Uncertainty Quantification of Elliptic Eigenvalue Problems

Dissertation

zur

Erlangung des Doktorgrades (Dr. rer. nat.)

der

Mathematisch-Naturwissenschaftlichen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

David Christoph Ebert

aus Heidelberg

Bonn 2025

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn.

Gutachter/Betreuer:	Prof. Dr. Jürgen Dölz
Gutachter:	Prof. Dr. Joscha Gedicke
Tag der Promotion:	13.03.2026
Erscheinungsjahr:	2026

Je me détourne avec effroi et horreur de cette plaie lamentable des fonctions continues qui n'ont point de dérivées.

I turn away in fright and horror from this lamentable plague of continuous functions that have no derivatives at all.

CHARLES HERMITE, *Letter to J. T. Stieltjes*, 20 May 1893

Acknowledgments

I would like to thank several people who supported me in writing this thesis.
First of all, I would like to express my gratitude to my doctoral supervisor Prof. Dr. Jürgen Dölz for the opportunity to work on this topic, his support, and scientific guidance. I would also like to thank Prof. Dr. Joscha Gedicke for co-reviewing this thesis.
Furthermore, I also would like to thank Prof. Dr. Sebastian Schöps and Dr. Anna Ziegler for the fruitful collaboration on the shape uncertainty quantification of TESLA cavities, as well as Dr. Jacopo Corno for providing data for our joint paper.
I also want to thank my current and former colleagues at the Institute for Numerical Simulation for many entertaining lunches and coffee breaks, and an all-around great atmosphere at the office. Special thanks go out to my (former) office mates Bartul Kovacic, Mikhail Kirilin, Franz Nowakowsky, and Dario Ferloni, as well as the members of my research group Bartul Kovacic, Allan Kuhn, and Konrad Böttger. I want to thank Andreas Hehl for his encouragement, as well as Allan Kuhn, Konstantinos Dimitriou, and Bartul Kovacic for proofreading parts of this thesis.
My work was partially supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)[1]. Additionally, I received support from the DFG under Germany's Excellence Strategy[2]. I also gratefully acknowledge the access to the Marvin cluster of the University of Bonn.
Lastly, I want to thank my parents, my girlfriend, and my friends for their emotional support and encouragement without which this thesis would not have been possible.

[1]**DFG project 501419255:** Datengetriebene Modellierung von elektromagnetischen Resonatoren mit unsicherer Form (data-driven modeling of electromagnetic resonators with uncertain shape).
[2]**DFG project 390685813:** EXC 2047: Hausdorff Center for Mathematics (HCM).

Abstract

This thesis considers the uncertainty quantification of elliptic eigenvalue problems (EVPs) with a special focus on degenerate eigenvalues. Elliptic EVPs are problems to find a pair of eigenvalues and eigenfunctions of an elliptic operator. We consider the stochastic moments of these eigenvalues and eigenfunctions to describe their uncertainty given a stochastic perturbation of the elliptic operator, for example, by material coefficients or shape deformations.

In a multiparametric stochastic model, the identification of eigenvalues and eigenfunctions as functions of the parameter is not trivial. Assuming analytic dependence of the elliptic operator with respect to the parameter, we first investigate the bifurcation behavior of these eigenvalue trajectories in a neighborhood of some reference parameter value. In the general degenerate case, these trajectories can only be defined with respect to the eigenspace of the eigenpairs. Assuming analyticity of the EVPs, these trajectories with respect to the eigenspace are also analytic and can thus be described by their derivatives. We characterize Fréchet derivatives of arbitrary order with respect to the eigenspace using saddle point equations. The trajectories in the traditional sense and with respect to the eigenspace are related via a pathwise-defined parameterized polarization matrix, which we also characterize including its derivatives.

Equipped with the (locally) well-defined and measurable trajectories of the eigenpairs with respect to the eigenspace, we investigate the uncertainty quantification of eigenvalues in a neighborhood of the reference point using a perturbation ansatz. We discuss the efficient implementation of this perturbation ansatz and benchmark it against other methods of calculating stochastic means and covariances of the eigenpairs like quasi-Monte Carlo. As an application of our setting, we consider stochastic shape deformation models of the Laplace and Maxwell EVP in more detail.

Lastly, we consider the possibility of incorporating measurement data into our model using a Bayesian inverse model. This leads to an adaptation of the perturbation approximations of the mean and correlation by amending terms that reflect the influence of the Radon–Nikodým derivative of the posterior measure with respect to the prior measure. We also consider the possibility of using the perturbation approximation of the posterior mean in an iteration to improve the parameter reference point. This iteration is related to the corresponding regularized inverse problem.

Contents

List of Figures

List of Tables

CHAPTER I

Introduction

Eigenvalue problems (EVPs) describe resonating states in natural sciences, from the frequency at which the string of a musical instrument can vibrate [**41, 53, 56**], to the frequency at which large structures such as skyscrapers or bridges swing due to environmental forces [**96**]. In addition, EVPs can also be used to describe the probability of the position of atomic particles [**17**], to find electromagnetic resonance frequencies to accelerate particles in TESLA cavities [**6**], or to describe the band structure of photonic crystals [**23**]. The mathematical modeling of any of these phenomena is subject to stochastic uncertainties, for example, in the form of material coefficients or shape deformations, which need to be quantified for safe and effective usage.
The above examples have in common that the solution of the EVP describes a resonating state. If a time-dependent system is excited according to a resonant frequency, it oscillates periodically over time. Then, the EVP arises by separation of variables and the solution of the EVP provides a time-harmonic solution of the original problem.

EXAMPLE I.1 ([**4**, 12.16]). The wave equation with Dirichlet boundary data is given by

$$\begin{aligned} -\Delta u(t,\mathbf{x}) + \tfrac{\mathrm{d}^2}{\mathrm{d}t^2} u(t,\mathbf{x}) &= 0 && (t,\mathbf{x}) \in (0,\infty) \times \mathscr{D} , \\ u(t,\mathbf{x}) &= 0 && (t,\mathbf{x}) \in (0,\infty) \times \partial\mathscr{D} \end{aligned}$$

with $\mathscr{D} \subset \mathbb{R}^n$, $n \in \mathbb{N}$ open and bounded and $\Delta v = \sum_{i=1}^n \frac{\mathrm{d}^2}{\mathrm{d}\mathbf{x}_i^2} v$ the Laplace operator. The Dirichlet boundary expresses that the amplitude u of the resonance vanishes on the boundary at all times, which leads to a *standing wave*, i.e., a wave whose peak does not move. A solution for this equation can be found by considering the Laplace EVP

$$\begin{aligned} -\Delta u(0,\mathbf{x}) &= u(0,\mathbf{x})\, \lambda && \mathbf{x} \in \mathscr{D} , \\ u(0,\mathbf{x}) &= 0 && \mathbf{x} \in \partial\mathscr{D} . \end{aligned}$$

Assuming without loss of generality that the wave is at its maximum amplitude for $t = 0$, the solution of the wave equation is then given by

$$u(t,\mathbf{x}) = \cos\left(\sqrt{\lambda}\, t\right)\, u(0,\mathbf{x}) .$$

Thus, the eigenvalue λ of the EVP relates to the resonance frequency $\frac{\sqrt{\lambda}}{2\pi}$ (or angular frequency $\sqrt{\lambda}$), i.e., one oscillation takes a (time) period of $\frac{2\pi}{\sqrt{\lambda}}$.

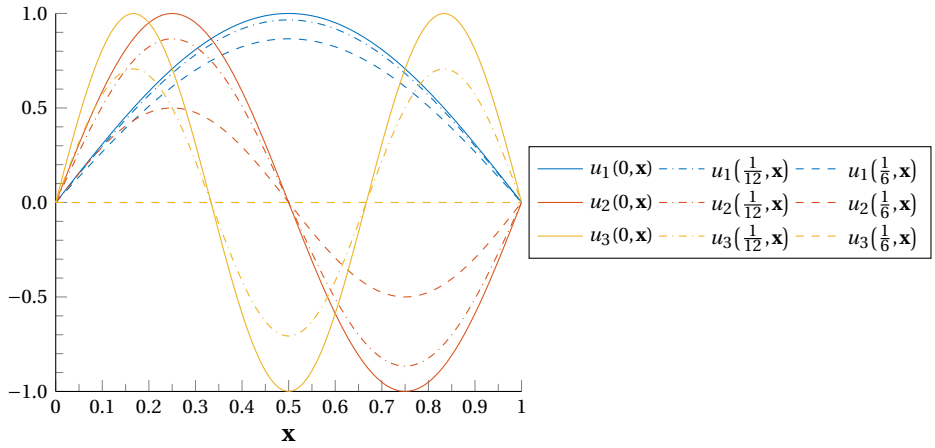

FIGURE I.1. Standing waves of example I.1 for domain $\mathscr{D} = (0,1)$ at times $t \in \{0, \frac{1}{12}, \frac{1}{6}\}$.

To make the example more specific, let $\mathscr{D} = (0,1)$, which can be thought of as an idealized model of a string, cf. [**53**], then there are infinite solutions corresponding to successively higher frequencies

$$u_n(t,\mathbf{x}) = \cos(n\pi t)\, \sin(n\pi \mathbf{x})\,, \qquad \lambda_n = (n\pi)^2\,, \qquad n \in \mathbb{N}\,.$$

The low-frequency solutions $n \in \{1,2,3\}$ are illustrated in fig. I.1 for $t \in \{0, \frac{1}{12}, \frac{1}{6}\}$. If the Laplace operator is perturbed by some parameter, so are the eigenvalues λ and eigenfunctions $u(0,\mathbf{x})$, and consequently the shape of the possible oscillations and their frequency, cf. [**59, 83**].

In addition to time-harmonic solutions, EVPs have applications in the analysis of linear operators [**4**, spectral decomposition], random fields [**62**, Karhunen–Loève expansion (KLE)], and statistical data [**100**, principal component analysis].

I.1. Setting

Let V be a Hilbert space $(V, \langle\cdot,\cdot\rangle_V)$. In its simplest form, an EVP is the problem of finding an *eigenpair* $(\lambda, u) \in \mathbb{K} \times V$ with $u \neq 0$, such that

$$Ku = u\,\lambda\,, \tag{I.1}$$

holds, where $K : V \to V$ is a linear operator, e.g., a differential operator given by a partial differential equation (PDE). In (I.1) $\lambda \in \mathbb{K}$ is called an *eigenvalue* and $u \in V$ an *eigenfunction*. Given the same operator, we can consider the related equation

$$Ku = f\,, \tag{I.2}$$

where $f \in V$ is some known function. Equation (I.2) clearly corresponds to EVP (I.1) if we set $f = u\lambda$.
In this thesis, we focus on EVPs where the operator is compact and self-adjoint, or it is elliptic and the related solution operator K^{-1} of (I.2) is compact and self-adjoint. Using the spectral theory of compact normal operators, one can show that a compact self-adjoint operator has a countable series of real eigenvalues $(\lambda_i)_{i\in\mathbb{N}}$. The subspace

$$\{u \in V \,:\, Ku = u\,\lambda\} \subset V$$

for a given eigenvalue λ is called its *eigenspace*, and the dimension m of this space is called the *multiplicity* of the eigenvalue λ. For compact normal operators, each eigenspace is finite-dimensional. If the multiplicity of the eigenvalue is larger than one, the eigenvalue is called *degenerate*, otherwise it is called *non-degenerate*. It is convenient to normalize the eigenfunctions and choose all eigenfunctions $(u_i)_{i\in\mathbb{N}}$ pairwise orthogonally with regard to the scalar product, i.e.,

$$\langle u_i, u_j\rangle_V = \delta_{ij} \coloneqq \begin{cases} 1 & i = j\,, \\ 0 & i \neq j\,, \end{cases} \qquad i, j \in \mathbb{N}\,, \tag{I.3}$$

where δ_{ij} is the Kronecker delta. If the eigenfunction u_i belongs to a non-degenerate eigenvalue, it is then uniquely determined up to the choice of sign. Otherwise, any linear combination of the eigenfunctions in the same eigenspace can be chosen.
Parameterized Eigenvalue Problems and the Identification of Eigenvalues. We consider a setting where the operator is analytically parameterized by a parameter $x \in U \subset X$, where X is a Banach space. Thus, we redefine K as an analytically parameterized operator

$$K : U \to \mathscr{L}(V)\,, \qquad x \mapsto K_x\,.$$

Then, the eigenpairs also depend on the parameter and the EVP for $x \in U$ is to find $(\lambda_x, u_x) \in \mathbb{K} \times V$ with $u_x \neq 0$, such that

$$K_x u_x = u_x\,\lambda_x\,. \tag{I.4a}$$

We again choose all eigenfunctions $((u_x)_i)_{i\in\mathbb{N}}$ pairwise orthogonally with regard to the scalar product for all $x \in U$, i.e.,

$$\langle (u_x)_i, (u_x)_j\rangle_V = \delta_{ij} \qquad i, j \in \mathbb{N}\,. \tag{I.4b}$$

Investigating the behavior of the parameterized eigenpair

$$(\lambda, u) : U \to \mathbb{K} \times V\,, \qquad x \mapsto (\lambda_x, u_x)$$

under the influence of a parameter $x \in U$, especially in the case of degenerate eigenvalues, is a central topic of this thesis. This mapping is the foundation for the later uncertainty quantification, where we consider an X-valued stochastic parameter, which consequently makes the eigenpair random.
Given the assumption of analyticity of the operator in the parameter $x \in U$, one can show that an individual eigenpair is locally analytic if the eigenvalue is isolated, i.e.,

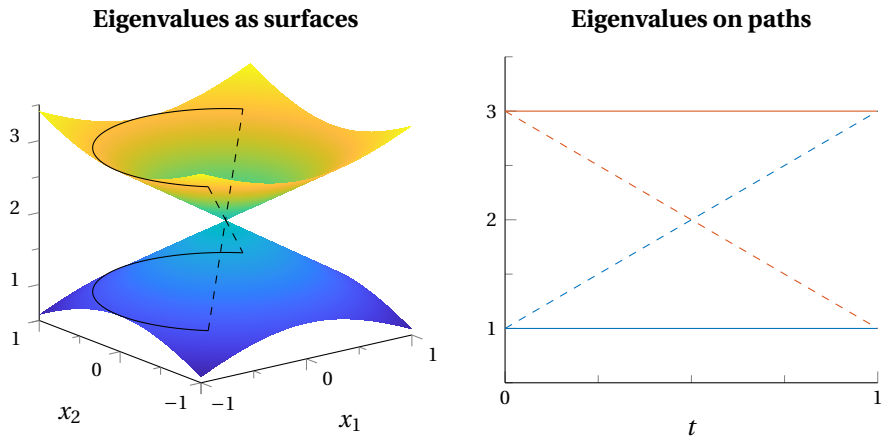

FIGURE I.2. Parameterized eigenvalues of example I.2 and suggested evolution on paths.

if it has a positive distance to the other eigenvalues and is non-degenerate, cf. [**83**]. However, this assumption is quite strong for the entire domain U. In general, the graph of eigenvalues typically shows crossings and bifurcations. It can still be proven that the eigenvalues are then continuous, cf. [**59**], but stating their regularity requires a more careful analysis. For a one-dimensional parameter space, i.e., $\dim(U) = 1$, it can be shown that the eigenpairs can be selected near and at degenerate points such that they are analytic. However, in the context of a multidimensional parameter space, this selection cannot always be generalized.

EXAMPLE I.2 ([**32**, p. 395]). We consider the eigenvalues of the parameterized normal matrix

$$K : \mathbb{R}^2 \to \mathbb{R}^{2\times 2}, \qquad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \mapsto K_x := \begin{bmatrix} 2+x_1 & -x_2 \\ -x_2 & 2-x_1 \end{bmatrix}. \tag{I.5}$$

Its eigenpairs at $x \in \mathbb{R}^2$ are given by

$$(\lambda_x)_1 = 2 - \sqrt{x_1^2 + x_2^2}, \qquad (u_x)_1 = \pm \begin{bmatrix} x_1 - \sqrt{x_1^2 + x_2^2} \\ -x_2 \end{bmatrix},$$

$$(\lambda_x)_2 = 2 + \sqrt{x_1^2 + x_2^2}, \qquad (u_x)_2 = \pm \begin{bmatrix} x_1 + \sqrt{x_1^2 + x_2^2} \\ -x_2 \end{bmatrix}.$$

and illustrated in fig. I.2. Consider two parameter paths

$$\gamma_{\text{circle}} : [0,1] \to \mathbb{R}^2 , \qquad \gamma_{\text{circle}}(t) \coloneqq \begin{bmatrix} \sin\left(\pi\left(t-\frac{3}{4}\right)\right) \\ \cos\left(\pi\left(t-\frac{3}{4}\right)\right) \end{bmatrix} , \tag{I.6a}$$

$$\gamma_{\text{direct}} : [0,1] \to \mathbb{R}^2 , \qquad \gamma_{\text{direct}}(t) \coloneqq \frac{1}{\sqrt{2}} \left(-\begin{bmatrix} 1 \\ 1 \end{bmatrix} + 2t \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right) , \tag{I.6b}$$

which coincide at their start and end, i.e.,

$$\gamma_{\text{circle}}(0) = \gamma_{\text{direct}}(0) = -\frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} , \qquad \gamma_{\text{circle}}(1) = \gamma_{\text{direct}}(1) = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} .$$

As highlighted in fig. I.2, following the eigenvalues on these paths such that their graph is differentiable leads to a different value of the eigenvalue at the final parameter point $t = 1$. From this simple example, we can already see that formulating Fréchet differentiable functions

$$\lambda_i : U \to \mathbb{R} , \qquad x \mapsto (\lambda_x)_i$$

is not possible if $(0,0) \in U \subset \mathbb{R}^2$ and $\dim(U) = 2$.

In this thesis, we consider a close substitute for the actual eigenpairs, their *trajectory with respect to the eigenspace* instead. These eigenpair trajectories with respect to the eigenspace retain the regularity of the underlying operator near the degenerate point and make uncertainty quantification of degenerate eigenpairs possible.

I.1.1. Uncertainty Quantification. The solution to EVP (I.1) is often known or can be approximated to adequate accuracy by numerical methods. Still, real-life conditions typically deviate from idealized models. In engineering, this is typically accounted for by introducing tolerances, i.e., maximum allowed deviations of the physical object compared to the ideal. These tolerances can be incorporated into the model by introducing stochastic model parameters that vary within certain bounds. Such parameters might model the degradation of materials by applying randomness to coefficients or by altering the geometry of the object to model shape uncertainty.
Given a probability space $(\Omega, \mathscr{A}, \mathbb{P})$, we let $\xi : \Omega \to X$ be a random variable for the parameter $x \in X$ and $\mathbb{P}_\xi$ its distribution. Naturally, replacing the deterministic parameter by a stochastic variable turns the eigenfunctions of EVP (I.4) into a V-valued and the eigenvalues into an $\mathbb{K}$-valued random variable. If these random variables are well behaved, their likely behavior can be described by their means

$$\mathbb{E}[\lambda] \coloneqq \int_X \lambda \, \mathrm{d}\mathbb{P}_\xi , \qquad \mathbb{E}[u] \coloneqq \int_X u \, \mathrm{d}\mathbb{P}_\xi , \tag{I.7a}$$

and the uncertainty of the solution can be expressed by their (co)variance

$$\mathbb{V}\mathrm{ar}[\lambda] \coloneqq \mathbb{E}[(\lambda - \mathbb{E}[\lambda]) \otimes (\lambda - \mathbb{E}[\lambda])] , \qquad \mathbb{C}\mathrm{ov}[u] \coloneqq \mathbb{E}[(u - \mathbb{E}[u]) \otimes (u - \mathbb{E}[u])] . \tag{I.7b}$$

We will discuss how the mean, correlation, and covariance of the eigenpair (trajectory with respect to their eigenspace) can be approximated.

In stochastic models, the distribution $\mathbb{P}_\xi$ is chosen in a way that reflects our prior knowledge of what we consider possible. Given measurements, i.e., realizations, of some measurable function $\mathsf{Q} : X \to \mathbb{R}^K$ subject to some additive stochastically independent Gaussian noise $\varepsilon \sim \mathcal{N}(0, \Sigma)$ with $\Sigma \in \mathbb{R}^{K\times K}$ its symmetric positive definite covariance matrix, we can update the distribution using Bayesian inversion. Let $\eta^\delta \in \mathbb{R}^K$ be the measurement data, then the new posterior distribution $\mathbb{P}_\xi^\delta$ is given by the Radon–Nikodým derivative

$$\frac{\mathrm{d}\mathbb{P}_\xi^\delta}{\mathrm{d}\mathbb{P}_\xi} \propto \exp(-\tfrac{1}{2}\|\eta^\delta - \mathsf{Q}(x)\|_\Sigma^2)\,, \qquad \|\cdot\|_\Sigma := \sqrt{\langle \Sigma^{-1}\cdot, \cdot\rangle_{\mathbb{R}^K}}\,.$$

This is called a *(Bayesian) inverse problem*, cf. [**93**]. In this regard, we refer to the mapping in the original direction (I.7) as the *forward problem*.

I.2. Related Work

I.2.1. Parametric Eigenvalue Problems. Seminal work on parameterized EVPs was done by Rellich in his series of articles [**78, 79, 80, 81, 82**], and later summarized in [**83**]. Rellich showed that the eigenpairs of an operator which is analytic with respect to a real parameter $t \in \mathbb{R}$ are locally analytic, but that the same does not hold in a multiparametric setting, when the eigenvalues are degenerate. Another more recent compilation of results on perturbed EVPs is given by Katō [**59**].

Later, Nelson [**68**] investigated the efficient computation of such eigenvalue trajectories, focusing on the case of non-degenerate eigenvalues. His method was refined by Dailey [**19**] and Mills-Curran [**65, 66**] for non-degenerate eigenvalues. The later articles [**32, 88**] point out that Dailey's method implicitly assumes that the derivatives of the eigenvalues are distinct ([**65, 66**] assume the same explicitly), however, they only characterize the slightly more general case where this is true for the second-order derivatives. Still, Dailey's method was later adapted to higher-order derivatives by [**55**]. To avoid these issues, some recent articles, e.g. [**42**], use non-linear characterizations instead.

More recently, the derivatives of eigenpairs have been under investigation for derivative-based tracking of eigenvalue trajectories[1], in particular for the Maxwell EVP on TESLA cavities[2] and similar resonators [**107, 108, 109**]. Derivative-free tracking algorithms were also proposed by [**3, 77**]. Moreover, [**63**] proposed a Chebyshev expansion, which also requires derivatives of the eigenpair.

[1] As in example I.2, such a tracking problem might not lead to path-independent eigenpair trajectories. Thus, either the parameter path must be deliberately chosen or an argument for path-independence must be made.

[2] For details on TeV-energy superconducting linear accelerator (TESLA) cavities see [**6**].

I.2.2. Uncertainty Quantification. The mathematical theory of uncertainty quantification has seen many advances in the last decades, primarily for PDEs, cf. [**35, 62, 90, 91, 94**] for introductory books on the matter. We provide an overview of relevant methods used in uncertainty quantification and then discuss recent advances and peculiarities for EVPs in particular.

One common approach to uncertainty quantification are sampling-based methods, such as Monte Carlo (MC), multilevel (ML)MC [**39**], quasi-Monte Carlo (QMC) [**14, 21, 70**] and Markov chain Monte Carlo (MCMC) methods [**24, 50, 52**]. Classical MC methods provide a relatively slow square-root convergence rate with respect to the number of function evaluations, independent of the dimension of the parameter space, using the law of large numbers. QMC methods try to improve the convergence rate by giving up the independence of the samples and replacing them by deterministic sequences in the parameter space that spread out more reliably than random samples. Their rate of convergence is, in general, not independent of the dimension of the parameter space. MCMC methods also abandon the independence of samples, but replace the independent samples with a Markov chain whose subsequent samples are correlated. In order to make MCMC robust with respect to the dimension of the parameter space, the challenge lies in finding an appropriate proposal kernel.

Alternatives to these sampling-based methods are quadrature-based methods, which evaluate the stochastic moments (I.7) as integrals, e.g., stochastic collocation like sparse grid methods [**12, 43, 44**] and polynomial chaos (PC) expansions [**34, 101, 102**]. PC methods use statistical methods to approximate the integrand using a polynomial surrogate, which can be evaluated using closed form solutions. Meanwhile, sparse gird methods use regularity of the integrand to reduce the number of function evaluations needed to achieve improved convergence rates.

For Bayesian inverse problems, in addition to the use of the aforementioned methods, Kálmán filters [**31, 58, 84**] are also used. These filters typically trade computational speed for accuracy, except if the forward mapping is linear and the prior distribution Gaussian.

For forward problems with relatively small scale but high-dimensional uncertainties, the perturbation method, also called local sensitivity analysis, is an alternative that is particularly popular for EVPs, cf. [**2, 7, 18, 89, 95**]. The idea here is to pick a reference point $x_0 \in X$ in the parameter space and apply a Taylor expansion of the variable of interest. For EVPs, the question whether Fréchet derivatives exist and how they can be characterized is therefore essential. Throughout this dissertation, our primary approximation method for the moments (I.7) is the perturbation method, since eigenpair trajectories are generally only locally well defined as functions.

With few exceptions, e.g. [**33, 43, 44, 95**] (and the articles [**25, 27**], which were created in preparation for this thesis), most articles consider only non-degenerate eigenvalues. Without this degeneracy, the eigenvalue and eigenfunctions are well defined as functions of the parameter $x \in X$, so that both sampling-based [**18, 36, 37, 38**] and

quadrature-based methods [**5, 43, 44, 33**] have been successfully implemented. In this thesis, we consider degenerate eigenpair trajectories with respect to the eigenspace using a perturbation ansatz.

I.3. Contributions

This dissertation relates to the recently published articles [**25, 27**] and the preprint article [**26**], providing context and further elaborations. Our special focus is to include degenerate eigenvalues in the quantification of uncertainties. Therefore, we discuss their behavior around points in the parameter space where an eigenvalue becomes degenerate in detail. In the degenerate case, we mostly consider their trajectories with respect to the eigenspace, as other forms are generally not available. In [**25**] a linear characterization of the derivatives of eigenpairs is presented for trajectories in the classical sense and for trajectories with respect to the eigenspace. An implementation of this characterization was recently published in the repository [**30**] to accompany this thesis. Given the linear characterization of the derivatives, efficient perturbation approximations of mean and covariance are also proposed in [**25**]. This framework was applied in [**27**] for uncertainty quantification of the modes of a TESLA cavity under shape uncertainty. In [**26**] the application of a general perturbation framework to Bayesian inversion was investigated for PDEs. Here, we present corresponding results in the context of EVPs.
The results can be summarized as

- characterization of eigenpair derivatives with respect to the eigenspace as a linear problem, as well as characterization of the relationship between trajectories with respect to the eigenspace and otherwise via a parametrized polarization,
- perturbation-based uncertainty quantification of eigenpairs and discussion of stochastic moments of eigenpairs in the presence of degenerate eigenvalues,
- expansion of the perturbation approximation framework to degenerate eigenspaces and adaptation to measurement data within Bayesian inverse problems.

I.4. Outline

The structure of this dissertation is as follows.

II. Preliminaries. We briefly introduce the definitions and theorems underlying the results developed throughout this thesis. This includes notations, Hilbert and Banach spaces, as well as various concepts related to linear operators, which are necessary for the formulation of EVPs. We also discuss derivatives for Banach space-valued functions, which are instrumental in our definition of well-defined eigenpairs trajectories, and measure theory as necessary for the variational formulation of EVPs. Next, we recall the spectral theory for compact normal operators and the variational formulation of EVPs. As examples, we state the Laplace and Maxwell EVP. Lastly, we recapitulate some aspects of probability theory.

III. Trajectories and Derivatives of Eigenpairs. This chapter is based on [**25**] though some results are expanded upon and presented in more detail. We prove the existence of trajectories of eigenpairs with respect to the eigenspace, given that the involved operators are sufficiently regular. Subsequently, we characterize the derivatives of eigenvalues and eigenfunctions with respect to the eigenspace as the solution of saddle point equations. We discuss how the eigenpair trajectories with respect to the eigenspace relate to the eigenvalues in the traditional sense via a parameterized polarization matrix. The insight into the derivatives can then be used to analyze crossings and bifurcations of eigenvalues to decide if and how they can be expressed as functions on a parameter space. We present some examples to illustrate possible bifurcation behavior. Lastly, we discuss the efficient implementation of the eigenvalue derivatives as saddle point equations and equivalent formulations. Numerical examples are presented to demonstrate the validity of the Taylor approximations suggested by the derivatives.

IV. Uncertainty Quantification. In this chapters we continue with the stochastic aspect of [**25**]. We consider uncertainty quantification of the eigenpairs with degenerate eigenvalues by considering a stochastic parameter for the parametric EVP of chapter III. We follow a perturbation ansatz which leads to second-order approximations of the mean, correlation, and covariance. The approximations have asymptotic convergence of $\mathcal{O}(t^3)$ with $t \geq 0$ being the amplitude of the perturbation. Given stronger assumptions the convergence can be improved to $\mathcal{O}(t^4)$. We also discuss the efficient implementation of the approximations given a decomposition of the random variable of KLE-type. Finally, we apply the approximation to the perturbed Laplace EVP discussed in chapter III.

V. Shape Uncertainty Quantification. This chapters is based on results of [**27**], applying the perturbation approximation of eigenpairs in the context of shape deformations. We build on the results of chapters III and IV by demonstrating how shape deformations can be transferred to the previous setting using matrix-valued deformation coefficients. To this end, we formulate these coefficients and their derivatives for the Laplace and Maxwell EVP and conclude the chapter with numerical examples for both.

VI. Bayesian Inverse Problems. Based on results of [**26**], we investigate a perturbational approach to Bayesian inverse problems. We interpret the results in the context of EVPs, thus chapters III and IV are prerequisites to define the forward model. The perturbation approximations of chapter IV are augmented by additional terms to approximate the posterior mean, correlation, and covariance of the eigenpairs with respect to the eigenspace. Additionally, we discuss how the perturbation approximation of the posterior mean can be used to motivate an iteration on the parameter space and point out a relationship to the corresponding regularized inverse problem. Finally, we provide numerical examples

applying the approximations for the Bayesian inverse problem and the iteration to EVPs.

VII. Conclusion and Outlook. Here, we summarize the results of this thesis, draw a conclusion and highlight potential leads for future research.

CHAPTER II

Preliminaries

This chapter recalls the preliminaries needed for the development of perturbation-based uncertainty quantification of elliptic EVP. We start by fixing some notations and recapitulate some definitions and properties that are used for the formulation of EVPs. The formulation of the perturbation approach requires (Fréchet) differentiability, therefore we will recall both Gâteaux and Fréchet differentiability and state Taylor's theorem, which leads to the definition of analyticity, as well as the implicit function theorem. In order to introduce a stochastic model, we recall some definitions of measure theory, leading up to Sobolev spaces and their embeddings. Then, we introduce EVPs and recall the spectral theorem of compact normal operators. We recall how the spectral theorem can also be used to derive the spectral properties of variational EVPs. Next, we discuss the EVPs that we use as examples in this thesis, i.e., the Laplace and Maxwell EVP. The Maxwell EVP requires the definition of some additional function spaces. Lastly, we dedicate a section to probability theory.

II.1. Notations, Conventions, and Vector Spaces

We denote the set of natural numbers by $\mathbb{N}$, set $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$, and let $\mathbb{K}$ be real numbers $\mathbb{R}$ or complex numbers $\mathbb{C}$. Complex numbers $x \in \mathbb{C}$ are decomposed into **real** and **imaginary part**

$$x = \Re(x) + \mathrm{i}\,\Im(x)\,.$$

Let $(u_i)_{i=1,\dots,m} \in V$ be an element of a vector space, then we use bold notation to denote a vector with entries

$$\boldsymbol{u} = [u_1, \dots, u_m] \in \bigtimes_{i=1}^{m} V =: V^m\,.$$

The space V^m is called a **product space**, cf. [**60**, Definition 14.1]. Let u be a vector, then we denote the **transposed** vector by $u^\top$, the **conjugate** vector by $\overline{u}$, and the **adjoint** vector by u^*, with the same notation also applied to matrices $K \in \mathbb{K}^{n\times m}$, i.e.,

$$[K^\top]_{ji} := [K]_{ij} \quad i = 1,\dots,n, \quad j = 1,\dots,m\,, \qquad \overline{K} := \Re(K) - \mathrm{i}\,\Im(K)\,, \qquad K^* := \overline{K}^\top\,.$$

For an invertible matrix $K \in \mathbb{K}^{n\times n}$ we also write $K^{-\top} := (K^{-1})^\top = (K^\top)^{-1}$. The **identity matrix** is denoted by $\boldsymbol{I} := \operatorname{diag}(1,\dots,1) \in \mathbb{R}^{n\times n}$ and the **identity operator** between two function spaces $X \subset Y$ as $\mathrm{Id} : X \to Y, x \mapsto x$.

II.1.1. Banach Spaces and Topological Properties. It is assumed that the reader is familiar with the Banach and Hilbert space as well as their commonly used topological properties, cf. [**4**, Chapter 2]. However, we still recall the definitions, which are prominently used for spectral theory and embeddings in the context of variational EVPs to clarify the notation.

DEFINITION II.1 ([**4**, 2.4 & 2.22]). Let X be a $\mathbb{K}$-vector space. The pair $(X, \|\cdot\|_X)$ is called a **normed space** if $\|\cdot\|_X : X \to \mathbb{R}$ satisfies the following conditions for $x, y \in X$ and $\alpha \in \mathbb{K}$:

(1) $\|x\|_X \geq 0$ and $\|x\|_X = 0 \iff x = 0$,
(2) $\|\alpha x\|_X = |\alpha| \|x\|_X$,
(3) $\|x + y\|_X \leq \|x\|_X + \|y\|_X$.

We then call $\|\cdot\|_X$ a **norm** on X.

We call $\|x_1 - x_2\|_X$ the **distance** of two points $x_1, x_2 \in X$ and fix the notation

$$B_r(x) := \left\{ y \in X \,:\, \|y - x\|_X < r \right\}$$

for an (open) **ball** with center $x \in X$ and radius $r > 0$. Keeping the radius unspecified, i.e., for some radius small enough, we may refer to it a as a **neighborhood** $B(x) \subset X$. For a set A, we denote

(1) the **interior** by $\operatorname{intr}(A) := \{ x \in X \,:\, B_\epsilon(x) \subset A, \epsilon > 0 \} \subset A$,
(2) the **closure** by $\overline{A} := \operatorname{clos}(A) := \{ x \in X \,:\, B_\epsilon(x) \cap A \neq 0, \forall \epsilon > 0 \} \supset A$, and
(3) the **boundary** by $\partial A := \operatorname{clos}(A) \setminus \operatorname{intr}(A)$.

A set with $A = \operatorname{intr}(A)$ is called **open** and a set $A = \overline{A}$ is **closed**. A **domain** is an open set $\mathscr{D} \subset \mathbb{R}^n$ with $n \in \mathbb{N}$.

DEFINITION II.2 ([**4**, 2.13]). Let $(X, \|\cdot\|_X)$ be a normed space.

(1) A subset $A \subset X$ is called **dense** in X, if $\overline{A} = X$,
(2) X is called **separable** if X contains a countable dense subset.

DEFINITION II.3 ([**4**, 2.21]). Let $(X, \|\cdot\|_X)$ be a normed space.

(1) A sequence $(x_k)_{k\in\mathbb{N}}$ in X is called a **Cauchy sequence** if for each $\epsilon > 0$ exists an $N \in \mathbb{N}$, such that for every $k, \ell \geq N$ holds $\|x_k - x_\ell\|_X < \epsilon$.
(2) If $(x_k)_{k\in\mathbb{N}}$ is a sequence in X, then a point $x \in X$ is called a **cluster point** of this sequence if there exists a subsequence $(x_{k_i})_{i\in\mathbb{N}}$ such that $x = \lim_{i\to\infty} x_{k_i}$.
(3) The normed space $(X, \|\cdot\|_X)$ is called **complete** if every Cauchy sequence in X has a cluster point in X. Since every Cauchy sequence can have at most one cluster point, this means that every Cauchy sequence in X has a limit in X.

DEFINITION II.4. A complete normed space is called a **Banach space**.

Recall that for every normed space, there exists a **completion** $(\overline{X}, \|\cdot\|_{\overline{X}})$, cf. [**4**, 2.24], with $\overline{X}$ defined by adding the missing limits of the Cauchy sequence to X and $\|\cdot\|_{\overline{X}}$ such that

$$\|x_1 - x_2\|_{\overline{X}} = \|x_1 - x_2\|_X \qquad \forall\, x_1, x_2 \in X\,.$$

We may indicate the norm, i.e., $\overline{X}^{\|\cdot\|}$, to stress with which norm the completion is performed.

II.1.2. Hilbert Spaces. Hilbert spaces provide the possibility of defining an angle of two of its elements, which allows us to define orthogonality. This is important for the investigation of eigenfunctions. The operator whose spectral properties determine our EVPs, will map from and to Hilbert spaces.

DEFINITION II.5 ([**4**, 2.1 & 2.22]). Let X be a $\mathbb{K}$-vector space.

(1) We call a map $(x_1, x_2) \mapsto \langle x_1, x_2\rangle_X$ from $X \times X$ to $\mathbb{K}$ a **sesquilinear form** if for all $\alpha \in \mathbb{K}$ and for all $x, x_1, x_2, y, y_1, y_2 \in X$ one has
   (a) $\langle \alpha x, y\rangle_X = \alpha\langle x, y\rangle_X$,
   $\langle x, \alpha y\rangle_X = \overline{\alpha}\langle x, y\rangle_X$,
   (b) $\langle x_1 + x_2, y\rangle_X = \langle x_1, y\rangle_X + \langle x_2, y\rangle_X$,
   $\langle x, y_1 + y_2\rangle_X = \langle x, y_1\rangle_X + \langle x, y_2\rangle_X$.
   This means $\langle\cdot,\cdot\rangle_X$ is **linear** in the first argument and **conjugate linear** in the second argument. The sesquilinear form is called
   (a) a **Hermitian form** if $\langle x, y\rangle_X = \overline{\langle y, x\rangle_X}$ for all $x, y \in X$,
   (b) **positive semidefinite** if $\langle x, x\rangle_X \geq 0$, $\langle x, x\rangle_X \in \mathbb{R}$ for all $x \in X$ and **positive definite** if additionally $\langle x, x\rangle_X = 0 \iff x = 0$ for all $x \in X$.
(2) A positive definite Hermitian form is called a **scalar product** or **inner product**.
(3) The pair $(X; \langle\cdot,\cdot\rangle_X)$ is called an **inner product space**.
(4) Every inner product space is also a normed space, when equipped with the **induced norm**

$$\|x\|_X := \sqrt{\langle x, x\rangle_X}\,. \tag{II.1}$$

(5) If an inner product space is complete via (II.1) we call it a **Hilbert space**.

From $\langle x, x\rangle_X = \overline{\langle x, x\rangle_X}$ follows $\langle x, x\rangle_X \in \mathbb{R}$. For $\mathbb{K} = \mathbb{R}$, the sesquilinear form becomes a **bilinear form** and a Hermitian form becomes **symmetric**, without the conjugation, respectively. Some authors also call Hermitian forms symmetric.
For an inner product spaces X, in addition to the inequalities of normed spaces, the **Cauchy–Schwarz inequality** holds, cf. [**4**, Lemma 2.2], $|\langle x, y\rangle_X| \leq \|x\|_X \|y\|_X$ for all $x, y \in X$.
Orthogonality. We recall orthogonality, which is used to formulate variational EVPs.

DEFINITION II.6 ([**4**, 2.3]). Let X be a inner product space and let $\|\cdot\|_X$ be the induced norm (II.1).

(1) Let $x, y \in X$. If $\langle x, y\rangle_X = 0$, we say that x and y are **orthogonal**. Then **Pythagoras' theorem** $\|x-y\|_X^2 = \|x\|_X^2 + \|y\|_X^2$, holds.
(2) If X is an inner product space, then two subspaces $Y, Z \subset X$ are called **orthogonal** if $\langle y, z\rangle_X = 0$ for all $y \in Y$ and $z \in Z$. Then $Y \cap Z = \{0\}$ holds and we write $Y \perp Z$.
(3) The **orthogonal complement** of a subspace Y is defined by
$$Y^\perp := \left\{ x \in X \,:\, \langle y, x\rangle_X = 0 \quad \forall\, y \in Y \right\}$$
It holds that $Y \cap Y^\perp = \{0\}$.
(4) If Y and Z are two subspaces of a vector space X, then
$$Y + Z := \left\{ y + z \,:\, y \in Y \text{ and } z \in Z \right\}$$
is again a subspace. If Y and Z are orthogonal, it is called the **direct sum** with notation $Y \oplus Z$.

DEFINITION II.7 ([**4**, 9.5 & 9.7]). Let X be an inner product space.

(1) A sequence $(e_k)_{k\in N}, N \in \mathbb{N}$, in X is called an **orthogonal system** if
   (a) $\langle e_k, e_l\rangle_X = 0$ for all $k \neq l$, and
   (b) $e_k \neq 0$ for all $k \in \mathbb{N}$.
   It is called an **orthonormal system** if $\langle e_k, e_l\rangle_X = \delta_{k,l}$ for all $k, l \in \mathbb{N}$.
(2) Let $(e_k)_{k\in\mathbb{N}}$ be an orthonormal system in a inner product space X. $(e_k)_{k\in\mathbb{N}}$ is called an **orthonormal basis** if one of the following equivalent conditions is satisfied:
   (a) $\operatorname{span}(\{ e_k \,:\, k \in \mathbb{N}\})$ is dense in X.
   (b) Every vector $x \in X$ can be represented as $x = \sum_{k=1}^\infty \langle x, e_k\rangle\, e_k$.
   (c) **Parseval's identity** holds:
   $$\langle x, y\rangle_X = \sum_{k=1}^\infty \langle x, e_k\rangle_X \overline{\langle y, e_k\rangle_X} \qquad \forall\, x, y \in X\,.$$
   (d) The **completeness relation** holds:
   $$\|x\|_X^2 = \sum_{k=1}^\infty |\langle x, e_k\rangle_X|^2 \qquad \forall\, x \in X\,.$$

LEMMA II.8 ([**4**, Theorem 9.8]). *For every infinite-dimensional Hilbert space X over $\mathbb{K}$ the following are equivalent:*

(1) *X is separable.*
(2) *X has an orthonormal basis.*

Tensor Products on Hilbert spaces. We introduce tensor products of Hilbert spaces following [**94**, Chapter 3.5] building on equivalence relations defined by quotient spaces, cf. [**4**, 2.4].

DEFINITION II.9 ([**94**, Definition 3.28 & 3.29]). Let X, Y be two Hilbert spaces over a common field $\mathbb{K}$.

(1) The **free vector space** $F_{X\times Y}$ on the Cartesian product $X \times Y$ is defined by

$$F_{X\times Y} := \left\{ \sum_{i=1}^{n} \alpha_i e_{(x_i,y_i)} \,:\, n\in\mathbb{N};\quad \alpha_i \in \mathbb{K},\quad (x_i,y_i)\in X\times Y \quad \forall\, i=1,\dots,n. \right\}$$

(2) Let Z be the subspace of $F_{X\times Y}$ that is generated by the equivalence relation

$$\begin{aligned} e_{(x_1+x_2,y)} &\sim e_{(x_1,y)} + e_{(x_2,y)}\,, \\ e_{(x,y_1+y_2)} &\sim e_{(x,y_1)} + e_{(x,y_2)}\,, \\ \alpha e_{(x,y)} &\sim e_{(\alpha x,y)} \sim e_{(x,\alpha y)}\,. \end{aligned}$$

Then the **(algebraic) tensor product** $X \otimes Y$ is the quotient space

$$X\otimes Y := \frac{F_{X\times Y}}{Z}\,.$$

The following equalities hold in the tensor product space

$$\begin{aligned} (x_1+x_2)\otimes y &= x_1\otimes y + x_2\otimes y\,, \\ x\otimes(y_1+y_2) &= x\otimes y_1 + x\otimes y_2\,, \\ \alpha(x\otimes y) &= (\alpha x)\otimes y = x\otimes(\alpha y) \end{aligned}$$

for all $x, x_1, x_2 \in X$, $y, y_1, y_2 \in Y$, and $\alpha \in \mathbb{K}$.
For vectors and matrices in the Hilbert spaces $\mathbb{R}^n, n \in \mathbb{N}$ the tensor product relates to the **Kronecker product** which we also denote by the symbol $\otimes$. We avoid tensor products in Banach spaces, which are more complicated, cf. [**54**].
Now we can define the Hilbert space tensor product of two Hilbert spaces as a completion.

DEFINITION II.10 ([**94**, Definition 3.30]). The **Hilbert space tensor product** of two Hilbert spaces X and Y over the same field $\mathbb{K}$ is given by defining an inner product on the algebraic tensor product $X \otimes Y$ by

$$\langle x_1\otimes y_1, x_2\otimes y_2\rangle_{X\otimes Y} := \langle x_1,x_2\rangle_X \langle y_1,y_2\rangle_Y \qquad \forall\, x_1,x_2\in X,\quad y_1,y_2\in Y\,,$$

extending this definition to all of the algebraic tensor product by sesquilinearity, and defining the Hilbert space tensor product $X \otimes Y$ to be the completion of the algebraic tensor product with respect to this inner product and its induced norm.

From now on, we use the notation $X \otimes Y$ only for the (complete) Hilbert space tensor product.

II.2. Operators

We now recall definitions concerning linear operators, especially projections and embeddings, which are used for spectral theory. In preparation for the following chapter on derivatives, we also introduce n-linear operators. Since for the spectral theorems

we want to rely on are formulated for compact (normal) operators, we also recall definitions and results for compact operators, which are a subset of bounded linear operators. Lastly, we recall results on the (bi-)dual space. This leads to Gelfand triples, which we use to formulate variational EVPs, and normal operators, which are again the subject of the spectral theorems presented later.

II.2.1. Linear Operators. We start with bounded linear operators, which are required for derivatives and spectral theory.

DEFINITION II.11. Let X, Y be normed vector spaces over $\mathbb{K}$.

(1) An operator $T : X \to Y$ is called **linear** if

$$T(\alpha x) = \alpha T(x) \qquad \forall \alpha \in \mathbb{K}, x \in X ,$$
$$T(x_1 + x_2) = T(x_1) + T(x_2) \qquad \forall x_1, x_2 \in X .$$

(2) An operator $T : X \to Y$ is called **bounded** if there exists a $C > 0$, such that

$$\|Tx\|_Y \leq C\|x\|_X \qquad \forall x \in X .$$

(3) An operator $T : X \to Y$ is called **continuous** at $x_1 \in X$ if

$$\lim_{x_2 \to x_1} \|T(x_1) - T(x_2)\|_Y = 0$$

and **continuous** if it is continuous at x_1 for all $x_1 \in X$.

A linear operator is continuous if and only if it is bounded, cf. [**4**, Lemma 5.1].

DEFINITION II.12 ([**4**, 5.2 & 5.5 (4)]). Let X, Y be normed vector spaces. We define the function space

$$\mathscr{L}(X;Y) := \{ T : X \to Y : T \text{ is linear and bounded} \}$$

as **bounded linear operators** or **continuous linear operators**. It forms a normed space with the **operator norm**

$$\|T\|_{\mathscr{L}(X;Y)} := \sup_{\|x\|_X \leq 1} \|Tx\|_Y . \tag{II.2}$$

We use shorthand notation $\mathscr{L}(X) := \mathscr{L}(X;X)$. For $T \in \mathscr{L}(X;Y)$ we denote by

$$\mathscr{N}(T) := \{ x \in X \,:\, Tx = 0 \}$$

the **null space** of T. The continuity of T yields that $\mathscr{N}(T)$ is a closed subspace. The **range** or **image** of T is defined by

$$\mathscr{R}(T) := \{ Tx \in Y \,:\, x \in X \} .$$

The subspace $\mathscr{R}(T)$ is in general not closed. We also use the notation $T(X) := \mathscr{R}(T)$.

$\mathscr{L}(X;Y)$ is a Banach space if Y is a Banach space, cf. [**4**, Theorem 5.3].

Linear Projections.

DEFINITION II.13 ([**4**, 5.5 (3)]). Let X, Y be normed vector spaces. An linear operator $P \in \mathscr{L}(X)$ is called a **(linear) projection** if $P^2 = P \circ P = P$. We denote the set of **continuous (linear) projections** by

$$\mathscr{P}(X) := \left\{ P \in \mathscr{L}(X) \,:\, P^2 = P \right\} .$$

The following theorem refines projections to orthogonal projections.

THEOREM II.14 (**Projection theorem**, [**4**, 4.3]). *Let X be a Hilbert space, and let $A \subset X$ be nonempty, closed, and convex. Then there exists a unique map $P : X \to A$ such that*

$$\|x - P(x)\|_X = \inf_{y \in A} \|x - y\|_X \qquad \forall x \in X .$$

For $x \in X$ an equivalent characterization of $P(x) \in A$ is given by

$$\Re\langle x - P(x), a - P(x)\rangle_X \le 0 \qquad \forall a \in A .$$

*The map $P : X \to A$ is called the **orthogonal projection** from X to A.*

Linear Embeddings.

DEFINITION II.15 ([**4**, 5.5 (5,6, & 7)]).

(1) Let X, Y be normed spaces. $T \in \mathscr{L}(X; Y)$ is called a **(continuous, linear) embedding** of X into Y if T is injective, i.e., if $\mathscr{N}(T) = \{0\}$. We use the notation $X \hookrightarrow Y$ to write that X is embedded in Y and call it **dense** if $T(X)$ is dense in Y.
(2) Let X and Y be Banach spaces. If $T \in \mathscr{L}(X; Y)$ is bijective, then $T^{-1} \in \mathscr{L}(Y; X)$, cf. *inverse mapping theorem* [**4**, theorem 7.8]. Then T is called an **invertible (linear) operator** or a **(continuous, linear) isomorphism**.
(3) Let X and Y be normed spaces. $T \in \mathscr{L}(X; Y)$ is called an **isometry** if $\|Tx\|_Y = \|x\|_X$ for all $x \in X$.
(4) Banach spaces X and Y are called **isometrically isomorph** if there is a mapping $T \in \mathscr{L}(X; Y)$, which is an isometry and an isomorphism. We then write $X \cong Y$.

The identity embedding $\mathrm{Id} : X \hookrightarrow \overline{X}$ of a normed vector space X to its completion is an example of a dense and isometric embedding, cf. [**4**, 2.24].

II.2.2. n-linear Operators.

DEFINITION II.16 ([**106**, Definition 4.15]). Let $X_i, i \in \mathbb{N}, Y$ be Banach spaces. The mapping

$$T : X_1 \times X_2 \times \cdots \times X_n \to Y ,$$

is called

(1) n-**linear** if and only if T is linear in each argument
(2) and **bounded** if there is a fixed $C \ge 0$, such that

$$\|T(x_1, x_2, \ldots, x_n)\|_Y \le C \|x_1\|_{X_1} \cdots \|x_n\|_{X_n} \qquad \forall x_i \in X_i, i = 1, \ldots, n . \tag{II.3a}$$

We denote the vector spaces of **bounded n-linear functions** as

$$\mathscr{L}^{(n)}(X_1 \times \cdots \times X_n; Y) \coloneqq \left\{ f : X_1 \times \cdots \times X_n \to Y \ : \ f \text{ is bounded } n\text{-linear.} \right\}$$

with shorthand $\mathscr{L}^{(n)}(X;Y) \coloneqq \mathscr{L}^{(n)}(X \times \cdots \times X; Y)$. The **operator norm** for T generalizes to

$$\| \cdot \|_{\mathscr{L}^{(n)}(X;Y)} \coloneqq \sup_{\|x_1\|_{X_1} = \ldots = \|x_n\|_{X_n} = 1} \| T(x_1, \ldots, x_n) \|_Y \tag{II.3b}$$

and $(\mathscr{L}^{(n)}(X;Y), \| \cdot \|_{\mathscr{L}^{(n)}(X;Y)})$ is a normed vector space.

The space of n-linear operators as well as n-times nested operators is isometrically isomorph, i.e.,

$$\underbrace{\mathscr{L}(X; \ldots \mathscr{L}(X;Y))}_{n \text{ times}} \cong \mathscr{L}^{(n)}(X;Y) \tag{II.4a}$$

via the isomorphism

$$(T x_n) \ldots x_1 \mapsto T(x_n, \ldots, x_1) \, . \tag{II.4b}$$

The result that $\mathscr{L}(X;Y)$ is a Banach space if Y is a Banach space (definition II.12) extends to the nested spaces via induction and to the n-linear spaces via isomorphism. We can also extend this notation to bounded sesquilinear forms.

DEFINITION II.17. Let X be a normed vector space. We define boundedness of sesquilinear forms as in (II.3a) and the space of **bounded sesquilinear forms**

$$\mathscr{L}^{(1.5)}(X;Y) \coloneqq \left\{ T : X \times X \to Y \ : \ T \text{ is bounded sesquilinear.} \right\}$$

Using the convention of definition II.5, the second argument is conjugate linear. The operator norm for the sesquilinear forms $\mathscr{L}^{(1.5)}(X;\mathbb{K})$ is given by (II.3).

II.2.3. Compact Operators.

DEFINITION II.18 ([**4**, 4.6 & 5.5 (2)]). Let X, Y be normed vector spaces.

(1) A subset $A \subset X$ is called **compact** if every sequence in A contains a convergent subsequence with limit in A.
(2) The set of **compact (linear) operators** from X to Y is defined by

$$\mathscr{K}(X;Y) \coloneqq \left\{ T \in \mathscr{L}(X;Y) \ : \ \overline{T(B_1(0))} \text{ is compact} \right\} .$$

Compact operators have the following useful properties.

LEMMA II.19 ([**4**, Lemma 10.2]). *Let X, Y be Banach spaces.*

(1) *$\mathscr{K}(X;Y)$ is a closed subspace of $\mathscr{L}(X;Y)$.*
(2) *If $T \in \mathscr{L}(X;Y)$ with $\dim(\mathscr{R}(T)) < \infty$, then $T \in \mathscr{K}(X;Y)$.*
(3) *If Y is a Hilbert space and $T \in \mathscr{L}(X;Y)$, then*

$$T \in \mathscr{K}(X;Y) \iff \begin{array}{l} \text{there exist } T_n \in \mathscr{L}(X;Y) \text{ with } \dim(\mathscr{R}(T_n)) < \infty, \\ \text{such that } \|T - T_n\| \to 0 \text{ as } n \to \infty. \end{array}$$

The following lemma characterizes compositions in which one of the mappings is compact.

LEMMA II.20 ([**4**, Lemma 10.3]). *Let X be a Banach space. For $T_1 \in \mathscr{L}(X;Y)$ and $T_2 \in \mathscr{L}(Y;Z)$ it holds that:*

$$T_1 \text{ or } T_2 \text{ are compact.} \qquad \Longrightarrow \qquad T_2 \circ T_1 \text{ is compact.}$$

The following result is also useful for the solution of systems of equations.

THEOREM II.21 (**Fredholm alternative**, [**4**, 11.11]). *Let X be a Banach space. If $T \in \mathscr{K}(X)$ and $\lambda \neq 0$, then it holds either*

- *the equation $Tx - \lambda x = y$ is uniquely solvable for every $y \in X$, or*
- *the equation $Tx - \lambda x = 0$ has nontrivial solutions.*

II.2.4. Dual Space. The dual space is required to define normal operators.

DEFINITION II.22 ([**4**, 5.5 (1)]). Let X be a normed vector space.

(1) The space $X' := \mathscr{L}(X;\mathbb{K})$ is the **dual space** to X. The elements of X' area also called **linear functionals**. If X is a normed space the norm (II.2) for $T \in X'$ becomes

$$\|T\|_{X'} := \sup_{\|x\|_X \leq 1} |Tx|. \tag{II.5}$$

(2) If $T \in \mathscr{L}(X;Y)$, then

$$(T'y')(x) := y'(Tx) \qquad \text{for } y' \in Y', x \in X$$

defines a map $T' \in \mathscr{L}(Y';X')$ (cf. proof in [**4**, 5.5]), the **adjoint map** of T. We also call T' the **adjoint operator** of T (also **dual operator** or **adjoint**).

The following result makes the relationship between a normed space and its dual clearer.

LEMMA II.23 ([**4**, 12.1]). *Let X, Y be normed spaces. Then the adjoint map*

$$(T'y')(x) := y'(Tx) \qquad \text{for } x \in X, y' \in Y'$$

defines an isometric embedding $T \mapsto T'$ from $\mathscr{L}(X;Y)$ to $\mathscr{L}(Y';X')$.

Compact operators between Banach spaces have a compact dual counterpart.

THEOREM II.24 (**Schauder's theorem**, [**4**, 12.6]). *Let X and Y be Banach spaces and $T \in \mathscr{L}(X;Y)$. Then*

$$T \in \mathscr{K}(X;Y) \qquad \Longleftrightarrow \qquad T' \in \mathscr{K}(Y';X').$$

The Riesz representation theorem defines an embedding from a Hilbert space to its dual space and back.

THEOREM II.25 (**Riesz representation theorem**, [**4**, 6.1]). *If X is a Hilbert space, then*

$$R_X(x)(y) \coloneqq \langle y, x \rangle_X \qquad \text{for } x, y \in X$$

defines an isometric conjugate linear isomorphism $R_X : X \to X'$.

The Riesz representation theorem yields that the dual space of a Hilbert space is also a Hilbert space. Its scalar product can be defined by

$$\langle x_1', x_2' \rangle_{X'} \coloneqq \langle R_X^{-1} x_1', R_X^{-1} x_2' \rangle_X \qquad x_1', x_2' \in X' .$$

The Lax–Milgram theorem uses the Riesz representation theorem to translate sesquilinear forms (or bilinear forms in real Hilbert spaces) into bounded invertible operators.

THEOREM II.26 (**Lax–Milgram theorem**, [**4**, 6.2]). *Let X be a Hilbert space over $\mathbb{K}$ and let $a : X \times X \to \mathbb{K}$ be sesquilinear. Assume that there exist constants c_0 and C_0 with $0 < c_0 \leq C_0 < \infty$ such that for all $x, y \in X$*

(1) ***continuous:*** $|a(x,y)| \leq C_0 \|x\|_X \|y\|_X$,
(2) ***coercive/elliptic:*** $\Re(a(x,x)) \geq c_0 \|x_0\|_X^2$.

Then there exists a unique map $A : X \to X$ with

$$a(y,x) = \langle y, Ax \rangle_X \qquad \forall x, y \in X .$$

In addition, $A \in \mathscr{L}(X)$ is an invertible operator with

$$\|A\|_{\mathscr{L}(X)} \leq C_0 , \qquad \|A^{-1}\|_{\mathscr{L}(X)} \leq \frac{1}{c_0} .$$

REMARK II.27. The Lax–Milgram theorem is best known for the more specific case of bilinear forms and a real Hilbert space V. Note also that we could compartmentalize some of its claims, cf. the proof of [**4**, 6.2]. If we only assume to have a sesquilinear form, without continuity or ellipticity, then

$$a(y,x) = \langle y, Ax \rangle_X \qquad \forall x, y \in X$$

still defines a linear operator $A : X \to X$, which is in general not bounded. If we only have continuity, then we still have a bonded linear operator $A \in \mathscr{L}(X)$ since

$$\|A(x)\|_X = \|a(\cdot, x)\|_{X'} \leq C_0 \|x\|_X ,$$

albeit in general not an invertible one.

In analogy to the operator defined in the Lax–Milgram theorem, we define ellipticity.

DEFINITION II.28. Let X be a Hilbert space. We call an operator $A \in \mathscr{L}(X)$ **coercive** or **elliptic** if there is a constant $c_0 > 0$, such that

$$\Re(\langle Ax, x \rangle_X) \geq c_0 \|x\|_X^2 \qquad \forall x \in X .$$

We emphasize it as X-elliptic if the norm is ambiguous.

Gelfand Triple. Variational EVPs are usually modeled on two Hilbert spaces V, H with $\mathrm{Id} : V \hookrightarrow H$. We introduce Gelfand triples that provide another embedding into the dual space of V. To this end, we need reflexive spaces.

DEFINITION II.29 ([**4**, 8.2 & 8.8]). Let X be a Banach space. Defining $(J_X x)(x') := x'(x)$ for $x \in X, x' \in X'$ yields an isometric map $J_X \in \mathscr{L}(X; X'')$. Here $X'' := (X')' = \mathscr{L}(X'; \mathbb{K})$ is called the **bidual space** of X. We call X **reflexive** if J_X is surjective.

Note that every Hilbert space V is reflexive, cf. [**4**, 8.11(1)]. We summarize the definitions of the Gelfand triple [**103**, § 17].

THEOREM II.30 ([**103**, Definition 17.1 & Theorem 17.3]). *Let X be a reflexive Banach space and Y a Hilbert space. Suppose that $i : X \hookrightarrow Y$, and that $\mathscr{R}(i)$ is dense in Y. Then $i' : Y' \to X'$ is continuous, injective and $\mathscr{R}(i')$ is dense in X'. Altogether we have*

$$X \overset{i}{\hookrightarrow} Y \overset{R_Y}{\to} Y' \overset{i'}{\hookrightarrow} X' , \tag{II.6}$$

*where both embeddings i, i' are continuous, injective, and have dense images in Y and X' respectively. A scheme of this kind is called a **Gelfand triple**. If the embedding operations i, i' are clear from context, they are omitted. The norm of X' has the property*

$$\|y\|_{X'} = \sup_{0 \neq x \in X} \frac{|\langle y, ix\rangle_Y|}{\|x\|_X} , \qquad y \in Y . \tag{II.7}$$

Applying Schauder's theorem (theorem II.24), if the embedding $i : X \hookrightarrow Y$ is compact, then so is the dual embedding $i' : Y' \hookrightarrow X'$.

II.2.5. Self-adjoint and Normal Operators.

DEFINITION II.31 ([**4**, 12.2 & 12.9]). Let X, Y be Hilbert spaces.

(1) Let $R_X : X \to X', R_Y : Y \to Y'$ be the isomorphism from theorem II.25. For $T \in \mathscr{L}(X; Y)$ we define the **Hilbert adjoint** as
$$T^* = R_X^{-1} \circ T' \circ R_Y .$$
Then we have $T^* \in \mathscr{L}(Y; X)$ and
$$\langle x, T^* y\rangle_X = \langle Tx, y\rangle_Y \qquad \forall x \in X, y \in Y .$$
In the case $Y = X$ we call $T \in \mathscr{L}(X)$ **self-adjoint** if $T^* = T$.
(2) An operator $T \in \mathscr{L}(X)$ is called **normal** if T and T^* commute, i.e., $T^* T = T T^*$.

We may call an operator $\tilde{A} \in \mathscr{L}(X; X'), x \mapsto \tilde{A}x = \langle \cdot, Ax\rangle_X$ self-adjoint if the corresponding operator $A \in \mathscr{L}(X)$ is self-adjoint. Every self-adjoint operator is normal. A special case are unitary operators which are useful to manipulate operators.

DEFINITION II.32. Let X be a Hilbert space.

(1) An operator $T \in \mathscr{L}(X)$ is called **unitary** if $T^* \cdot T = \mathrm{Id} = T \cdot T^*$.
(2) A matrix $T \in \mathbb{K}^{m \times m}$ is called **unitary** if $T^* \cdot T = \boldsymbol{I} = T \cdot T^*$.
(3) A matrix $T \in \mathbb{R}^{m \times m}$ is called **orthogonal** or **orthonormal** if $T^\top \cdot T = \boldsymbol{I} = T \cdot T^\top$.

II.3. Derivatives

Following [**4, 106**], we recall the definition of Fréchet and Gâteaux derivatives in Banach spaces and the space of differentiable functions. We assume that the reader is already familiar with derivatives of functions $f : \mathbb{R}^n \to \mathbb{R}^m$ with $n, m \in \mathbb{N}$, and start by recapitulating the multi-index notation for partial derivatives. Next we recall Landau's $\mathcal{O}$- and o-notation, which are convenient for the definition of first-order derivatives in Banach spaces and Taylor's theorem. We formulate the usual derivation rules and then show how higher-order and partial derivatives can be defined. Equipped with these definitions, we can then define the space of differentiable functions and state Taylor's theorem (theorem II.42) as well as the implicit function theorem (theorem II.45), which we need for the development of some of the central results of this thesis.

II.3.1. Multi-Index Notation. The multi-index notation is a convenient way to denote partial derivatives of higher order. For the convenience of the reader, we recall the multi-index notation and multinomial coefficients, which we use in chapter III. We call a tuple of non-negative integers

$$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_n) \in \mathbb{N}_0^n$$

a **multi-index**. We call $|\boldsymbol{\alpha}| = \sum_{i=1}^n \alpha_i$ the **order** of multi-index $\boldsymbol{\alpha}$. Let $\boldsymbol{\beta} \in \mathbb{N}_0^n$ be a second multi-index. We can add two multi-indices by adding elementwise, i.e.,

$$\boldsymbol{\alpha} + \boldsymbol{\beta} = (\alpha_1 + \beta_1, \dots, \alpha_n + \beta_n)$$

and find $|\boldsymbol{\alpha} + \boldsymbol{\beta}| = |\boldsymbol{\alpha}| + |\boldsymbol{\beta}|$. There is also an order of two multi-indices via the partial order of their elements, i.e.,

$$\begin{aligned} \boldsymbol{\alpha} \le \boldsymbol{\beta} \quad &\iff \quad \alpha_i \le \beta_i, \qquad i = 1, \dots, n\,, \\ \boldsymbol{\alpha} < \boldsymbol{\beta} \quad &\iff \quad \alpha_i < \beta_i, \qquad i = 1, \dots, n\,. \end{aligned}$$

The mixed derivatives of a function $f : \mathbb{R}^n \to \mathbb{R}$ can now be expressed as

$$\partial^{\boldsymbol{\alpha}} := \prod_{i=1}^n \partial_i^{\alpha_i} \qquad \text{with partial derivatives } \partial_i^{\alpha_i} := \tfrac{\partial^{\alpha_i}}{\partial x_i^{\alpha_i}}\,.$$

The **factorial** of a multi-index is defined as $\boldsymbol{\alpha}! = \prod_{i=1}^n (\alpha_i!)$. Then **multinomial coefficient** is defined as

$$\binom{k}{\alpha_1, \alpha_2, \dots, \alpha_n} = \frac{k!}{\prod_{i=1}^n \alpha_i!} = \frac{|\boldsymbol{\alpha}|!}{\boldsymbol{\alpha}!}\,, \tag{II.8}$$

where $k = |\boldsymbol{\alpha}| \in \mathbb{N}_0^n$. Let $f_1, \dots, f_n : \mathbb{R} \to \mathbb{R}$ be k-times differentiable functions, then, using the **product rule**, we can write the k-th partial derivative of their product as

$$\frac{\mathrm{d}^k \left(\prod_{i=1}^{n} f_i(x) \right)}{\mathrm{d}x^k} = \sum_{\alpha_1 + \cdots + \alpha_n = k} \binom{k}{\alpha_1, \dots, \alpha_n} \prod_{i=1}^{n} \frac{\mathrm{d}^{\alpha_i}}{\mathrm{d}x^{\alpha_i}} f_i(x) \, .$$

The above sum includes the $\binom{n+k-1}{n-1,k}$ multiindices $\boldsymbol{\alpha} \in \mathbb{N}^n$ of order k.

II.3.2. Derivatives in Banach Spaces. We recall the Landau notation, which we use to define derivatives and to quantify residues qualitatively throughout this thesis.

DEFINITION II.33. Let $f, g : \mathbb{R} \to \mathbb{R}$ be functions and $x_0 \in \mathbb{R} \cup \{-\infty, \infty\}$ be a finite limit. We use the following notations, as part of the **Landau notation**,

$$f \in o(g) \quad (x \to x_0) \qquad :\Longleftrightarrow \qquad \lim_{x \to x_0} \frac{|f(x)|}{|g(x)|} = 0 \, ,$$

$$f \in \mathcal{O}(g) \quad (x \to x_0) \qquad :\Longleftrightarrow \qquad \limsup_{x \to x_0} \frac{|f(x)|}{|g(x)|} < \infty \, .$$

In most cases, the limit point x_0 is clear from context and will not be specified. When we use the Landau symbols additively in equations, this is shorthand for the remaining terms in question being in the set of functions that the Landau notation describes.

Fréchet and Gâteaux derivatives. We recall Fréchet and Gâteaux derivatives, which will be useful in investigating the differentiability of EVPs.

DEFINITION II.34 ([**106**, Definition 4.5]). Let X, Y be Banach spaces with $B(x) \subset X$ a neighborhood of $x \in X$ and $f : B(x) \to Y$ be a map.

(1) The map f is **Fréchet differentiable** at x if there exists a map $\mathbf{D}_x f \in \mathscr{L}(X; Y)$ such that

$$f(x+h) - f(x) = \mathbf{D}_x f[h] + o(\|h\|_X) \, , \qquad h \to 0 \in X \, , \tag{II.9a}$$

for all $h \in B(0) \subset X$. If it exists, $\mathbf{D}_x f$ is called the **Fréchet derivative** at x.

(2) The map f is **Gâteaux differentiable** at x if there exists a map $\mathrm{D}_x f \in \mathscr{L}(X; Y)$ such that

$$f(x+th) - f(x) = t \, \mathrm{D}_x f[h] + o(t) \, , \qquad t \to 0 \in \mathbb{R} \, , \tag{II.9b}$$

for all h with $\|h\|_X = 1$ and all $t \in B(0) \subset \mathbb{R}$. $\mathrm{D}_x f$ is called the **Gâteaux derivative** of f at x.

(3) If the Fréchet derivatives $\mathbf{D}_x f$ exist for all $x \in U \subset X$, then the mapping

$$\mathbf{D} f : U \to \mathscr{L}(X; Y) \, , \qquad x \mapsto \mathbf{D}_x f \tag{II.9c}$$

is called the Fréchet derivative of f on U. The analogous holds, respectively, for Gâteaux derivatives $\mathrm{D}_x f$.

(4) **Higher-order derivatives** (Fréchet or Gâteaux) are defined inductively, i.e., $\mathbf{D}_x^2 f$ is the Fréchet derivative of $\mathbf{D} f$ at x.

The Gâteaux derivative can alternatively be defined as the limit

$$\mathrm{D}_x f[k] = \lim_{t\to 0} \frac{f(x+tk) - f(x)}{t} \tag{II.10}$$

and the Fréchet derivative can be defined implicitly as a bounded linear operator for which the limit

$$\lim_{\|h\|_X \to 0} \frac{\|f(x+h) - f(x) - \mathbf{D}_x f[h]\|_Y}{\|h\|_X} = 0$$

holds.

DEFINITION II.35. Let X be a Banach space and $f : X \to \mathbb{R}$ be Fréchet differentiable at $x \in X$, then we define the **gradient** $\operatorname{grad} f \in \mathscr{L}(X)$ of f at x such that

$$\mathbf{D}_x f[h] =: \big\langle \operatorname{grad} f(x), h \big\rangle_X \qquad\qquad \forall h \in X\,.$$

The following lemma makes clear why Fréchet differentiability is stronger than Gâteaux differentiability.

LEMMA II.36 ([**106**, Proposition 4.8]).

(1) *Every Fréchet derivative at x is also a Gâteaux derivative at x.*
(2) *A Gâteaux derivative at x for which the passage to limit* (II.10) *is uniform for all k with $\|k\|_X = 1$, is also a Fréchet derivative at x.*
(3) *If $\mathrm{D} f$ exists as a Gâteaux derivative in some neighborhood of x, and $\mathrm{D} f$ is continuous at x, then $\mathrm{D}_x f$ is also a Fréchet derivative at x.*
(4) *If $\mathbf{D}_x f$ exists as a Fréchet derivative at x, then $\mathbf{D} f$ is also continuous at x.*

The n-th Fréchet derivative $\mathbf{D}^n f(x)$ of $f : X \supset U \to Y$ at x has already been defined in definition II.34 by induction. This yields

$$\begin{aligned} f &: U \to Y\,, \\ \mathbf{D} f &: U \to \mathscr{L}(X;Y)\,, \\ \mathbf{D}^2 f &: U \to \mathscr{L}(X;\mathscr{L}(X;Y))\,, \\ &\vdots \\ \mathbf{D}^n f &: U \to \mathscr{L}(X;\dots\,\mathscr{L}(X;Y))\,. \end{aligned}$$

We can redefine higher-order derivatives of order n equivalently as n-linear forms using the isometric isomorphism (II.4) between the nested bounded linear forms and n-linear forms, cf. [**106**, Proposition 4.19], i.e.,

$$\mathbf{D}^n f : U \to \mathscr{L}^{(n)}(X;Y)\,, \qquad\qquad x \mapsto \mathbf{D}_x^n f \in \mathscr{L}^{(n)}(X;Y)\,.$$

We also introduce the more convenient notation

$$\mathbf{D}_x^n f[h] := \mathbf{D}_x^n f[h, \cdots, h] ,$$

Derivation Rules. We recall the derivation rules for the case of Fréchet and Gâteaux derivatives.

LEMMA II.37 ([**106**, Proposition 4.9 - 4.11]).

(1) ***Sum rule:***
Let X, Y be Banach spaces, $B(x) \subset X$ a neighborhood of $x \in X$, and $f, g : B(x) \to Y$ Fréchet differentiable mappings. Then it is true for all $\alpha, \beta \in \mathbb{K}$ that

$$\mathbf{D}_x(\alpha f + \beta g) = \alpha \mathbf{D}_x f + \beta \mathbf{D}_x g ,$$

where the expression of the left exists as a Fréchet derivative.

(2) ***Chain rule:***
Let X, Y, Z be Banach spaces, $B(x) \subset X$ a neighborhood of $x \in X$, and $V(y) \subset Y$ a neighborhood of $y \in Y$. Let x be fixed and set $y = f(x)$. Suppose we are given maps

$$f : B(x) \to Y , \qquad\qquad g : V(y) \to Z$$

with $f(B(x)) \subset V(y)$. This defines the composite map

$$h := g \circ f : B(x) \to Z .$$

Suppose that $\mathbf{D}_x f$ and $\mathbf{D}_{f(x)} g$ exist as Fréchet derivatives. Then, h is Fréchet differentiable at x and it holds

$$\mathbf{D}_x h = \mathbf{D}_{f(x)} g[\mathbf{D}_x f] .$$

(3) ***Product rule:***
Let X, X_1, X_2, Y be Banach spaces, $U_1, U_2 \subset X$, and $F \in \mathscr{L}^{(2)}(X_1 \times X_2; Y)$. Suppose further, that the maps

$$f_i : U_i(x) \to X_i , \qquad\qquad i = 1, 2$$

are Fréchet differentiable at x. Then $b(x) := F(f_1(x), f_2(x))$ is Fréchet differentiable at x and

$$\mathbf{D}_x b[h] = F(\mathbf{D}_x f_1[h], f_2(x)) + F(f_1(x), \mathbf{D}_x f_2[h]) .$$

If the derivatives only exist as Gâteaux derivatives, the above rules hold respectively replacing the Fréchet derivatives by Gâteaux derivatives.

Partial Fréchet and Gâteaux derivatives. Now we recall partial Fréchet and Gâteaux derivatives, i.e., derivatives in one of several Banach space-valued arguments. This generalizes the notion of partial derivatives towards elements of $\mathbb{R}^n$ we previously considered.

DEFINITION II.38 ([**106**, Definition 4.13]). Let X, Y, Z be Banach spaces with a neighborhood $B((x_0, y_0)) \subset X \times Y$ around $(x_0, y_0) \in X \times Y$ and consider a mapping

$$f : B((x_0, y_0)) \to Z\,, \qquad (x, y) \mapsto f(x, y)\,.$$

Let y be fixed and set $g(x) = f(x, y)$. If g has a Fréchet derivative at x, then we define the **partial Fréchet derivative** of f at (x, y) with respect to the first variable x to be $\partial_x f(x, y) = \mathbf{D}_x g$. The partial derivative $\partial_y f(x, y)$ is defined analogously by fixing x instead of y. **Partial Gâteaux derivatives** are defined analogously using Gâteaux derivatives instead.

The following lemma specifies the relationship between Fréchet derivatives with respect to the space $X \times Y$ and partial Fréchet derivatives.

LEMMA II.39 ([**106**, Proposition 4.14]).

(1) *If f is Fréchet differentiable at (x, y), then the partial Fréchet derivatives $\partial_x f$ and $\partial_y f$ exists at (x, y) and*

$$\mathbf{D}_{x,y} f[h, k] = \partial_x f(x, y)[h] + \partial_y f(x, y)[k] \tag{II.11}$$

holds for all $h \in X, k \in Y$.

(2) *Conversely, if f has partial Fréchet derivatives $\partial_x f$ and $\partial_y f$ in a neighborhood of (x, y), and if theses are continuous at (x, y), then $\mathbf{D}_{x,y} f$ exists as a Fréchet derivative and* (II.11) *holds.*

(3) *The map f is continuously Fréchet differentiable in a neighborhood of (x, y) if and only if all partial Fréchet derivatives are continuous in a neighborhood of (x, y).*

Definition II.38 and lemma II.39 can be generalized to maps with an arbitrary number of arguments

$$(x_1, \dots, x_n) \mapsto f(x_1, \dots, x_n)\,.$$

II.3.3. Regularity. We establish some notations for sets of functions that are Fréchet differentiable to a certain order.

