospectral approximations'. In: *SIAM Journal on Scientific Computing* 35.6 (2013), A2643–A2671. ISSN: 1064-8275. DOI: 10.1137/120890715.

[23] P. G. Constantine, M. S. Eldred, and E. T. Phipps. 'Sparse pseudospectral approximation method'. In: *Computer Methods in Applied Mechanics and Engineering* 229/232 (2012), pp. 1–12. ISSN: 0045-7825. DOI: 10.1016/j.cma.2012.03.019.

[24] F. Dalfovo, S. Giorgini, L. P. Pitaevskii, and S. Stringari. 'Theory of Bose-Einstein condensation in trapped gases'. In: *Rev. Mod. Phys.* 71 (3 Apr. 1999), pp. 463–512. DOI: 10.1103/RevModPhys.71.463.

[25] J. Dawes and M. Souza. 'A derivation of Holling's type I, II and III functional responses in predator–prey systems'. In: *Journal of Theoretical Biology* 327 (2013), pp. 11–22. ISSN: 0022-5193. DOI: 10.1016/j.jtbi.2013.02.017.

[26] S. Descombes and M. Ribot. 'Convergence of the Peaceman-Rachford approximation for reaction-diffusion systems'. In: *Numerische Mathematik* 95.3 (2003), pp. 503–525. ISSN: 0029-599X. DOI: 10.1007/s00211-002-0434-9.

[27] T. J. Dodwell, C. Ketelsen, R. Scheichl, and A. L. Teckentrup. 'A hierarchical multilevel Markov chain Monte Carlo algorithm with applications to uncertainty quantification in subsurface flow'. In: *SIAM/ASA Journal on Uncertainty Quantification* 3.1 (2015), pp. 1075–1108. DOI: 10.1137/130915005.

[28] K.-J. Engel and R. Nagel. *One-parameter semigroups for linear evolution equations*. Vol. 194. Graduate Texts in Mathematics. With contributions by S. Brendle, M. Campiti, T. Hahn, G. Metafune, G. Nickel, D. Pallara, C. Perazzoli, A. Rhandi, S. Romanelli and R. Schnaubelt. Springer-Verlag, New York, 2000. ISBN: 0-387-98463-1.

[29] O. G. Ernst, A. Mugler, H.-J. Starkloff, and E. Ullmann. 'On the convergence of generalized polynomial chaos expansions'. In: *ESAIM. Mathematical Modelling and Numerical Analysis* 46.2 (2012), pp. 317–339. ISSN: 0764-583X. DOI: 10.1051/m2an/2011045.

[30] P. Frauenfelder, C. Schwab, and R. A. Todor. 'Finite elements for elliptic problems with stochastic coefficients'. In: *Computer Methods in Applied Mechanics and Engineering* 194.2-5 (2005), pp. 205–228. ISSN: 0045-7825. DOI: 10.1016/j.cma.2004.04.008.

[31] M. R. Garvie, J. Burkardt, and J. Morgan. 'Simple Finite Element Methods for Approximating Predator–Prey Dynamics in Two Dimensions Using Matlab'. In: *Bulletin of Mathematical Biology* 77.3 (2015), pp. 548–578. ISSN: 1522-9602. DOI: 10.1007/s11538-015-0062-z.

[32] M. R. Garvie and C. Trenchea. 'Finite element approximation of spatially extended predator–prey interactions with the Holling type II functional response'. In: *Numerische Mathematik* 107.4 (2007), pp. 641–667. ISSN: 0945-3245. DOI: 10.1007/s00211-007-0106-x.

[33] L. Gauckler. 'Convergence of a split-step Hermite method for the Gross-Pitaevskii equation'. In: *IMA Journal of Numerical Analysis* 31.2 (2011), pp. 396–415. ISSN: 0272-4979. DOI: 10.1093/imanum/drp041.

