Title

Hessian Approximations for Large-Scale Inverse Problems Governed By Partial Differential Equations

Author

Vuchkov, Radoslav G.

Publication Date

2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA, MERCED

Hessian Approximations for Large-Scale Inverse Problems Governed By Partial Differential Equations

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Applied Mathematics

by

Radoslav G. Vuchkov

Committee in charge:

Professor Noemi Petra, Chair
Professor Boaz Ilan
Dr. Cosmin G. Petra
Professor Roummel Marcia

2022

The dissertation of Radoslav G. Vuchkov is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

(Professor Boaz Ilan)

(Dr. Cosmin G. Petra)

(Professor Roummel Marcia)

(Professor Noemi Petra, Chair)

University of California, Merced

2022

DEDICATION

To my wife and family.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to give my sincere gratitude to my advisor, Prof. Noemi Petra, for giving me the opportunity to work on many interesting problems. Throughout my graduate studies, Noemi has provided excellent advice, encouragement, and guidance, as well as opportunities to support my research and professional development. I would also like to thank my committee members, Prof. Boaz Ilan, Dr. Cosmin G. Petra, and Prof. Roummel Marcia, for serving as my committee members. Their careful reading of this thesis and many great suggestions have been critical to the success of this thesis. I am grateful to my internship mentors; Dr. Cosmin G. Petra from Lawrence Livermore National Laboratory, Dr. Boian Alexandrov, Dr. Gianmarco Manzini, and Dr. Duc Truong from Los Alamos National Laboratory, for the opportunities and support they provided. Furthermore, it has been a pleasure to collaborate with Prof. Ekkehard Sachs, Prof. Umberto Villa, and Prof. Ruanui Nicholson who have helped me craft a portion of this thesis and become a better mathematician. I want to thank my fellow graduate students and in particular Tucker Hartland and Prof. Johannes Brust.

Lastly, I would like to mention my appreciation to the Applied Mathematics Department and the Research and Training Grant (RTG) for funding the research of this dissertation in the form of the RTG Fellowship.

VITA

2013	B. S. in Mathematics, California State University Monterey Bay
2016	M. A. in Mathematics, San Francisco State University
2022	Ph. D. in Applied Mathematics, University of California, Merced

PUBLICATIONS

R. G. Vuchkov, C. G. Petra, and N. Petra,“*A Derivation of mesh-independent secant quasi-Newton formulas for optimization in function spaces*”, Journal of Numerical, Functional Analysis and Optimization, 2020.

R. G. Vuchkov, R. Nicholson, U. Villa, N. Petra,“*Variance reduction for the Bayesian approximation error (BAE) with application to the Stokes ice sheet model under uncertain thermal distribution*” (In preparation).

B. Alexandrov, G. Manzini, E. Skau, D. P. Truong and R. Vuchkov, “*Challenging the curse of dimensionality in multidimensional numerical integration by using a low-rank tensor-train format*” (In preparation)

E. Sachs, R. G. Vuchkov, and N. Petra, “*Inexact-Hessian-vector products for reduced space PDE-constrained optimization*” (In preparation).

G. Manzini, D. P. Truong, R. Vuchkov, and B. Alexandrov, “*A tensor-train low-rank mimetic finite difference method for three-dimensional Maxwell wave propagation problems*” (To be submitted - Technical Report LA-UR-21-24664).

ABSTRACT OF THE DISSERTATION

Hessian Approximations for Large-Scale Inverse Problems Governed By Partial Differential Equations

by

Radoslav G. Vuchkov
Doctor of Philosophy in Applied Mathematics

University of California, Merced, 2022

Professor Noemi Petra, Chair

Inverse problems abound in all areas of science, engineering, and beyond. These can be seen as tools that can be used to refine mathematical models using measurement data. Here by refine we mean, estimate unknown or uncertain input parameters that cannot directly be measured. This is an important task, since the quality and predictability of the mathematical models relies on the ability to estimate these parameters as accurately as possible. In this thesis, we focus on a particular class of inverse problems, namely on inverse problems governed by partial differential equations (PDEs). These inverse problems are formulated as nonlinear least squares optimization problems constrained by PDEs. The major part of the thesis is devoted to developing efficient computational strategies to solve these optimization problems. To this end, we focus on derivative-based optimization methods, *e.g.*, quasi-Newton and Newton. The first- and second-order (when applicable) derivative information is derived using adjoint-based techniques. For quasi-Newton, we provide a new derivation of well-known quasi-Newton formulas in an infinite-dimensional Hilbert space setting. We show numerical results that demonstrate the desired mesh-independence property and superior performance of the resulting quasi-Newton methods. For Newtons' method, we aim to reduce the computational cost (measured in PDE solves) per Newton iteration. There are a number of existing approaches in the literature that target this goal. For instance, via efficient preconditioning of the underlying Newton system, inexact Newton-CG solves, via low-rank approximations of the second-order derivative (Hessian) of the optimization objective, and via inexact Hessian-vector products (*i.e.*, inexact second-order adjoint solves). In this thesis we focus on the latter and derive bounds for tolerances for inexact PDE solves required by the Hessian apply. We apply these tolerances for an inverse problem governed by a Poisson problem and show that relaxing the Hessian apply can lead to an overall reduced number of PDE solves.

In the last part of the thesis, we go beyond a deterministic setup and quantify the uncertainties in the solution of inverse problems. To this end, we adopt the framework of Bayesian inference which allows us to systematically take into account noisy observations, uncertain models and prior knowledge about the unknown. The problem of interest is the estimation of the basal sliding coefficient field for an uncertain thermally-dependent nonlinear Stokes ice sheet model. The novelty in this inverse problem is the uncertainty in the forward model in addition to the uncertain basal sliding coefficient field. This additional uncertainty stems from the unknown temperature distribution within the ice, which is dictated by both the unknown thermal conductivity and unknown geothermal heat flux. To account for model uncertainties, we use the Bayesian approximation error (BAE) approach combined with a variance reduction technique. Preliminary results indicate that the BAE approach can be used to account for model uncertainties induced by the unknown thermal properties of the ice, and that failure to take into account these uncertainties can lead to erroneous estimates. In addition, we show that BAE combined with a variance reduction technique has the potential to reduce the offline costs of the BAE approach.

Chapter 1

Introduction

Mathematical models play a central role in all areas of sciences (e.g., physical, engineering, social sciences). These can help analyze and simulate systems and to ultimately make predictions. Models can take many forms, including dynamical systems, differential equations, statistical models, etc. In this thesis, we will focus on physics-based models governed by partial differential equations (PDEs). Once a mathematical model is designed and validated (typically against idealistic experimental setup and results), the next question is how to solve the underlying mathematical problem or equations. In other words, the problem at hand becomes: given inputs (e.g., source terms, coefficients fields, initial and/or boundary conditions, geometry, etc.) solve the underlying equations to obtain a specific output. This is the so-called forward (or state) problem. When one derives (or defines) the underlying mathematical equations, one needs to decide on the values for these inputs. While some of these input parameters may be known (i.e., can be directly observed or measured), often these are unknown or uncertain. The quality and predictability of the mathematical models depends on the ability to better estimate these parameters. When observations are available, these parameters can be inferred via solving an *inverse problem.*

Inverse problems come with various mathematical and computational challenges. The most typical one is the fact that these problems are ill-posed, that is, their solution is not unique and is highly sensitive to errors in the observations. The remedy for ill-posedness is regularization. Defining a suitable regularization is a modeling choice and will affect the solution of the inverse problem. Therefore care must be taken to properly choose this term for the inverse problem. The second challenge is the numerical solution for these problems. Inverse problems are typically formulated as nonlinear least squares minimization problems, in which the cost (or the objective) function is composed of a so-called misfit term (between the solution of the forward problem and observations) and a regularization term. One can solve this optimization problem with various methods. However, the dimension of optimization problems governed by PDEs are typically large-scale stemming from discretization. Therefore, efficient solution methods will require at least the first derivative of the cost function. For inverse problems governed by PDEs, deriving (and computing) derivatives is challenging. Typically, however, adjoint-based methods, the approach taken in this thesis, give a systematic means to derive derivatives. In addition, when solving an inverse problem governed by PDEs, the forward problem will need to be solved several times. Therefore, solving the inverse problem inherits all the computational challenges the forward problem has. Finally, the solution of (deterministic) inverse problems are highly sensitive to the noise in measurements and modeling errors. Therefore, it is not sufficient to solve the inverse problem, but one needs to quantify the uncertainties in its solution. To do so, we model the unknown as a random variable, which leads to a statistical inverse problem. The resulting solution to the statistical inverse problem is a posterior distribution. In this thesis, we adopt the framework of Bayesian inference. The deterministic inverse problem will be a subproblem in a Bayesian inverse problem, therefore, the algorithmic developments for Bayesian inversion will inherit all the computational challenges of the deterministic inverse and forward problems and have the additional challenge to explore possibly high-dimensional distributions.

In this thesis, we focus on computational methods for inverse problems governed by PDEs. In particular we focus on reducing the computational cost (measured in forward PDE solves) for second-order (and second-order-like) methods, such as Newton and quasi-Newton methods. In addition, in this thesis we also focus on inverse problems governed by uncertain forward models. In particular we account for model uncertainty additional to the uncertainty in inversion parameters and propose a technique to reduce the computational cost in the process.

Organization of the Dissertation. The overall structure of this dissertation is divided into: background material (Chapter 2), research contributions (Chapters 3-5), and conclusion and future work (Chapter 6). The chapters in this dissertation are organized as follows:

- **Chapter 2: Background: Large-Scale Inverse Problems governed by PDEs**

  In this section, we discuss the general formulation of inverse problems governed by PDEs, provide background on optimization methods to solve such problems, and introduce notations.

- **Chapter 3: Quasi-Newton Formulas for Optimization in Function Spaces**

  In this section, we discussed quasi-Newton alternatives to Newton's method, which are preferred when solving optimization problems due to its superior convergence properties. In our work, we provide a new derivation of well-known quasi-Newton formulas in an infinite-dimensional Hilbert space setting. It is known that quasi-Newton update formulas are solutions to certain variational problems over the space of symmetric matrices. In this thesis, we formulate similar variational problems over the space of bounded symmetric operators in Hilbert spaces. By changing the constraints of the variational problem we obtain updates (for the Hessian and Hessian inverse) not only for the Broyden-Fletcher-Goldfarb-Shanno (BFGS) quasi-Newton method but also for Davidon–Fletcher–Powell (DFP), Symmetric Rank One (SR1), and Powell-Symmetric-Broyden (PSB). In addition, for an inverse problem governed by a partial differential equation (PDE), we derive DFP and BFGS "structured" secant formulas that explicitly use the derivative of the regularization and only approximates the second derivative of the misfit term. We show numerical results that demonstrate the desired mesh-independence property and superior performance of the resulting quasi-Newton methods. This work was published in the Journal of Numerical Functional Analysis and Optimization [3].

- **Chapter 4: Second-Order Adjoints in Inexact Hessian-Vector Products**
  In this section, we focus on Newton's method to solve optimization problems governed by PDEs. Second-order, Newton-like algorithms exhibit convergence properties superior to gradient-based or derivative-free optimization algorithms. However, deriving and computing second-order derivatives– needed for the Hessian-vector product in a Krylov iteration for the Newton step– often is not trivial. Second-order adjoints provide a systematic and efficient tool to derive second derivative information. In this chapter of the thesis, we show that the efficiency of an inexact Newton-Conjugate Gradient (CG) approach to solve inverse problems governed by PDEs can be improved with inexact Hessian-vector products using approximate second-order adjoint solves. We showed numerical results for an inverse problem governed by an elliptic PDE. In particular we show that one can relax the tolerance of the second-order adjoint solves, which leads to reducing the number of inner CG iterations and overall computational effort. This is joint work with Prof. Ekkehard Sacks and Prof. Noemi Petra.

- **Chapter 5: Variance reduction for the Bayesian approximation error (BAE) with application to the Stokes ice sheet model under uncertain thermal distribution**
  In this chapter, we considered the problem of estimating the basal sliding coefficient field for an uncertain thermally-dependent nonlinear Stokes ice sheet model based on synthetic surface velocity measurements. The uncertainty in the forward model stems from the unknown temperature distribution within the ice which is dictated by both the unknown thermal conductivity and unknown geothermal heat flux. The estimation problem is considered within the Bayesian framework which allows for the incorporation and subsequent quantification of uncertainty. The Bayesian approximation error (BAE) approach is employed to simultaneously premarginalise over both the model uncertainties and measurement errors. The basal sliding parameter can then

be estimated independently of the thermal parameters. To reduce the computational costs associated with the BAE approach, we propose a linear Taylor-based control variate to reduce the variance of the model error. To quantify the posterior uncertainty in the basal sliding parameter we employ a local Gaussian approximation to the posterior centered at the maximum a posteriori (MAP) basal sliding coefficient estimate. Computation of the MAP estimate is carried out using inexact Newton-CG for which the gradient and (action of) the Hessian are computed using the adjoint method. The performance and computational costs of the BAE approach are assessed on a two-dimensional test problem taken from the Ice Sheet Model Intercomparison Project for Higher-Order Ice Sheet Models (ISMIP-HOM) benchmark study. We pay particular attention to the feasibility of the estimates, i.e., how well the posterior density supports the truth, as well as the computational costs associated with carrying out the premarginalisation. Our results indicate that the BAE approach can be used to account for model uncertainties induced by the unknown thermal properties of the ice, and that failure to take into account these uncertainties can lead to erroneous estimates. Furthermore, preliminary results show that the proposed control variate has the potential to reduce the offline costs of the BAE approach. This is joint work with Prof. Ruanui Nicholsons, Prof. Umberto Villa, and Prof. Noemi Petra.

- **Chapter 6: Conclusion and Potential Future Research Directions**
  In this chapter, we summarize the contributions of this thesis, discuss research work tangential to the thesis but parallel to possible future work, and future research directions.

Chapter 2

Background: Large-Scale Inverse Problems Governed by Partial Differential Equations (PDEs)

In what follows, we provide a brief introduction to the deterministic and Bayesian formulation of inverse problems and their connection to PDE-constrained optimization.

2.1 Deterministic Inverse Problems

In model parameter inversion (*i.e.,* inverse problems governed by partial differential equations), we seek to reconstruct an unknown spatially varying parameter field m from measurements d of a forward (or state) variable u that depend on the parameter implicitly through the solution of the underlying (forward) model. In this work we will assume the forward (state) equation is described by a PDE unless otherwise stated. Mathematically the inverse problem can be written as

$$d = \mathcal{F}(m) + \eta, \tag{2.1}$$

where $\mathcal{F} : \mathcal{M} \to \mathbb{R}^q$ is the so-called *parameter-to-observable* map, and the additive noise η can be modeled as $\eta \sim \mathcal{N}(0, \Gamma_{noise})$ [82]. The domain of the mapping $\mathcal{M}$ is assumed to be compact and unless otherwise specified $\mathcal{M} \subseteq L^2(\mathcal{D})$, where $\mathcal{D} \subset \mathbb{R}^d$, with $d \in \mathbb{N}$.