DEFINITION II.40 ([**106**, Definition 4.22]). Let X, Y be Banach spaces, $U \subset X$ a subset, $f : U \to Y$ a mapping, and $m \in \mathbb{N}$.

(1) If U is open, then f is called a C^m**-map** if and only if f has continuous Fréchet derivatives up to order r on U. We use the notation

$$C^m(U; Y) := \big\{ f : U \to Y \;:\; f \text{ is } m\text{-times continuously Fréchet differentiable.} \big\}$$

with shorthand $C^m(U) := C^m(U; \mathbb{R})$.

(2) If U is arbitrary, then f is called a C^m-map if and only if it can be extended locally to a C^m-map in the previous sense. That means that for $x \in U$, there exists an open neighborhood $B(x)$ such that f can be extended to a C^m-map on $B(x)$.

(3) We extend the notation to

$$C^0(U;Y) \coloneqq \left\{ f : X \to Y \,:\, f \text{ is continuous.} \right\},$$
$$C^\infty(U;Y) \coloneqq \bigcap_{m\geq 0} C^m(U;Y),$$

where we also write $C(U;Y) \coloneqq C^0(U;Y)$ and use the convention $\mathbf{D}^0 f \coloneqq f$. We call $f \in C^\infty(U;Y)$ **smooth**.

Lipschitz continuous functions are used later and lead to Lipschitz boundaries.

DEFINITION II.41 ([**4**, 3.7 & A8.2]). Let X, Y be Banach spaces with $U \subset X$.

(1) A function $f : U \to Y$ is called **Lipschitz continuous** if there exists a constant $L > 0$, such that

$$\|f(x_1) - f(x_2)\|_Y \leq L\|x_1 - x_2\|_X, \qquad \forall\, x_1, x_2 \in U.$$

(2) A bounded domain $\mathscr{D}$ is said to have **Lipschitz boundary** if its boundary $\partial\mathscr{D}$ can be covered by finitely many open sets $U_1, \ldots, U_l$ such that $\partial\mathscr{D} \cap U_j$ for $j = 1, \ldots, l$ is a graph of a Lipschitz continuous function with $\mathscr{D} \cap U_j$ in each case lying on one side of this graph.

II.3.4. Taylor's Theorem. The integrals in this section are to be understood as Riemann integrals, cf. [**4**, 6.22], which are well defined for continuous functions.

THEOREM II.42 (**Taylor's theorem**, [**106**, Theorem 4.A]). *Let X, Y be Banach spaces, $B(x) \subset X$ an (open, convex) neighborhood of $x \in X$, and $f : B(x) \to Y$ a mapping. If f is n-times Fréchet differentiable, then*

$$f(x+h) = f(x) + \sum_{k=1}^{n-1} \frac{1}{k!} \mathbf{D}_x^k f[h] + R_n$$

holds with

$$\|R_n\|_Y \leq \frac{1}{n!} \sup_{0<\tau<1} \|\mathbf{D}^n_{x+\tau h} f[h]\|_Y.$$

If $\mathbf{D}^n f$ is continuous on $B(x)$, i.e., $f \in C^n(B(x);Y)$, then

$$R_n = \int_0^1 \frac{(1-\tau)^{n-1}}{(n-1)!} \mathbf{D}^n_{x+\tau h} f[h] \, \mathrm{d}\tau.$$

As the integrand of the residue R_n is n-linear, we can describe its behavior qualitatively using the Landau notation.

COROLLARY II.43. *Let X, Y be Banach spaces and $B(x) \subset X$ an (open, convex) neighborhood of $x \in X$. If $f \in C^n(B(x); Y)$, $n \in \mathbb{N}$ and $x + h \in B(x)$, then*

$$f(x+h) = \sum_{k=0}^{n} \frac{\mathbf{D}_x^k f[h]}{k!} + o(\|h\|_X^n) \tag{II.12a}$$

$$f(x+h) = \sum_{k=0}^{n-1} \frac{\mathbf{D}_x^k f[h]}{k!} + \mathcal{O}(\|h\|_X^n) \tag{II.12b}$$

holds in the sense $\|h\|_X \to 0$.

PROOF. See [**22**, 8.14.3] for (II.12a). (II.12b) follows from the bounded n-linearity of the n-th derivative. □

We can define analyticity via the Taylor series.

DEFINITION II.44. Let X, Y be Banach spaces and $B(x) \subset X$ an (open, convex) neighborhood of $x \in X$.

(1) We call the expansions (II.12) with sum up to order n the **Taylor expansion** of order n.
(2) Let $f : B(x) \to Y$ be smooth. The infinite series

$$f(x+h) = \sum_{k=0}^{\infty} \frac{\mathbf{D}_x^k f[h]}{k!} \tag{II.13}$$

is called the **Taylor series** of f at x.
(3) We call a function $f \in C^\infty(X; Y)$ **(locally) analytic** on $B(x)$ if the Taylor series of f at x converges against $f(x+h)$ for $x + h \in B(x)$ and **analytic** if this holds on X. The vector space of (locally) analytic functions on an open, convex subset $U \subset X$ is denoted by

$$C^\omega(U; Y) := \left\{ f : U \to Y \text{ analytic on } U \right\}.$$

with shorthand $C^\omega(U) := C^\omega(U; \mathbb{R})$.

II.3.5. Implicit Function Theorem. The implicit function theorem will be useful in characterizing the derivatives of eigenpairs.

THEOREM II.45 (**Implicit function theorem**, [**106**, Theorem 4.B & Corollary 4.23]). *Let X, Y, Z be Banach spaces, $B((x_0, y_0)) \subset X \times Y$ an open neighborhood on $(x_0, y_0) \in X \times Y$, and $F : B((x_0, y_0)) \to Z$ a mapping with $F((x_0, y_0)) = 0$. Suppose that*

(1) *$\partial_y F : Y \to Z$ exists as a partial Fréchet derivative on $B((x_0, y_0))$ and is bijective in (x_0, y_0),*
(2) *F and $\partial_y F$ are continuous at (x_0, y_0).*

Then the following are true:

(1) ***Existence and uniqueness.***
There exist $r_0, r > 0$ such that for every $x \in X$ satisfying $\|x - x_0\|_X \leq r_0$, there exists one $y(x) \in Y$ for which $\|y(x) - y_0\|_Y \leq r$ and $F(x, y(x)) = 0$.

(2) ***Continuity.***
If F is continuous in $B((x_0, y_0))$, then $y : X \to Z$ is continuous in a neighborhood of x_0.
(3) ***Continuous differentiability.***
If F is a C^m-map, $1 \leq m \leq \infty$, on $B((x_0, y_0))$, then $y : X \to Z$ is a C^m-map on a neighborhood of x_0.
(4) ***Analyticity.***
If F is analytic on $B((x_0, y_0))$, then the solution $y : X \to Z$ is analytic in a neighborhood of x_0.

II.3.6. On Line Integrals. We have used derivatives in example I.2 to track eigenvalues through the parameter space, cf. fig. I.2. Now we want to formalize this investigation using line integrals. First, we need the following definitions to formalize paths.

DEFINITION II.46 ([**4**, p. 242], [**85**, 10.38]). Let X be a normed space.

(1) A **path** is a continuous function $\gamma : [t_0, t_1] \to X$.
(2) $A \subset X$ is called **path-connected** if there is a path $\gamma : [t_0, t_1] \to A$ between any points $x_0, x_1 \in A$, i.e.,

$$\gamma(t_0) = x_0 \,, \qquad \gamma(t_1) = x_1 \,.$$

(3) A **loop** is a continuous path $\gamma : [t_0, t_1] \to X$ with $\gamma(t_0) = \gamma(t_1)$.
(4) $A \subset X$ is called **simply connected** if it is path-connected and every loop can be contracted, i.e., continuously transformed, to a single point in X.

Let X, Y be Banach spaces, $U \subset X$ a path-connected subset, and consider a Gâteaux differentiable function $f : U \to Y$ as well as a continuously differentiable path $\gamma : [t_0, t_1] \to X$, such that $f \circ \gamma : [t_0, t_1] \to Y$ is Fréchet differentiable. Then by lemma II.37 holds

$$\mathbf{D}_t (f \circ \gamma)[s] = \mathrm{D}_{\gamma(t)} f [\mathbf{D}_t \gamma[s]]$$

and by the fundamental theorem of calculus, cf. [**4**, E3.6], we get

$$(f \circ \gamma(t_1)) - (f \circ \gamma(t_0)) = \int_{t_0}^{t_1} \mathrm{D}_{\gamma(t)} f [\mathbf{D}_t \gamma] \, \mathrm{d}t =: \int_\gamma \mathrm{D}_x f \, \mathrm{d}x \,. \tag{II.14}$$

We call such an integral over a path γ a **line integral**.
As seen in the counterexample I.2, the choice of path is often not arbitrary if we continue a function using its (continuous) derivative on a path. However, the choice of path $\gamma \subset U$ between two fixed points is arbitrary on a neighborhood $U \subset X$, where the

function $f : U \to Y$ is continuously Fréchet differentiable, cf. theorem II.42. Integrating over a (piecewise) differentiable loop $\gamma \subset U$ in the parameter space then yields

$$\int_\gamma \mathbf{D}_x f \, \mathrm{d}x = 0 \, .$$

If U is not simply connected, it is, however, not necessarily possible to construct a Fréchet differentiable function $f : U \to Y$ even if a continuous Fréchet derivative $\mathbf{D} f$ is *locally* well defined in U. This will become apparent when we observe the eigenfunctions of example I.2 in chapter III.

Line Integrals as Solutions of Ordinary Differential Equations. We recall some results on ordinary differential equations (ODEs) to show how the tracking problem can be formalized given a fixed path $\gamma : [t_0, t_1] \to U$. Thus, we assume that we know the function value $f(x_0) \in Y$ at some reference point $x_0 \in U$, which is the start of our path, as well as the derivative $\mathrm{D}_{\gamma(t)} f[\mathbf{D}_t \gamma]$ for $t \in I$ on the path. Then we can calculate the function value at the end of the path $\gamma(t_1) = x_1$ by rearranging (II.14) as

$$f(x_1) = f(x_0) + \int_{t_0}^{t_1} \mathrm{D}_{\gamma(t)} f[\mathbf{D}_t \gamma] \, \mathrm{d}t \, . \tag{II.15}$$

We can formalize the derivative into a function

$$g : [t_0, t_1] \times Y \to Y \, , \qquad t \times y \mapsto g(t, y) \, ,$$

where y is an additional argument[1] that takes the function value $f \circ \gamma(t)$. Then we have expressed the change of the function value in an **ordinary differential equation (ODE)** of first order, i.e., involving only the first-order derivative

$$\mathrm{D}_t \, y = g(t, y) \, , \tag{II.16a}$$

The problem to find a solution fitting the **initial value**

$$y(t_0) = f(x_0) \, . \tag{II.16b}$$

is called an **initial value problem (IVP)**. Its solution

$$y : [t_0, t_1] \to Y \, , \qquad t \mapsto y(t) = f \circ \gamma(t) \, ,$$

is called a **trajectory**.

The existence and uniqueness of a solution of an IVP can be proven using the Picard–Lindelöf theorem, cf. [**106**, Proposition 1.8]. In our case, we already know that there is a unique and continuously differentiable trajectory by construction, since we start with an EVP and build the ODE only to formalize the tracking problem.

[1] The function value is not necessarily needed to compute the derivative. For derivatives of eigenpairs, we will see in chapter III that the influence of the current values of eigenpairs on their derivatives is quite explicit, which makes this form useful.

The most prominent class of method to solve IVPs are *Runge–Kutta methods* [**13, 92**]. It is also possible to combine two Runge–Kutta schemes in order to build algorithms with step size control in parameter t, so that the number of function evaluations is controlled by a specified error tolerance.

II.4. Measure Theory and Integrals

In this section, we recall the fundamentals of measure theory. This gives us the opportunity to define the function spaces needed for the variational formulation of EVPs and provides several results needed for probability theory, which we delay until section II.7. The results of this section are taken mainly from [**4, 60**] and our definition of Sobolev spaces follows [**1**].

II.4.1. Measure Spaces. First, we need σ-algebras of a set Ω.

DEFINITION II.47 ([**60**, 1.1, 1.2, 1.16, & 1.21]).

(1) We denote by $P(\Omega)$ the **power set**, the set of all subsets of Ω, which includes the empty set $\emptyset$ and the set Ω.
(2) A set $\mathscr{A} \subset P(\Omega)$ is called a $\boldsymbol{\sigma}$**-algebra** if
   (a) $\emptyset \in \mathscr{A}$,
   (b) $A^{\complement} := \Omega \setminus A \in \mathscr{A}$ for all $A \in \mathscr{A}$, and
   (c) $\bigcup_{j\in\mathbb{N}} F_j \in \mathscr{A}$ for $F_j \in \mathscr{A}$.
(3) Let $\mathscr{E} \subset P(\Omega)$. Then there exists a smallest σ-algebra $\sigma(\mathscr{E})$ with $\mathscr{E} \subset \sigma(\mathscr{E})$

$$\sigma(\mathscr{E}) := \bigcap_{\substack{\mathscr{A}\subset P(\Omega) \text{ is a } \sigma\text{-algebra} \\ \mathscr{A}\supset\mathscr{E}}} \mathscr{A} ,$$

   which is called the σ-algebra **generated** by $\mathscr{E}$. $\mathscr{E}$ is called a **generator** of $\sigma(\mathscr{E})$.
(4) Let X be a normed space. The σ-algebra $\mathscr{B}(X)$ that is generated by the open sets in X is called the **Borel** $\boldsymbol{\sigma}$**-algebra** on Ω.

The Borel σ-algebra provides a suitable σ-algebra, which we can use whenever we encounter a Hilbert or Banach space-valued random variable.

DEFINITION II.48 ([**60**, 1.27, 1.28, & 1.29]). Let $\mathscr{A} \subset P(\Omega)$ and let $\mu : \mathscr{A} \to [0,\infty]$ be a set function. We say that μ is

(1) **additive** if for any choice of finitely many mutually disjoint sets $A_1, \dots, A_n \in \mathscr{A}$ with $\bigcup_{i=1}^n A_i \in \mathscr{A}$

$$\mu\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mu(A_i) .$$

   It is called $\boldsymbol{\sigma}$**-additive** if this holds for countably infinite sets, i.e., $n = \infty$.
(2) Let $\mathscr{A}$ be a σ-algebra and let $\mu : \mathscr{A} \to [0,\infty]$ be a σ-additive set function with $\mu(\emptyset) = 0$, then μ is called a **measure**.

(3) A measure μ is called **finite** if $\mu(A) < \infty$ for every $A \in \mathscr{A}$ and $\boldsymbol{\sigma}$**-finite** if there exists a sequence of sets $\Omega_1, \Omega_2, \ldots \in \mathscr{A}$ such that $\Omega = \bigcup_{n=1}^{\infty} \Omega_n$ and $\mu(\Omega_n) < \infty$ for all $n \in \mathbb{N}$.

Measures lead to the definition of a measure space.

DEFINITION II.49 ([**60**, 1.38]).

(1) A pair $(\Omega, \mathscr{A})$ consisting of a nonempty set Ω and a σ-algebra $\mathscr{A} \subset P(\Omega)$ is called a **measurable space**. The sets $A \in \mathscr{A}$ are called **measurable sets**.
(2) A triple $(\Omega, \mathscr{A}, \mu)$ is called a **measure space** if $(\Omega, \mathscr{A})$ is a measurable space and if μ is a measure on $\mathscr{A}$.

An important example is the Lebesgue measure, which we denote Leb, cf. [**60**, 1.55].

DEFINITION II.50 ([**60**, 1.68 & 1.69]). Let $(\Omega, \mathscr{A}, \mu)$ be a measure space.

(1) A set $A \in \mathscr{A}$ is called a $\boldsymbol{\mu}$**-null set**, or briefly a null set if $\mu(A) = 0$. By $\mathscr{N}_\mu$ we denote the class of μ-null sets.
(2) Let $E(\omega)$ be a property that a point $\omega \in \Omega$ can have or not. We say that E holds $\boldsymbol{\mu}$**-almost everywhere (a.e.)** if there exists a null set N such that $E(\omega)$ holds for every $\omega \in \Omega \setminus N$. If $A \in \mathscr{A}$ and if there exists a null set N such that $E(\omega)$ holds for every $\omega \in A \setminus N$, then we say that E holds almost everywhere on A.
(3) A measure space $(\Omega, \mathscr{A}, \mu)$ is called **complete** if $\mathscr{N}_\mu \subset \mathscr{A}$.

Note that for every measure space $(\Omega, \mathscr{A}, \mu)$ there exists a completion $(\Omega, \sigma(\mathscr{A} \cup \mathscr{N}_\mu), \tilde{\mu})$ with $\tilde{\mu}(A \cup N) = \mu(A)$ for all $A \in \mathscr{A}$ and $N \in \mathscr{N}_\mu$, cf. [**60**, Remark 1.70]. To avoid technicalities, we assume throughout the rest of this thesis that every measure space is complete.

DEFINITION II.51. [**4**, 3.11] Let $(\Omega, \mathscr{A}, \mu)$ be a measure space and Y a Banach space. A map $f : \Omega \to Y$ is called $\boldsymbol{\mu}$**-measurable** if

(1) $U \in \mathscr{B}(Y) \implies f^{-1}(U) \in \mathscr{A}$.
(2) There exists a μ-null set N such that $f(\Omega \setminus N)$ is separable.

If the space Y is itself separable, the second condition is trivially satisfied, leading to a concept of measurability dependent only on the measurable space.

Recall the following results, which are convenient to ensure measurability.

LEMMA II.52 ([**4**, 3.12]). *The following hold:*

(1) *If $f_1 : \Omega \to Y_1$ and $f_2 : \Omega \to Y_2$ are μ-measurable, then $(f_1, f_2) : \Omega \to Y_1 \times Y_2$ is also μ-measurable.*
(2) *If $f : \Omega \to Y$ is μ-measurable, Z is a Banach space, and $\phi : Y \to Z$ is continuous, then $\phi \circ f$ is also μ-measurable.*

A μ-measurable function connects two measurable spaces and induces a measure.

DEFINITION II.53 ([**60**, Definition 1.98]). Let $(\Omega, \mathscr{A}, \mu)$ be a measure space, Y a Banach space, and $f : \Omega \to Y$ be μ-measurable. The **image measure** (or **push-forward measure**) of μ under the map f is the measure $\mu \circ f^{-1}$ on $(Y, \mathscr{B}(Y))$ that is defined by

$$\mu \circ f^{-1} : \mathscr{B}(Y) \to [0, \infty] , \qquad A \mapsto \mu(f^{-1}(A)) .$$

II.4.2. Bochner–Lebesgue integral. We recall the following definitions.

DEFINITION II.54 ([**62**, 1.21]). Let Y be a Banach space and $(\Omega, \mathscr{A}, \mu)$ a measure space.

(1) Let $A \in \mathscr{A}$ be a set, then the **indicator function** of A is

$$\mathbb{1}_A(\omega) := \begin{cases} 1, & \omega \in A , \\ 0, & \text{else} . \end{cases}$$

(2) A function $s : \Omega \to Y$ is **simple** if there exist $s_j \in Y$ and $A_j \in \mathscr{A}$ for $i = 1, \dots, N$ with $\mu(A_j) < \infty$ such that $s(\omega) = \sum_{j=1}^{N} s_j \, \mathbb{1}_{A_j}(\omega), \omega \in \Omega$. The integral of a simple function with respect to a measure space $(\Omega, \mathscr{A}, \mu)$ is

$$\int_\Omega s(\omega) \, \mathrm{d}\mu(\omega) := \sum_{j=1}^{N} s_j \mu(A_j) .$$

(3) We say a μ-measurable function u is **(Bochner–Lebesgue-)integrable** with respect to μ if there exist simple functions u_n, such that $u_n(\omega) \to u(\omega)$ as $n \to \infty$ for almost all $\omega \in \Omega$ and u_n is a Cauchy sequence in the sense that for $\epsilon > 0$

$$\int_\Omega \| u_n(\omega) - u_m(\omega) \|_Y \, \mathrm{d}\mu(\omega) < \epsilon$$

holds for any n, m sufficiently large. Note that $\omega \mapsto \| u_n(\omega) - u_m(\omega) \|_Y$ is a simple function. We denote the space of (Bochner–Lebesgue-)integrable functions as

$$L_\mu(\Omega; Y) := \left\{ f : \Omega \to Y \ : \ f \text{ is (Bochner–Lebesgue-)integrable with respect to } \mu \right\}$$

with shorthand $L_\mu(\Omega) := L_\mu(\Omega; \mathbb{R})$.

(4) If $u \in L_\mu(\Omega; Y)$, we define

$$\int_\Omega u(\omega) \, \mathrm{d}\mu(\omega) := \lim_{n \to \infty} \int_\Omega u_n(\omega) \, \mathrm{d}\mu(\omega)$$

and for $A \in \mathscr{A}$ we define

$$\int_A u(\omega) \, \mathrm{d}\mu(\omega) := \int_\Omega u(\omega) \, \mathbb{1}_A(\omega) \, \mathrm{d}\mu(\omega) .$$

This integral is called the **Bochner–Lebesgue integral**.

For $\mathscr{D} \subset \mathbb{R}^d$, we write the integral with respect to the completion of the measure space $(\mathscr{D}, \mathscr{B}(\mathscr{D}), \text{Leb})$ as

$$\int_{\mathscr{D}} u(\mathbf{x}) \, \mathrm{d}\mathbf{x} := \int_{\mathscr{D}} u(\mathbf{x}) \, \mathrm{d}\,\mathrm{Leb}(\mathbf{x}) \, .$$

LEMMA II.55 (**Bochner's criterion**, [**4**, A3.19(1)]). *Let $(\Omega, \mathscr{A}, \mu)$ a measure space, Y a Banach space, and $f : \Omega \to Y$ a μ-measurable function, then*

$$f \in L_\mu(\Omega; Y) \iff f \text{ is } \mu\text{-measurable and } \|f\|_Y \in L_\mu(\Omega).$$

If both the Lebesgue integral and the Riemann integral of a function are well defined for a function $f : \mathbb{R} \supset I \to \mathbb{R}$, they coincide [**60**, Theorem 4.23]. For more details on the construction of integrals and important results, see [**4**, A3].

II.4.3. Radon–Nikodým derivatives. Another concept relevant later for conditional probabilities is the Radon–Nikodým derivative. We first require the concept of absolute continuity of measures.

DEFINITION II.56 ([**60**, 7.30]). Let μ and ν be two measures on the same measure space $(\Omega, \mathscr{A})$. Then μ is called **absolutely continuous**, with notation $\mu \ll \nu$, if $\mu(A) = 0$ for all $A \in \mathscr{A}$ with $\nu(A) = 0$. The measures are called **equivalent** if $\mu \ll \nu$ and $\nu \ll \mu$.

THEOREM II.57 (**Radon–Nikodým theorem**, [**93**, Theorem 6.2]). *Let $\mu, \nu : \mathscr{A} \to [0, \infty]$ be two measures on the same measure space $(\Omega, \mathscr{A})$. If $\mu \ll \nu$ and ν is σ-finite then there exists a ν-measurable function $f : \Omega \to [0, \infty]$ such that, for all ν-measurable sets $A \in \mathscr{A}$,*

$$\mu(A) = \int_A f(\omega) \, \mathrm{d}\nu(\omega) \, .$$

*The function f is called the **Radon–Nikodým derivative** of μ with respect to ν with notation*

$$\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(\omega) := f(\omega) \, .$$

II.4.4. Bochner–Lebesgue spaces. Using the integrals of definition II.54, we can now define the function spaces that we require for the variational formulations of EVPs and probability theory.

DEFINITION II.58 ([**4**, 3.15 & 3.16]). Let $(\Omega, \mathscr{A}, \mu)$ be a measure space and Y a Banach space over $\mathbb{K}$. For $p = \mathbb{N} \cup \{\infty\}$ the spaces

$$L^p_\mu(\Omega; Y) := \left\{ f \text{ is } \mu\text{-measurable and } \|f\|_{L^p_\mu(\Omega;Y)} < \infty \right\}$$

with norms

$$\|u\|_{L^p_\mu(\Omega;Y)} := \begin{cases} \left(\int_\Omega \|u(\omega)\|^p_Y \, \mathrm{d}\mu(\omega) \right)^{\frac{1}{p}} \, , & p < \infty \, , \\ \operatorname*{ess\,sup}_{\omega \in \Omega} \|u(\omega)\|_Y \, , & p = \infty \, , \end{cases}$$

are called **Bochner–Lebesgue spaces**. We consider two functions $f, g : \Omega \to Y$ to be equivalent in $L^p_\mu(\Omega; Y)$, i.e., $f = g$ in $L^p_\mu(\Omega; Y)$, if $f = g$ μ-a.e.. We use the shorthand notation $L^p_\mu(\Omega) \coloneqq L^p_\mu(\Omega; \mathbb{R})$ and drop the measure in the index if it is the standard Lebesgue measure.

THEOREM II.59 ([**4**, 3.16]). *Under the assumption of definition II.58 the following hold:*

(1) $(L^p_\mu(\Omega; Y), \|\cdot\|_{L^p(\Omega;Y)})$ *is a Banach space.*
(2) *For* $p = 1$, $L^p_\mu(\Omega; Y)$ *coincides with the space of Bochner–Lebesgue-integrable functions, i.e.,* $L^1_\mu(\Omega; Y) = L_\mu(\Omega; Y)$.
(3) *For* $p = 2$ *and if* Y *is a Hilbert space with scalar product* $\langle\cdot,\cdot\rangle_Y$*, the space* $L^2_\mu(\Omega; Y)$ *with*

$$\langle u, v\rangle_{L^2_\mu(\Omega;Y)} \coloneqq \int_{\mathscr{D}} \langle u(\omega), v(\omega)\rangle_Y \, \mathrm{d}\mu(\omega)$$

becomes a Hilbert space.

II.4.5. Sobolev Spaces. We also need Sobolev spaces for the variational formulation of EVPs.

DEFINITION II.60 ([**1**, 1.62, 3.1 & 3.2] & [**4**, 3.4, 3.6, & 5.13]). Let $\mathscr{D} \subset \mathbb{R}^n$ be a domain and $\boldsymbol{\alpha}$ a multi-index.

(1) The **support** of a function $f : \mathscr{D} \to Y$ is

$$\operatorname{supp}(f) \coloneqq \operatorname{clos}\left(\left\{ x \in \mathscr{D} \,:\, f(x) \neq 0 \right\}\right) \subset \operatorname{clos}(\mathscr{D}) \,.$$

We define the space of **smooth functions with compact support** as

$$C^\infty_0(\mathscr{D}; Y) \coloneqq \left\{ f \in C^\infty(\mathscr{D}; Y) \,:\, \operatorname{supp}(f) \text{ is a compact subset of } \mathscr{D} \right\} .$$

We use shorthand notation $C^\infty_0(\mathscr{D}) \coloneqq C^\infty_0(\mathscr{D}; \mathbb{R})$.

(2) For $1 \le p \le \infty$ we define the **locally integrable functions** as

$$L^p_{\mathrm{loc}}(\mathscr{D}) \coloneqq \left\{ f : \mathscr{D} \to \mathbb{K} \,:\, f|_U \in L^p(U) \text{ for all compact subsets } U \subset \mathscr{D} \right\} .$$

(3) Let $u \in L^1_{\mathrm{loc}}(\mathscr{D})$ and $\boldsymbol{\alpha} \in \mathbb{N}^n_0$ a multi-index. A function $u_{\boldsymbol{\alpha}} \in L^1_{\mathrm{loc}}(\mathscr{D})$ which satisfies

$$\int_{\mathscr{D}} u(\mathbf{x}) \partial^{\boldsymbol{\alpha}} \varphi(\mathbf{x}) \, \mathrm{d}\mathbf{x} = (-1)^{|\boldsymbol{\alpha}|} \int_{\mathscr{D}} u_{\boldsymbol{\alpha}}(\mathbf{x}) \varphi(\mathbf{x}) \, \mathrm{d}\mathbf{x} \,, \qquad \forall \varphi \in C^\infty_0(\mathscr{D})$$

is called a **weak derivative** of f with notation $\partial^{\boldsymbol{\alpha}} u = u_{\boldsymbol{\alpha}}$. For continuously differentiable functions $f \in C^m(\mathscr{D})$, the weak derivative is equal to the respective partial derivative.

(4) The **Sobolev space** of order $m \in \mathbb{N} \cup \{\infty\}$ with exponent p, $1 \le p \le \infty$ is defined by

$$W^{m,p}(\mathscr{D}) \coloneqq \left\{ f \,:\, f \text{ has weak derivatives } |\boldsymbol{\alpha}| \le m, \partial^{\boldsymbol{\alpha}} f \in L^p(\mathscr{D}) \right\} .$$

We also define norms

$$\|f\|_{W^{m,p}(\mathscr{D})} \coloneqq \left(\sum_{0\le|\boldsymbol{\alpha}|\le m} \|\partial^{\boldsymbol{\alpha}} f\|^p_{L^p(\mathscr{D})} \right)^{\frac{1}{p}} ,$$
$$\|f\|_{W^{m,\infty}(\mathscr{D})} \coloneqq \max_{0\le|\boldsymbol{\alpha}|\le m} \|\partial^{\boldsymbol{\alpha}} f\|_{L^\infty(\mathscr{D})} .$$

The **Sobolev space with zero boundary values**[2] of order m with exponent p is defined by

$$W^{m,p}_0(\mathscr{D}) \coloneqq \overline{C^\infty_0(\mathscr{D})}^{\|\cdot\|_{W^{m,p}(\mathscr{D})}} .$$

(5) Let $r \in \mathbb{N}$, then we define functions spaces

$$H^r(\mathscr{D}) \coloneqq W^{r,2}(\mathscr{D}) ,$$
$$H^r_0(\mathscr{D}) \coloneqq W^{r,2}_0(\mathscr{D}) ,$$

and a scalar product

$$\langle u, v\rangle_{H^r(\mathscr{D})} \coloneqq \sqrt{\sum_{0\le|\boldsymbol{\alpha}|\le r} \langle \partial^{\boldsymbol{\alpha}} u, \partial^{\boldsymbol{\alpha}} v\rangle_{L^2(\mathscr{D})}} .$$

Note that $W^{m,p}(\mathscr{D})$ can also be defined as the completion

$$W^{m,p}(\mathscr{D}) = \overline{\left\{ f \in C^m(\mathscr{D}) \,:\, \|f\|_{W^{m,p}} < \infty \right\}}^{\|\cdot\|_{W^{m,p}(\mathscr{D})}} .$$

The equivalence of the two definitions was proven in [**64**]. Also note that

$$W^{0,p}(\mathscr{D}) = L^p(\mathscr{D})$$

and for $p < \infty$

$$W^{0,p}_0(\mathscr{D}) = L^p(\mathscr{D}) .$$

Thus, the following inclusions hold

$$W^{m,p}_0(\mathscr{D}) \subset W^{m,p}(\mathscr{D}) \subset L^2(\mathscr{D}) . \tag{II.17}$$

The following result makes Sobolev spaces useful as function spaces.

THEOREM II.61 ([**1**, Theorem 3.3 & 3.6]).

(1) $(W^{m,p}(\mathscr{D}), \|\cdot\|_{W^{m,p}(\mathscr{D})})$ *are Banach spaces, that are separable if* $1 \le p < \infty, m \in \mathbb{N}$.
(2) $(H^r(\mathscr{D}), \langle\cdot,\cdot\rangle_{H^r(\mathscr{D})})$ *is a separable and reflexive Hilbert space for* $r \in \mathbb{N}$.

The above results extend to the closed subspaces $W^{m,p}_0(\mathscr{D})$ and $H^r_0(\mathscr{D})$.

LEMMA II.62 ([**4**, Theorem 10.9]).

(1) *For a bounded domain* $\mathscr{D} \subset \mathbb{R}^n$ *with Lipschitz boundary,* $\mathrm{Id} : H^1(\mathscr{D}) \to L^2(\mathscr{D})$ *is a compact embedding.*

[2]The boundary value is meant in the sense of the *boundary trace.* For the purpose of this introduction we omit traces and refer the interested reader to the result [**1**, 5.37].

(2) *For a bounded domain* $\mathscr{D} \subset \mathbb{R}^n$, $\mathrm{Id} : H_0^1(\mathscr{D}) \to L^2(\mathscr{D})$ *is a compact embedding.*

$C_0^\infty(\mathscr{D})$ is dense in $L^p(\mathscr{D})$ for $p < \infty$, cf. [**1**, Corollary 2.30]. This means that $H_0^1(\mathscr{D})$ must also be dense in $L^2(\mathscr{D})$ and due to (II.17) the same is true for $H^1(\mathscr{D})$. Thus, we can construct Gelfand triples (theorem II.30) of compact dense embeddings

$$H^1(\mathscr{D}) \overset{\mathrm{Id}}{\hookrightarrow} L^2(\mathscr{D}) \overset{R_{L^2(\mathscr{D})}}{\to} (L^2(\mathscr{D}))' \overset{\mathrm{Id}'}{\hookrightarrow} (H^1(\mathscr{D}))' , \tag{II.18a}$$

$$H_0^1(\mathscr{D}) \overset{\mathrm{Id}}{\hookrightarrow} L^2(\mathscr{D}) \overset{R_{L^2(\mathscr{D})}}{\to} (L^2(\mathscr{D}))' \overset{\mathrm{Id}'}{\hookrightarrow} (H_0^1(\mathscr{D}))' . \tag{II.18b}$$

II.5. Spectral Theory

We focus on the spectral theory for compact normal operators, following [**4**, Chapters 11 & 12], and recall how this spectral theory also informs variational EVPs, where a compact normal solution operator can be constructed.
For the definition of the spectrum, we consider a Banach space X over $\mathbb{C}$ and an operator $T \in \mathscr{L}(X)$. Without further assumptions on T, we cannot guarantee that its eigenvalues are real, even if the operator is defined between real Banach spaces X. For operators which are defined on real Banach spaces, we may consider their complexification instead.

DEFINITION II.63 ([**4**, 11.14]). Let X be a Banach space over $\mathbb{R}$ and $T \in \mathscr{K}(X)$, then its **complexification** is a complex Banach space $\tilde{X} = X \times X$ and for $x = (\Re(x), \Im(x)) \in \tilde{X}$, $a \in \mathbb{C}$ define

$$ax := (\Re(a)\Re(x) - \Im(a)\Im(x), \Re(a)\Im(x) + \Im(a)\Re(x)) , \qquad \overline{x} := (\Re(x), -\Im(x)) .$$

That is, we have represented complex vector elements as real tuples. The operator is translated into

$$\tilde{T} := (T\Re(x), T\Im(x)) \in \mathscr{K}(\tilde{X}) .$$

The formulation of spectral theory requires the spectrum and in particular the point spectrum.

DEFINITION II.64 ([**4**, 11.1]). Let X be a Banach space over $\mathbb{C}$. We define the **resolvent set** of $T \in \mathscr{L}(X)$ by

$$\varrho(T) := \{\lambda \in \mathbb{C} : \mathscr{N}(\lambda \mathrm{Id} - T) = \{0\} \text{ and } \mathscr{R}(\lambda \mathrm{Id} - T) = X\}$$

and the **spectrum** of T by $\mathrm{Spec}(T) := \mathbb{C} \setminus \varrho(T)$. The **point spectrum** is a subset of the spectrum defined by

$$\mathrm{Spec}_p(T) := \left\{\lambda \in \mathrm{Spec}(T) : \mathscr{N}(\lambda \mathrm{Id} - T) \neq \{0\}\right\} .$$

DEFINITION II.65 ([**4**, 11.2(2)]). Let X be a Banach space and $T \in \mathscr{L}(X)$. Then $\lambda \in \mathrm{Spec}_p(T)$ is equivalent to

$$\exists x \neq 0 \text{ such that } Tx = x\,\lambda \text{ with } \lambda \in \mathbb{K} .$$

Such a $\lambda \in \mathbb{K}$ is then called an **eigenvalue** and $x \in X$ an **eigenvector** of T. If X is a function space, then $x \in X$ is also called an **eigenfunction**. The pair $(\lambda, x) \in \mathbb{K} \times X$ is called an **eigenpair**. The subspace $\mathcal{N}(\lambda \mathrm{Id} - T)$ is the **eigenspace** of T corresponding to λ, its dimension $m = \dim(\mathcal{N}(\lambda \mathrm{Id} - T))$ is called the **multiplicity** of eigenvalue λ. The eigenspace is a T-invariant subspace, i.e.,

$$T(\mathcal{N}(\lambda \mathrm{Id} - T)) \subset \mathcal{N}(\lambda \mathrm{Id} - T) .$$

If the multiplicity of an eigenvalue λ is $m > 1$, we call the eigenvalue **degenerate**.

REMARK II.66.

(1) If an eigenvalue is real-valued, i.e., $\lambda \in \mathbb{R}$, then its eigenfunction $x \in X$ can be chosen to be real-valued, such that $\overline{x} = x$, cf. [**4**, 11.14]. Thus, if all eigenvalues are real, we can ignore the complexification of definition II.63 and consider the original operator $T \in \mathcal{K}(X)$ on a real Hilbert space X, cf. [**4**, 11.14].
(2) For a $\mathbb{K}$-valued Banach space X, the eigenvectors of a non-degenerate eigenvalue are unique up to a factor $c \in \mathbb{K}$, i.e., if u is an eigenfunction of eigenvalue λ, then so is cu. If the Hilbert space X is real, this means that the eigenvectors of a non-degenerate eigenvalue are unique up to sign and scaling.
(3) To reduce ambiguity, it is generally assumed that the eigenfunctions are normalized with respect to the norm of X. For real Banach spaces X, only the sign is arbitrary, while in the complex case the eigenfunctions can still be arbitrarily scaled by a factor $c \in \mathbb{C}$ with $|c| = 1$. For eigenvectors $(u_i)_{i=1,\dots,m}$ of a degenerate eigenvalue λ of multiplicity m any linear combination of the eigenfunctions is again an eigenfunction of λ, i.e.,

$$T\left(\sum_{i=1}^{m} c_i u_i\right) = \left(\sum_{i=1}^{m} c_i u_i\right) \lambda , \qquad c_i \in \mathbb{K} .$$

We assume that eigenfunctions within an eigenspace $(u_i)_{i=1,\dots,m}$ are chosen orthogonally on top of all eigenvalues being normalized, i.e., orthonormal.

We recall the spectral theorem for compact normal operators. The definition of normal operators requires us to upgrade X to a (complex) Hilbert space.

THEOREM II.67 (**Spectral theorem for compact normal operators**, [**4**, 12.12]). *If X is a Hilbert space over $\mathbb{C}$, $T \in \mathcal{K}(X)$ normal, $T \neq 0$, then:*

(1) *There exists an orthonormal system $(u_i)_{i \in N}$ in X and a sequence $(\lambda_i)_{i \in N}$ in $\mathbb{C}$ with $N \subset \mathbb{N} \cup \{\infty\}$ such that $\lambda_i \neq 0$ and*

$$Tu_i = u_i \lambda_i \text{ for } i \in N, \qquad \mathrm{Spec}(T) \setminus \{0\} = \{\lambda_i \,:\, i \in N\} .$$

Thus, the numbers λ_i are the nonzero eigenvalues of T with eigenfunctions u_i. If $N = \infty$, then $\lambda_i \to 0$ as $i \to \infty$.
(2) *For all eigenvalues λ holds $\mathcal{N}((\lambda \mathrm{Id} - T)^2) = \mathcal{N}(\lambda \mathrm{Id} - T)$.*
(3) $X = \mathcal{N}(T) \oplus \overline{\mathrm{span}\{u_i \,:\, i \in N\}}$.

(4) *The elements of the image of T can be decomposed as*

$$Tx = \sum_{i \in N} \lambda_k \langle x, u_i \rangle_X u_i \qquad \forall x \in X. \tag{II.19}$$

If the set of eigenfunctions form a basis of X, cf. definition II.7, we may call it an **eigenbasis** of T.
The adjoint operator to a compact normal operator T can be expressed as

$$T^* x = \sum_{i \in N} \overline{\lambda_k} \langle x, u_i \rangle_X u_i \qquad \forall x \in X\,.$$

Therefore, the eigenvalues of self-adjoint operators are real. Thus, a self-adjoint operator has an ordered sequence of real eigenvalues $(\lambda_i)_{i \in \mathbb{N}}$.
The difference between successive eigenvalues is then called an **(eigen)gap**. For a compact normal operator, all eigenvalues apart from an eigenvalue $\lambda \neq 0$ have a positive gap to its neighbors. Therefore, we can call them **isolated**, counting identical eigenvalues as one degenerate eigenvalue. For compact normal operators, an eigenvalue $\lambda = 0$ can possibly be an cluster point, so we cannot identify a closest neighboring eigenvalue, which makes it the only possible non-isolated eigenvalue in this setting.

II.5.1. Generalized Eigenvalue Problems. EVPs can be defined in a broader sense than that used in theorem II.67.

DEFINITION II.68. Let X be a Hilbert space, $K : X \to X$ be a linear operator, and $M : X \to X$ a bounded linear, elliptic, self-adjoint operator. Then the problem to find an **eigenpair** $(\lambda, u) \in \mathbb{K} \times X, u \neq 0$, such that

$$Ku = Mu\,\lambda\,, \tag{II.20a}$$

is called a **generalized operator EVP** with **eigenvalue** $\lambda \in \mathbb{K}$ and **eigenfunction** $u \in X$. Compared to the previous formulation Id is replaced by M, e.g., the **eigenspace** of the eigenvalue λ is $\mathcal{N}(\lambda M - K)$ and the multiplicity is again the dimension of the eigenspace. Eigenfunctions are called **orthonormal** if they are with respect to the scalar product induced by M, i.e.,

$$\left\langle Mu_i, u_j \right\rangle_X = \delta_{ij}\,, \qquad i, j \in 1, \ldots, m\,, \tag{II.20b}$$

where m is the multiplicity of eigenvalue λ and $(u_i)_{i=1,\ldots,m}$ its eigenfunctions.

REMARK II.69. Given two Hilbert spaces $V \subset H$ with $\mathrm{Id} : V \hookrightarrow H$ a dense embedding, we can replace K by a linear operator $\tilde{K} : V \to V'$ and M by an elliptic self-adjoint operator $\tilde{M} \in \mathscr{L}(H; H')$. The Hilbert spaces form a Gelfand triple, cf. theorem II.30,

$$V \overset{\mathrm{Id}}{\hookrightarrow} H \overset{R_H}{\to} H' \overset{\mathrm{Id}'}{\hookrightarrow} V'\,,$$

Then, $\tilde{K}u = \tilde{M}u \cdot \lambda$ can be considered an EVP in V'.

We can translate a generalized EVP into the simple EVP of definition II.65, cf. [**59**, Chapter VII.6]. The most straightforward formulation of a simple EVP is by inversion of M, otherwise decompositions of the inverse of M can be used. Since in general the combined operator is often not even bounded, i.e., theorem II.67 is not (directly) applicable. We briefly turn to matrix EVPs before discussing the solution operator.

II.5.2. Matrix Eigenvalue Problems. The special case in which the operators are matrices is relevant to numerical approximation. Recall that matrices are compact operators, cf. lemma II.19.

DEFINITION II.70. Let $K \in \mathbb{K}^{n\times n}$ be a matrix and $M \in \mathbb{K}^{n\times n}$ a self-adjoint positive definite matrix. Then the **generalized matrix EVP** is to find an **eigenpair** $(\lambda, u) \in \mathbb{K} \times \mathbb{K}^n, u \neq 0$, such that

$$Ku = Mu\,\lambda\,. \tag{II.21a}$$

It is comprised of **eigenvalue** λ and its **eigenvector** u, which is called **orthonormal** if

$$u_j^* M u_i = \delta_{ij}\,, \qquad i,j = 1,\dots,m\,. \tag{II.21b}$$

In case $M = \boldsymbol{I}$, the eigenpairs are called the eigenpairs (eigenvalues, eigenfunctions) of the matrix K.

The following result is helpful for estimating the eigenvalues of a block-diagonal matrix.

THEOREM II.71. [**40**, 7.2.1] *Let $A \in \mathbb{R}^{n\times n}$ with $X^{-1}\cdot A\cdot X = D + F$, where $X \in \mathbb{R}^{n\times n}$ is invertible, $D = \operatorname{diag}_{i=1,\dots,n} d_i$, and F has zeros as diagonal entries. Then for the eigenvalues of matrix A holds $(\lambda_i)_{i=1,\dots,n} \subset \bigcup_{i=1}^n D_i$, where*

$$D_i := \left\{ z \in \mathbb{C} \,:\, |z - d_i| \le \sum_{j=1}^{n} |f_{ij}| \right\}.$$

*The boundaries of D_i are called **Gershgorin circles**.*

See also [**40**, 8.1.3] for a version for real symmetric matrices.

II.5.3. Variational Eigenvalue Problem. We now turn to the variational formulation of EVPs. The variational formulation has the advantage that it is a convenient starting point to formulate a discretization, however, some explanation is due to find the underlying compact normal solution operator.

DEFINITION II.72. Let $V \subset H$ be Hilbert spaces with dense embedding $\mathrm{Id} : V \hookrightarrow H$ and

$a : V \times V \to \mathbb{K}$ a bounded sesquilinear form,

$b : H \times H \to \mathbb{K}$ a scalar product of H.

Then the problem to find an **eigenpair** $(\lambda, u) \in \mathbb{K} \times V, u \neq 0$, such that

$$a(u,v) = b(u,v)\,\lambda \qquad \forall\, v \in V \tag{II.22a}$$

is called the **variational EVP** with $\lambda \in \mathbb{K}$ the **eigenvalue** and $u \in V$ its **eigenfunction**. The definitions of spectral theory are adapted to the scalar product b, e.g., the **eigenspace** of λ generalizes to

$$\mathcal{N}\left(a(\cdot, v) - b(\cdot, v)\,\lambda\right)$$

and we call the eigenfunctions **orthonormal** if

$$b(u_i, u_j) = \delta_{ij} \qquad i, j = 1, \dots, m\,. \tag{II.22b}$$

If V, H are real Hilbert spaces, then the sesquilinear forms a, b become bilinear forms and the scalar product b is symmetric. A complexification (cf. definition II.63) can in turn be found by conjugating the second argument of the bilinear forms.
Let sesquilinear forms be defined for the EVP according to definition II.72, then we can construct linear operators

$$K \in \mathscr{L}(V)\,, \qquad M \in \mathscr{L}(H)\,,$$

using the Lax–Milgram theorem (theorem II.26, remark II.27), i.e.,

$$\langle u, Kv\rangle_V := a(u, v) \qquad \forall\, u, v \in V\,, \tag{II.23a}$$

$$\langle u, Mv\rangle_H := b(u, v) \qquad \forall\, u, v \in H\,, \tag{II.23b}$$

with $M \in \mathscr{L}(H)$ b-elliptic, self-adjoint. If $b \in \mathscr{L}^{(1.5)}(H; \mathbb{K})$ is the standard scalar product of H, then $M = \mathrm{Id}$. The following conclusions (remark II.27) are helpful:

(1) If the sesquilinear form is continuous, so is the operator.
(2) If the sesquilinear form is elliptic, so is the operator, and thus it is bounded invertible.
(3) If the sesquilinear form is Hermitian, so is the operator.

Of course, all of these implications apply to the standard scalar product b. An eigenpair $(\lambda, u) \in \mathbb{K} \times V$ of definition II.68 relates to an eigenpair $(\overline{\lambda}, u)$ of definition II.72 since

$$a(u, u) = \langle u, Ku\rangle_V = \langle u, Mu\,\lambda\rangle_H = b(u, u\,\lambda) = b(u, u)\,\overline{\lambda}\,.$$

In the self-adjoint case, we have $\lambda \in \mathbb{R}$, so the conjugation of the eigenvalues has no effect. Note that an operator $\tilde{K} : V \to V'$ according to remark II.27 is given by

$$v \mapsto \langle \cdot, Kv\rangle_V \in V'\,. \tag{II.24}$$

REMARK II.73. Matrices $K \in \mathbb{K}^{n\times n}$ can be translated into bilinear forms or sesquilinear forms

$$u \times v \mapsto a(u, v) := v^\top K u \in \mathscr{L}^{(2)}(\mathbb{R}^n; \mathbb{R}) \qquad \text{for } K \in \mathbb{R}^{n\times n}\,,$$

$$u \times v \mapsto a(u, v) := v^* K u \in \mathscr{L}^{(1.5)}(\mathbb{C}^n; \mathbb{C}) \qquad \text{for } K \in \mathbb{C}^{n\times n}\,.$$

This converts a matrix EVP into a variational EVP.

Solution Operators for Elliptic Eigenvalue Problems. So far we have described the relationship between general variational and operator EVPs in terms of the relationship of sesquilinear forms and associated operators. In order to apply the spectral theory of compact normal operators to a variational EVP, we need to reduce the two sesquilinear forms to a single operator. Given some stronger assumptions about the sesquilinear forms, we can construct a compact normal solution operator $T \in \mathscr{K}(H)$.

ASSUMPTION II.74. For the construction of the solution operator, let

(II.25a) $a \in \mathscr{L}^{(1.5)}(V;\mathbb{C})$ a continuous, V-elliptic, Hermitian sesquilinear form,

(II.25b) $b \in \mathscr{L}^{(1.5)}(H;\mathbb{C})$ an H-elliptic[3] scalar product of H,

(II.25c) $\mathrm{Id} : V \hookrightarrow H$ a compact dense embedding with Hilbert spaces $V \subset H$.

For real Hilbert spaces V, H the sesquilinear forms (II.25a) and (II.25b) can again be simplified to

$a \in \mathscr{L}^{(2)}(V;\mathbb{R})$ a continuous, V-elliptic, symmetric bilinear form,

$b \in \mathscr{L}^{(2)}(H;\mathbb{R})$ an H-elliptic[3] scalar product of H.

Due to the Lax–Milgram theorem, we can construct a **solution operator** $T : H \to H$ such that

$$a\big(Tf, v\big) = b\big(f, v\big) \qquad\qquad v \in V$$

holds for all $f \in H$ such that $Tf \in V$, cf. [**9**, Chapter 7]. We define operators as in (II.23) and get a V-elliptic, self-adjoint operator $K \in \mathscr{L}(V)$ and an H-elliptic, self-adjoint operator $M \in \mathscr{L}(H)$. According to (II.24), we define $\tilde{K} \in \mathscr{L}(V;V')$, which we can invert due to V-ellipticity, to get $\tilde{K}^{-1} \in \mathscr{L}(V';V)$. Since $\mathrm{Id} : V \hookrightarrow H$ is a compact embedding, we can define $\check{K}^{-1} \in \mathscr{K}(H)$ an equivalent compact, self-adjoint operator using lemma II.20, i.e.,

$$\check{K}^{-1} : H \overset{R_H}{\to} H' \overset{\mathrm{Id}'}{\hookrightarrow} V' \overset{\tilde{K}^{-1}}{\to} V \overset{\mathrm{Id}}{\hookrightarrow} H\,. \tag{II.26}$$

Thus, we can construct its inverse square root $\check{K}^{-\frac{1}{2}} \in \mathscr{K}(H)$ via spectral theorem II.67, replacing the eigenvalues of $\check{K}^{-1}$ in a decomposition of the form (II.19) by the square root of the same eigenvalues. Then for $M \in \mathscr{L}(H)$ as in (II.23b), a compact, self-adjoint solution operator is given by

$$T := \check{K}^{-\frac{1}{2}} M \check{K}^{-\frac{1}{2}} \in \mathscr{K}(H)\,. \tag{II.27}$$

The compactness of the solution operator T is again a consequence of lemma II.20. Due to the symmetric construction by square roots the operator is also self-adjoint, i.e.,

$$T^* = (\check{K}^{-\frac{1}{2}} M \check{K}^{-\frac{1}{2}})^* = (\check{K}^{-\frac{1}{2}})^* M^* (\check{K}^{-\frac{1}{2}})^* = \check{K}^{-\frac{1}{2}} M \check{K}^{-\frac{1}{2}} = T\,,$$

[3]The *standard* scalar product of H is H-elliptic, with constant $c_0 = 1$.

such that the spectral theorem II.67 applies. We can check that $Tf \in V$ for $f \in H$ since in (II.26) the last embedding $\mathrm{Id} : V \hookrightarrow H$ was only applied in order to return to H.
Each eigenpair of the solution operator T is given by $(\lambda^{-1}, \check{K}^{\frac{1}{2}} u)$, where $(\lambda, u) \in \mathbb{R} \times V$ is an eigenpair of (II.20), since

$$w \coloneqq \check{K}^{\frac{1}{2}} u = \check{K}^{-\frac{1}{2}} \check{K} u = \check{K}^{-\frac{1}{2}} M u \cdot \lambda = \check{K}^{-\frac{1}{2}} M \check{K}^{-\frac{1}{2}} w \cdot \lambda = Tw \cdot \lambda \,.$$

Thus, under assumption II.74 according to theorem II.67, we know that the reciprocals of the eigenvalues of (II.22) are the eigenvalues of the solution operator, which are bounded and accumulate at zero if $\dim(V) = \infty$, i.e.,

$$c^{-1} \geq \lambda_n^{-1} \to 0 \qquad\qquad n \to \infty \,,$$

In turn, if $\dim(V) = \infty$, the actual eigenvalues of (II.22) are bounded from below by the constant $c > 0$ and diverge, i.e.,

$$0 < c \leq \lambda_n \to \infty \qquad\qquad n \to \infty \,,$$

where eigenvalues $(\lambda_n)_{n \in \mathbb{N}}$ are counted multiple times according to their finite multiplicity. We infer that for infinite-dimensional Hilbert spaces V, H, a potential operator operator T^{-1} expressing the variational EVP as single-operator EVP is not bounded, i.e., $T^{-1} \notin \mathscr{L}(V)$ and $T^{-1} \notin \mathscr{L}(H)$.

REMARK II.75. We can alternatively construct a non-self-adjoint solution operator

$$S \coloneqq \check{K}^{-1} M \in \mathscr{K}(H) \tag{II.28}$$

which has eigenvalues (λ^{-1}, u), cf. [**4**, 10.14]. The conjugation of two self-adjoint operators is not necessarily self-adjoint itself. Self-adjointness must then be proven separately, cf. [**4**, 12.16]. For our proofs, it is convenient to first consider a self-adjoint solution operator (II.27) to find the eigenvalues using theorem II.67 and then consider a solution operator (II.28) with the same eigenvalues but simpler eigenfunctions.

Vector Notation. We now make use of the vector notation mentioned in section II.1, which is convenient when analyzing whole eigenspaces in later chapters. In analogy to the sesquilinear forms in definition II.72 and by abuse of notation, we define

$$\begin{aligned} a &: V^m \times V^m \to \mathbb{K}^{m \times m} \,, \qquad \boldsymbol{u} \times \boldsymbol{v} \mapsto a(\boldsymbol{u}, \boldsymbol{v}) \,, \qquad a(\boldsymbol{u}, \boldsymbol{v}) \coloneqq [a(u_j, v_i)]_{i,j=1}^m \,, \\ b &: H^m \times H^m \to \mathbb{K}^{m \times m} \,, \qquad \boldsymbol{u} \times \boldsymbol{v} \mapsto b(\boldsymbol{u}, \boldsymbol{v}) \,, \qquad b(\boldsymbol{u}, \boldsymbol{v}) \coloneqq [b(u_j, v_i)]_{i,j=1}^m \,. \end{aligned}$$

The variational EVP with orthonormal eigenfunctions can then be written in bold vector notation as the problem to find $(\boldsymbol{\lambda}, \boldsymbol{u}) \in \mathbb{K}^{m \times m} \times V^m$, such that

$$a(\boldsymbol{u}, \boldsymbol{v}) = b(\boldsymbol{u}, \boldsymbol{v}) \cdot \boldsymbol{\lambda} \qquad\qquad \forall \boldsymbol{v} \in V^m \,, \tag{II.29a}$$

$$b(\boldsymbol{u}, \boldsymbol{u}) = \boldsymbol{I} \,, \tag{II.29b}$$

where $\boldsymbol{\lambda}$ is diagonal, i.e., $\boldsymbol{\lambda} = \lambda \boldsymbol{I}$, and all eigenfunctions belong to the same eigenspace of $\lambda \in \mathbb{K}$.

II.5.4. Eigenvalue Problems as a Minimization Problems. Given assumption II.74, under which we have constructed the solution operator, the eigenpair can also be characterized as a minimization problem.

THEOREM II.76 (**Courant–Fischer theorem**, **Rayleigh quotient**, [**9**, eq. (7.3)]). *Given assumption II.74, the eigenvalues $\lambda_n \in \mathbb{R}$ and their eigenfunctions $u_n \in V$, $n \in \mathbb{N}$, admit the representation*

$$\lambda_1 = \min_{v \in V} \frac{a(v,v)}{b(v,v)}\,, \qquad u_1 = \operatorname*{arg\,min}_{v \in V} \frac{a(v,v)}{b(v,v)}\,,$$

$$\lambda_n = \min_{v \in \left(\bigoplus_{i=1}^{n-1} \operatorname{span}(u_i)\right)^\perp} \frac{a(v,v)}{b(v,v)}\,, \qquad u_n = \operatorname*{arg\,min}_{v \in \left(\bigoplus_{i=1}^{n-1} \operatorname{span}(u_i)\right)^\perp} \frac{a(v,v)}{b(v,v)}\,.$$

Of course, if assumption II.74 does not hold, we can still represent the eigenvalues by quotients, when the eigenfunctions are already known. If the eigenfunctions are normed according to b, the denominator vanishes.

II.5.5. Discretization and Solution of Eigenvalue Problems. The results of this thesis can be seen as independent of the discretization scheme. However, at this time, we want to give a short description of a typical implementation, as well as some remarks on solvers for matrix EVPs. Let $V_h \subset V$ be a finite-dimensional approximation space of a real Hilbert space V spanned by basis functions $(\varphi_i)_{i=1,\dots,n}$. For a finite element (FE) discretization of the bilinear forms, we define two matrices

$$\underline{K} = \left[a(\varphi_j, \varphi_i)\right]_{i,j=1}^n\,, \qquad \underline{M} = \left[b(\varphi_j, \varphi_i)\right]_{i,j=1}^n\,,$$

where $\underline{K} \in \mathbb{R}^{n\times n}$ is called the **stiffness matrix** and $\underline{M} \in \mathbb{R}^{n\times n}$ the **mass matrix**. The corresponding matrix EVP is then to find $(\underline{\lambda}, \underline{u}) \in \mathbb{K} \times \mathbb{K}^n$, such that

$$\underline{K}\,\underline{u} = \underline{M}\,\underline{u}\,\underline{\lambda}\,. \tag{II.30}$$

The eigenfunctions $u_h \in V_h$ are given by $u_h = \sum_{i=1}^n [\underline{u}]_i \varphi_i$ with eigenvalues $\lambda_h = \underline{\lambda}$. Usual choices of basis functions $(\varphi_i)_{i=1,\dots,n}$ lead to sparse matrices $\underline{K}, \underline{M} \in \mathbb{K}^{n\times n}$.
Classical algorithms to solve matrix EVP (II.30) approximately are the *power iteration* [**40**, Chapter 8.4], *QR iteration* [**40**, Chapter 8.3], and the *(restarted) Arnoldi iteration* [**40**, Chapter 10.5.2]. The QR iteration is feasible only on relatively small matrices. The Arnoldi iteration can make use of sparsity when only calculating some eigenvalues. It is possible to find the eigenvalues with the largest absolute value or, by inversion, the smallest.
If we are looking for specific eigenvalues of which we already have a good estimate, we can look for the eigenvalues with the smallest absolute difference to this estimate. Let $c \in \mathbb{K}$ be our educated guess of eigenvalue $\lambda \in \mathbb{K}$. The eigenpairs of the **shifted EVP** with shift $c \in \mathbb{K}$

$$(\underline{K} - c\underline{M})\underline{u} = \underline{M}\,\underline{u}\,(\underline{\lambda} - c)$$

are $(\underline{\lambda} - c, \underline{u}) \in \mathbb{K} \times \mathbb{K}^n$, where the $(\underline{\lambda}, \underline{u})$ are the eigenpairs of (II.30).
For EVPs with real symmetric matrices, the convergence rate of the eigenvalues is typically double the convergence rate of the eigenfunctions with respect to the mesh size $h \to 0$, cf. [**9**, Remark 2.1]. The interested reader is referred to [**9**, Chapter 9] for more results on the convergence of FE approximations of the EVPs considered later.

II.5.6. Parameterized Eigenvalue Problems. The objective of this thesis is uncertainty quantification of the eigenpairs of parameterized EVPs. So far, we have established EVPs in operator and variational form (II.22) without parameters. We now formulate a parameterized version of the variational EVP.
Let X be a Banach space for the parameters, $B \subset X$ an open subset, and consider parameterized sesquilinear forms

$$x \mapsto a(\cdot,\cdot;x) \in \mathscr{L}^{(1.5)}(V;\mathbb{K}) , \qquad x \in B ,$$
$$x \mapsto b(\cdot,\cdot;x) \in \mathscr{L}^{(1.5)}(H;\mathbb{K}) , \qquad x \in B .$$

Here, the parameter $x \in B$ determines the two sesquilinear forms such that assumption II.74 holds uniformly for all $x \in B$. This means that there are constants for continuity and ellipticity which hold for all $x \in B$. For the parameterized scalar product b H-ellipticity is no longer trivial, as it is no longer the standard scalar product.
We can state a **parameterized** variational EVP to find an eigenpair $(\lambda_x, u_x) \in \mathbb{R} \times V$, such that

$$a(u_x, v; x) = b(u_x, v; x)\, \lambda_x \qquad \forall v \in V , \tag{II.31a}$$
$$b((u_x)_i, (u_x)_j; x) = \delta_{ij} \qquad i, j = 1, \dots, m . \tag{II.31b}$$

We first use the index x naively to indicate that an eigenpair originates from a certain parameter $x \in B$ and keep the counting indices outside the bracket. Note that the test function $v \in V$ does not depend on the parameter $x \in B$.
Formulating functions

$$B \to \mathbb{R} , \qquad x \mapsto \lambda_x ,$$
$$B \to V , \qquad x \mapsto u_x ,$$

consistently is not trivial, and sometimes not even possible. However, the simpler case of $\dim(B) = 1$ is already well understood. Assuming that the dependence of the parameterized bilinear forms (or equivalent parameterized operators) is analytic, the eigenvalue and eigenfunctions can also be expressed as (locally) analytic functions, cf. [**59, 83**]. We have used this in the pathwise tracking in example I.2. Note that continuity of the eigenfunctions often enforces a specific choice of basis at points where the eigenfunctions are in a degenerate eigenspace. We continue this discussion in more detail in chapter III.

II.6. Examples of Variational Eigenvalue Problems

In this section, we introduce the Laplace and Maxwell EVP as instances of variational EVPs, which we use as examples throughout the rest of this thesis and for which assumption II.74 holds. There are more examples of variational EVP that fit assumption II.74, e.g. the shell EVP, cf. [**44, 99**]. We formulate each of our examples without parameterization in order not to clutter the notation. A parameterized version, as discussed previously in section II.5.6, is straightforward and can be expressed through parameterized material coefficients and shape deformations, both of which we encounter later in more concrete numerical examples. In our examples, we assume that the domain $\mathscr{D} \subset \mathbb{R}^n$ is open, bounded, simply connected, and Lipschitz.

II.6.1. The Laplace Eigenvalue Problem. As a simple example of a variational EVP, we consider the Laplace EVP with (zero) Dirichlet boundary data, i.e.,

$$\begin{aligned} -\Delta u(\mathbf{x}) &= u(\mathbf{x})\, \lambda && \mathbf{x} \in \mathscr{D}\,, \\ u(\mathbf{x}) &= 0 && \mathbf{x} \in \partial\mathscr{D}\,. \end{aligned}$$

The **Laplace operator** can be defined as

$$\Delta u(\mathbf{x}) := \operatorname{div}(\operatorname{grad} u(\mathbf{x})) = \sum_{i=1}^{n} \frac{\partial^2 u(\mathbf{x})}{\partial \mathbf{x}_i^2}$$

with the **gradient** $\operatorname{grad} u(\mathbf{x})$ and the **divergence** $\operatorname{div}(u(\mathbf{x}))$ defined by

$$\operatorname{grad} u(\mathbf{x}) = \begin{bmatrix} \frac{\partial u(\mathbf{x})}{\partial \mathbf{x}_1} \\ \vdots \\ \frac{\partial u(\mathbf{x})}{\partial \mathbf{x}_n} \end{bmatrix}, \qquad \operatorname{div}(u(\mathbf{x})) := \sum_{i=1}^{n} \frac{\partial u_i(\mathbf{x})}{\partial \mathbf{x}_i}\,.$$

Note that we already considered the Laplace operator in example I.1. The variational form of the Laplace EVP can be found by partial integration, cf. [**4**, 6.5(1)]. We can define the Laplace EVP concisely for the framework of section II.5 in terms of bilinear forms on real Hilbert spaces.

EXAMPLE II.77 (**Laplace EVP**). For Dirichlet boundary conditions consider the function spaces

$$V = H_0^1(\mathscr{D})\,, \qquad H = L^2(\mathscr{D})$$

with $\mathscr{D} \subset \mathbb{R}^n$ domain $\mathscr{D}$ is open, bounded, simply connected, and Lipschitz. Then the **Laplace EVP in variational form** (II.22) is given by the bilinear forms

$$a \in \mathscr{L}^{(2)}(V;\mathbb{R})\,, \qquad a\big(u,v\big) = \int_{\mathscr{D}} \big\langle \operatorname{grad} u(\mathbf{x}), \operatorname{grad} v(\mathbf{x}) \big\rangle_{\mathbb{R}^n}\, \mathrm{d}\mathbf{x}\,, \tag{II.32a}$$

$$b \in \mathscr{L}^{(2)}(H;\mathbb{R})\,, \qquad b\big(u,v\big) = \int_{\mathscr{D}} u(\mathbf{x})\, v(\mathbf{x})\, \mathrm{d}\mathbf{x}\,. \tag{II.32b}$$

The V-ellipticity of $a \in \mathscr{L}(V;\mathbb{R})$ is a result of the Poincaré inequality, cf. [**4**, 6.7].

Finite Element Implementation. To discretize the Laplace EVP, we may use standard FEs for the Laplace operator. Our examples will use triangular meshes and piecewise linear basis functions $(\varphi_i)_{i=1,\dots,n}$ with $\varphi_i \in H^1(\mathscr{D})$. The stiffness and mass matrix can then be calculated and the discretized problem solved as described in section II.5.5. The reader may find an introduction to FE methods in [**10**].
As an example of the Laplace EVP the four smallest eigenvalues and the associated eigenfunctions of the Laplace operator with zero Dirichlet boundary data for the unit square $\mathscr{D} = (0,1)^2$ can be seen in fig. II.1 on a mesh using 545 nodes. The second and third eigenpairs are an example of a degenerate eigenvalue, so we may also accept any linear combination of the depicted eigenfunctions, which were aligned to the coordinate axes.

II.6.2. The Maxwell Eigenvalue Problem. A further relevant example that we want to discuss is the Maxwell EVP. To this end, we first need to introduce some more operators and vector spaces. The summary here will stick to the EVP, which is derived from the *time-harmonic* Maxwell equations. The reader interested in more details is referred to [**51, 67**] for a more thorough discussion of Maxwell's equations and to [**9**] for a discussion focusing on the EVP.
To define the Maxwell EVP for a domain $\mathscr{D} \subset \mathbb{R}^3$, we need to introduce the **curl operator**

$$\mathbf{curl}(u(\mathbf{x})) \coloneqq \begin{bmatrix} \frac{\partial u_3(\mathbf{x})}{\partial \mathbf{x}_2} - \frac{\partial u_2(\mathbf{x})}{\partial \mathbf{x}_3} \\ \frac{\partial u_1(\mathbf{x})}{\partial \mathbf{x}_3} - \frac{\partial u_3(\mathbf{x})}{\partial \mathbf{x}_1} \\ \frac{\partial u_2(\mathbf{x})}{\partial \mathbf{x}_1} - \frac{\partial u_1(\mathbf{x})}{\partial \mathbf{x}_2} \end{bmatrix} .$$

Consider the **time-harmonic Maxwell equations** with angular frequency $\omega > 0$

$$\mathbf{curl}(E(t,\mathbf{x})) + \mathrm{i}\omega\mu(\mathbf{x})\, H(t,\mathbf{x}) = 0 \qquad (t,\mathbf{x}) \in (0,\infty)\times\mathscr{D}\,, \tag{II.33a}$$

$$\mathbf{curl}(H(t,\mathbf{x})) - \mathrm{i}\omega\varepsilon(\mathbf{x})\, E(t,\mathbf{x}) = 0 \qquad (t,\mathbf{x}) \in (0,\infty)\times\mathscr{D}\,, \tag{II.33b}$$

$$\operatorname{div}(\mu(\mathbf{x})\, H(t,\mathbf{x})) = 0 \qquad (t,\mathbf{x}) \in (0,\infty)\times\mathscr{D}\,, \tag{II.33c}$$

$$\operatorname{div}(\varepsilon(\mathbf{x})\, E(t,\mathbf{x})) = 0 \qquad (t,\mathbf{x}) \in (0,\infty)\times\mathscr{D}\,, \tag{II.33d}$$

where $\mu > 0$ is the **magnetic permeability** and $\varepsilon > 0$ the **electric permittivity**, both material coefficients, E is the **electric field** and H is the **magnetic field**, which oscillate in time $t \in [0,\infty)$ according to angular frequency $\omega > 0$. In order to solve the time-harmonic system, a time-independent EVP is formulated. The **electric Maxwell EVP**, cf. [**105**, (7.2)], is to find an eigenvalue $\lambda = \omega^2$ and an electric field E for fixed time $t = 0$ satisfying

$$\mathbf{curl}(\mu^{-1}(\mathbf{x})\ \mathbf{curl}(E(0,\mathbf{x}))) = \varepsilon(\mathbf{x})\, E(0,\mathbf{x})\, \lambda \qquad x \in \mathscr{D}\,, \tag{II.34a}$$

$$\operatorname{div}(\varepsilon(\mathbf{x})\, E(0,\mathbf{x})) = 0 \qquad x \in \mathscr{D}\,, \tag{II.34b}$$

$$E(0,\mathbf{x}) \times \boldsymbol{n} = 0 \qquad x \in \partial\mathscr{D}\,. \tag{II.34c}$$

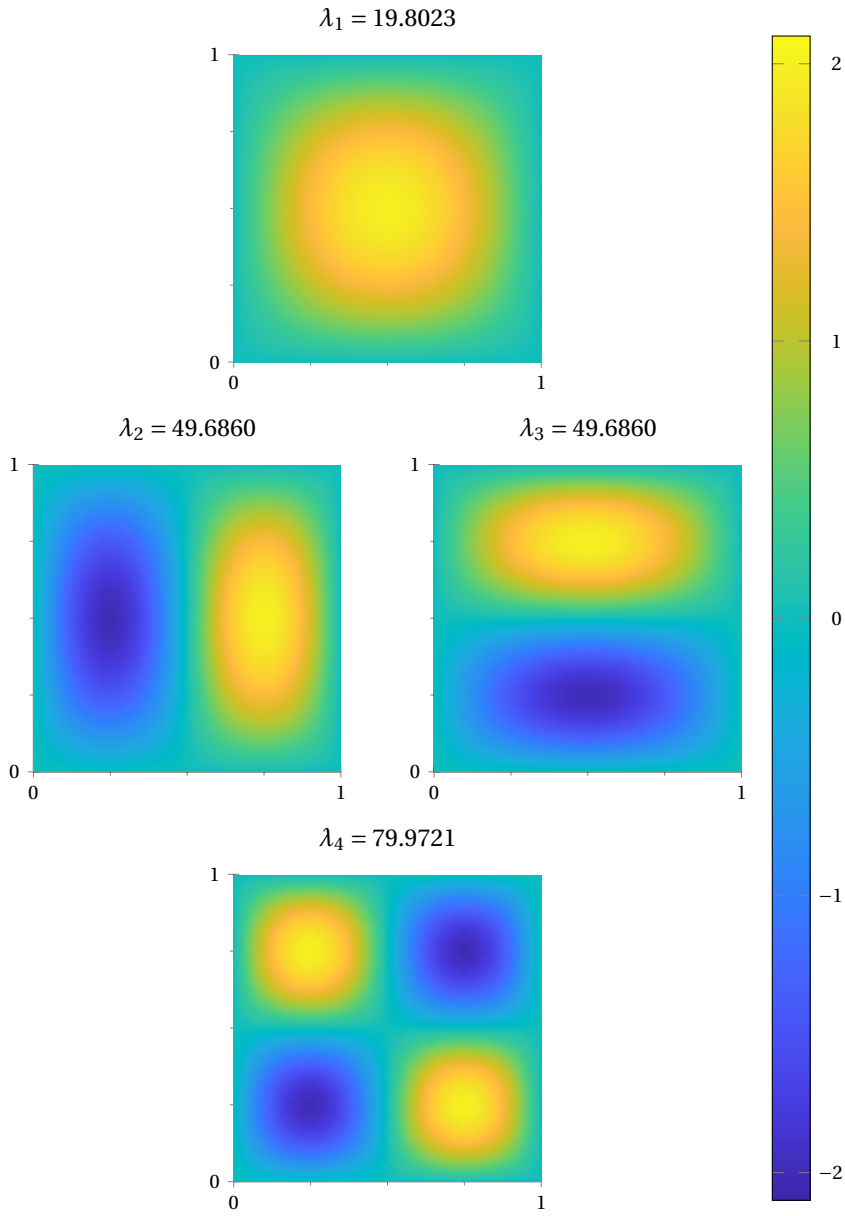

FIGURE II.1. Solutions of the Laplace EVP with zero Dirichlet boundary conditions on the unit square (example II.77).

where $\boldsymbol{n}$ is the unit outer normal vector of the boundary $\partial\mathscr{D}$. There is an equivalent *magnetic* Maxwell EVP in terms of the magnetic field H, cf. [**105**, (7.3)].
For the variational formulation, we define function spaces

$$H(\mathbf{curl};\mathscr{D}) := \left\{ u \in [L^2(\mathscr{D})]^3 \,:\, \mathbf{curl}(u) \in [L^2(\mathscr{D})]^3 \right\}, \qquad \mathscr{D} \subset \mathbb{R}^3 .$$

with scalar product

$$\langle u, v\rangle_{H(\mathbf{curl};\mathscr{D})} := \int_{\mathscr{D}} \langle \mathbf{curl}\, u(\mathbf{x}), \mathbf{curl}\, v(\mathbf{x})\rangle_{\mathbb{R}^3}\, \mathrm{d}\mathbf{x} + \langle u, v\rangle_{[L^2(\mathscr{D})]^3} . \tag{II.35}$$

$(H(\mathbf{curl};\mathscr{D}), \langle\cdot,\cdot\rangle_{H(\mathbf{curl};\mathscr{D})})$ is a Hilbert space with a closed subspace

$$H_0(\mathbf{curl},\mathscr{D}) := \overline{[C_0^\infty(\mathscr{D})]^3}^{\|\cdot\|_{H(\mathbf{curl};\mathscr{D})}}, \qquad \mathscr{D} \subset \mathbb{R}^3 .$$

As a closed subspace $H_0(\mathbf{curl},\mathscr{D}) \subset H(\mathbf{curl};\mathscr{D})$ is again a Hilbert space using (II.35). We can find compact dense embeddings, cf. [**105**, Problem 7.1],

$$\mathrm{Id} : H(\mathbf{curl};\mathscr{D}) \cap \mathscr{N}(\mathrm{div}) \hookrightarrow [L^2(\mathscr{D})]^3 , \quad \mathrm{Id}' : ([L^2(\mathscr{D})]^3)' \hookrightarrow (H(\mathbf{curl};\mathscr{D}) \cap \mathscr{N}(\mathrm{div}))'$$

which leads to Gelfand triples (theorem II.30)

$$H(\mathbf{curl};\mathscr{D}) \cap \mathscr{N}(\mathrm{div}) \overset{\mathrm{Id}}{\hookrightarrow} [L^2(\mathscr{D})]^3 \overset{R_{[L^2(\mathscr{D})]^3}}{\to} ([L^2(\mathscr{D})]^3)' \overset{\mathrm{Id}'}{\hookrightarrow} (H(\mathbf{curl};\mathscr{D}) \cap \mathscr{N}(\mathrm{div}))' ,$$

$$H_0(\mathbf{curl},\mathscr{D}) \cap \mathscr{N}(\mathrm{div}) \overset{\mathrm{Id}}{\hookrightarrow} [L^2(\mathscr{D})]^3 \overset{R_{[L^2(\mathscr{D})]^3}}{\to} ([L^2(\mathscr{D})]^3)' \overset{\mathrm{Id}'}{\hookrightarrow} (H_0(\mathbf{curl},\mathscr{D}) \cap \mathscr{N}(\mathrm{div}))' .$$

Now that we have established the vector spaces, we can formulate the variational form.