[34] T. Gerstner and M. Griebel. 'Numerical integration using sparse grids'. In: *Numerical Algorithms* 18.3-4 (1998), pp. 209–232. ISSN: 1017-1398. DOI: 10.1023/A:1019129717644.

[35] R. G. Ghanem and P. D. Spanos. *Stochastic finite elements: a spectral approach*. Springer-Verlag, New York, 1991. ISBN: 0-387-97456-3. DOI: 10.1007/978-1-4612-3094-6.

[36] M. B. Giles. 'Improved multilevel Monte Carlo convergence using the Milstein scheme'. In: *Monte Carlo and quasi-Monte Carlo methods 2006*. Springer, Berlin, 2008, pp. 343–358. DOI: 10.1007/978-3-540-74496-2_20.

[37] M. B. Giles. 'Multilevel Monte Carlo methods'. In: *Acta Numerica* 24 (2015), pp. 259–328. ISSN: 0962-4929. DOI: 10.1017/S096249291500001X.

[38] M. B. Giles. 'Multilevel Monte Carlo path simulation'. In: *Operations Research* 56.3 (2008), pp. 607–617. ISSN: 0030-364X. DOI: 10.1287/opre.1070.0496.

[39] P. Glasserman. *Monte Carlo Methods in Financial Engineering*. Springer New York, 2003. ISBN: 978-0-387-21617-1. DOI: 10.1007/978-0-387-21617-1.

[40] C. González, A. Ostermann, C. Palencia, and M. Thalhammer. 'Backward Euler discretization of fully nonlinear parabolic problems'. In: *Mathematics of Computation* 71.237 (2002), pp. 125–145. ISSN: 0025-5718. DOI: 10.1090/S0025-5718-01-01330-8.

[41] C. González and C. Palencia. 'Stability of time-stepping methods for abstract time-dependent parabolic problems'. In: *SIAM Journal on Numerical Analysis* 35.3 (1998), pp. 973–989. ISSN: 0036-1429. DOI: 10.1137/S0036142995283412.

[42] M. J. Grote, S. Michel, and F. Nobile. *Uncertainty Quantification by MLMC and Local Time-stepping For Wave Propagation*. 2021. arXiv: 2106.11117 [math.NA].

[43] E. Hairer, C. Lubich, and G. Wanner. *Geometric numerical integration*. Vol. 31. Springer Series in Computational Mathematics. Structure-preserving algorithms for ordinary differential equations, Reprint of the second (2006) edition. Springer, Heidelberg, 2010. ISBN: 978-3-642-05157-9.

[44] A.-L. Haji-Ali, F. Nobile, L. Tamellini, and R. Tempone. 'Multi-index stochastic collocation convergence rates for random PDEs with parametric regularity'. In: *Foundations of Computational Mathematics* 16.6 (2016), pp. 1555–1605. ISSN: 1615-3375. DOI: 10.1007/s10208-016-9327-7.

[45] A.-L. Haji-Ali, F. Nobile, L. Tamellini, and R. Tempone. 'Multi-index stochastic collocation for random PDEs'. In: *Computer Methods in Applied Mechanics and Engineering* 306 (2016), pp. 95–122. ISSN: 0045-7825. DOI: 10.1016/j.cma.2016.03.029.

[46] M. Hanke-Bourgeois. *Grundlagen der numerischen Mathematik und des wissenschaftlichen Rechnens*. Third edition. Vieweg + Teubner, Wiesbaden, 2009. ISBN: 978-3-8348-0708-3. DOI: 10.1007/978-3-8348-9309-3.

[47] A. Hansbo. 'Nonsmooth data error estimates for damped single step methods for parabolic equations in Banach space'. In: *Calcolo. A Quarterly on Numerical Analysis and Theory of Computation* 36.2 (1999), pp. 75–101. ISSN: 0008-0624. DOI: 10.1007/s100920050024.