Evaluating $\mathcal{F}$ is equivalent to solving the underlying forward problem for a given parameter m, followed by the application of an observation operator $\mathcal{B}$ to extract the solution at measurement locations. To be more precise

$$\mathcal{F}(m) = \mathcal{B}u \quad \text{s.t} \quad r(u, m) = 0, \tag{2.2}$$

where $u \in \mathcal{V}$ is the solution of the forward (state) problem, $\mathcal{V}$ is a Hilbert space, $\mathcal{B} : \mathcal{V} \to \mathbb{R}^q$ is the observation operator, which maps from the state space to the measurement data space, and $r : \mathcal{M} \times \mathcal{V} \to \mathcal{V}^*$ represents the PDE. A common problem that arises is when the data d is sparse and there are multiple parameters that can fit the model, leading to an ill-posed problem in the sense of Hadamard [26]. To cope with ill-posedness, additional assumptions on the inversion parameter, such as smoothness are included via regularization [83, 132]. The inverse problem is stated as a nonlinear least-squares minimization problem, namely

$$\min_{m \in \mathcal{M}} \mathcal{J}(m) := \frac{1}{2} \|\mathcal{F}(m) - d\|^2_{\Gamma^{-1}_{\text{noise}}} + \mathcal{R}(m), \tag{2.3}$$

where the first term on the right in Equation (2.3) represents the misfit term (between observations and the predicted forward solution), and $\mathcal{R}(m)$ is the regularization. Here $\|\cdot\|_{\Gamma^{-1}_{\text{noise}}}$ is a weighted norm defined in finite dimensions as $\|u\|^2_W = u^T W u$), and Γ_{noise} is the noise covariance matrix. In this thesis, unless otherwise stated, Equation (2.3) is our objective function with Tikhonov regularization [80].

2.1.1 Adjoint-based first- and second-order derivatives

We use the Lagrangian formalism [79] and follow [126] to derive abstract expressions for the first- (i.e., gradient) and second-order (i.e., Hessian) derivative information. The procedure is outlined below. First we write the Lagrangian functional as

$$\mathcal{L}(u,m,p) := \frac{1}{2}\|\mathcal{F}(m)-d\|_{\Gamma^{-1}_{\text{noise}}} + \mathcal{R}(m) + \langle p, r(u,m)\rangle_{\mathcal{V},\mathcal{V}^*}, \tag{2.4}$$

where $p \in \mathcal{V}$ is the Lagrange multiplier (which later will become the adjoint variable or so-called test function, depending on the context), and $\langle \cdot, \cdot \rangle_{\mathcal{V},\mathcal{V}^*}$ represents the duality pair or the variational (or weak) formulation of the PDE.

We remind the reader that the weak form of a PDE given in general form by $r(u,m) = 0$ can be written as $\langle p, r(u,m)\rangle_{\mathcal{V},\mathcal{V}^*} = r(u,m)p = \int_\Omega r(u,m)(x)p(x)dx$, for all test functions $p \in \mathcal{V}$. Returning to the derivation of the optimality system, the derivative of the Lagrangian function (2.4) in an arbitrary direction $\tilde{m} \in \mathcal{M}$ (in weak form) is given by

$$\mathcal{G}(u,m,p)(\tilde{m}) = (\mathcal{R}_m(m), \tilde{m}) + \langle p, r_m(u,m)[\tilde{m}]\rangle_{\mathcal{V},\mathcal{V}^*}, \quad \forall \tilde{m} \in \mathcal{M}, \tag{2.5}$$

where $(\mathcal{R}(m), \tilde{m})$ denotes the derivative of the regularization with respect to m in the direction $\tilde{m}$. Here we used the Euler-Lagrange formula from variational calculus [2], namely $G(u,m,p)(\tilde{m}) = \langle \nabla \mathcal{L}(u,m,p), \tilde{m}\rangle = \frac{d}{d\varepsilon}\mathcal{L}(u, m+\varepsilon, p)|_{\varepsilon=0}$. Similarly $r_m(u,m)[\tilde{m}]$ is the derivative of r with respect to m in a direction $\tilde{m}$. Furthermore, we can derive the Fréchet derivatives of the Lagrangian with respect to the adjoint p and the state variable u as follows

$$\mathcal{L}_p(u,m,p)(\tilde{p}) = \langle \tilde{p}, r(u,m)\rangle_{\mathcal{V}^*} \quad \forall \tilde{p} \in \mathcal{V}, \tag{2.6}$$

$$\mathcal{L}_u(u,m,p)(\tilde{u}) = \langle p, r_u(u,m)[\tilde{u}]\rangle_{\mathcal{V}^*} + \langle \mathcal{B}(u)[\tilde{u}], \mathcal{B}(u) - d\rangle_{\mathbb{R}^q} \quad \forall \tilde{u} \in \mathcal{V}, \tag{2.7}$$

where we call $\mathcal{L}_p(u,m,p)(\tilde{p}) = 0$, for all $\tilde{p} \in \mathcal{V}$, the forward (state) problem and $\mathcal{L}_u(u,m,p)(\tilde{u}) = 0$, for all $\tilde{u} \in \mathcal{V}$, is the adjoint problem. To obtain the second derivative of the objective function we can consider the gradient, forward, and adjoint together in a new Lagrangian (second order) functional

$$\mathcal{L}^H(u,m,p;\hat{u},\hat{,}m,\hat{,}p) : = (\mathcal{G}(m), \hat{m}) + \langle \hat{p}, r(u,m)\rangle_{\mathcal{V},\mathcal{V}^*} + \langle p, r_u(u,m)[\hat{u}]\rangle_{\mathcal{V},\mathcal{V}^*} \tag{2.8}$$

$$+ \langle \mathcal{B}(u)[\hat{u}], \mathcal{B}(u) - d\rangle_{\mathbb{R}^q}, \tag{2.9}$$

where the first term is the gradient expression, the second term stems from the forward problem, and the last two terms represent the adjoint problem. We can repeat the previous steps to obtain the new optimality system. The Hessian-apply (or action) in an arbitration direction $\hat{m}$ evaluated at $[p_0, u_0, m_0]$ can be written as [126, 125, 127]

$$\langle \tilde{m}, H(m_0)\hat{m}\rangle = \langle \tilde{m}, \mathcal{R}_{mm}(m_0)[\hat{m}]\rangle + \langle p_0, r_{mm}(u_0, m_0)[\tilde{m}, \hat{m}]\rangle_{\mathcal{V},\mathcal{V}^*} \tag{2.10}$$

$$+ \langle \hat{p}, r_m(u_0, m_0)[\tilde{m}]\rangle_{\mathcal{V},\mathcal{V}^*} + \langle \hat{p}_0, r_{um}(u_0, m_0)[\tilde{u}, \hat{m}]\rangle_{\mathcal{V},\mathcal{V}^*}. \tag{2.11}$$

We note that the cost of each Hessian-apply requires the solution of the so-called incremental state problem

$$\mathcal{L}^H_p(u, m, p; \hat{u}, \hat{,}m, \hat{,}p)(\tilde{p}) = \langle \tilde{p}, r_u(u, m)[\hat{u}]\rangle_{\mathcal{V},\mathcal{V}^*} = 0 \quad \forall \tilde{p} \in \mathcal{V}, \tag{2.12}$$

and incremental adjoint problem

$$\mathcal{L}^H_u(u, m, p; \hat{u}, \hat{,}m, \hat{,}p)(\tilde{u}) = \langle \hat{p}, r(u, m)[\tilde{u}]\rangle_{\mathcal{V},\mathcal{V}^*} + \langle \mathcal{B}(u)[\hat{u}], \mathcal{B}(u)[\tilde{u}] - d\rangle_{\mathbb{R}^q} = 0 \quad \forall \tilde{u} \in \mathcal{V}. \tag{2.13}$$

2.2 Bayesian Inverse Problems

In what follows we state the inverse problem in a Bayesian inference framework and formulate the inverse problem as a problem of statistical inference over the space of uncertain parameters, which are to be inferred from data and a physical-based model. This is done using Bayes formula, which in infinite dimensions reads

$$\frac{d\mu_{\text{post}}}{d\mu_{\text{prior}}} \propto \pi_{\text{like}}(\mathrm{d}|m). \tag{2.14}$$

Here, $d\mu_{\text{post}}/d\mu_{\text{prior}}$ denotes the Radon-Nikodym derivative [78] of the posterior measure μ_{post} with respect to μ_{prior}, and $\pi_{\text{like}}(\mathrm{d}|m)$ denotes the data likelihood [76].

Following [126] we assume an additive noise model, $\boldsymbol{d} = \mathcal{F}(m) + \boldsymbol{\eta}$, where $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Gamma}_{\text{noise}})$ is a centered Gaussian on $\mathbb{R}^q$. Therefore we define the likelihood as

$$\pi_{\text{like}}(\mathrm{d}|m) \propto \exp\Big\{ -\frac{1}{2}\|\mathcal{F}(m) - \boldsymbol{d}\|^2_{\mathbf{\Gamma}^{-1}_{\text{noise}}}\Big\}. \tag{2.15}$$

Also in line [126], we choose the prior to be Gaussian, i.e., $m \sim \mathcal{N}(m_{\text{pr}}, \mathcal{C}_{\text{prior}})$. This implies

$$d\mu_{\text{prior}}(m) \propto \exp\Big\{ -\frac{1}{2}\|m - m_{\text{pr}}\|^2_{\mathcal{C}^{-1}_{\text{prior}}}\Big\}, \tag{2.16}$$

where the covariance operator is defined using a Laplacian like operator [126, 76]. With the likelihood and prior chosen as above, the posterior distribution in (2.14) becomes

$$d\mu_{\text{post}} \propto \exp\Big\{ -\frac{1}{2}\|\mathcal{F}(m) - \boldsymbol{d}\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{noise}}} - \frac{1}{2}\|m - m_{\text{pr}}\|^2_{\mathcal{C}^{-1}_{\text{prior}}} \Big\}. \tag{2.17}$$

The *maximum a posteriori* (MAP) point m_{MAP} is defined as the parameter field that maximizes the posterior distribution, namely

$$m_{\text{MAP}} := \underset{m\in\mathcal{M}}{\arg\min}(-\log d\mu_{\text{post}}(m)) = \underset{m\in\mathcal{M}}{\arg\min}\,\frac{1}{2}\|\mathcal{F}(m) - \boldsymbol{d}\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{noise}}} + \frac{1}{2}\|m - m_{\text{pr}}\|^2_{\mathcal{C}^{-1}_{\text{prior}}}. \tag{2.18}$$

We note that when $\mathcal{F}$ is linear, due to the particular choice of prior and noise model, the posterior measure is Gaussian, $\mathcal{N}(m_{\text{MAP}}, \mathcal{C}_{\text{post}})$ with [76, Section 6.4],

$$\mathcal{C}_{\text{post}} = \mathcal{H}^{-1} = (\mathcal{F}^*\boldsymbol{\Gamma}^{-1}_{\text{noise}}\mathcal{F} + \mathcal{C}^{-1}_{\text{prior}})^{-1}, \qquad m_{\text{MAP}} = \mathcal{C}_{\text{post}}(\mathcal{F}^*\boldsymbol{\Gamma}^{-1}_{\text{noise}}\mathrm{d} + \mathcal{C}^{-1}_{\text{prior}}m_{\text{pr}}), \tag{2.19}$$

where $\mathcal{F}^* : \mathbb{R}^q \to \mathcal{M}$ is the adjoint of $\mathcal{F}$. In the following projects we focus on approximations and reduced order models for the Hessian $\mathcal{H}(m_{map})$ of the negative log-posterior evaluated at the maximum a posteriori. The driving force behind that is the Hessian plays a fundamental role in the inversion and the uncertainty quantification for the inferred parameter. According to [67] the Hessian (inverse) indicates which directions in the parameter space are most informed by the data.

Chapter 3

Quasi-Newton Formulas for Optimization in Function Space

3.1 Introduction

In optimization, quasi-Newton methods are a pragmatic alternative to Newton-type methods for problems where the Hessian of the objective function is difficult to derive (*e.g.*, for optimization problems constrained by differential equations, which requires a considerable amount of work to setup the numerical evaluation of the second-order derivatives) or is computationally expensive to evaluate. In computational practice, it is often the case that quasi-Newton performs similarly or even outperforms Newton's method: while the iteration count is generally higher for quasi-Newton methods than for Newton-type methods, the cost of one iteration of quasi-Newton methods is generally lower than the cost of a Newton iteration, which may offset the disadvantage of a higher iteration count.