EXAMPLE II.78 (**Maxwell EVP**). Consider the function spaces

$$V = H_0(\mathbf{curl};\mathscr{D}) \cap \mathscr{N}(\mathrm{div}) , \qquad H = [L^2(\mathscr{D})]^3$$

with open, bounded, simply connected, and Lipschitz domain $\mathscr{D} \subset \mathbb{R}^3$. Then the **Maxwell EVP in variational form** (II.22) is given by the bilinear forms

$$a(u,v) = \int_{\mathscr{D}} \left\langle \mu^{-1}(\mathbf{x})\ \mathbf{curl}(u(\mathbf{x})), \mathbf{curl}(v(\mathbf{x}))\right\rangle_{\mathbb{R}^3}\, \mathrm{d}\mathbf{x} , \tag{II.36a}$$

$$b(u,v) = \int_{\mathscr{D}} \langle \varepsilon(\mathbf{x})\ u(\mathbf{x}), v(\mathbf{x})\rangle_{\mathbb{R}^3}\, \mathrm{d}\mathbf{x} . \tag{II.36b}$$

The model assumption II.74 holds and V-ellipticity of a can be proven by a Poincaré-style inequality, cf. [**51**, Corollary 4.4].

REMARK II.79. The discretization of example II.78 is more convenient for Hilbert spaces

$$V = H_0(\mathbf{curl};\mathscr{D}) , \qquad H = [L^2(\mathscr{D})]^3 .$$

Since the fields $E \in \mathscr{N}(\mathrm{div})$ are in the null space of the **curl** operator, we get additional eigenvalues $\lambda = 0$ with associated eigenfunctions in $\mathscr{N}(\mathrm{div})$ when using these

function spaces. If we use a good estimate of the eigenvalue that we are looking for, we will not encounter the additional eigenpairs, when solving the discrete EVP, since they all have the same eigenvalue $\lambda = 0$. Otherwise, if we have to solve for all eigenvalues, we simply have to discard eigenvalues $\lambda = 0$, since the original EVP is elliptic and only contributes eigenvalues $\lambda > 0$, cf. [**105**, Problem 7.1].

Finite Element Implementation. The classical FE schemes are Nédélec's FEs of first [**71, 87**] and second kind [**72**]. For the Maxwell EVP, we rely on the isogeometric analysis (IGA) implementation of the GeoPDEs library [**98**] for which remark II.79 applies. IGA is a FE method, which uses non-uniform rational B-splines (NURBS) to describe the domain. This eliminates additional discretization errors in the representation of the domain, as the NURBS functions are also used for computer-aided design (CAD) software. For the FE basis functions, B-splines are chosen.
The eigenfunctions to the smallest eigenvalue with multiplicity $m = 3$ of the Maxwell EVP on the unit sphere (decomposed into seven patches) are illustrated in fig. II.2.

II.6.3. The two-dimensional Maxwell Eigenvalue Problem as a Laplace Eigenvalue Problem. If the domain $\mathscr{D} \subset \mathbb{R}^3$ is translation invariant in one coordinate direction, we can model the Maxwell EVP on a two-dimensional slice instead. We summarize the presentation of [**87**, Appendix A].
First, we have to adapt the operator **curl** to the new two-dimensional setting. For the curl of a two-dimensional vector $u \in \mathbb{R}^2$, we embed $\mathbb{R}^2 \hookrightarrow \mathbb{R}^3$ and apply $\mathbf{curl} : \mathbb{R}^3 \to \mathbb{R}^3$. The result in $\mathbb{R}$ is orthogonal to the embedded $\mathbb{R}^2$-plane, pointing in the additional third dimension. Thus, we can represent the two-dimensional curl by the scalar value

$$\operatorname{curl}(u(\mathbf{x})) \coloneqq \frac{\partial u_2(\mathbf{x})}{\partial \mathbf{x}_1} - \frac{\partial u_1(\mathbf{x})}{\partial \mathbf{x}_2} .$$

The **curl** of this scalar interpreted in $\mathbb{R}^3$ is again in the $\mathbb{R}^2$-plane and given by

$$\operatorname{Curl}(u(\mathbf{x})) \coloneqq \begin{bmatrix} \frac{\partial u(\mathbf{x})}{\partial \mathbf{x}_2} \\ -\frac{\partial u(\mathbf{x})}{\partial \mathbf{x}_1} \end{bmatrix} .$$

PROPOSITION II.80 ([**87**], Appendix A). *Let $\mathscr{D} \subset \mathbb{R}^2$ be an open, bounded, simply connected, and Lipschitz domain. Consider the solution (λ, w) of the EVP with Neumann boundary data*

$$-\Delta w(\mathbf{x}) = w(\mathbf{x})\, \lambda \qquad \mathbf{x} \in \mathscr{D} , \tag{II.37a}$$

$$n \cdot \operatorname{grad} w(\mathbf{x}) = 0 \qquad \mathbf{x} \in \partial\mathscr{D}. \tag{II.37b}$$

Then (λ, E) with $E = \operatorname{Curl}(w)$ is a solution to the 2D Maxwell EVP problem,

$$\operatorname{Curl}(\operatorname{curl}(E(\mathbf{x}))) = E(\mathbf{x})\, \lambda \qquad \mathbf{x} \in \mathscr{D} ,$$

$$E(\mathbf{x}) \times n = 0 \qquad \mathbf{x} \in \partial\mathscr{D} .$$

We state the Laplace EVP with the new Neumann boundary conditions again in terms of bilinear forms.

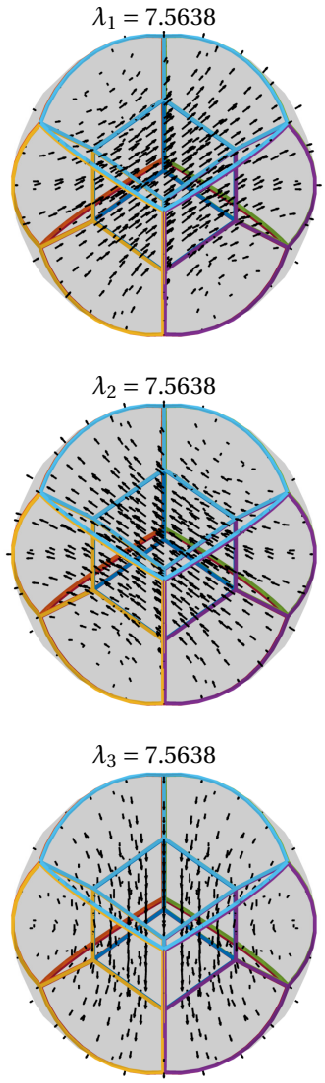

FIGURE II.2. Solution of the Maxwell EVP on the unit sphere with multiplicity $m = 3$. The eigenfunctions are aligned to the coordinate axes and the vectors not scaled for normalization. The boundaries of the patches are highlighted.

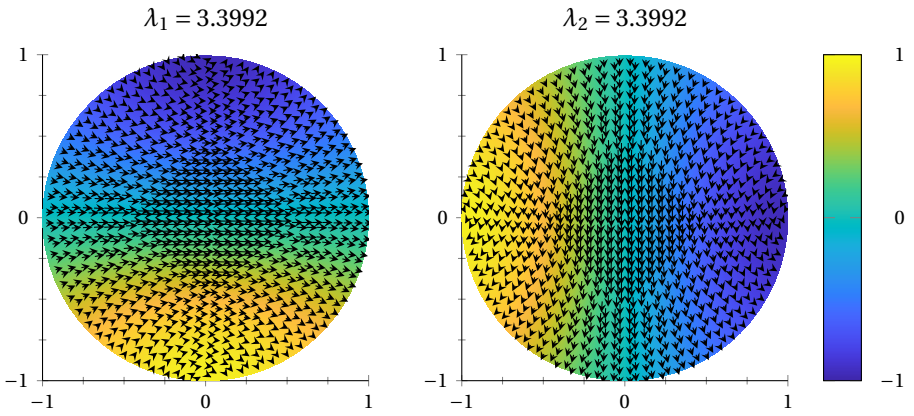

FIGURE II.3. First two Laplace eigenpairs with Neumann boundary condition on unit disk (example II.81). The vector fields relate to two eigenfunctions of fig. II.2.

EXAMPLE II.81 ([**4**, 6.5(2)]). For the **weak formulation of the Laplace EVP with Neumann boundary data**, consider the function spaces

$$V = H^1(\mathscr{D}) \setminus \mathcal{N}(\mathrm{grad}) , \qquad H = L^2(\mathscr{D}) ,$$

and the EVP (II.22) with bilinear forms and domain as in example II.77, then the assumption II.74 still holds.

For convenience of implementation, we can also consider the function spaces

$$V = H^1(\mathscr{D}) , \qquad H = L^2(\mathscr{D}) ,$$

which yields an additional eigenpair $(0, \pm 1) \in \mathbb{R} \times V$, similar to remark II.79. The implementation is analogous to the Laplace EVP with zero Dirichlet boundary, but with additional degrees of freedom associated with nodes on the boundary $\partial\mathscr{D}$. For $\mathscr{D}$ being the unit disk, the two eigenfunctions to the smallest (non-zero) eigenvalue with multiplicity $m = 2$ are illustrated in fig. II.3.

II.7. Probability Theory

We discuss the prerequisites of probability theory following [**4, 60**] and for random fields in particular [**62**]. Finally, we recall some essentials on quadrature rules and Monte Carlo (MC) methods.

II.7.1. Probability Spaces. We start with the definition of probability spaces.

DEFINITION II.82 ([**60**, 1.28, 1.38, & 1.68]). Let $(\Omega, \mathscr{A}, \mathbb{P})$ be a measure space. If $\mathbb{P} : \mathscr{A} \to [0,1]$ we call $\mathbb{P}$ a **probability measure** and $(\Omega, \mathscr{A}, \mathbb{P})$ a **probability space**. In this case, a set $A \in \mathscr{A}$ is called an **event** and $\mathbb{P}(A)$ is its **probability**. In the context of a probability space the expression $\mathbb{P}$-a.e. (definition II.50) is replaced by **$\mathbb{P}$-almost surely (a.s.)**.

We can define random variables on the structure of the probability space.

DEFINITION II.83 ([**60**, 1.102]). Let $(\Omega, \mathscr{A}, \mathbb{P})$ be a probability space and X a Banach space. A $\mathbb{P}$-measurable function

$$\xi : \Omega \to X\,, \qquad\qquad \omega \mapsto \xi(\omega)\,,$$

is then called a **random variable**. We call $\xi(\omega)$ the **realization** of the random variable. A random variable $\xi : \Omega \to \mathbb{R}$ is called a **real random variable**.

Random variables are described by their distribution.

DEFINITION II.84 ([**60**, 1.103]). Let ξ be a random variable of definition II.83.

(1) The probability measure $\mathbb{P}_\xi := \mathbb{P} \circ \xi^{-1}$ is called the **distribution** of ξ, cf. definition II.53. We write $\xi \sim \mu$ if $\mu = \mathbb{P}_\xi$ and say that ξ has distribution μ.
(2) For a real random variable ξ, the map $F_\xi : \mathbb{R} \to [0,1], x \mapsto \mathbb{P}(\xi \leq x)$ is called the **distribution function** of ξ.
(3) A family $(\xi)_{i\in I}$ of random variables is called **identically distributed** if $\mathbb{P}_{\xi_i} = \mathbb{P}_{\xi_j}$ for all $i, j \in I$. We also write $\xi_i \sim \xi_j$.

II.7.2. Stochastic Independence. We define stochastic independence for random variables, which is based on the definition of independent events, cf. [**60**, Chapter 2].

DEFINITION II.85 ([**60**, 2.15]). A family of random variables $(\xi_i)_{i\in I}$, $\xi_i : \Omega \to X_i$ with X_i Banach spaces is called **independent** if for any finite set $J \subset I$ and any choice of $A_j \in \mathscr{B}(X_j)$, $j \in J$, we have

$$\mathbb{P}\left(\bigcap_{j\in J} \{\xi_j \in A_j\}\right) = \prod_{j\in J} \mathbb{P}\left(\xi_j \in A_j\right)\,.$$

We write $(\xi_i)_{i\in I}$ **iid** if $(\xi_i)_{i\in I}$ are independent and identically distributed.

DEFINITION II.86 ([**60**, 2.20]). For any $i \in I$, let ξ_i be a real random variable. For any finite subset $J \subset I$, let

$$F_J := F_{(\xi_j)_{j\in J}} : \mathbb{R}^{|J|} \to [0,1]\,, \qquad x \mapsto \mathbb{P}\left(\xi_j \leq x_j, \forall\, j \in J\right) = \mathbb{P}\left(\bigcap_{j\in J} \xi_j^{-1}((-\infty, x_j])\right)\,.$$

Then F_J is called the **joint distribution function** of $(\xi_j)_{j\in J}$. The probability measure $\mathbb{P}_{(\xi_j)_{j\in J}}$ on $\mathbb{R}^{|J|}$ is called the **joint distribution** of $(\xi_j)_{j\in J}$.

Independence can be expressed by distributions using the following result.

LEMMA II.87 ([**60**, 2.21]). *A family $(\xi_i)_{i\in I}$ of real random variables is independent if and only if for every $J \subset I$ and every $x = (x_j)_{j\in J} \in \mathbb{R}^{|J|}$, the joint distribution function can be expressed as*

$$F_J(x) = \prod_{j\in J} F_j(x_j) .$$

II.7.3. Stochastic Moments. We now define the mean, correlation, and covariance, which are central quantities of interest for uncertainty quantification.

DEFINITION II.88 ([**60**, 5.1]). Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, X a Banach space, and $\mathcal{D} \subset \mathbb{R}^n$ a domain.

(1) For $\xi \in L^1_{\mathbb{P}}(\Omega; X)$

$$\mathbb{E}[\xi] := \int_\Omega \xi \, d\mathbb{P} \in X \tag{II.38a}$$

is called the **mean** or **expected value** of ξ. ξ is called **centered** if $\mathbb{E}[\xi] = 0$. If X is a function space of functions $\mathcal{D} \to \mathbb{R}$, then

$$\mathbb{E}[\xi] : \mathcal{D} \to \mathbb{R} , \qquad \mathbf{x} \mapsto \mathbb{E}[\xi](\mathbf{x}) .$$

(2) Let X, Y be two Hilbert spaces, $\xi \in L^2_{\mathbb{P}}(\Omega; X)$ and $\zeta \in L^2_{\mathbb{P}}(\Omega; Y)$, then

$$\mathbb{C}\mathrm{or}[\xi, \zeta] := \mathbb{E}[\xi \otimes \zeta] \in X \otimes Y , \tag{II.38b}$$

is called the **correlation** and

$$\mathbb{C}\mathrm{ov}[\xi, \zeta] := \mathbb{E}[(\xi - \mathbb{E}[\xi]) \otimes (\zeta - \mathbb{E}[\zeta])] \in X \otimes Y . \tag{II.38c}$$

is called the **covariance**. They are called **uncorrelated** if $\mathbb{C}\mathrm{ov}[\xi, \zeta] = 0$ and **correlated** otherwise. If X, Y are function spaces of functions $\mathcal{D} \to \mathbb{R}$

$$\mathbb{C}\mathrm{or}[\xi, \zeta] : \mathcal{D} \times \mathcal{D} \to \mathbb{R} , \qquad (\mathbf{x}, \mathbf{y}) \mapsto \mathbb{C}\mathrm{or}[\xi, \zeta](\mathbf{x}, \mathbf{y}) ,$$
$$\mathbb{C}\mathrm{ov}[\xi, \zeta] : \mathcal{D} \times \mathcal{D} \to \mathbb{R} , \qquad (\mathbf{x}, \mathbf{y}) \mapsto \mathbb{C}\mathrm{ov}[\xi, \zeta](\mathbf{x}, \mathbf{y}) .$$

We use the shorthand notations $\mathbb{C}\mathrm{or}[\xi] := \mathbb{C}\mathrm{or}[\xi, \xi]$ and $\mathbb{C}\mathrm{ov}[\xi] := \mathbb{C}\mathrm{ov}[\xi, \xi]$. The **variance (function)** is defined as

$$\mathbb{V}\mathrm{ar}[\xi] : \mathcal{D} \to \mathbb{R} , \qquad \mathbf{x} \mapsto \mathbb{V}\mathrm{ar}[\xi](\mathbf{x}) := \mathbb{C}\mathrm{ov}[\xi](\mathbf{x}, \mathbf{x}) .$$

(3) For random variables $\xi \in L^2_{\mathbb{P}}(\Omega; \mathbb{R}^n)$ the covariance $\mathbb{C}\mathrm{ov}[\xi]$ is a matrix $\Sigma \in \mathbb{R}^{n\times n}$ which we call the **covariance matrix** with entries $\Sigma_{ij} = \mathbb{C}\mathrm{ov}[\xi_i, \xi_j]$, $i, j = 1, \ldots, n$.

(4) For a real random variable $\xi \in L^2_{\mathbb{P}}(\Omega)$, the **variance** is $\mathbb{V}\mathrm{ar}[\xi] := \mathbb{C}\mathrm{ov}[\xi, \xi] \geq 0$ and $\sqrt{\mathbb{V}\mathrm{ar}[\xi]}$ is called the **standard deviation**.

(5) For $\xi \in L^3_{\mathbb{P}}(\Omega)$ we call $\mathbb{E}[(\xi - \mathbb{E}[\xi])^3] \, \mathbb{V}\mathrm{ar}[\xi]^{-\frac{3}{2}}$ the **skew** of ξ and call it **skewfree** if the skew vanishes.

If there is ambiguity about the probability measure of definition II.88, we may add it to the notation as an index. These quantities of interest can be generalized to the notion of stochastic moments.

DEFINITION II.89 ([**60**, 5.1]). Let $(\Omega, \mathscr{A}, \mathbb{P})$ be a probability space and X a Hilbert space. For $p \in \mathbb{N}$ and $\xi \in L^p_{\mathbb{P}}(\Omega; X)$ then

$$\mathbb{E}[\xi^{\otimes p}] \coloneqq \mathbb{E}[\underbrace{\xi \otimes \ldots \otimes \xi}_{p \text{ times}}]$$

is called the p-**th stochastic moment** and $\mathbb{E}[(\xi - \mathbb{E}[\xi])^{\otimes p}]$ the p-**th centered stochastic moment**.

Recall the fact that $\mathbb{V}\mathrm{ar}[\xi] = 0$ if and only if $\xi = \mathbb{E}[\xi]$ a.s.. The following rules for mean and covariance hold, assuming that they are well defined respectively for $\xi : \Omega \to X$ and $\zeta : \Omega \to Y$

$$\mathbb{E}[x + \xi] = x + \mathbb{E}[\xi] \,, \qquad x \in X \,, \tag{II.39a}$$

$$\mathbb{E}[a\xi] = a\mathbb{E}[\xi] \,, \qquad a \in \mathbb{R} \,, \tag{II.39b}$$

$$\mathbb{C}\mathrm{ov}[x + \xi, y + \zeta] = \mathbb{C}\mathrm{ov}[\xi, \zeta] \,, \qquad x \in X, y \in Y \,, \tag{II.39c}$$

$$\mathbb{C}\mathrm{ov}[a\xi, b\zeta] = ab\mathbb{C}\mathrm{ov}[\xi, \zeta] \,, \qquad a, b \in \mathbb{R} \,, \tag{II.39d}$$

$$\mathbb{C}\mathrm{ov}[\xi, \zeta] = \mathbb{C}\mathrm{or}[\xi, \zeta] - \mathbb{E}[\xi] \otimes \mathbb{E}[\zeta] \,. \tag{II.39e}$$

For uncorrelated random variables, (II.39e) can be simplified to

$$\mathbb{C}\mathrm{or}[\xi, \zeta] = \mathbb{E}[\xi] \otimes \mathbb{E}[\zeta] \,.$$

Recall that independent random $\xi \in L^2_{\mathbb{P}}(\Omega; X), \zeta \in L^2_{\mathbb{P}}(\Omega; Y)$ are uncorrelated. Note also that the indicator function expresses the probability of $A \in \mathscr{A}$ as a mean

$$\mathbb{P}(A) = \int_\Omega \mathbb{1}_A(\omega) \, \mathrm{d}\mathbb{P}(\omega) = \mathbb{E}_{\mathbb{P}}[\mathbb{1}_A] \,. \tag{II.40}$$

II.7.4. Examples of Distributions. In the finite-dimensional case, we can express the distribution of a random variable in terms of a density.

DEFINITION II.90 ([**60**, 2.22]). If the distribution function $F : \mathbb{R}^n \to [0, 1]$ is of the form

$$F(x) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(t_1, \ldots, t_n) \, \mathrm{d}t_n \cdots \mathrm{d}t_1$$

for some integrable function $f : \mathbb{R}^n \to [0, \infty)$, then f is called the **density** of the distribution.

We recall the uniform and normal distributions to establish their stochastic moments.

EXAMPLE II.91 ([**60**, 1.75, 1.105(vii, ix)]). Let $(\Omega, \mathscr{A}, \mathbb{P})$ be a probability space.

(1) Let $A \in \mathscr{B}(\mathbb{R}^n)$ be a measurable set and $\mathrm{Leb} : \mathscr{B}(\mathbb{R}^n) \to [0,\infty]$ the Lebesgue measure, such that $\mathrm{Leb}(A) < \infty$. Then we can define a probability measure $\mathbb{P}$ on $\mathscr{B}(\mathbb{R}^n)|_A$ by

$$\mathbb{P}(B) \coloneqq \frac{\mathrm{Leb}(B)}{\mathrm{Leb}(A)} \qquad \forall B \in \mathscr{B}(\mathbb{R}^n) \text{ with } B \subset A\,.$$

This measure $\mathbb{P}$ is called the **uniform distribution** on A and is denoted by $\mathscr{U}(A)$.

(2) Let $[a,b] \subset \mathbb{R}$ be an interval and $\xi \sim \mathscr{U}([a,b])$, then its density is

$$f(t) = \frac{1}{b-a}\,\mathbb{1}_{[a,b]}(t) \qquad t \in \mathbb{R}$$

and its mean, variance, and skew are given by

$$\mathbb{E}[\xi] = \frac{a+b}{2}\,, \qquad \mathbb{V}\mathrm{ar}[\xi] = \frac{(b-a)^2}{12}\,, \qquad \mathbb{E}[(\xi - \mathbb{E}[\xi])^3] = 0\,.$$

(3) Let $m \in \mathbb{R}$, $s > 0$ and ξ be a real random variable with density

$$f(t) = \frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{1}{2}\left(\frac{t-m}{s}\right)^2\right).$$

Then we call ξ **normally distributed** or **Gaussian** and write $\xi \sim \mathscr{N}(m, s^2)$. Its mean, variance, and skew are given by

$$\mathbb{E}[\xi] = m\,, \qquad \mathbb{V}\mathrm{ar}[\xi] = s^2\,, \qquad \mathbb{E}[(\xi - \mathbb{E}[\xi])^3] = 0\,.$$

We call $\mathscr{N}(0,1)$ the **standard normal distribution**.

(4) Let $m \in \mathbb{R}^n$, $\Sigma \in \mathbb{R}^{n\times n}$ a symmetric positive definite covariance matrix, and let $\xi : \Omega \to \mathbb{R}^n$ be a random variable with density

$$f(t) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \exp\left(-\frac{1}{2}\|t-m\|_\Sigma^2\right) \qquad t \in \mathbb{R}^n\,,$$

where

$$\langle u, v\rangle_\Sigma \coloneqq \left\langle \Sigma^{-1}u, v\right\rangle_{\mathbb{R}^n}\,, \qquad \|u\|_\Sigma \coloneqq \sqrt{\langle u,u\rangle_\Sigma}\,. \tag{II.41}$$

Then, we call ξ **multinomial normally distributed** and write $\xi \sim \mathscr{N}(m,\Sigma)$. Its first stochastic moments are given by

$$\mathbb{E}[\xi] = m\,, \qquad \mathbb{V}\mathrm{ar}[\xi] = \Sigma\,, \qquad \mathbb{E}[(\xi - \mathbb{E}[\xi])^{\otimes 3}] = 0\,.$$

Note, that if $\xi \sim \mathscr{N}(m,\Sigma)$, then for each entry ξ_i holds $\xi_i \sim \mathscr{N}(m_i, \Sigma_{ii})$ and for pairs of entries ξ_i, ξ_j holds $\mathbb{C}\mathrm{ov}[\xi_i,\xi_j] = \Sigma_{ij} = \Sigma_{ji}$.

We only give a short summary on the generation of random variables and refer the interested reader to [**62**, Chapter 4.4]. In algorithmic applications, continuously uniform random variables are usually replaced by *pseudorandom* sequences, which are

statically indistinct form actual probabilistic independent continuously uniform random variables. They are however entirely predictable, with knowledge of the employed algorithm and a starting value, called *seed*. There are multiple ways to generate standard normally distributed random variables $\xi \sim \mathcal{N}(0,1)$ from samples of $\mathcal{U}([0,1])$. Uniformly distributed random variables on different intervals or normally distributed random variables with different mean and variance can then be generated using scales and shifts according to (II.39). Random variables $\xi \sim \mathcal{N}(m,\Sigma)$ can be generated using the following lemma.

LEMMA II.92 ([**62**, 4.32]). *Let $\xi_i \in L^2_{\mathbb{P}}(\Omega)$, $i = 1,\dots,n$, be centered iid random variables with variance $\mathbb{V}\mathrm{ar}[\xi_i] = 1$ and $\Sigma \in \mathbb{R}^{n\times n}$ be a symmetric positive definite covariance matrix with decomposition $\Sigma = LL^\top$. Then for $\zeta = L\xi + m$ holds $\mathbb{E}[\zeta] = m$, $\mathbb{C}\mathrm{ov}[\zeta] = \Sigma$. If $\xi_i \sim \mathcal{N}(0,1)$, then $\zeta \sim \mathcal{N}(m,\Sigma)$.*

PROOF. The proof is done in [**62**, 4.32] directly for $\xi_i \sim \mathcal{N}(0,1)$, but the result on mean and covariance hold for more general distributions since

$$\begin{aligned}\mathbb{E}[\zeta] &= \mathbb{E}[\xi] + m = m\,,\\ \mathbb{C}\mathrm{ov}[\zeta] &= \mathbb{C}\mathrm{ov}[L\xi] = \mathbb{E}[(L\xi)(L\xi)^\top] = L\mathbb{E}[\xi\xi^\top]L^\top = LL^\top = \Sigma\,.\end{aligned}$$ □

II.7.5. Conditional Probabilities and Bayes' Theorem. In chapter VI we consider a Bayesian inverse problem. Bayesian inversion relies on Bayes' formula, for which we first have to define conditional probabilities.

DEFINITION II.93 ([**60**, 8.2]). Let $(\Omega,\mathcal{A},\mathbb{P})$ be a probability space and $B \in \mathcal{A}$. We define the **conditional probability** given B for any $A \in \mathcal{A}$ by

$$\mathbb{P}(A|B) = \begin{cases} \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)} & \text{if } \mathbb{P}(B) > 0\,,\\ 0 & \text{else}\,.\end{cases}$$

Conditional probabilities lead to conditional probability measures, which lets us state Bayes' formula.

LEMMA II.94 ([**60**, 8.4, 8.5, 8.6, & 8.7]). *Let the assumptions of definition II.93 hold.*

(1) *If $\mathbb{P}(B) > 0$, then $\mathbb{P}(\cdot|B)$ is a probability measure on $(\Omega,\mathcal{A})$.*
(2) *Let $A,B \in \mathcal{A}$ with $\mathbb{P}(A),\mathbb{P}(B) > 0$. Then*

$$A,B \text{ independent} \quad\Longleftrightarrow\quad \mathbb{P}(A|B) = \mathbb{P}(A) \quad\Longleftrightarrow\quad \mathbb{P}(B|A) = \mathbb{P}(B)\,.$$

(3) **Summation formula:**
Let I be a countable set and let $(B_i)_{i\in I}$ be pairwise disjoint sets with $\mathbb{P}(\bigcup_{i\in I} B_i) = 1$. Then for any $A \in \mathcal{A}$,

$$\mathbb{P}(A) = \sum_{i\in I} \mathbb{P}(A|B_i)\mathbb{P}(B_i)\,.$$

(4) **Bayes formula:**
Let I be a countable set and let $(B_i)_{i\in I}$ be pairwise disjoint sets with $\mathbb{P}(\bigcup_{i\in I} B_i) = 1$. Then for any $A \in \mathscr{A}$ with $\mathbb{P}(A) > 0$ and $k \in I$

$$\mathbb{P}(B_k|A) = \frac{\mathbb{P}(A|B_k)\mathbb{P}(B_k)}{\sum_{i\in I} \mathbb{P}(A|B_i)\mathbb{P}(B_i)}$$

Let X be a Banach space, $\xi \in L_{\mathbb{P}}(\Omega; X)$ a random variable, and $A \in \mathscr{A}$ an event, then we define the **conditional mean** as the mean under the probability measure of definition II.93 and lemma II.94

$$\mathbb{E}[\xi|A] := \mathbb{E}_{\mathbb{P}(\cdot|A)}[\xi] \,.$$

Let $\zeta : \Omega \to Y$ be a second random variable. Then we write $\mathbb{E}[\xi|\zeta = y]$ to condition ξ to the event that the realization of ξ is $y \in Y$. If ξ, ζ are independent, then

$$\mathbb{E}[\xi \mid \zeta = y] = \mathbb{E}[\xi] \,.$$

For Hilbert spaces X, Y, **conditional (co)variances** and **correlations** can be defined analogously to definition II.88 using the conditional probability measure.

II.7.6. Random Fields. Random fields can be seen as Hilbert space-valued random variable, however, we provide additional results, which are more specific to random fields, following [**62**, Chapter 7].

DEFINITION II.95 ([**62**, Definition 7.1 & 7.3]). Let $\mathscr{D} \subset \mathbb{R}^n$ be a domain.

(1) A real-valued **random field** $\{\xi(\mathbf{x}) : \mathbf{x} \in \mathscr{D}\}$ is a set of $\mathbb{R}$-valued random variables on a probability space $(\Omega, \mathscr{A}, \mathbb{P})$. Formally

$$\xi : \mathscr{D} \times \Omega \to \mathbb{R} \,, \qquad\qquad (\mathbf{x}, \omega) \mapsto \xi(\mathbf{x}, \omega) \,,$$

however, we often drop the argument ω for ease of notation.
(2) We call a random field **second-order** if $\xi(\mathbf{x}) \in L^2_{\mathbb{P}}(\Omega)$ for every $\mathbf{x} \in \mathscr{D}$.
(3) A second-order random field is called **Gaussian** if $u(\mathbf{x}) : \Omega \to \mathbb{R}$ is normally distributed for all $\mathbf{x} \in \mathscr{D}$.

Consistent with definition II.88, a second-order random field has a mean and covariance function

$$\mathbb{E}[\xi] : \mathscr{D} \to \mathbb{R} \,, \qquad\qquad \mathbb{C}\mathrm{ov}[\xi] : \mathscr{D} \times \mathscr{D} \to \mathbb{R} \,.$$

DEFINITION II.96 ([**62**, eq. (7.40)]). Let $\mathscr{D} \subset \mathbb{R}^n$ be a domain and $C : \mathscr{D} \times \mathscr{D} \to \mathbb{R}$ a covariance function. Then we define the **covariance operator** $\mathscr{C} : L^2(\mathscr{D}) \to L^2(\mathscr{D})$ by

$$(\mathscr{C}\phi)(\mathbf{x}) = \int_{\mathscr{D}} C(\mathbf{x}, \mathbf{y})\, \phi(\mathbf{y})\, \mathrm{d}\mathbf{y} \,, \qquad\qquad \forall \phi \in L^2(\mathscr{D}),\, \mathbf{x} \in \mathscr{D} \,. \tag{II.42}$$

If $C \in L^2(\mathscr{D} \times \mathscr{D})$ the covariance operator is a *Hilbert–Schmidt operator*. The space of Hilbert–Schmidt operators is a subset of compact operators, cf. [**62**, 1.64, 1.65, & 1.68]. Thus, $\mathscr{C} \in \mathscr{K}(L^2(\mathscr{D}))$ and additionally $\mathscr{C}$ is self-adjoint since C is symmetric. Using the covariance operator whose spectral properties we know by theorem II.67, we can formulate the following decomposition.

THEOREM II.97 ([**62**, 7.52]). *Consider a domain $\mathscr{D} \subset \mathbb{R}^n$ and a second-order random field $\xi \in L^2_{\mathbb{P}}(\Omega; L^2(\mathscr{D}))$. Then it holds*

$$\xi(\mathbf{x}, \omega) = m(\mathbf{x}) + \sum_{i=1}^{\infty} \sqrt{\lambda_i} \phi_i(\mathbf{x}) z_i(\omega) \tag{II.43}$$

where the sum converges in $L^2_{\mathbb{P}}(\Omega; L^2(\mathscr{D}))$. The real random variables $z_j \in L^2_{\mathbb{P}}(\Omega)$ are given by

$$\omega \mapsto z_j(\omega) = \frac{1}{\sqrt{\lambda_i}} \left\langle \xi(\mathbf{x}, \omega) - m(\mathbf{x}), \phi_j(\mathbf{x}) \right\rangle_{L^2(\mathscr{D})} ,$$

with $\mathbb{E}[z_i] = 0$, $\mathbb{C}\mathrm{ov}[z_i, z_j] = \delta_{ij}$. $(\lambda_i, \phi_i) \in \mathbb{R} \times L^2(\mathscr{D})$ are eigenpairs of the covariance operator (II.42) *with the convention of indexing eigenvalues with descending magnitude, i.e., $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$. We call* (II.43) *the **Karhunen–Loève expansion (KLE)** of random field ξ. If ξ is Gaussian, then $z_i \sim \mathcal{N}(0,1)$ iid.*

The following definitions are helpful for categorizing covariances of random fields.

DEFINITION II.98 ([**62**, 7.9 & 7.14]).

(1) We call a second-order random field ξ **stationary** if the mean $\mathbb{E}[\xi](\mathbf{x})$ is independent of $\mathbf{x} \in \mathscr{D}$, i.e., constant, and the covariance has the form

$$\mathbb{C}\mathrm{ov}[\xi](\mathbf{x}, \mathbf{y}) = c(\mathbf{x} - \mathbf{y})$$

for a function $c(x)$ called the **stationary covariance**.

(2) A stationary random field ξ is called **isotropic** if the stationary covariance $c(\mathbf{x})$ is invariant to rotations, i.e., if it is of the form

$$c(\mathbf{x}) = c^0(r) , \qquad r = \|\mathbf{x}\|_2 ,$$

for a function $c^0(r)$ called the **isotropic covariance**.

We present a concrete example of a covariance function, which we will use later.

EXAMPLE II.99. The we call the covariance defined by the isotropic covariance

$$c^0(r) = \exp\left(-\frac{r^2}{2\rho^2}\right) , \tag{II.44}$$

the **squared exponential covariance**. We call the parameter $\rho > 0$ the **correlation length**.

The squared exponential covariance is clearly isotropic. If $c(r)$ decays exponentially, it has the property that the eigenvalues λ_i decay exponentially for a finite-dimensional bounded domain $\mathscr{D} \subset \mathbb{R}^n$, cf. [**62**, 7.57].

Numerical Simulation. The KLE is useful for numerical simulation, since we can simulate a random field by reversing the theorem. In practice, we need to find the eigenpairs of the covariance operator (λ_i, ϕ_i) and assume the distribution for random variables $z_i \in L^2_{\mathbb{P}}(\Omega)$ iid. The sum is truncated after M terms, according to some tolerance. We then sample the random variables $(z_i)_{i=1,\dots,M}$ and apply the formula (II.43). The truncated random field ξ_M has slightly smaller variance, i.e., assuming $\mathbb{V}\mathrm{ar}[z_i] = 1$ and $\|\phi_i\|_{L^2(\mathscr{D})} = 1$, we get

$$\xi_M(\omega, \mathbf{x}) \coloneqq m(\mathbf{x}) + \sum_{i=1}^{M} \sqrt{\lambda_i}\, \phi_i(\mathbf{x})\, z_i(\omega)\,, \qquad \| \mathbb{V}\mathrm{ar}[\xi_M] \|_{L^2(\mathscr{D})} = \sum_{i=1}^{M} \lambda_i\,.$$

If we only require the values of the random field at certain points $(\mathbf{x}_i \in \mathscr{D})_{i=1,\dots,n}$ we can consider a random vector $\zeta \in L^2_{\mathbb{P}}(\Omega; \mathbb{R}^n)$ instead with the covariance matrix Σ such that

$$[\Sigma]_{ij} = \mathbb{C}\mathrm{ov}[\zeta_i, \zeta_j] = \mathbb{C}\mathrm{ov}[\xi](\mathbf{x}_i, \mathbf{x}_j)\,.$$

Thus, we can consider the matrix EVP to find eigenpairs of Σ instead of $\mathscr{C}$ as Σ is symmetric and positive definite. Although the number of eigenpairs is n is finite, we can still truncate according to some tolerance to reduce computational cost. This method becomes computationally expensive if n becomes too large, since Σ is densely populated. An alternative is to use a Cholesky decomposition of Σ and lemma II.92 instead. Larger vector sizes n can still be managed by using a *pivoted Cholesky decomposition*, cf. [**45, 47**], that provides a factorization of a low-rank approximation of Σ and only has to calculate entries of $[\Sigma]_{ij}$ as needed.

As an example, we simulate a random field according to example II.99 on the nodes of a triangular mesh on $\mathscr{D} = (0,1)^2$ with 545 degrees of freedom. Figure II.4 shows two samples of random fields with squared exponential covariance kernels according to

$$\rho \in \left\{ \tfrac{1}{4}, \tfrac{1}{10} \right\}\,, \qquad z_i \sim \mathscr{U}([-\sqrt{3}, \sqrt{3}])\ \text{iid}\,,$$

as well as the respective covariance kernels in terms of Euclidean distance r.

II.7.7. Evaluation of Stochastic Moments by Numerical Integration. The mean and covariance are integrals, so after truncating the dimension of the integration space, we get a finite-dimensional integral of the form

$$\mathbb{E}[\mathsf{R}] = \int_{\mathbb{R}^d} \mathsf{R}(x)\, \mathrm{d}x\,. \tag{II.45}$$

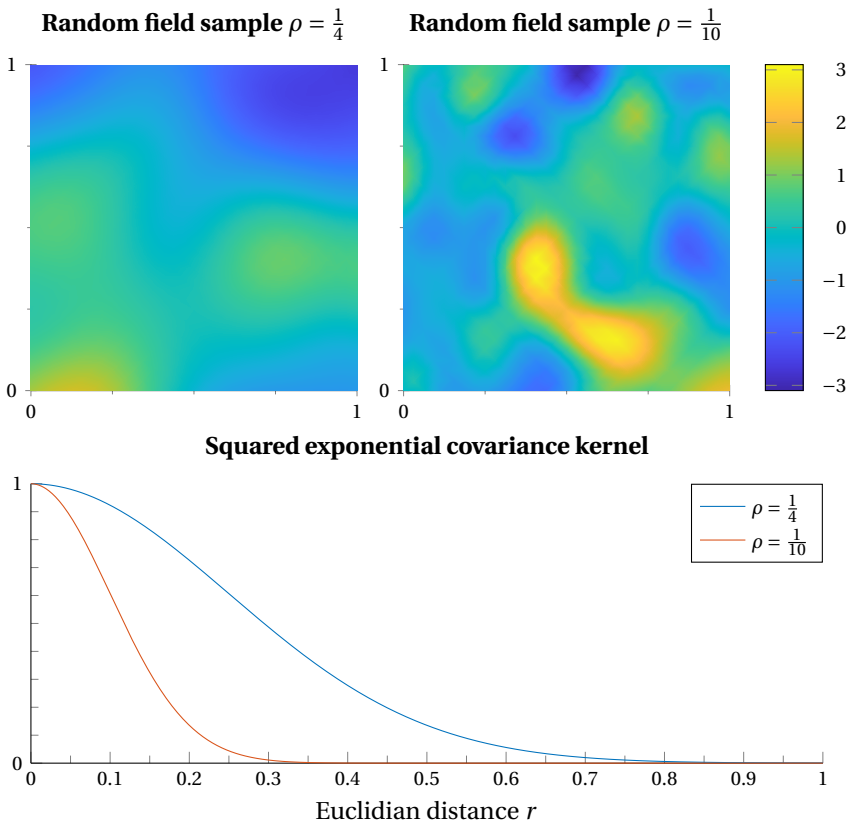

FIGURE II.4. Random fields of the squared exponential kernel (II.44) with different correlation lengths ρ and respective covariance kernels.

Gauß–Legendre quadrature. If the dimension $d \in \mathbb{N}$ of the parameter space in (II.45) is relatively small, we can apply quadrature formulas. Consider first $d = 1$, then an integral of a function on the interval $[a, b] \subset \mathbb{R}$ can be approximated by

$$\int_a^b \mathtt{R}(x)\,\mathrm{d}x \approx (b-a)\sum_{i=1}^{n} w_i \mathtt{R}(x_i) \tag{II.46}$$

where w_i are **weights** and x_i are **nodes**, i.e., evaluation points. The **Gauß–Legendre (GL) quadrature** is optimal in the sense that for n evaluations of the function, the approximation (II.46) is exact if $\mathtt{R}$ is a polynomial of degree $2n-1$, cf. [**92**, Chapter 3.6].

	$n=1$
i	1
x_i	$\frac{a+b}{2}$
w_i	1

	$n=2$	
i	1	2
x_i	$\frac{a+b}{2} - \frac{b-a}{2}\sqrt{\frac{1}{3}}$	$\frac{a+b}{2} + \frac{b-a}{2}\sqrt{\frac{1}{3}}$
w_i	$\frac{1}{2}$	$\frac{1}{2}$

	$n=3$		
i	1	2	3
x_i	$\frac{a+b}{2} - \frac{b-a}{2}\sqrt{\frac{3}{5}}$	$\frac{a+b}{2}$	$\frac{a+b}{2} + \frac{b-a}{2}\sqrt{\frac{3}{5}}$
w_i	$\frac{5}{18}$	$\frac{8}{18}$	$\frac{5}{18}$

TABLE II.1. Nodes and weights of the Gauß–Legendre quadrature with $n \in \{1,2,3\}$ nodes.

The nodes and weights of the GL quadrature for $n \in \{1,2,3\}$ are given in table II.1. For $d > 1$, we can tensorize the quadrature formulas, i.e., we apply them to each dimension consecutively, which means that we get n^d nodes and weights in total.
Using classical tensor product quadrature rules, we can find numerical approximations such that the error converges in $\mathcal{O}(n^{-\frac{2}{d}})$ as the number n of function evaluations increases, cf. [**70**, Chapter 1]. As the dimension d increases, the necessary number of evaluations of f grows exponentially. This unfortunate phenomenon is known as the *curse of dimensionality*.
Monte Carlo Integration. We recall the main ideas of Monte Carlo (MC) integration as a versatile method to approximate the mean and covariance of random variables, which does not suffer from the curse of dimensionality. Our discussion of MC methods is only brief, stating the estimators and their order of convergence. The reader interested in more details is referred to [**14, 21, 70**]. Lastly, we give some references to QMC methods, which are algorithmically very similar and often offer better estimates.

DEFINITION II.100. Let $(\xi_i)_{i=1,\dots,n}$ be iid random variables with $\xi_i \sim \xi \in L^1_{\mathbb{P}}(\Omega; X)$ with X a Banach space. Then the **Monte Carlo (MC) estimator of the mean** is given by

$$M_n[\xi] := \frac{1}{n}\sum_{i=1}^{n} \xi_i \qquad \approx \mathbb{E}[\xi]\,. \tag{II.47a}$$

Let $(\xi_i)_{i=1,\dots,n}$ iid and $(\zeta_i)_{i=1,\dots,n}$ iid be random variables with $\xi_i \sim \xi \in L^2_{\mathbb{P}}(\Omega; X)$ and $\zeta_i \sim \zeta \in L^2_{\mathbb{P}}(\Omega; Y)$ with X, Y Hilbert spaces, then the **MC estimator of the covariance** is

$$V_n[\xi,\zeta] \coloneqq \frac{1}{n-1}\sum_{i=1}^{n}(\xi_i - M_n[\xi]) \otimes (\zeta_i - M_n[\zeta]) \qquad \approx \mathbb{C}\mathrm{ov}[\xi,\zeta]\,, \tag{II.47b}$$

Both estimators are **bias-free**, which means that the expected value of the estimate is exactly the quantity that it is supposed to estimate, i.e.,

$$\mathbb{E}[M_n[\xi]] = \mathbb{E}[\xi]\,, \qquad\qquad \mathbb{E}[V_n[\xi,\zeta]] = \mathbb{C}\mathrm{ov}[\xi,\zeta]\,.$$

Note that the estimator (II.47b) with the factor $(n-1)^{-1}$ only applies when the mean is itself estimated using (II.47a). If the means are known exactly, we can use

$$M_n[(\xi - \mathbb{E}[\xi]) \otimes (\zeta - \mathbb{E}[\zeta])] \approx \mathbb{C}\mathrm{ov}[\xi,\zeta]$$

instead. In analogy to *Bienaymé's formula* [**60**, 5.7] and using independence of samples, we can show that

$$\|M_n[\xi] - \mathbb{E}[\xi]\|_{L^2_{\mathbb{P}}(\Omega;X)} = \frac{\|\xi - \mathbb{E}[\xi]\|_{L^2_{\mathbb{P}}(\Omega;X)}}{\sqrt{n}} \in \mathcal{O}\left(n^{-\frac{1}{2}}\right)\,, \qquad\qquad n \to \infty\,.$$

Therefore, we say that the MC estimator has *square-root convergence*. This is a relatively slow rate of convergence compared to classical quadrature rules, however, independent of the dimension d of the parameter space.

Quasi-Monte Carlo Methods. We only give a short review of known results, which motivate the use of quasi-Monte Carlo (QMC) methods in later numerical experiments. QMC methods use the same estimators (II.47), however, they replace the pseudorandom samples $(\xi_i)_{i=1,\dots,n}$ by a deterministic sequence of points $(x_i)_{i=1,\dots,n}$.

First, the mean must be given as an integral over a hypercube with dimension $d \in \mathbb{N}$, i.e., consider an integral over a function $\mathtt{R} : [0,1]^d \to \mathbb{R}$,

$$\mathbb{E}[\mathtt{R}] = \int_{[0,1]^d} \mathtt{R}(x)\,\mathrm{d}x\,.$$

There are several established QMC sequences. In our numerical experiments, we will use the Halton sequence[4]. See fig. II.5 for a comparison of 10^3 pseudorandom samples of $x_i \sim \mathcal{U}([0,1)^2)$ iid and the corresponding Halton sequence.

Under mild assumptions, the convergence rate of the QMC estimator can be bounded by $\mathcal{O}((\log(n))^d n^{-1})$, cf. [**14**, 5.1]. Thus, in general, the convergence rate cannot be considered dimension-independent. Faster convergence rates, independent of the dimension d, can sometimes be found given stricter assumptions, cf. [**48**]. The interested reader may find a more in depth discussion of QMC methods in [**14, 21, 70**].

[4]The Halton sequence depends on a set of primes as constants. We use the implementation of Matlab's `haltonset`-function with primes of [**61**].

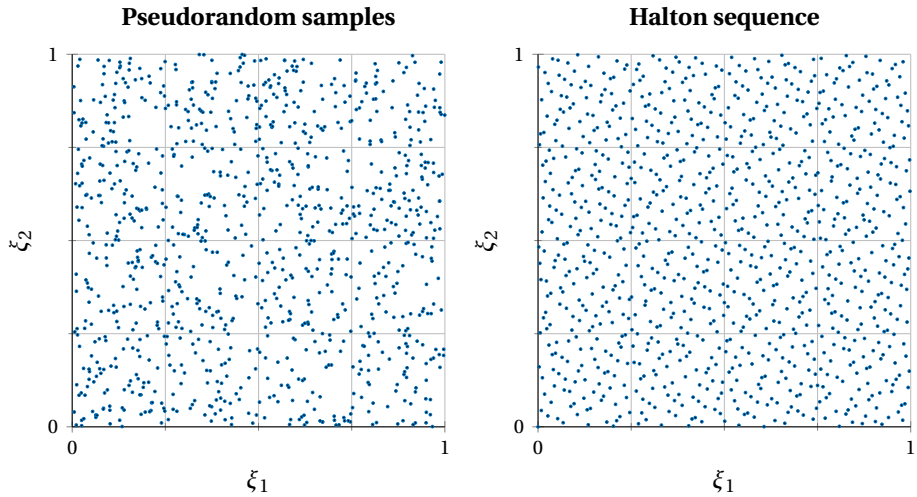

FIGURE II.5. Iid samples of the uniform distribution on the unit square and the entries of the corresponding Halton sequence.

We will use QMC estimates instead of MC estimates as reference solutions in our numerical experiments, when their observed precision is superior, without further investigation of their convergence rates.

CHAPTER III

Trajectories and Derivatives of Eigenpairs

The contents of this chapter are largely based on [**25**] and provide extensions thereof. We formulate the results on the regularity of parameterized EVPs with respect to bilinear forms on real Hilbert spaces according to assumption II.74. As discussed in section II.5, in this setting, solution operators (II.27) and (II.28) exist for each parameter choice $x \in X$.

We also recall the result of [**25**] that there exist trajectories *with respect to the eigenspace*, which are locally analytic around a reference point $x_0 \in X$, if the parameterized bilinear forms of the variational EVP are analytic with respect to the parameter. Thus, the perturbed eigenpair can be represented locally by Taylor series

$$\boldsymbol{\lambda}_{x_\xi} = \boldsymbol{\lambda}_{x_0} + \sum_{k=1}^{\infty} \frac{1}{k!} \mathbf{D}^k_{x_0} \boldsymbol{\lambda}[\xi] \,, \qquad \boldsymbol{u}_{x_\xi} = \boldsymbol{u}_{x_0} + \sum_{k=1}^{\infty} \frac{1}{k!} \mathbf{D}^k_{x_0} \boldsymbol{u}[\xi] \tag{III.1}$$

with $x_\xi = x_0 + \xi \in B(x_0)$ in a neighborhood of $x_0 \in X$. Recall that the bold vector notation (II.29) refers to the m eigenfunctions of a common eigenspace of multiplicity $m \in \mathbb{N}$ at $x_0 \in X$ and its eigenvalues are represented by a matrix $\boldsymbol{\lambda} \in \mathbb{R}^{m\times m}$. The eigenpair $(\boldsymbol{\lambda}_{x_0}, \boldsymbol{u}_{x_0})$ is associated to the reference point $x_0 \in X$. We shall refer to it as the **reference** or **initial eigenpair**. As discussed in section II.5, stating such a reference eigenpair requires choices for the eigenfunctions, i.e., a choice of sign and, if the eigenvalue is degenerate, a choice of orthogonal basis of the eigenspace. The eigenpair $(\boldsymbol{\lambda}_{x_\xi}, \boldsymbol{u}_{x_\xi})$ is called the **perturbed eigenpair**, associated with $x_\xi \in B(x_0)$ and defined by the series (III.1), i.e., the eigenfunction $\boldsymbol{u}_{x_\xi}$ depends on the choice of eigenfunctions we use for the reference eigenfunction at $x_0 \in X$ and the change prescribed by the derivatives, which we are yet to determine.

Following [**25**], we characterize derivatives of eigenpairs with respect to eigenspaces, however, we extend this characterization to derivatives of arbitrary order. As in [**25**], we first recall the characterization of the derivatives in terms of a saddle point equation. In order to provide a better understanding of the concept of derivatives with respect to eigenspaces, we additionally present the explicit representation of derivatives of the i-th eigenfunction $\mathbf{D}^k_{x_0}(u)_i$ with respect to its eigenspace in terms of the eigenbasis $((u_{x_0})_j)_{j\in\mathbb{N}}$ of all eigenfunctions at x_0.

After characterizing the derivatives with respect to the eigenspace, we discuss the derivatives of eigenpairs in the traditional sense, i.e., where each eigenvalue is a scalar,

on a one-dimensional, analytic parameter path $\gamma : [t_0, t_1] \to B(x_0)$. To the best knowledge of the author, such derivatives have only been characterized in cases where either the first- or second-order derivatives of the eigenvalues do not repeat, cf. [**19, 32, 55, 65, 66, 68, 73, 88**][1], or resorting to non-linear iteration approximations, cf. [**42**]. This is despite the eigenvalues being analytic in such a setting, cf. [**59, 83**], and despite k-th derivatives depending k-linearly on the perturbation, cf. section II.3. The procedure of finding a linear combination of the eigenfunctions, so that continuity on the parameter path is preserved, is called **polarization** [**25, 42**]. Extending the characterization of [**25**], we characterize the derivatives of the polarization up to an arbitrary order, providing a way to compute the derivatives of the eigenvalues without the above limitations.
The approximation of this pathwise analytic polarization links analytic trajectories of eigenpairs with respect to the eigenspace and otherwise. We discuss some immediate consequences for the local bifurcation behavior of analytic eigenvalues that can be inferred from the polarization.
Finally, we discuss the implementation of the characterization of derivatives with respect to the eigenspace as saddle point problems, following [**8**]. We end this section with numerical examples that demonstrate the validity and efficiency of the characterizations of the derivatives of the eigenpairs with respect to the eigenspaces and of the polarization. The discussion of stochastic parameters and results in terms of the uncertainty quantification of [**25**] follows in chapter IV.

III.1. Fréchet Derivatives of Eigenpairs with Respect to the Eigenspace

We start with the existence result for EVPs in real Hilbert spaces that have a solution operator of the form (II.27) and (II.28) given assumption II.74. For the convenience of the reader, we restate assumption II.74 for analytically parameterized bilinear forms.

ASSUMPTION III.1. Consider X a Banach space, $B(x_0) \subset X$ an open neighborhood of a reference point $x_0 \in X$, and $V \subset H$ real Hilbert spaces with embedding

$$\mathrm{Id} : V \hookrightarrow H \text{ compact, dense.}$$

Consider

$$\begin{aligned} a &: B(x_0) \to \mathscr{L}^{(2)}(V;\mathbb{R}) \text{ analytic,} \\ &\qquad x \mapsto a(\cdot,\cdot;x) \text{ a } V\text{-elliptic, continuous, symmetric bilinear form,} \\ b &: B(x_0) \to \mathscr{L}^{(2)}(H;\mathbb{R}) \text{ analytic,} \\ &\qquad x \mapsto b(\cdot,\cdot;x) \text{ an } H\text{-elliptic, continuous, symmetric scalar product,} \end{aligned}$$

[1] [**55**] provides an algorithm for higher-order derivatives, which, however, does not cover the general case, where degenerate eigenvalues also share their lowest-order derivatives, extending the approach of [**19**].

where the properties are to be understood as uniformly over $B(x_0)$ Then, the parameterized EVP is to find $(\lambda_x, u_x) \in \mathbb{R} \times V$ such that

$$a\big(u_x, v; x\big) = b\big(u_x, v; x\big)\, \lambda_x \qquad \forall\, v \in V \tag{III.2a}$$

holds for all $x \in B(x_0)$. From section II.5 we know that for each $x \in B(x_0)$, the compact normal solution operator (II.27) and the solution operator (II.28) can be constructed. We also assume the eigenfunctions u_i to be orthonormal

$$b\big((u_x)_i, (u_x)_j; x\big) = \delta_{ij} \qquad \forall\, x \in B(x_0), \quad i, j \in \mathbb{N}\,. \tag{III.2b}$$

Fixing $x = x_0$ we can collect all eigenpairs of the same eigenspace with multiplicity m at $x_0 \in X$ using the vector notation as

$$a\big(\boldsymbol{u}_x, \boldsymbol{v}; x\big) = b\big(\boldsymbol{u}_x, \boldsymbol{v}; x\big) \cdot \boldsymbol{\lambda}_x \qquad \forall\, \boldsymbol{v} \in V^m\,, \tag{III.3a}$$

$$b\big(\boldsymbol{u}_x, \boldsymbol{u}_x; x\big) = \boldsymbol{I}\,, \tag{III.3b}$$

with $\boldsymbol{\lambda}_x = \lambda_x \boldsymbol{I} \in \mathbb{R}^{m\times m}$.

From the discussion of section II.5 we know that we can analyze the eigenvalues of the variational EVP using the solution operator (II.27) which is compact and self-adjoint, so theorem II.67 applies. It can be shown that the eigenvalues of a continuously perturbed operator are also locally continuous, cf. [**59**, Chapter 4.3.5]. Via the solution operator, this also holds for the variational EVP if the bilinear forms are continuously perturbed. As we vary the parameter, the eigenvalues may split into eigenvalues with smaller multiplicity. Thus, the sum of the multiplicities of eigenvalues that emerged by splitting is the multiplicity of the previously degenerate eigenvalue. In reverse, different eigenvalues can also combine to form a degenerate eigenvalue of higher multiplicity. We have already observed this in example I.2. The following result, in contrast to this separate point of view, considers the trajectories with respect to the eigenspaces in analogy to [**25**]. It takes inspiration from [**43, 44, 95**], which also considered isolated eigenspaces to preserve regularity.

THEOREM III.2 ([**25**, Theorem 2.4 & Corollary 2.5]). *Consider the setting of the parameterized EVP in assumption III.1. Let λ_{x_0} be an m-fold eigenvalue of* (III.2) *at x_0 with eigenspace U_{x_0} and $b\big(\cdot,\cdot; x_0\big)$-orthonormal eigenbasis $\boldsymbol{u}_{x_0}$. Then there exists a local, analytic trajectory on a neighborhood*[2] *$B(x_0) \subset X$ of $x_0 \in X$*

$$(\boldsymbol{\lambda}, \boldsymbol{u}) : B(x_0) \to \mathbb{R}^{m\times m} \times V^m \qquad x \mapsto (\boldsymbol{\lambda}_x, \boldsymbol{u}_x)$$

such that $(\boldsymbol{\lambda}_x, \boldsymbol{u}_x)$ satisfies (III.3) *(with $\boldsymbol{\lambda}_x$ not necessarily being diagonal) and, at x_0 it holds*

$$(\boldsymbol{\lambda}_{x_0}, \boldsymbol{u}_{x_0}) = (\lambda_{x_0} \boldsymbol{I}, \boldsymbol{u}_{x_0})\,.$$

[2] This neighborhood is a subset of the neighborhood for which assumption III.1 holds.

Moreover, unique trajectories may be selected by choosing

$$[\boldsymbol{u}]_i - [\boldsymbol{u}_{x_0}]_i \in U_{x_0}^{\perp} \cup \operatorname{span}([\boldsymbol{u}_{x_0}]_i) \qquad \forall i = 1, \dots, m \,, \tag{III.4}$$

i.e., such that the trajectories of the eigenfunction trajectories are locally $b(\cdot,\cdot;x_0)$-orthogonal to other eigenfunctions of the degenerate eigenspace.

PROOF. Recall the one-to-one correspondence of $a \in C^\omega(B(x_0); \mathscr{L}^{(2)}(V;\mathbb{R}))$ to

$$K \in C^\omega(B(x_0); \mathscr{L}(V;V')) \,, \qquad x \mapsto K_x$$

and $b \in C^\omega(B(x_0); \mathscr{L}^{(2)}(H;\mathbb{R}))$ to

$$M \in C^\omega(B(x_0); \mathscr{L}(H)) \,, \qquad x \mapsto M_x$$

and the corresponding generalized operator EVP in V', cf. section II.5.
Let $U_{x_0}^{\perp}$ be the $b(\cdot,\cdot;x_0)$-orthogonal complement of U_{x_0} in V. We split the operators into the action on $U_{x_0} \oplus U_{x_0}^{\perp} = V$ onto $V' = U_{x_0}' \oplus (U_{x_0}^{\perp})'$,

$$K_x : U_{x_0} \oplus U_{x_0}^{\perp} \to (U_{x_0})' \oplus (U_{x_0}^{\perp})' \,, \qquad (u_{x_0}, u_{x_0}^{\perp}) \mapsto \begin{bmatrix} K_x^{0,0} & K_x^{0,\perp} \\ K_x^{\perp,0} & K_x^{\perp,\perp} \end{bmatrix} \begin{bmatrix} u_{x_0} \\ u_{x_0}^{\perp} \end{bmatrix} \,,$$

$$M_x : U_{x_0} \oplus U_{x_0}^{\perp} \to (U_{x_0})' \oplus (U_{x_0}^{\perp})' \,, \qquad (u_{x_0}, u_{x_0}^{\perp}) \mapsto \begin{bmatrix} M_x^{0,0} & M_x^{0,\perp} \\ M_x^{\perp,0} & M_x^{\perp,\perp} \end{bmatrix} \begin{bmatrix} u_{x_0} \\ u_{x_0}^{\perp} \end{bmatrix} \,,$$

and remark that $K_x^{0,\perp} = (K_x^{\perp,0})'$ and $M_x^{0,\perp} = (M_x^{\perp,0})'$ and the diagonal blocks are self-adjoint in the H-inner product due to the self-adjointness of K_x and M_x. Similarly, we remark that a finite system of eigenvalues separated from the rest of the spectrum changes locally continuous under perturbation, cf. [**59**, Theorem 3.16 & Chapter 4.3.5.], applied to solution operator (II.27). That is, there is a sufficiently small neighborhood of x_0 for which there is a continuous mapping $x \mapsto (\boldsymbol{\lambda}_x, \boldsymbol{u}_x)$ with $x_0 \mapsto (\lambda_{x_0} \boldsymbol{I}, \boldsymbol{u}_{x_0})$ and $(\boldsymbol{\lambda}_x, \boldsymbol{u}_x)$ being the solution to (III.3). Moreover, in this neighborhood of x_0, there are, counting multiplicities, exactly m solutions to (III.2) with an eigenvalue in a neighborhood of λ_{x_0}. We denote the m-dimensional space spanned by the corresponding eigenfunctions as U_x and its $b(\cdot,\cdot;x)$-orthogonal complement as $U_x^{\perp}$. The $b(\cdot,\cdot;x)$-orthogonality of U_x and $U_x^{\perp}$ implies that K_x and M_x have at least one block-diagonal representation in $U_x \oplus U_x^{\perp}$ and we claim that in a neighborhood of x_0 there exist

$$W_x \in \mathscr{L}((U_{x_0})'; (U_x^{\perp})') \,, \qquad Z_x \in \mathscr{L}(U_x; U_{x_0}^{\perp}) \,,$$

depending on x such that

(III.5a)
$$\begin{bmatrix} (P^0)' & (Z_x)' \\ W_x & (P^\perp)' \end{bmatrix} \begin{bmatrix} K_x^{0,0} & K_x^{0,\perp} \\ K_x^{\perp,0} & K_x^{\perp,\perp} \end{bmatrix} \begin{bmatrix} P^0 & (W_x)' \\ Z_x & P^\perp \end{bmatrix} = \begin{bmatrix} K_x^{(1)} & 0 \\ 0 & K_x^{(2)} \end{bmatrix} : U_x \oplus U_x^\perp \to (U_x)' \oplus (U_x^\perp)',$$

(III.5b)
$$\begin{bmatrix} (P^0)' & (Z_x)' \\ W_x & (P^\perp)' \end{bmatrix} \begin{bmatrix} M_x^{0,0} & M_x^{0,\perp} \\ M_x^{\perp,0} & M_x^{\perp,\perp} \end{bmatrix} \begin{bmatrix} P^0 & (W_x)' \\ Z_x & P^\perp \end{bmatrix} = \begin{bmatrix} M_x^{(1)} & 0 \\ 0 & M_x^{(2)} \end{bmatrix} : U_x \oplus U_x^\perp \to (U_x)' \oplus (U_x^\perp)',$$

with the $b(\cdot,\cdot;x_0)$-orthogonal projections $P^0 : U_x \to U_{x_0}$ and $P^\perp : U_x^\perp \to U_{x_0}^\perp$ onto U_{x_0} and $U_{x_0}^\perp$ and

$$\begin{aligned}
K_x^{(1)} &= (P^0)' K_x^{0,0} P^0 + (Z_x)' K_x^{\perp,0} P^0 + (P^0)' K_x^{0,\perp} Z_x + (Z_x)' K_x^{\perp,\perp} Z_x \,, \\
K_x^{(2)} &= W_x K_x^{0,0} (W_x)' + W_x K_x^{0,\perp} P^\perp + (P^\perp)' K_x^{\perp,0} (W_x)' + (P^\perp)' K_x^{\perp,\perp} P^\perp \,, \\
M_x^{(1)} &= (P^0)' M_x^{0,0} P^0 + (Z_x)' M_x^{\perp,0} P^0 + (P^0)' M_x^{0,\perp} Z_x + (Z_x)' M_x^{\perp,\perp} Z_x \,, \\
M_x^{(2)} &= W_x M_x^{0,0} (W_x)' + W_x M_x^{0,\perp} P^\perp + (P^\perp)' M_x^{\perp,0} (W_x)' + (P^\perp)' M_x^{\perp,\perp} P^\perp \,.
\end{aligned}$$

Using the fact that K_x and M_x are self-adjoint, we directly note that the necessary condition for W_x and Z_x for such a diagonal representation to hold is that the off-diagonal blocks of the matrix products must vanish. That is,

$$\mathfrak{F}(x,W,Z) := \begin{bmatrix} F(x,W,Z) \\ G(x,W,Z) \end{bmatrix} = 0 \,,$$

where

$$F, G : X \times \mathscr{L}((U_{x_0})'; (U_x^\perp)') \times \mathscr{L}(U_x; U_{x_0}^\perp) \to \mathscr{L}(U_x; (U_x^\perp)')$$

are given by

$$\begin{aligned}
F(x,W,Z) &= (P^\perp)' K_x^{\perp,0} P^0 + (P^\perp)' K_x^{\perp,\perp} Z + W K_x^{0,0} P^0 + W K_x^{0,\perp} Z \,, \\
G(x,W,Z) &= (P^\perp)' M_x^{\perp,0} P^0 + (P^\perp)' M_x^{\perp,\perp} Z + W M_x^{0,0} P^0 + W M_x^{0,\perp} Z \,.
\end{aligned}$$

Since $\mathscr{F}(x_0,0,0) = 0$, we have shown the claim if we verify the assumptions of the implicit function theorem, cf. theorem II.45. To that end, we note that

$$\mathbf{D}_{(x_0,0,0)} \mathfrak{F}[0,\cdot,\cdot] : \mathscr{L}((U_{x_0})'; (U_{x_0}^\perp)') \times \mathscr{L}(U_{x_0}; U_{x_0}^\perp) \to \mathscr{L}(U_{x_0}; (U_{x_0}^\perp)') \times \mathscr{L}(U_{x_0}; (U_{x_0}^\perp)')$$

with

$$(\Xi,\Upsilon) \mapsto \mathbf{D}_{(x_0,0,0)} \mathfrak{F}[0,\Xi,\Upsilon] = \begin{bmatrix} \Xi K_{x_0}^{0,0} & K_{x_0}^{\perp,\perp}\Upsilon \\ \Xi M_{x_0}^{0,0} & M_{x_0}^{\perp,\perp}\Upsilon \end{bmatrix} = \begin{bmatrix} \lambda_{x_0}(\,\cdot\, M_{x_0}^{0,0}) & K_{x_0}^{\perp,\perp} \\ \cdot\, M_{x_0}^{0,0} & M_{x_0}^{\perp,\perp} \end{bmatrix} \begin{bmatrix} \Xi \\ \Upsilon \end{bmatrix} , \tag{III.6}$$

since P^0 acts as the identity on U_{x_0} and $P^\perp$ as the identity on $U_x^\perp$. To show that (III.6) is an isomorphism, we note that solving (III.2) at x_0 is equivalent to computing the eigenpairs of the compact solution operator $S_{x_0} = K_{x_0}^{-1} M_{x_0} : H \to H$, cf. (II.28). Fredholm's alternative (theorem II.21) applied to $S_{x_0} - \lambda_{x_0}^{-1}$ implies that $K_{x_0}^{\perp,\perp} - \lambda_{x_0} M_{x_0}^{\perp,\perp}$ is

bounded invertible and, using Gaussian elimination, this means that (III.6) is an isomorphism. Upon noting that $x \mapsto K_x$ and $x \mapsto M_x$ are analytic, the implicit function theorem yields that W_x and Z_x locally exist and are analytic.
From the construction of (III.5), we can infer that we can find a representation (III.4). To this end, first, consider the possibility that the scalar product $b(\cdot,\cdot;x)$ is constant. Then, we can confirm from the mappings in (III.5) that we can find a unique representation

$$[\boldsymbol{u}]_i - [\boldsymbol{u}_{x_0}]_i \in U_{x_0}^{\perp} .$$

Now, if the scalar product $b(\cdot,\cdot;x)$ is not constant, each eigenfunction $[\boldsymbol{u}]_i$ must also be able to rescale to remain normed with respect to the norm induced by the scalar product, which yields (III.4).
We turn to the analyticity of $\boldsymbol{\lambda}$, which has yet to be proven. Symmetry and ellipticity of $b(\cdot,\cdot;x)$ imply that, in $B(x_0)$, $b(\cdot,\cdot;x) \in \mathbb{R}^{m\times m}$ is an invertible matrix such that analyticity of $\boldsymbol{\lambda}$ follows from testing (III.3) with $\boldsymbol{u}$ and solving for $\boldsymbol{\lambda}$. □

In the non-degenerate case, we get locally analytic standard eigenpair trajectories, cf. [**5, 59, 69, 83**].

COROLLARY III.3 ([**25**, Corollary 2.6]). *Let the assumptions of theorem III.2 hold and* (λ_{x_0}, u_{x_0}) *be an eigenpair of* (III.2) *at* x_0 *with non-degenerate eigenvalue. Then there exists a unique local analytic trajectory defined on a neighborhood* $B(x_0) \subset X$ *of* x_0*,*

$$(\lambda, u) : B(x_0) \to \mathbb{R} \times V , \qquad x \mapsto (\lambda_x, u_x)$$

such that (λ_x, u_x) *satisfies* (III.2) *and* $(\lambda_x, u_x) = (\lambda_{x_0}, u_{x_0})$ *at* x_0*.*

The following generalizations apply assuming less regularity.

COROLLARY III.4 ([**25**, Corollary 2.5]). *Let the assumptions of theorem III.2 hold, but with lower regularity of the bilinear forms. Then the regularity of the mapping* $x \mapsto (\boldsymbol{\lambda}_x, \boldsymbol{u}_x)$ *is the least regularity of the two bilinear forms. The same applies to the non-degenerate case of corollary III.3.*

PROOF. The combined regularity of the two bilinear forms is their least regularity. Revisiting the implicit function theorem II.45, we find that given less than analyticity, the solution still exists, has the regularity of the implicit function. The rest is analogous to theorem III.2. □

To illustrate trajectories with respect to the eigenspace, we return to example I.2.

EXAMPLE III.5. We show the trajectories of the eigenpairs of example I.2 with respect to the eigenspace with reference point $x_0 = (0,0)$. For direct comparison with the eigenvalues trajectories illustrated in fig. I.2, we consider the same paths (I.6). As

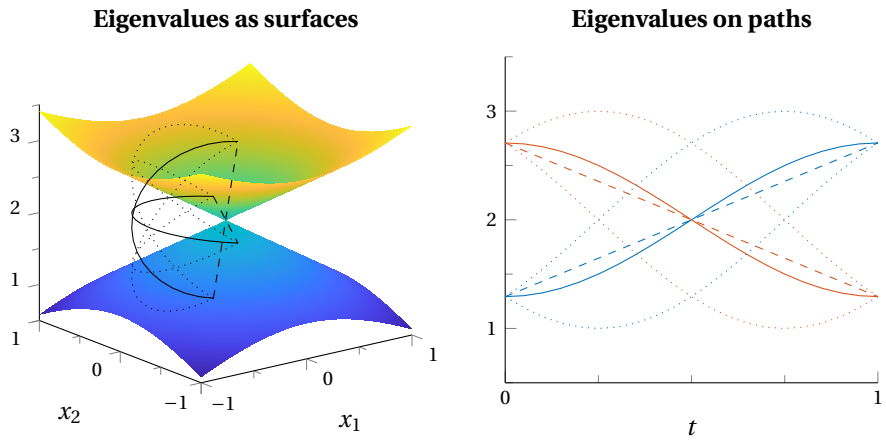

FIGURE III.1. Eigenvalue trajectories of example I.2 with respect to the eigenspace. The off-diagonal entries of $\boldsymbol{\lambda}$ are represented by intervals (dotted) relating to Gershgorin circles. For the actual eigenvalues see fig. I.2.

there are only two eigenvalues whose eigenspace we consider, the scalar product is constant, so we may simplify the eigenpair trajectories to

$$\boldsymbol{\lambda}_x = \boldsymbol{u}_{x_0}^\top \cdot K_x \cdot \boldsymbol{u}_{x_0} \,, \qquad\qquad \boldsymbol{u}_x = \boldsymbol{u}_{x_0} \,.$$

Note that the arbitrary choice $\boldsymbol{u}_{x_0}$ determines the final trajectory. The off-diagonal entries of $\boldsymbol{\lambda}_x$ do not vanish, except on the line $x_1 = x_2$. We can represent the off-diagonal entries on the circular path by Gershgorin circles, cf. theorem II.71, which are evaluated at each parameter value on the paths. The trajectories in terms of Gershgorin circles are illustrated in fig. III.1.

This example is rather extreme, since the degenerate eigenspace includes all eigenvalues of the matrix. In the general case, there are other eigenvalues that influence the evolution of the trajectories with respect to the eigenspace so that the trajectories are less trivial.

III.2. Characterization of Derivatives

An early characterization of the derivatives of non-degenerate eigenpairs was given by Nelson [**68**]. Later, the series of articles [**19, 65, 66, 73**] considered the case of degenerate eigenvalues of which Dailey's method [**19**] is still used frequently. As pointed out by [**32, 88**], Dailey's method, however, implicitly assumes that the derivatives of the eigenvalues are distinct. The eigenvalue derivatives appear in an auxiliary EVP as

eigenvalues themselves and, as such, are assumed to be non-degenerate. In [**32, 88**], the more explicit assumption is made that at least second-order derivatives are non-degenerate. To sidestep these issues, some authors, e.g. [**42**], use a non-linear characterization and iterative solvers instead. In [**55**] higher-order derivatives of eigenpairs are characterized, but the approach to degenerate eigenvalues follows [**19**] and thus cannot hold in the general case.
Following [**25**], we characterize the derivatives with respect to the eigenspace, which for non-degenerate eigenpairs coincides with [**55**]. In section III.3 we provide a general linear characterization to determine the derivatives of the eigenpair, given the derivatives with respect to the eigenspace. This two-stage characterization provides a clearer presentation of the additional conditions for degenerate eigenpairs in the general case than [**19, 32, 88**].
As in [**25**], we start with the simpler problem of finding derivatives to non-degenerate eigenvalues and their eigenfunctions. We develop the characterization of derivatives using the variational formulation of assumption III.1. From our discussion in section II.5, it is clear that an equivalent operator or matrix version can also be formulated.

III.2.1. non-degenerate Eigenvalues. The following formula on the derivatives of non-degenerate eigenvalues are well-known, cf. [**76, 86**]. In quantum mechanics it is called the Hellmann–Feynman theorem, cf. [**17**, Chapter 11.G].

THEOREM III.6 (**Hellmann–Feynman theorem**, [**25**, 2.7]). *Given assumption III.1, let* (λ_{x_0}, u_{x_0}) *be a non-degenerate eigenpair of* (III.2). *Then it holds that*

$$\mathbf{D}_{x_0}\lambda = \big(\mathbf{D}_{x_0} a\big(u_{x_0}, u_{x_0}; \cdot\big)\big) - \big(\mathbf{D}_{x_0} b\big(u_{x_0}, u_{x_0}; \cdot\big)\big)\ \lambda_{x_0}\ . \tag{III.7}$$

PROOF. We consider the derivative of the Rayleigh quotient, cf. theorem II.76, using the quotient and product rule and the linearity of the first two arguments to calculate

$$\begin{aligned}
\mathbf{D}_{x_0}\lambda &= \mathbf{D}_{x_0}\frac{a\big(u,u;\cdot\big)}{b\big(u,u;\cdot\big)} \\
&= \frac{\big(a\big(\mathbf{D}_{x_0}u, u_{x_0}; x_0\big) + a\big(u_{x_0}, \mathbf{D}_{x_0}u; x_0\big) + \mathbf{D}_{x_0} a\big(u_{x_0}, u_{x_0}; \cdot\big)\big)\, b\big(u_{x_0}, u_{x_0}; x_0\big)}{b\big(u_{x_0}, u_{x_0}; x_0\big)^2} \\
&\quad - \frac{a\big(u_{x_0}, u_{x_0}; x_0\big)\big(b\big(\mathbf{D}_{x_0}u, u_{x_0}; x_0\big) + b\big(u_{x_0}, \mathbf{D}_{x_0}u; x_0\big) + \mathbf{D}_{x_0} b\big(u_{x_0}, u_{x_0}; \cdot\big)\big)}{b\big(u_{x_0}, u_{x_0}; x_0\big)^2} \\
&= a\big(\mathbf{D}_{x_0}u, u_{x_0}; x_0\big) + a\big(u_{x_0}, \mathbf{D}_{x_0}u; x_0\big) + \mathbf{D}_{x_0} a\big(u_{x_0}, u_{x_0}; \cdot\big) \\
&\quad - \big(b\big(\mathbf{D}_{x_0}u, u_{x_0}; x_0\big) + b\big(u_{x_0}, \mathbf{D}_{x_0}u; x_0\big) + \mathbf{D}_{x_0} b\big(u_{x_0}, u_{x_0}; \cdot\big)\big)\ \lambda_{x_0}\ .
\end{aligned}$$

In the last line, we identify the eigenvalue again by the Rayleigh coefficient. Realizing that $\mathbf{D}_{x_0}u \in V$ lets the second terms cancel as we find the main condition of the EVP. Due to symmetry, the first terms cancel as well, so we arrive at the formula. □

III.2.2. Eigenpairs with non-degenerate Eigenvalues. In order to also determine the derivative of the eigenfunctions, we calculate the derivative of the main condition (III.2a) of the variational EVP by using the derivation rules and the linearity of the first two arguments, i.e.,

$$\begin{aligned} a\big(\mathbf{D}_{x_0} u, v; x_0\big) - b\big(\mathbf{D}_{x_0} u, v; x_0\big)\, \lambda_{x_0} - b\big(u_{x_0}, v; x_0\big)\, (\mathbf{D}_{x_0} \lambda) \\ = -\big(\mathbf{D}_{x_0} a\big(u_{x_0}, v; \cdot\big)\big) + \big(\mathbf{D}_{x_0} b\big(u_{x_0}, v; \cdot\big)\big)\, \lambda_{x_0} \qquad \forall v \in V \end{aligned} \tag{III.8a}$$

Note that setting $v = u_{x_0}$ recovers the Hellmann–Feynman formula (III.7). We call (III.8a) the **main condition** of the derivatives of the EVP since it is derived from the main condition of the EVP. This main condition does not yet uniquely characterize the derivative of the eigenfunction $\mathbf{D}_{x_0} u$ since

$$\mathbf{D}_{x_0} u + c\, u_{x_0}$$

solves (III.8a) for all $c \in \mathbb{R}$. Taking the derivative of the normalization condition of the EVP (II.22b) we get

$$-b\big(\mathbf{D}_{x_0} u, u_{x_0}; x_0\big) - b\big(u_{x_0}, \mathbf{D}_{x_0} u; x_0\big) = \mathbf{D}_{x_0} b\big(u_{x_0}, u_{x_0}; \cdot\big)\,.$$

Note that $\mathbf{D}_{x_0} b\big(u_{x_0}, u_{x_0}; x_0\big) \in \mathbb{R}$ and $b\big(u_{x_0}, \mathbf{D}_{x_0} u; x_0\big) = b\big(\mathbf{D}_{x_0} u, u_{x_0}; x_0\big)$, since $b\big(\cdot, \cdot; x\big)$ is a scalar product of a real Hilbert space. Thus the equation only holds, when the linear **normalization condition**

$$-b\big(\mathbf{D}_{x_0} u, u_{x_0}; x_0\big) = \frac{\mathbf{D}_{x_0} b\big(u_{x_0}, u_{x_0}; \cdot\big)}{2} = -b\big(u_{x_0}, \mathbf{D}_{x_0} u; x_0\big) \tag{III.8b}$$

is met. The main and normalization condition form a saddle point equation. We gather the terms in auxiliary functions for a variational saddle point equation form and to prove their unique solvability.