[48] E. Hansen and E. Henningsson. 'A convergence analysis of the Peaceman-Rachford scheme for semilinear evolution equations'. In: *SIAM Journal on Numerical Analysis* 51.4 (2013), pp. 1900–1910. ISSN: 0036-1429. DOI: 10.1137/120890570.

[49] H. Harbrecht, M. Peters, and M. Siebenmorgen. 'On multilevel quadrature for elliptic stochastic partial differential equations'. In: *Sparse grids and applications*. Vol. 88. Lect. Notes Comput. Sci. Eng. Springer, Heidelberg, 2013, pp. 161–179. DOI: 10.1007/978-3-642-31703-3.

[50] M. Hardy. 'Combinatorics of partial derivatives'. In: *Electronic Journal of Combinatorics* 13.1 (2006). DOI: 10.37236/1027.

[51] S. Heinrich. 'Multilevel Monte Carlo Methods'. In: *Large-Scale Scientific Computing*. Ed. by S. Margenov, J. Waśniewski, and P. Yalamov. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 58–67. ISBN: 978-3-540-45346-8.

[52] V. H. Hoang and C. Schwab. 'Sparse tensor Galerkin discretization of parametric and random parabolic PDEs – analytic regularity and generalized polynomial chaos approximation'. In: *SIAM Journal on Mathematical Analysis* 45.5 (2013), pp. 3050–3083. ISSN: 0036-1410. DOI: 10.1137/100793682.

[53] M. Hochbruck and A. Ostermann. 'Exponential integrators'. In: *Acta Numerica* 19 (2010), pp. 209–286. ISSN: 0962-4929. DOI: 10.1017/S0962492910000048.

[54] C. S. Holling. 'The Functional Response of Predators to Prey Density and its Role in Mimicry and Population Regulation'. In: *Memoirs of the Entomological Society of Canada* 97.S45 (1965), pp. 5–60. DOI: 10.4039/entm9745fv.

[55] E. E. Holmes, M. A. Lewis, J. E. Banks, and R. R. Veit. 'Partial Differential Equations in Ecology: Spatial Interactions and Population Dynamics'. In: *Ecology* 75.1 (1994), pp. 17–29. DOI: 10.2307/1939378.

[56] W. Hundsdorfer. 'Trapezoidal and midpoint splittings for initial-boundary value problems'. In: *Mathematics of Computation* 67.223 (1998), pp. 1047–1062. ISSN: 0025-5718. DOI: 10.1090/S0025-5718-98-00984-3.

[57] W. Hundsdorfer and J. Verwer. *Numerical solution of time-dependent advection-diffusion-reaction equations*. Vol. 33. Springer Series in Computational Mathematics. Springer-Verlag, Berlin, 2003. ISBN: 3-540-03440-4. DOI: 10.1007/978-3-662-09017-6.

[58] B. A. Ibrahimoglu. 'Lebesgue functions and Lebesgue constants in polynomial interpolation'. In: *Journal of Inequalities and Applications* (2016), Paper No. 93, 15. DOI: 10.1186/s13660-016-1030-3.

[59] T. Jahnke and C. Lubich. 'Error bounds for exponential operator splittings'. In: *BIT. Numerical Mathematics* 40.4 (2000), pp. 735–744. ISSN: 0006-3835. DOI: 10.1023/A:1022396519656.

[60] T. Jahnke, M. Mikl, and R. Schnaubelt. 'Strang splitting for a semilinear Schrödinger equation with damping and forcing'. In: *Journal of Mathematical Analysis and Applications* 455.2 (2017), pp. 1051–1071. ISSN: 0022-247X. DOI: 10.1016/j.jmaa.2017.06.004.

[61] T. Jahnke and B. Stein. 'A multi-level stochastic collocation method for Schrödinger equations with a random potential'. In: *SIAM/ASA Journal on Uncertainty Quantification (accepted, not published yet)* (2022).