Quasi-Newton methods received considerable attention in the optimization community in the last decades [111]. When applied to the minimization of a twice continuously differentiable function $f(x)\colon \mathbb{R}^n \to \mathbb{R}$, that is

$$\min_{x\in\mathbb{R}^n} f(x), \tag{3.1}$$

a quasi-Newton method generates a sequence of iterates x_k, $k = \{0, 1, \ldots\}$, by computing a search direction of the form $\Delta x_k = \alpha_k B_k^{-1} \nabla f(x_k)$, and by choosing an appropriate scalar step size α_k that ensures a minimum decrease of the objective $f(x)$ along the direction Δx_k. Alternatively, the search direction can be in the form $\Delta x_k = \alpha_k H_k \nabla f(x_k)$. The $n \times n$ matrices B_k and H_k are approximations of the Hessian $\nabla^2 f(x_k)$ and its inverse, respectively. The salient idea of quasi-Newton methods is to maintain these approximations by enforcing the secant condition in the form $B_k s_k = y_k$ or $H_k y_k = s_k$, where

$$s_k = x_{k+1} - x_k \text{ and } y_k = \nabla f(x_{k+1}) - \nabla f(x_k).$$

The Davidon–Fletcher–Powell (DFP) and the Broyden–Fletcher–Goldfarb–Shanno (BFGS) rank-two update formulas have emerged in the last decades [111] as the most efficient and, as a consequence, most commonly used Hessian approximations in a quasi-Newton framework. These formulas have closed-form algebraic forms, namely,

$$\begin{aligned}
B_{k+1}^{DFP} &= (I - \gamma_k y_k s_k^T) B_k (I - \gamma_k s_k y_k^T) + \gamma_k y_k y_k^T, \\
H_{k+1}^{DFP} &= H_k - \frac{H_k y_k y_k^T H_k}{y_k^T H_k y_k} + \gamma_k s_k s_k^T, \\
B_{k+1}^{BFGS} &= B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \gamma_k y_k y_k^T, \text{ and} \\
H_{k+1}^{BFGS} &= (I - \gamma_k s_k y_k^T) H_k (I - \gamma_k y_k s_k^T) + \gamma_k s_k s_k^T,
\end{aligned}$$

where $\gamma_k = \frac{1}{s_k^T y_k}$. Other secant formulas that have been proposed and investigated in the past are the symmetric rank-one (SR1) [111] and Powell–Symmetric–Broyden (PSB) updates [96]. They also have closed-form expressions in the form of

$$\begin{aligned}
B_{k+1}^{PSB} &= B_k + \frac{s_k (y_k - B_k s_k)^T + (y_k - B_k s_k) s_k^T}{\langle s_k, s_k \rangle} - \frac{\langle y_k - B_k s_k, s_k \rangle}{\langle s_k, s_k \rangle^2} s_k s_k^T, \\
H_{k+1}^{PSB} &= H_k + \frac{y_k (s_k - H_k y_k)^T + (s_k - H_k y_k) y_k^T}{\langle y_k, y_k \rangle} - \frac{\langle s_k - H_k y_k, y_k \rangle}{\langle y_k, y_k \rangle^2} y_k y_k^T, \\
B_{k+1}^{SR1} &= B_k + \frac{(y_k - B_k s_k)(y_k - B_k s_k)^T}{\langle s_k, y_k - B_k s_k \rangle}, \text{ and} \\
H_{k+1}^{SR1} &= H_k + \frac{(s_k - H_k y_k)(s_k - H_k y_k)^T}{\langle y_k, s_k - H_k y_k \rangle}.
\end{aligned}$$

Related work In this thesis we consider the optimization problem (3.1) over a separable Hilbert space $\mathcal{H}$, possibly infinite-dimensional, *e.g.*, a function space such as L^2, and derive infinite-dimensional versions of the update formulas above. Central to our derivation is the use of a variational, least-squares approach that was first introduced by Güler et al. for finite-dimensional optimization problems [103]. To this extent the present work can be seen as a generalization of the work presented in [103] to an infinite-dimensional optimization setting. Some of the infinite-dimensional quasi-Newton formulas we derive in this work have previously appeared in the literature. In [109], for example, the authors use the class of variable metric methods, of which the BFGS, DFP, and Symmetric Rank One (SR1) formulas are members, for control problems over function spaces, while Broyden updates are proposed in [122] for solving nonlinear operator equations in Hilbert spaces. In [102] the authors derive the BFGS formula in infinite dimension starting from finite-rank updates and by imposing symmetry and positivity to arrive to desired form. More recently, a survey of quasi-Newton methods in Hilbert spaces is given in [87] with a case study for Riccati matrix equations. The BFGS and DFP formulas are used for optimization problems in a Hilbert space setting also in [124, 123, 37]. In [124, 123, 107] the authors present an instructive example of the impact of taking into account the infinite-dimensional nature of the underlying optimization problem on the performance of the numerical algorithm leading to mesh-independence. However, these quasi-Newton formulas are typically simply constructed/conjectured in analogy with the finite-dimensional counterparts.

Contributions To the best of our knowledge, the present work is the first to introduce a formal derivation of BFGS, DFP, SR1, and PSB formulas for infinite-dimensional optimization problems. We note that our derivation can be also used to formally generalize the finite-dimensional limited-memory compact quasi-Newton representations of Byrd et al. [92] to Hilbert spaces. We succinctly do so in Section 3.4.1. Furthermore, in this thesis we also illustrate how the infinite-dimensional least-square variational approach can be used to derive new and improved quasi-Newton formulas that exploit structured Hessians present in some specific classes of optimization problems; in particular, we look at inverse problems governed by partial-differential equations, derive new structured updates that explicitly incorporate the computationally affordable part of Hessian, and show that the new "structured" quasi-Newton formulas improve considerably over the unstructured counterparts.

The remaining sections of this chapter of the thesis are organized as follows. After presenting the requisite background material in section 3.2, we derive a series of technical results that are crucial for the main results in section 3.3. In section 3.4, we derive formally the update formulas for various standard secant formulas over infinite-dimensional Hilbert spaces. In the same section we also show that the limited-memory compact representations for BFGS and DFP can be generalized to Hilbert spaces using the technical results presented in section 3.3. Finally, in section 3.5 we exploit the structure present in certain classes of infinite-dimensional inverse problems and show how *structured*, more efficient secant formulas, can be obtained using the variational approach developed in sections 3.3 and 3.4. Here we also show numerical results. Section 3.6 provides concluding remarks.

3.2 Preliminaries

In this section, we summarize the terminology and background material required for the derivation of the quasi-Newton formulas in infinite-dimensional setting. In what follows, we consider $\mathcal{H}$ and $\mathcal{K}$ separable Hilbert spaces, *i.e.*, they have a countable basis [93].

Definition 1. *[106, p. 187] The space of all bounded linear operators from $\mathcal{H}$ to $\mathcal{K}$ is denoted by $\mathcal{B}(\mathcal{H},\mathcal{K})$. In particular, the space of all bounded linear operators from $\mathcal{H}$ to itself is denoted by $\mathcal{B}(\mathcal{H})$.*

Definition 2. *[110, p. 60] Let $\mathcal{H}$ be a separable Hilbert space and $\{e_i\}_{i\in I}$ be an orthonormal basis for $\mathcal{H}$. A bounded operator $A \in \mathcal{B}(\mathcal{H})$ is a Hilbert–Schmidt (HS) operator if*

$$\|A\|_{HS} = \sum_{i\in I} \|Ae_i\|^2 < \infty. \tag{3.2}$$

We denote the set of all Hilbert–Schmidt operators by $\mathcal{B}_{00}(\mathcal{H})$.

Definition 3. *[110, p. 60] For any A and $B \in \mathcal{B}_{00}(\mathcal{H})$, the Hilbert–Schmidt inner product is defined as*

$$\langle A, B\rangle_{HS} = \sum_{i \in I} \langle Ae_i, Be_i\rangle, \tag{3.3}$$

where $\{e_i\}_{i \in I}$ is an orthonormal basis of $\mathcal{H}$.

Definition 4. *[120, p. 97] The adjoint of an operator $A \in \mathcal{B}(\mathcal{H})$ is denoted by A^* and is defined as an operator from $\mathcal{B}(\mathcal{H})$ that allows the transformation $\langle Ax, y\rangle = \langle x, A^*y\rangle$ for all x and y in $\mathcal{H}$.*

Proposition 1. *[110, p. 62] The Hilbert–Schmidt operators form a two-sided ideal in the Banach algebra of bounded operators on $\mathcal{H}$, that is, for any $A \in \mathcal{B}_{00}(\mathcal{H})$ and $B \in \mathcal{B}(\mathcal{H})$, one must necessarily have $AB \in \mathcal{B}_{00}(\mathcal{H})$, $BA \in \mathcal{B}_{00}(\mathcal{H})$, and $A^* \in \mathcal{B}_{00}(\mathcal{H})$.*

Definition 5. *[118, p. 132] A linear bounded operator $A \in \mathcal{B}(\mathcal{H})$ is positive if $\langle x, Ax\rangle \geq 0$ for all $x \in \mathcal{H}$.*

Definition 6. *[118, p. 263] The square root operator R of symmetric positive A is defined as a symmetric operator such that $R^2 = A$.*

Theorem 1. *[118, p. 265] If $A \in \mathcal{B}(\mathcal{H})$ is a symmetric positive operator, then there exists a unique positive square root R of A. Furthermore, R commutes with any bounded operator that commutes with A.*

Theorem 2. *[118, p. 266] Given any $A \in \mathcal{B}(\mathcal{H})$, the following conditions are equivalent:*

i) A is invertible;

ii) there exists a constant $\alpha > 0$ such that $A^*A \geq \alpha I_{\mathcal{H}}$ and $AA^* \geq \alpha I_{\mathcal{H}}$;

iii) there exists a constant $\alpha > 0$ such that

$$\langle A^*Ax, x\rangle \geq \alpha \|x\| \;\; \textit{and} \; \langle AA^*x, x\rangle \geq \alpha \|x\|;$$

iv) both operators $A^*A \in \mathcal{B}(\mathcal{H})$ and $AA^* \in \mathcal{B}(\mathcal{H})$ are invertible.

Following [87, 122, 105], we next define the outer (dyadic) product, which is the correspondent of the rank-one update used with finite-dimensional secant formulas.

Definition 7. *[110, p. 55] Let $x, y \in \mathcal{H}$. The outer or dyadic product of x and y is the (linear) operator, denoted by $x \otimes y$, that satisfies*

$$(x \otimes y)z = \langle y, z \rangle x, \forall z \in \mathcal{H}. \tag{3.4}$$

We note that $x \otimes y$ is a bounded linear operator.

Definition 8. *[94, p. 41] An operator T is of finite-rank if its range is finite-dimensional.*

Example 1. *The operator $x \otimes y$ is a rank-one operator since it has the range equal to the one-dimensional subspace of $\mathcal{H}$ that is spanned by x.*

Remark 1. *For the vectors $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$, the operator $\sum_{i=1}^n x_i \otimes y_i$ has finite rank of most n.*

Remark 2. *It can be proven that every finite-rank operator is a Hilbert–Schmidt operator [110].*

Definition 9. *[106, p. 110] A linear operator $T : \mathcal{H} \to \mathcal{K}$ is compact if and only if for every bounded sequence $\{x_n\} \in \mathcal{H}$ there exists a subsequence $\{x_{n_k}\}$ such that $T(\{x_{n_k}\})$ converges in $\mathcal{K}$.*

Theorem 3. *[94, p. 41] If T is a compact operator, then there exists a sequence of finite rank operators $\{T_n\}$ such that $\|T - T_n\| \to 0$.*

We now state the Hilbert Projection Theorem, which is one of the key results used by our least-squares variational approach.

Theorem 4. *[108, p. 50] Let $\mathcal{H}$ be a Hilbert space and M a closed subspace of $\mathcal{H}$. For any vector $x \in \mathcal{H}$, there is a unique vector $m_0 \in M$ such that $\|x - m_0\| \leq \|x - m\|$ for all $m \in M$. Furthermore, a necessary and sufficient condition to characterize $m_0 \in M$ is that $x - m_0$ is orthogonal to M.*

Finally, the following Theorem states the Sherman–Morrison–Woodbury formula in Banach spaces of linear operators [95]; for compactness, we consider such linear operators to be defined over Hilbert spaces $\mathcal{H}$ and $\mathcal{K}$, however, they can be defined in general over Banach spaces.

Theorem 5. *[95, p. 1] Let $A \in \mathcal{B}(\mathcal{H})$ and $G \in \mathcal{B}(\mathcal{K})$ both be invertible and $Y, Z \in \mathcal{B}(\mathcal{K}, \mathcal{H})$. The operator $A+YGZ^*$ is invertible if and only if $G^{-1}+Z^*A^{-1}Y$ is invertible. Furthermore,*

$$(A + YGZ^*)^{-1} = A^{-1} - A^{-1}Y(G^{-1} + Z^*A^{-1}Y)^{-1}Z^*A^{-1}. \tag{3.5}$$

3.3 Least-squares variational characterization framework for deriving quasi-Newton updates

This section derives intermediary results needed in Section 3.4 to derive various quasi-Newton update formulas as analytical solutions to infinite-dimensional variational problems. Let us first denote by $\mathcal{B}^s(\mathcal{H})$ the set of bounded linear operators that are *self-adjoint* and consider the linear subspace $\mathcal{L} = \{X \in \mathcal{B}^s(\mathcal{H}) : Xs = 0\}$, which corresponds to the affine subspace given by the secant equation, namely to $\mathcal{A} = \{X \in \mathcal{B}^s(\mathcal{H}) : Xs = y\}$. Furthermore, we define the operators $S_i = s \otimes e_i + e_i \otimes s$ for each $i \in I$, where $\{e_i\}_{i \in I}$ is an (countable) orthonormal basis of the (separable) Hilbert space $\mathcal{H}$.

Lemma 1. *If $s, y \in \mathcal{H}$ with $s \neq 0$, then the following statements are true:*

(i) $\mathcal{L} = \{X \in \mathcal{B}^s(\mathcal{H}) : \langle X, S_i \rangle_{HS} = 0, \forall i \in I\}$;

(ii) $\mathcal{L}^\perp = \operatorname{span}\{\{S_i\}_{i \in I}\}$;

(iii) $\mathcal{L}^\perp = \{s \otimes \lambda + \lambda \otimes s \,:\, \lambda \in \mathcal{H}\}$.

Proof. (i) We first remark that

$$XS_ie_j = X(s \otimes e_i)e_j + X(e_i \otimes s)e_j = \langle e_i, e_j \rangle Xs + \langle s, e_j \rangle Xe_i \tag{3.6}$$

for any $X \in \mathcal{B}^s(\mathcal{H})$ and $i \in I$. Since $\langle X, S_i\rangle_{HS} = \sum_{j=1}^{\infty}\langle Xe_j, S_ie_j\rangle = \sum_j\langle e_j, XS_ie_j\rangle$, identity (3.6) allows us to write

$$\langle X, S_i\rangle_{HS} = \sum_{j=1}^{\infty}\big[\langle e_i, e_j\rangle\langle e_j, Xs\rangle + \langle s, e_j\rangle\langle e_j, Xe_i\rangle\big]. \tag{3.7}$$

Since X is self-adjoint, $\langle e_i, e_j\rangle = 0$ for $i \neq j$, and $\langle e_j, e_j\rangle = 1$, one can subsequently write that

$$\begin{aligned}\langle X, S_i\rangle_{HS} &= \langle e_i, Xs\rangle + \langle\sum_{j=1}^{\infty}\langle s, e_j\rangle e_j, Xe_i\rangle \\ &= \langle e_i, Xs\rangle + \langle s, Xe_i\rangle = \langle e_i, Xs\rangle + \langle Xs, e_i\rangle = 2\langle Xs, e_i\rangle.\end{aligned}$$

This shows that $X \in \mathcal{L}$ if and only if $\langle X, S_i\rangle_{HS} = 0$ for all $i \in I$.