LEMMA III.7 ([**25**, Lemma 2.8]). *Consider the bilinear forms*

$$\begin{aligned} &\mathfrak{A} : V \times V \to \mathbb{R}\,, && u \times v \mapsto \mathfrak{A}(u, v) \coloneqq a\big(u, v; x_0\big) - b\big(u, v; x_0\big)\, \lambda_{x_0}\,, \\ &\mathfrak{B} : \mathbb{R} \times V \to \mathbb{R}\,, && \zeta \times v \mapsto \mathfrak{B}(\zeta, v) \coloneqq \zeta\, b\big(u_{x_0}, v; x_0\big)\,. \end{aligned}$$

Then the saddle point problem

$$\begin{aligned} \mathfrak{A}(\mathbf{D}_{x_0} u, v) - \mathfrak{B}(\mathbf{D}_{x_0} \lambda, v) &= -\big(\mathbf{D}_{x_0} a\big(u_{x_0}, v; \cdot\big)\big) + \big(\mathbf{D}_{x_0} b\big(u_{x_0}, v; \cdot\big)\big)\, \lambda_{x_0} && \forall v \in V\,, \\ -\mathfrak{B}(\zeta, \mathbf{D}_{x_0} u) &= \frac{\zeta\, \mathbf{D}_{x_0} b\big(u_{x_0}, u_{x_0}; \cdot\big)}{2} && \forall \zeta \in \mathbb{R} \end{aligned}$$

is uniquely solvable.

PROOF. We first show the Ladyzhenskaya–Babuška–Brezzi-condition (LBB-condition) of $\mathfrak{B}$, cf. [**11**]. To this end, we estimate

$$\inf_{0 \neq \zeta \in \mathbb{R}} \sup_{0 \neq v \in V} \frac{\mathfrak{B}(\zeta, v)}{|\zeta|\, \|v\|_V} = \inf_{0 \neq \zeta \in \mathbb{R}} \sup_{0 \neq v \in V} \frac{\zeta b\big(u_{x_0}, v; x_0\big)}{|\zeta|\, \|v\|_V} \geq \frac{b\big(u_{x_0}, u_{x_0}; x_0\big)}{\|u_{x_0}\|_V} \geq \frac{1}{C\sqrt{\lambda_{x_0}}} > 0\,,$$

where we set $v = u_{x_0}$, used the normalization constraint $b\big(u_{x_0}, u_{x_0}; x_0\big) = 1$, and used the V-ellipticity of $a\big(\cdot,\cdot; x_0\big)$ through

$$\|u_{x_0}\|_V \leq C\sqrt{a\big(u_{x_0}, u_{x_0}; x_0\big)} = C\sqrt{\lambda_{x_0} b\big(u_{x_0}, u_{x_0}; x_0\big)} = C\sqrt{\lambda_{x_0}}\,.$$

It remains to show that finding $w \in \mathcal{N}(\mathfrak{B})$ such that

$$\mathfrak{A}(w, v) = \ell(v)$$

for all $v \in \mathcal{N}(\mathfrak{B})$ with

$$\mathcal{N}(\mathfrak{B}) = \{\, v \in V \,:\, \mathfrak{B}(\zeta, v) = 0 \text{ for all } \zeta \in \mathbb{R}\} = \operatorname{span}\Big(\big\{\, u_{x_0}\big\}^{\perp}\Big)$$

is uniquely solvable for all $\ell \in (\mathcal{N}(\mathfrak{B}))'$. Similarly to the proof of theorem III.2, this follows from the Fredholm alternative applied to $S_{x_0} - \lambda_{x_0}^{-1}$ with S_{x_0} the solution operator, cf. (II.28). □

III.2.3. Eigenpairs with Degenerate Eigenvalues. We now move on to degenerate eigenvalues and their derivatives with respect to the eigenspace. First, we show that for individual eigenpairs, which are part of a degenerate subspace, the saddle point equation of lemma III.7 does not provide a unique solution. Then we show how we can instead characterize the derivatives with respect to the eigenspace according to theorem III.2, which are unique.
Taking the derivative of the main condition of the EVP in vectorized form for all eigenpairs in an eigenspace at x_0 yields the vectorized equation

$$\begin{aligned}\text{(III.11a)}\qquad & a\big(\mathbf{D}_{x_0}\boldsymbol{u}, \boldsymbol{v}; x_0\big) - b\big(\mathbf{D}_{x_0}\boldsymbol{u}, \boldsymbol{v}; x_0\big)\cdot\boldsymbol{\lambda}_{x_0} - b\big(\boldsymbol{u}_{x_0}, \boldsymbol{v}; x_0\big)\cdot(\mathbf{D}_{x_0}\boldsymbol{\lambda}) \\ & \qquad = -\big(\mathbf{D}_{x_0} a\big(\boldsymbol{u}_{x_0}, \boldsymbol{v}; \cdot\big)\big) + \big(\mathbf{D}_0 b\big(\boldsymbol{u}_{x_0}, \boldsymbol{v}; \cdot\big)\big)\cdot\boldsymbol{\lambda}_{x_0}\,, \qquad \forall \boldsymbol{v} \in V^m\,.\end{aligned}$$

This **main condition** relates to m equations of the form (III.8a). Analogously, we can formulate m **normalization conditions** of the form (III.8b), i.e.,

$$\text{(III.11b)}\qquad -b\big(\mathbf{D}_{x_0}[\boldsymbol{u}]_i, [\boldsymbol{u}_{x_0}]_i; x_0\big) = \frac{\mathbf{D}_{x_0} b\big([\boldsymbol{u}_{x_0}]_i, [\boldsymbol{u}_{x_0}]_i; \cdot\big)}{2}\,, \qquad i = 1, \ldots, m\,.$$

Due to the degeneracy of the eigenspace, after including the normalization condition, each underlying saddle point equation still has a kernel of size $m-1$, relating to $\operatorname{span}(\{[\boldsymbol{u}_{x_0}]_j, j \neq i = 1, \ldots, m\})$. That is, if $(\mathbf{D}_{x_0}[\boldsymbol{u}]_i)_{i=1,\ldots,m}$ satisfy (III.11a) and (III.11b), then so do

$$\mathbf{D}_{x_0}[\boldsymbol{u}]_i + \sum_{\substack{j=1 \\ j\neq i}}^{m} c_j [\boldsymbol{u}_{x_0}]_j \qquad\qquad c_j \in \mathbb{R}, \quad i = 1, \ldots, m\,.$$

From theorem III.2 we know that choosing the trajectories orthogonal with respect to the other eigenfunctions in the same eigenspace yields a unique representation of

derivatives with respect to the eigenspace. This is expressed by **orthogonality conditions**

$$b\big(\mathbf{D}_{x_0}[\boldsymbol{u}]_i, [\boldsymbol{u}_{x_0}]_j; x_0\big) = 0 \qquad i \neq j, \quad i,j = 1,\dots,m. \tag{III.11c}$$

In order to keep the vectorized notation we summarize (III.11b) and (III.11c) as an **orthonormality condition**

$$-b\big(\mathbf{D}_{x_0}\boldsymbol{u}, \boldsymbol{u}_{x_0}; x_0\big) = \operatorname*{diag}_{i=1,\dots,m} \frac{\mathbf{D}_{x_0} b\big([\boldsymbol{u}_{x_0}]_i, [\boldsymbol{u}_{x_0}]_i; \cdot\big)}{2}\,. \tag{III.11d}$$

Here, the right hand side is to be read as a diagonal matrix with entries as in (III.11b). This vectorized system of equations is again in saddle point equation form. We formally prove its solvability in the following theorem.

THEOREM III.8 ([**25**, Theorem 2.9]). *Given assumption III.1, let λ_{x_0} be an eigenvalue of multiplicity m at x_0 of* (III.2) *with $b(\cdot,\cdot;x_0)$-orthonormal eigenbasis $\boldsymbol{u}_{x_0}$. Let $x \mapsto (\boldsymbol{\lambda}, \boldsymbol{u})$ be the unique locally analytic trajectory such that $(\boldsymbol{\lambda}_x, \boldsymbol{u}_x)$ satisfies* (III.3) *at x and $(\boldsymbol{\lambda}_x, \boldsymbol{u}_x) = (\boldsymbol{\lambda}_{x_0}, \boldsymbol{u}_{x_0})$ holds at x_0. Consider the auxiliary functions*

$$\mathfrak{A} : V \times V \to \mathbb{R}\,, \qquad u \times v \mapsto \mathfrak{A}(u,v) \coloneqq a\big(u,v;x_0\big) - b\big(u,v;x_0\big)\,\lambda_{x_0}\,, \tag{III.12a}$$

$$\mathfrak{B} : \mathbb{R}^m \times V \to \mathbb{R}\,, \qquad \boldsymbol{\zeta} \times v \mapsto \mathfrak{B}(\boldsymbol{\zeta}, v) \coloneqq \sum_{j=1}^{m} [\boldsymbol{\zeta}]_j\, b\big([\boldsymbol{u}_{x_0}]_j, v; x_0\big)\,. \tag{III.12b}$$

Then the derivatives $\mathbf{D}_{x_0}\boldsymbol{u}$ and $\mathbf{D}_{x_0}\boldsymbol{\lambda}$ are uniquely determined via the solutions of the saddle point equation to find $([\mathbf{D}_{x_0}\boldsymbol{\lambda}]_{:i}, [\mathbf{D}_{x_0}\boldsymbol{u}]_i) \in \mathbb{R}^m \times V$ such that

$$\begin{aligned}\mathfrak{A}(\mathbf{D}_{x_0}[u]_i, v) - \mathfrak{B}(\mathbf{D}_{x_0}[\boldsymbol{\lambda}]_{:i}, v)& \\ = -\big(\mathbf{D}_{x_0} a\big([\boldsymbol{u}_{x_0}]_i, v; \cdot\big)\big) + \big(\mathbf{D}_{x_0} b\big([\boldsymbol{u}_{x_0}]_i, v; \cdot\big)\big)\,\lambda_{x_0}\,, &\quad \forall v \in V\,,\end{aligned} \tag{III.13a}$$

$$\mathfrak{B}(\boldsymbol{\zeta}, \mathbf{D}_{x_0}[\boldsymbol{u}]_i) = -\frac{[\boldsymbol{\zeta}]_i\, \mathbf{D}_{x_0} b\big([\boldsymbol{u}_{x_0}]_i, [\boldsymbol{u}_{x_0}]_i; \cdot\big)}{2}\,, \qquad \forall \boldsymbol{\zeta} \in \mathbb{R}^m \tag{III.13b}$$

holds. Therein, by $\mathbf{D}_0[\boldsymbol{\lambda}]_{:i}$ we refer to the i-th column of $\mathbf{D}_0\boldsymbol{\lambda} \in \mathbb{R}^{m\times m}$.

PROOF. The characterization of (III.13) follow from (III.11) where (III.11c) was derived from theorem III.2. In analogy to lemma III.7 we prove the solvability of the saddle point equations (III.13). To this end, setting

$$v_{\boldsymbol{\zeta}} = \sum_{j=1}^{m} [\boldsymbol{\zeta}]_j [\boldsymbol{u}_{x_0}]_j$$

yields

$$\begin{aligned}\|v_{\boldsymbol{\zeta}}\|_V &\le C\sqrt{a\big(v_{\boldsymbol{\zeta}}, v_{\boldsymbol{\zeta}}; x_0\big)} = C\sqrt{\lambda_{x_0} b\big(v_{\boldsymbol{\zeta}}, v_{\boldsymbol{\zeta}}; x_0\big)} \\ &= C\sqrt{\lambda_{x_0} \sum_{i,j=1}^{m} \zeta_i \zeta_j\, b\big([\boldsymbol{u}_{x_0}]_j, [\boldsymbol{u}_{x_0}]_i; x_0\big)} = C\sqrt{\lambda_{x_0}}\|\boldsymbol{\zeta}\|_{\mathbb{R}^m}\end{aligned}$$

with $C > 0$. The LBB-condition of $\mathfrak{B}$ follows from

$$\begin{aligned}\inf_{0\neq\boldsymbol{\zeta}\in\mathbb{R}^m}\sup_{0\neq\boldsymbol{v}\in V}\frac{\mathfrak{B}(\boldsymbol{\zeta},v)}{\|\boldsymbol{\zeta}\|_{\mathbb{R}^m}\|v\|_V} &= \inf_{0\neq\boldsymbol{\zeta}\in\mathbb{R}^m}\sup_{0\neq\boldsymbol{v}\in V}\frac{\sum_{j=1}^m[\boldsymbol{\zeta}]_j\, b\big([\boldsymbol{u}_{x_0}]_j,v;x_0\big)}{\|\boldsymbol{\zeta}\|_{\mathbb{R}^m}\|v\|_V}\\ &\geq \inf_{0\neq\boldsymbol{\zeta}\in\mathrm{R}^m}\frac{\sum_{i,j=1}^m[\boldsymbol{\zeta}]_i[\boldsymbol{\zeta}]_j\, b\big([\boldsymbol{u}_{x_0}]_j,[\boldsymbol{u}_{x_0}]_j;x_0\big)}{\|\boldsymbol{\zeta}\|_{\mathbb{R}^m}\|v\|_V}\\ &\geq \inf_{0\neq\boldsymbol{\zeta}\in\mathrm{R}^m}\frac{\sum\limits_{j=1}^m[\boldsymbol{\zeta}]_j^2}{C\sqrt{\lambda_{x_0}}\|\boldsymbol{\zeta}\|_{\mathbb{R}^m}^2}\geq\frac{1}{C\sqrt{\lambda_{x_0}}}>0\,.\end{aligned}$$

In analogy to lemma III.7, it remains to show that $\mathfrak{A}$ is invertible on the kernel of $\mathfrak{B}$

$$\mathcal{N}(\mathfrak{B})=\big\{\,v\in V\,:\,\mathfrak{B}(\boldsymbol{\zeta},v)=0\text{ for all }\boldsymbol{\zeta}\in\mathbb{R}^m\big\}=\operatorname{span}(\big\{[\boldsymbol{u}_{x_0}]_1,\dots,[\boldsymbol{u}_{x_0}]_m\big\}^{\perp})\,.$$

The unique solvability of $\mathfrak{A}(w,v)=\ell(v)$ for all $\ell\in(\mathcal{N}(\mathfrak{B}))'$ follows again from the Fredholm alternative. □

III.2.4. Higher-order Derivatives. We continue with the systems of equations for the higher-order derivatives, which extend (III.11) using multinomial coefficients (II.8). For ease of notation, we restrict ourselves to non-mixed higher-order derivatives. The system of equations can be formulated in complete analogy for mixed partial derivatives by replacing the scalar orders of the derivatives with multi-indices. We again formulate the system of equations in terms of the vector notation of the eigenspace of multiplicity m. The case $m=1$ again covers derivatives in the traditional sense for the non-degenerate case.

The higher-order derivatives of the eigenpairs can be found in an iterative way. For the characterization of the k-th-order derivatives of the eigenpair, we derive both the main condition (II.29a) and the normalization condition (II.29b) k-times.

In the k-th iteration, the derivatives of the eigenpair up to order $k-1$ are known and we only need to determine the derivatives of order k. Sorting the terms including the derivatives of the highest order k to the left-hand side of the equation yields the same left-hand side as in (III.11) only with new unknown derivatives. The right-hand side changes in each iteration and is given by

$$\begin{aligned}\mathbf{R}(k,\boldsymbol{u},\boldsymbol{v}) \coloneqq &-\Big(\mathbf{D}_{x_0}^k\, a\big(\boldsymbol{u},\boldsymbol{v};\cdot\big)\Big)+\Big(\mathbf{D}_{x_0}^k\, b\big(\boldsymbol{u},\boldsymbol{v};\cdot\big)\Big)\;\lambda_{x_0}\\ &-\sum_{k_u+k_x=k-1}\binom{k}{k_u,k_x}\Big(\mathbf{D}_{x_0}^{k_x}\, a\big(\mathbf{D}_{x_0}^{k_u}\,\boldsymbol{u},\boldsymbol{v};\cdot\big)\Big)\\ &+\sum_{k_u+k_x+k_\lambda=k-1}\binom{k}{k_u,k_x,k_\lambda}\Big(\mathbf{D}_{x_0}^{k_x}\, b\big(\mathbf{D}_{x_0}^{k_u}\,\boldsymbol{u},\boldsymbol{v};\cdot\big)\Big)\cdot\big(\mathbf{D}_{x_0}^{k_\lambda}\,\boldsymbol{\lambda}\big)\end{aligned}\tag{III.14}$$

Thus, the **main condition** is given by

(III.15a)
$$a\big(\mathbf{D}^k_{x_0}\boldsymbol{u},\boldsymbol{v};x_0\big) - b\big(\mathbf{D}^k_{x_0}\boldsymbol{u},\boldsymbol{v};x_0\big)\,\lambda_{x_0} - b\big(\boldsymbol{u}_{x_0},\boldsymbol{v};x_0\big)\cdot\big(\mathbf{D}^k_{x_0}\boldsymbol{\lambda}\big) = \mathbf{R}(k,\boldsymbol{u}_{x_0},\boldsymbol{v}) \quad \forall \boldsymbol{v}\in V^m\,.$$

In complete analogy to the first-order derivative, the normalization condition only provides m conditions on the diagonal, and the other m^2-m orthogonality conditions are provided by our ansatz to consider derivatives with respect to the eigenspace. This leaves us with an **orthonormality condition**

(III.15b)
$$-b\big(\mathbf{D}^k_{x_0}\boldsymbol{u},\boldsymbol{u}_{x_0};x_0\big) = \operatorname*{diag}_{i=1,\dots,m}\frac{\mathbf{D}^k_{x_0}\,b\big([\boldsymbol{u}_{x_0}]_i,[\boldsymbol{u}_{x_0}]_i;\cdot\big)}{2} + \operatorname*{diag}_{i=1,\dots,m}\sum_{k_{u_1}+k_{u_2}+k_x=k-1}\binom{k}{k_{u_1},k_{u_2},k_x}\frac{\mathbf{D}^{k_x}_{x_0}\,b\big(\mathbf{D}^{k_{u_1}}_{x_0}[\boldsymbol{u}]_i,\mathbf{D}^{k_{u_2}}_{x_0}[\boldsymbol{u}]_i;\cdot\big)}{2}\,.$$

Since the left-hand sides of (III.15) are the same with the exception of the unknown derivatives for each iteration $k\in\mathbb{N}$, the solvability of (III.15) follows from the solvability of the first-order derivatives, cf. theorem III.8.

III.2.5. Implicit Conditions and Eigenbasis Representation. The derivative of any individual eigenfunction can be decomposed using the eigenbasis $(u_{x_0})_{i\in\mathbb{N}}$ of eigenfunctions at x_0, i.e.,

(III.16)
$$\mathbf{D}^k_{x_0}(u)_i = \sum_{j=1}^{\infty} b\big(\mathbf{D}^k_{x_0}(u)_i,(u_{x_0})_j;x_0\big)(u_{x_0})_j\,.$$

This basis representation is also sometimes called the *modal superposition method*, cf. [**42**]. The basis coefficients with respect to the eigenfunctions, which are in the same eigenspace as $(u_{x_0})_i$, are given by orthonormality constraints (III.15b). To properly understand the contribution of the main condition of the saddle point equation (III.15a), we derive the remaining basis coefficients.
Considering the main condition in (III.15a) for only the i-th eigenpair, i.e.,

$$a\big(\mathbf{D}^k_{x_0}(u)_i,v;x_0\big) - b\big(\mathbf{D}^k_{x_0}(u)_i,v;x_0\big)\,(\lambda_{x_0})_i - b\big((u_{x_0})_i,v;x_0\big)\cdot\mathbf{D}^k_{x_0}(\lambda)_{iv} = \mathbf{R}(k,(u_{x_0})_i,v)$$

for all $v\in V$, where $\mathbf{D}^k_{x_0}(\lambda)_{iv}\in\mathbb{R}$ is a placeholder. Consider eigenpairs $((\lambda_{x_0})_j,(u_{x_0})_j)$ that are not in the same eigenspace as the i-th eigenpair, i.e., $(\lambda_{x_0})_j\neq(\lambda_{x_0})_i$, and set $v=(u_{x_0})_j$. Since the eigenfunctions are $b\big(\cdot,\cdot;x_0\big)$-orthogonal, the term with the placeholder vanishes, and due to

$$a\big(\mathbf{D}^k_{x_0}(u)_i,(u_{x_0})_j;x_0\big) = b\big(\mathbf{D}^k_{x_0}(u)_i,(u_{x_0})_j;x_0\big)\,(\lambda_{x_0})_j$$

the above equation reduces to

$$b\big(\mathbf{D}^k_{x_0}(u)_i,(u_{x_0})_j;x_0\big)\,\big((\lambda_{x_0})_j-(\lambda_{x_0})_i\big) = \mathbf{R}(k,(u_{x_0})_i,(u_{x_0})_j)\,.$$

Thus, we can summarize the coefficients for (III.16) as

(III.17)
$$b\big(\mathbf{D}^k_{x_0}(u)_i,(u_{x_0})_j;x_0\big)=\begin{cases}\frac{\mathbf{R}(k,(u_{x_0})_i,(u_{x_0})_j)}{(\lambda_{x_0})_j-(\lambda_{x_0})_i}, & \text{for } (\lambda_{x_0})_i\neq(\lambda_{x_0})_j\,,\\ i\text{-th diagonal entry of (III.15b)} & i=j\\ 0, & \text{else}\,.\end{cases}$$

Hence, we have found a more explicit characterization of the derivatives of the eigenfunctions in terms of basis coefficients. The formula (III.17) makes explicit how the gaps $(\lambda_{x_0})_j-(\lambda_{x_0})_i$ control the behavior of the derivatives with respect to the eigenspace. Note that, since we have already proven the existence of such derivatives in theorems III.2 and III.8, we do not have to prove that the sum in (III.16) converges. The downside of this formulation for calculations is that we need to know all unperturbed eigenfunctions, which is only feasible for rather low-dimensional spaces.
If we use (III.16) for the derivatives of the eigenfunctions, we are still missing the corresponding derivatives of the eigenvalues with respect to the eigenspace. Considering the m eigenpairs of the same eigenspace, i.e., $(\lambda_{x_0})_j=(\lambda_{x_0})_i$, provides the missing derivatives as a generalization of the Hellmann–Feynman formula, cf. theorem III.6.

COROLLARY III.9. *The k-th-order derivatives of the eigenvalue with respect to the eigenspace of* (III.15) *are given by*

(III.18)
$$\mathbf{D}^k_{x_0}\boldsymbol{\lambda}=-\mathbf{R}(k,\boldsymbol{u}_{x_0},\boldsymbol{u}_{x_0})\,.$$

PROOF. Set $\boldsymbol{v}=\boldsymbol{u}_{x_0}$ in (III.15a), use the orthonormality of eigenfunctions

$$b\big(\boldsymbol{u}_{x_0},\boldsymbol{u}_{x_0};x_0\big)=\boldsymbol{I}\,,$$

and discard the terms relating to the main condition of the EVP, i.e.,

$$a\big(\mathbf{D}_{x_0}\boldsymbol{u},\boldsymbol{u}_{x_0};x_0\big)=b\big(\mathbf{D}_{x_0}\boldsymbol{u},\boldsymbol{u}_{x_0};x_0\big)\cdot\boldsymbol{\lambda}_{x_0}\,. \qquad \square$$

III.2.6. Complex-valued Hilbert spaces. As discussed in section II.5, the spectral theorem for compact normal operators includes the complexification of real Hilbert spaces, cf. definition II.63. We verify that the previous results are still valid in the context of the complexification.

ASSUMPTION III.10. Let V,H be complex-valued Hilbert spaces. As pointed out in assumption II.74, in this case the bilinear forms must be replaced by sesquilinear forms, i.e., we then consider

$a\big(\cdot,\cdot;x\big):B(x_0)\to\mathscr{L}^{(1.5)}(V;\mathbb{C})$ analytic,
$\qquad x\mapsto a\big(\cdot,\cdot;x\big)$ a V-elliptic, continuous, Hermitian sesquilinear form,
$b\big(\cdot,\cdot;x\big):B(x_0)\to\mathscr{L}^{(1.5)}(H;\mathbb{C})$ analytic,
$\qquad x\mapsto b\big(\cdot,\cdot;x\big)$ an H-elliptic , continuous, Hermitian scalar product,

which form EVPs (III.2) and (III.3) verbatim.

Recall that after normalization of the eigenfunctions, the choice of eigenfunction even for an eigenfunction relating to a non-degenerate eigenvalue is arbitrary up to a complex factor $c \in \mathbb{C}$ with $|c| = 1$, i.e., if we choose u_{x_0}, then cu_{x_0} is also a valid choice. This choice of factor could vary over $x \in B(x_0)$. Keeping the eigenfunctions orthogonal with respect to other eigenfunctions in the degenerate eigenspace due to theorem III.2 is not sufficient to define a unique trajectory. Let $\mathbf{D}^k_{x_0} u$ be a derivative of the eigenfunctions defined as in the real-valued case, then

$$\mathbf{D}^k_{x_0} u + \mathrm{i}\, c u_{x_0} \qquad\qquad c \in \mathbb{R} \tag{III.19}$$

is also an eigenfunction that obeys the normalization condition, since

$$b\big(\mathbf{D}^k_{x_0} u, u_{x_0}; x_0\big) = \overline{b\big(u_{x_0}, \mathbf{D}^k_{x_0} u; x_0\big)}\,.$$

If we select $c = 0$ in (III.19), we again arrive at the characterization that we derived for real-valued Hilbert spaces. Sticking to this characterization has the desirable property that if the unperturbed eigenfunctions u_{x_0} can be chosen to be real-valued, and the right-hand side of the saddle point problem remains real-valued, the derivative, and thus the trajectory will also remain real-valued. Thus, if we take an EVP that obeys the assumptions of the real-valued setting of assumption III.1 and consider its complexification, this choice will lead us to the same result as if we ignored the complexification.

III.3. Polarization

As discussed in the beginning of this chapter, we know that there is also a locally analytic trajectory for each eigenpair in the traditional sense, if the parameter space is one-dimensional, cf. [**83**]. We shall use the convention to denote this version of the eigenvalue trajectory by an *accent grave*, i.e., $\grave{\boldsymbol{\lambda}} \in \mathbb{R}^{m\times m}$, which is in reference to the fact that the eigenvalues in question are on the diagonal of a diagonal matrix, i.e., $\grave{\boldsymbol{\lambda}} = \mathrm{diag}_{i=1,\dots,m} \grave{\lambda}_i$. The corresponding eigenfunctions are also marked with an accent, i.e., $\grave{u}_i$. For non-degenerate eigenpairs, this diagonal representation $(\grave{\lambda}, \grave{u})$ is the same as the representation (λ, u) with respect to the eigenspace, cf. corollary III.3.
If $m > 1$ for each $x \in B(x_0)$, there must be an orthogonal matrix $\boldsymbol{P}_x \in \mathbb{R}^{m\times m}$, which we call the **polarization matrix**, cf. [**42**], such that

$$\boldsymbol{P}_x^\top \cdot \boldsymbol{\lambda}_x \cdot \boldsymbol{P}_x = \grave{\boldsymbol{\lambda}}_x\,.$$

For convenience of notation, we include the path $\gamma : [t_0, t_1] \to X$ in the index to denote the Fréchet derivatives on the parameter path, i.e., to denote compositions

$$\boldsymbol{\lambda}_{\gamma(t)} : [t_0, t_1] \to \mathbb{R}^{m\times m}\,, \qquad \boldsymbol{\lambda}_{\gamma(t)} \coloneqq \boldsymbol{\lambda} \circ \gamma(t)\,, \tag{III.20a}$$

$$\boldsymbol{u}_{\gamma(t)} : [t_0, t_1] \to V^m\,, \qquad \boldsymbol{u}_{\gamma(t)} \coloneqq \boldsymbol{u} \circ \gamma(t)\,, \tag{III.20b}$$

with derivatives

$$\mathbf{D}^k_{\gamma(t_0)}\,\boldsymbol{\lambda} \coloneqq \mathbf{D}^k_{t_0}\,(\boldsymbol{\lambda}\circ\gamma)\,, \tag{III.20c}$$

$$\mathbf{D}^k_{\gamma(t_0)}\,\boldsymbol{u} \coloneqq \mathbf{D}^k_{t_0}\,(\boldsymbol{u}\circ\gamma) \tag{III.20d}$$

according to the chain rule, the derivatives with respect to the eigenspace, and the derivatives of the path. Then there must be a locally analytic trajectory

$$\boldsymbol{P}:[t_0,t_1]\to\mathbb{R}^{m\times m}, \qquad t\mapsto \boldsymbol{P}_t = \boldsymbol{P}_{t_0} + \sum_{k=1}^{\infty}\frac{t^k}{k!}\,\mathrm{D}^k_{t_0}\,\boldsymbol{P}\,, \tag{III.21a}$$

such that

$$\begin{aligned}\dot{\boldsymbol{\lambda}}_t &= \boldsymbol{P}_t^\top\cdot\boldsymbol{\lambda}_{\gamma(t)}\cdot\boldsymbol{P}_t \\ &= \boldsymbol{\lambda}_{x_0} + \sum_{k=1}^{\infty}\frac{t^k}{k!}\underbrace{\sum_{k_\lambda+k_{P_1}+k_{P_2}=k}\binom{k}{k_{P_1},k_{P_2},k_\lambda}(\mathrm{D}^{k_{P_1}}_{t_0}\,\boldsymbol{P})^\top\cdot(\mathbf{D}^{k_\lambda}_{\gamma(t_0)}\,\boldsymbol{\lambda})\cdot(\mathrm{D}^{k_{P_2}}_{t_0}\,\boldsymbol{P})}_{=\mathrm{D}^k_{t_0}\,\dot{\boldsymbol{\lambda}}}\,,\end{aligned} \tag{III.21b}$$

$$\dot{\boldsymbol{u}}_t = \boldsymbol{u}_{\gamma(t)}\cdot\boldsymbol{P}_t = \boldsymbol{u}_{x_0}\cdot\boldsymbol{P}_{t_0} + \sum_{k=1}^{\infty}\frac{t^k}{k!}\underbrace{\sum_{k_u+k_P=k}\binom{k}{k_P,k_u}(\mathbf{D}^{k_u}_{\gamma(t_0)}\,\boldsymbol{u})\cdot(\mathrm{D}^{k_P}_{t_0}\,\boldsymbol{P})}_{=\mathrm{D}^k_{t_0}\,\dot{\boldsymbol{u}}}\,. \tag{III.21c}$$

We know that $\boldsymbol{\lambda}_{x_0}$ is already diagonal with identical entries on the diagonal, thus $\boldsymbol{P}_{t_0}$ is arbitrary if we consider only the information given by the unperturbed eigenpairs, i.e.,

$$\dot{\boldsymbol{\lambda}}_{t_0} \coloneqq \boldsymbol{P}_{t_0}^\top\cdot\lambda_{x_0}\boldsymbol{I}\cdot\boldsymbol{P}_{t_0} = \boldsymbol{\lambda}_{x_0}\,. \tag{III.22}$$

This corresponds to the fact that linear combinations of the eigenfunctions of a degenerate eigenspace are also valid eigenfunctions. Therefore, we must include the information given by the derivatives according to the path to determine $\boldsymbol{P}_{t_0}$.

III.3.1. Initial Polarization Matrix. Using the product rule on the parameterized matrix EVP

$$\boldsymbol{\lambda}_{\gamma(t)}\cdot\boldsymbol{P}_t = \boldsymbol{P}_t\cdot\dot{\boldsymbol{\lambda}}_t \tag{III.23}$$

yields the equation

$$(\mathbf{D}_{\gamma(t_0)}\,\boldsymbol{\lambda})\cdot\boldsymbol{P}_{t_0} + \boldsymbol{\lambda}_{x_0}\cdot(\mathrm{D}_{t_0}\,\boldsymbol{P}) = \boldsymbol{P}_{t_0}\cdot(\mathrm{D}_{t_0}\,\dot{\boldsymbol{\lambda}}) + (\mathrm{D}_{t_0}\,\boldsymbol{P})\cdot\dot{\boldsymbol{\lambda}}_{t_0}\,. \tag{III.24}$$

Note that the term $\mathrm{D}_{t_0}\,\boldsymbol{P}$ is arbitrary in (III.24) as it is paired with $\boldsymbol{\lambda}_{x_0} = \lambda_{x_0}\boldsymbol{I} = \dot{\boldsymbol{\lambda}}_{t_0}$, analogously to $\boldsymbol{P}_{t_0}$ in (III.22). Thus, we may simplify the equation to

$$(\mathrm{D}_{\gamma(t_0)}\,\boldsymbol{\lambda})\cdot\boldsymbol{P}_{t_0} = \boldsymbol{P}_{t_0}\cdot(\mathrm{D}_{t_0}\,\dot{\boldsymbol{\lambda}}) \tag{III.25a}$$

and solve the matrix EVP to find $(\mathrm{D}_{t_0}(\dot{\lambda})_i, [\boldsymbol{P}_{t_0}]_i) \in \mathbb{R} \times \mathbb{R}^m$

$$(\mathrm{D}_{\gamma(t_0)}\,\boldsymbol{\lambda}) \cdot [\boldsymbol{P}_{t_0}]_i = [\boldsymbol{P}_{t_0}]_i \cdot \mathrm{D}_{t_0}(\dot{\lambda})_i \,.$$

The standard normalization condition

$$\boldsymbol{P}_{t_0}^\top \cdot \boldsymbol{P}_{t_0} = \boldsymbol{I} \tag{III.25b}$$

makes $\boldsymbol{P}_{t_0}$ orthogonal and is consistent with the normalization of the eigenfunctions on the path

$$b\big(\boldsymbol{u}_{\gamma(t)} \cdot \boldsymbol{P}_t, \boldsymbol{u}_{\gamma(t)} \cdot \boldsymbol{P}_t; \gamma(t)\big) = \boldsymbol{I}\,, \qquad \forall\, t \in [t_0, t_1] \tag{III.26}$$

at $t = t_0$. If the eigenvalues of $\mathrm{D}_{\gamma(t_0)}\,\boldsymbol{\lambda}$, i.e., the polarized eigenvalue derivatives $\mathrm{D}_{t_0}(\dot{\lambda})_i$, are non-degenerate, we have determined $\boldsymbol{P}_{t_0}$, up to sign.

REMARK III.11. Dailey's method [**19**] also assumes that the (polarized) derivatives are non-degenerate, as pointed out by [**32, 88**], who in their respective proposals assume that the second-order derivatives $\mathrm{D}^2_{t_0}(\dot{\lambda})_i$ do not repeat.

Although the EVP (III.25) in practice often has m non-degenerate eigenvalues, we want to characterize the more general case that the eigenvalue trajectories are pairwise distinguishable at some order of derivative, i.e., not necessarily the first or second order. We continue to derive (III.23) so that we get the sequence of equations

$$\sum_{k_\lambda + k_P = k} \binom{k}{k_\lambda, k_P} (\mathbf{D}^{k_\lambda}_{\gamma(t_0)}\,\boldsymbol{\lambda}) \cdot (\mathrm{D}^{k_P}_{t_0}\,\boldsymbol{P}) = \sum_{k_\lambda + k_P = k} \binom{k}{k_\lambda, k_P} (\mathrm{D}^{k_P}_{t_0}\,\boldsymbol{P}) \cdot (\mathrm{D}^{k_\lambda}_{t_0}\,\dot{\boldsymbol{\lambda}}) \qquad k \in \mathbb{N}\,. \tag{III.27}$$

In analogy to what we have found for $k = 1$, cf. (III.24), the highest order derivative $(\mathrm{D}^k_{t_0}\,\boldsymbol{P})$ is already arbitrary in the equation of order k since it is paired with $\boldsymbol{\lambda}_{x_0} = \lambda_{x_0}\,\boldsymbol{I}$. We iterate through the equations (III.27) to determine the initial polarization. To this end, we start at $k = 1$ and solve the EVP (III.25). We find an initial polarization $\boldsymbol{P}_{t_0}$ even if it is not unique (yet) and replace the Fréchet derivatives of the eigenvalues with respect to the eigenspace in (III.27) by

$$\mathbf{D}^k_{\gamma(t_0)}\,\tilde{\boldsymbol{\lambda}} = \boldsymbol{P}_{t_0}^\top \cdot (\mathbf{D}^k_{\gamma(t_0)}\,\tilde{\boldsymbol{\lambda}}) \cdot \boldsymbol{P}_{t_0}\,,$$

which are the trajectories corresponding to the basis $\tilde{\boldsymbol{u}}_{\gamma(t_0)} = \boldsymbol{u}_{\gamma(t_0)} \cdot \boldsymbol{P}_{t_0}$. Then we sort the degenerate polarized eigenvalue derivatives into disjoint index sets $(I_i)_{i=1,2,\ldots}$ indicating their degeneracy in the current iteration. To fully determine the initial polarization, we consider the submatrix $[\tilde{\boldsymbol{\lambda}}]_{I_i,I_i} \in \mathbb{R}^{|I_i|\times|I_i|}$ of the trajectory with respect to the eigenspace that relates to one degenerate index set I_i with $|I_i| > 1$ and check the next equation of (III.27). This next equation (of order k) reduces to an EVP

$$(\mathbf{D}^k_{\gamma(t_0)}\,\boldsymbol{\lambda}) \cdot \boldsymbol{P}_{t_0} = \boldsymbol{P}_{t_0} \cdot (\mathrm{D}^k_{t_0}\,\dot{\boldsymbol{\lambda}})$$

since only the terms involving $\boldsymbol{P}_{t_0}$ remain. Checking degeneracy of the eigenvalue of the k-th-order derivatives and adjusting $\boldsymbol{P}_{t_0}$, we iterate through (III.27) until we have found the lowest-order non-degenerate derivative for each eigenvalue.
Throughout the iterations, we thus consider matrix EVPs of declining size, which determine the initial polarization more uniquely. We denote by $[\grave{\boldsymbol{k}}]_{ij} \in \mathbb{N}$ the order of the equation that separates the eigenvalues $(\grave{\lambda})_i, (\grave{\lambda})_j$ by non-degenerate derivatives, as well as $[\grave{\boldsymbol{k}}]_{ii} \in \mathbb{N}$ the lowest-order non-degenerate derivatives for eigenvalue of $(\grave{\lambda})_i$, and collect this information in a **decision matrix** $\grave{\boldsymbol{k}} \in \mathbb{N}^{m\times m}$. This symmetric matrix encodes a tree structure, so other data structures can also be used.
Algorithm III.1 summarizes the procedure to determine the initial polarization matrix explicitly, given a sequence of derivatives of eigenvalues with respect to the eigenspace. In practice, these input derivatives should, of course, only be computed as needed, i.e., up to some order $d \in \mathbb{N}$. For $d = \infty$, the algorithm terminates, if $\boldsymbol{P}_{t_0}$ can be determined uniquely up to sign and permutations. Infinite recursions occur only if there are two or more eigenvalues whose derivatives are all identical.

REMARK III.12. We can determine $\boldsymbol{P}_{t_0}$, up to sign, if on the path γ, which we select, none of the eigenvalues are identical. If two or more eigenvalues stick together on the whole path, their eigenspace degenerate on the whole path. Thus, $\boldsymbol{P}_{t_0}$ remains arbitrary with respect to their eigenspace, cf. (III.22). This edge case cannot be caught by iterating through (III.27). In practice, one might abort the iteration for some large k. If the eigenvalues are actually degenerate on the path, using the initial polarization $\boldsymbol{P}_{t_0}$ that is determined as far as possible is sufficient.

In practice, if we encounter $[\boldsymbol{k}]_{ii} > 1$ in a FE approximation of an EVP, the algorithm is prone to instabilities due to the use of EVP-solvers on the derivatives of the eigenvalues with respect to the eigenspace. When floating-point arithmetic approximates zeros by numbers close to machine precision, the EVP-solver may not recognize that the input matrix is symmetric. We can stabilize the algorithm by first cleaning up these artifacts, up to some tolerance.

REMARK III.13. After determining $\boldsymbol{P}_{t_0}$, it may be convenient to apply the initial polarization matrix $\boldsymbol{P}_{t_0}$ for a change of basis and to use the derivatives

$$\mathbf{D}^k_{\gamma(t_0)} \tilde{\boldsymbol{\lambda}} \coloneqq \boldsymbol{P}_{t_0}^\top \cdot (\mathbf{D}^k_{\gamma(t_0)} \boldsymbol{\lambda}) \cdot \boldsymbol{P}_{t_0} \,, \qquad \forall k \in \mathbb{N} \,,$$
$$\mathbf{D}^k_{\gamma(t_0)} \tilde{\boldsymbol{u}} \coloneqq (\mathbf{D}^k_{\gamma(t_0)} \boldsymbol{u}) \cdot \boldsymbol{P}_{t_0} \,, \qquad \forall k \in \mathbb{N} \,.$$

This means that the new reference basis $\tilde{\boldsymbol{u}}_{x_0} = \boldsymbol{u}_{x_0} \cdot \boldsymbol{P}_{t_0}$ is chosen correctly for a continuous trajectory of the eigenfunctions and that multiplication by $\tilde{\boldsymbol{P}}_{t_0} \coloneqq \boldsymbol{I}$ can be implemented as a selection of vector entries.

III.3.2. Derivatives of the Polarization Matrix. We formulate a characterization of the derivatives of the polarization matrix given the result of algorithm III.1. It is convenient to consider the derivatives columnwise a in a basis representation, similar

Algorithm III.1 Initial Polarization

Input: $(\mathbf{D}^l_{\gamma(t_0)}\boldsymbol{\lambda})_{l=1,\dots,d}$ series of eigenvalue derivatives with respect to the eigenspace of multiplicity m;
k current level of recursion (if not given, set default $k \leftarrow 1$);
`tol` tolerance for identification of degeneracy.
Output: $\boldsymbol{P}_{t_0}$ initial polarization; $\grave{\boldsymbol{k}}$ decision matrix.

function POLARIZE($(\mathbf{D}^l_{\gamma(t_0)}\boldsymbol{\lambda})_{l=1,\dots,d}$, k)
  infer $d \leftarrow$ SERIESLENGTH($(\mathbf{D}^l_{\gamma(t_0)}\boldsymbol{\lambda})_{l=1,\dots,d}$), $m \leftarrow$ LENGTH($\mathbf{D}^l_{\gamma(t_0)}\boldsymbol{\lambda}$)
  initialize $\grave{\boldsymbol{k}} \leftarrow k \cdot \mathbf{1} \in \mathbb{R}^{m\times m}$
  solve EVP $(\mathrm{D}^k_{\gamma(t_0)}\boldsymbol{\lambda}) \cdot \boldsymbol{P}_{t_0} = \boldsymbol{P}_{t_0} \cdot (\mathrm{D}^k_{t_0}\grave{\boldsymbol{\lambda}})$ subject to $\boldsymbol{P}^\top_{t_0} \cdot \boldsymbol{P}_{t_0} = \boldsymbol{I}$ ▷ k-th EVP
  $(\mathbf{D}^l_{\gamma(t_0)}\boldsymbol{\lambda})_{l=1,\dots,d} \leftarrow (\boldsymbol{P}^\top_{t_0} \cdot \mathbf{D}^l_{\gamma(t_0)}\boldsymbol{\lambda} \cdot \boldsymbol{P}_{t_0})_{l=1,\dots,d}$
  sort $\mathrm{D}^k_{t_0}\grave{\boldsymbol{\lambda}}$ into index sets $(I_i)_{i=1,2,\dots}$ according to degeneracy up `tol`
  $\boldsymbol{p} \leftarrow \boldsymbol{I} \in \mathbb{R}^{m\times m}$
  for $\mathrm{D}_{t_0}(\grave{\lambda})_i$ eigenvalue with index set I_i **do**
    $m_i \leftarrow |I_i|$
    if $m_i > 1$ **then**
      if $k < d$ **then**
        $[[\boldsymbol{p}]_{I_i,I_i}, [\grave{\boldsymbol{k}}]_{I_i,I_i}] \leftarrow$ POLARIZE($([\mathrm{D}^l_{\gamma(t_0)}\boldsymbol{\lambda}]_{I_i,I_i})_{l=1,\dots,d}$, $k+1$) ▷ recursion
      else
        $[\grave{\boldsymbol{k}}]_{I_i,I_i} \leftarrow \infty$ ▷ assume $\mathrm{D}^k_{t_0}(\grave{\lambda})_i = \mathrm{D}^k_{t_0}(\grave{\lambda})_j, k \in \mathbb{N}_0$ for $i, j \in I_i$
        throw **warning**
      end if
    end if
  end for
  $\boldsymbol{P}_{t_0} \leftarrow \boldsymbol{P}_{t_0} \cdot \boldsymbol{p}$
  return $[\boldsymbol{P}_{t_0}, \grave{\boldsymbol{k}}]$
end function

to (III.16), i.e.,

$$\mathrm{D}^k_{t_0}[\boldsymbol{P}]_i = \sum_{j=1}^{m} [\boldsymbol{P}_{t_0}]_j \cdot ([\boldsymbol{P}_{t_0}]^\top_j \cdot \mathrm{D}^k_{t_0}[\boldsymbol{P}]_i)\,. \tag{III.28}$$

We now have to determine the basis coefficients $[\boldsymbol{P}_{t_0}]_j^\top \cdot \mathrm{D}_{t_0}^k[\boldsymbol{P}]_i$. The **normalization condition** for the case $i = j$ is deduced from the diagonal argument applied to derivatives of (III.26), i.e.,

(III.29)

$$
\begin{aligned}
&-[\boldsymbol{P}_{t_0}]_i^\top \cdot \mathrm{D}_{t_0}^k[\boldsymbol{P}]_i \\
&= \frac{1}{2} \sum_{\substack{k_{P_1},k_{P_2}=0 \\ \sum_{j=1}^2 k_{P_j}+k_{u_j}+k_x=k}}^{k-1} \binom{k}{k_{P_1},k_{P_2},k_{u_1},k_{u_2},k_x} \mathrm{D}_{t_0}^{k_{P_1}}[\boldsymbol{P}]_i^\top \cdot \mathbf{D}_{t_0}^{k_x} b\big(\mathbf{D}_{\gamma(t_0)}^{k_{u_1}} \boldsymbol{u}, \mathbf{D}_{\gamma(t_0)}^{k_{u_2}} \boldsymbol{u}; \gamma(\cdot)\big) \cdot \mathrm{D}_{t_0}^{k_{P_2}}[\boldsymbol{P}]_i \,.
\end{aligned}
$$

Note that it is independent of the decision matrix $\grave{\boldsymbol{k}}$.
The rest of the coefficients derive from different equations of (III.27). Since the column of the initial polarization $[\boldsymbol{P}_{t_0}]_i$ was derived from the equation (III.27) of order $[\grave{\boldsymbol{k}}]_{ij}$, for $k > 0$ the coefficient $[\boldsymbol{P}_{t_0}]_j^\top \cdot \mathrm{D}_{t_0}^k[\boldsymbol{P}]_i$ is derived from equation $[\grave{\boldsymbol{k}}]_{ij} + k$, i.e.,

$$
\sum_{k_\lambda+k_P=[\grave{\boldsymbol{k}}]_{ij}+k} \binom{[\grave{\boldsymbol{k}}]_{ij}+k}{k_\lambda, k_P} \Big((\mathbf{D}_{\gamma(t_0)}^{k_\lambda} \boldsymbol{\lambda}) \cdot (\mathrm{D}_{t_0}^{k_P} \boldsymbol{P}) - (\mathrm{D}_{t_0}^{k_P} \boldsymbol{P}) \cdot (\mathrm{D}_{t_0}^{k_\lambda} \grave{\boldsymbol{\lambda}}) \Big) = 0 \,.
$$

In order to focus on the i-th column of the polarization matrix, we only consider the i-th column of this equation. Since $\grave{\boldsymbol{\lambda}}$ and its derivatives are diagonal, we get

$$
\sum_{k_\lambda+k_P=[\grave{\boldsymbol{k}}]_{ij}+k} \binom{[\grave{\boldsymbol{k}}]_{ij}+k}{k_\lambda, k_P} \Big((\mathbf{D}_{\gamma(t_0)}^{k_\lambda} \boldsymbol{\lambda}) \cdot \mathrm{D}_{t_0}^{k_P}[\boldsymbol{P}]_i - \mathrm{D}_{t_0}^{k_P}[\boldsymbol{P}]_i \cdot \mathrm{D}_{t_0}^{k_\lambda}(\grave{\lambda})_i \Big) = 0 \,.
$$

Now we multiply $[\boldsymbol{P}_{t_0}]_j^\top$ from the left to get

$$
\sum_{k_\lambda+k_P=[\grave{\boldsymbol{k}}]_{ij}+k} \binom{[\grave{\boldsymbol{k}}]_{ij}+k}{k_\lambda, k_P} [\boldsymbol{P}_{t_0}]_j^\top \cdot \Big(\mathbf{D}_{\gamma(t_0)}^{k_\lambda} \boldsymbol{\lambda} - \mathrm{D}_{t_0}^{k_\lambda}(\grave{\lambda})_i \boldsymbol{I} \Big) \cdot \mathrm{D}_{t_0}^{k_P}[\boldsymbol{P}]_i = 0 \,.
$$

Observe that the terms for $k_\lambda < [\grave{\boldsymbol{k}}]_{ij}$, i.e., $k_P > k$, vanish due to the degeneracy of the eigenvalues and their derivatives . Thus, we arrive at the condition

$$
\sum_{\substack{k_\lambda=[\grave{\boldsymbol{k}}]_{ij} \\ k_\lambda+k_P=[\grave{\boldsymbol{k}}]_{ij}+k}}^{[\grave{\boldsymbol{k}}]_{ij}+k} \binom{[\grave{\boldsymbol{k}}]_{ij}+k}{k_\lambda, k_P} [\boldsymbol{P}_{t_0}]_j^\top \cdot \Big(\mathbf{D}_{\gamma(t_0)}^{k_\lambda} \boldsymbol{\lambda} - \mathrm{D}_{t_0}^{k_\lambda}(\grave{\lambda})_i \boldsymbol{I} \Big) \cdot \mathrm{D}_{t_0}^{k_P}[\boldsymbol{P}]_i = 0 \,. \tag{III.30}
$$

The unknown variables are $\mathrm{D}^k_{t_0}[\boldsymbol{P}]_i$ and possibly $\mathrm{D}^{[\dot{\boldsymbol{k}}]_{ij}+k}_{t_0}(\dot{\lambda})_i$, if $[\dot{\boldsymbol{k}}]_{ij} = [\dot{\boldsymbol{k}}]_{ii}$. The alternative is $[\dot{\boldsymbol{k}}]_{ij} < [\dot{\boldsymbol{k}}]_{ii}$, in which case we can formulate the condition

$$\begin{aligned}(\text{III.31}) \qquad &-\binom{[\dot{\boldsymbol{k}}]_{ij}+k}{[\dot{\boldsymbol{k}}]_{ij},k}\left(\mathrm{D}^{[\dot{\boldsymbol{k}}]_{ij}}_{t_0}(\dot{\lambda})_j - \mathrm{D}^{[\dot{\boldsymbol{k}}]_{ij}}_{t_0}(\dot{\lambda})_i\right)[\boldsymbol{P}_{t_0}]_j^\top \cdot \mathrm{D}^k_{t_0}[\boldsymbol{P}]_i \\ &= \sum_{\substack{k_\lambda=[\dot{\boldsymbol{k}}]_{ij}+1 \\ k_\lambda+k_P=[\dot{\boldsymbol{k}}]_{ij}+k}}^{[\dot{\boldsymbol{k}}]_{ij}+k} \binom{[\dot{\boldsymbol{k}}]_{ij}+k}{k_\lambda,k_P}[\boldsymbol{P}_{t_0}]_j^\top \cdot \left(\mathbf{D}^{k_\lambda}_{\gamma(t_0)}\boldsymbol{\lambda} - \mathrm{D}^{k_\lambda}_{t_0}(\dot{\lambda})_i \boldsymbol{I}\right)\cdot \mathrm{D}^{k_P}_{t_0}[\boldsymbol{P}]_i\,.\end{aligned}$$

Note that we again found a dependency on a gap of eigenvalues, this time those of the $[\dot{\boldsymbol{k}}]_{ij}$-th EVP of algorithm III.1.

For the case $[\dot{\boldsymbol{k}}]_{ij} = [\dot{\boldsymbol{k}}]_{ii}$ we first need to find a formula for $\mathrm{D}^{k_\lambda}_{t_0}(\dot{\lambda})_i$ with $k_\lambda = [\dot{\boldsymbol{k}}]_{ij}+k$. To this end, we multiply $[\boldsymbol{P}_{t_0}]_i^\top$ from the left-hand side to the k_λ-th equation of (III.27) and consider only the i-th column. Analogously to the simplifications previously used, we arrive at

$$\sum_{\substack{l_\lambda=[\dot{\boldsymbol{k}}]_{ii} \\ l_\lambda+l_P=k_\lambda}}^{k_\lambda}\binom{k_\lambda}{l_\lambda,l_P}[\boldsymbol{P}_{t_0}]_i^\top \cdot \left((\mathrm{D}^{l_\lambda}_{\gamma(t_0)}\boldsymbol{\lambda})\cdot(\mathrm{D}^{l_P}_{t_0}[\boldsymbol{P}]_i) - (\mathrm{D}^{l_P}_{t_0}[\boldsymbol{P}]_i)\cdot(\mathrm{D}^{l_\lambda}_{t_0}(\dot{\lambda})_i)\right) = 0\,.$$

As $\mathrm{D}^{k_\lambda}_{t_0}(\dot{\lambda})_i$ is paired with $[\boldsymbol{P}_{t_0}]_i$, we can reformulate this condition to

$$\begin{aligned}(\text{III.32}) \qquad \mathrm{D}^{k_\lambda}_{t_0}(\dot{\lambda})_i &= [\boldsymbol{P}_{t_0}]_i^\top \cdot \mathrm{D}^{k_\lambda}_{\gamma(t_0)}\boldsymbol{\lambda}\cdot[\boldsymbol{P}_{t_0}]_i \\ &\quad + \sum_{\substack{l_\lambda=[\dot{\boldsymbol{k}}]_{ii} \\ l_\lambda+l_P=k_\lambda}}^{k_\lambda-1}\binom{k_\lambda}{l_\lambda,l_P}[\boldsymbol{P}_{t_0}]_i^\top \cdot \left(\mathrm{D}^{l_\lambda}_{\gamma(t_0)}\boldsymbol{\lambda}\cdot\mathrm{D}^{l_P}_{t_0}[\boldsymbol{P}]_i - \mathrm{D}^{l_P}_{t_0}[\boldsymbol{P}]_i\cdot\mathrm{D}^{l_\lambda}_{t_0}(\dot{\lambda})_i\right)\,.\end{aligned}$$

At first glance, it seems that we have just reproduced the previous issue, since this formula includes $\mathrm{D}^{l_P}_{t_0}[\boldsymbol{P}]_i$ with $l_P = k$. However, it turns out that this polarization derivative can be chosen partially arbitrary in (III.32), that is, we only need to obey the previously determined conditions. This version of $\mathrm{D}^{l_P}_{t_0}[\boldsymbol{P}]_i$ is not (yet) correct for (III.21), as the derivatives of $\dot{\boldsymbol{\lambda}}$ will not be diagonal, so it will not lead to the correct eigenfunction derivatives. However, the diagonal entries of the derivatives of $\dot{\boldsymbol{\lambda}}$ will be correct, as well as the formula (III.32).

Given (III.32) and the partially correct derivative of the polarization used to compute it, we can then calculate the missing conditions for the case $[\dot{\boldsymbol{k}}]_{ij} = [\dot{\boldsymbol{k}}]_{ii}$ using the

formula

(III.33)

$$-[\boldsymbol{P}_{t_0}]_j^\top \cdot \mathrm{D}_{t_0}^k[\boldsymbol{P}]_i \leftarrow \frac{\sum\limits_{k_\lambda=[\grave{\boldsymbol{k}}]_{ij}}^{[\grave{\boldsymbol{k}}]_{ij}+k} \binom{[\grave{\boldsymbol{k}}]_{ij}+k}{k_\lambda,k_P}[\boldsymbol{P}_{t_0}]_j^\top \cdot \left(\mathrm{D}_{\gamma(t_0)}^{k_\lambda}\boldsymbol{\lambda} - \mathrm{D}_{t_0}^{k_\lambda}(\grave{\lambda})_i \boldsymbol{I}\right)\cdot \mathrm{D}_{t_0}^{k_P}[\boldsymbol{P}]_i}{\binom{[\grave{\boldsymbol{k}}]_{ij}+k}{[\grave{\boldsymbol{k}}]_{ij},k}\left(\mathrm{D}_{\gamma(t_0)}^{([\grave{\boldsymbol{k}}]_{ij})}(\grave{\lambda})_j - \mathrm{D}_{t_0}^{[\grave{\boldsymbol{k}}]_{ij}}(\grave{\lambda})_i\right)}.$$

To reduce algorithmic complexity, if we initialize the derivative of the polarization matrix using zero matrices, we can replace (III.31) with (III.33).
In summary, the proposed algorithm iterates through $k_\lambda = 1,2,\ldots$ to calculate the derivatives of the polarized eigenvalues $\mathrm{D}_{t_0}^{k_\lambda}(\grave{\lambda})_{i=1,\ldots,m}$. To this end, the derivative of the polarization matrix is first solved only partially correctly, in that some basis coefficients are still missing and set to zero. Then, with the help derivative of the eigenvalue correctly identified, it is solved finally and actually correctly. The complete algorithm for the derivative of the polarization is formalized in algorithm III.2. For the convenience of the reader, we summarize the whole procedure of calculating the derivatives of the eigenpair using the polarization in algorithm III.3. An implementation of the code is provided in the repository [**30**].

REMARK III.14. In the simple case, $\grave{k} := [\grave{\boldsymbol{k}}]_{ij}$ uniformly for $i,j = 1,\ldots,m$, we may also formulate a **main condition** from (III.30) solving both $\mathrm{D}_{t_0}^k[\boldsymbol{P}]_i$ and the $\mathrm{D}_{t_0}^{k_\lambda}(\lambda)_i$ at the same time as

(III.34)

$$\begin{aligned}&\left(\left(\mathbf{D}_{\gamma(t_0)}^{\grave{k}}\boldsymbol{\lambda}\right) - \boldsymbol{I}\cdot \mathrm{D}_{t_0}^{\grave{k}}(\grave{\lambda})_i\right)\cdot\binom{\grave{k}+k}{\grave{k},k}\mathrm{D}_{t_0}^k[\boldsymbol{P}]_i - [\boldsymbol{P}_{t_0}]_i \cdot \mathrm{D}_{t_0}^{\grave{k}+k}(\grave{\lambda})_i\\ &= -\left(\mathbf{D}_{\gamma(t_0)}^{\grave{k}+k}\boldsymbol{\lambda}\right)\cdot[\boldsymbol{P}_{t_0}]_i + \sum_{\substack{k_\lambda=\grave{k}+1\\ k_P+k_\lambda=\grave{k}+k}}^{\grave{k}+k-1}\binom{\grave{k}+k}{k_\lambda,k_P}\left(-\left(\mathbf{D}_{\gamma(t_0)}^{(k_\lambda)}\boldsymbol{\lambda}\right)\cdot \mathrm{D}_{t_0}^{k_P}[\boldsymbol{P}]_i + \mathrm{D}_{t_0}^{k_P}[\boldsymbol{P}]_i \cdot \mathrm{D}_{t_0}^{k_\lambda}(\grave{\lambda})_i\right).\end{aligned}$$

The main condition is again complemented by the normalization condition (III.26). Given an appropriate scaling of (III.26), the problem is then again in symmetric saddle point equation form. This less general characterization is still useful, as in practice we may often encounter $\grave{k} = 1$ for certain kinds of perturbations. However, in the general case, this main condition may become singular.
The multinomial coefficients for (III.34) can be read from Pascal's triangle by ignoring the first $\grave{k}$ columns, which relate to the arbitrary terms, cf. table III.1.

Algorithm III.2 Polarization Derivatives

Input: $(\mathbf{D}^l_{\gamma(t_0)}\boldsymbol{\lambda}, \mathbf{D}^l_{\gamma(t_0)}\boldsymbol{u})_{l=0,\dots,d}$ series of eigenpairs and its derivatives with respect to the eigenspace of multiplicity m up to desired order d of $\mathrm{D}^d_{t_0}\dot{\boldsymbol{\lambda}}$;
$\boldsymbol{P}_{t_0}$ initial polarization (algorithm III.1);
$\dot{\boldsymbol{k}}$ decision matrix (algorithm III.1);
$(\mathbf{D}^l_{x_0} b)_{l=0,\dots,d}$ series of the scalar product and its derivatives.

Output: $(\mathrm{D}^l_{t_0}\dot{\boldsymbol{\lambda}}_{t_0})_{l=1,\dots,d}$ polarized eigenvalue derivatives;
$(\mathrm{D}^l_{t_0}\boldsymbol{P}_{t_0})_{l=1,\dots,d-1}$ partially determined polarization derivatives.

infer $d \leftarrow \textsc{SeriesLength}((\mathbf{D}^l_{\gamma(t_0)}\boldsymbol{\lambda})_{l=1,\dots,d})$, $m \leftarrow \textsc{Length}(\mathbf{D}^l_{\gamma(t_0)}\boldsymbol{\lambda})$
initialize $(\mathrm{D}^k_{t_0}\dot{\lambda})_{k=1,\dots,d} \leftarrow \mathbf{0} \in \mathbb{R}^m$, $(\mathrm{D}^k_{t_0}\boldsymbol{P})_{k=1,\dots,d-1} \leftarrow \mathbf{0} \in \mathbb{R}^{m\times m}$
for $k_\lambda = 1,\dots,d$ **do**
  for $i = 1,\dots,m$ **do**
    if $k_\lambda \le [\dot{\boldsymbol{k}}]_{ii}$ **then**
      $\mathrm{D}^{k_\lambda}_{t_0}(\dot{\lambda})_i \leftarrow [\boldsymbol{P}_{t_0}]_i^\top \cdot (\mathrm{D}^{k_\lambda}_{\gamma(t_0)}\boldsymbol{\lambda}) \cdot [\boldsymbol{P}_{t_0}]_i$ ▷ determined by $\boldsymbol{P}_{t_0}$
    else
      $k \leftarrow k_\lambda - [\dot{\boldsymbol{k}}]_{ii}$ ▷ $\mathrm{D}^k_{t_0}[\boldsymbol{P}]_i$ associated to $\mathrm{D}^{k_\lambda}_{t_0}(\dot{\lambda})_i$
      $\boldsymbol{r} \leftarrow \mathbf{0} \in \mathbb{R}^m$ ▷ **Part I**: determine $\mathrm{D}^{k_\lambda}_{t_0}(\dot{\lambda})_i$
      for $j = 1,\dots,m$ **do**
        if $[\dot{\boldsymbol{k}}]_{ij} < [\dot{\boldsymbol{k}}]_{ii}$ **then** ▷ if $\mathrm{D}^k_{t_0}[\boldsymbol{P}]_j$ previously determined
          compute $[\boldsymbol{r}]_j \leftarrow [\boldsymbol{P}_{t_0}]_j^\top \cdot (\mathrm{D}^k_{t_0}[\boldsymbol{P}]_i)$ using (III.33)
        end if
      end for
      $\mathrm{D}^k_{t_0}[\boldsymbol{P}]_i \leftarrow \boldsymbol{P}_{t_0} \cdot \boldsymbol{r}$ ▷ $\mathrm{D}^k_{t_0}[\boldsymbol{P}]_i$ determined sufficiently for $\mathrm{D}^{k_\lambda}_{t_0}(\dot{\lambda})_i$
      determine $\mathrm{D}^{k_\lambda}_{t_0}(\dot{\lambda})_i$ using (III.32)
      for $j = 1,\dots,m$ **do** ▷ **PART II**: determine $\mathrm{D}^k_{t_0}[\boldsymbol{P}]_i$
        if $[\dot{\boldsymbol{k}}]_{ij} = [\dot{\boldsymbol{k}}]_{ii}$ **then**
          if $i \ne j$ **then**
            compute $[\boldsymbol{r}]_j \leftarrow [\boldsymbol{P}_{t_0}]_j^\top \cdot (\mathrm{D}^k_{t_0}[\boldsymbol{P}]_i)$ using (III.33)
          else
            compute $[\boldsymbol{r}]_i \leftarrow [\boldsymbol{P}_{t_0}]_i^\top \cdot (\mathrm{D}^k_{t_0}[\boldsymbol{P}]_i)$ using (III.29)
          end if
        end if
      end for
      $\mathrm{D}^k_{t_0}[\boldsymbol{P}]_i \leftarrow \boldsymbol{P}_{t_0} \cdot \boldsymbol{r}$ ▷ $\mathrm{D}^k_{t_0}[\boldsymbol{P}]_i$ determined finally
    end if
  end for
end for

Algorithm III.3 Eigenpair Derivatives

Input: $((\mathbf{D}^k_{\gamma(t_0)}\, a)_{k=1,2,\dots}, (\mathbf{D}^k_{\gamma(t_0)}\, b)_{k=1,2,\dots}$ series of derivatives of the bilinear forms (to sufficient order so that the polarization can be determined, and then additional orders for each derivative of the eigenfunction required);
$\gamma : [t_0, t_1] \to X$ the analytic path;
$(\boldsymbol{\lambda}_{x_0}, \boldsymbol{u}_{x_0})$ the unperturbed eigenpair.

Output: $(\mathrm{D}^k_{t_0}\, \grave{\boldsymbol{\lambda}}, \mathrm{D}^k_{t_0}\, \grave{\boldsymbol{u}})_{k=1,2,\dots}$ polarized eigenpair derivatives.

1: calculate $(\mathbf{D}^k_{x_0}\, \boldsymbol{\lambda}, \mathbf{D}^k_{x_0}\, \boldsymbol{u})_{k=1,2,\dots}$ using (III.15) iteratively
2: determine $(\mathbf{D}^k_{\gamma(t_0)}\, \boldsymbol{\lambda}, \mathbf{D}^k_{\gamma(t_0)}\, \boldsymbol{u})_{k=1,2,\dots}$ on the path using (III.20)
3: (set entries of $(\mathbf{D}^k_{\gamma(t_0)}\, \boldsymbol{\lambda})_{k=1,2,\dots}$ close to machine precision to zero for stability)
4: calculate $\boldsymbol{P}_{t_0}$ and $\grave{\boldsymbol{k}}$ using algorithm III.1
5: calculate $(\mathrm{D}^k_{t_0}\, \boldsymbol{P})_{k=1,2,\dots}$ and $(\mathrm{D}^k_{t_0}\, \grave{\boldsymbol{\lambda}})_{k=1,2,\dots}$ using algorithm III.2
6: calculate $(\mathrm{D}_{t_0}\, \grave{\boldsymbol{u}})_{k=1,2,\dots}$ using (III.21c)

$$\binom{k_\lambda + k}{k_\lambda, k} =$$

$k_\lambda + k$ \ k_λ	0	1	2	3	4	5	6
0	1						
1	1	1					
2	1	2	1				
3	1	3	3	1			
4	1	4	6	4	1		
5	1	5	10	10	5	1	
6	1	6	15	20	15	6	1

TABLE III.1. Pascal's triangle with entries for (III.34). The irrelevant entries $k_\lambda < \grave{k}$ are grayed out for $\grave{k} \geq 1$. The first $\grave{k}$ columns are irrelevant for $\grave{k} \in \mathbb{N}$.

REMARK III.15. Given the characterization summarized in algorithms III.1 and III.2, we can check that in the special case in which all derivatives with respect to the eigenspace $(\mathbf{D}^k_{\gamma(t_0)}\, \boldsymbol{\lambda})_{k\in\mathbb{N}}$ are diagonal, $\boldsymbol{P}_{t_0} = \boldsymbol{I}$ is a valid choice and (III.33) yields

$$[\boldsymbol{P}_{t_0}]_j^\top \cdot \mathrm{D}^k_{t_0} [\boldsymbol{P}]_i = 0 \qquad\qquad i \neq j, \quad k \in \mathbb{N}\,.$$

Thus, since the eigenfunctions with respect to the eigenspace are already orthonormal on $B(x_0)$, (III.26) also yields

$$[\boldsymbol{P}_{t_0}]_i^\top \cdot \mathrm{D}^k_{t_0} [\boldsymbol{P}]_i = 0 \qquad\qquad k \in \mathbb{N}\,.$$

So in this case, we get $\mathrm{D}^k_{t_0}\, \boldsymbol{P} = \boldsymbol{0}$ for $k \in \mathbb{N}$.

III.4. Mapping Eigenpairs to the Reference Eigenspace

So far, we have characterized the derivatives of eigenpairs with respect to the eigenspace and characterized the polarization matrix, which relates the derivatives with respect to the eigenspace and the actual derivatives of the eigenpair by (III.21). To calculate Taylor approximations of eigenfunctions of order n, we need to determine the derivatives of the polarization matrix up to order $n-1$. That is, if we formulate Taylor approximations

$$\boldsymbol{\lambda}_{t,\mathrm{appr},n} \coloneqq \sum_{k=0}^{n} \frac{t^k}{k!} \mathrm{D}^k_{\gamma(t_0)} \boldsymbol{\lambda} \,, \quad \boldsymbol{u}_{t,\mathrm{appr},n} \coloneqq \sum_{k=0}^{n} \frac{t^k}{k!} \mathrm{D}^k_{\gamma(t_0)} \boldsymbol{u} \,, \quad \boldsymbol{P}_{t,\mathrm{appr},n} \coloneqq \sum_{k=0}^{n} \frac{t^k}{k!} \mathrm{D}^k_{t_0} \boldsymbol{P} \,,$$

then we can formulate approximations

$$\begin{aligned} \grave{\boldsymbol{\lambda}}_t &= \boldsymbol{P}^{-1}_{t,\mathrm{appr},n-1} \cdot \boldsymbol{\lambda}_{t,\mathrm{appr},n} \cdot \boldsymbol{P}_{t,\mathrm{appr},n-1} + \mathscr{O}(t^{n+1}) \,, \\ \grave{\boldsymbol{u}}_t &= \boldsymbol{u}_{t,\mathrm{appr},n} \cdot \boldsymbol{P}_{t,\mathrm{appr},n} + \mathscr{O}(t^{n+1}) \,. \end{aligned}$$

Note that these approximations include some additional terms compared to the direct Taylor approximations of the polarized eigenpairs, i.e., truncations of (III.21). We use the explicit inversion instead of the transposed matrix as the approximation of the polarization matrix is only approximately orthogonal.
Assume that we have some solution of the (polarized) eigenpair $(\grave{\boldsymbol{\lambda}}_t, \grave{\boldsymbol{u}}_t)$ by directly solving the perturbed EVP, which in terms of sign and basis matches the Taylor approximation of the eigenpair suggested by the Taylor approximation. Then, we can calculate approximations of samples with respect to the eigenspace by inversion, i.e.,

$$\boldsymbol{\lambda}_{\gamma(t)} = \boldsymbol{P}_{t,\mathrm{appr},n-1} \cdot \grave{\boldsymbol{\lambda}}_t \cdot \boldsymbol{P}^{-1}_{t,\mathrm{appr},n-1} + \mathscr{O}(t^{n+1}) \,, \tag{III.35a}$$

$$\boldsymbol{u}_{\gamma(t)} = \grave{\boldsymbol{u}}_t \cdot \boldsymbol{P}^{-1}_{t,\mathrm{appr},n} + \mathscr{O}(t^{n+1}) \,. \tag{III.35b}$$

In summary, the polarization matrix $\boldsymbol{P}$ describes an (invertible) orthogonal projection between the trajectories with respect to the eigenspace and otherwise, which can be approximated by its Taylor approximation.

III.4.1. Approximate Projection by Singular Value Decomposition. Projecting samples to the eigenspace via an approximation of the polarization matrix is a process that, at least implicitly, requires the identification of a path γ to which the polarization pertains, as well as the implementation of algorithms III.1 and III.2 . Since the trajectory of the eigenpair with respect to the eigenspace does not require the selection of a path, we provide a projection in a more direct manner via a **singular value decomposition (SVD)**.
This ansatz was presented in [**25**, Chapter 3.3] and uses the SVD

$$b\big(\boldsymbol{u}_x, \boldsymbol{u}_{x_0}; x_0\big) =: \boldsymbol{U} \cdot \boldsymbol{\Lambda} \cdot \boldsymbol{V} \tag{III.36}$$

where $\mathbf{\Lambda}$ is a diagonal matrix of **singular values** and $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{m\times m}$ are orthogonal matrices. We omit the theoretical background and implementation of the SVD and refer the interested reader to [**40**, Chapter 2.4]. Given the SVD (III.36), we may equivalently write

$$b\big(\boldsymbol{u}_x \cdot \boldsymbol{V}^\top, \boldsymbol{u}_{x_0} \cdot \boldsymbol{U}^\top; x_0\big) = \boldsymbol{U}^\top \cdot b\big(\boldsymbol{u}_x, \boldsymbol{u}_{x_0}; x_0\big) \cdot \boldsymbol{V}^\top = \mathbf{\Lambda}\,.$$

For $x \approx x_0$ it holds $\mathbf{\Lambda} \approx \boldsymbol{I}$, as the norm induced by the scalar product has changed little. Therefore, if we combine both orthogonal matrices in the first argument, we get

$$b\big(\boldsymbol{u}_x \cdot \boldsymbol{V}^\top \cdot \boldsymbol{U}, \boldsymbol{u}_{x_0}; x_0\big) = b\big(\boldsymbol{u}_x, \boldsymbol{u}_{x_0}; x_0\big) \cdot \boldsymbol{V}^\top \cdot \boldsymbol{U} \approx \boldsymbol{I}\,.$$

This suggests that the orthogonal matrix

$$\check{\boldsymbol{P}} := \boldsymbol{V}^\top \cdot \boldsymbol{U} \tag{III.37}$$

relates $(\check{\boldsymbol{P}}^\top \cdot \boldsymbol{\lambda}_x \cdot \check{\boldsymbol{P}}, \boldsymbol{u}_x \cdot \check{\boldsymbol{P}})$ to $(\boldsymbol{\lambda}_{x_0}, \boldsymbol{u}_{x_0})$ for x in a neighborhood $B(x_0)$ of x_0. Near the reference point, (III.37) approximates the polarization matrix.

REMARK III.16. As an alternative projection method, [**43, 44**] proposed a spectral projection.

The numerical calculation of the SVD introduces some numerical error to the projection matrix (III.37). The stopping criterion for SVD-solvers is usually formulated by convergence of the singular values $\mathbf{\Lambda}_x$. The calculated orthogonal matrices may thus be accurate[3], however, still not accurate enough to calculate (III.37) sufficiently accurately to observe certain error rates predicted by Taylor's theorem. Since Taylor approximations of the polarization matrix (III.35) yield better results in this aspect, we rely on this method in the following experiments.