[62] T. Jahnke and B. Stein. *Stochastic Galerkin-collocation splitting for PDEs with random parameters*. CRC 1173 Preprint 2018/28. URL: https://www.waves.kit.edu/downloads/CRC1173_Preprint_2018-28.pdf.

[63] J. Klaers, J. Schmitt, F. Vewinger, and M. Weitz. 'Bose-Einstein-Kondensat aus Licht'. In: *Physik in unserer Zeit* 42.2 (2011), pp. 58–59. DOI: 10.1002/piuz.201190007.

[64] A. Kunoth and C. Schwab. 'Analytic regularity and GPC approximation for control problems constrained by linear parametric elliptic and parabolic PDEs'. In: *SIAM Journal on Control and Optimization* 51.3 (2013), pp. 2442–2471. ISSN: 0363-0129. DOI: 10.1137/110847597.

[65] J. Lang, R. Scheichl, and D. Silvester. 'A fully adaptive multilevel stochastic collocation strategy for solving elliptic PDEs with random data'. In: *Journal of Computational Physics* 419 (2020). ISSN: 0021-9991. DOI: 10.1016/j.jcp.2020.109692.

[66] B. Lapeyre, É. Pardoux, and R. Sentis. *Introduction to Monte-Carlo methods for transport and diffusion equations*. Vol. 6. Oxford Texts in Applied and Engineering Mathematics. Translated from the 1998 French original by Alan Craig and Fionn Craig. Oxford University Press, 2003. ISBN: 0-19-852593-1.

[67] S. Larsson, V. Thomée, and L. B. Wahlbin. 'Finite-element methods for a strongly damped wave equation'. In: *IMA Journal of Numerical Analysis* 11.1 (1991), pp. 115–142. ISSN: 0272-4979. DOI: 10.1093/imanum/11.1.115.

[68] O. P. Le Maître and O. M. Knio. *Spectral Methods for Uncertainty Quantification*. Scientific Computation. Springer, New York, 2010. ISBN: 978-90-481-3519-6. DOI: 10.1007/978-90-481-3520-2.

[69] M. Loève. *Probability theory. II*. Fourth edition. Graduate Texts in Mathematics, Vol. 46. Springer-Verlag, New York-Heidelberg, 1978. ISBN: 0-387-90262-7.

[70] C. Lubich and A. Ostermann. 'Linearly implicit time discretization of non-linear parabolic equations'. In: *IMA Journal of Numerical Analysis* 15.4 (1995), pp. 555–583. ISSN: 0272-4979. DOI: 10.1093/imanum/15.4.555.

[71] C. Lubich and A. Ostermann. 'Runge-Kutta methods for parabolic equations and convolution quadrature'. In: *Mathematics of Computation* 60.201 (1993), pp. 105–131. ISSN: 0025-5718. DOI: 10.2307/2153158.

[72] C. Lubich. *From quantum to classical molecular dynamics: reduced models and numerical analysis*. Zurich Lectures in Advanced Mathematics. European Mathematical Society (EMS), Zürich, 2008. ISBN: 978-3-03719-067-8. DOI: 10.4171/067.

[73] C. Lubich. 'On splitting methods for Schrödinger-Poisson and cubic nonlinear Schrödinger equations'. In: *Mathematics of Computation* 77.264 (2008), pp. 2141–2153. ISSN: 0025-5718. DOI: 10.1090/S0025-5718-08-02101-7.

[74] C. Lubich and O. Nevanlinna. 'On resolvent conditions and stability estimates'. In: *BIT. Numerical Mathematics* 31.2 (1991), pp. 293–313. ISSN: 0006-3835. DOI: 10.1007/BF01931289.

[75] A. Lunardi. *Analytic semigroups and optimal regularity in parabolic problems*. Modern Birkhäuser Classics. Birkhäuser/Springer Basel AG, 1995. ISBN: 978-3-0348-0556-8.

[76] P. A. Markowich, P. Pietra, and C. Pohl. 'Numerical approximation of quadratic observables of Schrödinger-type equations in the semi-classical limit'. In: *Numerische Mathematik* 81.4 (1999), pp. 595–630.