(ii) Let $Y \in \operatorname{span}\{\{S_i\}_{i=1}^{\infty}\}$, namely $Y = \sum_{i=1}^{\infty}\alpha_iS_i$. Then consider $\langle X, Y\rangle$ for any $X \in \mathcal{L}$, one can write

$$\langle X, Y\rangle = \langle X, \sum_{i=1}^{\infty}\alpha_iS_i\rangle = \sum_{i=1}^{\infty}\alpha_i\langle X, S_i\rangle = 0$$

obtaining $\mathcal{L}^{\perp} \supseteq \operatorname{span}\{\{S_i\}_{i=1}^{\infty}\}$. For the other inclusion, let $Y \in \operatorname{span}\{\{S_i\}_{i=1}^{\infty}\}^{\perp}$, this implies $\langle Y, S_i\rangle = 0$ for all S_i, *i.e.*, $Y \in \mathcal{L}$. This shows that $\operatorname{span}\{\{S_i\}_{i=1}^{\infty}\}^{\perp} \subseteq \mathcal{L}$, taking the orthogonal complement we obtain $\operatorname{span}\{\{S_i\}_{i=1}^{\infty}\} \supseteq \mathcal{L}^{\perp}$.

(iii) Consider $Y \in \mathcal{L}^{\perp}$, $Y = \sum_{i=1}^{\infty}\alpha_iS_i$, which we can rewrite as

$$\sum_{i=1}^{\infty}\alpha_iS_i = \sum_{i=1}^{\infty}\alpha_i(s \otimes e_i + e_i \otimes s) = (s \otimes \lambda + \lambda \otimes s) \tag{3.8}$$

for some $\lambda = \sum_{i=1}^{\infty}\alpha_ie_i$. This shows that $\mathcal{L}^{\perp} \subseteq \{s \otimes \lambda + \lambda \otimes s \,:\, \lambda \in \mathcal{H}\}$. On the other hand, if $Y \in \{s \otimes \lambda + \lambda \otimes s \,:\, \lambda \in \mathcal{H}\}$ and $X \in \mathcal{L}$ we have

$$\begin{aligned}\langle X, Y\rangle = \langle X, s \otimes \lambda + \lambda \otimes s\rangle &= \langle X, s \otimes \sum_{i=1}^{\infty}\alpha_ie_i + \sum_{i=1}^{\infty}\alpha_ie_i \otimes s\rangle \\ &= \sum_{i=1}^{\infty}\alpha_i\langle X, s \otimes e_i + e_i \otimes s\rangle = \sum_{i=1}^{\infty}\alpha_i\langle X, S_i\rangle = 0.\end{aligned}$$

This completes the proof that $\mathcal{L}^{\perp} = \{s \otimes \lambda + \lambda \otimes s \,:\, \lambda \in \mathcal{H}\}$.

□

We now consider a generic least-squares problem that is closely related to the variational problem used to derive the various quasi-Newton formulas in Sections 3.4 and 3.5. This is motivated in the spirit of the work in [37].

Theorem 6. *Given $s, y \in \mathcal{H}$, the variational problem*

$$\min_{X \in \mathcal{B}(\mathcal{H})} \frac{1}{2} \|X\|_{HS}^2 \tag{3.9}$$

$$\text{s.t. } Xs = y \tag{3.10}$$

has a self-adjoint solution operator $\overline{X} \in \mathcal{B}_{00}(\mathcal{H})$ given by

$$\overline{X} = \frac{s \otimes y + y \otimes s}{\langle s, s \rangle} - \frac{\langle y, s \rangle}{\langle s, s \rangle^2} s \otimes s. \tag{3.11}$$

Proof. We note that the set $\mathcal{A} = \{X \in \mathcal{B}(\mathcal{H}) \,|\, Xs = y\}$ is closed. Let $\overline{X}$ denote a solution of (3.9)-(3.10); such solution necessarily exists per Hilbert projection Theorem [121, p. 80]. We remark that for any $A \in \mathcal{L}$ and for any $t \in \mathbb{R}$, the function $\overline{X} + tA$ satisfies the secant equation (3.10). Let us consider an arbitrary $A \in \mathcal{L}$. Then we obtain by the minimality of $\overline{X}$ that $\|\overline{X}\|_{HS}^2 \leq \|\overline{X} + tA\|_{HS}^2$, or, equivalently, that $\langle \overline{X}, \overline{X} \rangle_{HS} \leq \langle \overline{X} + tA, \overline{X} + tA \rangle_{HS}$ for any $t \in \mathbb{R}$. A simple manipulation of this inequality reveals that one must necessarily have $-2t\langle \overline{X}, A \rangle_{HS} \leq t^2 \langle A, A \rangle_{HS}$ for any $t \in \mathbb{R}$. For positive t, the previous inequality is equivalent to $\langle \overline{X}, A \rangle_{HS} \geq -\frac{t}{2} \langle A, A \rangle_{HS}$ and can hold for arbitrarily small t only if $\langle \overline{X}, A \rangle_{HS} \geq 0$. Similarly, by taking t to be negative and arbitrarily close to zero, one must necessarily have $\langle \overline{X}, A \rangle_{HS} \leq 0$. Therefore, we have that $\langle \overline{X}, A \rangle_{HS} = 0$. Since A was chosen arbitrary from $\mathcal{L}$, this implies that $\overline{X} \in L^\perp$ and thus, based on iii) of Lemma 1 that $\overline{X} = s \otimes \lambda + \lambda \otimes s$, for some $\lambda \in \mathcal{H}$.

Next we find an explicit expression for λ. Since $\overline{X}s = y$, we can write

$$\begin{aligned} \langle y, s \rangle = \langle \overline{X}s, s \rangle &= \langle [s \otimes \lambda + \lambda \otimes s]s, s \rangle \\ &= \langle [s \otimes \lambda]s, s \rangle + \langle [\lambda \otimes s]s, s \rangle = \langle \langle \lambda, s \rangle s, s \rangle + \langle \langle s, s \rangle \lambda, s \rangle \\ &= \langle \langle \lambda, s \rangle s, s \rangle + \|s\|^2 \langle \lambda, s \rangle = 2\|s\|^2 \langle \lambda, s \rangle, \end{aligned}$$

to obtain that $\langle \lambda, s \rangle = \frac{\langle y, s \rangle}{2\|s\|^2}$. This can be used in conjunction with the secant equation to write that $y = \overline{X}s = [s \otimes \lambda + \lambda \otimes s]s = \langle s, s \rangle \lambda + \langle \lambda, s \rangle s = \|s\|^2 \lambda + \frac{\langle y, s \rangle}{2\|s\|^2} s$, from which λ is obtained to be

$$\lambda = \frac{1}{\|s\|^2} y - \frac{\langle y, s\rangle}{2\|s\|^4} s.$$

Equation (3.11) follows readily by substituting the above expression for λ in $\overline{X} = \lambda \otimes s + s \otimes \lambda$. Finally, we remark that $\overline{X}$ given by (3.11) is self-adjoint; also, one can easily verify that has rank two, which implies that $\overline{X} \in \mathcal{B}_{00}$ [97]. □

The following corollary offers an analytical expression for the solution of a prototype variational problem and will be the basis of the derivation of quasi-Newton update formulas in generic Hilbert spaces.

Corollary 1. *For any given operator $X_0 \in \mathcal{B}^s(\mathcal{H})$ and positive and invertible "weight" operator $W \in \mathcal{B}^s(\mathcal{H})$, the variational problem*

$$\min_{X \in \mathcal{B}(\mathcal{H})} \frac{1}{2} \left\| W^{1/2}(X - X_0)W^{1/2} \right\|_{HS}^2 \tag{3.12}$$

$$\text{s.t. } Xs = y \tag{3.13}$$

admits a solution $\overline{X} \in \mathcal{B}^s(\mathcal{H})$ in the form

$$\begin{aligned}\overline{X} = & X_0 + \frac{W^{-1}s \otimes (y - X_0 s) + (y - X_0 s) \otimes W^{-1}s}{\langle s, W^{-1}s\rangle} - \\ & \frac{\langle y - X_0 s, s\rangle}{\langle s, W^{-1}s\rangle^2} W^{-1}s \otimes W^{-1}s.\end{aligned}$$

Furthermore, the operator $\overline{X} - X_0$ lies in $\mathcal{B}_{00}(\mathcal{H})$.

Proof. The corollary is a direct consequence of Theorem 6. More specifically, since W is invertible and positive, we can write the secant equation (3.13) as

$$W^{1/2}(X - X_0)W^{1/2}(W^{-1/2}s) = W^{1/2}(y - X_0 s).$$

Then Theorem 6 implies that a minimizer $\overline{X}$ of (3.12)-(3.13) exists and satisfies

$$\begin{aligned}W^{1/2}(\overline{X} - X_0)W^{1/2} = & \frac{W^{-1/2}s \otimes W^{1/2}(y - X_0 s) + W^{1/2}(y - X_0 s) \otimes W^{-1/2}s}{\langle W^{-1/2}s, W^{-1/2}s\rangle} \\ & - \frac{\langle W^{1/2}(y - X_0 s), W^{-1/2}s\rangle}{\langle W^{-1/2}s, W^{-1/2}s\rangle^2} W^{-1/2}s \otimes W^{-1/2}s.\end{aligned}$$

The form of $\overline{X}$ from the corollary follows from the above identity by multiplying from left and right with $W^{-1/2}$ and performing appropriate simple algebraic manipulations. We remark that Theorem 6 also implies that $W^{1/2}(\overline{X} - X_0)W^{1/2} \in \mathcal{B}_{00}(\mathcal{H})$. This implies that $\overline{X} - X_0 = W^{-1/2}W^{1/2}(\overline{X} - X_0)W^{1/2}W^{-1/2} \in \mathcal{B}_{00}(\mathcal{H})$ since Hilbert–Schmidt operators form an ideal in $\mathcal{B}(\mathcal{H})$ (see Proposition 1). □

3.4 Derivation of various secant update formulas

In this section we derive the quasi-Newton update formulas for approximating a second-order derivative operators defined over generic Hilbert spaces. Corollary 1 is used with specific choices for the "weight" operator W to obtain in this section the classical BFGS, DFP, PSB, and SR1 formulas in their operator form.

Proposition 2 (DFP formula for Hessian operator)**.** *Let us consider an operator $B_k \in \mathcal{B}^s(\mathcal{H})$, a positive and invertible operator $W \in \mathcal{B}^s(\mathcal{H})$ such that $Wy_k = s_k$, and s_k and y_k nonzero elements of $\mathcal{H}$.*

(i) The solution to the variational problem

$$\begin{aligned} \min_{B\in\mathcal{B}(\mathcal{H})} \ & \frac{1}{2} \left\| W^{1/2}(B - B_k)W^{1/2} \right\|_{HS}^2 \\ s.t \ \ & Bs_k = y_k \end{aligned} \tag{3.14}$$

is given by

$$B_{k+1} = (I - \gamma(y_k \otimes s_k))B_k(I - \gamma(s_k \otimes y_k)) + \gamma(y_k \otimes y_k), \tag{3.15}$$

where $\gamma = \frac{1}{\langle s_k, y_k\rangle}$; in addition, $B_{k+1} \in \mathcal{B}^s(\mathcal{H})$ and $B_{k+1} - B_k \in \mathcal{B}_{00}(\mathcal{H})$.

(ii) If B_k is positive and invertible, and the positive curvature condition $\langle s_k, y_k\rangle > 0$ holds, then B_{k+1} is positive and invertible.

Proof. (i) By Corollary 1, we have that

$$\begin{aligned} B_{k+1} = B_k + & \frac{W^{-1}s_k \otimes (y_k - B_k s_k) + (y_k - B_k s_k) \otimes W^{-1}s_k}{\langle s_k, W^{-1}s_k\rangle} \\ & - \frac{\langle y_k - B_k s_k, s_k\rangle}{\langle s_k, W^{-1}s_k\rangle^2} W^{-1}s_k \otimes W^{-1}s_k. \end{aligned}$$

Since $y_k = W^{-1}s_k$ and by letting $\gamma = \frac{1}{\langle s_k, y_k\rangle}$, the above identity becomes

$$\begin{aligned} B_{k+1} = B_k + \gamma[y_k \otimes (y_k - B_k s_k) + (y_k - B_k s_k) \otimes y_k] \\ -\gamma^2 \langle y_k - B_k s_k, s_k\rangle (y_k \otimes y_k). \end{aligned} \tag{3.16}$$

We note that the last term in the above equality can be simplified as follows

$$\begin{aligned} \gamma^2 \langle y_k - B_k s_k, s_k\rangle (y_k \otimes y_k) &= \gamma^2 [\langle y_k, s_k\rangle - \langle B_k s_k, s_k\rangle](y_k \otimes y_k) \\ &= \gamma(y_k \otimes y_k) - \gamma^2 \langle B_k s_k, s_k\rangle (y_k \otimes y_k). \end{aligned} \tag{3.17}$$

With the above simplification, equation (3.16) above can be manipulated to obtain the following

$$\begin{aligned} B_{k+1} = B_k &+ \gamma[y_k \otimes (y_k - B_k s_k) + (y_k - B_k s_k) \otimes y_k - (y_k \otimes y_k)] \\ &+ \gamma^2 \langle B_k s_k, s_k\rangle (y_k \otimes y_k), \end{aligned}$$

and hence

$$\begin{aligned} B_{k+1} = [B_k - \gamma(y_k \otimes B_k s_k) - \gamma(B_k s_k \otimes y_k) + \gamma^2 \langle B_k s_k, s_k\rangle (y_k \otimes y_k)] \\ + \gamma(y_k \otimes y_k). \end{aligned} \tag{3.18}$$

One can further manipulate the last identity to get the desired equation (3.15) as follows

$$\begin{aligned} B_{k+1} &= (B_k - \gamma y_k \otimes B_k s_k)(I - \gamma s_k \otimes y_k) + \gamma(y_k \otimes y_k) \\ &= (I - \gamma y_k \otimes s_k) B_k (I - \gamma s_k \otimes y_k) + \gamma(y_k \otimes y_k). \end{aligned} \tag{3.19}$$

From (3.18) we note that $B_{k+1} - B_k$ is a finite rank operator as it has at most rank four, therefore it is a Hilbert-Schmidt operator [110]. Finally, since $y_k \otimes B_k s_k$ is the adjoint of $B_k s_k \otimes y_k$, and $y_k \otimes y_k$ is self-adjoint, by using the properties of the dyadic product and the fact that B_k is self-adjoint we conclude that B_{k+1} is self-adjoint.