III.5. Local Identification of Analytically Perturbed Eigenpairs

In this section, we discuss the local bifurcation behavior of the polarized eigenvalues. The behavior of perturbed eigenpairs has been investigated, for example, in [**20, 59, 74, 83**]. The objective in this section is to point out what the polarization matrix $\boldsymbol{P}$ and the decision matrix $\grave{\boldsymbol{k}}$, newly introduced in algorithm III.1, tell us about the local bifurcation behavior of the degenerate eigenvalue. We then investigate when it is possible to describe individual eigenvalues and eigenfunctions in a neighborhood $B(x_0) \subset X$, which motivates the use of trajectories with respect to the eigenspace.

III.5.1. Identification of the Eigenvalue on a Path. First, consider a one-dimensional parameter domain $t \in B(t_0)$, i.e., an interval where t_0 is an inner point. This can again be interpreted as a composition with an analytic path $\gamma : B(t_0) \to X$. Due to (III.21), we know that the eigenpair can be thought of as a locally analytic function on $B(t_0)$.

[3] That is in terms of singular values, and so that (III.36) holds up to machine precision.

If the eigenpair is degenerate at t_0, this involves an analytic polarization matrix. An edge case is again where a pair of eigenvalues remains identically degenerate on the path. In this case, the formulation of degenerate eigenvalues is trivial, since the eigenvalues have the same function values and derivatives. Their eigenfunctions remain arbitrary with respect to their polarization, but a choice of eigenfunction trajectories can easily be found. If they are not identical, the following lemma holds.

LEMMA III.17. *Consider the setting of the parameterized EVP in assumption III.1 with* $\dim(X) = 1$, $t_0 \in X$. *Let* $\grave{\lambda}_i, \grave{\lambda}_j : B(t_0) \to \mathbb{R}$ *be eigenfunctions that are not identical. If* $[\grave{\boldsymbol{k}}]_{ij}$ *is even, the eigenvalue trajectories are such that the eigenvalues touch and then diverge, (locally) retaining their order. Otherwise, they switch their order.*

PROOF. The difference of two eigenvalue trajectories is described by a Taylor series of the form

$$\grave{\lambda}_i - \grave{\lambda}_j = \sum_{n=0}^{\infty} \frac{t^n}{n!} \underbrace{(\mathrm{D}_{t_0}^n \grave{\lambda}_i - \mathrm{D}_{t_0}^n \grave{\lambda}_j)}_{=:c_n},$$

such that $c_{[\grave{\boldsymbol{k}}]_{ij}}$ the first non-zero coefficient. □

DEFINITION III.18. In lemma III.17, if the eigenvalues switch their order, we call this behavior a **crossing**. If they retain their order, we call it a **deflection**.

III.5.2. Multidimensional Parameter Spaces. For the case of multidimensional parameter spaces X, we have seen in example I.2 that, in general, no function can be formulated to identify eigenvalues or eigenfunctions in a neighborhood $B(x_0)$ when they are part of a degenerate subspace in $x_0 \in X$. Extending the one-dimensional perspective and still assuming pathwise analyticity, we formalize what we mean by identifying a function in $B(x_0)$.

DEFINITION III.19. Consider the setting of the parameterized EVP in assumption III.1 and an open neighborhood $B(x_0)$ around a reference point x_0. We call an eigenvalue $\grave{\lambda}_{x_0}$ **identifiable as a function** $\grave{\lambda} : B(x_0) \to \mathbb{R}$, if we can select an arbitrary analytic path

$$\gamma : [t_0, t_1] \to B(x_0)$$

with $\gamma(t_0) = x_0$ and follow its one-dimensional locally analytic polarized eigenvalue trajectory

$$\grave{\lambda} : [t_0, t_1] \to \mathbb{R}$$

with initial value $\grave{\lambda}(t_0) = \grave{\lambda}_{x_0}$ such that the final eigenvalue $\grave{\lambda}(t_1)$ is unique. We also call eigenfunction $\grave{u}_{x_0}$ **identifiable as a function** $\grave{u} : B(x_0) \to V$ if the same holds for the eigenfunction trajectory such that

$$\grave{u} : [t_0, t_1] \to V$$

with a path-independent initial value $\grave{u}(t_0) = \grave{u}_{x_0}$ and such that the final eigenfunction $\grave{u}(t_1)$ is unique.

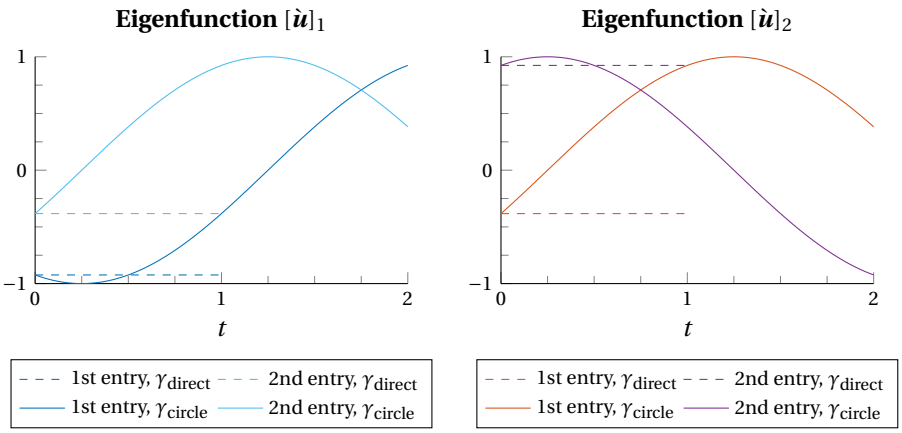

FIGURE III.2. Eigenvectors of example I.2 on paths (I.6) with the circular path prolonged on $t \in [1,2]$ for a full circle. For the eigenvalues compare fig. I.2.

EXAMPLE III.20 (Continuation of example I.2). As an example of eigenvalues that are not identifiable as a function $B(x_0) \to \mathbb{R}$ with $x_0 = \mathbf{0} \in \mathbb{R}^2$, we have already pointed out the eigenvalues of example I.2. Similarly to the highlighted paths, every path that passes through x_0 once will lead to a crossing, and any path around x_0 will retain the order of the eigenvalues, as they only become degenerate in x_0.
Now that we know that the eigenvalues cannot be identified as a function on $B(x_0)$, we also consider their eigenfunctions. In fig. III.2 the values of the eigenfunctions $[\grave{\boldsymbol{u}}]_{i=1,2} \in \mathbb{R}^2$ are presented on the respective paths of example I.2 for $t \in [0,1]$. Comparing the eigenvalues following a straight path and a path on a half-circle, the order of the eigenvalues is switched in the end. For this reason, on the half-circular path, each eigenfunction u_i must evolve into what was initially the other eigenfunction, although one has a switched sign compared to the initial eigenfunctions. We continue the trajectory of the eigenfunction to a complete circle on $t \in (1,2]$. If we observe the evolution of the eigenfunctions while following the circular path in two encirclements of the degenerate point, and denoting the state of the eigenfunctions with respect to the initial state at $t_0 = 0$ after each half-circle, the eigenfunctions pass through the following states

$$\begin{bmatrix} [\boldsymbol{u}_{t_0}]_1 \\ [\boldsymbol{u}_{t_0}]_2 \end{bmatrix} \rightsquigarrow \begin{bmatrix} [\boldsymbol{u}_{t_0}]_2 \\ -[\boldsymbol{u}_{t_0}]_1 \end{bmatrix} \rightsquigarrow \begin{bmatrix} -[\boldsymbol{u}_{t_0}]_1 \\ -[\boldsymbol{u}_{t_0}]_2 \end{bmatrix} \rightsquigarrow \begin{bmatrix} -[\boldsymbol{u}_{t_0}]_2 \\ [\boldsymbol{u}_{t_0}]_1 \end{bmatrix} \rightsquigarrow \begin{bmatrix} [\boldsymbol{u}_{t_0}]_1 \\ [\boldsymbol{u}_{t_0}]_2 \end{bmatrix}.$$

So, after a full circle the sign is flipped and, after another encirclement, we arrive at the original eigenfunctions. Of course, following a circular path in the opposite direction reverses the order of these evolutions. In this example, the eigenfunctions are

constant with respect to the radius of the circle, so we can conclude that the eigenfunctions are clearly not identifiable as functions on $B(x_0)$ either, since we cannot choose the sign of the eigenfunction consistently.

The effect that eigenfunctions rotate in the presence of degenerate parameter points is well known and is referred to by various names, e.g. *coupling* [**17**] or *veering* [**74**].

DEFINITION III.21. We will call an eigenpair (λ_i, u_i) **uncoupled** if it can be identified as a *locally analytic function*. Otherwise, we will call it **coupled**.

This condition is stronger than just identifiable as a function, definition III.19, in that the pathwise analytic functions must now be analytic in a multidimensional sense, cf. Fréchet differentiability. Once we have found one subspace in which the eigenpairs are coupled, they are coupled in the context of the larger parameter space, i.e., in example I.2 the parameter space $X = \mathbb{R}^2$.

REMARK III.22. In quantum mechanics, similar effects are often described using **spinors**, which also capture the effect that after one rotation of the reference space their sign is flipped, cf. [**15**, Definition 52]. A formal discussion of spinors exceeds the scope of this thesis. The interested reader is referred to [**15, 17**].

Uncoupled Eigenpairs. Uncoupled eigenpairs constitute a benign setting, which we discuss in more detail. From the viewpoint of individual eigenpairs, we can formulate the following conclusion.

LEMMA III.23. *Consider an eigenpair $(\grave{\lambda}, \grave{u})$ in the setting of the parameterized EVP of assumption III.1. The eigenvalue can be identified as an analytic function if the eigenfunctions can be identified as an analytic function.*

PROOF. This follows from the Rayleigh quotient, cf. theorem II.76. Due to normalization $b\big(u_x, u_x; x\big) = 1$ we get

$$\grave{\lambda} = a\big(\grave{u}, \grave{u}; \cdot\big) : B(x_0) \to \mathbb{R}$$

with the first two arguments linear and a analytic in x. Since the composition of analytic functions is again analytic, we have found the analytic trajectory of the eigenvalue. □

In accordance with theorem III.2, the same can be stated for trajectories with respect to the eigenspace, i.e.,

$$\boldsymbol{\lambda} = a\big(\boldsymbol{u}, \boldsymbol{u}; \cdot\big) : B(x_0) \to \mathbb{R}^{m\times m} .$$

REMARK III.24. In some cases, it is possible to generalize the (pathwise) analytic polarization matrix $\boldsymbol{P} : [t_0, t_1] \to \mathbb{R}^{m\times m}$ of section III.3 to the notion of an analytic polarization matrix

$$\boldsymbol{P} : B(x_0) \to \mathbb{R}^{m\times m} , \qquad x = x_0 + h \mapsto \boldsymbol{P}_x = \boldsymbol{P}_{x_0} + \sum_{k=1}^{\infty} \frac{1}{k!} \mathbf{D}^k_{x_0} \boldsymbol{P}[h] , \tag{III.38a}$$

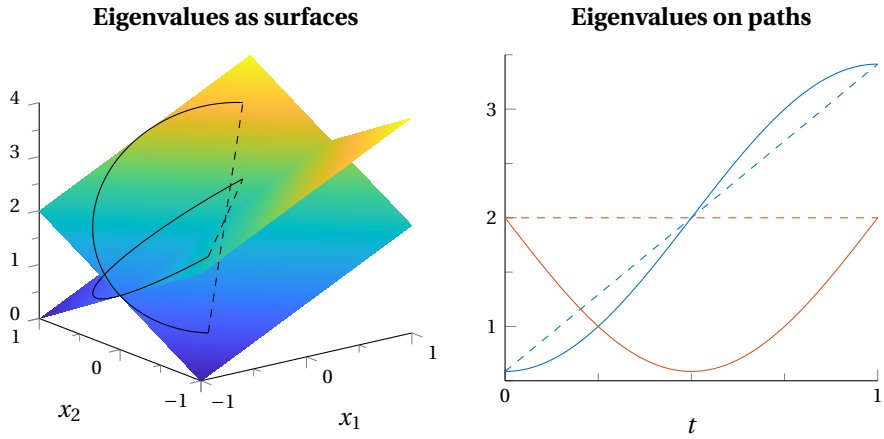

FIGURE III.3. Eigenvalues of example III.25 (and III.27) with evolution on paths of example I.2 highlighted.

such that

$$\grave{\boldsymbol{\lambda}} \coloneqq \boldsymbol{P}^\top \cdot \boldsymbol{\lambda} \cdot \boldsymbol{P}\,, \qquad \grave{\boldsymbol{u}} \coloneqq \boldsymbol{u} \cdot \boldsymbol{P}\,, \tag{III.38b}$$

defines an analytic polarized trajectory $(\grave{\boldsymbol{\lambda}}, \grave{\boldsymbol{u}}) : B(x_0) \to \mathbb{R}^{m\times m} \times V^m$ with $\grave{\boldsymbol{\lambda}}$ diagonal and $\grave{\boldsymbol{u}}$ subject to a normalization

$$b\big(\grave{\boldsymbol{u}}, \grave{\boldsymbol{u}}; x\big) = b\big(\boldsymbol{u}\cdot\boldsymbol{P}, \boldsymbol{u}\cdot\boldsymbol{P}; x\big) = \boldsymbol{I}\,. \tag{III.38c}$$

The following example shows that this is indeed possible with non-trivial Fréchet derivatives for the analytic polarization matrix.

EXAMPLE III.25. The parameterized matrix

$$K : \mathbb{R}^2 \to \mathbb{R}^{2\times 2}\,, \qquad x \mapsto K_x = \begin{bmatrix} 2 + x_1 + x_2\cos(2x_2) & -x_2\sin(2x_2) \\ -x_2\sin(2x_2) & 2 + x_1 - x_2\cos(2x_2) \end{bmatrix} \tag{III.39}$$

has eigenpairs given by

$$\grave{\boldsymbol{\lambda}}_x = \begin{bmatrix} 2 + x_1 - x_2 & 0 \\ 0 & 2 + x_1 + x_2 \end{bmatrix}\,, \qquad \grave{\boldsymbol{u}}_x = \pm \begin{bmatrix} \cos(x_2) & -\sin(x_2) \\ \sin(x_2) & \cos(x_2) \end{bmatrix}\,.$$

The eigenvalue surfaces and trajectories on the paths of example I.2 are illustrated in fig. III.3. The eigenfunctions follow a rotation matrix parameterized by x_2. In one point on each path the eigenvalues are degenerate, so the eigenfunctions are arbitrary there. However, continuity suggests the given trajectory at these points.

The following lemma is a special case of remark III.24.

LEMMA III.26. *Let the setting of remark III.24 hold such that the path-independent initial polarization $\boldsymbol{P}_{x_0} \in \mathbb{R}^{m\times m}$ exists. Then the following equivalence holds:*

(1) *The polarization $\boldsymbol{P} : B(x_0) \to \mathbb{R}^{m\times m}$ is constant, i.e., $x \mapsto \boldsymbol{P}_x = \boldsymbol{P}_{x_0}$.*
(2) *The orthogonality condition $b\big(\mathbf{D}_{x_0}\grave{\boldsymbol{u}}, \grave{\boldsymbol{u}}_{x_0}; x_0\big) = 0$ holds.*

PROOF. The orthogonality condition holds for trajectories with respect to the eigenspace given the choice of initial basis $\grave{\boldsymbol{u}}_{x_0} = \boldsymbol{u}_{x_0} \cdot \boldsymbol{P}_{x_0}$, cf. remark III.13. For the equivalence to hold, this trajectory with respect to the eigenspace must already be the polarized trajectory. Thus, the polarization derivatives vanish, cf. also remark III.15. □

According to lemma III.26, the eigenpair trajectories (in the polarized sense)

$$\grave{\boldsymbol{\lambda}} : B(x_0) \to \mathbb{R}^{m\times m} \text{ (diagonal)}\,, \qquad\qquad \grave{\boldsymbol{u}} : B(x_0) \to V^m\,,$$

coincide with the trajectories with respect to the eigenspace, when the orthogonality condition (III.11c) is a natural choice and not an evasive choice to construct Fréchet differentiable surrogate trajectories.

EXAMPLE III.27. By switching only one sign in the parameterized matrix of example I.2, we consider the parameterized matrix

$$K : \mathbb{R}^2 \to \mathbb{R}^{2\times 2}\,, \qquad\qquad x \mapsto K_x = \begin{bmatrix} 2+x_1 & -x_2 \\ -x_2 & 2+x_1 \end{bmatrix}. \tag{III.40}$$

Its eigenpairs are given by

$$\grave{\boldsymbol{\lambda}}_x = \begin{bmatrix} 2+x_1-x_2 & 0 \\ 0 & 2+x_1+x_2 \end{bmatrix}, \qquad\qquad \grave{\boldsymbol{u}}_x = \pm\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

The eigenvalues are the same as for example III.25 and are illustrated in fig. III.3, but the polarization is constant. As there are only two eigenfunctions, they are also constant.

The prerequisites of remark III.24 and lemma III.26 are unfortunately infeasible to check numerically. For a more involved setting than that of examples III.25 and III.27, an analytical investigation of the perturbation model is required. A more elaborate example of such a model can be found in [**44**].

Coupled, yet Identifiable Eigenvalues. Non-analytic eigenvalues trajectories can sometimes also be found, without analyticity of eigenfunction trajectories. The following corollary extends lemma III.17.

COROLLARY III.28. *Consider the setting of the parameterized EVP in assumption III.1. If all eigenvalues have even order $[\grave{\boldsymbol{k}}]_{ij}$, $i, j = 1, \ldots, m$ for all paths $\gamma : [t_0, t_1] \to B(x_0)$, the eigenfunctions are identifiable as continuous functions*

$$\grave{\lambda}_i : B(x_0) \to \mathbb{R}\,.$$

PROOF. Under the assumptions of the lemma, the result of lemma III.17 can be applied pathwise and between each pair of eigenvalues. This yields a consistent identification of the eigenvalues on $B(x_0)$ according to deflection. □

EXAMPLE III.29. Altering example I.2 such that the dependence on the parameters is quadratic, we consider the parameterized matrix

$$K : \mathbb{R}^2 \to \mathbb{R}^{2\times 2}, \qquad x \mapsto K_x = \begin{bmatrix} 2+x_1^2 & -x_2^2 \\ -x_2^2 & 2-x_1^2 \end{bmatrix}. \tag{III.41}$$

Its eigenvalues and (unnormalized) eigenfunctions are given by

$$\grave{\boldsymbol{\lambda}}_x = \begin{bmatrix} 2-\sqrt{x_1^4+x_2^4} & 0 \\ 0 & 2+\sqrt{x_1^4+x_2^4} \end{bmatrix}, \quad \grave{\boldsymbol{u}}_x = \pm \begin{bmatrix} x_1^2-\sqrt{x_1^4+x_2^4} & x_1^2+\sqrt{x_1^4+x_2^4} \\ -x_2^2 & -x_2^2 \end{bmatrix}.$$

The surfaces and trajectories of the eigenvalues on the paths of example I.2 are illustrated in fig. III.4. According to lemma III.17, the correct identification of the eigenvalues in a neighborhood $B(\mathbf{0})$ is such that the eigenvalue functions are deflected. We can therefore find a vanishing first-order Fréchet derivative for each eigenvalue at the origin and second-order Gâteaux derivatives according to the upper and lower surface of the graph.
The evolution of the eigenfunctions on the paths is also illustrated in fig. III.4. Again, the eigenfunction trajectories on the circle are independent of the radius. Although the function can be identified as a function on $B(\mathbf{0}) \setminus \{\mathbf{0}\}$, since the sign is now consistent, it is not possible to find a unique continuation of the eigenfunctions for $x = \mathbf{0}$. Thus, the eigenfunctions are not identifiable as a function on $B(\mathbf{0})$.

III.5.3. Implications and Alternative Approaches. We have seen that eigenpairs often cannot be identified as functions in the neighborhood surrounding a point where the eigenvalues are degenerate. One might argue that the identification by locally analyticity is but one possibility to construct functions.
An alternative ansatz is to order the eigenvalues $\lambda \in \mathbb{R}$ by their real value. This may be justified since the smallest eigenvalue sometimes has a certain physical interpretation, e.g., a ground state of lowest energy. Obviously, this in general yields a non-differentiable eigenvalue function.
In general, this ansatz does not work for complex eigenvalues, as their order is not uniquely defined. As we have seen in the previous examples, this ansatz neither works for coupled eigenfunctions, as the basis may be arbitrary at the degenerate point. If we accept that the eigenfunctions are ill-defined in points where the eigenvalues become degenerate, we still have to check if the sign can be chosen consistently. Lastly, in some applications, it makes sense to consider a variable of interest which depends on the eigenfunction but is oblivious to its sign. For example, in quantum physics the absolute value of the eigenfunction is interpreted as a probability of location, cf. [**17**].

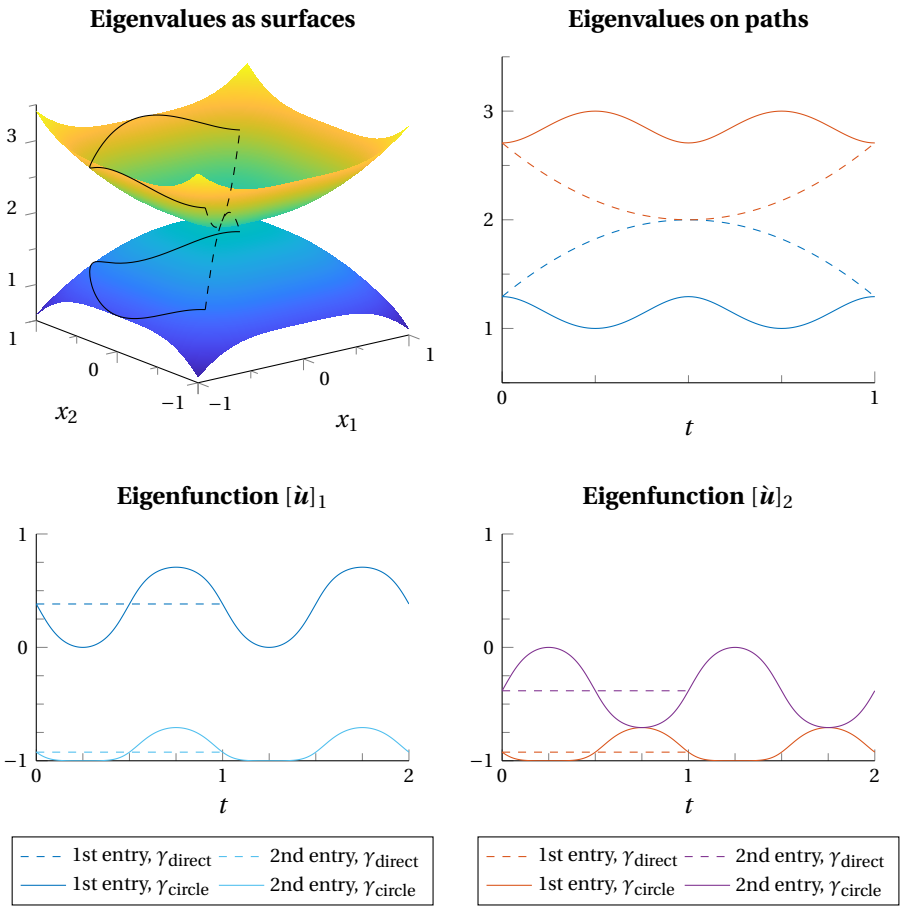

FIGURE III.4. Eigenpairs of example III.29 with evolution on paths of example I.2 highlighted. The eigenvector trajectories are illustrated on the paths (I.6) with the circular path prolonged on $t \in [1,2]$ for a full circle.

In summary, only the ansatz of trajectories with respect to the eigenspace is, at least locally, always well defined, which is why we will use it for uncertainty quantification. If the equivalence of lemma III.26 holds, the eigenspace ansatz is equivalent to the actual eigenvalues up to a change of basis.

III.6. Implementation

In this section, we discuss the implementation of the derivatives with respect to the eigenspace. For the numerical approximation of variational EVPs, we must translate the variational problem into a matrix EVP. To this end, we consider a Galerkin discretization of the Hilbert spaces.
We also discuss how saddle point equations of different, i.e., otherwise non-degenerate eigenvalues, can be joined in larger saddle point equations. Although multiple smaller saddle point equations are more efficient to solve, this representation is convenient when calculating the derivatives with respect to the eigenspace in an environment $B(x_0)$ where at points $x \in B(x_0) \setminus \{x_0\}$ the eigenvalues are non-degenerate, while at x_0 they are degenerate. We then discuss solution strategies for saddle point equations following [**8**] and point out which are particularly suitable to us.

III.6.1. Discretization. Some aspects of the FE discretization of the Laplace and Maxwell EVP were discussed in section II.6. Here, we focus on the discretization of the derivatives and the formulation of discrete saddle point equations. The solvability of the discrete saddle point equations can be shown analogously to the variational case. As noted in [**25**], it is equivalent to discretize the variational EVP and then calculate derivatives thereof or to discretize the variational characterization of the derivatives, up to a consistency error. This consistency error is introduced, since we do not have the exact eigenvalue but its discrete approximation.
We assume that a finite-dimensional subspace $V_h \subset V$ spanned by basis functions $(\varphi_i)_{i=1,\dots,n}$ is given. Then we define the usual stiffness and mass matrices

$$\underline{K}_{x_0} = \left[a(\varphi_j, \varphi_i; x_0)\right]_{i,j=1}^n , \qquad \underline{M}_{x_0} = \left[b(\varphi_j, \varphi_i; x_0)\right]_{i,j=1}^n$$

as well as their derivatives

$$\mathbf{D}_{x_0}^k \underline{K} = \left[\mathbf{D}_{x_0}^k a(\varphi_j, \varphi_i; \cdot)\right]_{i,j=1}^n , \qquad \mathbf{D}_{x_0}^k \underline{M} = \left[\mathbf{D}_{x_0}^k b(\varphi_j, \varphi_i; \cdot)\right]_{i,j=1}^n , \qquad k \in \mathbb{N} .$$

Both stiffness and mass matrix are elliptic if their respective bilinear forms are. The matrices, as well as their derivatives, are symmetric for the same reason. Quite often, given an appropriate choice of basis functions, the stiffness and mass matrix as well as their derivatives are sparse.
The discrete EVP is then to find $(\underline{\lambda}_x, \underline{u}_x) \in \mathbb{R} \times \mathbb{R}^n$

$$\underline{K}_x \underline{u}_x = \underline{M}_x \underline{u}_x \, \underline{\lambda}_x \tag{III.42}$$

where $\lambda_{x,h} = \underline{\lambda}_x$ is the discretized eigenvalue at $x \in X$ and

$$u_{x,h} = \sum_{j=1}^n [\underline{u}_x]_j \varphi_j \in V_h$$

is the discretized eigenfunction. We again use the vector notation

$$\underline{\boldsymbol{u}}_x = [(\underline{u}_x)_1, \dots, (\underline{u}_x)_m] \in \mathbb{R}^{n \times m}$$

with m the multiplicity. The discrete version of the saddle point equation of the first-order derivative (III.11) is then given by

(III.43)
$$\underbrace{\begin{bmatrix} \underline{K}_{x_0} - \underline{M}_{x_0}\,\underline{\lambda}_{x_0} & -\underline{M}_{x_0}\cdot\underline{\boldsymbol{u}}_{x_0} \\ -\underline{\boldsymbol{u}}_{x_0}^{\top}\cdot\underline{M}_{x_0} & \end{bmatrix}}_{=:\mathfrak{S}\in\mathbb{R}^{(n+m)\times(n+m)}} \underbrace{\begin{bmatrix} \mathbf{D}_{x_0}\underline{\boldsymbol{u}} \\ \mathbf{D}_{x_0}\underline{\boldsymbol{\lambda}} \end{bmatrix}}_{\in\mathbb{R}^{(n+m)\times m}} = \begin{bmatrix} -(\mathbf{D}_{x_0}\underline{K})\cdot\underline{\boldsymbol{u}}_{x_0} + (\mathbf{D}_{x_0}\underline{M})\cdot\underline{\boldsymbol{u}}_{x_0}\,\underline{\lambda}_{x_0} \\ \operatorname{diag}_{i=1,\dots,m}\frac{[\underline{\boldsymbol{u}}_{x_0}]_i^{\top}\cdot(\mathbf{D}_{x_0}\underline{M})\cdot[\underline{\boldsymbol{u}}_{x_0}]_i}{2} \end{bmatrix}.$$

The blank part of the system matrix $\mathfrak{S}$ represents $m\times m$ zero entries. We omit further derivatives (III.15), which can be formulated in complete analogy using $\mathfrak{S}$. The first-order formula of corollary III.9 can also be computed as

$$\mathbf{D}_{x_0}\underline{\boldsymbol{\lambda}} = \underline{\boldsymbol{u}}_{x_0}^{\top}\cdot(\mathbf{D}_{x_0}\underline{K})\cdot\underline{\boldsymbol{u}}_{x_0} - \underline{\boldsymbol{u}}_{x_0}^{\top}\cdot(\mathbf{D}_{x_0}\underline{M})\cdot\underline{\boldsymbol{u}}_{x_0}\,\underline{\lambda}_{x_0}\,. \tag{III.44}$$

III.6.2. Joining Multiple Saddle Point Equations by Vectorization. We may find it convenient to calculate the derivatives of eigenpairs with different eigenvalues in one system of equations. Thus, we consider a vector notation with $\underline{\boldsymbol{u}}\in\mathbb{R}^{n\times(m_1+m_2+\cdots)}$ to join the systems for eigenvalues $(\lambda_{x_0})_1,(\lambda_{x_0})_2,\dots$ with multiplicities $m_1,m_2,\dots$. When considering the main condition

$$\begin{aligned}\underline{K}_{x_0}\cdot(\mathbf{D}_{x_0}\underline{\boldsymbol{u}}) - \underline{M}_{x_0}\cdot(\mathbf{D}_{x_0}\underline{\boldsymbol{u}})\cdot\underline{\boldsymbol{\lambda}}_{x_0} - \underline{M}_{x_0}\cdot\underline{\boldsymbol{u}}_{x_0}\cdot(\mathbf{D}_{x_0}\underline{\boldsymbol{\lambda}}) \\ = -(\mathbf{D}_{x_0}\underline{K})\cdot\underline{\boldsymbol{u}}_{x_0} + (\mathbf{D}_{x_0}\underline{M})\cdot\underline{\boldsymbol{u}}_{x_0}\cdot\underline{\boldsymbol{\lambda}}_{x_0}\,,\end{aligned}$$

we find that the unknown $\mathbf{D}_{x_0}\underline{\boldsymbol{u}}$ is located to the left of $\underline{\boldsymbol{\lambda}}_{x_0}$, which now does not repeat the same eigenvalue on its diagonal. Thus, we first have to use the Kronecker product and the vectorization operator $\operatorname{vec}:\mathbb{R}^{n\times m}\to\mathbb{R}^{nm\times 1}$, cf. [**76**, eq. (273)], in order to isolate $\operatorname{vec}(\mathbf{D}_{x_0}\underline{\boldsymbol{u}})$ as a right factor

$$\begin{aligned}\left(\boldsymbol{I}\otimes\underline{K}_{x_0} - \underline{\boldsymbol{\lambda}}_{x_0}^{\top}\otimes\underline{M}_{x_0}\right)\cdot\operatorname{vec}(\mathbf{D}_{x_0}\underline{\boldsymbol{u}}) - \left(\boldsymbol{I}\otimes(\underline{M}_{x_0}\cdot\underline{\boldsymbol{u}}_{x_0})\right)\cdot\operatorname{vec}(\mathbf{D}_{x_0}\underline{\boldsymbol{\lambda}}) \\ = \operatorname{vec}\left(-(\mathbf{D}_{x_0}\underline{K})\cdot\underline{\boldsymbol{u}}_{x_0} + (\mathbf{D}_{x_0}\underline{M})\cdot\underline{\boldsymbol{u}}_{x_0}\cdot\underline{\boldsymbol{\lambda}}_{x_0}\right).\end{aligned} \tag{III.45}$$

Then we can formulate an equivalent saddle point equation

$$\begin{aligned}&\begin{bmatrix} \boldsymbol{I}\otimes\underline{K}_{x_0} - \underline{\boldsymbol{\lambda}}_{x_0}^{\top}\otimes\underline{M}_{x_0} & \boldsymbol{I}\otimes(-\underline{M}_{x_0}\cdot\underline{\boldsymbol{u}}_{x_0}) \\ \boldsymbol{I}\otimes(-\underline{\boldsymbol{u}}_{x_0}^{\top}\cdot\underline{M}_{x_0}) & \end{bmatrix}\begin{bmatrix}\operatorname{vec}(\mathbf{D}_{x_0}\underline{\boldsymbol{u}}) \\ \operatorname{vec}(\mathbf{D}_{x_0}\underline{\boldsymbol{\lambda}})\end{bmatrix} \\ &\qquad = \begin{bmatrix}\operatorname{vec}\left(-(\mathbf{D}_{x_0}\underline{K})\cdot\underline{\boldsymbol{u}}_{x_0} + (\mathbf{D}_{x_0}\underline{M})\cdot\underline{\boldsymbol{u}}_{x_0}\cdot\underline{\boldsymbol{\lambda}}_{x_0}\right) \\ \clubsuit\end{bmatrix},\end{aligned}$$

where $\clubsuit$ is a placeholder for the orthonormality conditions.
Note that in this formulation $\mathbf{D}_{x_0}\underline{\boldsymbol{\lambda}}\in\mathbb{R}^{(m_1+m_2+\cdots)\times(m_1+m_2+\cdots)}$. Previously, its block-diagonal nature was implicit, since we solved multiple smaller saddle point equations, which in total had fewer degrees of freedom than the larger saddle point equation. Now, in order to obtain the old derivatives such that $\mathbf{D}_{x_0}\underline{\boldsymbol{\lambda}}$ is block-diagonal, we have to include the implicit conditions (III.17) that control the scalar products relating to the joined eigenvalues explicitly in $\clubsuit$.

If we want the new degrees of freedom in $\mathbf{D}_{x_0} \underline{\boldsymbol{\lambda}}$ to vanish anyway, it is clear that this formulation is less efficient than considering only one eigenspace. However, if we set the new orthogonality conditions to zeros, this formulation lets us consider the eigenvalues with respect to the joint eigenspace. This is useful when we want to calculate the derivatives of the eigenvalue trajectories with respect to the eigenspace at a point $x \in B(x_0)$ in the neighborhood of the degenerate point x_0, where the actual eigenvalues have split up and are no longer degenerate.

III.6.3. Solution Strategies for Saddle Point Equations. We give a brief subsumption of solution strategies to solve saddle point equations (III.43) following the review article [**8**]. In [**8**, Chapter 4] solvers are classified into *segregated* methods that solve parts of the equation and *coupled* methods that solve the equation all at once.

Calculating Basis Coefficients. We have already seen a segregated ansatz in the basis representation of eigenfunctions (III.16). If the matrices of (III.42) are relatively small, i.e., calculating all eigenvectors is feasible, we can use the coefficient representation (III.16) and calculate all coefficients (III.17) explicitly. The derivatives of the eigenvalues can then be calculated using corollary III.9.

The Pseudoinverse. Consider the matrix resulting from the main and normalization conditions of a single eigenpair $(\underline{\lambda}_{x_0}, [\underline{\boldsymbol{u}}_{x_0}]_i) \in \mathbb{R} \times V$

$$\tilde{\mathfrak{S}} := \begin{bmatrix} \underline{K}_{x_0} - \underline{M}_{x_0} \underline{\lambda}_{x_0} & -\underline{M}_{x_0} \cdot [\underline{\boldsymbol{u}}_{x_0}]_i \\ -[\underline{\boldsymbol{u}}_{x_0}]_i^\top \cdot \underline{M}_{x_0} & \end{bmatrix} \in \mathbb{R}^{(n+1)\times(n+1)} .$$

If the eigenvalue is part of a degenerate eigenvalue, we have already seen that $\tilde{\mathfrak{S}}$ is singular and rank deficient by $m-1$, which we fixed by considering the derivative with respect to the eigenspace and postulating orthogonality conditions. The orthogonality conditions are equivalent to picking a minimum-norm solution, since we set the coefficients with respect to the null space to zero, cf. (III.17).

This approach can be equivalently formulated using the pseudoinverse of $\tilde{\mathfrak{S}}$.

DEFINITION III.30. Let $\tilde{\mathfrak{S}} = \boldsymbol{U} \cdot \boldsymbol{\Lambda} \cdot \boldsymbol{V}$ be a SVD of a matrix with $\boldsymbol{U}, \boldsymbol{V}$ orthogonal and $\boldsymbol{\Lambda}$ diagonal, then we can define its **pseudoinverse** as

$$\tilde{\mathfrak{S}}^\dagger := \boldsymbol{V}^\top \cdot \boldsymbol{\Lambda}^\dagger \cdot \boldsymbol{U}^\top . \tag{III.46}$$

where $\boldsymbol{\Lambda}^\dagger$ is a diagonal matrix with

$$[\boldsymbol{\Lambda}^\dagger]_{ii} = \begin{cases} [\boldsymbol{\Lambda}]_{ii}^{-1} & \text{for } [\boldsymbol{\Lambda}]_{ii} \neq 0 , \\ 0 & \text{else} . \end{cases}$$

If $\tilde{\mathfrak{S}}$ is not singular, $\tilde{\mathfrak{S}}^\dagger$ coincides with the inverse. Thus, for the first-order derivative, we can write

$$\begin{bmatrix} \mathbf{D}_{x_0} [\underline{\boldsymbol{u}}]_i \\ \mathbf{D}_{x_0} \underline{\lambda} \end{bmatrix} = \tilde{\mathfrak{S}}^\dagger \cdot \begin{bmatrix} -(\mathbf{D}_{x_0} \underline{K}) \cdot [\underline{\boldsymbol{u}}_{x_0}]_i + (\mathbf{D}_{x_0} \underline{M}) \cdot [\underline{\boldsymbol{u}}_{x_0}]_i \, \underline{\lambda}_{x_0} \\ \frac{[\underline{\boldsymbol{u}}_{x_0}]_i^\top \cdot (\mathbf{D}_{x_0} \underline{M}) \cdot [\underline{\boldsymbol{u}}_{x_0}]_i}{2} \end{bmatrix}$$

for degenerate and non-degenerate eigenvalues alike. In this case, we may also write

$$\mathbf{D}_{x_0}[\underline{\boldsymbol{u}}]_i = \begin{bmatrix} \underline{K}_{x_0} - \underline{M}_{x_0}\underline{\lambda}_{x_0} \\ -[\underline{\boldsymbol{u}}_{x_0}]_i^\top \cdot \underline{M}_{x_0} \end{bmatrix}^\dagger \cdot \begin{bmatrix} -(\mathbf{D}_{x_0}\underline{K}) \cdot [\underline{\boldsymbol{u}}_{x_0}]_i + (\mathbf{D}_{x_0}\underline{M}) \cdot [\underline{\boldsymbol{u}}_{x_0}]_i \, \underline{\lambda}_{x_0} \\ \frac{[\underline{\boldsymbol{u}}_{x_0}]_i^\top \cdot (\mathbf{D}_{x_0}\underline{M}) \cdot [\underline{\boldsymbol{u}}_{x_0}]_i}{2} \end{bmatrix},$$

If the normalization entry vanishes, e.g., when the scalar product is constant, we can further simplify by expressing the normalization condition using the pseudoinverse, cf. [**76**, Chapter 2.3]. Usually stating the orthogonality conditions explicitly is more computationally efficient than using the pseudoinverse for numerical implementation. For more information on the pseudoinverse, the interested reader is referred to [**40**, Chapter 5.5.2].

Direct solvers. The most straightforward way to solve the saddle point equation is to use the solver library for linear equations which are usually based on a factorization of the system matrix. The **LDL factorization** [**8**, Chapter 7] is an algorithm that can take advantage of the symmetry of the system matrix $\mathfrak{S}$ and factorize it as

$$\mathfrak{S} =: L \cdot D \cdot L^\top,$$

where D is a diagonal matrix and L is a lower triangular matrix, cf. [**40**, Chapter 4.1.2]. If $\mathfrak{S}$ is sparse, there is a special version of the LDL factorization, which yields a factorization

$$\mathfrak{S} =: Q^\top \cdot L \cdot D \cdot L^\top \cdot Q \tag{III.47}$$

where Q is a sparse permutation matrix such that L is sparse. This method is often most efficient for large, symmetric, and sparse stiffness and mass matrices. If the formulation of (III.43) is used, the system matrix needs to be factorized only once and the factor matrices of (III.47) can be reused for all m eigenvalues and higher-order derivatives.

III.7. Numerical Examples

In the last section, we present numerical examples for the characterization of derivatives with respect to the eigenspace and the subsequent polarization. We first consider the Laplace EVP with a sample perturbation by random fields. Then, we discuss two matrix EVP in order to construct more examples where the polarization is more interesting.

III.7.1. A Perturbed Laplace Eigenvalue Problem. We demonstrate the tracking and approximation of a subset of eigenvalues of a variational EVP. To this end, we build on example II.77, i.e., the domain is $\mathscr{D} = (0,1)^2$. As the parameterized bilinear forms,

we consider

$$a(u,v;x) = \int_{\mathscr{D}} \langle \mu_x(\mathbf{x}) \operatorname{grad} u(\mathbf{x}), \operatorname{grad} v(\mathbf{x}) \rangle_{\mathbb{R}^2} \, \mathrm{d}\mathbf{x} \,,$$
$$b(u,v;x) = \int_{\mathscr{D}} \varepsilon_x(\mathbf{x}) u(\mathbf{x}) v(\mathbf{x}) \, \mathrm{d}\mathbf{x} \,,$$

where we take μ, ε to be smooth scalar coefficient fields which depend linearly on $x \in X$, i.e.,

$$\mu \in \mathscr{L}(X; C^\infty(\mathscr{D})) \,, \qquad \varepsilon \in \mathscr{L}(X; C^\infty(\mathscr{D})).$$

We choose $x_0 = 0$ as the reference point with

$$x_0 \mapsto \mu_{x_0} = 1 \,, \qquad x_0 \mapsto \varepsilon_{x_0} = 1 \,.$$

Thus, the reference point $x_0 = 0$ relates to example II.77, where the smallest eigenvalue λ_1 is non-degenerate, and the second and third eigenvalue form a degenerate eigenvalue of multiplicity $m = 2$.
We again discretize $V = H_0^1(\mathscr{D})$ using piecewise linear basis functions on a uniform triangular mesh with 545 nodes, 64 of which are boundary points, leaving 481 degrees of freedom for the solution of the matrix EVP.
To construct a perturbation, we sample a random field according to the squared-exponential covariance operator, cf. example II.99, using correlation length $\rho = \frac{1}{4}$. For the discretization of the random fields, we calculate the covariance matrix $\Sigma \in \mathbb{R}^{545\times 545}$ of the covariances at all nodes and use an EVP-solver to calculate its eigenpairs. Using an absolute tolerance of 10^{-5}, we select the 57 largest eigenpairs $(\nu_i, \phi_i)_{i=1,\dots,57}$ of the covariance matrix. We then express both parameters as independent samples of this random field

$$x_\xi \mapsto \mu_{x_\xi} = \mu_{x_0} + \sum_{i=1}^{57} \sqrt{\nu_i} \phi_i \xi_i \,,$$
$$x_\xi \mapsto \varepsilon_{x_\xi} = \varepsilon_{x_0} + \sum_{i=58}^{114} \sqrt{\nu_i} \phi_i \xi_i \,,$$

such that $X = [-\frac{1}{2}, \frac{1}{2}]^{114}$, where we independently sample according to a uniform distribution

$$\xi_i \sim \mathscr{U}([-\tfrac{1}{2}, \tfrac{1}{2}])$$

The perturbation is bounded, so we can ensure ellipticity if the perturbation is scaled sufficiently small.
Since we have a sample of the random field at the nodes of the grid, we can approximate its value at the center of the element by its mean as a first-order approximation. Using this value for the approximation of the entries of the FE matrices is consistent

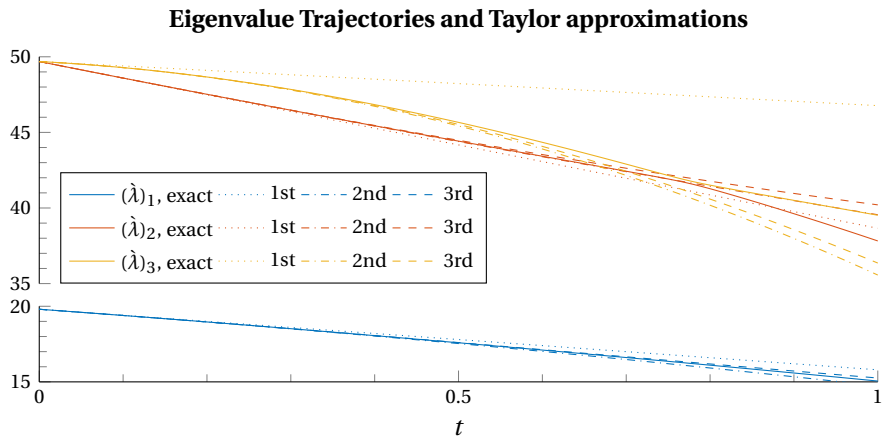

FIGURE III.5. Sample trajectory of $(\grave{\lambda}_i, \grave{u}_i)_{i=1,2,3}$ of the Laplace EVP and first- to third-order Taylor approximations.

with the use of piecewise linear basis functions and yields a first-order FE approximation like the model without perturbation.
We draw a sample $\xi \in X$ and observe the evolution of the three smallest eigenvalues and their associated eigenfunctions on the path

$$\gamma : [0,1] \to X\,, \qquad\qquad t \mapsto \gamma(t) = x_0 + t\xi\,.$$

Drawing two samples of the random fields will almost surely lead to a different initial polarization $\boldsymbol{P}_{t_0}$. In this perturbation model, the second and third eigenvalue are thus coupled.
Sampled Trajectory. In order to test the characterization of derivatives with respect to the eigenspace and of the polarization matrix, we calculate the Taylor approximations (cf. theorem II.42) of the eigenvalues up to seventh-order. For the eigenfunctions, we calculate Taylor approximations up to seventh-order for the first non-degenerate eigenfunctions and up to sixth-order for the degenerate eigenfunctions. We choose these orders, since, for the degenerate eigenpairs, we calculate the deciding order $\grave{k} = 1$.
The comparison of the eigenvalue trajectories with the first- to third-order Taylor approximations can be seen in fig. III.5. The degenerate eigenspace splits into two separate trajectories and is well approximated around the reference point t_0. In this sample, near $t \approx \frac{3}{4}$, the eigenvalue trajectories of the formerly degenerate eigenvalues approach each other again. They do not actually become degenerate again on this path, but repel each other. Here, the quality of the Taylor approximation deteriorates rapidly since the Taylor approximation follows the other eigenvalue. It is reasonable

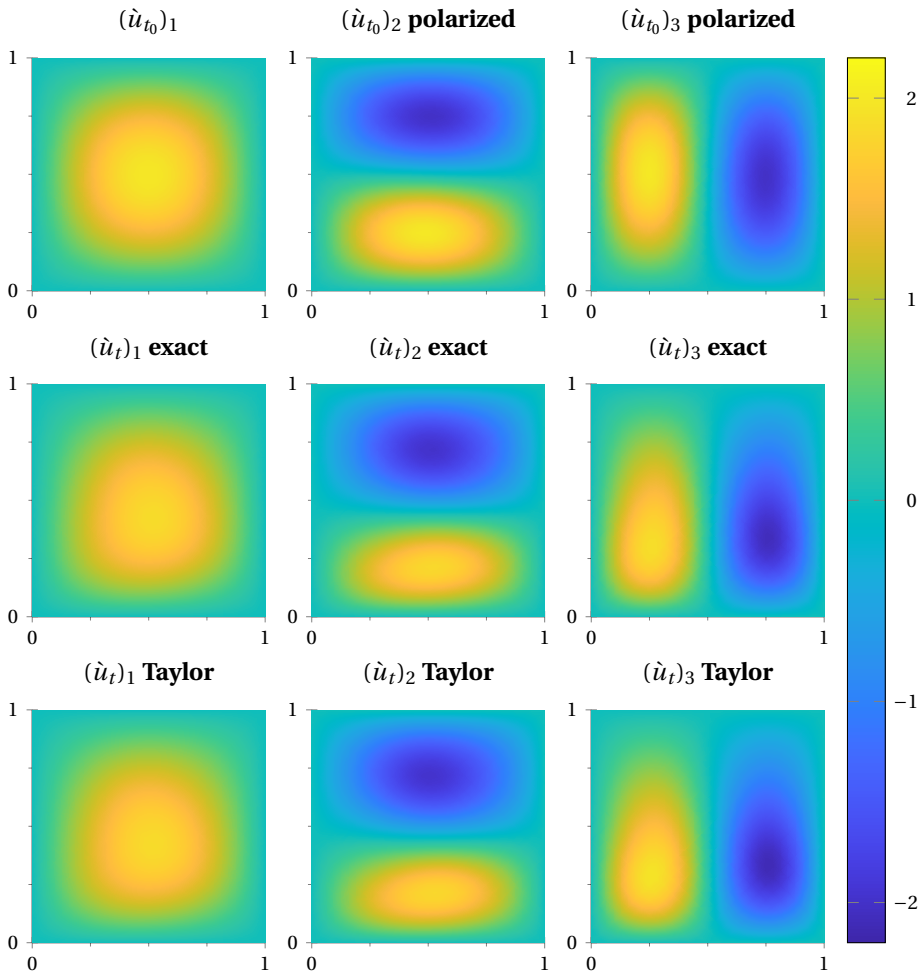

FIGURE III.6. Comparison of unperturbed and perturbed eigenfunctions of Laplace EVP (exact and second-order Taylor approximations) at $t = \frac{1}{2}$.

to speculate that for some very similar sample, they would actually have become degenerate and crossed, as this is hyperbolic behavior is usually encountered for a close miss of the degenerate points at the tip of two cone-like structures, compare fig. I.2. The eigenfunction of the three eigenvalues at $t \approx \frac{1}{2}$ can be seen in fig. III.6. It compares

(1) their unperturbed state, for degenerate eigenvalues polarized according to the calculated initial polarization $\boldsymbol{P}_{t_0}$,
(2) the exact perturbed eigenfunctions,
(3) the second-order Taylor approximation of the perturbed eigenfunctions.

The sampled polarization is not aligned to the coordinate axis. Compared to the eigenfunction basis of fig. II.1, the initial polarization is

$$\boldsymbol{P}_{t_0} \approx \begin{bmatrix} 0.0295 & -0.9996 \\ -0.9996 & -0.0295 \end{bmatrix}$$

and for other samples any orthogonal matrix may occur as the initial polarization. Similarly as for the eigenvalues, we can see that the approximations of the eigenfunctions are qualitatively good for $t \approx \frac{1}{2}$. For $t > \frac{3}{4}$ the approximation closely resamples the wrong eigenfunction (one also with the wrong sign), since the exact eigenfunction trajectories behave similarly to fig. III.2.

Verification of the Order of Convergence. We compare the residue term of the calculated Taylor approximations for the eigenpairs at

$$t \in \left\{ 2^n \,:\, n = -10, -9, \dots, 0 \right\} .$$

We evaluate the residues by comparing the Taylor approximations with the exact solutions. Then we calculate the absolute value of the residue for the eigenvalues and the $L^2(\mathscr{D})$-norm for the eigenfunctions. The results are illustrated in fig. III.7 in comparison to the desired convergence rates. We can see that the error terms all decay in accordance with corollary II.43, confirming the validity of the characterization of the derivatives with respect to the eigenspaces, as well as the polarization matrix and its derivatives.

III.7.2. Non-uniform Polarization Order. In order to verify the characterization of the polarization matrix more thoroughly, we consider two additional parameterized matrix EVPs. Our goal is to find simple examples where the decision matrix $\grave{\boldsymbol{k}}$ is a matrix that is not filled uniformly as in the previous example. To this end, we consider a parameterized matrix model

$$K : \mathbb{R}^2 \to \mathbb{R}^2 , \qquad x \mapsto K_x = 2\boldsymbol{I} + \sum_{k=1}^{4} \frac{x_1^k}{k!} \mathbf{D}_0^k K[x_1] + \sum_{k=1}^{4} \frac{x_2^k}{k!} \mathbf{D}_0^k K[x_2] \tag{III.48a}$$

and a path

$$\gamma : \mathbb{R} \to \mathbb{R}^2 \qquad t \mapsto \gamma(t) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} + t \begin{bmatrix} 1 \\ 1 \end{bmatrix} . \tag{III.48b}$$

In order to verify the order of convergence, we consider the points

$$t \in \left\{ 2^n \,:\, n = -15, -9, \dots, 1 \right\}$$

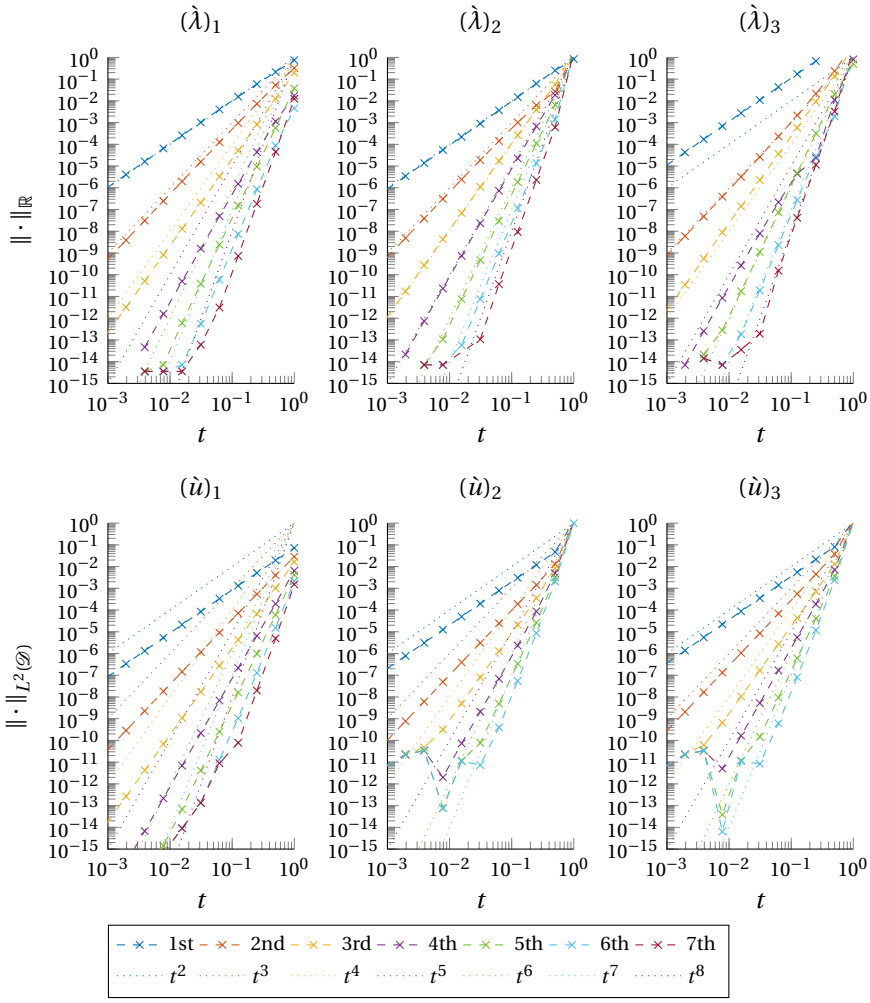

FIGURE III.7. Convergence of the residue terms of the Taylor approximation (1st to 6th/7th-order) for first three eigenvalues and respective eigenfunctions of the parameterized Laplace EVP.

on the path. We do not perturb the a *mass matrix* for a generalized EVP in these examples, to keep these examples concise[4].

[4]The polarization depends on the intermediate results of the derivatives with respect to the eigenspace, so the only difference in terms of the polarization derivatives is the normalization condition. We encounter

Crossing and Deflection of Three Eigenvalues. For the first example, we consider a matrix in $\mathbb{R}^{3\times3}$ and set (III.48a) to

$$\mathbf{D}_{x_0} K[x_1] = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}, \qquad \mathbf{D}_{x_0} K[x_2] = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

$$\mathbf{D}^2_{x_0} K[x_1] = 2\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}, \qquad \mathbf{D}^2_{x_0} K[x_2] = 2\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix},$$

$$\mathbf{D}^k_{x_0} K[x_1] = k!\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}, \qquad \mathbf{D}^k_{x_0} K[x_2] = k!\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}, \qquad k = 3,4\,.$$

For path (III.48b), the decision matrix is

$$\grave{\boldsymbol{k}} = \begin{bmatrix} 2 & 2 & 1 \\ 2 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix}. \tag{III.49}$$

In fig. III.8 the trajectories of the eigenvalues are illustrated. The second-order Taylor approximations inform the local bifurcation behavior at the degenerate point, cf. lemma III.17, and are also included. Taylor approximations of the eigenfunctions are given up to the order of the derivative of the polarization matrix which corresponds to the fourth-order derivative of the eigenvalue according to (III.49). The orders of convergence of the residues of the Taylor approximations are each confirmed.

Deflection in Pairs while Crossing as Pairs. For the second example, we consider a matrix in $\mathbb{R}^{4\times4}$ and set (III.48a) to

$$\mathbf{D}_{x_0} K[x_1] = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \qquad \mathbf{D}_{x_0} K[x_2] = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

$$\mathbf{D}^2_{x_0} K[x_1] = 2\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}, \qquad \mathbf{D}^2_{x_0} K[x_2] = 2\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix},$$

$$\mathbf{D}^k_{x_0} K[x_1] = k!\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}, \qquad \mathbf{D}^k_{x_0} K[x_2] = k!\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}, \qquad k = 3,4\,.$$

---

examples where first-order derivatives may coincide for a generalized EVP in chapter V when considering shape deformations.

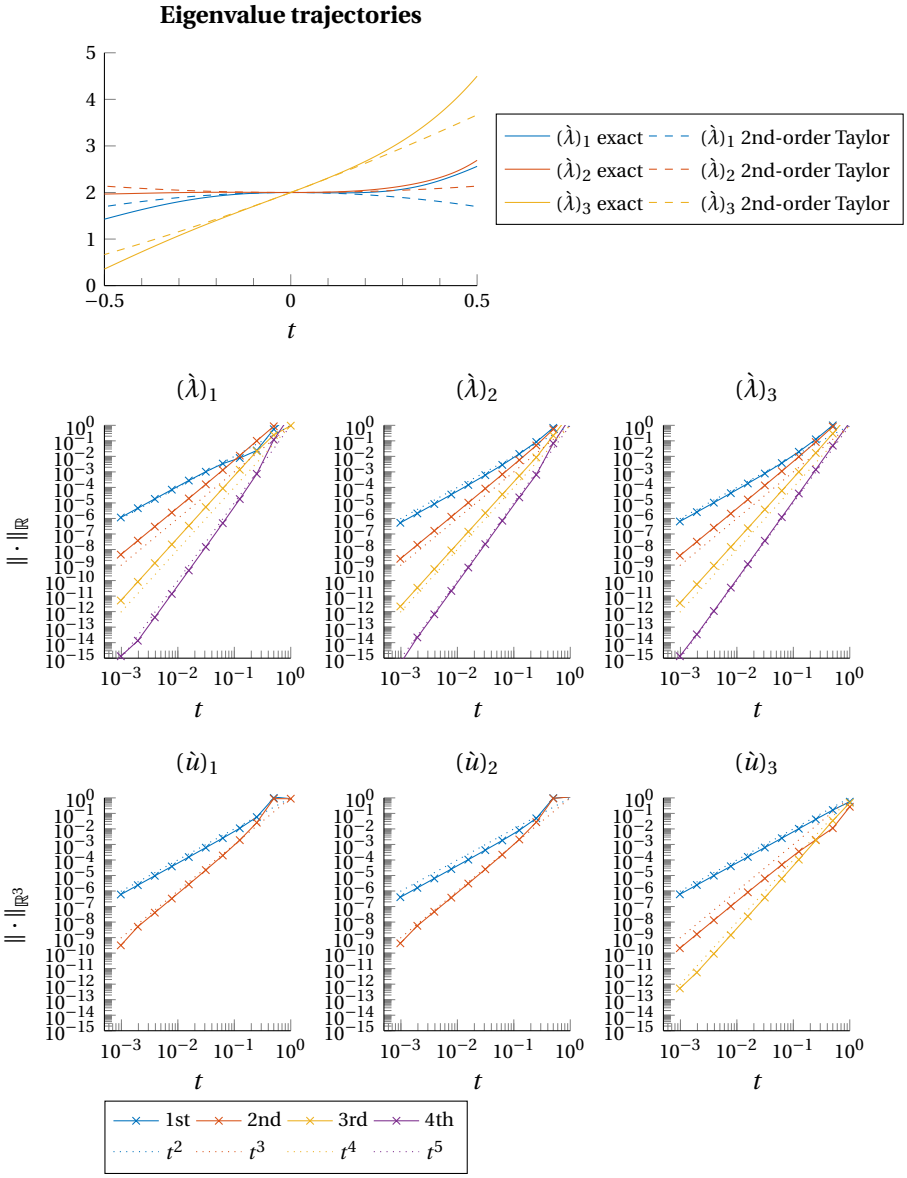

FIGURE III.8. Trajectories of eigenvalues and convergence of eigenpair approximations for the parameterized matrix EVP with crossing and deflection coinciding in one point.

For path (III.48b), the decision matrix is

$$\grave{\boldsymbol{k}} = \begin{bmatrix} 2 & 2 & 1 & 1 \\ 2 & 2 & 1 & 1 \\ 1 & 1 & 2 & 2 \\ 1 & 1 & 2 & 2 \end{bmatrix}. \tag{III.50}$$

The trajectories of the eigenvalues and the second-order Taylor approximations are provided in fig. III.9 according to (III.50). According to lemma III.17 we observe that the pair of first and second eigenvalue cross the trajectory of the third and fourth eigenvalue, while each of them retains their order in their respective pairing. The convergence plots confirm the validity of the derivatives.

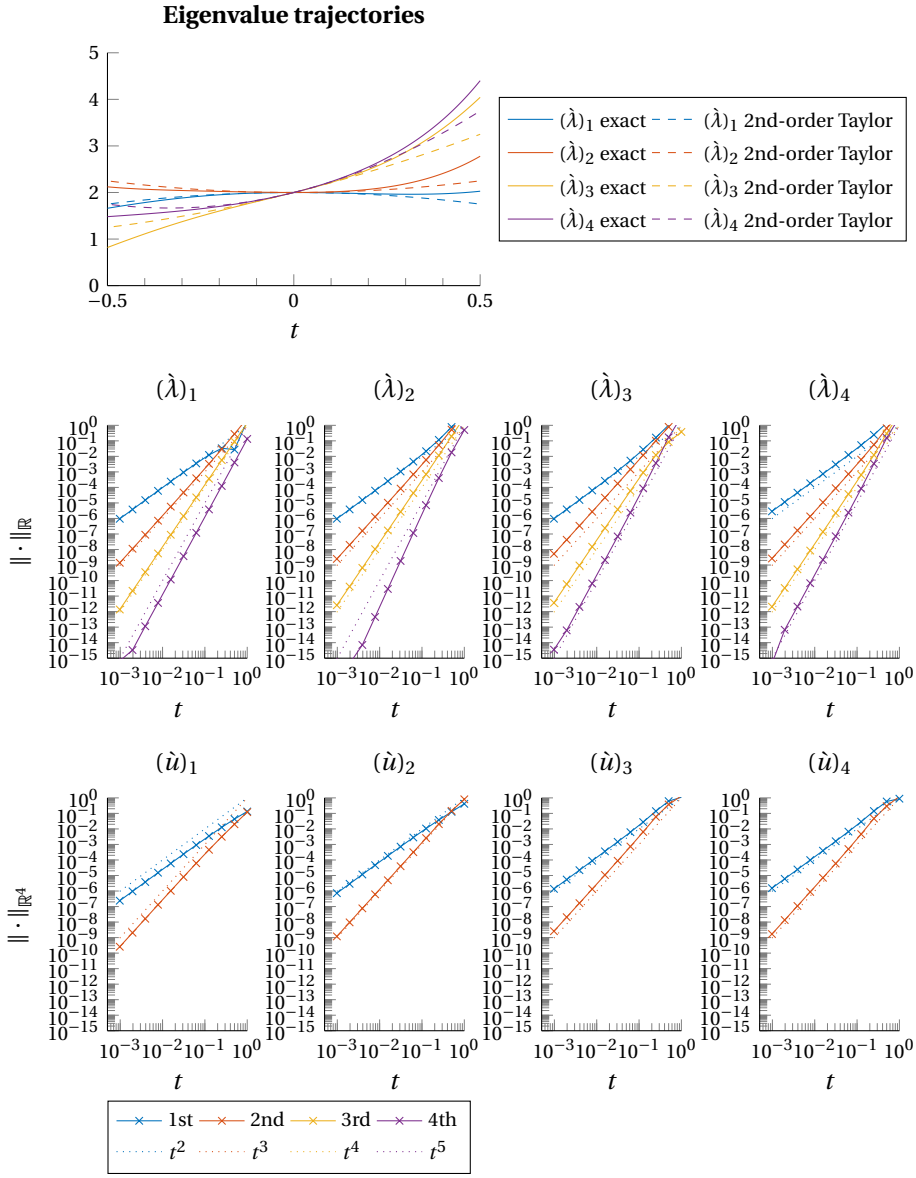

FIGURE III.9. Trajectories of eigenvalues and convergence of eigenpair approximations for the parameterized matrix EVP with deflection in pairs while crossing as pairs.

CHAPTER IV

Uncertainty Quantification

In this chapter, we characterize the mean, correlation, and covariance of eigenpairs with respect to their eigenspace as described in chapter III, where the EVP depends on a multidimensional parameter $x \in X$. The results of this chapter correspond to the stochastic aspects of [**25**], extending the approximations proposed therein to second-order approximations. These second-order approximations are made possible due to the characterization of second-order derivatives with respect to the eigenspace in chapter III.
We discuss uncertainty quantification for eigenpair trajectories with respect to the eigenspace exclusively, since they are always locally well defined, cf. theorem III.2 and section III.5. For non-degenerate eigenpairs $(\lambda, u) \in \mathbb{R} \times V$, this ansatz is again equal to the eigenpairs in the traditional sense, cf. corollary III.3. In the simpler case of one-dimensional parameter spaces, the presented ansatz can also be used on the polarized eigenpairs $(\grave{\lambda}, \grave{u}) \in \mathbb{R} \times V$ instead. This is also true for uncoupled eigenpairs, cf. remark III.24, an example of which is presented in [**44**].

IV.1. Stochastic Eigenvalue Problem

We first state the stochastic setting used in this chapter.

ASSUMPTION IV.1. Let $(\Omega, \mathscr{A}, \mathbb{P})$ be a probability space and consider the EVP (III.2) given assumption III.1. We express the stochastic parameter $x \in X$ via a stochastic perturbation model of the form

$$x_\xi(\omega) = x_0 + \xi(\omega) \,, \qquad \omega \in \Omega \,, \tag{IV.1}$$

where $x_0 \in X$ is a deterministic reference point and $\xi : \Omega \to X$ is a random variable. In this chapter, we relax the analyticity of the bilinear forms to

$$a \in C^3(X; \mathscr{L}^{(2)}(V;\mathbb{R})) \,, \qquad b \in C^3(X; \mathscr{L}^{(2)}(H;\mathbb{R})) \,.$$

As in corollary III.4, the regularity of the eigenpair trajectories with respect to the eigenspace of multiplicity $m \in \mathbb{N}$ is then only a C^3-mapping. Since the derivatives of the eigenpair are thus bounded in some neighborhood of $x_0 \in X$, there is a neighborhood

$B(x_0) \subset X$ of x_0 such that

$$\boldsymbol{\lambda} \in C^3(B(x_0); \mathbb{R}^{m\times m}) \cap L^3_{\mathbb{P}_\xi}(B(x_0); \mathbb{R}^{m\times m}) \,, \tag{IV.2a}$$

$$\boldsymbol{u} \in C^3(B(x_0); V^m) \cap L^3_{\mathbb{P}_\xi}(B(x_0); V^m) \,. \tag{IV.2b}$$

Thus, the mean, correlation, and covariance, cf. definition II.88, of the trajectories of the eigenvalue and eigenfunction with respect to the eigenspace are well defined in this neighborhood $B(x_0)$. For some results, we will require stronger C^4-regularity and L^4-integrability. The latter can analogously be assumed to hold in some neighborhood $B(x_0)$.

Given the assumed regularity, the eigenpair trajectories with respect to the eigenspace can be expressed as Taylor expansions

$$\boldsymbol{\lambda}_{x_\xi}(\omega) = \boldsymbol{\lambda}_{x_0} + \mathbf{D}_{x_0}\boldsymbol{\lambda}[\xi(\omega)] + \tfrac{1}{2}\mathbf{D}^2_{x_0}\boldsymbol{\lambda}[\xi(\omega)] + \mathcal{O}(\|\xi(\omega)\|^3_X) \,, \tag{IV.3a}$$

$$\boldsymbol{u}_{x_\xi}(\omega) = \boldsymbol{u}_{x_0} + \mathbf{D}_{x_0}\boldsymbol{u}[\xi(\omega)] + \tfrac{1}{2}\mathbf{D}^2_{x_0}\boldsymbol{u}[\xi(\omega)] + \mathcal{O}(\|\xi(\omega)\|^3_X) \tag{IV.3b}$$

for all $\omega \in \Omega$.
We have already seen in chapter III that in general polarized eigenpairs are not Fréchet differentiable. In this stochastic setting, the polarization[1] $\boldsymbol{P}(\omega)$ which maps the trajectories with respect to the eigenspace to the actual eigenvalues is also a stochastic variable. Therefore, for the perturbation ansatz, we investigate the uncertainty quantification of eigenpairs exclusively with respect to the eigenspace.
We omit arguments $\omega \in \Omega$ and $\xi(\omega) \in X$ for improved readability.

IV.2. Perturbation Approximations of Stochastic Moments

Since stochastic moments are only defined locally in a neighborhood $B(x_0)$ of $x_0 \in X$, it is sensible to approximate the integrand by Taylor approximations at x_0. This approach is called **perturbation approximation** or **local sensitivity analysis** and is already frequently used to approximate stochastic moments, cf. [**2, 7, 18, 89, 95**].

THEOREM IV.2. *Let λ_{x_0} be an eigenvalue with multiplicity $m \in \mathbb{N}$ of EVP* (III.2) *at $x_0 \in B(x_0)$ with $b(\cdot,\cdot;x_0)$-orthogonal eigenbasis $\boldsymbol{u}_{x_0} \in V^m$ and let assumption IV.1 hold. Let $(\boldsymbol{\lambda}, \boldsymbol{u})$ be the unique locally analytic trajectories with respect to the eigenspace of multiplicity m, cf. theorem III.2. Then for the eigenvalue and the eigenfunction trajectory with respect to the eigenspace, i.e., $\mathrm{R} \in \{\boldsymbol{\lambda}, \boldsymbol{u}\}$, hold the approximations*

$$\mathbb{E}[\mathrm{R}] = \mathrm{R}_{x_0} + \mathbb{E}[\mathbf{D}_{x_0}\mathrm{R}] + \tfrac{1}{2}\mathbb{E}[\mathbf{D}^2_{x_0}\mathrm{R}] + \mathcal{O}(\|\xi\|^3_{L^3_{\mathbb{P}}(\Omega;X)}) \,, \tag{IV.4a}$$

$$\begin{aligned}\mathbb{C}\mathrm{or}[\mathrm{R}] = {} & \mathrm{R}_{x_0} \otimes \mathrm{R}_{x_0} + \big(\mathbb{E}[\mathbf{D}_{x_0}\mathrm{R}] \otimes \mathrm{R}_{x_0} + \mathrm{R}_{x_0} \otimes \mathbb{E}[\mathbf{D}_{x_0}\mathrm{R}]\big) \\ & + \tfrac{1}{2}\big(\mathbb{E}[\mathbf{D}^2_{x_0}\mathrm{R}] \otimes \mathrm{R}_{x_0} + \mathrm{R}_{x_0} \otimes \mathbb{E}[\mathbf{D}^2_{x_0}\mathrm{R}]\big) + \mathbb{C}\mathrm{or}[\mathbf{D}_{x_0}\mathrm{R}] + \mathcal{O}(\|\xi\|^3_{L^3_{\mathbb{P}}(\Omega;X)}) \,,\end{aligned} \tag{IV.4b}$$

$$\mathbb{C}\mathrm{ov}[\mathrm{R}] = \mathbb{C}\mathrm{ov}[\mathbf{D}_{x_0}\mathrm{R}] + \mathcal{O}(\|\xi\|^3_{L^3_{\mathbb{P}}(\Omega;X)}) \,. \tag{IV.4c}$$

[1] For example, given by the non-arbitrary choice of path $\gamma : [0,1] \to B(x_0), t \mapsto \gamma(t) = x_0 + t\xi(\omega)$.

PROOF. We apply the stochastic moments to the expansions (IV.3). Due to linearity of the mean, we only have to consider the mean of every individual term in the expansion. The unperturbed value $\mathtt{R}_{x_0}$ is deterministic. Assumption IV.1 allow us to bound the third-order (and higher-order) derivatives $\mathbb{P}$-a.s.. The residue follows from Taylor's theorem, cf. corollary II.43.
For (IV.4b), we consider the expansion

$$\begin{aligned}\mathtt{R}_{x_\xi} \otimes \mathtt{R}_{x_\xi} &= \mathtt{R}_{x_0} \otimes \mathtt{R}_{x_0} + \big(\mathbf{D}_{x_0}\mathtt{R} \otimes \mathtt{R}_{x_0} + \mathtt{R}_{x_0} \otimes \mathbf{D}_{x_0}\mathtt{R}\big) \\ &\quad + \tfrac{1}{2}\big(\mathbf{D}^2_{x_0}\mathtt{R} \otimes \mathtt{R}_{x_0} + 2\,\mathbf{D}_{x_0}\mathtt{R} \otimes \mathbf{D}_{x_0}\mathtt{R} + \mathtt{R}_{x_0} \otimes \mathbf{D}^2_{x_0}\mathtt{R}\big) + \mathscr{O}(\|\xi\|^3_{L^3_{\mathbb{P}}(\Omega;X)})\end{aligned}$$

for $\mathtt{R} \in \{\boldsymbol{\lambda}, \boldsymbol{u}\}$ in complete analogy. The formula (IV.4c) is found by combining (IV.4a) and (IV.4b) using (II.39e) on $\mathtt{R}$, then again in reverse on $\mathbf{D}_{x_0}\mathtt{R}$ and simplifying. □

Note that covariances and correlations between trajectories of eigenvalues and eigenfunctions with respect to the eigenspace as well as across different eigenspaces can be considered in complete analogy.
Often, the reference point $x_0 \in X$ is the mean of the perturbation (IV.1), i.e., ξ is centered. In this case, we get simplified perturbation approximations.

COROLLARY IV.3. *Let the assumptions of theorem IV.2 hold and assume that ξ is centered, i.e., $\mathbb{E}[\xi] = 0$. Then the perturbation approximation of theorem IV.2 with $\mathtt{R} \in \{\boldsymbol{\lambda}, \boldsymbol{u}\}$ simplifies to*

$$\mathbb{E}[\mathtt{R}] = \mathtt{R}_{x_0} + \tfrac{1}{2}\,\mathbb{E}[\mathbf{D}^2_{x_0}\mathtt{R}] + \mathscr{O}(\|\xi\|^3_{L^3_{\mathbb{P}}(\Omega;X)})\,, \tag{IV.5a}$$

$$\begin{aligned}\mathbb{C}\mathrm{or}[\mathtt{R}] &= \mathtt{R}_{x_0} \otimes \mathtt{R}_{x_0} \\ &\quad + \tfrac{1}{2}\big(\mathbb{E}[\mathbf{D}^2_{x_0}\mathtt{R}] \otimes \mathtt{R}_{x_0} + \mathtt{R}_{x_0} \otimes \mathbb{E}[\mathbf{D}^2_{x_0}\mathtt{R}]\big) + \mathbb{C}\mathrm{or}[\mathbf{D}_{x_0}\mathtt{R}] + \mathscr{O}(\|\xi\|^3_{L^3_{\mathbb{P}}(\Omega;X)})\,,\end{aligned} \tag{IV.5b}$$

$$\mathbb{C}\mathrm{ov}[\mathtt{R}] = \mathbb{C}\mathrm{or}[\mathbf{D}_{x_0}\mathtt{R}] + \mathscr{O}(\|\xi\|^3_{L^3_{\mathbb{P}}(\Omega;X)})\,, \tag{IV.5c}$$

If, in addition, $\xi \in L^4_{\mathbb{P}}(\Omega;X)$ is also skewfree, i.e., $\mathbb{E}[\xi \otimes \xi \otimes \xi] = 0$, the convergence rates improve to $\mathscr{O}(\|\xi\|^4_{L^4_{\mathbb{P}}(\Omega;X)})$.