[77] L. Mathelin, M. Y. Hussaini, and T. A. Zang. 'Stochastic approaches to uncertainty quantification in CFD simulations'. In: *Numerical Algorithms* 38.1-3 (2005), pp. 209–236. ISSN: 1017-1398. DOI: 10.1007/s11075-004-2866-z.

[78] R. I. McLachlan and G. R. W. Quispel. 'Splitting methods'. In: *Acta Numerica* 11 (2002), pp. 341–434. ISSN: 0962-4929. DOI: 10.1017/S0962492902000053.

[79] J. Mendonça and H. Terças. *Physics of Ultra-Cold Matter: Atomic Clouds, Bose-Einstein Condensates and Rydberg Plasmas*. First edition. Springer Series on Atomic, Optical, and Plasma Physics. Springer, 2013. ISBN: 9781461454137. DOI: 10.1007/978-1-4614-5413-7.

[80] M. Motamed, F. Nobile, and R. Tempone. 'A stochastic collocation method for the second order wave equation with a discontinuous random speed'. In: *Numerische Mathematik* 123.3 (2013), pp. 493–536. ISSN: 0029-599X. DOI: 10.1007/s00211-012-0493-5.

[81] M. Motamed, F. Nobile, and R. Tempone. 'Analysis and computation of the elastic wave equation with random coefficients'. In: *Computers & Mathematics with Applications. An International Journal* 70.10 (2015), pp. 2454–2473. ISSN: 0898-1221. DOI: 10.1016/j.camwa.2015.09.013.

[82] H. Niederreiter. *Random number generation and quasi-Monte Carlo methods*. Vol. 63. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1992. ISBN: 0-89871-295-5. DOI: 10.1137/1.9781611970081.

[83] F. Nobile, R. Tempone, and C. G. Webster. 'A sparse grid stochastic collocation method for partial differential equations with random input data'. In: *SIAM Journal on Numerical Analysis* 46.5 (2008), pp. 2309–2345. ISSN: 0036-1429. DOI: 10.1137/060663660.

[84] F. Nobile, R. Tempone, and C. G. Webster. 'An anisotropic sparse grid stochastic collocation method for partial differential equations with random input data'. In: *SIAM Journal on Numerical Analysis* 46.5 (2008), pp. 2411–2442. ISSN: 0036-1429. DOI: 10.1137/070680540.

[85] F. Nobile and R. Tempone. 'Analysis and implementation issues for the numerical approximation of parabolic equations with random coefficients'. In: *International Journal for Numerical Methods in Engineering* 80.6-7 (2009), pp. 979–1006. ISSN: 0029-5981. DOI: 10.1002/nme.2656.

[86] E. Novak and K. Ritter. 'High-dimensional integration of smooth functions over cubes'. In: *Numerische Mathematik* 75.1 (1996), pp. 79–97. ISSN: 0029-599X. DOI: 10.1007/s002110050231.

[87] E. Novak and K. Ritter. 'Simple cubature formulas with high polynomial exactness'. In: *Constructive Approximation. An International Journal for Approximations and Expansions* 15.4 (1999), pp. 499–522. ISSN: 0176-4276. DOI: 10.1007/s003659900119.

[88] E. Novak and K. Ritter. 'The curse of dimension and a universal method for numerical integration'. In: *Multivariate approximation and splines*. Vol. 125. Internat. Ser. Numer. Math. Birkhäuser, 1997, pp. 177–187.

[89] B. Øksendal. *Stochastic differential equations*. Fifth edition. Universitext. An introduction with applications. Springer-Verlag, Berlin, 1998. ISBN: 3-540-63720-6. DOI: 10.1007/978-3-662-03620-4.

[90] A. Okulov. 'Cold matter trapping via slowly rotating helical potential'. In: *Physics Letters A* 376.4 (2012), pp. 650–655. ISSN: 0375-9601. DOI: 10.1016/j.physleta.2011.11.033.