(ii) Let us write $B_{k+1} = G^*G + F$ where $G = B_k^{1/2}(I - \gamma s_k \otimes y_k)$, and $F = \gamma(y_k \otimes y_k)$. Since B_k is positive, one can prove that $\langle x, G^*Gx\rangle \geq 0$ for all $x \in \mathcal{H}$. Furthermore, $0 = \langle x, G^*Gx\rangle$, or equivalently, $0 = \langle Gx, Gx\rangle$ if and only if $Gx = 0$, which in turn holds if and only if $x - \gamma(s_k \otimes y_k)x = 0$

by the positiveness of B_k. We leave the proof of the fact that $\{x \in \mathcal{H} : x - \gamma(s_k \otimes y_k)x = 0\} = \{\alpha s_k : \alpha \in \mathbb{R}\}$ as an exercise to the reader, and conclude that $0 = \langle x, G^* G x\rangle$ if and only if $x = \alpha s_k$ for some real scalar α. On the other hand, it is straightforward to prove that $\langle x, Fx\rangle \geq 0$ for all $x \in \mathcal{H}$ when the positive curvature holds (and, as a result, $\gamma > 0$). Furthermore, we remark that $\langle \alpha s_k, F\alpha s_k\rangle = \alpha^2\langle s_k, y_k\rangle > 0$ for all nonzero $\alpha \in \mathbb{R}$.

With the above, we have that $\langle x, B_{k+1}x\rangle > 0$ for all nonzero x, which shows the positive definiteness of B_{k+1}. Finally, the invertibility of B_{k+1} follows from the Sherman–Morrison–Woodbury formula. We note that the latter is shown in detail in the proof of Proposition 6.

□

Proposition 3 (BFGS formula for the inverse Hessian operator)**.** *Let us consider an operator $H_k \in \mathcal{B}^s(\mathcal{H})$, a positive and invertible operator $W \in \mathcal{B}^s(\mathcal{H})$ such that $Ws_k = y_k$, and s_k and y_k nonzero elements of $\mathcal{H}$.*

(i) The solution to the variational problem

$$\min_{H \in \mathcal{B}(\mathcal{H})} \frac{1}{2}\left\|W^{1/2}(H - H_k)W^{1/2}\right\|_{HS}^2 \tag{3.20}$$

$$s.t \quad Hy_k = s_k \tag{3.21}$$

is given by

$$H_{k+1} = (I - \gamma(s_k \otimes y_k))H_k(I - \gamma(y_k \otimes s_k)) + \gamma(s_k \otimes s_k), \tag{3.22}$$

where $\gamma = \frac{1}{\langle s_k, y_k\rangle}$; in addition, $H_{k+1} \in B^s(\mathcal{H})$ and $H_{k+1} - H_k$ lies in $\mathcal{B}_{00}(\mathcal{H})$.

(ii) If H_k is positive and invertible, and the positive curvature condition $\langle y_k, s_k\rangle > 0$ holds, then H_{k+1} is positive and invertible.

Proof. The proof is identical to the proof of Proposition 2. □

We next show that different choices of the "weight" operator W inside the Hilbert-Schmidt norm lead to different quasi-Newton formulas for the Hessian or its inverse. The Powell–Symmetric–Broyden formula is obtained using the trivial weight $W = I$ as we show next in Proposition 4. Surprisingly, the symmetric rank-one update can be also obtained (when it exists) with a particular choice of W, as shown in Proposition 5. Furthermore, notable from these two examples is that W does not have to satisfy the secant equation (as it does in Propositions 2 and 3 for the DFP and BFGS formulas).

Proposition 4 (Powell–Symmetric–Broyden Update)**.** *Let us consider an operator $B_k \in \mathcal{B}^s(\mathcal{H})$ and s_k and y_k nonzero elements of $\mathcal{H}$. The solution to the variational problem*

$$\min_{B\in\mathcal{B}(\mathcal{H})} \frac{1}{2}\left\|B - B_k\right\|_{HS}^2$$
$$s.t \quad Bs_k = y_k$$

is given by

$$B_{k+1} = B_k + \frac{s_k \otimes (y_k - B_k s_k) + (y_k - B_k s_k) \otimes s_k}{\langle s_k, s_k\rangle} - \frac{\langle y_k - B_k s_k, s_k\rangle}{\langle s_k, s_k\rangle^2} s_k \otimes s_k.$$

Furthermore, B_{k+1} is self-adjoint and $B_{k+1} - B_k \in \mathcal{B}_{00}(\mathcal{H})$.

Proof. The proof follows by taking W to be the identity in Corollary 1. □

Proposition 5 (Symmetric Rank-One Update)**.** *Let us consider an operator $B_k \in \mathcal{B}^s(\mathcal{H})$ and assume that a positive and invertible operator $W \in \mathcal{B}^s(\mathcal{H})$ exists such that $W^{-1}s_k = y_k - B_k s_k$ for s_k and y_k nonzero elements of $\mathcal{H}$. The solution to the variational problem*

$$\min_{B\in\mathcal{B}(\mathcal{H})} \frac{1}{2}\left\|W^{1/2}\left(B - B_k\right)W^{1/2}\right\|_{HS}^2$$
$$s.t \quad Bs_k = y_k$$

is the operator

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k) \otimes (y_k - B_k s_k)}{\langle s_k, y_k - B_k s_k\rangle},$$

which is self-adjoint and satisfies $B_{k+1} - B_k \in \mathcal{B}_{00}(\mathcal{H})$.

Proof. By Corollary 1 we have that

$$\begin{aligned}B_{k+1} = B_k &+ \frac{W^{-1}s_k \otimes (y_k - B_k s_k) + (y_k - B_k s_k) \otimes W^{-1}s_k}{\langle s_k, W^{-1}s_k\rangle} \\ &- \frac{\langle y_k - B_k s_k, s_k\rangle}{\langle s_k, W^{-1}s_k\rangle^2} W^{-1}s_k \otimes W^{-1}s_k.\end{aligned}$$

Since $W^{-1}s_k = y_k - B_k s_k$, we simplify the above identity as follows:

$$\begin{aligned}B_{k+1} = B_k &+ \frac{(y_k - B_k s_k) \otimes (y_k - B_k s_k) + (y_k - B_k s_k) \otimes (y_k - B_k s_k)}{\langle s_k, y_k - B_k s_k\rangle} \\ &- \frac{\langle y_k - B_k s_k, s_k\rangle}{\langle s_k, y_k - B_k s_k\rangle^2}(y_k - B_k s_k) \otimes (y_k - B_k s_k) \\ = B_k &+ \frac{(y_k - B_k s_k) \otimes (y_k - B_k s_k)}{\langle s_k, y_k - B_k s_k\rangle},\end{aligned}$$

which completes the proof. □

In the remainder of this section we derive the inverse formulas for the DFP and BFGS formulas presented above in Propositions 2 and 3 using a generalization of Sherman–Morrison–Woodbury formula [95] given in Theorem 5.

Proposition 6 (BFGS formula for Hessian operator)**.** *Let us consider the positive definite and invertible operators $B_k \in \mathcal{B}^s(\mathcal{H})$ and $W \in \mathcal{B}^s(\mathcal{H})$ such that $Wy_k = s_k$, where s_k and y_k are nonzero elements of $\mathcal{H}$.*

(i) The solution to the variational problem

$$\begin{aligned}\min_{B\in\mathcal{B}(\mathcal{H})} \quad & \frac{1}{2}\left\| W^{1/2}(B^{-1} - B_k^{-1})W^{1/2}\right\|_{HS}^2 \\ s.t \quad & Bs_k = y_k\end{aligned} \tag{3.23}$$

is given by

$$B_{k+1} = B_k - \frac{B_k s_k \otimes B_k s_k}{\langle s_k, B_k s_k\rangle} + \frac{y_k \otimes y_k}{\langle s_k, y_k\rangle}; \tag{3.24}$$

in addition, $B_{k+1} \in \mathcal{B}^s(\mathcal{H})$, $B_{k+1} - B_k \in \mathcal{B}_{00}(\mathcal{H})$, and B_{k+1} is invertible.

(ii) If the positive curvature condition $\langle s_k, y_k\rangle > 0$ holds, then B_{k+1} is positive.

Proof. (i) The salient idea of the proof is to obtain (3.24) by inverting the inverse Hessian BFGS formula $B_{k+1}^{-1} = H_{k+1}$ from Proposition 3 using the Sherman–Morrison–Woodbury (SMW) formula of Theorem 5.

Let the linear operator $Y : \mathbb{R}\times\mathbb{R} \to \mathcal{H}$ be defined by $Y(\alpha, \beta) = \alpha s_k + \beta H_k y_k$. We remark that $Y \in \mathcal{B}(\mathbb{R} \times \mathbb{R}, \mathcal{H})$ and that the adjoint operator $Y^* \in \mathcal{B}(\mathcal{H}, \mathbb{R} \times \mathbb{R})$ is given by

$$Y^* x = \begin{bmatrix} \langle s_k, x\rangle \\ \langle H_k y_k, x\rangle \end{bmatrix}. \tag{3.25}$$

Also, let $G : \mathbb{R} \times \mathbb{R} \to \mathbb{R} \times \mathbb{R}$ be given by

$$G(\alpha, \beta) = \begin{bmatrix} \gamma\alpha + \gamma^2\langle H_k y_k, y_k\rangle\alpha - \gamma\beta \\ -\gamma\alpha \end{bmatrix}.$$

Above, we used the notation $\gamma = (\langle s_k, y_k\rangle)^{-1}$. We remark that G is a linear bounded invertible operator and has a bounded inverse in the form of

$$G^{-1}(\omega, \nu) = \begin{bmatrix} -\frac{\nu}{\gamma} \\ -\frac{\omega}{\gamma} - \frac{\nu}{\gamma}(1 + \gamma\langle H_k y_k, y_k\rangle) \end{bmatrix}.$$

One can show that $[G^{-1} + Y^* H_k^{-1} Y](\alpha, \beta) = \begin{bmatrix} \alpha\langle s, B_k s\rangle & -\frac{\beta}{\gamma} \end{bmatrix}$, which implies that $G^{-1} + Y^* H_k^{-1} Y$ is invertible and

$$(G^{-1} + Z^* H_k^{-1} Y)^{-1} = \begin{bmatrix} \frac{1}{\langle s_k, B_k s_k\rangle} & 0 \\ 0 & -\gamma \end{bmatrix}. \tag{3.26}$$

Note we chose $Z = Y$. In the above we used the matrix notation to be close to the the standard finite dimensional notation, but one can write this in operator notation as well. We next notice that

$$
\begin{aligned}
[H_k + YGY^*]x &= H_k x + Y(G(Y^* x)) = H_k x + Y(G\left(\begin{bmatrix} \langle s_k, x\rangle \\ \langle H_k y_k, x\rangle \end{bmatrix}\right) \\
&= H_k x + Y\left(\begin{bmatrix} \gamma\langle s_k, x\rangle + \gamma^2\langle H_k y_k, y_k\rangle\langle s_k, x\rangle - \gamma\langle H_k y, x\rangle \\ -\gamma\langle s_k, x\rangle \end{bmatrix}\right) \\
&= H_k x + \big(\gamma\langle s_k, x\rangle + \gamma^2\langle H_k y_k, y\rangle\langle s_k, x\rangle - \gamma\langle H_k y_k, x\rangle\big)\, s_k \\
&\quad - \gamma\langle s_k, x\rangle H_k y \\
&= H_k x - \gamma(s_k \otimes H_k y_k)(x) - \gamma(H_k y_k \otimes s_k)(x) \\
&\quad + (s_k \otimes \gamma^2\langle H_k y_k, y_k\rangle s_k)(x) + \gamma(s_k \otimes s_k)(x).
\end{aligned}
$$

On the other hand, the inverse Hessian BFGS formula from Proposition 3 can be manipulated as follows:

$$
\begin{aligned}
H_{k+1} &= (I - \gamma(s_k \otimes y_k))H_k(I - \gamma(y_k \otimes s_k)) + \gamma(s_k \otimes s_k) \\
&= H_k - \gamma(s_k \otimes y_k)H_k - \gamma(H_k y_k \otimes s_k) + \gamma^2(s_k \otimes y_k)H_k(y_k \otimes s_k) + \gamma(s_k \otimes s_k).
\end{aligned}
$$

Therefore $H_k + YGY^* = H_{k+1}$. Finally, from the Sherman–Morrison–Woodbury formula formula and by using the fact that $H_k^{-1} = B_k$ we obtain that

$$
\begin{aligned}
B_{k+1} = H_{k+1}^{-1} &= H_k^{-1} - H_k^{-1}Y(G^{-1} + Y^* H_k^{-1} Y)^{-1} Y^* H_k^{-1} \\
&= B_k - B_k Y(G^{-1} + Y^* B_k Y)^{-1} Y^* B_k.
\end{aligned}
$$

The definition of Y, equations (3.25) and (3.26), and the properties of dyadic products can be used to write for any $x \in \mathcal{H}$ that

$$\begin{aligned}
B_{k+1}x &= B_k x - B_k Y \begin{bmatrix} \frac{1}{\langle s_k, B_k s_k\rangle} & 0 \\ 0 & -\gamma \end{bmatrix} \begin{bmatrix} \langle s_k, B_k x\rangle \\ \langle H_k y_k, B_k x\rangle \end{bmatrix} \\
&= B_k x - B_k Y \begin{bmatrix} \frac{\langle s_k, B_k x\rangle}{\langle s_k, B_k s_k\rangle} \\ -\frac{\langle H_k y_k, B_k x\rangle}{\langle s_k, y_k\rangle} \end{bmatrix} \\
&= B_k x - B_k \left(\frac{\langle s_k, B_k x\rangle}{\langle s_k, B_k sv\rangle} s_k - \frac{\langle H_k y_k, B_k x\rangle}{\langle s_k, y_k\rangle} H_k y_k \right) \\
&= B_k x - \frac{\langle s_k, B_k x\rangle}{\langle s_k, B_k s\rangle} B_k s_k + \frac{\langle H_k y_k, B_k x\rangle}{\langle s_k, y_k\rangle} y_k \\
&= B_k x - \frac{B_k s_k \otimes s_k}{\langle s_k, B_k s_k\rangle} B_k x + \frac{y_k \otimes y_k}{\langle s_k, y_k\rangle} x \\
&= B_k x - \frac{B_k s_k \otimes B_k s_k}{\langle s_k, B_k s_k\rangle} x + \frac{y_k \otimes y_k}{\langle s_k, y_k\rangle} x.
\end{aligned}$$

This completes the proof of (3.24) and also shows that B_{k+1} is invertible. It remains to show $B_{k+1} \in \mathcal{B}^s(\mathcal{H})$, $B_{k+1} - B_k \in \mathcal{B}_{00}(\mathcal{H})$ and part *(ii)* *i.e.*, to show B_{k+1} is positive. Both are consequence of Proposition 3 which gives us $H_{k+1} \in \mathcal{B}^s(\mathcal{H})$, $H_{k+1} - H_k \in \mathcal{B}_{00}(\mathcal{H})$ and H_{k+1} is positive. One way to show both in one step is to use the fact $B_{00}(\mathcal{H})$ and the space of positive bounded linear operators are an ideal in the space of bounded operators.