PROOF. Since $\mathbb{E}[\xi] = 0$ and the first-order Fréchet derivatives are linear mappings, i.e.,

$$\mathbf{D}_{x_0}\boldsymbol{\lambda} \in \mathscr{L}(B(x_0);\mathbb{R}^{m\times m})\,, \qquad \mathbf{D}_{x_0}\boldsymbol{u} \in \mathscr{L}(B(x_0);V^m)\,,$$

we get $\mathbb{E}[\mathbf{D}_{x_0}\boldsymbol{\lambda}[\xi]] = \mathbf{D}_{x_0}\boldsymbol{\lambda}[\mathbb{E}[\xi]] = \mathbf{0}$ and $\mathbb{E}[\mathbf{D}_{x_0}\boldsymbol{u}[\xi]] = \mathbf{D}_{x_0}\boldsymbol{u}[\mathbb{E}[\xi]] = \mathbf{0}$, which simplifies the approximations. For the improved convergence rate, consider the third-order Taylor approximations

$$\begin{aligned}\boldsymbol{\lambda}_{x_\xi}(\omega) &= \boldsymbol{\lambda}_{x_0} + \mathbf{D}_{x_0}\boldsymbol{\lambda}[\xi(\omega)] + \tfrac{1}{2}\,\mathbf{D}^2_{x_0}\boldsymbol{\lambda}[\xi(\omega)] + \tfrac{1}{6}\,\mathbf{D}^3_{x_0}\boldsymbol{\lambda}[\xi(\omega)] + \mathscr{O}(\|\xi(\omega)\|^4_X)\,, \\ \boldsymbol{u}_{x_\xi}(\omega) &= \boldsymbol{u}_{x_0} + \mathbf{D}_{x_0}\boldsymbol{u}[\xi(\omega)] + \tfrac{1}{2}\,\mathbf{D}^2_{x_0}\boldsymbol{u}[\xi(\omega)] + \tfrac{1}{6}\,\mathbf{D}^3_{x_0}\boldsymbol{u}[\xi(\omega)] + \mathscr{O}(\|\xi(\omega)\|^4_X)\end{aligned}$$

for $\omega \in \Omega$. Then we can use the fact that the third-order Fréchet derivatives are trilinear, cf. section II.3, i.e.,

$$\mathbf{D}^3_{x_0} \boldsymbol{\lambda} \in \mathscr{L}^{(3)}(B(x_0); \mathbb{R}^{m\times m}) , \qquad \mathbf{D}^3_{x_0} \boldsymbol{u} \in \mathscr{L}^{(3)}(B(x_0); V^m) .$$

Similarly, the products

$$\begin{aligned} &\mathbf{D}^2_{x_0} \boldsymbol{\lambda} \otimes \mathbf{D}_{x_0} \boldsymbol{\lambda},\ \mathbf{D}_{x_0} \boldsymbol{\lambda} \otimes \mathbf{D}^2_{x_0} \boldsymbol{\lambda} \in \mathscr{L}^{(3)}(B(x_0); \mathbb{R}^{m\times m} \otimes \mathbb{R}^{m\times m}) , \\ &\mathbf{D}^2_{x_0} \boldsymbol{u} \otimes \mathbf{D}_{x_0} \boldsymbol{u},\ \mathbf{D}_{x_0} \boldsymbol{u} \otimes \mathbf{D}^2_{x_0} \boldsymbol{u} \in \mathscr{L}^{(3)}(B(x_0); V^m \otimes V^m) \end{aligned}$$

are trilinear. The trilinear terms, which previously dominated the convergence rate, cancel if $\mathbb{E}[\xi \otimes \xi \otimes \xi] = 0$ and the fourth-order derivatives determine the new convergence rate as $\mathscr{O}(\|\xi\|^4_{L^4_{\mathbb{P}}(\Omega;X)})$. □

REMARK IV.4. As the perturbation approximations are Taylor expansions of the stochastic moments, cf. theorem II.42, we can also consider first-order approximations by considering only first-order derivatives, i.e,

$$\mathbb{E}[\mathbb{R}] = \mathbb{R}_{x_0} + \mathbb{E}[\mathbf{D}_{x_0} \mathbb{R}] + \mathscr{O}(\|\xi\|^2_{L^2_{\mathbb{P}}(\Omega;X)}) , \tag{IV.6a}$$

$$\mathbb{C}\mathrm{or}[\mathbb{R}] = \mathbb{R}_{x_0} \otimes \mathbb{R}_{x_0} + \mathbb{E}[\mathbf{D}_{x_0} \mathbb{R}] \otimes \mathbb{R}_{x_0} + \mathbb{R}_{x_0} \otimes \mathbb{E}[\mathbf{D}_{x_0} \mathbb{R}] + \mathscr{O}(\|\xi\|^2_{L^2_{\mathbb{P}}(\Omega;X)}) , \tag{IV.6b}$$

$$\mathbb{C}\mathrm{ov}[\mathbb{R}] = 0 + \mathscr{O}(\|\xi\|^2_{L^2_{\mathbb{P}}(\Omega;X)}) . \tag{IV.6c}$$

This requires only C^2-regularity and L^2-integrability. In the setting of corollary IV.3, the mean of the first-order derivative again vanishes. The approximations then correspond to the first-order GL approximation, cf. table II.1.
Given sufficient regularity and integrality, we can theoretically also formulate higher-order approximations. However, when implemented naively, the computational costs grow exponentially in the order of approximation.
The interested reader is referred to [**16, 46**] for approximations of higher-order moments.

IV.3. Covariance and Correlation of Derivatives

In order to use the approximations for the covariance and correlations, we need to characterize the correlations and covariances of the derivatives of the eigenpairs.

THEOREM IV.5 ([**25**, Theorem 3.6]). *Consider the setting of theorem IV.2 and a vectorized version of the auxiliary functions* (III.12)*, i.e.,*

$$\begin{aligned} &\mathfrak{A} : V^m \times V^m \to \mathbb{R}^{m\times m} , && \boldsymbol{u} \times \boldsymbol{v} \mapsto \mathfrak{A}(\boldsymbol{u}, \boldsymbol{v}) := a\big(\boldsymbol{u}, \boldsymbol{v}; x_0\big) - b\big(\boldsymbol{u}, \boldsymbol{v}; x_0\big)\ \lambda_{x_0} , \\ &\mathfrak{B} : \mathbb{R}^m \times V^m \to \mathbb{R}^m , && \boldsymbol{\zeta} \times \boldsymbol{v} \mapsto \mathfrak{B}(\boldsymbol{\zeta}, \boldsymbol{v}) := \sum_{j=1}^{m} [\boldsymbol{\zeta}]_j\, b\big([\boldsymbol{u}_{x_0}]_j, \boldsymbol{v}; x_0\big) . \end{aligned}$$

Then it holds that

$$\begin{aligned}(\text{IV.7})\quad &\begin{bmatrix}\mathfrak{A}(\cdot,\boldsymbol{v})\otimes\mathrm{Id} & -\mathfrak{B}(\cdot,\boldsymbol{v})\otimes\mathrm{Id}\\ \mathfrak{B}(\boldsymbol{\zeta},\cdot)\otimes\mathrm{Id} & \end{bmatrix}\cdot\mathbb{C}\mathrm{or}\begin{bmatrix}\mathbf{D}_{x_0}\boldsymbol{u}\\ \mathbf{D}_{x_0}\boldsymbol{\lambda}\end{bmatrix}\cdot\begin{bmatrix}\mathrm{Id}\otimes\mathfrak{A}(\cdot,\boldsymbol{w}) & \mathrm{Id}\otimes\mathfrak{B}(\boldsymbol{\vartheta},\cdot)\\ -\mathrm{Id}\otimes\mathfrak{B}(\cdot,\boldsymbol{w}) & \end{bmatrix}\\ &\qquad=\mathbb{C}\mathrm{or}\begin{bmatrix}-\mathbf{D}_{x_0}a\big(\boldsymbol{u}_{x_0},\cdot;\cdot\big)+\mathbf{D}_{x_0}b\big(\boldsymbol{u}_{x_0},\cdot;\cdot\big)\cdot\boldsymbol{\lambda}_{x_0}\\ -\operatorname*{diag}_{i=1,\dots,m}\frac{\cdot\left(\mathbf{D}_{x_0}b\left([\boldsymbol{u}_{x_0}]_i,[\boldsymbol{u}_{x_0}]_i;\cdot\right)\right)}{2}\end{bmatrix}(\boldsymbol{v}\otimes\boldsymbol{\zeta},\boldsymbol{w}\otimes\boldsymbol{\vartheta})\end{aligned}$$

PROOF. The existence of the covariances on the right-hand side follows from the Cauchy–Schwarz inequality in $L^2_{\mathbb{P}}$, given (IV.2). From theorem III.8 follows

$$\begin{aligned}&\mathbb{C}\mathrm{or}\begin{bmatrix}-\mathbf{D}_{x_0}a\big(\boldsymbol{u}_{x_0},\cdot;\cdot\big)+\mathbf{D}_{x_0}b\big(\boldsymbol{u}_{x_0},\cdot;\cdot\big)\cdot\boldsymbol{\lambda}_{x_0}\\ -\operatorname*{diag}_{i=1,\dots,m}\frac{\cdot\left(\mathbf{D}_{x_0}b\left([\boldsymbol{u}_{x_0}]_i,[\boldsymbol{u}_{x_0}]_i;\cdot\right)\right)}{2}\end{bmatrix}(\boldsymbol{v}\otimes\boldsymbol{\zeta},\boldsymbol{w}\otimes\boldsymbol{\vartheta})\\ &\quad=\mathbb{C}\mathrm{or}\begin{bmatrix}\mathfrak{A}(\mathbf{D}_{x_0}\boldsymbol{u},\cdot)-\mathfrak{B}(\mathbf{D}_{x_0}\boldsymbol{\lambda},\cdot)\\ \mathfrak{B}(\cdot,\mathbf{D}_{x_0}\boldsymbol{u})\end{bmatrix}(\boldsymbol{v}\otimes\boldsymbol{\zeta},\boldsymbol{w}\otimes\boldsymbol{\vartheta})\\ &\quad=\int_\Omega\begin{bmatrix}\mathfrak{A}(\cdot,\boldsymbol{v})\otimes\mathrm{Id} & -\mathfrak{B}(\cdot,\boldsymbol{v})\otimes\mathrm{Id}\\ \mathfrak{B}(\boldsymbol{\zeta},\cdot)\otimes\mathrm{Id} & \end{bmatrix}\cdot\begin{bmatrix}\mathbf{D}_{x_0}\boldsymbol{u}\otimes\mathbf{D}_{x_0}\boldsymbol{u} & \mathbf{D}_{x_0}\boldsymbol{u}\otimes\mathbf{D}_{x_0}\boldsymbol{\lambda}\\ \mathbf{D}_{x_0}\boldsymbol{\lambda}\otimes\mathbf{D}_{x_0}\boldsymbol{u} & \mathbf{D}_{x_0}\boldsymbol{\lambda}\otimes\mathbf{D}_{x_0}\boldsymbol{\lambda}\end{bmatrix}\\ &\qquad\cdot\begin{bmatrix}\mathrm{Id}\otimes\mathfrak{A}(\cdot,\boldsymbol{w}) & \mathrm{Id}\otimes\mathfrak{B}(\boldsymbol{\vartheta},\cdot)\\ -\mathrm{Id}\otimes\mathfrak{B}(\cdot,\boldsymbol{w}) & \end{bmatrix}\mathrm{d}\mathbb{P}\\ &\quad=\begin{bmatrix}\mathfrak{A}(\cdot,\boldsymbol{v})\otimes\mathrm{Id} & -\mathfrak{B}(\cdot,\boldsymbol{v})\otimes\mathrm{Id}\\ \mathfrak{B}(\boldsymbol{\zeta},\cdot)\otimes\mathrm{Id} & \end{bmatrix}\mathbb{C}\mathrm{or}\begin{bmatrix}\mathbf{D}_{x_0}\boldsymbol{u}\\ \mathbf{D}_{x_0}\boldsymbol{\lambda}\end{bmatrix}\begin{bmatrix}\mathrm{Id}\otimes\mathfrak{A}(\cdot,\boldsymbol{w}) & \mathrm{Id}\otimes\mathfrak{B}(\boldsymbol{\vartheta},\cdot)\\ -\mathrm{Id}\otimes\mathfrak{B}(\cdot,\boldsymbol{w}) & \end{bmatrix}.\end{aligned}$$

□

The covariance of the first-order derivatives can be characterized by (II.39e). In the setting of corollary IV.3, the covariance and correlation of the first-order derivatives coincide.

The characterization of the correlation of the derivatives of the eigenpair with respect to the eigenspace is due to the structure of the saddle point equation. In analogy to the Hellmann–Feynman formula for the derivative of the eigenvalue, we can formulate the correlation of the eigenvalues only.

LEMMA IV.6 ([**25**, Lemma 3.7]). *Consider the setting of theorem IV.2. Let $\boldsymbol{\lambda}$ be the trajectory of the eigenvalue in the sense eigenspace of multiplicity m. Then it holds that*

$$\mathbb{C}\mathrm{or}[\mathbf{D}_{x_0}\boldsymbol{\lambda}]=\mathbb{C}\mathrm{or}\left[\mathbf{D}_{x_0}a\big(\boldsymbol{u}_{x_0},\boldsymbol{u}_{x_0};\cdot\big)-\mathbf{D}_{x_0}b\big(\boldsymbol{u}_{x_0},\boldsymbol{u}_{x_0};\cdot\big)\,\lambda_{x_0}\right].$$

PROOF. This follows by applying the correlation to the formula of corollary III.9. □

IV.4. Implementation

The considerations for the discretization of the saddle point equation presented in section III.6 also hold for the stochastic version of the model. Therefore, in this section, we focus on the implementation of the covariances of the derivatives outlined

in the previous section, in particular given a KLE-type decomposition. We again abbreviate the system matrix of the saddle point equation (III.43) as

$$\mathfrak{S} \in \mathbb{R}^{(n+m)\times(n+m)} .$$

IV.4.1. Covariance and Correlation Equation. Given a Galerkin discretization of the saddle point equation (III.43), the variational covariance equation of theorem IV.5 translates into the matrix equation

$$\mathfrak{S} \cdot \mathbb{C}\mathrm{or}\begin{bmatrix} \mathbf{D}_{x_0}\underline{\boldsymbol{u}} \\ \mathbf{D}_{x_0}\underline{\boldsymbol{\lambda}} \end{bmatrix} \cdot \mathfrak{S} = \mathbb{C}\mathrm{or}\begin{bmatrix} -(\mathbf{D}_{x_0}\underline{K})\cdot\underline{\boldsymbol{u}}_{x_0} + (\mathbf{D}_{x_0}\underline{M})\cdot\underline{\boldsymbol{u}}_{x_0}\,\underline{\lambda}_{x_0} \\ \operatorname*{diag}_{i=1,\dots,m} \frac{[\underline{\boldsymbol{u}}_{x_0}]_i^\top\cdot(\mathbf{D}_{x_0}\underline{M})\cdot[\underline{\boldsymbol{u}}_{x_0}]_i}{2} \end{bmatrix} . \tag{IV.8}$$

Unfortunately, the right-hand side of (IV.8) is usually densely populated, even if the underlying FE matrices and their derivatives are sparse. Therefore, simply evaluating (IV.8) as a matrix equation becomes computationally unfeasible for EVPs with many degrees of freedom. Several methods can be applied to approximate the solution of such *correlation equations*, e.g., sparse grid methods, global low-rank approximations, or hierarchical matrices [**12, 28, 29, 45, 49, 75**]. One way to obtain a global low-rank approximation of the right-hand side is the pivoted Cholesky decomposition [**45, 47**].

IV.4.2. Karhunen–Loève-type Decomposition. In the following section, we assume that the perturbation model is given as a KLE, cf. theorem II.97. Truncation after M terms yields a suitable low-rank approximation, i.e., (IV.1) is given in the more explicit form

$$x_\xi(\omega) = x_0 + \xi(\omega) = x_0 + \sum_{i=1}^{M} \phi_i z_i(\omega) \tag{IV.9}$$

with $z_i \in L^3_{\mathbb{P}}(\Omega;\mathbb{R})$ iid random variables, $\{\phi_i\}_{i=1}^M \subset X$, and $M \in \mathbb{N}\cup\{\infty\}$. Given this decomposition, we can write the series representation of the eigenpairs (IV.3) as

(IV.10a)

$$\boldsymbol{\lambda}_{x_\xi}(\omega) = \boldsymbol{\lambda}_{x_0} + \sum_{i=1}^{M} \mathbf{D}_{x_0}\boldsymbol{\lambda}[\phi_i] z_i(\omega) + \frac{1}{2}\sum_{i,j=1}^{M} \mathbf{D}^2_{x_0}\boldsymbol{\lambda}[\phi_i,\phi_j] z_i(\omega) z_j(\omega) + \mathcal{O}(\|\xi(\omega)\|_X^3) ,$$

(IV.10b)

$$\boldsymbol{u}_{x_\xi}(\omega) = \boldsymbol{u}_{x_0} + \sum_{i=1}^{M} \mathbf{D}_{x_0}\boldsymbol{u}[\phi_i] z_i(\omega) + \frac{1}{2}\sum_{i,j=1}^{M} \mathbf{D}^2_{x_0}\boldsymbol{u}[\phi_i,\phi_j] z_i(\omega) z_j(\omega) + \mathcal{O}(\|\xi(\omega)\|_X^3)$$

for $\omega \in \Omega$ and express the approximations of theorem IV.2 more explicitly.

COROLLARY IV.7. *Assuming that the perturbation of the parameter is of the form* (IV.9)*, the approximations of theorem IV.2 for the stochastic moments of the eigenvalue and*

eigenfunction trajectories with respect to the eigenspace $\mathrm{R} \in \{\boldsymbol{\lambda}, \boldsymbol{u}\}$ *can be expressed as*

$$\mathbb{E}[\mathrm{R}] = \mathrm{R}_{x_0} + \mathbf{D}_{x_0}\mathrm{R}\left[\sum_{i=1}^{M} \phi_i \, \mathbb{E}[z_i]\right] + \frac{1}{2}\left(\sum_{i=1}^{M} \mathbb{V}\mathrm{ar}[z_i]\, \mathbf{D}^2_{x_0}\mathrm{R}[\phi_i] + \mathbf{D}^2_{x_0}\mathrm{R}\left[\sum_{i=1}^{M} \mathbb{E}[z_i]\phi_i\right]\right) + \mathscr{O}(\|\xi\|^3_{L^3_{\mathbb{P}}(\Omega;X)}) \,, \tag{IV.11a}$$

$$\begin{aligned}\mathbb{C}\mathrm{or}[\mathrm{R}] = {} & \mathrm{R}_{x_0} \otimes \mathrm{R}_{x_0} + \left(\mathbf{D}_{x_0}\mathrm{R}\left[\sum_{i=1}^{M} \mathbb{E}[z_i]\phi_i\right] \otimes \mathrm{R}_{x_0} + \mathrm{R}_{x_0} \otimes \mathbf{D}_{x_0}\mathrm{R}\left[\sum_{i=1}^{M} \mathbb{E}[z_i]\phi_i\right]\right) \\ & + \frac{1}{2}\left(\sum_{i=i}^{M} \mathbb{V}\mathrm{ar}[z_i]\, \mathbf{D}^2_{x_0}\mathrm{R}\left[\phi_i\right] \otimes \mathrm{R}_{x_0} + \mathrm{R}_{x_0} \otimes \sum_{i=1}^{M} \mathbb{V}\mathrm{ar}[z_i]\, \mathbf{D}^2_{x_0}\mathrm{R}\left[\phi_i\right]\right. \\ & \left. + \mathbf{D}^2_{x_0}\mathrm{R}\left[\sum_{i=1}^{M} \phi_i \, \mathbb{E}[z_i]\right] \otimes \mathrm{R}_{x_0} + \mathrm{R}_{x_0} \otimes \mathbf{D}^2_{x_0}\mathrm{R}\left[\sum_{i=1}^{M} \phi_i \, \mathbb{E}[z_i]\right]\right) \\ & + \sum_{i=1}^{M} \mathbb{V}\mathrm{ar}[z_i]\, \mathbf{D}_{x_0}\mathrm{R}\left[\phi_i\right] \otimes \mathbf{D}_{x_0}\mathrm{R}\left[\phi_i\right] \\ & + \mathbf{D}_{x_0}\mathrm{R}\left[\sum_{i=1}^{M} \mathbb{E}[z_i]\phi_i\right] \otimes \mathbf{D}_{x_0}\mathrm{R}\left[\sum_{i=1}^{M} \mathbb{E}[z_i]\phi_i\right] + \mathscr{O}(\|\xi\|^3_{L^3_{\mathbb{P}}(\Omega;X)}) \,,\end{aligned} \tag{IV.11b}$$

$$\mathbb{C}\mathrm{ov}[\mathrm{R}] = \sum_{i=1}^{M} \mathbb{V}\mathrm{ar}[z_i]\, \mathbf{D}_{x_0}\mathrm{R}[\phi_i] \otimes \mathbf{D}_{x_0}\mathrm{R}[\phi_i] + \mathscr{O}(\|\xi\|^3_{L^3_{\mathbb{P}}(\Omega;X)}) \,. \tag{IV.11c}$$

As in corollary IV.3, the first-order terms vanish if we assume that ξ *is centered, i.e.,* $\mathbb{E}[z_i] = 0$. *Assuming* $\xi \in L^4_{\mathbb{P}}(\Omega;X)$ *is centered and also skewfree, i.e.,* $\mathbb{E}[z_i \otimes z_i \otimes z_i] = 0$, *the convergence rates improve to* $\mathscr{O}(\|\xi\|^4_{L^4_{\mathbb{P}}(\Omega;X)})$ *as in corollary IV.3.*

PROOF. We arrive at the proposed expressions by inserting the formulation of the derivatives found in (IV.10) into the approximations of theorem IV.2 and corollary IV.3 and using (II.39e) and the (bi-)linearity of the Fréchet derivatives. □

REMARK IV.8. We can consider multiple stochastically independent perturbations in analogy to the above decomposition, cf. [**25**].

IV.5. Numerical Experiments

We continue with the example of the perturbed Laplace EVP of section III.7. Previously, we considered a single sample of $x_\xi(\omega)$, i.e., a deterministic problem. Now we keep the random field stochastic and determine the stochastic moments of the eigenpair trajectories with respect to the eigenspace. The perturbation is again given as the same truncated KLE so the explicit formulas of corollary IV.7 can be used.
There is no unique polarization $\boldsymbol{P}_{x_0}$ for the eigenfunctions of the degenerate eigenspace of the second and third eigenvalue (counting individually), since the eigenpairs

that emerge from the degenerate eigenspace are coupled. Thus, considering their trajectories with respect to the eigenspace, we are free to choose any basis. We align them to the coordinate axes as in fig. II.1.

IV.5.1. Parallelization of Calculations. As the coefficients are linear, we can save some computational effort in building FE matrices by calculating the derivatives of the stiffness and mass matrix

$$\mathbf{D}_{x_0} \underline{K}[\phi_i] \,, \qquad \mathbf{D}_{x_0} \underline{M}[\phi_i] \tag{IV.12}$$

according to (IV.9), which can be done in parallel. We add a more explicit scaling parameter $t \geq 0$ to the perturbation model (IV.9), i.e.,

$$x_\xi(t) = x_0 + t\xi = x_0 + t \sum_{i=1}^{M} \phi_i z_i. \tag{IV.13}$$

In each sample, we can then build the stiffness and mass matrix as

$$\underline{K}_{x_\xi} = \underline{K}_{x_0} + t \sum_{i=1}^{M} \mathbf{D}_{x_0} \underline{K}[\phi_i] z_i \,, \qquad \underline{M}_{x_\xi} = \underline{M}_{x_0} + t \sum_{i=1}^{M} \mathbf{D}_{x_0} \underline{M}[\phi_i] z_i \,,$$

which is faster than building them from the ground up in every sample of z. We use uniform distributions for z_i, so that we can bound the stochastic perturbation, which allows us to guarantee ellipticity of $\underline{K}_{x_\xi}$ and $\underline{M}_{x_\xi}$ for sufficiently small $t \geq 0$.
The prepared derivatives (IV.12) can also be used to calculate the derivatives of the eigenpair trajectories with respect to the eigenspace for the formulas according to corollary IV.7. These calculations can be performed independently and in parallel with respect to index $i = 1, \ldots, M$.
The parallel computations were performed by the Marvin cluster of the University of Bonn. Each task was executed on a single node with two Intel Xeon Platinum 8468 with forty-eight 2.1 GHz cores, hyper-threading enabled, and 1024 GB RAM.

IV.5.2. Uncertainty Quantification with Respect to the Eigenspace. Since the perturbation we consider is centered, we can use corollary VI.5 to approximate the variance of the eigenfunction trajectories with respect to the eigenspace. The second-order approximations can be seen in fig. IV.1 for $t = 1$. As we have aligned the unperturbed basis to the coordinate axes, the variances with respect to the eigenspace are aligned as well. We can see that eigenfunctions belonging to the eigenspace of the next smaller or larger eigenvalue have the largest impact on the shape of the variance, cf. fig. II.1 This can be explained by the influence of the gap of the eigenvalues on the derivatives and by the fact that due to our random field, all eigenvalues are coupled.

IV.5.3. Verification of Convergence Rates of the Approximations. To confirm the order of convergence of our approximations, we calculate all approximations of theorem VI.3 for the eigenpair trajectories with respect to the eigenspace. In addition to

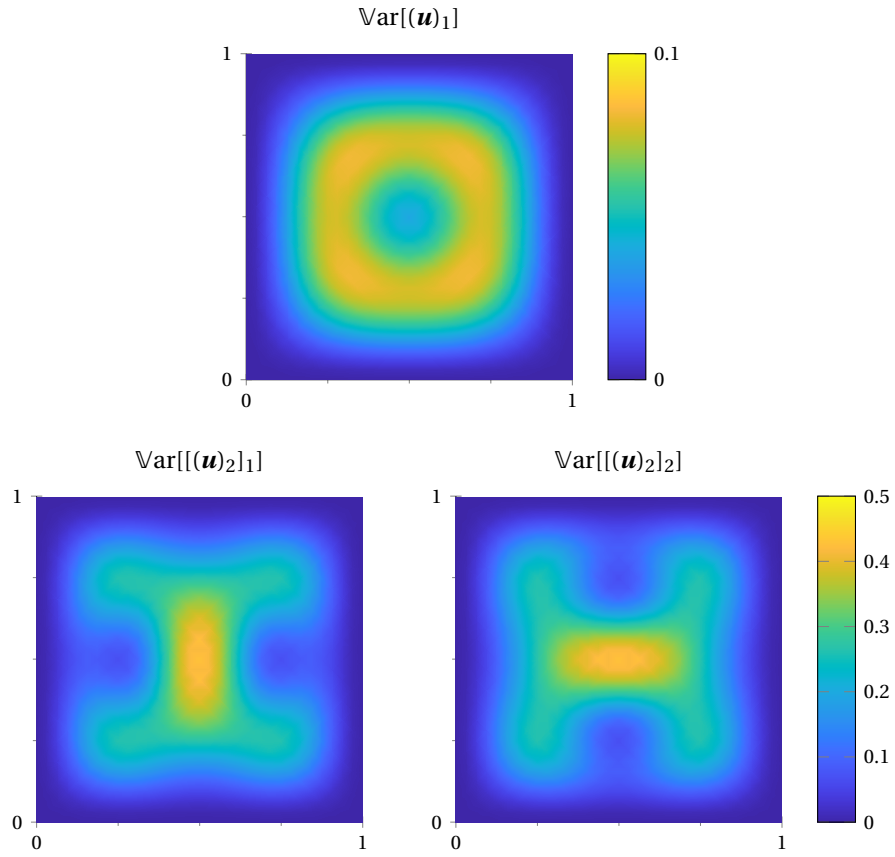

FIGURE IV.1. Second-order approximations of the variance of the eigenfunctions $\mathbb{V}\mathrm{ar}[(\boldsymbol{u})_i], i = 1,2$ with respect to the eigenspace for the two smallest eigenvalues with multiplicities $m_1 = 1$ and $m_2 = 2$ at $t = 1$.

the second-order approximations (theorem VI.3), we also test the first-order approximations (IV.6).

Benchmarking the Order or Convergence. Given a perturbation model (IV.13), the second-order perturbation approximations converge as $\mathcal{O}(t^3)$ and if z_i are centered and skewless the rate of convergence improves to $\mathcal{O}(t^4)$.

Since we use uniformly distributed random variables z_i, we can treat the mean as an integration over a hypercube with the amplitude of the perturbation $t \geq 0$ scaling the sides of the hypercube uniformly. The (tensorized) GL approximations with

$n = 2$ nodes in each dimension (table II.1) approximate polynomials of degree three exactly, cf. section II.7.7. Due to our perturbation model of the parameter (IV.13), the third-order perturbation approximations are third-degree polynomials in t, so the GL approximation with $n = 2$ points per dimension has a convergence rate of $\mathcal{O}(t^4)$ as $t \to 0$. The difference between two approximations that converge in $\mathcal{O}(t^k)$, $k \in \mathbb{N}$, also converges in $\mathcal{O}(t^k)$. Note that for large perturbation amplitudes $t \geq 0$, the GL approximation is likely quite inaccurate. However, $n = 2$ points per dimension are sufficient to validate the convergence rate of our perturbation approximations.
Using GL approximations, we need $n^{\dim(X)}$ function evaluations, which becomes untenable for $\dim(X) > \log_n(10^7)$. Therefore, for higher-dimensional parameter spaces, we compute QMC estimators using the Halton sequence of fixed length 10^7, cf. section II.7.7. Unfortunately, the QMC approximations are less accurate for very small perturbation amplitudes $t \geq 0$.
In the following, we present the computational results of the full model of chapter III with $\dim(X) = 114$ dimensions compared to QMC approximations. Additionally, we present the results compared to a GL quadrature, based on a premature truncation of the KLE after $M = 4$ terms, i.e., $\dim(X) = 8$ for both random fields combined. As discussed in chapter III, selecting such a small subset of the parameter space may lead to different coupling behavior of the eigenvalues. However, in this case, $M = 4$ is high enough for the degenerate eigenvalues to be coupled in the same way as they are for the high-dimensional random fields.
For each evaluation of the model with parameter configuration x_ξ, either according to the QMC sequence or the nodes of the GL approximation, we use an EVP-solver to solve the perturbed EVP directly. We then align the sampled eigenpairs with the (polarized) Taylor approximations according to the implied path $\gamma(t) = x_0 + t\xi$. Then we apply the inverse polarization as described section III.4 to approximate the evaluation of the trajectory with respect to the eigenspace. Each parameter configuration is computed in parallel on a single node of the cluster.
We compare the results of the perturbation approximation and the QMC approximations by calculating the Euclidean (matrix) norm $\|\cdot\|_{\mathbb{R}^{m\times m}}$ for the mean of the eigenvalue trajectory matrices and the $\|\cdot\|_{[L^2(\mathscr{D})]^m}$-norm for the mean of the eigenfunctions according to the multiplicity m of the eigenspace. For the correlation and covariance, we consider the norms of definition II.10. To this end, the eigenvalue matrices are vectorized first.
Uncentered Perturbation. For the uncentered case, we consider random variables $z_i \sim \mathcal{U}([0,1])$ iid with different amplitudes of perturbation t according to

$$t \in \left\{ t^n \,:\, n = -15, -14, \ldots, 0 \right\} .$$

The residues calculated from the difference of the perturbation approximations compared to GL approximations are given in fig. IV.2 and the residues compared to QMC estimates are illustrated in fig. IV.3. We can see that for large to intermediate per-

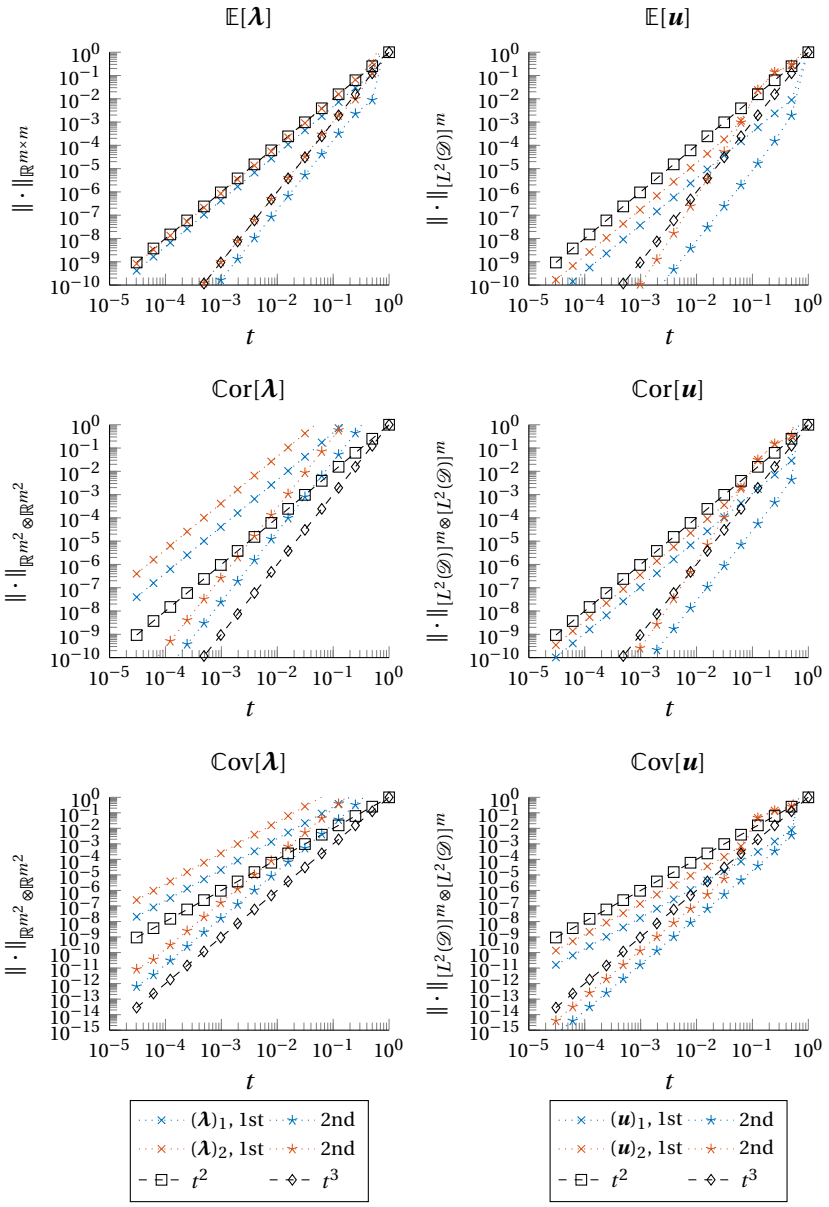

FIGURE IV.2. Convergence rates of uncentered approximations compared to GL estimate for the Laplace EVP with coefficients perturbed by random fields.

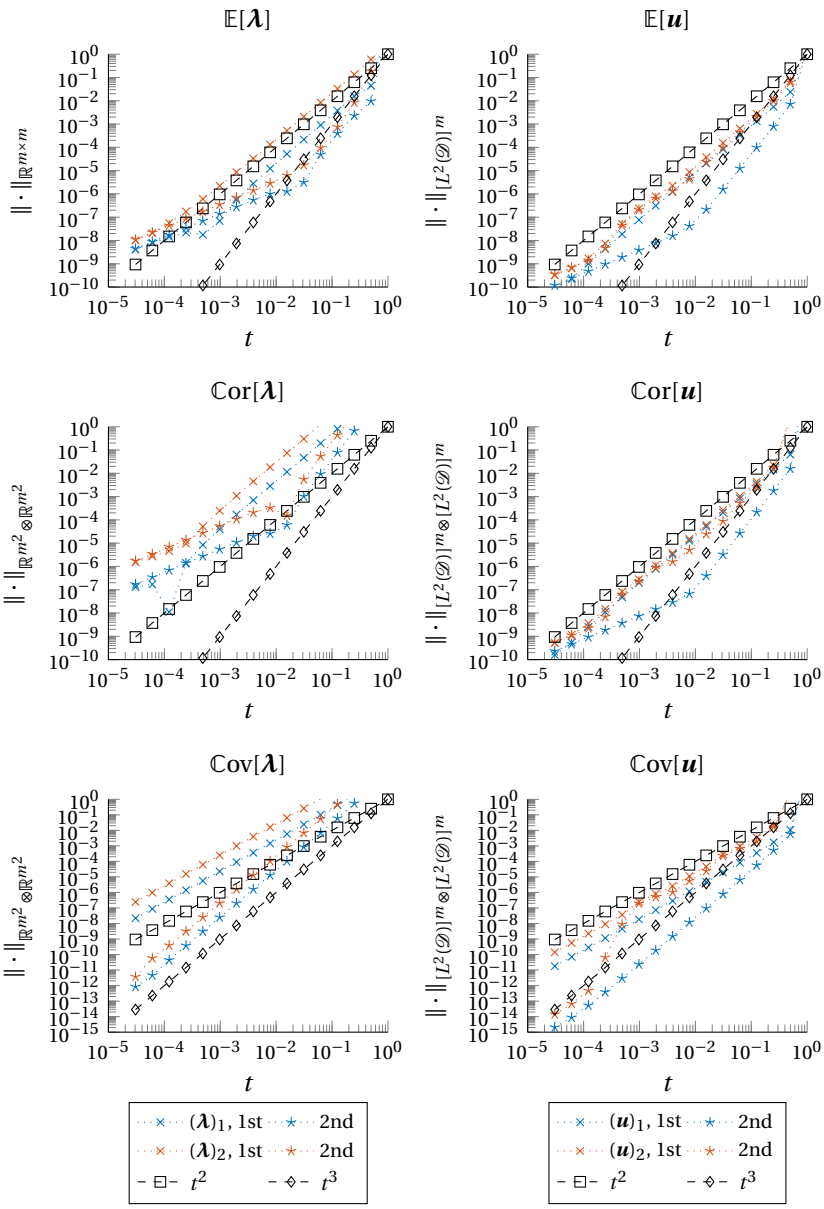

FIGURE IV.3. Convergence rates of uncentered approximations compared to QMC estimate for the Laplace EVP with coefficients perturbed by random fields.

turbation amplitudes the predicted rate of convergence can be confirmed. For the smallest perturbations the QMC approximations are less accurate than the perturbation approximations. Therefore, we can only observe the error decay of the QMC approximation with respect to the perturbation amplitude t, which is linear for the mean and correlation and quadratic for the covariance. The QMC method struggles to confirm the order of convergence for the eigenfunction with respect to the degenerate eigenspace. Meanwhile, the results compared to the GL quadrature suffer from numerical artifacts for small perturbation amplitudes, likely caused by the machine precision constraints during the transformation to the eigenspace and the accuracy of the eigenvalue solver. These artifacts, the limited precision of the QMC approximations given the fixed number of samples, and the abstract neighborhood $B(x_0)$ possibly being small may distort the observed orders of convergence.
Calculating the derivatives needed for the second-order approximations takes less than a second, while sampling and evaluation of 10^7 samples takes roughly 45 minutes using the setup discussed in section IV.5.1. For small approximations and high-dimensional parameter spaces, our perturbation approximation is thus clearly advantageous.
Centered and Skewfree Perturbation. For the centered and skewfree case, we repeat the experiment with random variables $z_i \sim \mathcal{U}([-\frac{1}{2}, \frac{1}{2}])$ iid. According to corollary IV.3 the convergence rate for the second-order perturbation approximations should improve by one order of convergence. This can be confirmed in figs. IV.4 and IV.5 with similar caveats as in the previous experiments.

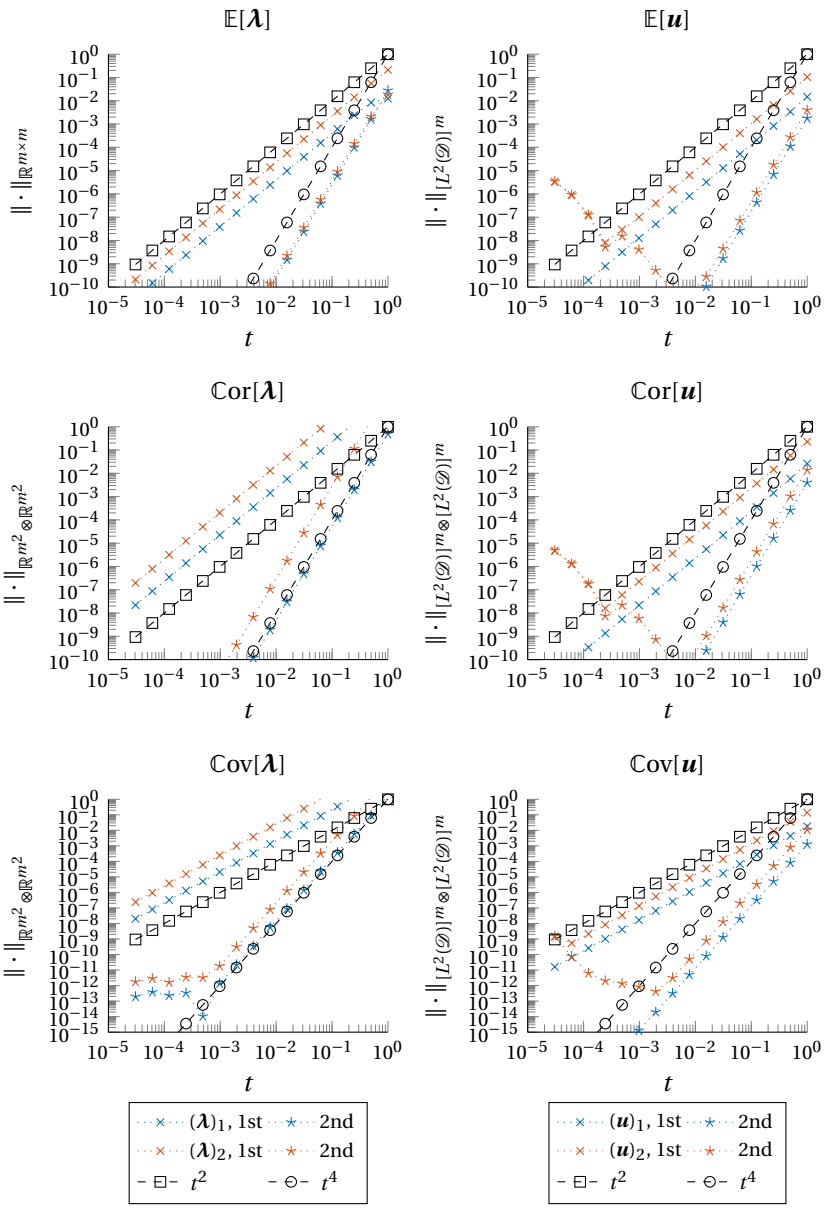

FIGURE IV.4. Convergence rates of centered approximations compared to GL estimate for the Laplace EVP with coefficients perturbed by random fields.

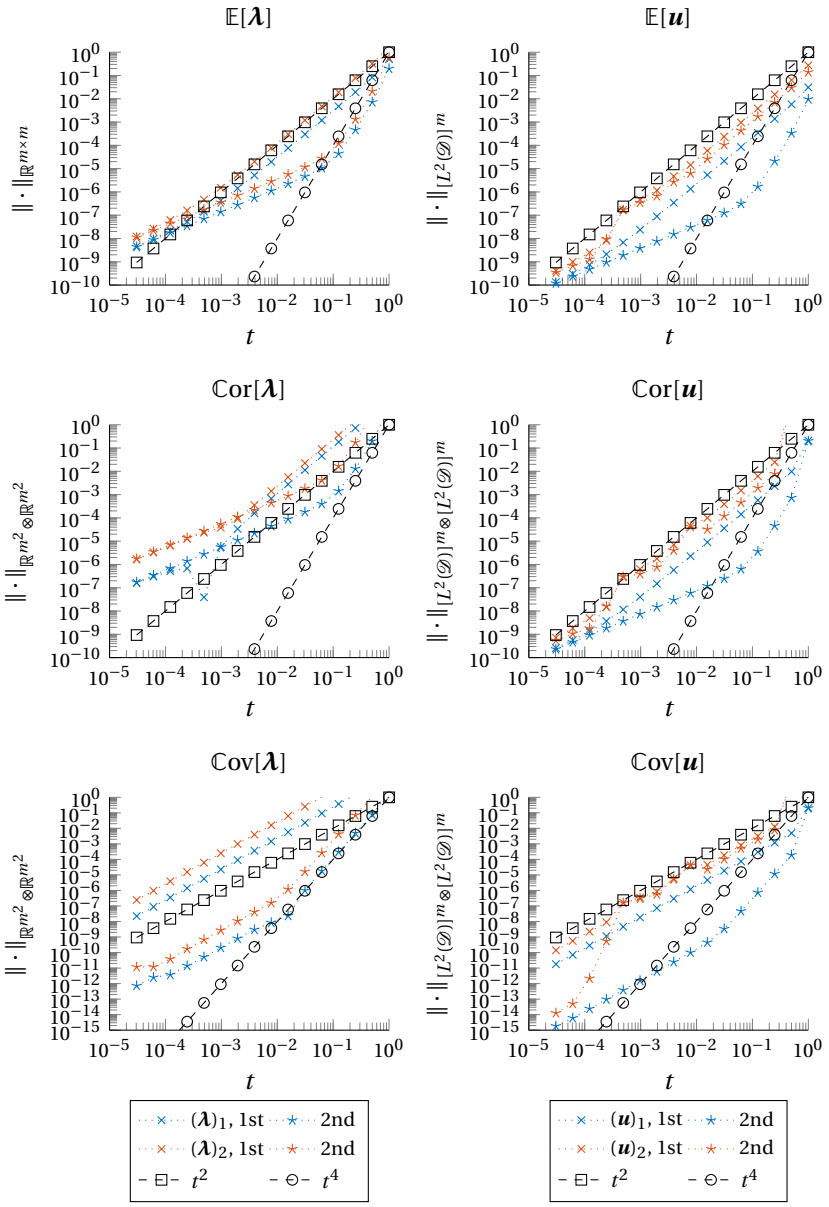

FIGURE IV.5. Convergence rates of centered approximations compared to QMC estimate for the Laplace EVP with coefficients perturbed by random fields.

CHAPTER V

Shape Uncertainty Quantification

In this chapter, we consider how shape deformations can be investigated using the perturbation approximations of chapter IV. We focus on shape deformations for the Maxwell and Laplace EVP that have previously been discussed. To this end, we recall how smooth deformations of an unperturbed reference domain can be represented by introducing matrix-valued deformation coefficients.
The results presented in this chapter expand on the article [**27**], where the first-order approximations of the eigenpairs with respect to the eigenspace, cf. chapters III and IV or [**25**], were combined with first-order derivatives of the stiffness and mass matrix, cf. [**108**]. The resulting first-order approximation was used to evaluate a shape uncertainty model for a TESLA cavity based on deformation data of cavities measured at DESY [**33, 104**].
The results of [**27**] are expanded here to second-order approximations as in chapter IV. This additionally requires the second-order derivatives of the stiffness and mass matrix of [**109**]. We also include the deformation coefficients for the Laplace EVP of [**108**]. In turn, we deemphasize TESLA cavities, to avoid an extensive discussion of the underlying technology. The interested reader is referred to [**6, 27**].
Nevertheless, an example demonstrating shape uncertainty for a TESLA cavity is included at the end of this chapter to demonstrate the efficacy of the perturbation approximations for variational Maxwell EVPs with many degrees of freedom. To confirm the validity of our approximations for trajectories with respect to a degenerate eigenspace, we also consider a deformation of the a Laplace EVP seen in previous chapters, as well as a deformation for eigenpairs of the Maxwell EVP with multiplicity $m = 3$ on a cube domain.

V.1. Representation of Shape Deformations as Deformation Coefficients

We start by stating the model assumptions, expanding on the assumptions of chapters III and IV.

ASSUMPTION V.1. We consider domains in $\mathbb{R}^n$, $n = 2,3$, that are open, bounded, simply connected, and Lipschitz and consider a deformation model defined by a mapping

$$G \in C^3(X; C^1(\mathscr{D}_{x_0}; \mathbb{R}^n)) , \qquad x \mapsto G_x \in C^1(\mathscr{D}_{x_0}; \mathbb{R}^n) . \tag{V.1}$$

The mapping

$$G_x : \mathscr{D}_{x_0} \to \mathbb{R}^n , \qquad \mathbf{x} \mapsto G_x(\mathbf{x})$$

maps the reference domain $\mathscr{D}_{x_0} \subset \mathbb{R}^n$ to the deformed domain

$$\mathscr{D}_x := G_x(\mathscr{D}_{x_0})$$

such that $\mathscr{D}_x$ satisfies the above properties that we assumed of the domain. The reference point $x_0 \in X$ relates to the reference domain $\mathscr{D}_{x_0}$, i.e.,

$$G_{x_0} = \mathrm{Id} .$$

As in assumption IV.1, for some results, we replace the C^3-mapping by a C^4-mapping. We assume a stochastic perturbation model of the form (IV.1), which makes the eigenpair trajectories with respect with to the eigenspace locally L^3- or L^4-integrable as discussed in assumption IV.1.

To prepare for the transformation rules used in this chapter, we recall the Jacobi matrix.

DEFINITION V.2. The **Jacobi matrix** is the matrix of partial derivatives

$$\mathrm{J}\, G_x := \left[\frac{\partial [G_x]_i(\mathbf{x})}{\partial \mathbf{x}_j} \right]_{i,j=1}^{n} .$$

We assume that the Jacobi matrix is invertible, which is sensible for some neighborhood $B(x_0)$ as $\mathrm{J}\, G_{x_0} = \boldsymbol{I}$. Using the Jacobi matrix, we can formulate the following well-known transformation rules, cf. [**67, 108**].

LEMMA V.3 ([**108**, Lemma 1]). *Consider a deformation according to assumption V.1.*

(1) *Let $u \in H^1(\mathscr{D}_x)$, then it holds*

$$(\operatorname{grad} u) \circ G_x = (\mathrm{J}\, G_x)^{-\top} \cdot \operatorname{grad}(u \circ G_x) ,$$

so $u \circ G_x \in H^1(\mathscr{D}_x)$.

(2) *Consider domains in $\mathbb{R}^3$ and let $u \in H(\mathbf{curl}; \mathscr{D}_x)$, then it holds*

$$(\mathbf{curl}\, u) \circ G_x = \frac{1}{\det(\mathrm{J}\, G_x)} (\mathrm{J}\, G_x) \cdot \mathbf{curl}((\mathrm{J}\, G_x)^{\top} \cdot (u \circ G_x)) .$$

Thus, $u \in H(\mathbf{curl}; \mathscr{D}_x)$ if and only if $(\mathrm{J}\, G_x)^{\top} \cdot u \circ G_x \in H(\mathbf{curl}; \mathscr{D}_{x_0})$.

V.1.1. Deformation coefficients. The Laplace and Maxwell EVP on the deformed domain can be formulated as a variational EVP over the perturbed domain $\mathscr{D}_x$. However, in order to use perturbation approximations, it is more convenient to first express them as an integral over the reference domain $\mathscr{D}_{x_0}$. In this formulation, the deformation G_x is represented by a **deformation coefficient** included in each bilinear form. We first consider the simpler case of the Laplace EVP.

LEMMA V.4 ([**108**, Lemma 2]). *Consider the case of assumption V.1 with domains in* $\mathbb{R}^2$. *Then the bilinear forms of the Laplace EVP (examples II.77 and II.81) on the deformed domain* $\mathscr{D}_x$ *are given by*

$$a(u,v;x) := \int_{\mathscr{D}_{x_0}} \left\langle \mathfrak{G}_x(\mathbf{x}) \cdot \operatorname{grad} u(\mathbf{x}), \operatorname{grad} v(\mathbf{x}) \right\rangle_{\mathbb{R}^2} \mathrm{d}\mathbf{x}\,, \tag{V.2a}$$

$$b(u,v;x) := \int_{\mathscr{D}_{x_0}} \mathfrak{s}_x(\mathbf{x})\, u(\mathbf{x})\, v(\mathbf{x})\, \mathrm{d}\mathbf{x} \tag{V.2b}$$

with matrix-valued deformation coefficient $\mathfrak{G}_x(\mathbf{x}) \in \mathbb{R}^{2\times 2}$ *and scalar deformation coefficient* $\mathfrak{s}_x(\mathbf{x}) \in \mathbb{R}$ *given by*

$$\mathfrak{G}_x(\mathbf{x}) := \det(\mathrm{J}\, G_x(\mathbf{x}))\, (\mathrm{J}\, G_x(\mathbf{x}))^{-1} \cdot (\mathrm{J}\, G_x(\mathbf{x}))^{-\top}\,, \tag{V.3a}$$

$$\mathfrak{s}_x(\mathbf{x}) := \det(\mathrm{J}\, G_x(\mathbf{x}))\,. \tag{V.3b}$$

Now consider the case of the Maxwell EVP.

LEMMA V.5 ([**108**, Lemma 3]). *Consider the case of assumption V.1 with domains in* $\mathbb{R}^3$. *Then the bilinear forms of the Maxwell EVP (example II.78) on the domain* $\mathscr{D}_x$ *are given by*

$$a(u,v;x) := \int_{\mathscr{D}_{x_0}} \left\langle \mu^{-1}(\mathbf{x})\, \mathfrak{C}_x(\mathbf{x}) \cdot \mathbf{curl}(u(\mathbf{x})), \mathbf{curl}(v(\mathbf{x})) \right\rangle_{\mathbb{R}^3} \mathrm{d}\mathbf{x}\,, \tag{V.4a}$$

$$b(u,v;x) := \int_{\mathscr{D}_{x_0}} \langle \varepsilon(\mathbf{x})\, \mathfrak{G}_x(\mathbf{x}) \cdot u(\mathbf{x}), v(\mathbf{x}) \rangle_{\mathbb{R}^3} \mathrm{d}\mathbf{x} \tag{V.4b}$$

with matrix-valued coefficients $\mathfrak{C}_x(\mathbf{x}), \mathfrak{G}_x(\mathbf{x}) \in \mathbb{R}^{3\times 3}$

$$\mathfrak{C}_x(\mathbf{x}) := \frac{(\mathrm{J}\, G_x(\mathbf{x}))^{\top} \cdot \mathrm{J}\, G_x(\mathbf{x})}{\det(\mathrm{J}\, G_x(\mathbf{x}))}\,, \tag{V.5a}$$

$$\mathfrak{G}_x(\mathbf{x}) := \det(\mathrm{J}\, G_x(\mathbf{x}))\, (\mathrm{J}\, G_x(\mathbf{x}))^{-1} \cdot (\mathrm{J}\, G_x(\mathbf{x}))^{-\top}\,. \tag{V.5b}$$

REMARK V.6. Note that the coefficients (V.3b) and (V.5a) are identical in terms of the Jacobi matrix. Thus, we denote them with the same symbol, although in the case of the Laplace EVP, it refers to a $\mathbb{R}^{2\times 2}$-matrix, while for the Maxwell EVP, it refers to a $\mathbb{R}^{3\times 3}$-matrix. We omit the argument of the domain variable $\mathbf{x} \in \mathscr{D}$ for improved readability, as it is only relevant in the integrals (V.2) and (V.4).

V.1.2. Derivatives of the Deformation Coefficients. The derivatives of the bilinear forms are determined by inserting the derivatives of the coefficients (V.2) and (V.4) instead of the respective coefficients, i.e., for the Laplace EVP,

$$\mathbf{D}_{x_0}^k a\big(u,v;\cdot\big) = \int_{\mathscr{D}_{x_0}} \Big\langle (\mathbf{D}_{x_0}^k \mathfrak{G}(\mathbf{x})) \cdot \operatorname{grad} u(\mathbf{x}), \operatorname{grad} v(\mathbf{x}) \Big\rangle_{\mathbb{R}^2} \,\mathrm{d}\mathbf{x}\,, \qquad k \in \mathbb{N}\,,$$

$$\mathbf{D}_{x_0}^k b\big(u,v;\cdot\big) = \int_{\mathscr{D}_{x_0}} (\mathbf{D}_{x_0}^k \mathfrak{s}(\mathbf{x}))\, u(\mathbf{x})\, v(\mathbf{x}) \,\mathrm{d}\mathbf{x}\,, \qquad k \in \mathbb{N}\,,$$

and for the Maxwell EVP,

$$\mathbf{D}_{x_0}^k a\big(u,v;\cdot\big) = \int_{\mathscr{D}_{x_0}} \Big\langle \mu^{-1}(\mathbf{x})\, (\mathbf{D}_{x_0}^k \mathfrak{C}(\mathbf{x})) \cdot \mathbf{curl}(u(\mathbf{x})), \mathbf{curl}(v(\mathbf{x})) \Big\rangle_{\mathbb{R}^3} \,\mathrm{d}\mathbf{x}\,, \qquad k \in \mathbb{N}\,,$$

$$\mathbf{D}_{x_0}^k b\big(u,v;\cdot\big) = \int_{\mathscr{D}_{x_0}} \Big\langle \varepsilon(\mathbf{x})\, (\mathbf{D}_{x_0}^k \mathfrak{G}(\mathbf{x})) \cdot \mathbf{curl}(u(\mathbf{x})), \mathbf{curl}(v(\mathbf{x})) \Big\rangle_{\mathbb{R}^3} \,\mathrm{d}\mathbf{x}\,, \qquad k \in \mathbb{N}\,.$$

Due to remark V.6, we refer to the derivatives of the deformation coefficients without distinguishing between the Laplace and Maxwell EVP.

LEMMA V.7 ([**109**, Chapter 4]). *Given assumption V.1, the first-order derivatives of the deformation coefficients of* (V.3) *and* (V.5) *are given by*

$$\text{(V.6a)} \quad \mathbf{D}_x \mathfrak{C} = -\operatorname{tr}\big((\mathbf{D}_x \mathrm{J}\, G)\cdot(\mathrm{J}\, G_x)^{-1}\big)\,\mathfrak{C}_x + \frac{(\mathbf{D}_x \mathrm{J}\, G)^\top \cdot (\mathrm{J}\, G_x)}{\det(\mathrm{J}\, G_x)} + \frac{\big((\mathbf{D}_x \mathrm{J}\, G)^\top \cdot (\mathrm{J}\, G_x)\big)^\top}{\det(\mathrm{J}\, G_x)}\,,$$

$$\text{(V.6b)} \quad \mathbf{D}_x \mathfrak{G} = \operatorname{tr}\big((\mathbf{D}_x \mathrm{J}\, G)\cdot(\mathrm{J}\, G_x)^{-1}\big)\,\mathfrak{G}_x - (\mathrm{J}\, G_x)^{-1}\cdot(\mathbf{D}_x \mathrm{J}\, G)\cdot\mathfrak{G}_x - \big((\mathrm{J}\, G_x)^{-1}\cdot(\mathbf{D}_x \mathrm{J}\, G)\cdot\mathfrak{G}_x\big)^\top\,,$$

$$\text{(V.6c)} \quad \mathbf{D}_x \mathfrak{s} = \operatorname{tr}\big((\mathbf{D}_x \mathrm{J}\, G)\cdot(\mathrm{J}\, G_x)^{-1}\big)\,\mathfrak{s}_x\,.$$

The second-order derivatives are given by

$$\begin{aligned}\text{(V.7a)} \quad \mathbf{D}_x^2 \mathfrak{C} = {} & -\operatorname{tr}\big((\mathbf{D}_x^2 \mathrm{J}\, G)\cdot(\mathrm{J}\, G_x)^{-1}\big)\,\mathfrak{C}_x + \frac{(\mathbf{D}_x^2 \mathrm{J}\, G)^\top \cdot (\mathrm{J}\, G_x)}{\det(\mathrm{J}\, G_x)} + \frac{\big((\mathbf{D}_x^2 \mathrm{J}\, G)^\top \cdot (\mathrm{J}\, G_x)\big)^\top}{\det(\mathrm{J}\, G)} \\ & + \operatorname{tr}\big(((\mathbf{D}_x \mathrm{J}\, G)\cdot(\mathrm{J}\, G_x)^{-1})^2\big)\,\mathfrak{C}_x - \operatorname{tr}\big((\mathbf{D}_x \mathrm{J}\, G)\cdot(\mathrm{J}\, G_x)^{-1}\big)\,(\mathbf{D}_x \mathfrak{C}) \\ & - \operatorname{tr}\big((\mathbf{D}_x \mathrm{J}\, G)\cdot(\mathrm{J}\, G_x)^{-1}\big)\,\frac{((\mathbf{D}_x \mathrm{J}\, G)\cdot(\mathrm{J}\, G_x)^{-1})^\top + ((\mathbf{D}_x \mathrm{J}\, G)\cdot(\mathrm{J}\, G_x)^{-1})}{\det(\mathrm{J}\, G_x)} \\ & + 2\,\frac{(\mathbf{D}_x \mathrm{J}\, G)^\top \cdot (\mathbf{D}_x \mathrm{J}\, G)}{\det(\mathrm{J}\, G_x)}\,,\end{aligned}$$

$$\begin{aligned}(\text{V.7b})\quad \mathbf{D}_x^2\,\mathfrak{G} &= \operatorname{tr}\left((\mathbf{D}_x^2\,\mathrm{J}\,G)\cdot(\mathrm{J}\,G_x)^{-1}\right)\mathfrak{G}_x - (\mathrm{J}\,G_x)^{-1}\cdot(\mathbf{D}_x^2\,\mathrm{J}\,G)\cdot\mathfrak{G}_x \\ &\quad - \left((\mathrm{J}\,G_x)^{-1}\cdot(\mathbf{D}_x^2\,\mathrm{J}\,G)\cdot\mathfrak{G}_x\right)^\top - \operatorname{tr}\left(((\mathbf{D}_x\,\mathrm{J}\,G)\cdot(\mathrm{J}\,G_x)^{-1})^2\right)\cdot\mathfrak{G}_x \\ &\quad + \operatorname{tr}\left((\mathbf{D}_x\,\mathrm{J}\,G)\cdot(\mathrm{J}\,G_x)^{-1}\right)\cdot(\mathbf{D}_x\,\mathfrak{G}) \\ &\quad + ((\mathrm{J}\,G_x)^{-1}\cdot(\mathbf{D}_x\,\mathrm{J}\,G))^2\cdot\mathfrak{G}_x - (\mathrm{J}\,G_x)^{-1}\cdot(\mathbf{D}_x\,\mathrm{J}\,G)\cdot(\mathbf{D}_x\,\mathfrak{G}) \\ &\quad + \left(((\mathrm{J}\,G_x)^{-1}\cdot(\mathbf{D}_x\,\mathrm{J}\,G))^2\cdot\mathfrak{G}_x - (\mathrm{J}\,G_x)^{-1}\cdot(\mathbf{D}_x\,\mathrm{J}\,G)\cdot(\mathbf{D}_x\,\mathfrak{G})\right)^\top, \\ (\text{V.7c})\quad \mathbf{D}_x^2\,\mathfrak{s} &= \operatorname{tr}\left((\mathbf{D}_x^2\,\mathrm{J}\,G)\cdot(\mathrm{J}\,G_x)^{-1}\right)\mathfrak{s}_x \\ &\quad + \operatorname{tr}\left((\mathbf{D}_x\,\mathrm{J}\,G)\cdot(\mathrm{J}\,G_x)^{-1}\right)(\mathbf{D}_x\,\mathfrak{s}) - \operatorname{tr}\left(((\mathbf{D}_x\,G)\cdot(\mathrm{J}\,G_x)^{-1})^2\right)\mathfrak{s}_x\,.\end{aligned}$$

PROOF. The calculation of the derivatives can be found in [**109**, Chapter 4] with the assumption $\mathbf{D}_x^2\,\mathrm{J}\,G = 0$. For the second-order derivatives (V.7), the slightly more general case where $\mathbf{D}_x^2\,\mathrm{J}\,G$ does not vanish adds some terms according to the product rule. □

Linear Deformation Model. A common choice for a deformation model is an affine model with displacement field

$$V[\xi]:\mathbb{R}^n\to\mathbb{R}^n\,,\qquad\qquad \xi\in X\,,$$

such that

$$(\text{V.8a})\qquad G_{x_\xi} = \mathbf{x} + V[\xi]\,,\qquad\qquad V[\xi] \coloneqq \mathbf{D}_{x_0}\,G[\xi]\,.$$

The Jacobi matrix and its derivative can then be simplified to

$$(\text{V.8b})\qquad \mathrm{J}\,G_{x_\xi} = \boldsymbol{I} + \mathrm{J}\,V[\xi]\,,\qquad\qquad \mathbf{D}_{x_0}\,\mathrm{J}\,G[\xi] = \mathrm{J}\,V[\xi]\,.$$

COROLLARY V.8. *Given the linear shape deformation model* (V.8) *the first-order derivatives of lemma V.7 simplify to*

$$\begin{aligned}(\text{V.9a})\qquad \mathbf{D}_{x_0}\,\mathfrak{C} &= -\operatorname{tr}(\mathrm{J}\,V)\cdot\boldsymbol{I} + (\mathrm{J}\,V)^\top + \mathrm{J}\,V\,, \\ (\text{V.9b})\qquad \mathbf{D}_{x_0}\,\mathfrak{G} &= \operatorname{tr}(\mathrm{J}\,V)\cdot\boldsymbol{I} - (\mathrm{J}\,V)^\top - \mathrm{J}\,V \qquad\qquad = -\mathbf{D}_{x_0}\,\mathfrak{C}\,, \\ (\text{V.9c})\qquad \mathbf{D}_{x_0}\,\mathfrak{s} &= \operatorname{tr}(\mathrm{J}\,V)\,,\end{aligned}$$

and the second-order derivatives simplify to

$$\begin{aligned}(\text{V.10a})\qquad \mathbf{D}_{x_0}^2\,\mathfrak{C} &= \operatorname{tr}\left((\mathrm{J}\,V)^2\right)\cdot\boldsymbol{I} - \operatorname{tr}(\mathrm{J}\,V)\cdot(\mathbf{D}_{x_0}\,\mathfrak{C}) - \operatorname{tr}(\mathrm{J}\,V)\left((\mathrm{J}\,V)^\top + \mathrm{J}\,V\right) \\ &\quad + 2\left((\mathrm{J}\,V)^\top\cdot(\mathrm{J}\,V)\right), \\ (\text{V.10b})\qquad \mathbf{D}_{x_0}^2\,\mathfrak{G} &= -\operatorname{tr}\left((\mathrm{J}\,V)^2\right)\cdot\boldsymbol{I} + \operatorname{tr}(\mathrm{J}\,V)\cdot(\mathbf{D}_{x_0}\,\mathfrak{G}) + (\mathrm{J}\,V)^2 - (\mathrm{J}\,V)\cdot(\mathbf{D}_{x_0}\,\mathfrak{G}) \\ &\quad + \left((\mathrm{J}\,V)^2 - (\mathrm{J}\,V)\cdot(\mathbf{D}_{x_0}\,\mathfrak{G})\right)^\top, \\ (\text{V.10c})\qquad \mathbf{D}_{x_0}^2\,\mathfrak{s} &= (\operatorname{tr}(\mathrm{J}\,V))^2 - \operatorname{tr}\left((\mathrm{J}\,V)^2\right).\end{aligned}$$

PROOF. Consider the derivatives of lemma V.7 and simplify using (V.8) such that

$$\mathbf{D}_{x_0}^2\,\mathrm{J}\,G = 0\,,\qquad \mathfrak{C}_{x_0} = \boldsymbol{I}\,,\qquad \mathfrak{G}_{x_0} = \boldsymbol{I}\,,\qquad \mathfrak{s}_{x_0} = 1\,.$$ □

The discussed derivatives of the deformation coefficient are sufficient to calculate the derivatives of the eigenpairs with respect to the eigenspaces up to the second order, i.e., for the perturbation theorems of chapter IV. Given sufficient smoothness of G, formulas for higher-order derivatives can also be found.

V.2. Stochastic Deformation Model

Now that the deformation coefficients (Laplace EVP (V.2), Maxwell EVP (V.4)) and their derivatives have been established, we again consider a linear stochastic model (V.8). That is, given a probability space $(\Omega, \mathscr{A}, \mathbb{P})$, we consider a perturbation model

$$G_{x_\xi}(\omega) = \mathbf{x} + V[\xi(\omega)] , \qquad \omega \in \Omega , \tag{V.11}$$

with deterministic reference point x_0 and V depending linearly on the random variable $\xi \in X$. As in chapter IV, we use assumption IV.1 which guarantees that the mean, correlation, and covariance of $\mathrm{J}\, G_x(\mathbf{x})$ exist locally and that the required moments exist locally for $x_\xi \in B(x_0)$.
Thus, we can apply the perturbation approximations of chapter IV.

COROLLARY V.9. *Let λ_{x_0} be an eigenvalue of multiplicity m of EVP* (III.2) *at $x_0 \in B(x_0)$ with $b(\cdot,\cdot;x_0)$-orthogonal eigenbasis $\boldsymbol{u}_{x_0} \in V^m$ and let assumptions IV.1 and V.1 hold. Consider the bilinear forms for the Laplace EVP given by bilinear forms* (V.2) *or the Maxwell EVP given by bilinear forms* (V.4). *Assume a linear perturbation of the reference domain according to* (V.11). *Let $(\boldsymbol{\lambda}, \boldsymbol{u})$ be the unique locally analytic trajectories with respect to the eigenspace of multiplicity m, cf. theorem III.2. Then the approximations of theorem IV.2 hold in some neighborhood $B(x_0)$ of x_0.*
If $\xi \in B(x_0)$ is centered, then the simplified approximations of corollary IV.3 hold. Similarly, if G is a C^4-mapping, ξ is centered and skewfree, i.e., $\mathbb{E}[\xi \otimes \xi \otimes \xi] = 0$, the improved convergence rate of corollary IV.3 holds.

PROOF. We apply theorem IV.2 to the bilinear forms (V.2) and (V.4). The deformation coefficients, which describe the deformation on the reference domain $\mathscr{D}_{x_0}$, inherit their regularity from the mapping G. Thus, it is also the regularity of the bilinear forms with derivatives characterized by the derivatives of the deformation coefficients.
Since these derivatives of order k are all k-linear bounded operators, simplifications of corollary IV.3 hold in complete analogy given the stronger set of assumptions. □

V.3. Implementation

In order to implement the perturbation approximations, the considerations of sections III.6 and IV.4 apply to the derived bilinear forms (V.2) and (V.4). In order to avoid repetition, we only provide some remarks on the implementation of the deformation coefficients and notes on KLE-type decompositions of the deformation model.

V.3.1. Implementation of the Deformation. For the implementation of the bilinear forms, we use a FE discretization with basis functions $(\varphi_i)_{i=1,\dots,n}$ on the undeformed domain. For the implementation of the matrix-valued basis coefficients we assume that depending on the EVP either $\operatorname{grad}(\varphi_i)$ (Laplace) or $\mathbf{curl}(\varphi_i)$ (Maxwell) are calculated explicitly in the existing implementation at nodes according to a quadrature rule. Then we apply the deformation coefficients to the basis functions in the same evaluation points to calculate the FE-matrices of the deformed domain.
Since such a this design pattern is found in in most FE, also in geoPDEs [**98**], which we use for the Maxwell EVP, we can implement the deformation coefficients with only minor modifications.
Note that applying the map G_x directly to the nodes of the grid leads to a different domain. The same is true when moving the coordinates of the NURBS coefficients if the domain is implemented in an IGA library.

V.3.2. Karhunen–Loève-type Stochastic Models. Let the deformation model (V.11) be given by a KLE-type decomposition

$$G_{x_\xi}(\omega) = \mathbf{x} + V[\xi(\omega)] = \mathbf{x} + \sum_{i=1}^{M} V[\phi_i] z_i(\omega)\,, \qquad \omega \in \Omega\,,$$

with $z_i \in L^4_{\mathbb{P}}(\Omega)$ iid and $M = \mathbb{N} \cup \{\infty\}$. Such a shape deformation model can be compiled using statistical methods and measurement data, cf. [**33**].
The decomposition of the shape deformation model implies that the deformation coefficients can similarly be decomposed as expansions of the form

$$\mathfrak{C}_{x_\xi} = \boldsymbol{I} + \sum_{i=1}^{M} \mathbf{D}_{x_0} \mathfrak{C}[\phi_i] z_i(\omega) + \frac{1}{2} \sum_{i,j=1}^{M} \mathbf{D}^2_{x_0} \mathfrak{C}[\phi_i, \phi_j] z_i(\omega) z_j(\omega) + \mathcal{O}(\|\xi\|^3_{L^3_{\mathbb{P}}(\Omega;X)})\,,$$

$$\mathfrak{G}_{x_\xi} = \boldsymbol{I} + \sum_{i=1}^{M} \mathbf{D}_{x_0} \mathfrak{G}[\phi_i] z_i(\omega) + \frac{1}{2} \sum_{i,j=1}^{M} \mathbf{D}^2_{x_0} \mathfrak{G}[\phi_i, \phi_j] z_i(\omega) z_j(\omega) + \mathcal{O}(\|\xi\|^3_{L^3_{\mathbb{P}}(\Omega;X)})\,,$$

$$\mathfrak{s}_{x_\xi} = 1 + \sum_{i=1}^{M} \mathbf{D}_{x_0} \mathfrak{s}[\phi_i] z_i(\omega) + \frac{1}{2} \sum_{i,j=1}^{M} \mathbf{D}^2_{x_0} \mathfrak{s}[\phi_i, \phi_j] z_i(\omega) z_j(\omega) + \mathcal{O}(\|\xi\|^3_{L^3_{\mathbb{P}}(\Omega;X)})\,.$$

Each Fréchet derivative of the deformation coefficients toward $\phi_i \in X$ yields respective Fréchet derivatives of the bilinear forms according to section V.1.2. By insertion of these into the bilinear forms and calculation of the derivatives of the eigenpair trajectories with respect to the eigenspace of chapter III, we arrive at the setting of section IV.4.2. Thus, we can use the explicit formulas of corollary IV.7 for corollary V.9.

V.4. Numerical Experiments

In this section, we consider deformations of the Laplace EVP of chapters III and IV and Maxwell EVPs in order to demonstrate the implementation and validate the convergence rates of corollary V.9.

In the following examples, the shape deformation models will be constructed using trigonometric functions. This can be seen as an idealized version of a deformation model, similar in shape to the model of [**27, 33**]. Using such idealized deformations has the benefit that the Jacobi matrix can be calculated in closed form. Additionally, we may observe crossing behavior in the trajectories that we might not see if the perturbations included statistical artifacts.
We again consider a model with low-dimensional parameter space in order to benchmark the convergence rates of the perturbation approximations against a GL approximation as in section IV.5. In the following example, we only consider the centered and skewless setting with improved convergence rates. Given a perturbation model of the forms (IV.13) with $t \geq 0$ the amplitude of the perturbation and z_i centered and skewless, the second-order perturbation approximation and (tensorized) GL approximation for $n = 2$ converge in $\mathcal{O}(t^4)$.

V.4.1. Laplace Eigenvalue Problem. Consider the Laplace EVP on the unit square with zero Dirichlet boundary condition using the discretization of section III.7. For the deformation mapping, we first consider the one-dimensional case

$$G : [0,1] \to C^\infty(\mathscr{D}_{t_0}; \mathscr{D}_t) , \qquad t \mapsto G_t = \mathbf{x} + t \begin{bmatrix} 0 \\ \frac{1}{10}\sin(k\pi \mathbf{x}_1) \end{bmatrix} , \qquad k = 2 . \tag{V.13}$$

As seen in previous examples of chapters III and IV, the first eigenvalue is non-degenerate while the second and third form a degenerate eigenspace of multiplicity $m = 2$. These eigenpairs are illustrated for $t = 0$ and $t = 1$ in fig. V.1. The degenerate unperturbed eigenfunctions are polarized according to the deformation. In this example, the polarization is given by deciding order $\grave{k} = 1$. Note that if we change (V.13) so that k is odd, we instead get $\grave{k} = 2$ even, i.e., pathwise deflecting instead of crossing trajectories.
Verification of the Derivatives. Before considering the stochastic case, we verify the derivatives of the FE matrices and the derivatives of the polarized eigenpairs. To this end, we approximate each by first- and second-order Taylor approximations. In the case of the eigenfunctions of the degenerate eigenvalue, we restrict ourselves to first-order approximations due to $\grave{k} = 1$. We calculate each at perturbation amplitudes

$$t \in \left\{ t^n \,:\, n = -15, \ldots, 0 \right\} . \tag{V.14}$$

We again use the absolute value for the eigenvalues and the $\|\cdot\|_{L^2(\mathscr{D})}$-norm for the eigenfunctions. For the FE matrices we use the $\|\cdot\|_1$-norm, i.e., the maximum absolute column sum, which can be evaluated faster for sparse matrices. The resulting convergence rates can be seen in fig. V.2. Note that for this deformation, the mass matrix is constant and the stiffness matrix is exactly approximated by the second-order Taylor approximation. This can be confirmed empirically in the fig. V.2 up to machine precision and is convenient for the following discussion, since we did not discuss the formulas for higher-order derivatives of the deformation coefficients.