[91] A. Ostermann and K. Schratz. 'Error analysis of splitting methods for inhomogeneous evolution equations'. In: *Applied Numerical Mathematics. An IMACS Journal* 62.10 (2012), pp. 1436–1446. ISSN: 0168-9274. DOI: 10.1016/j.apnum.2012.06.002.

[92] A. Ostermann and M. Thalhammer. 'Convergence of Runge-Kutta methods for nonlinear parabolic equations'. In: *Applied Numerical Mathematics. An IMACS Journal* 42.1-3 (2002), pp. 367–380. ISSN: 0168-9274. DOI: 10.1016/S0168-9274(01)00161-1.

[93] C. Palencia. 'Stability of rational multistep approximations of holomorphic semigroups'. In: *Mathematics of Computation* 64.210 (1995), pp. 591–599. ISSN: 0025-5718. DOI: 10.2307/2153441.

[94] S. H. Paskov and J. F. Traub. 'Faster Valuation of Financial Derivatives'. In: *The Journal of Portfolio Management* 22.1 (1995), pp. 113–123. ISSN: 0095-4918. DOI: 10.3905/jpm.1995.409541.

[95] F. Riesz and B. Sz.-Nagy. *Functional analysis*. Dover Books on Advanced Mathematics. Translated from the second French edition by Leo F. Boron, Reprint of the 1955 original. Dover Publications, Inc., New York, 1990. ISBN: 0-486-66289-6.

[96] W. Rudin. *Real and complex analysis*. Third edition. McGraw-Hill Book Co., New York, 1987. ISBN: 0-07-054234-1.

[97] M. Růžička. *Nichtlineare Funktionalanalysis: Eine Einführung*. Springer-Verlag Berlin, 2020. ISBN: 978-3-662-62191-2. DOI: 10.1007/978-3-662-62191-2.

[98] R. Scheichl, A. M. Stuart, and A. L. Teckentrup. 'Quasi-Monte Carlo and multilevel Monte Carlo methods for computing posterior expectations in elliptic inverse problems'. In: *SIAM/ASA Journal on Uncertainty Quantification* 5.1 (2017), pp. 493–518. DOI: 10.1137/16M1061692.

[99] C. Schwab and R. A. Todor. 'Karhunen-Loève approximation of random fields by generalized fast multipole methods'. In: *Journal of Computational Physics* 217.1 (2006), pp. 100–122. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2006.01.048.

[100] *SciPy Quasi-Monte Carlo submodule*. URL: https://docs.scipy.org/doc/scipy/reference/stats.qmc.html#module-scipy.stats.qmc (visited on 02/02/2022).

[101] K. Sengstock, K. Bongs, and J. Reichel. 'Das ideale Quantenlabor: Bose-Einstein-Kondensation'. In: *Physik in unserer Zeit* 34.4 (2003), pp. 168–176. DOI: 10.1002/piuz.200301016.

[102] I. H. Sloan. 'Quasi-Monte Carlo Methods'. In: *Encyclopedia of Applied and Computational Mathematics*. Springer, 2015, pp. 1201–1203. ISBN: 978-3-540-70529-1. DOI: 10.1007/978-3-540-70529-1_391.

[103] I. H. Sloan and H. Woźniakowski. 'When are quasi-Monte Carlo algorithms efficient for high-dimensional integrals?' In: *Journal of Complexity* 14.1 (1998), pp. 1–33. ISSN: 0885-064X. DOI: 10.1006/jcom.1997.0463.

[104] S. A. Smolyak. 'Quadrature and interpolation formulas for tensor products of certain classes of functions'. In: *Dokl. Akad. Nauk SSSR* 148.5 (1963). Transl.: Soviet Math. Dokl. 4:240-243, 1963, pp. 1042–1053.