□

Proposition 7 (DFP formula for the inverse Hessian operator)**.** *Let us consider the positive definite and invertible operators $H_k \in \mathcal{B}^s(\mathcal{H})$ and $W \in \mathcal{B}^s(\mathcal{H})$ such that $Wy_k = s_k$, where s_k and y_k are nonzero elements of $\mathcal{H}$.*

(i) The solution to the variational problem

$$\min_{H \in \mathcal{B}(\mathcal{H})} \frac{1}{2} \left\| W^{1/2}(H^{-1} - H_k^{-1})W^{1/2} \right\|_{HS}^2 \tag{3.27}$$

$$s.t \quad Hy_k = s_k \tag{3.28}$$

is given by

$$H_{k+1} = H_k - \frac{(H_k y_k \otimes H_k y_k)}{\langle y_k, H_k y_k\rangle} + \frac{(s_k \otimes s_k)}{\langle s_k, y_k\rangle}. \tag{3.29}$$

Furthermore, H_{k+1} is invertible and $H_{k+1} - H_k$ lies in $\mathcal{B}_{00}(\mathcal{H})$.

(ii) If the curvature condition is met $\langle y_k, s_k\rangle > 0$, then $H_{k+1} \in \mathcal{B}_{00}(\mathcal{H})$ is self-adjoint and positive definite.

Proof. The proof is similar to the proof of Proposition 6 and uses Theorem 5 for the inverse Hessian operator given by Proposition 2. □

3.4.1 Note on the limited-memory compact representation formulas

The popular limited-memory compact representations introduced by Byrd et al. [92] have similar forms for Hessian operators defined over general Hilbert spaces. For completeness, we succintly present them below. Their derivation is analogous to the finite-dimensional case [92] and relies on the Sherman–Morrison–Woodbury formula (Theorem 5) along the lines of the proof of Proposition 6. In what follows, given $s_i \in \mathcal{H}$ and $y_i \in \mathcal{H}$, $i = \{0, 1, \ldots, l-1\}$, let $S_l : \mathbb{R}^l \to \mathcal{H}$ be given by $S_l(v) = \sum_{i=0}^{l-1} v_{i+1}s_i$, and $Y_l : \mathbb{R}^l \to \mathcal{H}$ given by $Y_l(v) = \sum_{i=0}^{l-1} v_{i+1}y_i$, where v_i denotes the i^{th} component of a vector $v \in \mathbb{R}^l$. Furthermore, define R_l as a $l \times l$ matrix as

$$(R_l)_{ij} = \begin{cases} \langle s_{i-1,} y_{j-1}\rangle & \text{if } i \leq j, \\ 0 & \text{otherwise.} \end{cases}$$

Theorem 7. *Let $H_0 \in B^s(\mathcal{H})$ be positive and invertible. Furthermore, let H_l be given by updating H_0 l times using the inverse BFGS formula obtained in Proposition 3. If all the pairs $\{s_i, y_i\}_{i=0}^{l-1}$ satisfy the positive curvature condition $\langle s_i, y_i\rangle > 0$, then*

$$H_l = H_0 + \begin{bmatrix} S_l & H_0Y_l \end{bmatrix} \left(\begin{bmatrix} R_l^{-T}(D_l + (H_0Y_l)^*Y_l)R_l^{-1} & -R_l^{-T} \\ -R_l^{-1} & 0 \end{bmatrix} \begin{bmatrix} S_l^* \\ (H_0Y_l)^* \end{bmatrix} \right),$$

where D_l is the $l \times l$ diagonal matrix given by

$$(D_l)_{ij} = \begin{cases} \langle s_{i,} y_j\rangle & \textit{if } i = j, \\ 0 & \textit{otherwise.} \end{cases}$$

Theorem 8. *Let $B_0 \in B^s(\mathcal{H})$ be positive and invertible. Furthermore, let B_l be given by updating B_0 l times using the BFGS formula obtained in Proposition 6. If all the pairs $\{s_i, y_i\}_{i=0}^{l-1}$ satisfy the positive curvature condition $\langle s_i, y_i \rangle > 0$, then*

$$B_l = B_0 - \begin{bmatrix} B_0 S_l & Y_l \end{bmatrix} \left(\begin{bmatrix} S_l^* B_0 S_l & L_l \\ L_l^T & -D_l \end{bmatrix}^{-1} \begin{bmatrix} (B_0 S_l)^* \\ Y_l^* \end{bmatrix} \right),$$

where L_l is the $l \times l$ matrix with entries

$$(L_l)_{ij} = \begin{cases} \langle s_{i-1,} y_{j-1} \rangle & \text{if } i > j, \\ 0 & \text{otherwise.} \end{cases}$$

3.5 Incorporating Hessian structure in quasi-Newton formulas: a case study for inverse problems governed by partial differential equations

As an illustration of potential uses of the results introduced by this thesis, we consider the class of regularized inverse problems governed by partial differential equations (PDEs) and derive DFP and BFGS "structured" secant formulas that explicitly use the derivative of the regularization and only approximates the second derivative of the misfit term. To this end, we consider the inversion of a coefficient field in an elliptic PDE. Depending on the interpretation of the inputs and the type of measurements, this problem arises, for instance, in inversion for the permeability field in a subsurface flow problem, for the conductivity field in a heat transfer problem, or the stiffness parameter field in a membrane deformation problem [117].

We formulate the inverse problem over $\Omega = [0,1] \times [0,1]$ as follows: given possibly noisy observations $\mathrm{d} \in \mathbb{R}^q$ of the state solution u, we wish to infer the coefficient field m that best reproduces the observations. Mathematically, this can be formulated as the nonlinear least-squares minimization problem

$$\min_m \mathcal{J}(m) := \frac{1}{2} \langle Ou(m) - \mathrm{d}, Ou(m) - \mathrm{d} \rangle_{\mathbb{R}^q} + \frac{\gamma}{2} \langle \nabla m, \nabla m \rangle_{L^2}, \tag{3.30}$$

$$\text{s.t.} \quad \underline{m} \leq m \leq \overline{m}, \tag{3.31}$$

where u solves the state (or forward) problem

$$-\nabla \cdot (m\nabla u) = f \text{ in } \Omega \text{ and } u = 0 \text{ on } \partial\Omega. \tag{3.32}$$

Above, $\mathrm{d} \in \mathbb{R}^q$ denotes the observations, with q denoting the number of observation spoints, $f \in H^{-1}(\Omega)$ is a given volume force, $O : L^2(\Omega) \to \mathbb{R}^q$ is a linear observation operator that extracts measurements from u, and $\underline{m}, \overline{m} \in L^\infty(\Omega)$ are the lower and upper bounds of the unknown coefficient field m, respectively. The first term in the objective of (3.30) is the data misfit term, which we will denote by $\mathcal{M}(m)$, and the second term, which we will denote by $\mathcal{R}(m)$, is a regularization term with regularization parameter $\gamma > 0$ added to render the inverse problem well-posed [132, 98]. We note that when we discretize the regularization term, this will take the form of $\mathbf{m}^T \boldsymbol{K} \mathbf{m}$, where $\mathbf{m}$ is the vector of finite element coefficients of the parameter field m, and $\boldsymbol{K}$ is the stiffness matrix [100, 113].

We solve (3.30) using a quasi-Newton interior-point method [112]. We assume that only the second derivative of the regularization term is available, while the second derivative of the misfit term is not (e.g., we target application problems for which this terms is expensive to evaluate). Therefore, to take advantage of this structure, in what follows, we derive and apply so-called *structured* DFP and BFGS formulas.

3.5.1 Derivation of structured DFP and BFGS formulas

To derive structured DFP and BFGS formulas for the Hessian matrix, we consider a structured variant of Proposition 2. More specifically, since we are looking for a DFP update in the form $B = R + A$, where A approximates the second-derivative of the misfit term $\mathcal{M}$, in the spirit of Proposition 2 we require that a formula for A satisfies

$$\begin{aligned} &\min_{A} \ \frac{1}{2} \left\| W^{1/2} \left(A - A_k\right) W^{1/2} \right\|_{HS}^2 \\ &\text{s.t} \quad A s_k = \bar{y}_k, \end{aligned} \tag{3.33}$$

where $\bar{y}_k = \nabla\mathcal{M}(m_{k+1}) - \nabla\mathcal{M}(m_k)$. In words, the variational form (3.33) builds the structured update A based only on the change in the gradient of the misfit.

Analogous to the proof of Proposition 2, one can show that the *structured DFP formula for the Hessian* is

$$A_{k+1} = (I - \bar{\gamma}(\bar{y}_k \otimes s_k))A_k(I - \bar{\gamma}(s_k \otimes \bar{y}_k)) + \bar{\gamma}(\bar{y}_k \otimes \bar{y}_k), \tag{3.34}$$

where $\bar{\gamma} = 1/\langle s_k, \bar{y}_k\rangle$. Similarly, the *structured BFGS formula for the Hessian* can be obtained by considering the structured version of Proposition 6 in the form of

$$\begin{aligned} &\min_{A} \frac{1}{2} \left\| W^{1/2} \left[(A+R)^{-1} - (A_k+R)^{-1} \right] W^{1/2} \right\|_{HS}^2, \\ &\text{s.t} \quad As_k = \bar{y}_k \end{aligned} \tag{3.35}$$

which gives the structured BFGS formula

$$A_{k+1} = A_k - \frac{(A_k+R)s_k \otimes (A_k+R)s_k}{\langle s_k, Rs_k + \bar{y}_k\rangle} + \frac{(\bar{y}_k + Rs_k) \otimes (\bar{y}_k + Rs_k)}{\langle s_k, Rs_k + \bar{y}_k\rangle}. \tag{3.36}$$

We remark that the Hessian formula $B_k = R + A_k$ with A_k given by (3.36) above is identical to the unstructured BFGS given by Proposition 6 as long as the two formulas are initialized with $B_0 = R + A_0$ and A_0. This is not the case for the structured and unstructured DFP formulas (3.34) and (3.15), respectively.

3.5.2 Numerical results

We compare the performance of structured update formulas derived in Section 3.5.1 with their unstructured counterparts for the inverse problem governed by the Poisson equation given by (3.30)-(3.32). The numerical algorithm we use is a filter line-search interior-point method for constrained optimization problems [134, 133] in which we replace the Hessian of the objective (3.30) with quasi-Newton approximations similarly to the state-of-the-art Ipopt solver [135]. The stopping criteria for the interior-point method consist of a stringent 10^{-8} tolerance for the norm of gradient (of the Lagrangian function of (3.30), for more details see [135]) and a maximum number of 100 iterations. We derive the gradient (i.e., the firs-derivative information) using an adjoint-based approach [113, 88, 126]. The underlying PDEs are solved with the finite element method using COMSOL with Matlab, while the interior-point method is implemented in Matlab. The problem was solved on five uniform 2D meshes and on one nonuniform 2D mesh with rectangular elements. For the discretization of the state and ajoint variables we used quadratic and for the parameter linear finite elements. The state dimension was increased form 441 to 5227 and the parameter dimension (i.e., the dimension of the optimization problem) from 121 to 1328. The numerical experiments were performed on an Intel Ivy Bridge 2.5GHz 8-Core Linux machine with 128 GB RAM memory.

In what follows, the structured quasi-Newton formulas are denoted with acronyms starting with "S-". These are compared with unstructured counterparts, which are prefixed by "U-". For a both fair play and comprehensive comparison, the unstructured quasi-Newton formulas are used with an *uninformed* (suffixed by "-U") and *informed* (suffixed by "-I") initial Hessian approximations. The uninformed initial approximations correspond to a plain, fully unstructured formula, while the informed initial approximation correspond to unstructured formulas that take into account the known part of the Hessian (that is, the Hessian of the regularization term). Table 3.1 summarizes this discussion and presents the algorithmic parameters used in the numerical experiments. The parameter multiple of the identity σ_k is the Barzilai-Borwein spectral estimate [86] that changes at each optimization iteration according to $\sigma_k = \langle s_k, s_k \rangle / \langle s_k, y_k \rangle$. This estimate is also used in Ipopt; in our experiments it gave the smallest number of iterations for all formulas from Table 3.1.

In Table 3.2, we report on the number of iterations for unstructured informed and uninformed and structured BFGS and DFP formulas. We have used these formulas with (s_k, y_k) pairs from the last ℓ iteration for $\ell = 8$ (a), $\ell = 16$ (b), and $\ell = 32$ (c). Our numerical experiments reveal that the standard unstructured updates with *uninformed* initialization, namely U-BFGS-U and U-DFP-U, exhibit a number of iterations that increases for finer or non-uniform meshes. This mesh dependence behavior is present for all three memory sizes $\ell = 8$, $\ell = 16$, and $\ell = 32$ we have used. On the other hand, the standard unstructured formulas with *informed* initialization, namely U-BFGS-I and U-DFP-I, do not show this mesh dependent behavior; instead, the iteration count for these updates remains relatively

Acronym	Formula for B_k	Initial Hessian	Notes
U-BFGS-U	(3.24)	$B_0 = \sigma_k M$	unstructured BFGS with uninformed initialization
U-BFGS-I	(3.24)	$B_0 = \sigma_k(M + R)$	unstructured BFGS with informed initialization
S-BFGS	$B_k = A_k + R$ A_k given by (3.36)	$A_0 = \sigma_k M$	structured BFGS
U-DFP-U	(3.15)	$B_0 = \sigma_k M$	unstructured DFP with uninformed initialization
U-DFP-I	(3.15)	$B_0 = \sigma_k(M + R)$	unstructured DFP with informed initialization
S-DFP	$B_k = A_k + R$ A_k given by (3.33)	$A_0 = \sigma_k M$	structured DFP

Table 3.1: Summary of the formulas investigated numerically in this section. The algorithmic parameter σ_k is the Barzilai-Borwein spectral estimate [86] discussed in the text.

constant for all meshes. Our point is that in order to obtain mesh independence, one needs not only to use the infinite-dimensional BFGS and DFP formulas but also to carefully choose the initial quasi-Newton approximation operator. Intuitively, for the inverse problem we solve here, the use of an informed initialization

Mesh	(a) Number of iterations for $\ell = 8$					
	U-BFGS-U	U-BFGS-I	S-BFGS	U-DFP-U	U-DFP-I	S-DFP
10×10	41	35	37	39	30	34
20×20	87	43	39	95	38	37
30×30	>100	41	38	>100	39	36
40×40	>100	42	39	>100	45	52
50×50	>100	46	39	>100	44	36
non-unif.	>100	43	40	>100	46	39
	(b) Number of iterations for $\ell = 16$					
	U-BFGS-U	U-BFGS-I	S-BFGS	U-DFP-U	U-DFP-I	S-DFP
10×10	39	29	30	38	27	29
20×20	78	36	34	90	35	32
30×30	>100	36	33	>100	39	32
40×40	>100	36	33	>100	39	32
50×50	>100	39	35	>100	38	34
non-unif.	>100	36	34	>100	38	39
	(c) Number of iterations for $\ell = 32$					
	U-BFGS-U	U-BFGS-I	S-BFGS	U-DFP-U	U-DFP-I	S-DFP
10×10	37	28	28	37	26	29
20×20	77	34	29	87	33	32
30×30	>100	35	30	>100	37	30
40×40	>100	38	30	>100	37	35
50×50	>100	37	30	>100	37	37
non-unif.	>100	37	30	>100	36	34

Table 3.2: Shown are the number of optimization iterations obtained with formulas from Table 3.1 with quasi-Newton memory for $\ell = 8$ (a), $\ell = 16$ (b), and $\ell = 32$ (c).

with U-BFGS-I and U-DFP-I, namely a multiple of the identity operator plus the stiffness operator, circumvents the need to approximate the stiffness operator; instead these formulas approximate only the Hessian of the misfit, which is known to be compact [99, 89] and, therefore, can be approximated relatively well (both in a mesh independent manner and within a relatively small number of iterations) by the finite-rank operators built using the infinite-dimensional BFGS and DFP formulas derived in this thesis.