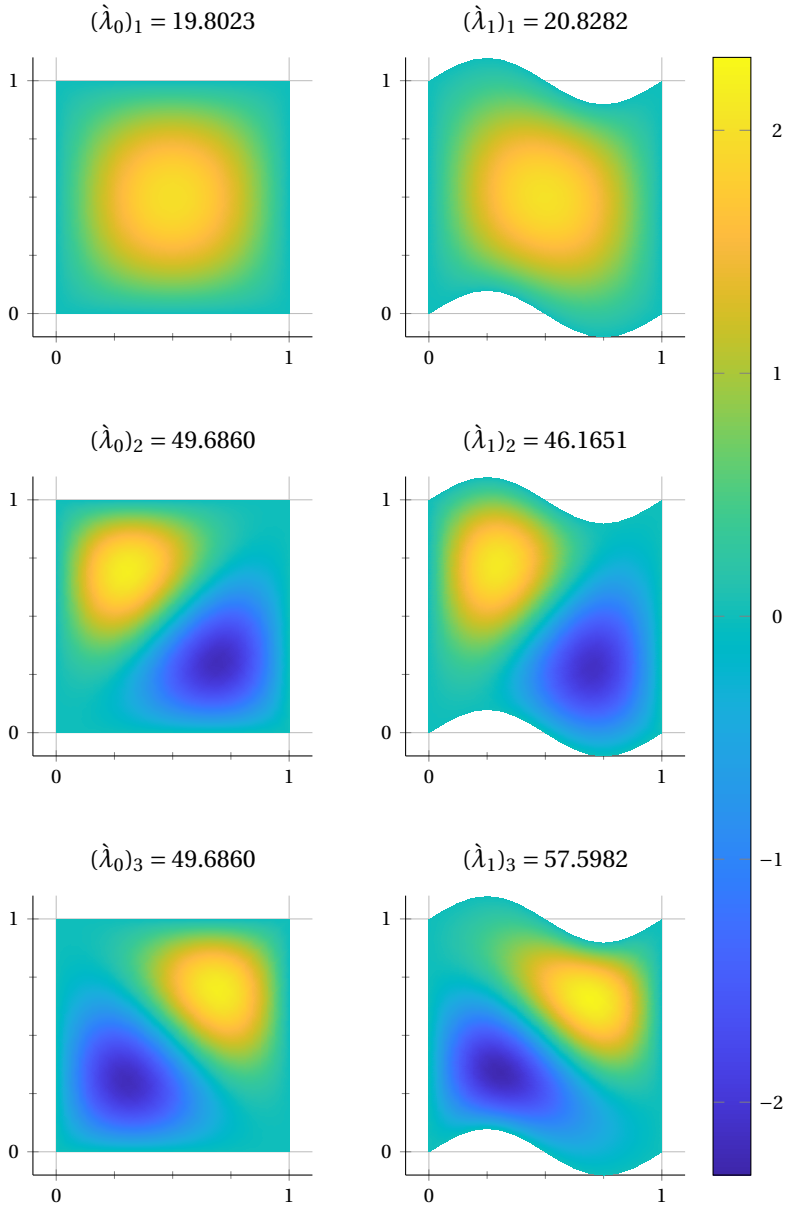

FIGURE V.1. Unperturbed and perturbed Laplace eigenpairs according to deformation (V.13).

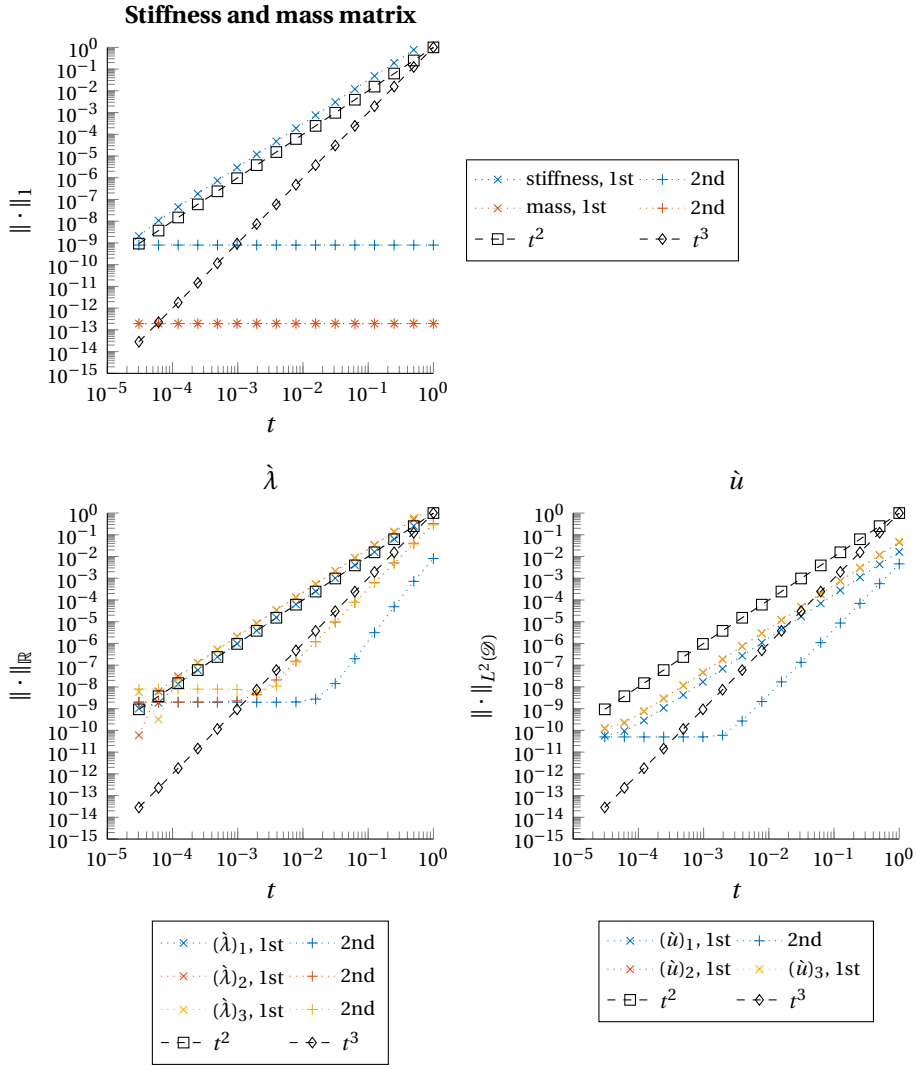

FIGURE V.2. Convergence of the Taylor approximations of the FE matrices and polarized eigenpairs for the Laplace EVP example.

Uncertainty Quantification. We now consider a stochastic model of the form

$$G_{x_\xi} = \mathbf{x} + V[\xi] = \mathbf{x} + t \sum_{k=1}^{M} V[\phi_k] z_i \,,$$

where we choose

$$V[\phi_k] = \begin{bmatrix} 0 \\ \exp(-k)\sin(k\pi\mathbf{x}_1) \end{bmatrix}$$

with $M = 5$ and $z_i = \mathscr{U}([-\frac{1}{2}, \frac{1}{2}])$ iid. Since we cannot expect a common polarization for all samples, we choose the basis of the degenerate eigenspace by aligning them to the coordinate axis as in fig. II.1.
Clearly, the deformation is centered and skewfree, so we expect convergence rates of $\mathscr{O}(t^4)$ according to corollary V.9. The approximations can be implemented using the formulas of corollary IV.7 and parallelization as in section IV.5.
Since the deformation model is not very high-dimensional, we use the GL quadrature with $n = 2$ points per dimension, cf. table II.1. For each evaluation, we calculate the third-order approximation of the polarization matrix as $t \mapsto [0,1]$ to calculate the evaluation of the trajectories with respect to the eigenspace as discussed in section III.4. To this end, we use the favorable circumstance that the derivatives of the FE-matrices vanish for orders higher than two. Since we did not proof this conjecture analytically, we check it up to machine precision for each evaluation. We calculate the approximations and the GL quadrature at perturbation amplitudes (V.14). The convergence rates are then calculated with the same norms as in section IV.5.
The convergence rates are illustrated and confirmed in fig. V.3. For the eigenfunctions in the degenerate eigenspace, the mapping to the eigenspace introduces some numerical errors to the GL approximation for small perturbations.

V.4.2. Maxwell Eigenvalue Problem. We now consider the Maxwell EVP on the unit cube $\mathscr{D} = (0,1)^3$, which is provided as a one-patch IGA example in geoPDEs [**98**]. For the basis, we use second-degree B-splines, let the library perform two subdivisions of the patch, and use the included GL quadrature with $n = 3$ points per dimension. We consider the smallest eigenvalue with multiplicity $m = 3$, which is approximated as $\lambda = 20$.
Similarly to the previous example, we first consider a one-dimensional parameter space and a deformation

$$G : [0,1] \to C^\infty(\mathscr{D}_{x_0}; \mathscr{D}_t)\,, \qquad t \mapsto G_t = \mathbf{x} + t \begin{bmatrix} 0 \\ \frac{1}{10}\sin(k\pi\mathbf{x}_1) \\ 0 \end{bmatrix}\,, \qquad k = 2\,. \tag{V.15}$$

The eigenpairs for $t = 0$ and $t = 1$ can be seen in fig. V.4. The unperturbed eigenfunctions belong to an eigenspace with multiplicity $m = 3$ and are polarized with respect to the deformation. Note that since we calculated the vector field on the reference domain, we need to reverse the transformation of lemma V.3 in order to arrive at the vector field in $H_0^1(\mathbf{curl}; \mathscr{D}_x)$.
The decision matrix in this case has uniform order $\grave{k} = 1$. Therefore, the eigenvalues are sorted according to their first-order derivatives of the eigenvalues. The second

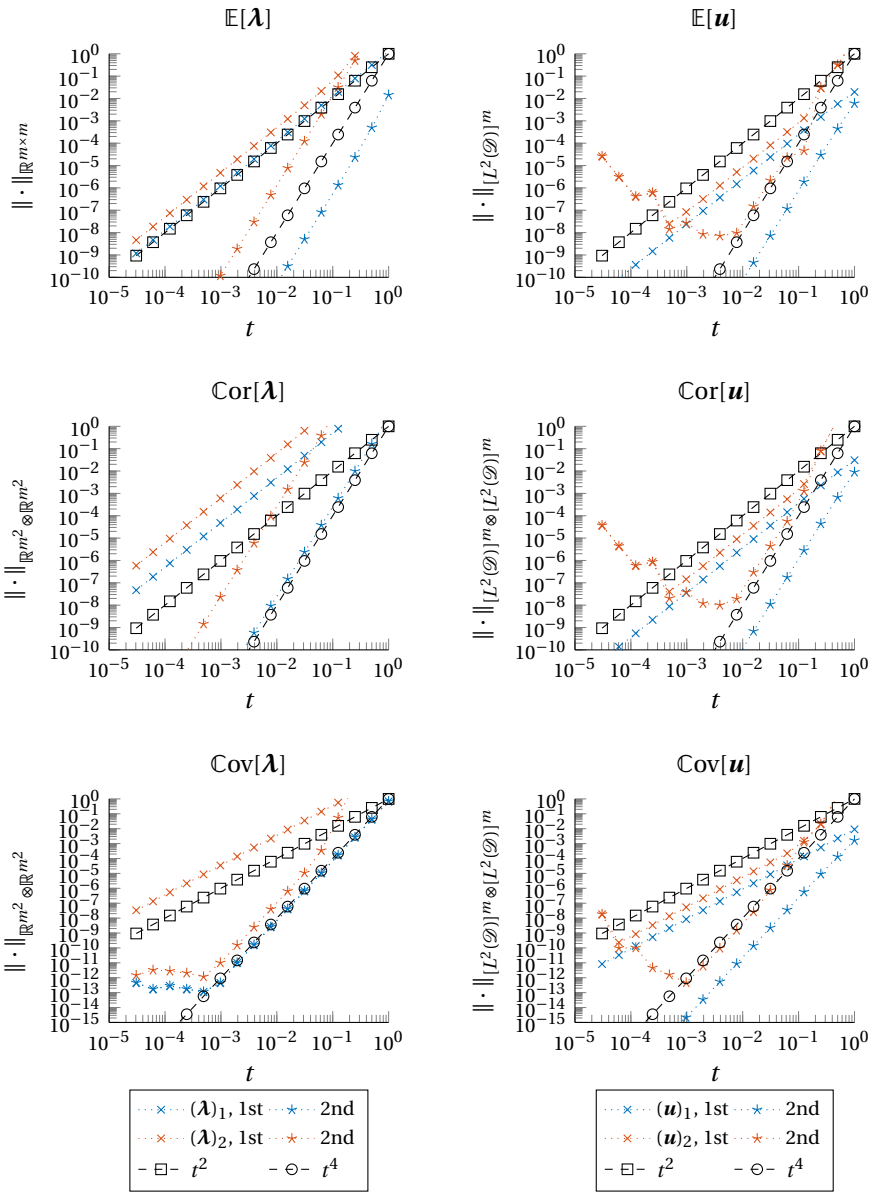

FIGURE V.3. Convergence rates of perturbation approximations for stochastic moments of the Laplace EVP with stochastic deformation.

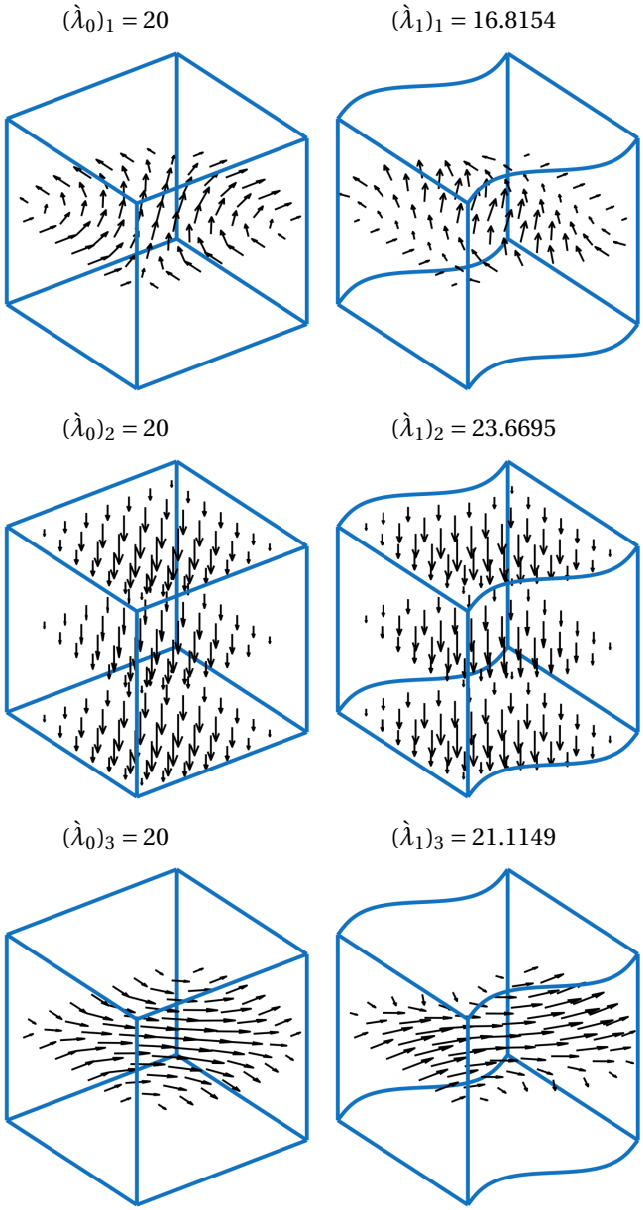

FIGURE V.4. Unperturbed and perturbed eigenfunctions of the Maxwell eigenpairs according to deformation (V.15).

and third eigenvalue cross again for some value $t < 1$, thus the eigenpairs are not ordered according to their eigenvalues at $t = 1$. If we choose $k = 1$ in (V.15) instead, we get deciding order $\grave{k} = 2$.

Verification of the Derivatives. As in the previous example, we check the derivatives of the FE matrices and the polarized eigenpairs by calculating the Taylor approximation at (V.14). For the eigenvalues we consider first- and second-order approximations, while for the eigenfunctions we consider only the first-order Taylor approximations as $\grave{k} = 1$. The norms are chosen in analogy to the previous example for the Laplace EVP. Similarly to the previous example, the derivatives of the FE matrices of order higher than two vanish for this specific deformation.

Uncertainty Quantification. Also in analogy to the previous example, we now consider a stochastic model of the form

$$G_{x_\xi} = \mathbf{x} + V[\xi] = \mathbf{x} + t \sum_{k=1}^{M} V[\phi_k] z_i \ ,$$

where we choose

$$V[\phi_k] = \begin{bmatrix} 0 \\ \exp(-k)\sin(k\pi \mathbf{x}_1) \\ 0 \end{bmatrix}$$

with $M = 5$ and $z_i = \mathscr{U}([-\frac{1}{2}, \frac{1}{2}])$ iid. The unperturbed eigenfunctions are again aligned to the coordinate axes, since the initial polarization varies.

We again use the GL approximation ($n = 2$) as the reference approximation for comparison with the perturbation approximation. For each parameter configuration, we perform a transformation of the eigenpairs to the trajectories with respect to the eigenspace using the third-order approximation of the polarization matrix. This is again possible since the higher-order derivatives of the FE matrices vanish, which we also check empirically up to machine precision.

The convergence rates are illustrated fig. V.6 using norms analogous to the previous example for fixed multiplicity $m = 3$. We can thus confirm the convergence rates, however, we again observe numerical errors in the reference approximations for small perturbation amplitudes.

V.4.3. Variance of Accelerating Mode of a Nine-cell TESLA Cavity. For the last example of the chapter, we consider the nine-cell TESLA cavity domain of [**98**]. We choose the same B-spline basis and quadrature rules as in the previous example to compute the FE matrices. The domain is constructed from 27 patches and, given our FE basis, the discrete EVP has 21652 degrees of freedom. Since the TESLA cavity has nine cells, the first nine eigenvalues are non-degenerate and clustered relatively close together, cf. a similar model of a TESLA cavity in [**27**]. The eigenfunction corresponding to the ninth eigenvalue is called the **accelerating mode** and is of the most interest for the

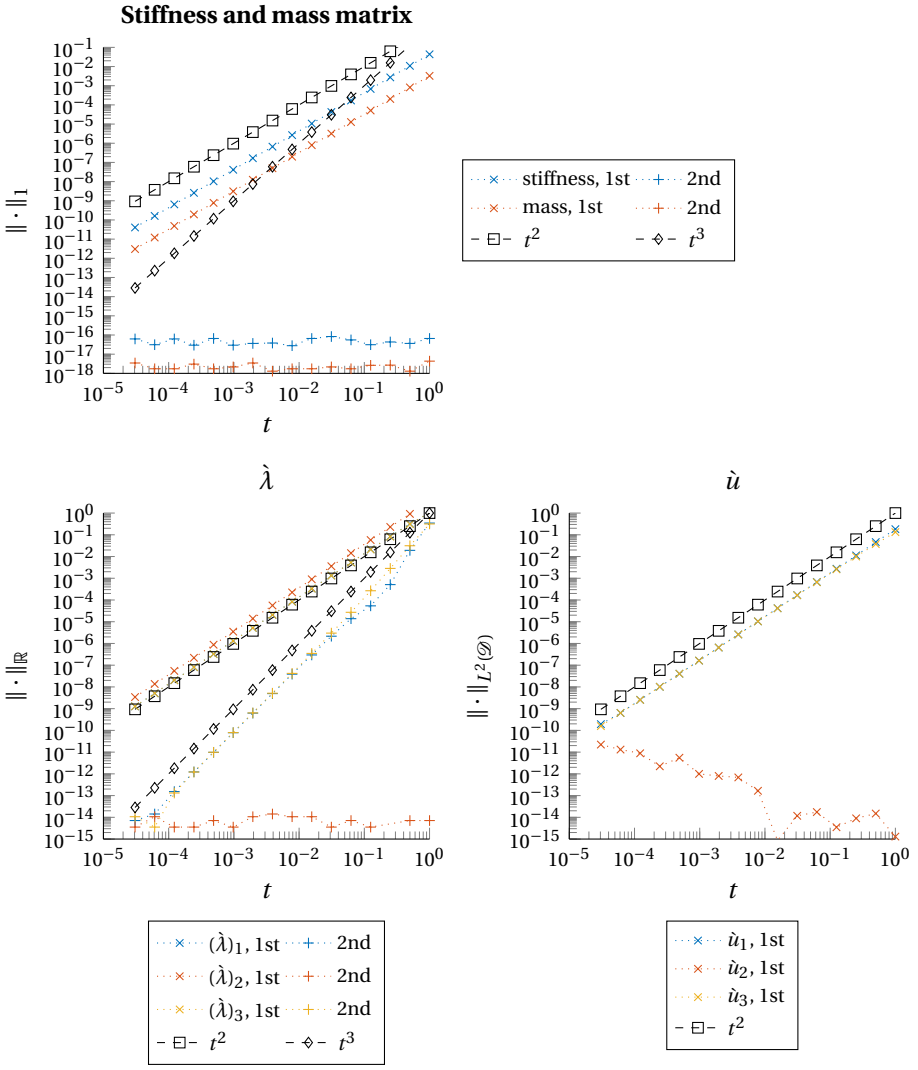

FIGURE V.5. Convergence of the Taylor approximations of the FE matrices and polarized eigenpairs for the Maxwell EVP example.

operation as a particle accelerator. This eigenfunction has the property that the E-field changes direction when comparing adjacent cells of the TESLA cavity, cf. fig. V.7.

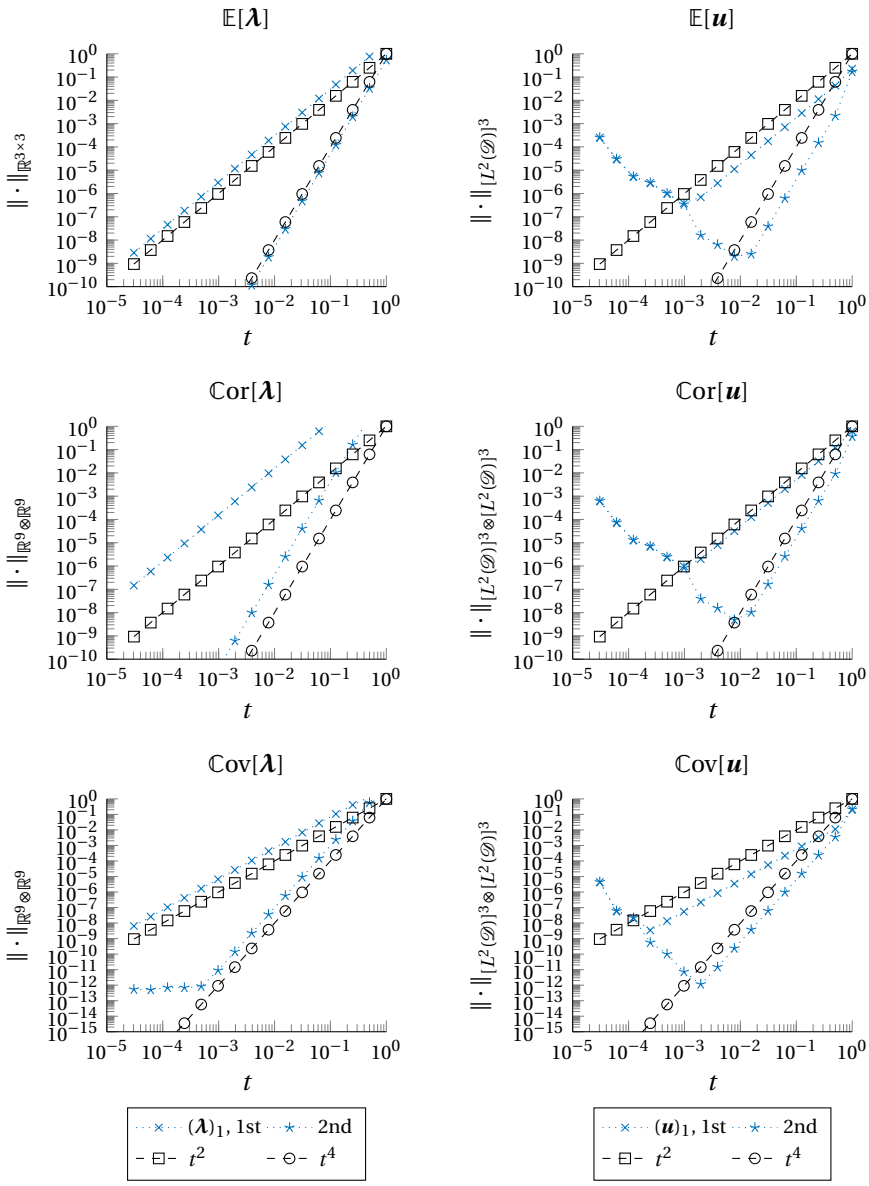

FIGURE V.6. Convergence rates of perturbation approximations for stochastic moments of the Maxwell EVP with stochastic deformation.

We want to quantify its variance for a stochastic model

$$G_{x_\xi} = \mathbf{x} + V[\xi] = \mathbf{x} + t \sum_{i=1}^{4} \sum_{k=1}^{M} V_i[\phi_k] z_i \,,$$

with

$$V_1[\phi_k] = \begin{bmatrix} 0 \\ \exp(-k)\sin(k\pi \mathbf{x}_1) \\ 0 \end{bmatrix}, \qquad V_2[\phi_k] = \begin{bmatrix} 0 \\ \exp(-k)\cos(k\pi \mathbf{x}_1) \\ 0 \end{bmatrix},$$

$$V_3[\phi_k] = \begin{bmatrix} \exp(-k)\sin(k\pi \mathbf{x}_2) \\ 0 \\ 0 \end{bmatrix}, \qquad V_4[\phi_k] = \begin{bmatrix} \exp(-k)\cos(k\pi \mathbf{x}_2) \\ 0 \\ 0 \end{bmatrix},$$

for $k = 1, \ldots, 5$, i.e., $\dim(X) = 20$. Thus, the deformation is a displacement along the symmetry axis $\mathbf{x}_3$ on which the particle beam is ideally accelerated. In fig. V.7 the accelerating mode is depicted on the unperturbed TESLA cavity and for a sample deformation. The last illustration shows the second-order approximation of the variance of the accelerating mode. It turns out that the variance is largest near the connections of two cells.

We consider the computation cost on the Marvin cluster[1] of the University of Bonn. Each assembly of a stiffness and mass matrices for one deformation sample takes around 6 seconds and the solution of the EVP takes around a second.

Calculating one derivative of a FE matrix for the perturbation approximations takes roughly the same time as calculating one perturbed FE matrix. We incur this computational cost for both the stiffness and the mass matrix, for first- and second-order derivatives, and for M perturbation dimensions. The solution of the saddle point equations itself takes less than a second.

Compared to sampling based approaches, the perturbation approximation is again advantageous for high-dimensional parameter spaces and when the amplitude of the perturbation is relatively small. This is the case for shape deformation models, where the deformation is small, so that the overall shape of the domain varies only slightly.

[1] The hardware specifics can be found in section IV.5.

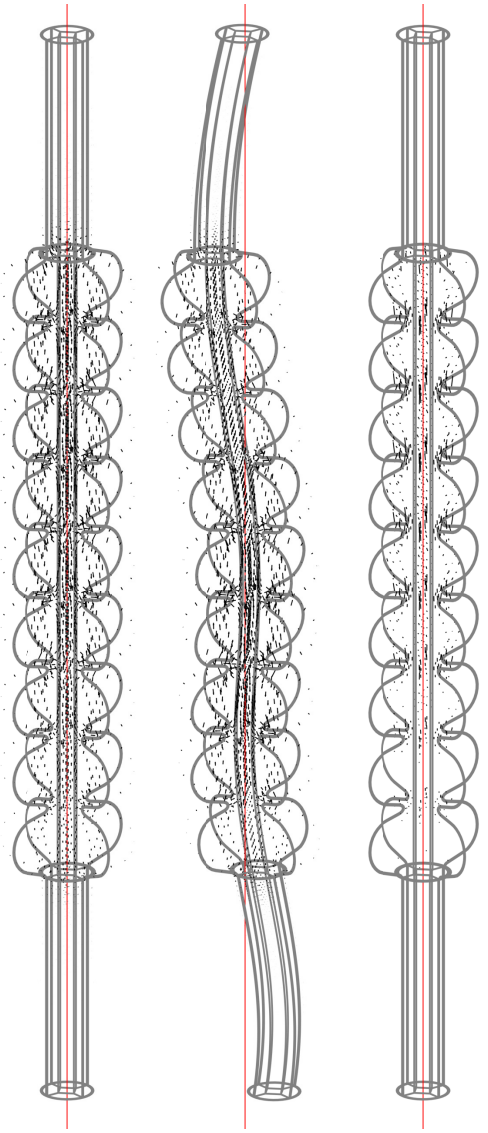

FIGURE V.7. Accelerating mode u_9 of in undeformed TESLA cavity and in a sampled deformation for $t = \frac{1}{10}$ (left and middle, scaled by factor 10^{-3}) as well as the second-order approximation of the variance $\mathbb{V}\mathrm{ar}[u_9]$ for $t = \frac{1}{10}$ (right and only scaled by factor 10^{-2}). The center line of the unperturbed cavity is marked in red.

CHAPTER VI

Bayesian Inverse Problems

In this chapter, we consider how we can incorporate measurement data into perturbation approximations of the stochastic moments of theorem IV.2, in the setting of a Bayesian inverse problem. To this end, we transfer the results of [**26**] to the setting of EVPs. In [**26**] generic Fréchet differentiable mappings are considered, thus the results are applicable locally to the trajectories of the eigenpairs with respect to their eigenspace, cf. theorem III.2. This again includes the eigenvalues trajectories in the traditional sense if the eigenvalues are non-degenerate, cf. corollary III.3, or uncoupled, cf. remark III.24.
Bayesian inversion is a method of estimating a stochastic model parameter by incorporating measurement data into the stochastic model, assuming that the measurement data are also perturbed, usually by an additive Gaussian noise. The assumed probability measure before the measurement is called *prior* and is updated after the measurement to a *posterior* probability measure. This update of the probability measure leads to new posterior stochastic moments, which we describe in this chapter using perturbation approximations with respect to the prior measure. As in [**26**], we also discuss an iterative improvement of the parameter estimate by repeated application of the perturbation approximation of the posterior mean of the perturbed parameter.

VI.1. Bayesian Inversion

We specify the model assumptions for the Bayesian inverse problem. The interested reader may find more information on Bayesian inverse problems in [**57, 94, 93**].

ASSUMPTION VI.1. Consider a **forward response map** $\mathsf{G} : X \to Y$ between two Banach spaces X, Y and refer to X as before as the parameter space and Y the observable space. Since Y is possibly infinite-dimensional, we consider an **observation operator** $\mathsf{O} \in \mathscr{L}(Y;\mathbb{R}^K)$ that returns $K < \infty$ measurements. For brevity, we introduce the composition $\mathsf{Q} \coloneqq \mathsf{O} \circ \mathsf{G} : X \to \mathbb{R}^K$ and call it the **measurement operator**. Let $(\Omega, \mathscr{A}, \mathbb{P})$ be a probability space. As is common for Bayesian inverse problems, we assume that the measurements are subject to additive Gaussian **noise** $\varepsilon \sim \mathcal{N}(0,\Sigma)$, i.e., $\varepsilon : \Omega \to \mathbb{R}^K$, with $\Sigma \in \mathbb{R}^{K\times K}$ a symmetric positive definite covariance matrix. We also assume to have an X-valued random variable $\xi : \Omega \to X$ as in previous chapters, where we call

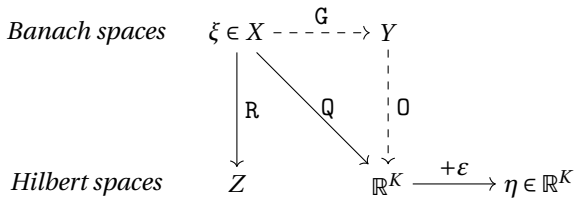

FIGURE VI.1. Setting of the Bayesian inverse problem.

the image measure $\mathbb{P}_\xi$ the **prior distribution** of ξ. It is assumed that the random variables ξ and ε are independent. Thus, the available noisy measurements η^δ are given as realizations of a random variable

$$\eta = \mathsf{Q}(\xi) + \varepsilon\,, \qquad\qquad (\xi,\varepsilon) \sim \mathbb{P}_\xi \otimes \mathcal{N}(0,\Sigma)\,.$$

Given a realization $\eta^\delta \in \mathbb{R}^K$ of this random variable, we define the (unnormalized) **likelihood**

$$\Theta : X \to \mathbb{R}\,, \qquad\qquad x \mapsto \exp(-\Phi(x))\,,$$

and refer to

$$\Phi(x) = \tfrac{1}{2}\|\eta^\delta - \mathsf{Q}(x)\|_\Sigma^2 \tag{VI.1}$$

as the **potential** of the likelihood using the notation (II.41). We assume that (VI.1) is $\mathbb{P}_\xi$-measurable.

In the setting of assumption VI.1, Bayes' rule (lemma II.94) can be expressed using Radon–Nikodým derivatives.

THEOREM VI.2 ([**93**, Theorem 6.29 & 6.31]). *Assume that the potential* (VI.1) *is $\mathbb{P}_\xi$-measurable. Then the posterior distribution $\mathbb{P}_\xi^\delta := \mathbb{P}_{\xi|\eta=\eta^\delta}$ of ξ conditioned on the realization $\eta = \eta^\delta$ exists. It is absolutely continuous with respect to $\mathbb{P}_\xi$, i.e., $\mathbb{P}_\xi^\delta \ll \mathbb{P}_\xi$, and given by the Radon–Nikodým derivative*

$$\frac{\mathrm{d}\mathbb{P}_\xi^\delta}{\mathrm{d}\mathbb{P}_\xi}(x) = \frac{\Theta(x)}{\int_X \Theta(x)\,\mathrm{d}\mathbb{P}_\xi(x)}\,. \tag{VI.2}$$

VI.1.1. Posterior Stochastic Moments. Let $\mathtt{R} : X \to Z$ be a **prediction function** taking values in some Hilbert space Z, where $R(x)$ is some variable of interest, e.g. eigenpair trajectories. The mappings of assumption VI.1 and $\mathtt{R}$ are summarized in fig. VI.1.
For $\mathtt{R} \in L^1_{\mathbb{P}_\xi}(X;Z)$ the **prior mean** is defined as the mean with respect to the prior distribution

$$\mathbb{E}[\mathtt{R}] = \int_X \mathtt{R}\,\mathrm{d}\mathbb{P}_\xi\,.$$

The **posterior mean** is defined with respect to the posterior probability measure $\mathbb{P}^\delta$, that is the mean conditioned on the data η^δ, i.e.,

$$\mathbb{E}_{\mathbb{P}^\delta}[\mathtt{R}] = \mathbb{E}\left[\mathtt{R} \mid \eta = \eta^\delta\right] = \mathbb{E}\left[\mathtt{R}\frac{\Theta}{\mathbb{E}[\Theta]}\right] . \tag{VI.3a}$$

For $\mathtt{R}, \mathtt{R}_1, \mathtt{R}_2 \in L^2_{\mathbb{P}_\xi}(X;Z)$ we can also define a **posterior correlation**

$$\mathbb{C}\mathrm{or}_{\mathbb{P}^\delta}[\mathtt{R}_1, \mathtt{R}_2] = \mathbb{E}_{\mathbb{P}^\delta}[\mathtt{R}_1 \otimes \mathtt{R}_2] , \tag{VI.3b}$$

and a **posterior covariance**

$$\mathbb{C}\mathrm{ov}_{\mathbb{P}^\delta}[\mathtt{R}_1, \mathtt{R}_2] = \mathbb{E}_{\mathbb{P}^\delta}\left[(\mathtt{R}_1 - \mathbb{E}_{\mathbb{P}^\delta}[\mathtt{R}_1]) \otimes (\mathtt{R}_2 - \mathbb{E}_{\mathbb{P}^\delta}[\mathtt{R}_2])\right] . \tag{VI.3c}$$

We use the usual shorthand notations

$$\mathbb{C}\mathrm{or}_{\mathbb{P}^\delta}[\mathtt{R}] = \mathbb{C}\mathrm{or}_{\mathbb{P}^\delta}[\mathtt{R}, \mathtt{R}] , \qquad \mathbb{C}\mathrm{ov}_{\mathbb{P}^\delta}[\mathtt{R}] = \mathbb{C}\mathrm{ov}_{\mathbb{P}^\delta}[\mathtt{R}, \mathtt{R}] ,$$

and $\mathbb{V}\mathrm{ar}_{\mathbb{P}^\delta}[\mathtt{R}] : \mathscr{D} \to \mathbb{R}, x \mapsto \mathbb{V}\mathrm{ar}_{\mathbb{P}^\delta}[\mathtt{R}](x) = \mathbb{C}\mathrm{ov}_{\mathbb{P}^\delta}[\mathtt{R}](x,x)$ which we call the **posterior variance (function)**.
In the following, if the stochastic moments are not indexed, they are to be understood with respect to the prior distribution $\mathbb{P}$ as before.

VI.2. Perturbation Approximations of Posterior Moments

For the following approximations, we require that the mappings $\mathtt{Q} : X \to \mathbb{R}^K$ and $\mathtt{R} : X \to Z$ also have at least C^3-regularity with respect to the parameter $x \in X$. We also require local L^3-integrability as discussed in assumption IV.1.
We intend $\mathtt{Q}$ and $\mathtt{R}$ to be or to depend upon (without deteriorating the above properties) the trajectories of eigenpairs with respect to the eigenspace. The stochastic parameter $x_\xi \in X$ is again given by a perturbational model (IV.1). As discussed in chapter III and assumption IV.1, the trajectories of the eigenpairs with respect to the eigenspace satisfy the above properties locally in a neighborhood $B(x_0)$ around the reference point $x_0 \in X$, cf. theorem III.2. This can be extended to trajectories in the traditional sense for non-degenerate eigenpairs, cf. corollary III.3, or when the degenerate eigenvalues are uncoupled, cf. remark III.24. Since $\mathtt{Q}$ and $\mathtt{R}$ can be interpreted in many ways in the context of EVPs, we stick to this generic notation for the following discussion.
Given a perturbation model of the parameter $x_\xi \in B(x_0) \subset X$ according to (IV.1), $\mathtt{Q} \in C^4(B(x_0);\mathbb{R}^K)$, and $\mathtt{R} \in C^4(B(x_0);Z)$, the mappings of our Bayesian setting can be expressed as expansions

$$\mathtt{Q}_{x_\xi} = \mathtt{Q}_{x_0} + \mathbf{D}_{x_0}\mathtt{Q}[\xi] + \tfrac{1}{2}\mathbf{D}^2_{x_0}\mathtt{Q}[\xi] + \tfrac{1}{6}\mathbf{D}^3_{x_0}\mathtt{Q}[\xi] + \mathscr{O}(\|\xi\|^4_X) , \tag{VI.4a}$$

$$\mathtt{R}_{x_\xi} = \mathtt{R}_{x_0} + \mathbf{D}_{x_0}\mathtt{R}[\xi] + \tfrac{1}{2}\mathbf{D}^2_{x_0}\mathtt{R}[\xi] + \tfrac{1}{6}\mathbf{D}^3_{x_0}\mathtt{R}[\xi] + \mathscr{O}(\|\xi\|^4_X) . \tag{VI.4b}$$

Accordingly, the likelihood $\Theta \in C^4(X;\mathbb{R})$ can be expressed as

$$\Theta_{x_\xi} = \Theta_{x_0} + \mathbf{D}_{x_0}\Theta[\xi] + \tfrac{1}{2}\mathbf{D}^2_{x_0}\Theta[\xi] + \tfrac{1}{6}\mathbf{D}^3_{x_0}\Theta[\xi] + \mathscr{O}(\|\xi\|^4_X) \tag{VI.4c}$$

with

$$\Theta_{x_0} = \exp\Big(-\tfrac{1}{2}\|\eta^\delta - \mathsf{Q}_{x_0}\|_\Sigma^2\Big), \tag{VI.5a}$$

$$\mathbf{D}_{x_0}\Theta[\xi] = \Theta_{x_0}\Big\langle \eta^\delta - \mathsf{Q}_{x_0}, \mathbf{D}_{x_0}\mathsf{Q}[\xi]\Big\rangle_\Sigma, \tag{VI.5b}$$

$$\mathbf{D}^2_{x_0}\Theta[\xi] = \Theta_{x_0}\Big(\Big\langle \eta^\delta - \mathsf{Q}_{x_0}, \mathbf{D}_{x_0}\mathsf{Q}[\xi]\Big\rangle_\Sigma^2 - \|\mathbf{D}_{x_0}\mathsf{Q}[\xi]\|_\Sigma^2 + \Big\langle \eta^\delta - \mathsf{Q}_{x_0}, \mathbf{D}^2_{x_0}\mathsf{Q}[\xi]\Big\rangle_\Sigma\Big), \tag{VI.5c}$$

$$\begin{aligned}\mathbf{D}^3_{x_0}\Theta[\xi] &= \mathbf{D}^2_{x_0}\Theta[\xi]\Big\langle \eta^\delta - \mathsf{Q}_{x_0}, \mathbf{D}_{x_0}\mathsf{Q}[\xi]\Big\rangle_\Sigma \\ &\quad + \Theta_{x_0}\Big(2\Big\langle \eta^\delta - \mathsf{Q}_{x_0}, \mathbf{D}_{x_0}\mathsf{Q}[\xi]\Big\rangle_\Sigma\Big(\Big\langle \eta^\delta - \mathsf{Q}_{x_0}, \mathbf{D}^2_{x_0}\mathsf{Q}[\xi]\Big\rangle_\Sigma - \big\|\mathbf{D}_{x_0}\mathsf{Q}[\xi]\big\|_\Sigma^2\Big) \\ &\quad - 3\big\langle \mathbf{D}_{x_0}\mathsf{Q}[\xi], \mathbf{D}^2_{x_0}\mathsf{Q}[\xi]\big\rangle_\Sigma + \Big\langle \eta^\delta - \mathsf{Q}_{x_0}, \mathbf{D}^3_{x_0}\mathsf{Q}[\xi]\Big\rangle_\Sigma\Big).\end{aligned} \tag{VI.5d}$$

Note that only the derivatives eqs. (VI.5b) to (VI.5d) are random. For ease of notation, we omit the stochastic argument ξ.

The following result can be seen as a generalization of theorem IV.2 for Bayesian inverse problems if we let $\mathsf{R} \in \{\boldsymbol{\lambda}, \boldsymbol{u}\}$.

THEOREM VI.3 ([**26**, Theorem 3.2]). *Let $B(x_0)$ be a neighborhood of $x_0 \in X$ such that $x_\xi \in B(x_0)$ and assume that $\xi \in L^3_{\mathbb{P}}(\Omega; X)$, $\mathsf{Q} \in C^3(B(x_0); \mathbb{R}^K) \cap L^3_{\mathbb{P}_\xi}(\Omega; \mathbb{R}^K)$, and also $\mathsf{R} \in C^3(B(x_0); Z) \cap L^3_{\mathbb{P}_\xi}(\Omega; Z)$. Then it holds that*

$$\begin{aligned}\mathbb{E}_{\mathbb{P}^\delta}[\mathsf{R}_{x_\xi}] &= \mathsf{R}_{x_0} + \mathbb{E}[\mathbf{D}_{x_0}\mathsf{R}] \\ &\quad + \tfrac{1}{2}\Big(\mathbb{E}[\mathbf{D}^2_{x_0}\mathsf{R}] + 2\,\mathbb{C}\mathrm{ov}\Big[\mathbf{D}_{x_0}\mathsf{R}, \Big\langle \eta^\delta - \mathsf{Q}_{x_0}, \mathbf{D}_{x_0}\mathsf{Q}\Big\rangle_\Sigma\Big]\Big) \\ &\quad + \mathcal{O}(\|\xi\|^3_{L^3_{\mathbb{P}}(\Omega;X)}),\end{aligned} \tag{VI.6a}$$

$$\begin{aligned}\mathbb{C}\mathrm{or}_{\mathbb{P}^\delta}[\mathsf{R}_{x_\xi}] &= \mathsf{R}_{x_0} \otimes \mathsf{R}_{x_0} + \Big(\mathbb{E}[\mathbf{D}_{x_0}\mathsf{R}] \otimes \mathsf{R}_{x_0} + \mathsf{R}_{x_0} \otimes \mathbb{E}[\mathbf{D}_{x_0}\mathsf{R}]\Big) \\ &\quad + \tfrac{1}{2}\Big(\mathbb{E}[\mathbf{D}^2_{x_0}\mathsf{R}] \otimes \mathsf{R}_{x_0} + \mathsf{R}_{x_0} \otimes \mathbb{E}[\mathbf{D}^2_{x_0}\mathsf{R}] + 2\,\mathbb{C}\mathrm{or}[\mathbf{D}_{x_0}\mathsf{R}] \\ &\qquad + 2\,\mathbb{C}\mathrm{ov}\Big[(\mathbf{D}_{x_0}\mathsf{R}) \otimes \mathsf{R}_{x_0} + \mathsf{R}_{x_0} \otimes (\mathbf{D}_{x_0}\mathsf{R}), \Big\langle \eta^\delta - \mathsf{Q}_{x_0}, \mathbf{D}_{x_0}\mathsf{Q}\Big\rangle_\Sigma\Big]\Big) \\ &\quad + \mathcal{O}(\|\xi\|^3_{L^3_{\mathbb{P}}(\Omega;X)}),\end{aligned} \tag{VI.6b}$$

$$\mathbb{C}\mathrm{ov}_{\mathbb{P}^\delta}[\mathsf{R}_{x_\xi}] = \mathbb{C}\mathrm{ov}[\mathbf{D}_{x_0}\mathsf{R}] + \mathcal{O}(\|\xi\|^3_{L^3_{\mathbb{P}}(\Omega;X)}). \tag{VI.6c}$$

PROOF. In order to prove (VI.6a), we need to calculate the derivatives of

$$\mathsf{R}_{x_\xi}\frac{\Theta_{x_\xi}}{\mathbb{E}[\Theta_{x_\xi}]} \in C^3(B(x_0); \mathbb{R})$$

with respect to the parameter in order to then apply the mean in analogy to theorem IV.2. The first- and second-order derivatives are

$$\mathbf{D}_{x_0}\Big(\mathsf{R}\frac{\Theta}{\mathbb{E}[\Theta]}\Big) = (\mathbf{D}_{x_0}\mathsf{R})\frac{\Theta_{x_0}}{\mathbb{E}[\Theta_{x_0}]} + \mathsf{R}_{x_0}\frac{\mathbf{D}_{x_0}\Theta}{\mathbb{E}[\Theta_{x_0}]} - \mathsf{R}_{x_0}\frac{\Theta_{x_0}(\mathbf{D}_{x_0}\mathbb{E}[\Theta])}{\mathbb{E}[\Theta_{x_0}]^2},$$

$$\begin{aligned}\mathbf{D}_{x_0}^2\left(\mathtt{R}\frac{\Theta}{\mathbb{E}[\Theta]}\right) &= (\mathbf{D}_{x_0}^2\,\mathtt{R})\frac{\Theta_{x_0}}{\mathbb{E}[\Theta_{x_0}]} + 2(\mathbf{D}_{x_0}\,\mathtt{R})\frac{\mathbf{D}_{x_0}\,\Theta}{\mathbb{E}[\Theta_{x_0}]} - 2(\mathbf{D}_{x_0}\,\mathtt{R})\frac{\Theta_{x_0}(\mathbf{D}_{x_0}\,\mathbb{E}[\Theta])}{\mathbb{E}[\Theta_{x_0}]^2}\\ &\quad + \mathtt{R}_{x_0}\frac{\mathbf{D}_{x_0}^2\,\Theta}{\mathbb{E}[\Theta_{x_0}]} - 2\mathtt{R}_{x_0}\frac{\mathbf{D}_{x_0}\,\Theta(\mathbf{D}_{x_0}\,\mathbb{E}[\Theta])}{\mathbb{E}[\Theta_{x_0}]^2}\\ &\quad - \mathtt{R}_{x_0}\frac{\Theta_{x_0}(\mathbf{D}_{x_0}^2\,\mathbb{E}[\Theta])}{\mathbb{E}[\Theta_{x_0}]^2} + 2\mathtt{R}_{x_0}\frac{\Theta_{x_0}(\mathbf{D}_{x_0}\,\mathbb{E}[\Theta])^2}{\mathbb{E}[\Theta_{x_0}]^3}\,.\end{aligned}$$

Since the k-th Fréchet derivative can be seen as an bounded k-linear operator in $\xi \in X$, cf. section II.3, it holds

$$\mathbf{D}_{x_0}^k\,\mathbb{E}[\Theta] = \mathbf{D}_{x_0}^k \int_X \Theta\,\mathrm{d}\mathbb{P}_\xi = \int_X \mathbf{D}_{x_0}^k\,\Theta\,\mathrm{d}\mathbb{P}_\xi = \mathbb{E}[\mathbf{D}_{x_0}^k\,\Theta]\,, \qquad k = 1,2\,,$$

and $\xi \in L^2_{\mathbb{P}}(\Omega;X)$ implies

$$|\mathbf{D}_{x_0}^k\,\mathbb{E}[\Theta]| \le \mathbb{E}[|\mathbf{D}_{x_0}^k\,\Theta|] \le \|\mathbf{D}_{x_0}^k\,\Theta\|_{\mathscr{L}^{(k)}(X;\mathbb{R})}\|\xi\|_{L^k_{\mathbb{P}}(\Omega;X)} < \infty\,, \qquad k = 1,2\,.$$

so these quantities are well defined. Furthermore, (VI.5a) implies $\mathbb{E}[\Theta_{x_0}] = \Theta_{x_0}$. Thus, applying the mean, exploiting $\mathbf{D}_{x_0}\,\mathbb{E}[\Theta] = \mathbb{E}[\mathbf{D}_{x_0}\,\Theta]$, and using (VI.5b), we get

$$\mathbb{E}_{\mathbb{P}^\delta}[\mathbf{D}_{x_0}\,\mathtt{R}] = \mathbb{E}\left[(\mathbf{D}_{x_0}\,\mathtt{R})\frac{\Theta_{x_0}}{\mathbb{E}[\Theta_{x_0}]}\right] + \mathbb{E}\left[\mathtt{R}_{x_0}\frac{\mathbf{D}_{x_0}\,\Theta}{\mathbb{E}[\Theta_{x_0}]}\right] - \mathbb{E}\left[\mathtt{R}_{x_0}\frac{\Theta_{x_0}(\mathbf{D}_{x_0}\,\mathbb{E}[\Theta])}{\mathbb{E}[\Theta_{x_0}]^2}\right] = \mathbb{E}[\mathbf{D}_{x_0}\,\mathtt{R}]\,,$$

and

$$\begin{aligned}\mathbb{E}_{\mathbb{P}^\delta}[\mathbf{D}_{x_0}^2\,\mathtt{R}] &= \mathbb{E}\left[(\mathbf{D}_{x_0}^2\,\mathtt{R})\frac{\Theta_{x_0}}{\mathbb{E}[\Theta_{x_0}]}\right] + 2\,\mathbb{E}\left[(\mathbf{D}_{x_0}\,\mathtt{R})\frac{\mathbf{D}_{x_0}\,\Theta}{\mathbb{E}[\Theta_{x_0}]}\right] - 2\,\mathbb{E}\left[(\mathbf{D}_{x_0}\,\mathtt{R})\frac{\Theta_{x_0}(\mathbf{D}_{x_0}\,\mathbb{E}[\Theta])}{\mathbb{E}[\Theta_{x_0}]^2}\right]\\ &\quad + \mathbb{E}\left[\mathtt{R}_{x_0}\frac{\mathbf{D}_{x_0}^2\,\Theta}{\mathbb{E}[\Theta_{x_0}]}\right] - 2\,\mathbb{E}\left[\mathtt{R}_{x_0}\frac{(\mathbf{D}_{x_0}\,\Theta)(\mathbf{D}_{x_0}\,\mathbb{E}[\Theta])}{\mathbb{E}[\Theta_{x_0}]^2}\right]\\ &\quad - \mathbb{E}\left[\mathtt{R}_{x_0}\frac{\Theta_{x_0}(\mathbf{D}_{x_0}^2\,\mathbb{E}[\Theta])}{\mathbb{E}[\Theta_{x_0}]^2}\right] + 2\,\mathbb{E}\left[\mathtt{R}_{x_0}\frac{\Theta_{x_0}(\mathbf{D}_{x_0}\,\mathbb{E}[\Theta])^2}{\mathbb{E}[\Theta_{x_0}]^3}\right]\\ &= \mathbb{E}[\mathbf{D}_{x_0}^2\,\mathtt{R}] + 2\,\mathbb{E}\left[(\mathbf{D}_{x_0}\,\mathtt{R})\frac{\mathbf{D}_{x_0}\,\Theta}{\Theta_{x_0}}\right] - 2\,\mathbb{E}\left[(\mathbf{D}_{x_0}\,\mathtt{R})\frac{\mathbf{D}_{x_0}\,\mathbb{E}[\Theta]}{\Theta_{x_0}}\right]\\ &= \mathbb{E}[\mathbf{D}_{x_0}^2\,\mathtt{R}] + 2\,\mathbb{E}\left[(\mathbf{D}_{x_0}\,\mathtt{R})\left\langle \eta^\delta - \mathtt{Q}_{x_0}, \mathbf{D}_{x_0}\,\mathtt{Q}\right\rangle_\Sigma\right] - 2\,\mathbb{E}[\mathbf{D}_{x_0}\,\mathtt{R}]\left\langle \eta^\delta - \mathtt{Q}_{x_0}, \mathbb{E}[\mathbf{D}_{x_0}\,\mathtt{Q}]\right\rangle_\Sigma\\ &= \mathbb{E}[\mathbf{D}_{x_0}^2\,\mathtt{R}] + 2\,\mathbb{C}\mathrm{ov}\left[\mathbf{D}_{x_0}\,\mathtt{R}, \left\langle \eta^\delta - \mathtt{Q}_{x_0}, \mathbf{D}_{x_0}\,\mathtt{Q}\right\rangle_\Sigma\right]\,.\end{aligned}$$

Therefore, applying the mean to the Taylor expansion of $\mathbb{E}_{\mathbb{P}^\delta}[\mathtt{R}_{x_\xi}]$, we arrive at (VI.6a). The approximation (VI.6b) is obtained by considering $\tilde{\mathtt{R}}_\xi := \mathtt{R}_{x_\xi} \otimes \mathtt{R}_{x_\xi}$ and the approximation (VI.6c) is found using (II.39e) in analogy to theorem IV.2. □

Note that compared to theorem IV.2, the approximations for mean and correlation have some extra terms that introduce the influence of the likelihood function Θ. The second-order approximation of the covariance is the same as in theorem IV.2.

REMARK VI.4. Using only C^2-regularity and L^2-integrability, similarly to remark IV.4, the following first-order approximations hold

$$\mathbb{E}_{\mathbb{P}^\delta}[\mathsf{R}_{x_\xi}] = \mathsf{R}_{x_0} + \mathbb{E}[\mathbf{D}_{x_0}\mathsf{R}] + \mathscr{O}(\|\xi\|^2_{L^2_{\mathbb{P}}(\Omega;X)}) , \tag{VI.7a}$$

$$\mathbb{C}\mathrm{or}_{\mathbb{P}^\delta}[\mathsf{R}_{x_\xi}] = \mathsf{R}_{x_0}\otimes\mathsf{R}_{x_0} + \Big(\mathbb{E}[\mathbf{D}_{x_0}\mathsf{R}]\otimes\mathsf{R}_{x_0} + \mathsf{R}_{x_0}\otimes\mathbb{E}[\mathbf{D}_{x_0}\mathsf{R}]\Big) + \mathscr{O}(\|\xi\|^2_{L^3_{\mathbb{P}}(\Omega;X)}) , \tag{VI.7b}$$

$$\mathbb{C}\mathrm{ov}_{\mathbb{P}^\delta}[\mathsf{R}_{x_\xi}] = 0 + \mathscr{O}(\|\xi\|^2_{L^2_{\mathbb{P}}(\Omega;X)}) . \tag{VI.7c}$$

Thus, the first-order approximations for the prior and posterior stochastic moments are identical and independent of the measurement data η^δ.
Given sufficient regularity and integrability, higher-order approximations are again possible, however, when implemented naively, the computational cost grows exponentially. Similarly to the case of the prior distribution, generalizations for higher-order moments are also possible in analogy to [**16, 46**].

In analogy to corollary IV.3, we can find simplifications and improved convergence rates given additional assumptions.

COROLLARY VI.5 ([**26**, Corollary 3.3]). *Let the assumptions of theorem VI.3 hold and assume also that $\xi \in L^3_{\mathbb{P}}(\Omega;X)$ is centered. Then the approximations of theorem VI.3 simplify to*

$$\mathbb{E}_{\mathbb{P}^\delta}[\mathsf{R}_{x_\xi}] = \mathsf{R}_{x_0} + \tfrac{1}{2}\Big(\mathbb{E}[\mathbf{D}^2_{x_0}\mathsf{R}] + 2\,\mathbb{C}\mathrm{or}\Big[\mathbf{D}_0\mathsf{R}, \Big\langle \eta^\delta - \mathsf{Q}_{x_0}, \mathbf{D}_{x_0}\mathsf{Q}\Big\rangle_\Sigma\Big]\Big) + \mathscr{O}(\|\xi\|^3_{L^3_{\mathbb{P}}(\Omega;X)}) , \tag{VI.8a}$$

$$\begin{aligned}\mathbb{C}\mathrm{or}_{\mathbb{P}^\delta}[\mathsf{R}_{x_\xi}] &= \mathsf{R}_{x_0}\otimes\mathsf{R}_{x_0} \\ &\quad + \tfrac{1}{2}\Big(\mathbb{E}[\mathbf{D}_{x_0}\mathsf{R}]\otimes\mathsf{R}_{x_0} + \mathsf{R}_{x_0}\otimes\mathbb{E}[\mathbf{D}_{x_0}\mathsf{R}] + 2\,\mathbb{C}\mathrm{or}[\mathbf{D}_{x_0}\mathsf{R}] \\ &\qquad + 2\,\mathbb{C}\mathrm{or}\Big[\big((\mathbf{D}_{x_0}\mathsf{R})\otimes\mathsf{R}_{x_0} + \mathsf{R}_{x_0}\otimes(\mathbf{D}_{x_0}\mathsf{R})\big), \Big\langle \eta^\delta - \mathsf{Q}_{x_0}, \mathbf{D}_{x_0}\mathsf{Q}\Big\rangle_\Sigma\Big]\Big) \\ &\quad + \mathscr{O}(\|\xi\|^3_{L^3_{\mathbb{P}}(\Omega;X)}) ,\end{aligned} \tag{VI.8b}$$

$$\mathbb{C}\mathrm{ov}_{\mathbb{P}^\delta}[\mathsf{R}_{x_\xi}] = \mathbb{C}\mathrm{or}[\mathbf{D}_{x_0}\mathsf{R}] + \mathscr{O}(\|\xi\|^3_{L^3_{\mathbb{P}}(\Omega;X)}) . \tag{VI.8c}$$

If $\xi \in L^4_{\mathbb{P}}(\Omega;X)$ is centered and skewfree, i.e., $\mathbb{E}[\xi\otimes\xi\otimes\xi] = 0$, $\mathsf{Q} \in C^4(X;\mathbb{R}^K)$, and $\mathsf{R} \in C^4(X;Z) \cap L^4_{\mathbb{P}_\xi}(X;Z)$, then the convergence rate improves to $\mathscr{O}(\|\xi\|^4_{L^4_{\mathbb{P}}(\Omega;X)})$.

PROOF. Again, since the first-order Fréchet derivative is linear in ξ, $\mathbb{E}[\xi] = 0$ implies

$$\mathbb{E}[\mathbf{D}_{x_0}\mathsf{R}] = 0 , \qquad\qquad \mathbb{E}[\mathbf{D}_{x_0}\mathsf{Q}] = 0 ,$$

and thus also $\mathbb{E}[\mathbf{D}_{x_0}\Theta] = 0$ so arrive at the simplified approximations.

For the improved convergence rate, we continue the derivations of theorem VI.3 and get a third-order derivative

$$\begin{aligned}
\mathbf{D}^3_{x_0}\left(\mathtt{R}\frac{\Theta}{\mathbb{E}[\Theta]}\right) &= (\mathbf{D}^3_{x_0}\,\mathtt{R})\frac{\Theta_{x_0}}{\mathbb{E}[\Theta_{x_0}]} + 3(\mathbf{D}^2_{x_0}\,\mathtt{R})\frac{\mathbf{D}_{x_0}\Theta}{\mathbb{E}[\Theta_{x_0}]} - 3(\mathbf{D}^2_{x_0}\,\mathtt{R})\frac{\Theta_{x_0}(\mathbf{D}_{x_0}\mathbb{E}[\Theta])}{\mathbb{E}[\Theta_{x_0}]^2} \\
&\quad + 3(\mathbf{D}_{x_0}\,\mathtt{R})\frac{\mathbf{D}^2_{x_0}\Theta}{\mathbb{E}[\Theta_{x_0}]} - 6(\mathbf{D}_{x_0}\,\mathtt{R})\frac{(\mathbf{D}_{x_0}\Theta)(\mathbf{D}_{x_0}\mathbb{E}[\Theta])}{\mathbb{E}[\Theta_{x_0}]^2} - 3(\mathbf{D}_{x_0}\,\mathtt{R})\frac{\Theta_{x_0}(\mathbf{D}^2_{x_0}\mathbb{E}[\Theta])}{\mathbb{E}[\Theta_{x_0}]^2} \\
&\quad + 6(\mathbf{D}_{x_0}\,\mathtt{R})\frac{\Theta_{x_0}(\mathbf{D}_{x_0}\mathbb{E}[\Theta])^2}{\mathbb{E}[\Theta_{x_0}]^3} + \mathtt{R}_{x_0}\frac{\mathbf{D}^3_{x_0}\Theta}{\mathbb{E}[\Theta_{x_0}]} - 3\mathtt{R}_{x_0}\frac{(\mathbf{D}^2_{x_0}\Theta)(\mathbf{D}_{x_0}\mathbb{E}[\Theta])}{\mathbb{E}[\Theta_{x_0}]^2} \\
&\quad - 3\mathtt{R}_{x_0}\frac{(\mathbf{D}_{x_0}\Theta)(\mathbf{D}^2_{x_0}\mathbb{E}[\Theta])}{\mathbb{E}[\Theta_{x_0}]^2} + 6\mathtt{R}_{x_0}\frac{(\mathbf{D}_{x_0}\Theta)(\mathbf{D}_{x_0}\mathbb{E}[\Theta])^2}{\mathbb{E}[\Theta_{x_0}]^3} - \mathtt{R}_{x_0}\frac{\Theta_{x_0}\mathbf{D}^3_{x_0}\mathbb{E}[\Theta]}{\mathbb{E}[\Theta_{x_0}]^2} \\
&\quad - 6\mathtt{R}_{x_0}\frac{\Theta_{x_0}(\mathbf{D}_{x_0}\mathbb{E}[\Theta])^3}{\mathbb{E}[\Theta_{x_0}]^4} + 6\mathtt{R}_{x_0}\frac{\Theta_{x_0}(\mathbf{D}_{x_0}\mathbb{E}[\Theta])(\mathbf{D}^2_{x_0}\mathbb{E}[\Theta])}{\mathbb{E}[\Theta_{x_0}]^3}\,.
\end{aligned}$$

Since the third-order derivative is trilinear in ξ, the same argument as in theorem VI.3 yields $\mathbb{E}[\mathbf{D}_{x_0}\,\mathtt{R}] = 0$ and $\mathbb{E}[\mathbf{D}_{x_0}\,\mathtt{Q}] = 0$. Similarly, the multiplication of second-order derivatives (bilinear in ξ) and first-order derivatives (linear in ξ) is also trilinear in ξ. Then, if $\mathbb{E}[\xi\otimes\xi\otimes\xi] = 0$, we find

$$\mathbb{E}\left[(\mathbf{D}^2_{x_0}\,\mathtt{R})\frac{\mathbf{D}_{x_0}\Theta}{\mathbb{E}[\Theta_{x_0}]}\right] = 0\,, \qquad \mathbb{E}\left[(\mathbf{D}_{x_0}\,\mathtt{R})\frac{\mathbf{D}^2_{x_0}\Theta}{\mathbb{E}[\Theta_{x_0}]}\right] = 0\,.$$

Thus, the third-order derivatives also vanish under these stricter assumptions. The approximations for correlation (VI.8b) and covariance (VI.8c) follow in complete analogy. □

VI.3. Variation of the Reference Point

The following section reflects on the possibility of changing the reference point $x_0 \in X$ to a more suitable position in the parameter space. For EVPs we need to observe the following caveats.

REMARK VI.6. Since the reference point is linked to the notion of which eigenvalues we consider degenerate, this variation of the reference point requires that we deliberately consider the eigenspace of multiple eigenvalues in a point where they are actually not degenerate. Alternatively, we must consider different eigenspaces with respect to the new reference point. If the eigenpairs that we consider remain non-degenerate throughout the variation or are uncoupled with constant polarization, this is not an issue, since they are Fréchet differentiable.

VI.3.1. Comparison to the Laplace Approximation. We compare the perturbation approximation of the mean of theorem VI.3 to the Laplace approximation, cf. [**97**], which is also based on a second-order Taylor approximation, to highlight the difference between both approaches. To keep the discussion concise, let $X = \mathbb{R}^d$ and

assume that the prior probability is given by a density $f(x) : \mathbb{R}^d \to [0,\infty)$ with $S = \operatorname{supp}(f)$. Let the posterior probability be **unimodal**, i.e., it has a unique local maximum, and let it be given by the **posterior density**

$$f^\delta(x) := \begin{cases} \frac{\exp(-\Phi(x))f(x)}{\int_X \exp(-\Phi(x))f(x)\,\mathrm{d}x} = \frac{\exp(-\mathtt{l}(x))}{\int_X \exp(-\mathtt{l}(x))\,\mathrm{d}x} & x \in S\,, \\ 0 & \text{else}\,, \end{cases}$$

with $\mathtt{l} : S \to \mathbb{R}$ the negative log-posterior density

$$\mathtt{l}(x) := \Phi(x) - \log(f(x))\,.$$

We assume $\Phi, f \in C^2(S)$, such that $\mathtt{l} \in C^2(S)$. Since the posterior density is unimodal, the **maximum a posteriori (MAP) point** is given by

$$x^{\mathrm{MAP}} := \operatorname*{arg\,max}_{x\in S} f^\delta(x) = \operatorname*{arg\,min}_{x\in S} \mathtt{l}(x)\,.$$

We require this specific point as the reference point. If we assume that x^{MAP} is in the interior of S, then $\mathbf{D}_{x^{\mathrm{MAP}}}\mathtt{l}[x] = 0$. Moreover, since X is finite-dimensional, we can represent the second-order derivative $\mathbf{D}^2_{x^{\mathrm{MAP}}}\mathtt{l}$ using the **Hessian matrix**

$$\mathtt{H}_{x^{\mathrm{MAP}}}\mathtt{l} := \Big[\tfrac{\partial^2 \mathtt{l}}{\partial x_i \partial x_j}(x^{\mathrm{MAP}})\Big]_{i,j=1}^n\,.$$

We assume that $\mathtt{H}^2_{x^{\mathrm{MAP}}}\mathtt{l}$ is symmetric positive definite. The Laplace approximation is derived by replacing $\mathtt{l}$ with its second-order Taylor approximation, cf. [**97**], i.e.,

$$\frac{\exp(-\mathtt{l}(x))}{\int_X \exp(-\mathtt{l}(x))\,\mathrm{d}x} \approx \frac{\exp(-\mathtt{l}_{x^{\mathrm{MAP}}} - \frac{1}{2}x^\top(\mathtt{H}_{x^{\mathrm{MAP}}}\mathtt{l})x)}{\int_X \exp(-\mathtt{l}_{x^{\mathrm{MAP}}} - \frac{1}{2}x^\top(\mathtt{H}_{x^{\mathrm{MAP}}}\mathtt{l})x)\,\mathrm{d}x} = \frac{\exp(-\frac{1}{2}x^\top(\mathtt{H}_{x^{\mathrm{MAP}}}\mathtt{l})x)}{\sqrt{(2\pi)^d \det(\mathtt{H}_{x^{\mathrm{MAP}}}\mathtt{l})}}\,.$$

Thus, the Laplace approximation of f^δ is the Gaussian distribution

$$\mathcal{N}(x^{\mathrm{MAP}}, (\mathtt{H}_{x^{\mathrm{MAP}}}\mathtt{l})^{-1})\,.$$

In comparison, the Laplace approximation needs the (unique) MAP point as the reference point, so the caveats of remark VI.6 apply. In general, it is also not necessarily the case for our setting that the posterior is unimodal at all. The Laplace approximation works well when the posterior is close to a multivariate Gaussian distribution and can also be sampled directly. Meanwhile, the point of reference of the perturbation approximation can be chosen more freely and performs best for small perturbations or when the means of the higher-order Fréchet derivatives vanish.

VI.3.2. Iterative Improvement of the Reference Value. The reference point x_0 is an important parameter of the stochastic model and informs the start of the perturbation approximations of theorem VI.3. For the local sensitivity analysis of the forward problem, x_0 is often considered an inherent property of the perturbation model (IV.1).

However, it might sometimes make sense to change x_0 and in turn introduce a bias in the probability distribution of ξ, i.e., for any $\tilde{x}_0$ we may write

$$x_\xi(\omega) = x_0 + \xi(\omega) = \tilde{x}_0 + (\xi(\omega) + x_0 - \tilde{x}_0) =: \tilde{x}_0 + \tilde{\xi}(\omega) \,. \tag{VI.9}$$

We investigate how we can use this flexibility in choosing the reference point to improve our asymptotic expansions. First, we state the following corollary on the posterior mean of the stochastic parameter.

COROLLARY VI.7 ([**26**, Corollary 4.1]). *Under the assumptions of theorem VI.3 holds*

$$\mathbb{E}_{\mathbb{P}^\delta}[x_\xi] = x_0 + \mathbb{E}[\xi] + \mathbb{C}\mathrm{ov}\left[\xi, \left\langle \eta^\delta - \mathbb{Q}_{x_0}, \mathbf{D}_{x_0}\mathbb{Q}[\xi]\right\rangle_\Sigma\right] + \mathscr{O}(\|\xi\|^3_{L^3_{\mathbb{P}}(\Omega;X)}) \,. \tag{VI.10}$$

PROOF. Set $\mathbb{R}_{x_\xi} = x_\xi$ in (VI.6a) from theorem VI.3. □

Since this approximation (VI.10) is hopefully a better estimate of the posterior mean of the stochastic parameter than x_0, we may use the result of this series approximation as the new reference point. This ansatz can be repeated and thus leads to the iteration

$$x^{(n+1)} := x^{(n)} + \mathbb{E}[\xi^{(n)}] + \mathbb{C}\mathrm{ov}\left[\xi^{(n)}, \left\langle \eta^\delta - \mathbb{Q}_{x^{(n)}}, \mathbf{D}_{x^{(n)}}\mathbb{Q}[\xi^{(n)}]\right\rangle_\Sigma\right], \quad n \in \mathbb{N}_0 \,, \tag{VI.11a}$$

with starting value $x^{(0)} := x_0$ and due to (VI.9) we update perturbation model by

$$x_\xi := x^{(n+1)} + \xi^{(n+1)} \,, \qquad \xi^{(n+1)} := \xi + x_0 - x^{(n+1)} \,. \tag{VI.11b}$$

The following corollary adapts Taylor's theorem (theorem II.42 and corollary II.43) to this shift of the reference point.

COROLLARY VI.8 ([**26**, Lemma 2.3]). *Let X, Z be Banach spaces, $U \subset X$ an open convex subset, $\xi \in X$, $x, \tilde{x} \in U$ and $f \in C^k(U;Z)$. Then it holds for $\tilde{\xi} := \xi - (\tilde{x}_0 - x_0) \in X$*

$$f(x_0 + \xi) = \sum_{k=0}^{n-1} \mathbf{D}_{\tilde{x}_0} f[\tilde{\xi}] + \mathscr{O}\left((\max\{\|\xi\|_X, \|\tilde{x}_0 - x_0\|_X\})^n\right) .$$

Combining corollaries VI.7 and VI.8, provided that $x^{(n)} \in B(x_0)$, we can state for n-th iteration that

$$\begin{aligned} \mathbb{E}_{\mathbb{P}^\delta}[x_\xi] = x^{(n)} + \mathbb{E}[\xi^{(n)}] + \mathbb{C}\mathrm{ov}\left[\xi^{(n)}, \left\langle \eta^\delta - \mathbb{Q}_{x^{(n)}}, \mathbf{D}_{x^{(n)}}\mathbb{Q}[\xi^{(n)}]\right\rangle_\Sigma\right] \\ + \mathscr{O}\left(\max\left\{\|\xi\|_{L^3_{\mathbb{P}}(\Omega;X)}, \|x^{(0)} - x^{(n)}\|_X\right\}^3\right), \qquad n \in \mathbb{N}_0 \,. \end{aligned} \tag{VI.12}$$

If the iteration (VI.11) converges, then the approximation simplifies as follows.

COROLLARY VI.9 ([**26**, Corollary 4.2]). *Let the assumptions of theorem VI.3 hold and let $x^\star$ be a fixed point of iteration* (VI.11). *Then it holds*

$$\mathbb{E}_{\mathbb{P}^\delta}[x_\xi] = x^\star + \mathscr{O}\left(\max\left\{\|\xi\|_{L^3_{\mathbb{P}}(\Omega;X)}, \|x^{(0)} - x^\star\|_X\right\}^3\right) .$$

Connection to Tikhonov regularization. Although we cannot guarantee the existence and uniqueness of a fixed point of iteration (VI.11) in general, we can relate the iteration to classical inverse problems. To this end, we assume that the parameter space X is a Hilbert space, to restrict ourselves to tensor products thereof.
Let $\mathbb{C}\mathrm{ov}[\xi] \in X \otimes X$ be a covariance (function), then the covariance operator, cf. definition II.96, can be expressed as

$$\mathscr{C} : X \to X\,, \qquad\qquad \mathscr{C}\,x = \big(\mathrm{Id} \otimes \langle \cdot, x\rangle_X\big)\,\mathbb{C}\mathrm{ov}[\xi]\,.$$

Recall that $\mathscr{C}$ is continuous and self-adjoint with respect to $\langle\cdot,\cdot\rangle_X$. Assume that $\mathscr{C}$ is invertible and consider the Tikhonov-regularized classical inverse problem

$$\min_{x\in X} F(x)\,, \qquad\qquad F(x) \coloneqq \tfrac{1}{2}\left(\|\eta^\delta - \mathsf{Q}(x)\|_\Sigma^2 + \|x - x_0 - \mathbb{E}[\xi]\|_{\mathscr{C}}^2\right)\,, \tag{VI.13}$$

with $\|\cdot\|_{\mathscr{C}} \coloneqq \sqrt{\langle \mathscr{C}^{-1}\cdot,\cdot\rangle_X}$. The Fréchet derivative of F is given by

$$\mathbf{D}_x F[y] = -\left\langle \eta^\delta - \mathsf{Q}_x, \mathbf{D}_x\mathsf{Q}[y]\right\rangle_\Sigma + \left\langle \mathscr{C}^{-1}(x - x_0 - \mathbb{E}[\xi]), y\right\rangle_X\,,$$

implying that the gradient $\operatorname{grad} F : X \to X$, cf. definition II.35, is given by

$$\operatorname{grad} F(x) = -(\mathbf{D}_x\mathsf{Q})^*\,\Sigma^{-1}\left(\eta^\delta - \mathsf{Q}_x\right) + \mathscr{C}^{-1}\left(x - x_0 - \mathbb{E}[\xi]\right)\,.$$

Here, $(\mathbf{D}_x\mathsf{Q})^*$ is the adjoint with respect to the $\mathbb{R}^K$-inner product. Now, observing that (VI.11b) implies $\mathbb{C}\mathrm{ov}[\xi^{(n)}] = \mathbb{C}\mathrm{ov}[\xi]$ and $\mathbb{E}[\xi^{(n)}] = \mathbb{E}[\xi] + x_0 - x^{(n)}$, the descent algorithm

$$x^{(n+1)} = x^{(n)} + d^{(n)} \tag{VI.14}$$

with

$$\begin{aligned} d^{(n)} &= -\,\mathscr{C}\operatorname{grad} F(x^{(n)}) \\ &= \left(\mathrm{Id} \otimes \left\langle \cdot, (\mathbf{D}_{x^{(n)}}\mathsf{Q})^*\,\Sigma^{-1}(\eta^\delta - \mathsf{Q}_{x^{(n)}})\right\rangle_X\right)\mathbb{C}\mathrm{ov}[\xi] + \mathbb{E}[\xi] + x_0 - x^{(n)} \\ &= \mathbb{E}[\xi^{(n)}] + \mathbb{C}\mathrm{ov}\left[\xi^{(n)}, \left\langle \eta^\delta - \mathsf{Q}_{x^{(n)}}, \mathbf{D}_{x^{(n)}}\mathsf{Q}[\xi^{(n)}]\right\rangle_\Sigma\right] \end{aligned}$$

coincides with (VI.11a). Our assumptions on $\mathscr{C}$ guarantee that

$$\left\langle d^{(n)}, \operatorname{grad} F(x^{(n)})\right\rangle_X = -\left\langle \mathscr{C}\operatorname{grad} F(x^{(n)}), \operatorname{grad} F(x^{(n)})\right\rangle_X < 0\,,$$

implying that $d^{(n)}$ is a descent direction for F. Even with $d^{(n)}$ being a descent direction, convergence of the iteration can only be proven if $d^{(n)}$ is scaled with a sufficiently small step length. However, if the iteration does converge, i.e., $x^{(n)} \to x^\star$, then so does $d^{(n)} \to 0$ and $\operatorname{grad} F(x^{(n)}) \to 0$. Thus, if our iteration (VI.11) converges, then its limit $x^\star$ is a (local) minimizer of (VI.13) assuming $(\mathbf{D}^2_{x^\star} F)[y] > 0$ for all $0 \neq y \in X$. Vice versa, if the descent algorithm (VI.14) of the minimization problem (VI.13) converges, then it is a fixed point of the iteration (VI.11), for which the improved approximation estimate of corollary VI.9 holds. Again, the caveats of remark VI.6 apply to degenerate eigenpairs. In the presence of non-degenerate or uncoupled eigenspaces, we need to

consider how large the neighborhood $B(x_0)$ can be chosen and if the iteration moves beyond this frame of reference.