[105] M. K. Stoyanov. *User Manual: TASMANIAN Sparse Grids*. Technical report ORNL/TM-2015/596. Oak Ridge National Laboratory, 2015.

[106] M. K. Stoyanov, D. Lebrun-Grandie, J. Burkardt, and D. Munster. *Tasmanian*. Sept. 2013. DOI: 10.11578/dc.20171025.on.1087. URL: https://github.com/ORNL/Tasmanian.

[107] M. K. Stoyanov and C. G. Webster. 'A dynamically adaptive sparse grids method for quasi-optimal interpolation of multidimensional functions'. In: *Computers & Mathematics with Applications. An International Journal* 71.11 (2016), pp. 2449–2465. ISSN: 0898-1221. DOI: 10.1016/j.camwa.2015.12.045.

[108] T. J. Sullivan. *Introduction to Uncertainty Quantification*. Vol. 63. Texts in Applied Mathematics. Springer, 2015. ISBN: 978-3-319-23394-9. DOI: 10.1007/978-3-319-23395-6.

[109] A. L. Teckentrup, P. Jantsch, C. G. Webster, and M. Gunzburger. 'A multilevel stochastic collocation method for partial differential equations with random input data'. In: *SIAM/ASA Journal on Uncertainty Quantification* 3.1 (2015), pp. 1046–1074. DOI: 10.1137/140969002.

[110] M. Thalhammer. 'Convergence analysis of high-order time-splitting pseudospectral methods for nonlinear Schrödinger equations'. In: *SIAM Journal on Numerical Analysis* 50.6 (2012), pp. 3231–3258. ISSN: 0036-1429. DOI: 10.1137/120866373.

[111] R. A. Todor and C. Schwab. 'Convergence rates for sparse chaos approximations of elliptic problems with stochastic coefficients'. In: *IMA Journal of Numerical Analysis* 27.2 (2007), pp. 232–261. ISSN: 0272-4979. DOI: 10.1093/imanum/drl025.

[112] L. Wang. *Karhunen-Loeve expansions and their applications*. Thesis (Ph.D.) – London School of Economics and Political Science (United Kingdom). ProQuest LLC, Ann Arbor, MI, 2008. ISBN: 978-1321-35484-3.

[113] G. W. Wasilkowski and H. Woźniakowski. 'Explicit cost bounds of algorithms for multivariate tensor product problems'. In: *Journal of Complexity* 11.1 (1995), pp. 1–56. ISSN: 0885-064X. DOI: 10.1006/jcom.1995.1001.

[114] N. Wiener. 'The Homogeneous Chaos'. In: *American Journal of Mathematics* 60.4 (1938), pp. 897–936. ISSN: 0002-9327. DOI: 10.2307/2371268.

[115] H.-W. van Wyk. *Multilevel Sparse Grid Methods for Elliptic Partial Differential Equations with Random Coefficients*. 2014. arXiv: 1404.0963 [math.NA].

[116] D. Xiu. 'Fast numerical methods for stochastic computations: a review'. In: *Communications in Computational Physics* 5.2-4 (2009), pp. 242–272. ISSN: 1815-2406.

[117] D. Xiu and J. S. Hesthaven. 'High-order collocation methods for differential equations with random inputs'. In: *SIAM Journal on Scientific Computing* 27.3 (2005), pp. 1118–1139. ISSN: 1064-8275. DOI: 10.1137/040615201.

[118] D. Xiu and G. E. Karniadakis. 'The Wiener-Askey polynomial chaos for stochastic differential equations'. In: *SIAM Journal on Scientific Computing* 24.2 (2002), pp. 619–644. ISSN: 1064-8275. DOI: 10.1137/S1064827501387826.

[119] C. Zenger. 'Sparse grids'. In: *Parallel algorithms for partial differential equations*. Vol. 31. Notes Numer. Fluid Mech. Friedr. Vieweg, Braunschweig, 1991, pp. 241–251. ISBN: 3-528-07631-3.