We now turn to the *structured* BFGS and DFP formulas, *i.e.*, S-BFGS and S-DFP, which we derived in this section to explicitly incorporate additional Hessian information (namely the stiffness operator). We remark from Tables 3.2 (a)–(c) that these structured formulas improve over the unstructured informed formulas U-BFGS-I and U-DFP-I in terms of number of iterations (by up to 20%) and, also, exhibit mesh independence behavior. In particular, we remark that S-BFGS shows a more consistent iteration count over all meshes when compared to S-DFP; and, for larger quasi-Newton memory sizes ($\ell = 32$), S-BFGS seems slightly faster than S-DFP, while for smaller memory sizes the two compare similarly.

3.6 Conclusions

We have presented a new derivation of well-known quasi-Newton formulas in an infinite-dimensional Hilbert space setting needed for example for solving optimization problems governed by differential equations. In particular, we have generalized the variational, least-squares framework of Güler et al. [103] to operators defined over general separable Hilbert spaces. The framework we present was used to derive classical BFGS, DFP, PSB, and SR1 formulas in operator form. Furthermore, we illustrated how the variational framework can be employed to derive improved DFP and BFGS updates for a class of inverse problems governed by PDEs. To illustrate the importance of using these infinite-dimensional quasi-Newton formulas we formulated and solved an inverse problem governed by partial differential equations (PDEs) via a quasi-Newton interior-point method on progressively finer uniform meshes and on a nonuniform mesh. In addition, we derived structured DFP and BFGS formulas for the Hessian operator, where we considered parts of the Hessian known and only approximate the remaining part (e.g., the second-derivative of the term corresponding to the misfit). Numerical results showed that in order to obtain mesh independence, it is essential not only to use the infinite-dimensional BFGS and DFP formulas but also to carefully choose the initial quasi-Newton approximation operator. In addition, we compared the performance of the structured update formulas with their unstructured counterparts and found that taking into account the structure of the problem leads to reducing further the computational cost.

Acknowledgments

This work was partially supported by the U.S. National Science Foundation, Division of Mathematical Sciences, Computational Mathematics under grant CAREER-1654311 and the summer internship at Lawrence Livermore National Laboratory under Dr. Cosmin G. Petra.

Chapter 4

Inexact Hessian-applies for Inverse Problems Governed by PDEs

4.1 Introduction

Second-order, Newton-like algorithms exhibit convergence properties superior to gradient-based or derivative-free optimization algorithms [111]. However, deriving and computing second-order derivatives needed for the Hessian-vector products in a Krylov iteration for the Newton step often is not trivial. As shown and discussed in Chapter 1, second-order adjoints (also called incremental state and adjoints) provide a systematic and efficient means to derive second derivative information for solving optimization problems efficiently. For inverse problems governed by PDEs, one Hessian-vector product costs two (linear) PDE solves. For many applications and problems, a high number of such Hessian-vector products may be needed for convergence. For large-scale problems often we cannot afford a very large number of such solves. For instance in [114], the authors solve a reasonable size ice sheet inverse problem with inexact Newton-CG and show that the cost of solving the optimization problem (up to tolerance 10^{-5}) was about 7000 PDE solves. When solving the same inverse problem on the continental/Antarctica scale, the computational cost per Newton iteration increased to about 100,000 PDE solves and this for a much higher tolerance and with a maximum number of CG iterations set to 250 [9]. For complex problems, such as the ice sheet inverse problem, there is a need to develop methods that reduce the number of PDE solves required for convergence.

There are a number of ways to reduce the computational cost. For instance, via efficient preconditioners for the Newton system (e.g., [90], via inexact Newton-CG solves [113], via low-rank approximations of the Hessian [43], and inexact Hessian-vector products (i.e., inexact second-order adjoint solves) [40]. In [40] the author shows bounds for the tolerances for solving the second-order adjoints inexactly that ensure the Hessian-vector product remains sufficiently accurate for inexact-Krylov methods. These bounds are of the typical inexact-Newton type where further in the iteration we are more computational effort we exert, in practice, this could be a problem for large-scale inversion. Their numerical results have been obtained with Newton-GMRES algorithm [61, 60]. In this work we investigate means of setting the tolerance for inexactness dynamically, while retaining robustness.

In this thesis, we follow the framework presented in [77] to obtain analytical Hessian-vector products. In section 4.2 we give the necessary background that leads to the Hessian-vector product. The main theorem that shows the dynamic tolerances is given in Section 4.4.1. This result shows that the inexactness can increase with the number of iterations of the inverse problem. While increase of the tolerance (inexactness) sounds counterintuitive, this has been previously observed in the following previous related works [49, 48, 47]. In [45] the authors also explored the application of Krylov methods to $(A + E)x = b$ and devolved a framework for inexact Krylov methods with increasing perturbation. They observe the rate of convergence and stability of Krylov methods as the norm of the perturbation matrix E grows. One of their conclusions was that the stopping criteria for the Krylov method is also allowed to increase as the norm of E increases. While their work was in the realm of linear algebra we found it inspirational to this project. We show numerical results for an inverse problem governed by a Poisson problem in Section 4.6. Our results reveal that close to the solution of the inverse problem, the tolerance of the second-order adjoint solves can be relaxed, which leads to reducing the number of inner Krylov iterations.

Problem formulation To set the stage, we choose a general framework: let Y, U, Z be Banach spaces, e.g. Y is the space variable or dependent variable, U the space of control or design variables, and Z the range space of the equality constraint. The optimization problem is formulated as follows:

Problem 4.1.1.

$$\begin{aligned} \min_{y,u} \quad & \phi(y,u), \quad (y,u) \in Y \times U && (4.1)\\ s.t. \quad & g(y,u) = 0 && (4.2)\\ \textit{where} \quad & \phi : Y \times U \to I\!R, \quad g : Y \times U \to Z. && (4.3) \end{aligned}$$

If we assume that for each control variable u, we have a unique solution $y = s(u)$ of the equality constraint $g(s(u), u) = 0$, then we can rewrite the constrained optimization problem as an unconstrained optimization problem, namely

$$\min_{u} \quad \phi(s(u), u), \quad u \in U.$$

It is well known that the gradient of this function can be expressed in two ways [77]. Either using

- the sensitivity equations or
- the adjoint equations.

The first approach is considered reasonable for a small number of variables, whereas the second one requires more analysis in the derivation, but shows to be highly efficient for large scale problems. If we turn to the second derivative applied to a vector as this required for iterative solvers like conjugate gradient (CG) or generalized minimal residual method (GMRES), we are free to choose for this purpose again either the adjoint or sensitivity approach. It was shown in [77] that this results in two possible implementable schemes: one with a second order sensitivity equation or one with a second order adjoint equation. We concentrate in this chapter of the thesis on the approach using a second order adjoint. The effort per iteration is comparable to that of an inexact Newton's method, where the matrix-vector multiplication is approximated by a finite difference quotient, yet it gives the precise result rather than an approximation.

4.2 First- and Second-Order Fréchet-Derivative

For notational purposes we recall that a map $g : X \to Z$ from a Banach space X to another Banach space Z is called Fréchet-differentiable at $x \in X$, if there exists a linear operator denoted by $g'(x) : X \to Z$ such that

$$\|g(x+h) - g(x) - g'(x)h\|_Z \le \alpha(\|h\|_X)\|h\|_X,$$

with a function $\alpha(r)$ satisfying $\alpha(r) \to 0$ for $r \to 0$. The partial Fréchet-derivatives of e.g., $g(y,u)$ is denoted by $g_y(y,u)$ or $g_u(y,u)$ with a subscript indicating the variable the derivative is taken. The adjoint operator of $g'(x)$ is denoted by $g'(x)^* : X^* \to Z^*$. We note that Fréchet-derivatives of second order like $g_{yu}(y,u)$ are linear operators in the spaces $L(U, L(Y,Z)) = L(U \times Y, Z)$. In what follows, we impose the following smoothness assumptions on the functions in the problem formulation.

Assumption 4.2.1. *Let the function ϕ and the mapping g be continuously Fréchet-differentiable on $Y \times U$.*

Furthermore, we assume the following constraint qualification to hold at a later to be specified point $(y,u) \in Y \times U$.

Assumption 4.2.2. *For $(y,u) \in Y \times U$ let the partial Fréchet-derivative $g_y(y,u) : Y \to Z$ be surjective and invertible.*

With these assumptions we can apply the implicit function theorem.

Theorem 4.2.1. *Let assumptions 4.2.1 and 4.2.2 hold at $(y_*, u_*) \in Y \times U$. Then there exist neighborhoods $B_Y \subset Y$ at y_* and $B_U \subset U$ at u_* and a Fréchet-differentiable map $s : B_U \to B_Y$ such that*

$$g(s(u),u) = 0 \quad \textit{and} \quad g_y(s(u),u)s'(u) = -g_u(s(u),u). \tag{4.4}$$

This theorem can be used to reformulate the constrained optimization problem from above as an unconstrained optimization problem, namely

$$\min_{u \in B_U} \Phi(u), \quad \Phi(u) := \phi(s(u), u). \tag{4.5}$$

The necessary optimality conditions of first order require various derivatives which are well defined under the statements above. Therefore, the first derivative of the objective function of the unconstrained problem can be computed as follows using the adjoint approach.

Theorem 4.2.2. *Let assumptions 4.2.1 and 4.2.2 hold at $(y, u) \in Y \times U$. Then we obtain*

$$\Phi'(u) = g_u(s(u), u)^* p + \phi_u(s(u), u) \in U^*, \tag{4.6}$$

where $p \in Z^*$ is defined as the solution of the adjoint equation

$$g_y(s(u), u)^* p = -\phi_y(s(u), u) \in Y^*. \tag{4.7}$$

Alternatively, the action of the adjoint variable $p \in Z^*$ is given by

$$p(z) = -\phi_y(s(u), u) g_y(s(u), u)^{-1} z \quad \forall z \in Z. \tag{4.8}$$

In the following we give the second derivative of the function Φ. A rigorous derivation of both the gradient and second derivative can be found in [77]. Since in many algorithmic applications, *e.g.*, the use of iterative solvers for the Newton step, the complete Hessian information $\Phi''(u)$ is not needed, we concentrate on the computation of the Hessian-vector product $\Phi''(u)\Delta v$. Obviously, we need to strengthen assumption 4.2.1 to

Assumption 4.2.3. *Let the functional ϕ and the map g be twice continuously Fréchet-differentiable on $Y \times U$.*

As noted before, the notation using second derivatives can be somewhat complex, since the second partial derivative $g_{yu}(y, u)$ can be interpreted as a map in $L(U, L(Y, Z))$ or $L(U \times Y, Z)$, which gives multiple perspectives of how to view second derivatives.

Theorem 4.2.3. *The second derivative $\Phi''(u)$ applied to Δv can be written as*

$$\Phi''(u)\Delta v = g_u(y,u)^*\pi + (g_{yu}(y,u)\xi)^*p + \phi_{yu}(y,u)\xi + (g_{uu}(y,u)\Delta v)^*p + \phi_{uu}(y,u)\Delta v,$$

where y, u satisfy $g(y,u) = 0$ and, p solves the adjoint equation (4.7). Furthermore, ξ solves the sensitivity equation of first order

$$g_y(y,u)\xi = -g_u(y,u)\Delta v, \tag{4.9}$$

and π the second order adjoint equation

$$g_y(y,u)^*\pi = -(g_{yy}(y,u)\xi)^*p - \phi_{yy}(y,u)\xi - (g_{uy}(y,u)\Delta v)^*p - \phi_{uy}(y,u)\Delta v. \tag{4.10}$$

Let us compare this with an inexact Newton method, where the Hessian-vector multiplication $\Phi''(u)\Delta v$ is approximated by one finite difference quotient in direction Δv:

$$\Phi''(u)\Delta v \approx \frac{1}{h}[\Phi'(u + h\Delta v) - \Phi'(u)].$$

First, a proper choice of h can be quite delicate, since an h too small magnifies any error tolerances in the gradient computations. Second, the additional gradient evaluation requires a nonlinear system solve for y and an additional adjoint solve for p. In our approach we need a linear instead of a nonlinear system solve for ξ and only one additional adjoint solve (4.10) for π.

4.3 Inexact Newton

It is well accepted in the inverse community the use of an inexact Newton's method, in particular when one solves the linear systems which involve the Hessian by iterative solvers. Here we take the opportunity to discuss some aspects in connection with the second order adjoints.