VI.4. Implementation

We again consider a more explicit setting where x_ξ is given in a KLE-like decomposition of the form (IV.9). Then,the series representation of (VI.4) can be decomposed as

$$\mathtt{Q}_{x_\xi} = \mathtt{Q}_{x_0} + \sum_{i=1}^{M} \mathbf{D}_{x_0} \mathtt{Q}[\phi_i] z_i + \frac{1}{2} \sum_{i,j=1}^{M} \mathbf{D}^2_{x_0} \mathtt{Q}[\phi_i, \phi_j] z_i z_j + \mathcal{O}(\|\xi\|_X^3) \,, \tag{VI.15a}$$

$$\Theta_{x_\xi} = \Theta_{x_0} + \sum_{i=1}^{M} \mathbf{D}_{x_0} \Theta[\phi_i] z_i + \frac{1}{2} \sum_{i,j=1}^{M} \mathbf{D}^2_{x_0} \Theta[\phi_i, \phi_j] z_i z_j + \mathcal{O}(\|\xi\|_X^3) \,, \tag{VI.15b}$$

$$\mathtt{R}_{x_\xi} = \mathtt{R}_{x_0} + \sum_{i=1}^{M} \mathbf{D}_{x_0} \mathtt{R}[\phi_i] z_i + \frac{1}{2} \sum_{i,j=1}^{M} \mathbf{D}^2_{x_0} \mathtt{R}[\phi_i, \phi_j] z_i z_j + \mathcal{O}(\|\xi\|_X^3) \,. \tag{VI.15c}$$

If $\mathtt{Q}$ and $\mathtt{R}$ are derivatives of the eigenpairs (with respect to the eigenspace), the derivatives can be calculated as described in chapter III. The following corollary adapts corollary IV.7 to the setting of Bayesian inverse problems.

COROLLARY VI.10 ([**26**, Corollary 5.1]). *Let the assumptions of theorem VI.3 hold and assume that the prior distribution of x_ξ is given in the form as in* (IV.9) *with $(z_j)_{j\in\mathbb{N}}$ being pairwise uncorrelated random variables satisfying $z_j \in L^2_{\mathbb{P}}(\Omega)$. Then the expansions in theorem VI.3 read*

(VI.16a)

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}^\delta}[\mathtt{R}_{x_\xi}] = \mathtt{R}_{x_0} &+ \mathbf{D}_{x_0} \mathtt{R}\left[\sum_{i=1}^{M} \phi_i \, \mathbb{E}[z_i]\right] \\
&+ \frac{1}{2}\left(\sum_{i=1}^{M} \mathbb{V}\mathrm{ar}[z_i] \, \mathbf{D}^2_{x_0} \mathtt{R}[\phi_i] + \mathbf{D}^2_{x_0} \mathtt{R}\left[\sum_{i=1}^{M} \phi_i \, \mathbb{E}[z_i]\right]\right) \\
&+ \sum_{i=1}^{M} \mathbb{V}\mathrm{ar}[z_i] \, \mathbf{D}_{x_0} \mathtt{R}[\phi_i] \left\langle \eta^\delta - \mathtt{Q}_{x_0}, \mathbf{D}_{x_0} \mathtt{Q}[\phi_i] \right\rangle_\Sigma \\
&+ \mathcal{O}(\|\xi\|^3_{L^3_{\mathbb{P}}(\Omega;X)}) \,,
\end{aligned}$$

(VI.16b)

$$\begin{aligned}\mathbb{C}\mathrm{or}_{\mathbb{P}^\delta}[\mathtt{R}_{x_\xi}] &= \mathtt{R}_{x_0}\otimes\mathtt{R}_{x_0} + \mathbf{D}_{x_0}\mathtt{R}\left[\sum_{i=1}^M \phi_i\,\mathbb{E}[z_i]\right]\otimes\mathtt{R}_{x_0} + \mathtt{R}_{x_0}\otimes\mathbf{D}_{x_0}\mathtt{R}\left[\sum_{i=1}^M \phi_i\,\mathbb{E}[z_i]\right]\\ &\quad+\frac{1}{2}\left(\sum_{i=1}^M \mathbb{V}\mathrm{ar}[z_i]\,\mathbf{D}^2_{x_0}\mathtt{R}[\phi_i]\otimes\mathtt{R}_{x_0} + \mathtt{R}_{x_0}\otimes\sum_{i=1}^M \mathbb{V}\mathrm{ar}[z_i]\,\mathbf{D}^2_{x_0}\mathtt{R}[\phi_i]\right.\\ &\qquad\left.+\mathbf{D}^2_{x_0}\mathtt{R}\left[\sum_{i=1}^M \phi_i\,\mathbb{E}[z_i]\right]\otimes\mathtt{R}_{x_0} + \mathtt{R}_{x_0}\otimes\mathbf{D}^2_{x_0}\mathtt{R}\left[\sum_{i=1}^M \phi_i\,\mathbb{E}[z_i]\right]\right)\\ &\quad+\sum_{i=1}^M \mathbb{V}\mathrm{ar}[z_i]\,\mathbf{D}_{x_0}\mathtt{R}[\phi_i]\otimes\mathbf{D}_{x_0}\mathtt{R}[\phi_i]\\ &\quad+\mathbf{D}_{x_0}\mathtt{R}\left[\sum_{i=1}^M \phi_i\,\mathbb{E}[z_i]\right]\otimes\mathbf{D}_{x_0}\mathtt{R}\left[\sum_{i=1}^M \phi_i\,\mathbb{E}[z_i]\right]\\ &\quad+\sum_{i=1}^M \mathbb{V}\mathrm{ar}[z_i]\left(\mathbf{D}_{x_0}\mathtt{R}[\phi_i]\otimes\mathtt{R}_{x_0} + \mathtt{R}_{x_0}\otimes\mathbf{D}_{x_0}\mathtt{R}[\phi_i]\right)\left\langle \eta^\delta - \mathtt{Q}_{x_0}, \mathbf{D}_{x_0}\mathtt{Q}[\phi_i]\right\rangle_\Sigma\\ &\quad+\mathscr{O}(\|\xi\|^3_{L^3_{\mathbb{P}}(\Omega;X)})\,,\end{aligned}$$

(VI.16c)

$$\mathbb{C}\mathrm{ov}_{\mathbb{P}^\delta}[\mathtt{R}_{x_\xi}] = \sum_{i=1}^M \mathbb{V}\mathrm{ar}[z_i]\,\mathbf{D}_{x_0}\mathtt{R}[\phi_i]\otimes\mathbf{D}_{x_0}\mathtt{R}[\phi_i] + \mathscr{O}(\|\xi\|^3_{L^3_{\mathbb{P}}(\Omega;X)})\,.$$

As in corollary VI.5, the first-order terms vanish, if we assume that the z_j are centered, i.e., $\mathbb{E}[z_j]=0$. Furthermore, if $\xi\in L^4_{\mathbb{P}}(\Omega;X)$ is centered and skewfree, i.e., $\mathbb{E}[\xi\otimes\xi\otimes\xi]=0$, $\mathtt{Q}\in C^4(B(x_0);\mathbb{R}^K)$, and $\mathtt{R}\in C^4(B(x_0);Z)\cap L^2_{\mathbb{P}_\xi}(X;Z)$, then the error term improves to $\mathscr{O}(\|\xi\|^4_{L^4_{\mathbb{P}}(\Omega;X)})$.

PROOF. Insert (IV.9) and (VI.15) into theorem VI.3 and corollary VI.5. □

VI.4.1. Iterative Approximation of the Posterior Mean. Now we consider how the decomposition of the parameter (IV.9) can be applied to the iteration (VI.11) that repeatedly uses the second-order approximation of the posterior mean of the parameter. According to (IV.9), let

$$x_\xi(\omega) = x_0 + \sum_{i=1}^M \phi_i z_i(\omega) = \sum_{i=1}^M \phi_i\big(z_i(\omega) + \underbrace{\langle\phi_i, x_0\rangle_X}_{=:y_i}\big) = \sum_{i=1}^M \phi_i\big(z_i(\omega)+y_i\big)\,. \tag{VI.17}$$

Assuming $\|\phi_i\|_X = 1$, for any $\alpha,\beta\in\mathbb{R}^M$, it holds

$$\langle\alpha,\beta\rangle_{\ell^2} = \left\langle\sum_{i=1}^M \phi_i\alpha_i, \sum_{i=1}^M \phi_i\beta_i\right\rangle_X\,.$$

So, the iteration of section VI.3.2 can be transferred to $\mathbb{R}^M$ equipped with the Euclidean ℓ^2-inner product. To this end, set the initial value $y^{(0)} \in \mathbb{R}^M, z^{(0)} : \Omega \to \mathbb{R}^M$ according to x_0. Then, iteration (VI.11) can be expressed as elementwise updates on the vector $y \in \mathbb{R}^M$, i.e.,

$$y^{(n+1)} = y^{(n)} + \mathbb{E}[z^{(n)}] + \left[\mathbb{V}\mathrm{ar}[z_i^{(n)}] \left\langle \eta^\delta - \mathsf{Q}_{x^{(n)}}, \mathbf{D}_{x^{(n)}} \mathsf{Q}[\phi_i] \right\rangle_\Sigma \right]_{i=1}^M , \tag{VI.18}$$

with updates of the perturbation model according to

$$z^{(n+1)} = z^{(0)} + y^{(0)} - y^{(n+1)} , \qquad x^{(n)} = x_0 + \sum_{i=1}^M \phi_i z_i^{(n)} .$$

VI.5. Numerical Examples

In this section, we provide two examples to demonstrate the perturbation approximation of theorem VI.3 and the iteration (VI.11), both of which are based on the Laplace EVP. The first example again builds on section IV.5, which makes it possible to directly compare the perturbation approximations of the prior and posterior stochastic moments.

The second example builds on example I.1, where we choose the coefficient according to a perturbation that is not differentiable in the spatial variable. Note that the perturbation is still analytic with respect to the stochastic parameterization. For this example, we consider measurements that depend locally smoothly on the eigenvalues. We also show the effect of scaling the measurement noise.

The perturbation approximations are applicable to other forward models as long as the differentiability and integrability requirements are met, e.g., models based on PDEs or ODEs. Since the focus of this thesis is on EVPs, we refer the interested reader to the examples in [**26**].

VI.5.1. Extension of previous Laplace Eigenvalue Problem on Unit Square. Consider the Laplace model of section IV.5 as a forward model. We take measurements of the eigenfunction u_1 that relates to the smallest non-degenerate eigenvalue λ_1 at the $K = 5$ points

$$\left\{ \left(\tfrac{1}{2}, \tfrac{1}{2}\right), \left(\tfrac{1}{4}, \tfrac{1}{4}\right), \left(\tfrac{3}{4}, \tfrac{1}{4}\right), \left(\tfrac{3}{4}, \tfrac{3}{4}\right), \left(\tfrac{1}{4}, \tfrac{3}{4}\right) \right\} \subset \mathscr{D} = (0,1)^2 .$$

Formally, this can be expressed as mappings $\mathsf{G} = u_1 : X \to V$, $\mathsf{O} : V \to \mathbb{R}^K$. The unperturbed eigenfunction and the five measurement points are illustrated in fig. VI.2. In

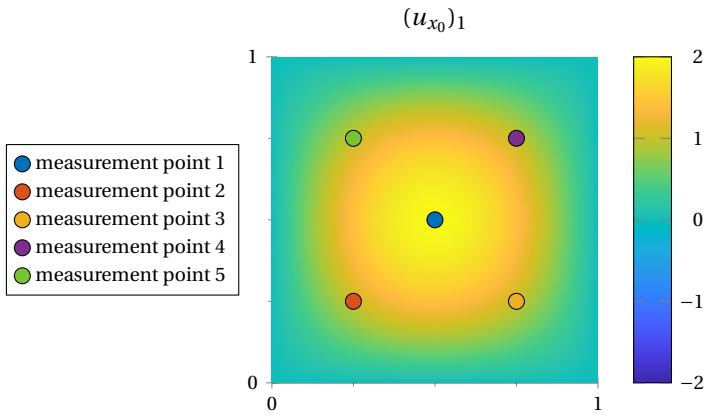

FIGURE VI.2. Measurement points on eigenfunction u_1 of the Laplace EVP ($\mathscr{D} = (0,1)^2$).

the following experiments, we assume that the measurement covariance of the measurement noise is

$$\Sigma = \begin{bmatrix} 1 & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \\ \frac{1}{5} & 1 & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \\ \frac{1}{5} & \frac{1}{5} & 1 & \frac{1}{5} & \frac{1}{5} \\ \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & 1 & \frac{1}{5} \\ \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & 1 \end{bmatrix}$$

and that we have measured the noisy data

$$\eta^\delta = \begin{bmatrix} 1.99 \\ 1.01 \\ 1.01 \\ 1.10 \\ 1.01 \end{bmatrix}.$$

Thus, the data are relatively close to the unperturbed state, but the fourth measurement is slightly higher than the unperturbed state.
For `R`, we consider the trajectories of the eigenpairs with respect to their eigenspace, which belong to the three smallest eigenvalues as in section IV.5. Additionally, we consider the posterior moments of $x_\xi : X \to X$. Since the parameter x_ξ is the representation of the two random fields, we evaluate it as a pair $(\mu, \varepsilon) \in [L^2(\mathscr{D})]^2$, which informs the norm we use to calculate the convergence rates.

Confirmation of Convergence Rate of Uncentered Approximations. We now want to confirm the convergence rates of theorem VI.3 and corollary VI.7 as well as the respective first-order approximations, which are the same for the prior stochastic moments.
First, as in section IV.5, we test a prematurely truncated KLE, such that $\dim(X) = 8$, to benchmark the perturbation approximation against a GL approximation. The perturbation amplitudes are also chosen as in section IV.5 and evaluated for perturbation amplitudes

$$t \in \left\{2^n \,:\, n = -15, \dots, 0\right\} .$$

We use the same norms for the eigenpairs and translate the parameters x_ξ into random fields to use the $\|\cdot\|_{[L^2(\mathscr{D})]^2}$-norm and its tensorized version. The predicted convergence rates are confirmed in fig. VI.3. In fig. VI.4 we test the perturbation approximations against QMC approximations using a Halton sequence of length 10^7, where the more accurate truncation is used for the random fields, such that $\dim(X) = 114$. The calculations were again conducted on the Marvin cluster[1] of the University of Bonn. In analogy to section IV.5, the calculations were parallelized with respect to the eigenfunctions of the covariance operator, cf. corollary VI.10. We can confirm the convergence of the approximations in general, however, the QMC method again struggles for small perturbation amplitudes. For this method, the results for mean and corelation of the degenerate eigenfunctions are not as good as predicted. The convergence rate of the covariance estimate suggests that the actual convergence rate would only have been visible for very small perturbation amplitudes, where the QMC approximation is not accurate enough for our fixed computation budget.
The calculation of 10^7 samples took approximately an hour, slightly longer than the calculations of section IV.5. This is due to the additional effort to evaluate the likelihood function for each sample as a weight and to calculate the normalization of the posterior distribution.
Confirmation of Convergence Rate of Centered Approximations. As in section IV.5, we also confirm the improved convergence rates of corollary VI.5 for a centered and skewfree prior distribution. We repeat the experiment with the respective centered and skewfree parameters of section IV.5 and the data of the previous experiment. In fig. VI.5 the convergence rates of the perturbation approximation compared with the GL approximations are illustrated for the prematurely truncated KLE. The results for the more accurately truncated KLE with comparison against a QMC approximation are illustrated in fig. VI.6. The convergence rates are confirmed with the same caveats regarding the precision of the QMC method for small perturbation amplitudes, especially in combination with the projection to the eigenspace for degenerate eigenfunctions.

[1] The hardware specifics can be found in section IV.5.

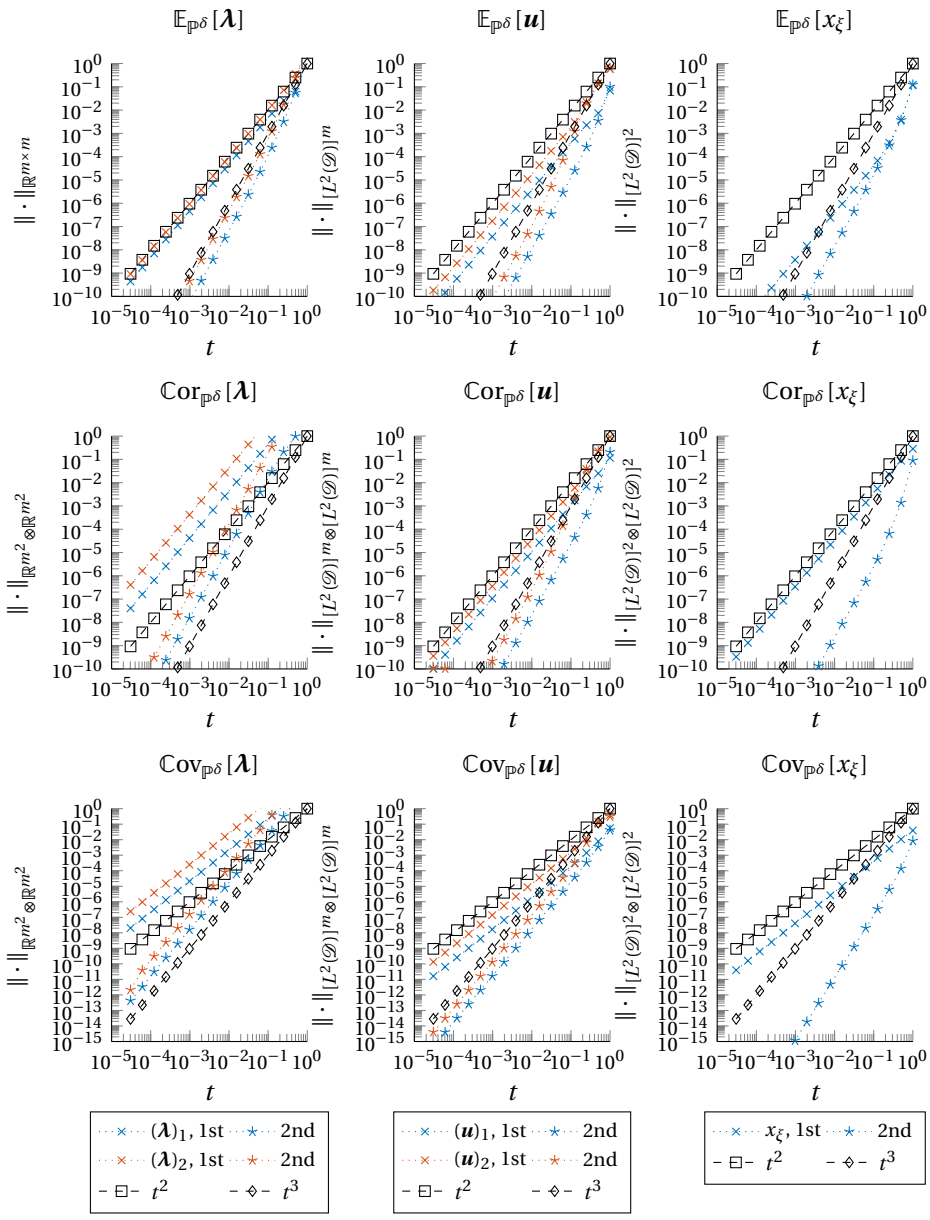

FIGURE VI.3. Convergence rates of uncentered approximations compared to GL estimate for the Laplace EVP ($\mathscr{D} = (0,1)^2$) with coefficients perturbed by random fields.

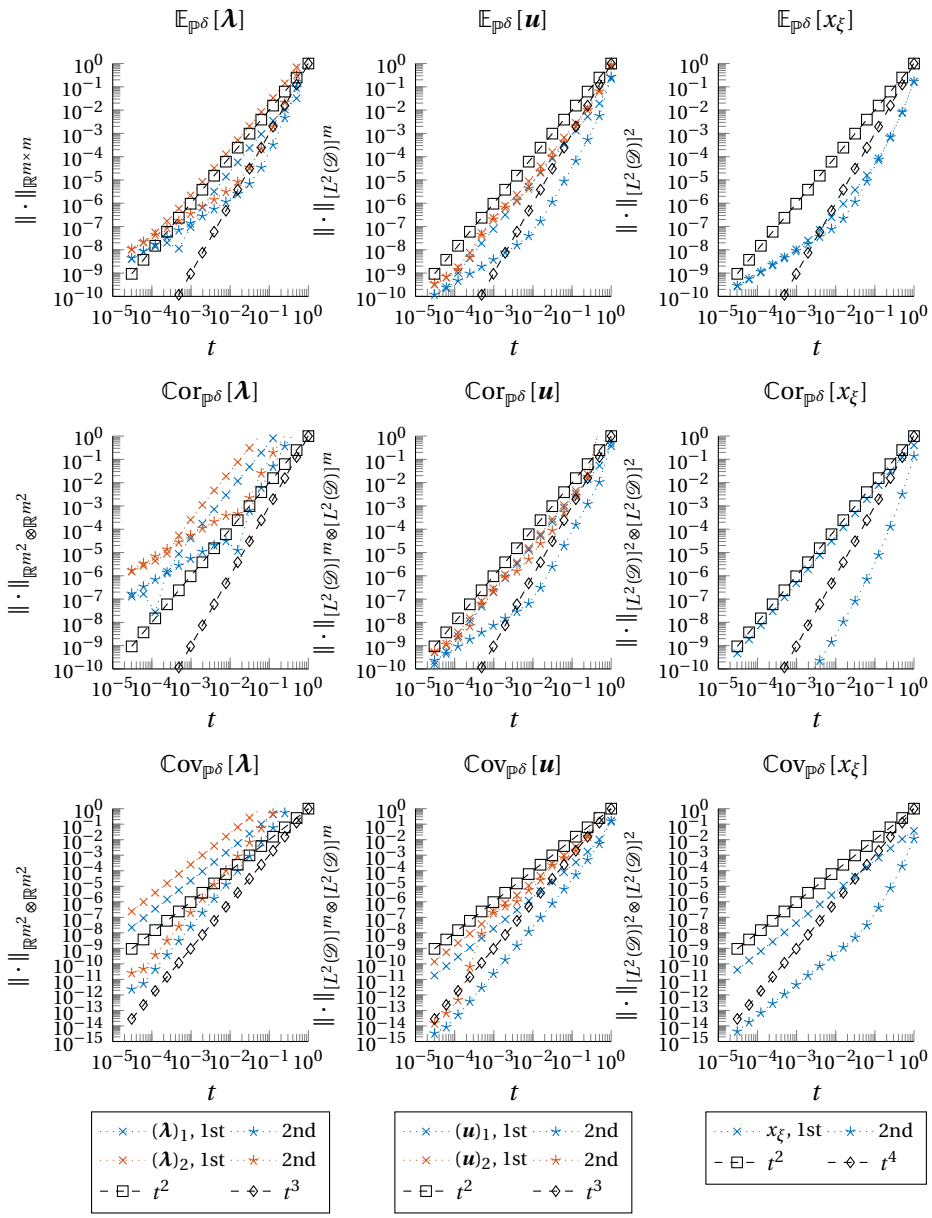

FIGURE VI.4. Convergence rates of uncentered approximations compared to QMC estimate for the Laplace EVP ($\mathscr{D} = (0,1)^2$) with coefficients perturbed by random fields.

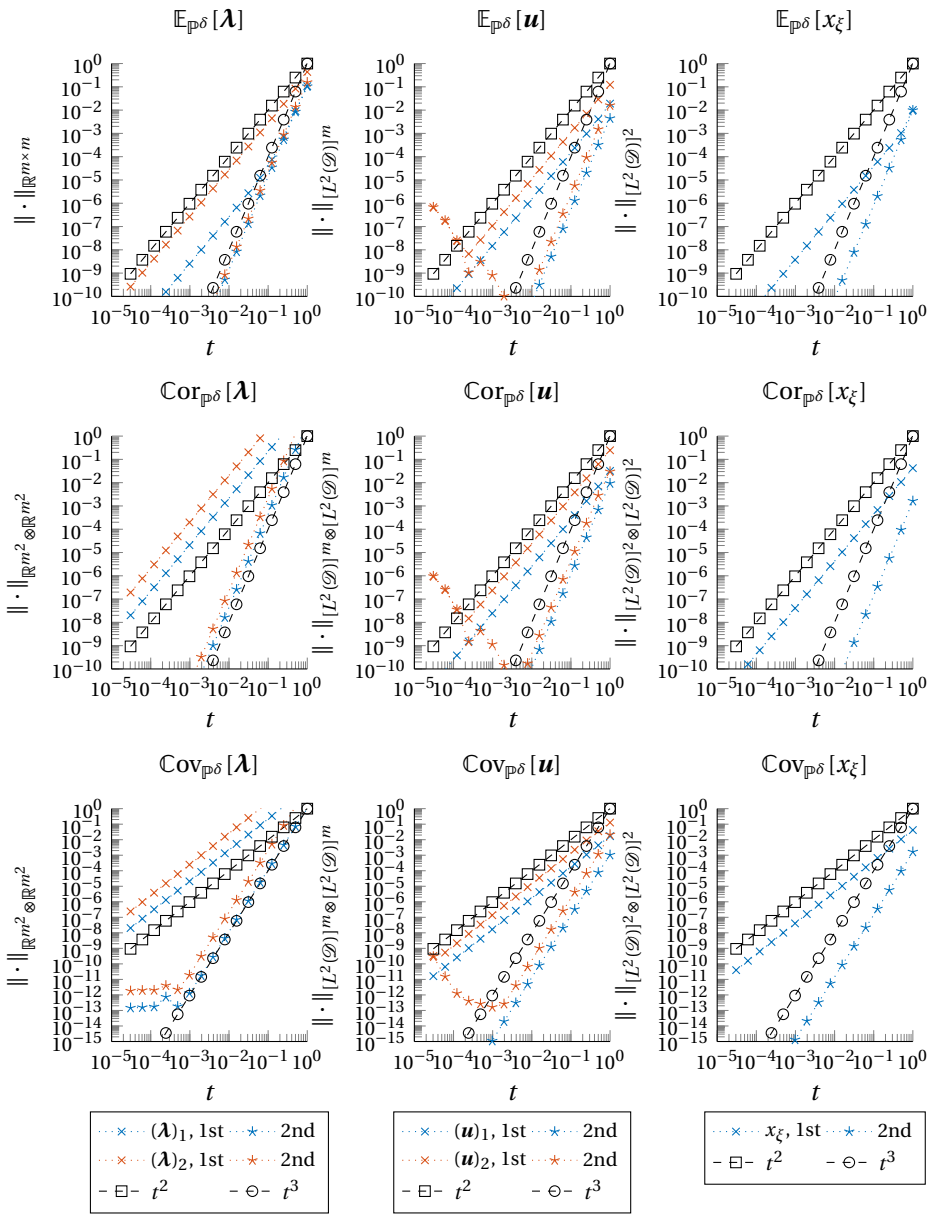

FIGURE VI.5. Convergence rates of centered approximations compared to GL estimate for the Laplace EVP ($\mathscr{D} = (0,1)^2$) with coefficients perturbed by random fields.

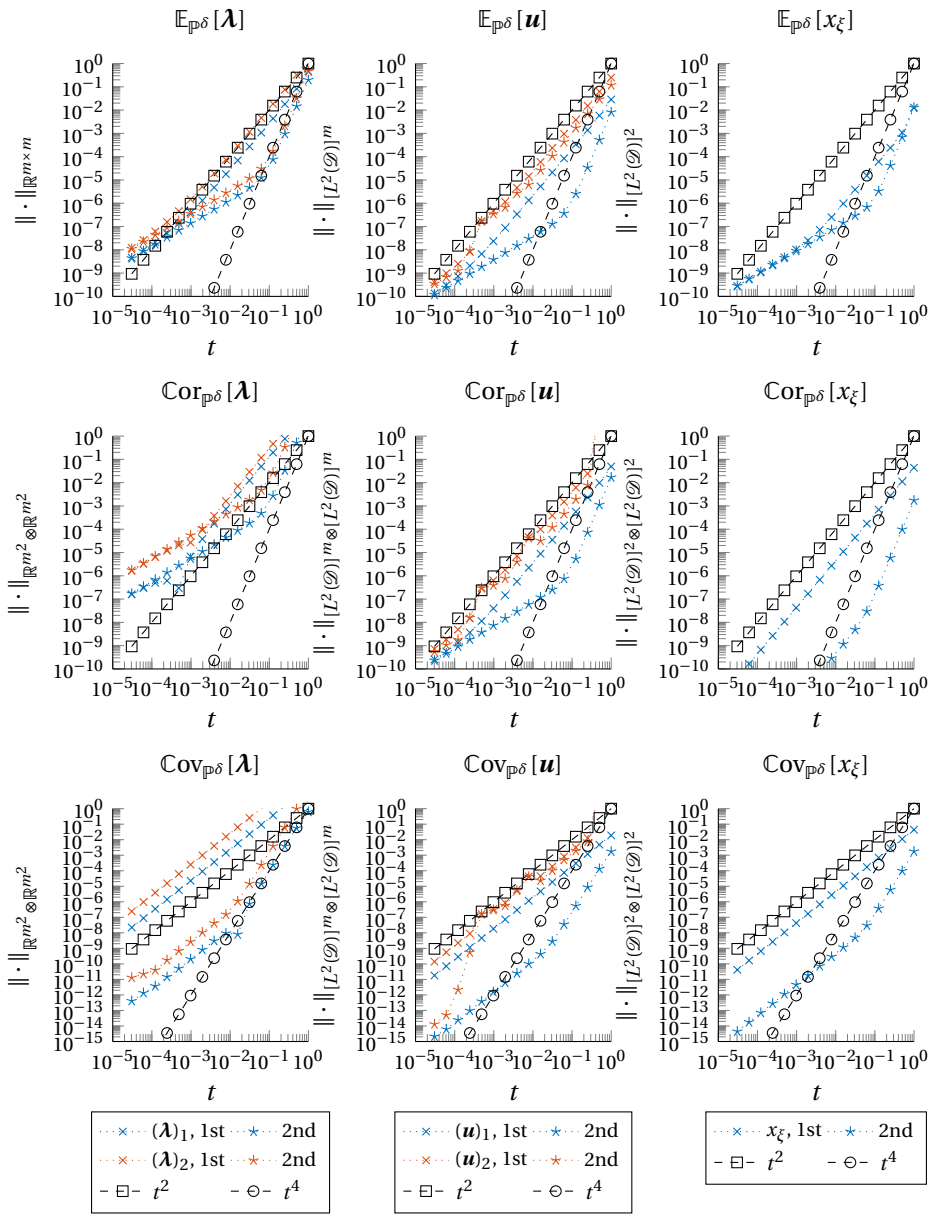

FIGURE VI.6. Convergence rates of centered approximations compared to QMC estimate for the Laplace EVP ($\mathscr{D} = (0,1)^2$) with coefficients perturbed by random fields.

Convergence of Iteration in the Parameter Space. Now we also test the iteration (VI.11). As discussed in section VI.4, since the random fields are decomposed in a truncated KLE, we can use the elementwise updates of (VI.18). For our test, we draw a sample of the random fields, calculate the perturbed eigenfunction $(u_{x_\xi})_1$ and draw a sample of the random noise to simulate the measurement data η. Using this realization η^δ, we perform 100 iterations for the same perturbation amplitudes as before. We stop the iteration if the norm of the update is suitably small, i.e., we have arrived at a fixed point. The norm of the updates is illustrated in fig. VI.7. It also shows the convergence rate of corollary VI.9 at the reference point after 100 iterations, compared to a QMC approximation of the posterior mean of x_ξ using 10^7 samples. The iterations converge for each perturbation amplitude, however, since the perturbation amplitude implicitly serves as a step size, the convergence is quite slow for small perturbation amplitudes. The predicted convergence rate of corollary VI.9 is confirmed, for perturbation amplitudes where the QMC method is accurate enough.

VI.5.2. One-dimensional Laplace Eigenvalue Problem. For the second example, we consider a modified version of example I.1. The one-dimensional Laplace EVP has the advantage that all its unperturbed eigenvalues are non-degenerate and locally Fréchet differentiable. We introduce a coefficient to example I.1 with $\mathscr{D} = (0,1)$, so that we need to find $(\lambda_x, u_x) \in \mathbb{R} \times H_0^1(\mathscr{D})$, $u \neq 0$, such that

$$\int_{\mathscr{D}} \mu_x(\mathbf{x}) \operatorname{grad} u_x(\mathbf{x}) \operatorname{grad} v(\mathbf{x}) \, \mathrm{d}\mathbf{x} = \int_{\mathscr{D}} u_x(\mathbf{x}) \, v(\mathbf{x}) \, \mathrm{d}\mathbf{x} \, \lambda_x \qquad \forall v \in H_0^1(\mathscr{D})$$

with the standard normalization $\|u_x\|_{L^2(\mathscr{D})} = 1$. The coefficient μ is given by the unperturbed state $\mu_{x_0} = 5$ and a linear perturbation

$$\mu = \mu_{x_0} + t\mu_1 \,,$$

where $t \geq 0$ and we let μ_1 be a random field given by a KLE similar to a Brownian bridge, i.e.,

$$\mu_1(\mathbf{x}, \omega) = \sum_{i=1}^{M} \frac{\sqrt{2}}{k\pi} \sin(k\pi\mathbf{x}) z_i(\omega)$$

that we truncate after $M = 100$ and let $z_i \sim \mathscr{U}([-\frac{1}{2}, \frac{1}{2}])$ iid. We choose this unperturbed value and z_i bounded to preserve the ellipticity of the bilinear form a. For this example, we consider the coefficient $\mu \in C([0,1])$ with its norm the maximum norm. Due to $\mu_0 = 5$, we can check that the unperturbed eigenfunctions are the same as in example I.1, and the eigenvalues are scaled to $\mu_0(n\pi)^2$ for $n \in \mathbb{N}$.
The unperturbed eigenfunctions, a sampled perturbed solution, and its first-order Taylor approximation are illustrated in fig. VI.8. We use a standard FE approximation with piecewise linear basis functions on 501 equidistant nodes.

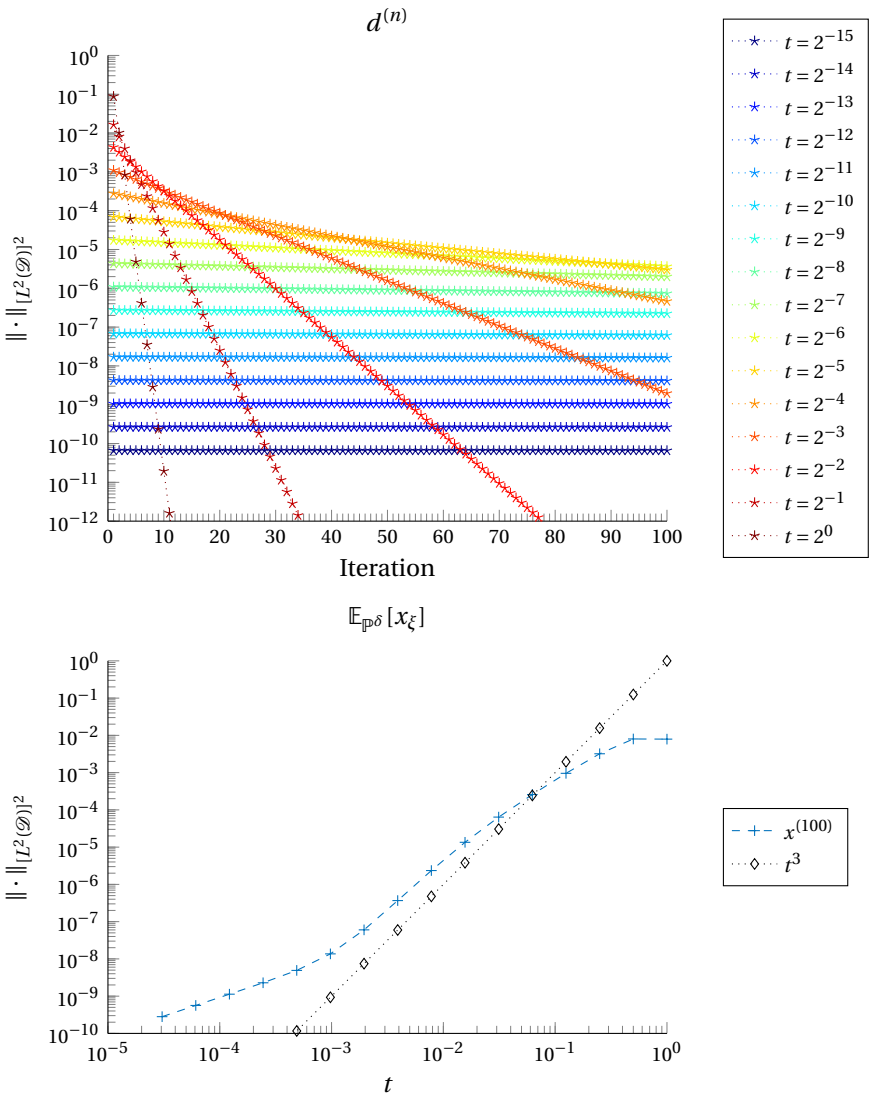

FIGURE VI.7. Norm of updates $d^{(n)}$ for 100 iterations for the Laplace EVP ($\mathscr{D} = (0,1)^2$) and convergence rate of $x^{(100)}$ compared to QMC estimate of posterior mean.

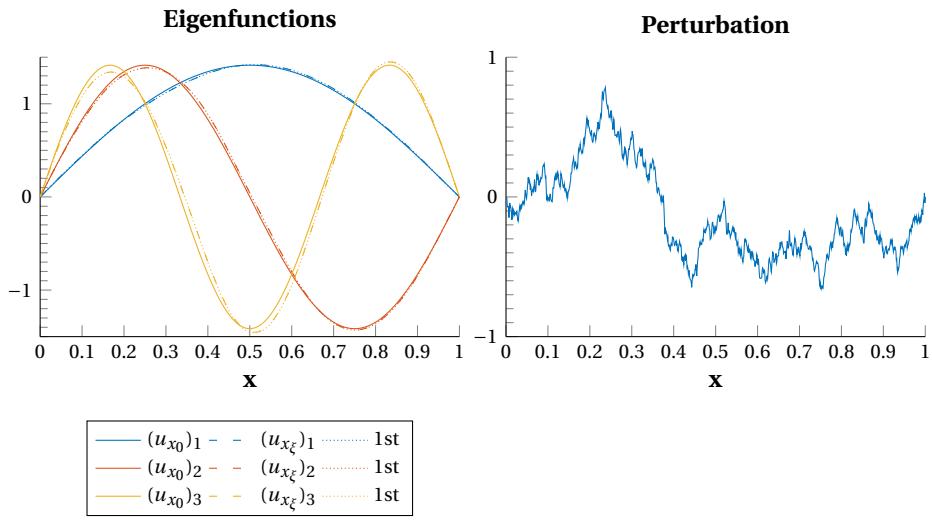

FIGURE VI.8. Unperturbed eigenfunctions, sample of the perturbation, and perturbed solution with first-order Taylor approximation of the Laplace EVP ($\mathscr{D} = (0,1)$).

Consider the function $g(\lambda) := \lambda^{-\frac{1}{2}}$ with Fréchet derivative given by

$$\mathbf{D}_{x_0} g = -\tfrac{1}{2}(\mathbf{D}_{x_0}\lambda)\lambda_{x_0}^{-\frac{3}{2}} .$$

As the measurement for this example, we consider $g(\lambda)$ corresponding to the three smallest eigenvalues λ, i.e., $K = 3$, $\mathtt{G} : X \to Y = \mathbb{R}^3$. Let the measured data be given by

$$\eta^\delta = \begin{bmatrix} 1.5000 \\ 0.0700 \\ 0.0480 \end{bmatrix}$$

and assume a covariance of the noise ε according to

$$\Sigma = \sigma\Sigma_0 , \qquad \Sigma_0 = \begin{bmatrix} 1 & \frac{1}{5} & \frac{1}{5} \\ \frac{1}{5} & 1 & \frac{1}{5} \\ \frac{1}{5} & \frac{1}{5} & 1 \end{bmatrix} ,$$

where we vary $\sigma \in \mathbb{R}$ to scale the measurement noise.
For brevity, we only investigate the posterior mean for $\mathtt{R} = x_\xi$ in the following experiments, consider its representation as a coefficient in $C^0(\mathscr{D})$, and measure the convergence rates using the maximum norm.

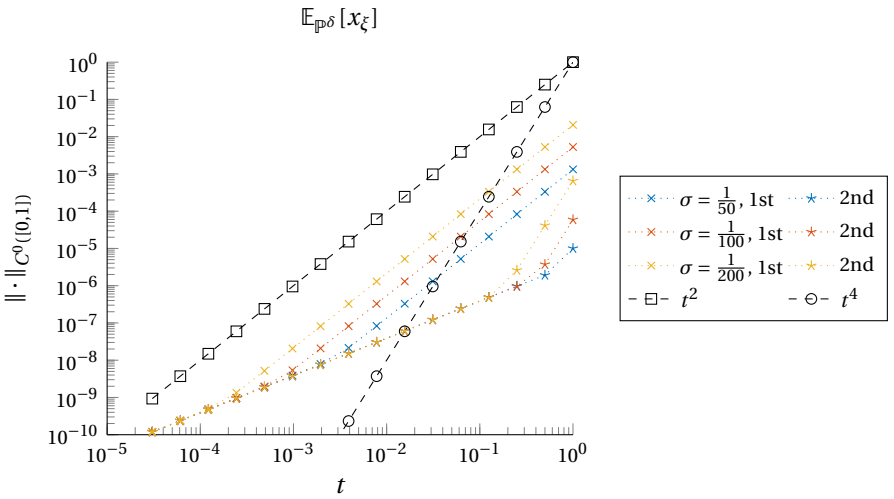

FIGURE VI.9. Convergence rates of centered approximations compared to QMC estimate for the Laplace EVP ($\mathscr{D} = (0,1)$) with perturbed coefficient and varied noise level σ.

Confirmation of Convergence Rates. We confirm the convergence rate for the centered approximation in fig. VI.9, where we compare the perturbation approximation against a QMC approximation using 10^7 samples for each perturbation amplitude. In order to illustrate the influence of the noise ε on the accuracy of the approximation, we repeat the experiment for $\sigma \in \left\{\frac{1}{50}, \frac{1}{100}, \frac{1}{200}\right\}$. We can see that the approximation realizes a larger error if the measurement noise is small, i.e., if the posterior measure becomes very concentrated. The calculation of the 10^7 samples was performed again on a single node of the Marvin cluster of the University of Bonn and took 5 minutes, using the same parallelization as before. The improvement in calculation speed compared to the previous example is partly due to the fact that we can skip the calculation of the perturbed eigenfunctions and subsequent transformation to the eigenspace.
Convergence of Iteration in Parameter Space. We also test the iteration (VI.11) for this second example. To this end, we again sample the perturbation and observation noise to get a representative sample of a measurement. For the iteration, we consider $\sigma = \frac{1}{100}$. Then, for the same perturbation amplitudes, we perform 100 iterations and test the convergence rate at $x^{(100)} \approx x^\star$ against a QMC estimate using a Halton sequence of length 10^7. In fig. VI.10 the size of the norm of the updates is illustrated, as well as the convergence rates predicted by corollary VI.9. Again, we can summarize that the iteration performs well for some intermediate perturbation amplitudes. For small perturbation amplitudes, the iteration converges quite slowly and for the

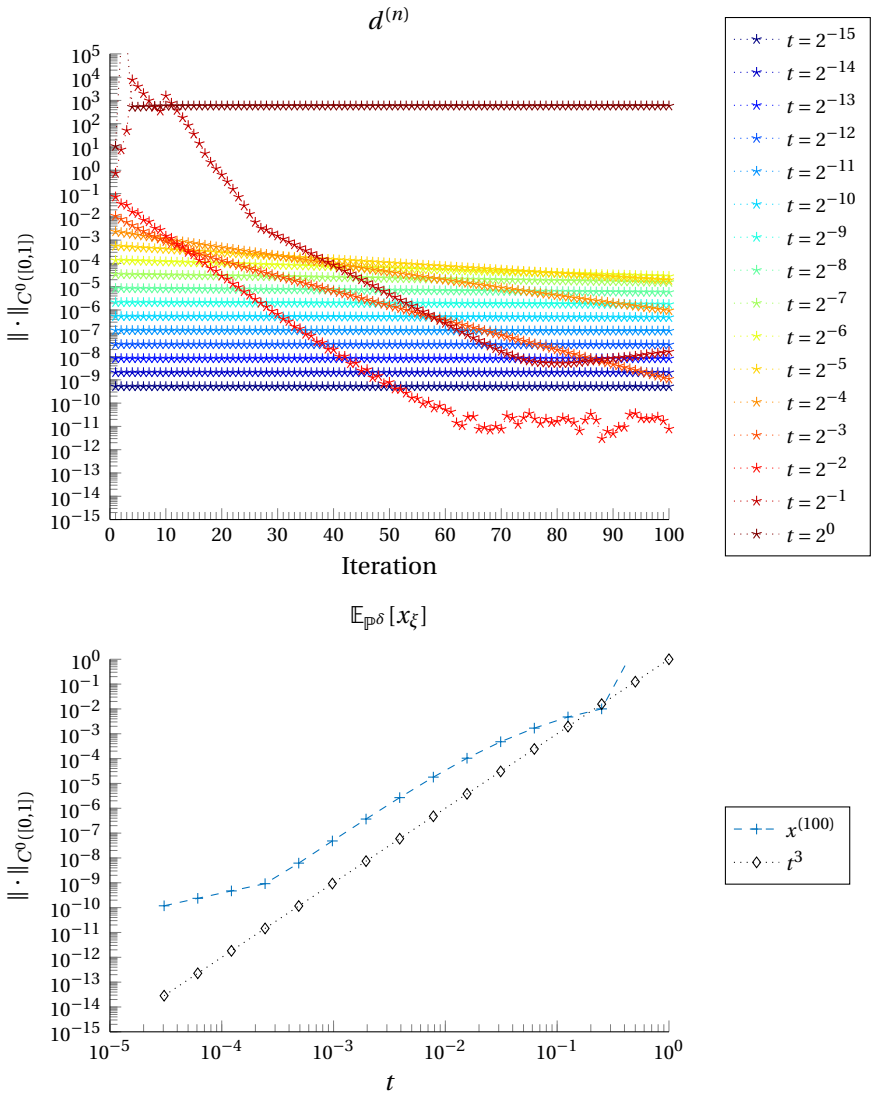

FIGURE VI.10. Norm of updates $d^{(n)}$ for 100 iterations for the Laplace EVP ($\mathscr{D} = (0,1)$) and convergence rate of $x^{(100)}$ compared to QMC estimate of posterior mean.

largest ones, the implied step size is too large, such that the iteration diverges. For intermediate step sizes, the predicted convergence rate is confirmed.

CHAPTER VII

Conclusion and Outlook

In this thesis, we have investigated the uncertainty quantification of eigenpairs of elliptic EVPs where the eigenvalues may be degenerate. The stochastic model may be given as a parameterized elliptic compact normal operator or in variational form by parameterized elliptic, continuous, and symmetric bilinear forms on a compact Gelfand triple.

VII.1. Eigenpair Trajectories and Their Regularity

Since stochastic models are often expressed by analytic dependence on a stochastic parameter, e.g., a KLE, we have discussed how eigenvalue trajectories can be identified in this setting. In a local neighborhood, where the eigenvalues remain non-degenerate, they are easy to isolate and analytic. The eigenpairs can be continued analytically on a pathwise basis to points where they become degenerate. However, this ansatz often leads to inconsistent matching of the eigenvalues across a multidimensional parameter space. In this sense, the eigenvalue trajectories can be considered only Gâteaux differentiable at points where they become degenerate.
We have therefore investigated the trajectories of the eigenpairs with respect to the eigenspace. The eigenspace of finite multiplicity at some reference point is thus fixed as a frame of reference. This ansatz preserves the analyticity of the EVP for the eigenpair trajectories, which is a useful property for numerical integration methods.
The eigenpair trajectories are not eigenpairs in a strict sense, as we define them by postulating an orthogonality condition on the derivatives of the eigenfunction trajectories. In turn, the eigenvalue trajectories are in general not diagonal.
In order to express the relationship between the trajectories with respect to the eigenspace and otherwise, we have characterized a parameterized polarization matrix, which in general can only be defined with respect to a parameter path, i.e., in the sense of Gâteaux derivatives. The main reason why this generally only works pathwise is that it is generally not possible to find a unique continuation of the eigenfunction trajectories at the point where their eigenvalues become degenerate. However, in the special case where the eigenvalues are uncoupled, such a basis as well as Fréchet differentiable eigenpair trajectories can be found.

Even when only pathwise defined, our two-stage characterization of the eigenvalue derivatives via the eigenspace and a subsequent polarization constitutes a generalization of previous methods to determine eigenvalue derivatives, which needed first- or second-order derivatives of the eigenvalues to be distinct when the eigenvalues are degenerate. We have characterized the derivatives of the eigenpairs with respect to the eigenspace and the derivatives of the polarization (for a given parameter path) by linear conditions. The derivatives of the former are given in saddle point equation form, which is convenient to prove unique solvability and to implement their calculation in variational form.

VII.2. Perturbation-based Uncertainty Quantification

Since the trajectories with respect to the eigenspace are, in general, only defined locally, and to make use of the linear characterization of eigenvalue derivatives, we have primarily investigated perturbation methods for the approximation of stochastic moments of these eigenpair trajectories. Extending previously available results, we stated first- and second-order approximations of the first and second stochastic moment, i.e. mean, correlation, and covariance, for the trajectories of eigenpairs with respect to the eigenspace. We also considered the possibility of incorporating measurement data into the stochastic model using Bayesian inversion. To this end, the perturbation approximations were augmented by additional terms to reflect the influence of the likelihood of the measurements on the posterior stochastic moments. These perturbation approximations of the posterior moments are expressed in terms of the prior distribution and have the same convergence rates as the perturbation approximations of the prior stochastic moments. Given some stronger assumptions on the perturbations, the second-order perturbation approximations of prior and posterior moments improve to third-order approximations.
Given the Bayesian inversion model, we considered the possibility of iteratively using the approximation of the posterior mean of the parameter to find a better reference point. It should be stressed that, for degenerate eigenvalues, this often leads to a different reference frame with respect to the eigenspace. Nevertheless, a correspondence can be found to an associated regularized inverse problem.
As examples of elliptic EVPs, we discuss the Laplace and Maxwell EVP. In addition to the possibility of including stochastic behavior through perturbation of material coefficients, we also discuss stochastic shape deformations of the Laplace and Maxwell EVP. These can be characterized equivalently by stochastic deformation coefficients which fit into the existing framework. We tested the convergence rates of the perturbation approximations given stochastic coefficient and shape deformations compared to GL and QMC approximations. This confirmed the usual property of perturbations to outperform other state of the art methods for high-dimensional parameter spaces and small scale perturbations.

VII.3. Outlook

Reflecting on our reliance on the spectral theorem of compact normal operators and ellipticity, it may also be possible to extend our results to more general EVPs in the future. For example, it might be possible to transfer some of our results to eigenpairs of compact non-normal operators or their Jordan normal form if defects occur. The derivatives for non-symmetric real matrices were already included in [**68**] and it might be possible to generalize these results.
Furthermore, we hope that the contributions regarding the linear characterization of Fréchet derivatives of eigenpairs are useful to future investigations of eigenpair trajectories. In particular, the parameterized polarization matrix and its derivatives, which were introduced and characterized in this thesis, may be a useful tool for understanding the regularity of degenerate eigenpairs. So far, degenerate eigenvalues have often been explicitly excluded in articles investigating the uncertainty quantification of eigenpairs. For practical use of the discussed results on the bifurcation behavior of eigenpair trajectories and their regularity, it may be possible to arrive at more concrete predictions by applying them to more specific operators, domains, and perturbations. If eigenpair trajectories for certain stochastic EVP models can be proven to be regular, this property can be used for perturbation approximations. However, this is also useful for various other numerical methods relying on regularity, e.g., sparse grids and PC expansions, which were already successfully applied to non-degenerate eigenvalues. To this end, it would also be useful to expand the local analyticity discussed in this thesis to analyticity in a global sense of the parameter space.
In this thesis, we identified eigenpair trajectories via (pathwise) analyticity. It might be worthwhile to instead consider other ways in which eigenpairs can be identified. For real eigenvalues, we could fix their order throughout the parameter space to consider the smallest or largest eigenvalue if they are well defined, respectively. Lastly, we could also consider deferred variables of interest, such as absolute values of eigenfunctions or spectral gaps of eigenvalues. For some applications, such approaches may be more suitable than strict adherence to analyticity.

Bibliography

[1] R. A. Adams and J. J. F. Fournier, *Sobolev Spaces*, vol. 140 of Pure and Applied Mathematics, Academic Press, New York, 2nd ed., 2003.

[2] S. Adhikari and M. I. Friswell, *Random matrix eigenvalue problems in structural dynamics*, International Journal for Numerical Methods in Engineering, 69 (2007), pp. 562–591.

[3] M. Alghamdi, D. Boffi, and F. Bonizzoni, *A greedy MOR method for the tracking of eigensolutions to parametric elliptic PDEs*, Journal of Computational and Applied Mathematics, 457 (2025), p. 116270.

[4] H. W. Alt, *Linear Functional Analysis*, Springer, London, 2016.

[5] R. Andreev and C. Schwab, *Sparse tensor approximation of parametric eigenvalue problems*, Numerical Analysis of Multiscale Problems, 83 (2012), pp. 203–241.

[6] B. Aune, R. Bandelmann, D. Bloess, B. Bonin, A. Bosotti, M. Champion, C. Crawford, G. Deppe, B. Dwersteg, D. A. Edwards, H. T. Edwards, M. Ferrario, M. Fouaidy, P.-D. Gall, A. Gamp, A. Gössel, J. Graber, D. Hubert, M. Hüning, M. Juillard, T. Junquera, H. Kaiser, G. Kreps, M. Kuchnir, R. Lange, M. Leenen, M. Liepe, L. Lilje, A. Matheisen, W.-D. Möller, A. Mosnier, H. Padamsee, C. Pagani, M. Pekeler, H.-B. Peters, O. Peters, D. Proch, K. Rehlich, D. Reschke, H. Safa, T. Schilcher, P. Schmüser, J. Sekutowicz, S. Simrock, W. Singer, M. Tigner, D. Trines, K. Twarowski, G. Weichert, J. Weisend, J. Wojtkiewicz, S. Wolff, and K. Zapfe, *Superconducting TESLA cavities*, Physical Review Special Topics - Accelerators and Beams, 3 (2000), p. 092001.

[7] H. Benaroya and M. Rehak, *Finite element methods in probabilistic structural analysis: A selective review*, Applied Mechanics Reviews, 41 (1988), pp. 201–213.

[8] M. Benzi, G. H. Golub, and J. Liesen, *Numerical solution of saddle point problems*, Acta Numerica, 14 (2005), pp. 1–137.

[9] D. Boffi, *Finite element approximation of eigenvalue problems*, Acta Numerica, 19 (2010), pp. 1–120.

[10] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, vol. 15 of Texts in Applied Mathematics, Springer, New York, 2008.

[11] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, no. 15 in Springer Series in Computational Mathematics, Springer, New York, 1991.

[12] H.-J. Bungartz and M. Griebel, *Sparse grids*, Acta Numerica, 13 (2004), pp. 147–269.

[13] J. C. Butcher, *Numerical Methods for Ordinary Differential Equations*, Wiley, Chicester, 2nd ed., 2008.

[14] R. E. Caflisch, *Monte Carlo and quasi-Monte Carlo methods*, Acta Numerica, 7 (1998), pp. 1–49.

[15] É. Cartan, *The Theory of Spinors*, Hermann, Paris, 1966.

[16] A. Chernov and C. Schwab, *First order k-th moment finite element analysis of nonlinear operator equations with stochastic data*, Mathematics of Computation, 82 (2013), pp. 1859–1888.

[17] C. Cohen-Tannoudji, B. Diu, and F. Laloë, *Quantum Mechanics.*, vol. 2, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, 2nd ed., 2019.

[18] J. D. COLLINS AND W. T. THOMSON, *The eigenvalue problem for structural systems with statistical properties.*, AIAA Journal, 7 (1969), pp. 642–648.
[19] R. L. DAILEY, *Eigenvector derivatives with repeated eigenvalues*, AIAA Journal, 27 (1989), pp. 486–491.
[20] C. DAVIS AND W. M. KAHAN, *The rotation of eigenvectors by a perturbation. III*, SIAM Journal on Numerical Analysis, 7 (1970), pp. 1–46.
[21] J. DICK, F. Y. KUO, AND I. H. SLOAN, *High-dimensional integration: The quasi-Monte Carlo way*, Acta Numerica, 22 (2013), pp. 133–288.
[22] J. A. DIEUDONNÉ, *Treatise on Analysis*, vol. 1 of Pure and Applied Mathematics, Academic Press, New York, 3rd ed., 1969.
[23] W. DÖRFLER, A. LECHLEITER, M. PLUM, G. SCHNEIDER, AND C. WIENERS, *Photonic Crystals: Mathematical Analysis and Numerical Approximation*, Springer, Basel, 2011.
[24] S. DUANE, A. KENNEDY, B. J. PENDLETON, AND D. ROWETH, *Hybrid Monte Carlo*, Physics Letters B, 195 (1987), pp. 216–222.
[25] J. DÖLZ AND D. EBERT, *On uncertainty quantification of eigenvalues and eigenspaces with higher multiplicity*, SIAM Journal on Numerical Analysis, 62 (2024), pp. 422–451.
[26] ———, *Local sensitivity analysis for Bayesian inverse problems*, Mar. 2025. Preprint available on `www.doi.org/10.48550/arXiv.2503.20526`. To appear in SIAM/ASA Journal on Uncertainty Quantification.
[27] J. DÖLZ, D. EBERT, S. SCHÖPS, AND A. ZIEGLER, *Shape uncertainty quantification of Maxwell eigenvalues and -modes with application to TESLA cavities*, Computer Methods in Applied Mechanics and Engineering, 428 (2024), p. 117108.
[28] J. DÖLZ, H. HARBRECHT, AND M. D. PETERS, *$\mathscr{H}$-matrix based second moment analysis for rough random fields and finite element discretizations*, SIAM Journal on Scientific Computing, 39 (2017), pp. B618–B639.
[29] J. DÖLZ, H. HARBRECHT, AND C. SCHWAB, *Covariance regularity and $\mathscr{H}$-matrix approximation for rough random fields*, Numerische Mathematik, 135 (2017), pp. 1045–1071.
[30] D. EBERT, *eigen_derivatives*. Zenodo, Oct. 2025. `www.doi.org/10.5281/ZENODO.17478322`.
[31] G. EVENSEN, *Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics*, Journal of Geophysical Research: Oceans, 99 (1994), pp. 10143–10162.
[32] M. I. FRISWELL, *The derivatives of repeated eigenvalues and their associated eigenvectors*, Journal of Vibration and Acoustics, 118 (1996), pp. 390–397.
[33] N. GEORG, W. ACKERMANN, J. CORNO, AND S. SCHÖPS, *Uncertainty quantification for Maxwell's eigenproblem based on isogeometric analysis and mode tracking*, Computer Methods in Applied Mechanics and Engineering, 350 (2019), pp. 228–244.
[34] R. GHANEM AND D. GHOSH, *Efficient characterization of the random eigenvalue problem in a polynomial chaos decomposition*, International Journal for Numerical Methods in Engineering, 72 (2007), pp. 486–504.
[35] R. GHANEM, D. HIGDON, AND H. OWHADI, eds., *Handbook of Uncertainty Quantification*, Springer International Publishing, Cham, 2017.
[36] A. D. GILBERT, I. G. GRAHAM, F. Y. KUO, R. SCHEICHL, AND I. H. SLOAN, *Analysis of quasi-Monte Carlo methods for elliptic eigenvalue problems with stochastic coefficients*, Numerische Mathematik, 142 (2019), pp. 863–915.
[37] A. D. GILBERT AND R. SCHEICHL, *Multilevel quasi-Monte Carlo for random elliptic eigenvalue problems I: Regularity and error analysis*, IMA Journal of Numerical Analysis, 44 (2024), pp. 466–503.
[38] ———, *Multilevel quasi-Monte Carlo for random elliptic eigenvalue problems II: Efficient algorithms and numerical results*, IMA Journal of Numerical Analysis, 44 (2024), pp. 504–535.
[39] M. B. GILES, *Multilevel Monte Carlo methods*, Acta Numerica, 24 (2015), pp. 259–328.

[40] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, Johns Hopkins Studies in the Mathematical Sciences, The Johns Hopkins University Press, Baltimore, 4th ed., 2013.

[41] C. GORDON, D. L. WEBB, AND S. WOLPERT, *One cannot hear the shape of a drum*, Bulletin of the American Mathematical Society, 27 (1992), pp. 134–138.

[42] S. GORGIZADEH, T. FLISGEN, AND U. VAN RIENEN, *Eigenmode computation of cavities with perturbed geometry using matrix perturbation methods applied on generalized eigenvalue problems*, Journal of Computational Physics, 364 (2018), pp. 347–364.

[43] L. GRUBIŠIĆ, M. SAARIKANGAS, AND H. HAKULA, *Stochastic collocation method for computing eigenspaces of parameter-dependent operators*, Numerische Mathematik, 153 (2023), pp. 85–110.

[44] H. HAKULA AND M. LAAKSONEN, *Multiparametric shell eigenvalue problems*, Computer Methods in Applied Mechanics and Engineering, 343 (2019), pp. 721–745.

[45] H. HARBRECHT, M. PETERS, AND R. SCHNEIDER, *On the low-rank approximation by the pivoted Cholesky decomposition*, Applied Numerical Mathematics, 62 (2012), pp. 428–440.

[46] H. HARBRECHT, M. PETERS, AND M. SIEBENMORGEN, *Combination technique based k-th moment analysis of elliptic problems with random diffusion*, Journal of Computational Physics, 252 (2013), pp. 128–141.

[47] ———, *Efficient approximation of random fields for numerical applications*, Numerical Linear Algebra with Applications, 22 (2015), pp. 596–617.

[48] ———, *Analysis of the domain mapping method for elliptic diffusion problems on random domains*, Numerische Mathematik, 134 (2016), pp. 823–856.

[49] H. HARBRECHT, R. SCHNEIDER, AND C. SCHWAB, *Multilevel frames for sparse tensor product spaces*, Numerische Mathematik, 110 (2008), pp. 199–220.

[50] W. K. HASTINGS, *Monte Carlo sampling methods using Markov chains and their applications*, Biometrika, 57 (1970), pp. 97–109.

[51] R. HIPTMAIR, *Finite elements in computational electromagnetism*, Acta Numerica, 11 (2002), pp. 237–339.

[52] M. D. HOFFMAN AND A. GELMAN, *The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo*, Journal of Machine Learning Research, 15 (2014), pp. 1593–1623.

[53] V. HOWLE AND L. N. TREFETHEN, *Eigenvalues and musical instruments*, Journal of Computational and Applied Mathematics, 135 (2001), pp. 23–40.

[54] S. JANSON AND S. KAIJSER, *Higher moments of Banach space valued random variables*, Memoirs of the American Mathematical Society, 238 (2015).

[55] P. JORKOWSKI, *Zur numerischen Berechnung von parametrischen und nichtlinearen Eigenwertproblemen in der elektromagnetischen Feldsimulation*, PhD thesis, TU Berlin, 2021.

[56] M. KAC, *Can one hear the shape of a drum?*, The American Mathematical Monthly, 73 (1966), pp. 1–23.

[57] J. KAIPIO AND E. SOMERSALO, *Statistical and Computational Inverse Problems*, Springer, New York, 2005.

[58] R. E. KÁLMÁN, *A new approach to linear filtering and prediction problems*, Journal of Basic Engineering, 82 (1960), pp. 35–45.

[59] T. KATŌ, *Perturbation Theory for Linear Operators*, Classics in Mathematics, Springer, Berlin Heidelberg, 1995.

[60] A. KLENKE, *Probability Theory*, Universitext, Springer International Publishing, Cham, 2020.

[61] L. KOCIS AND W. J. WHITEN, *Computational investigations of low-discrepancy sequences*, ACM Transactions on Mathematical Software, 23 (1997), pp. 266–294.

[62] G. J. LORD, C. E. POWELL, AND T. SHARDLOW, *An Introduction to Computational Stochastic PDEs*, Cambridge Texts in Applied Mathematics, Cambridge University Press, 2014.

[63] T. MACH AND M. A. FREITAG, *Solving the parametric eigenvalue problem by Taylor series and Chebyshev expansion*, SIAM Journal on Matrix Analysis and Applications, 46 (2025), pp. 957–983.

[64] N. G. MEYERS AND J. SERRIN, *$H = W$*, Proceedings of the National Academy of Sciences, 51 (1964), pp. 1055–1056.

[65] W. C. MILLS-CURRAN, *Calculation of eigenvector derivatives for structures with repeated eigenvalues*, AIAA Journal, 26 (1988), pp. 867–871.

[66] ———, *Comment on "Eigenvector Derivatives with Repeated Eigenvalues"*, AIAA Journal, 28 (1990), pp. 1846–1846.

[67] P. MONK, *Finite Element methods for Maxwell's equations*, Numerical Mathematics and Scientific Computation, Clarendon Press ; Oxford University Press, Oxford : New York, 2003.

[68] R. B. NELSON, *Simplified calculation of eigenvector derivatives*, AIAA Journal, 14 (1976), pp. 1201–1205.

[69] V. K. NGUYEN, *Analyticity of parametric elliptic eigenvalue problems and applications to quasi-Monte Carlo methods*, Complex Variables and Elliptic Equations, 69 (2024), pp. 1–21.

[70] H. NIEDERREITER, *Random Number Generation and Quasi-Monte Carlo Methods*, no. 63 in Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1992.

[71] J.-C. NÉDÉLEC, *Mixed finite elements in $\mathbb{R}^3$*, Numerische Mathematik, 35 (1980), pp. 315–341.

[72] ———, *A new family of mixed finite elements in $\mathbb{R}^3$*, Numerische Mathematik, 50 (1986), pp. 57–81.

[73] I. U. OJALVO, *Efficient computation of mode-shape derivatives for large dynamic systems*, AIAA Journal, 25 (1987), pp. 1386–1390.

[74] N. PERKINS AND C. MOTE, *Comments on curve veering in eigenvalue problems*, Journal of Sound and Vibration, 106 (1986), pp. 451–463.

[75] T. VON PETERSDORFF AND C. SCHWAB, *Sparse finite element methods for operator equations with stochastic data*, Applications of Mathematics, 51 (2006), pp. 145–180.

[76] K. B. PETERSEN AND M. S. PEDERSEN, *The matrix cookbook*, Nov. 2012. `http://www2.compute.dtu.dk/pubdb/pubs/3274-full.html`.

[77] D. PRADOVERA AND A. BORGHI, *Match-based solution of general parametric eigenvalue problems*, Journal of Computational Physics, 519 (2024), p. 113384.

[78] F. RELLICH, *Störungstheorie der Spektralzerlegung. I. Mitteilung. Analytische Störung der isolierten Punkteigenwerte eines beschränkten Operators*, Mathematische Annalen, 113 (1937), pp. 600–619.

[79] ———, *Störungstheorie der Spektralzerlegung. II. Mitteilung. Stetige Abhängigkeit der Spektralschar von einem Parameter*, Mathematische Annalen, 113 (1937), pp. 677–685.

[80] ———, *Störungstheorie der Spektralzerlegung. III. Mitteilung. Analytische, nicht notwendig beschränkte Störung*, Mathematische Annalen, 116 (1939), pp. 555–570.

[81] ———, *Störungstheorie der Spektralzerlegung. IV*, Mathematische Annalen, 117 (1940), pp. 356–382.

[82] ———, *Störungstheorie der Spektralzerlegung. V*, Mathematische Annalen, 118 (1941), pp. 462–484.

[83] ———, *Perturbation Theory of Eigenvalue Problems*, Notes on mathematics and its applications, Gordon and Breach Science Publishers Inc., New York, 1969.

[84] B. V. ROSIĆ, A. LITVINENKO, O. PAJONK, AND H. G. MATTHIES, *Sampling-free linear Bayesian update of polynomial chaos representations*, Journal of Computational Physics, 231 (2012), pp. 5761–5787.

[85] W. RUDIN, *Real and Complex Analysis*, McGraw-Hill, New York, NY, 3rd ed., 1987.

[86] Y. SAAD, *Numerical Methods for Large Eigenvalue Problems*, no. 66 in Classics in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia, rev. ed., 2011.

[87] A. SCHNEEBELI, *An $H(\mathrm{curl},\Omega)$-conforming FEM: Nédélec's elements of first type*, May 2003.

[88] J. SHAW AND S. JAYASURIYA, *Modal sensitivities for repeated eigenvalues and eigenvalue derivatives*, AIAA Journal, 30 (1992), pp. 850–852.

[89] M. SHINOZUKA AND C. J. ASTILL, *Random eigenvalue problems in structural analysis*, AIAA Journal, 10 (1972), pp. 456–462.

[90] R. C. SMITH, *Uncertainty Quantification: Theory, Implementation, and Applications*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2013.

[91] C. SOIZE, *Uncertainty Quantification: An Accelerated Course with Advanced Applications in Computational Engineering*, vol. 47 of Interdisciplinary Applied Mathematics, Springer International Publishing, Cham, 2017.

[92] J. STOER AND R. BULIRSCH, *Introduction to Numerical Analysis*, no. 12 in Texts in Applied Mathematics, Springer, New York, 2nd ed., 1996.

[93] A. M. STUART, *Inverse problems: A Bayesian perspective*, Acta Numerica, 19 (2010), p. 451–559.

[94] T. J. SULLIVAN, *Introduction to Uncertainty Quantification*, Springer International Publishing, Switzerland, 2015.

[95] J.-G. SUN, *Sensitivity analysis of multiple eigenvalues (I)*, Journal of Computational Mathematics, 6 (1988), pp. 28–38.

[96] W. T. THOMSON, *Theory of Vibration with Applications*, CRC Press, Boca Raton, 4th ed., Feb. 2018.

[97] L. TIERNEY AND J. B. KADANE, *Accurate approximations for posterior moments and marginal densities*, Journal of the American Statistical Association, 81 (1986), pp. 82–86.

[98] R. VÁZQUEZ, *A new design for the implementation of isogeometric analysis in Octave and Matlab: GeoPDEs 3.0*, Computers & Mathematics with Applications, 72 (2016), pp. 523–554.

[99] L. B. DA VEIGA, H. HAKULA, AND J. PITKÄRANTA, *Asymptotic and numerical analysis of the eigenvalue problem for a clamped cylindrical shell*, Mathematical Models and Methods in Applied Sciences, 18 (2008), pp. 1983–2002.

[100] M. J. WAINWRIGHT, *Principal component analysis in high dimensions*, in High-Dimensional Statistics: A Non-Asymptotic Viewpoint, Cambridge University Press, 1st ed., Feb. 2019.

[101] M. WILLIAMS, *A method for solving stochastic eigenvalue problems*, Applied Mathematics and Computation, 215 (2010), pp. 3906–3928.

[102] ———, *A method for solving stochastic eigenvalue problems II*, Applied Mathematics and Computation, 219 (2013), pp. 4729–4744.

[103] J. WLOKA, *Partial Differential Equations*, Cambridge University Press, Cambridge, 1987.

[104] S. YASAR, P. D. GALL, V. GUBAREV, J. IVERSEN, AND A. SULIMOV, *A database for the European XFEL*, in Proceedings of SRF2013, Paris, France, 2013, pp. 205–207.

[105] S. ZAGLMAYR, *High Order Finite Element Methods for Electromagnetic Field Computation*, PhD thesis, Johannes Kepler Universität Linz, 2006.

[106] E. ZEIDLER, *Fixed-point Theorems*, vol. 1 of Nonlinear Functional Analysis and Its Applications, Springer, New York, 1992.

[107] A. ZIEGLER, N. GEORG, W. ACKERMANN, AND S. SCHÖPS, *Mode recognition by shape morphing for Maxwell's eigenvalue problem in cavities*, IEEE Transactions on Antennas and Propagation, 71 (2023), pp. 4315–4325.

[108] A. ZIEGLER, M. MERKEL, P. GANGL, AND S. SCHÖPS, *On the computation of analytic sensitivities of eigenpairs in isogeometric analysis*, Computer Methods in Applied Mechanics and Engineering, 409 (2023), p. 115961.

[109] A. M. ZIEGLER, *Efficient Methods for Parameterized Eigenvalue Problems in Electromagnetism*, PhD thesis, TU Darmstadt, 2025.

Acronyms

a.e.: almost everywhere
a.s.: almost surely

B-spline: basis spline

CAD: computer-aided design

DESY: Deutsches Elektronen-Synchrotron

EVP: eigenvalue problem

FE: finite element

GL: Gauß–Legendre

IGA: isogeometric analysis
iid: independent and identically distributed
IVP: initial value problem

KLE: Karhunen–Loève expansion

LBB: Ladyzhenskaya–Babuška–Brezzi

MAP: maximum a posteriori
MC: Monte Carlo
MCMC: Markov chain Monte Carlo
ML: multilevel

NURBS: non-uniform rational B-splines

ODE: ordinary differential equation

PC: polynomial chaos
PDE: partial differential equation

QMC: quasi-Monte Carlo

SVD: singular value decomposition

TESLA: TeV-energy superconducting linear accelerator
TeV: teraelectronvolt

Symbols

$\hookrightarrow$: embedding, 17
$\ll$: absolutely continuous, 34
$\|\cdot\|_X$: norm of normed space X, 12
$\langle\cdot,\cdot\rangle_X$: scalar/inner product of inner product space X, 13
$\langle\cdot,\cdot\rangle_\Sigma$: scalar product induced by covariance matrix Σ, 56
$\frac{\mathrm{d}\mu}{\mathrm{d}\nu}$: Radon–Nikodým derivative, 34
$\times$: product, 11
$\cong$: isometrically isomorph, 17
$\cdot^*$: adjoint vector, 11 ; Hilbert adjoint, 21
$\grave{\cdot}$: polarized eigenvalue/ eigenfunction, 79 ; decision matrix (polarization), 82
$\bar{\cdot}$: conjugate, 11 ; closure, 12
$\cdot^\perp$: orthogonal complement, 14
$\cdot^\top$: transposed, 11
$\oplus$: direct sum, 14
$\partial\mathscr{D}$: boundary of $\mathscr{D}$, 12
$\partial^{\boldsymbol{\alpha}}$: partial derivative with multi-index $\boldsymbol{\alpha}$, 22
$\langle\cdot,\cdot\rangle_X$: inner product of inner product space X, 13
$\otimes$: algebraic tensor product, 15 ; Kronecker product, 15

$\mathbb{1}$: indicator function, 33

$B(x)$: neighborhood of x, 12
$B_r(x)$: ball with radius r and center x, 12
$\mathscr{B}(Y)$: Borel σ-algebra, 31

$A^\complement$: complement of A, 31
C : continuous functions, 27
C^0 : continuous functions, 27
C_0^∞ : smooth functions with compact support, 35
C^m : m-times continuously Fréchet differentiable functions, 26
C^∞ : smooth functions, 27
$\mathbb{C}\mathrm{or}_\mathbb{P}$: correlation (resp. probability measure $\mathbb{P}$), 54
$\mathbb{C}\mathrm{ov}_\mathbb{P}$: covariance (resp. probability measure $\mathbb{P}$), 54
C^ω : analytic functions, 28
Curl : curl $\mathbb{R}\to\mathbb{R}^2$, 50
curl : curl $\mathbb{R}^2\to\mathbb{R}$, 50
curl : curl $\mathbb{R}^3\to\mathbb{R}^3$, 47

$\mathscr{D}$: domain, 12
δ_{ij} : Kronecker delta, 3

$\mathbb{E}_\mathbb{P}$: mean/expected value (resp. probability measure $\mathbb{P}$), 54
η^δ : data (realization), 146
η : data (random variable), 146

$\mathtt{G}$: forward response map, 145
grad : gradient, 24 ; gradient, 46

$H^r(\mathscr{D})$: Sobolev space $W^{r,2}$, 36
$H_0^r(\mathscr{D})$: Sobolev space $W_0^{r,2}$, 36
$H(\mathbf{curl};\mathscr{D})$: , 49
$H_0(\mathbf{curl};\mathscr{D})$: , 49

$\boldsymbol{I}$: identity matrix, 12
i : imaginary unit ($\mathrm{i}=\sqrt{-1}$), 11
Id : identity operator, 12
$\Im$: imaginary part, 11
intr : interior, 12

$\mathbb{K}$: $\mathbb{R}$ or $\mathbb{C}$, 11
$\mathscr{K}$: compact (linear) operator, 18

L : (Bochner–Lebesgue) integrable functions, 33
$\mathscr{L}^{(1.5)}(X;Y)$: bounded sesquilinear form, 18
$\mathscr{L}^{(n)}(X;Y)$: bounded n-linear form, 18

Index