Consider Hessian-vector multiplication needed for a Krylov-type solver in the Newton equation, can be approximated by

$$\Phi''(u_k)v \approx (\Phi'(u_k + h_k v) - \Phi'(u_k))/h_k,$$

for h_k sufficiently small. Hence each Krylov iteration requires the solution of another state and adjoint equation. However, if one uses the concept of second-order adjoints, one does not need to approximate the Hessian-vector product, but can compute it exactly at the expense of the solution of one sensitivity equation and the computation of a second order adjoint. One advantage of our approach is that while the state and first order adjoint equation must be solved to full accuracy in order to obtain the correct gradient information, the solution of the sensitivity and the second order adjoint equation could be performed at a lower accuracy. This introduces only an error in the Hessian-vector product which can be monitored within the framework of an inexact Newton approach. If one wants to avoid the computation of certain second derivatives for the functions ϕ or the mapping g, then one can replace these by finite difference quotients. For example, in the right-hand side of the second order adjoint equation for the first term

$$\text{replace} \quad g_{yy}(y,u)\xi \quad \text{by} \quad [g_y(y+h\xi,u) - g_y(y,u)]/h.$$

The general expectation is as we relax the solves of the second adjoint and sensitivity in 4.2.3 we introduce more inexactness into the Hessian apply leading us to more Newton iterations. However, in our work given some assumptions we can retain good convergence rates while relaxing those solves moderately. In our numerical experiments we observed the same number of Newton iterations for solving the system exactly and using relaxed (dynamic) tolerances. While the number of Newton iterations was the same the computational effort in the latter case was less reflected in the number of Krylov iterations.

However, one can use the framework also for inexact gradient information, which would allow also to compute the solution of the system equation y and the adjoint equation p up to a certain accuracy. This is formulated in the following theorem, where in contrast to the usual theory on inexact Newton's method, the measure for the inexact solves is not required to be driven to zero by the size of the residual but rather by an independent given sequence of numbers converging to zero.

Theorem 4.3.1. *Assume that each Newton step is solved inexactly*

$$\Phi''(u_k)\Delta u_k = -\Phi'(u_k) + r_k, \qquad u_{k+1} = u_k + \Delta u_k.$$

If for some $c \in (0,1)$ we have for all k

$$\|r_k\| \leq \rho_k, \qquad \rho_{k+1} \leq c\rho_k,$$

then there exists an $\epsilon > 0$ such that if $\|u_0 - u_*\| \leq \epsilon$ and $\Phi''(u_*)$ is invertible, the sequence u_k converges to u_* at a r-linear rate of convergence.

Proof. The convergence analysis for Newton's method yields

$$u_{k+1} = \Phi''(u_k)^{-1}[\Phi''(u_k)u_k - \Phi'(u_k) + r_k].$$

With $e_k := u_k - u_*$ and the boundedness of $\|\Phi''(u)^{-1}\|$ for $\|u - u_*\| \leq \epsilon$ for sufficiently small ϵ, we obtain for $\|u_k - u_*\| \leq \epsilon$

$$\|e_{k+1}\| \leq L_1\|\Phi''(u_k)(u_k - u_*) - \Phi'(u_k) + \Phi'(u_*) + r_k\| \leq L_2(\|e_k\|^2 + \rho_k).$$

Choose $\epsilon > 0$ so small such that also $\epsilon \leq 1/(2L_2)$ and k_0 so large such that $c^k \leq \epsilon/(2L_2\rho_0)$. Hence

$$\|e_{k+1}\| \leq L_2\epsilon^2 + L_2c^k\rho_0 \leq \epsilon/2 + \epsilon/2 = \epsilon,$$

which completes the induction argument that u_k stay in an $\epsilon-$neighborhood of u_*.

We turn to the proof of the rate of convergence. We choose $\epsilon > 0$ so small that $L_2\epsilon + c < 1$. Then define

$$\chi_{k+1} = L_2(\epsilon\chi_k + \rho_k), \quad k \geq 1, \quad \chi_0 = \rho_0 L_2/c,$$

that converges at a linear rate to zero, because by induction

$$\begin{aligned}\chi_{k+1}/\chi_k &= L_2\epsilon + L_2\rho_k/\chi_k = L_2\epsilon + \rho_k/(\epsilon\chi_{k-1} + \rho_{k-1})\\ &\leq L_2\epsilon + \rho_k/\rho_{k-1} \leq L_2\epsilon + c < 1.\end{aligned}$$

Finally, we show that $\|e_k\|$ is bounded by χ_k which proves the r-linear rate. If this is true for k, then for $k+1$

$$\|e_{k+1}\| \leq L_2(\|e_k\|^2 + \rho_k) \leq L_2(\|e_k\|\chi_k + \rho_k) \leq \chi_{k+1}.$$

In order to see for the initiation of the induction that $\|e_0\| \leq \chi_0 = \rho_0 L_2/c$ holds, we can choose ϵ so small, that $\|e_0\| \leq \epsilon \leq \chi_0 = \rho_0 L_2/c$ is true. □

This is a typical result one may find in the literature of inexact Newton. We choose to prove this result as it is in the form most suited for our work, however this Theorem can be viewed as restatement of Theorems 6.1.1 and 6.1.2 from [52].

4.4 Inexact Richardson Iteration

As an example, for an iterative solver of linear equations, in our case the Newton system, we look at the Richardson iteration, since it can be analyzed in details without much effort. Most of the final estimates also hold for general Krylov methods such as Generalized Minimal Residual Method (GMRES) and Conjugate Gradient Method (CG), as proved in the works by the authors in [45, 56, 55]. Consider a symmetric positive definite matrix $A \in I\!R^{n\times n}$ and find $x_* \in I\!R^n$ for some $b \in I\!R^n$ with

$$Ax_* = b$$

or equivalently for any $\alpha \in I\!R$ with $\alpha \neq 0$

$$x_* = x_* + \alpha(b - Ax_*).$$

Richardson's method is an iterative procedure with

$$x_i = x_{i-1} + \alpha(b - Ax_{i-1}), \quad i \in I\!N, \quad x_0 = 0.$$

This results using $Ax_* = b$ in

$$x_i - x_* = x_{i-1} - x_* + \alpha(b - Ax_{i-1}) = (I - \alpha A)(x_{i-1} - x_*),$$

which is convergent for

$$\alpha = \frac{2}{\lambda_- + \lambda_+}, \tag{4.11}$$

where λ_-, and λ_+ are the smalles and largest eigenvalues of A, respectively. This is because for spectral matrix norm we get

$$\|I - \alpha A\| = 1 - \frac{2\lambda_-}{\lambda_- + \lambda_+} = \frac{\lambda_+ - \lambda_-}{\lambda_+ + \lambda_-} = \frac{cond_2(A) - 1}{cond_2(A) + 1} < 1,$$

where $cond_2(A) = \|A\|_2\|A^{-1}\|_2$ is the condition number of A. Updating the residual by $r_i = b - Ax_i$, as usual for Krylov methods, we can rewrite one step of the Richardson iteration as

$$\begin{aligned} r_i &= r_{i-1} - \alpha A r_{i-1}, \quad i \in I\!N, \quad r_0 = b \\ x_{i+1} &= x_i + \alpha r_i, \qquad\qquad i \in I\!N, \quad x_1 = \alpha b. \end{aligned} \tag{4.12}$$

For the further analysis let us assume that the matrix vector multiplication exhibits an error quantified by g_i. To be more precise, we assume at iteration i that the matrix vector product has an error: A applied to an arbitrary vector z is not Az, but rather $Az + g_i$. Here the error term g_i, in particular its size, could change from iteration to iteration, namely

$$\begin{aligned} \tilde{r}_i &= \tilde{r}_{i-1} - \alpha(A\tilde{r}_{i-1} + g_{i-1}), \quad i \in I\!N, \quad \tilde{r}_0 = b \\ \tilde{x}_{i+1} &= \tilde{x}_i + \alpha\tilde{r}_i, \qquad\qquad\qquad i \in I\!N, \quad \tilde{x}_1 = \alpha b. \end{aligned} \tag{4.13}$$

It is to be noted that similar results can be found in [51, 45] other works, but the authors keep the error in the matrix vector product not as $Ax + g$ with an error vector g, but in matrix form $(A + E)x$ with an error matrix E instead.

Lemma 2. *The true residual $b - A\tilde{x}_i$ and the computed residual $\tilde{r}_i$ of the inexact iteration imply the following error estimate*

$$\|\tilde{r}_i - (b - A\tilde{x}_i)\| \leq \alpha \sum_{j=0}^{i-1} \|g_j\|. \tag{4.14}$$

Proof. If we define

$$z_i = \tilde{r}_i - (b - A\tilde{x}_i), \quad i \in I\!N, \quad z_0 = 0$$

then

$$z_i = \tilde{r}_{i-1} - \alpha(A\tilde{r}_{i-1} + g_{i-1}) - (b - A(\tilde{x}_{i-1} + \alpha\tilde{r}_{i-1}) = z_{i-1} - \alpha g_{i-1}$$

and hence

$$z_i = z_{i-1} - \alpha g_{i-1} = ... = z_0 - \alpha \sum_{j=0}^{i-1} g_j.$$

From this, the statement of the lemma follows immediately. □

Lemma 2 states that the error in the perturbed residuals can accumulate as the iteration progresses. This is not the case, if we compare it to the residuals of the exact iteration. The following Lemma is explores how much is the difference between the residual of the exact iteration and the perturbed.

Lemma 3. *We have the following error estimate for the residuals of the inexact and exact iteration:*

$$\|\tilde{r}_i - r_i\| \leq \alpha \| \sum_{j=0}^{i-1} (I - \alpha A)^{i-1-j} g_j \|. \tag{4.15}$$

Proof. If we define

$$z_i = \tilde{r}_i - r_i, \quad i \in I\!N, \quad z_0 = 0$$

then

$$z_i = \tilde{r}_{i-1} - \alpha(A\tilde{r}_{i-1} + g_{i-1}) - r_{i-1} + \alpha A r_{i-1} = (I - \alpha A) z_{i-1} - \alpha g_{i-1}.$$

This leads to

$$z_i = (I - \alpha A)^i z_0 - \sum_{j=1}^{i} (I - \alpha A)^{j-1} \alpha g_{i-j}$$

and to (4.15) by rearranging the indices. □

If one compares the estimates (4.14) and (4.15) one realizes that the error in (4.15) contains a damping factor. This can be used to improve the estimate for $\|\tilde{r}_i - r_i\|$ such that it stays bounded in contrast to the estimate in $\|\tilde{r}_i - r_i\|$ where an accumulation of error cannot be avoided.

Lemma 4. *If the error g_i follows*

$$\|g_{i-1}\| \leq \epsilon \quad i \in I\!N,$$

then we obtain

$$\|\tilde{r}_i - (b - A\tilde{x}_i)\| \leq \alpha i \epsilon,$$

where as

$$\|\tilde{r}_i - r_i\| \leq \|A^{-1}\| \epsilon. \tag{4.16}$$

with α chosen by (4.11).

Proof. From the proof of the previous lemma we have

$$\|z_i\| \leq \alpha \| \sum_{j=1}^{i} (I - \alpha A)^{j-1} \| \| g_{i-j} \| \leq \alpha \sum_{j=1}^{i} \|I - \alpha A\|^{j-1} \epsilon$$

and by the geometric series

$$\|z_i\| \leq \alpha \sum_{j=1}^{\infty} \left(\frac{\lambda_+ - \lambda_-}{\lambda_+ + \lambda_-} \right)^{j-1} \epsilon = \alpha \frac{\lambda_- + \lambda_+}{2\lambda_-} \epsilon = \frac{1}{\lambda_-} \epsilon.$$

Similarly we have

$$\|z_i\| = \|z_{i-1} - \alpha g_{i-1}\| = \|z_0 - \alpha \sum_{j=0}^{i-1} g_j\| \leq \alpha \sum_{j=0}^{i-1} \epsilon, \tag{4.17}$$

which concludes our result. □

In [56, 55] the error g_j is measured as a relative error. This leads to the following corollary.

Corollary 2. *If the relative error for g_i satisfies*

$$\|g_i\| / \|r_i\| \leq \eta_i \quad i \in I\!N$$

with

$$\eta_i \leq \|A\| \frac{\epsilon}{\|r_i\|}$$

then

$$\|\tilde{r}_i - r_i\| \leq cond_2(A) \epsilon. \tag{4.18}$$

Furthermore the true residual r_i can be estimated by the computed residual $\tilde{r}_i$ through

$$\|r_i\| \leq \|\tilde{r}_i\| + cond_2(A) \epsilon.$$

Similar results can be seen in both the work of by Eshof and Sleijpen [56, 55] and by Szyld and Simoncini [45].

4.4.1 Application to finite dimensional case

Here we assume that the function spaces are all finite dimensional, while the results hold true in the continuous setting as well. We consider the original problem 4.1.1 in the following setting.

Problem 4.4.1.

$$\min \quad \phi(y,u), \quad (y,u) \in I\!R^m \times I\!R^n \tag{4.19}$$

$$s.t. \quad g(y,u) = 0 \tag{4.20}$$

$$where \quad \phi : I\!R^m \times I\!R^n \to I\!R, \quad g : I\!R^m \times I\!R^n \to I\!R^m \tag{4.21}$$

Theorem 4.4.1. *We have for the second derivative applied to a vector*

$$\Phi''(u)\Delta v = g_u(y,u)^T\pi + (g_{yu}(y,u)\xi)^T p + \phi_{yu}(y,u)\xi + (g_{uu}(y,u)\Delta v)^T p + \phi_{uu}(y,u)\Delta v$$

where y,u satisfy $g(y,u)=0$ and $p \in I\!R^m$ solves the adjoint equation (4.7). Furthermore ξ solves the sensitivity equation of first order

$$g_y(y,u)\xi = -g_u(y,u)\Delta v \tag{4.22}$$

and π the second order adjoint equation

$$g_y(y,u)^T\pi = -(g_{yy}(y,u)\xi)^T p - \phi_{yy}(y,u)\xi - (g_{uy}(y,u)\Delta v)^T p - \phi_{uy}(y,u)\Delta v. \tag{4.23}$$

Note that Theorem 4.4.1 is the finite dimensional counterpart of Theorem 4.2.3.

Theorem 4.4.2. *Let assumptions 4.2.1 and 4.2.2 hold at $(y,u) \in Y \times U = I\!R^m \times I\!R^n$ and equations* (4.22) *and* (4.23) *be solved inexactly i.e,*

$$g_y(y,u)\xi + g_u(y,u)\Delta v = e_s \tag{4.24}$$

$$g_y(y,u)^T\pi + (g_{yy}(y,u)\xi)^T p + \phi_{yy}(y,u)\xi + (g_{uy}(y,u)\Delta v)^T p + \phi_{uy}(y,u)\Delta v = e_a. \tag{4.25}$$

If we assume that the errors are proportional to the corresponding residuals

$$\|e_s\| \le \eta_u \|g_u(y,u)\Delta v\|,$$

$$\|e_a\| \le \eta_p \left\|(g_{yy}(y,u)\xi)^T p + \phi_{yy}(y,u)\xi + (g_{uy}(y,u)\Delta v)^T p + \phi_{uy}(y,u)\Delta v\right\|,$$

then there exist constants $\eta_u, d_u, c_u, \eta_p, d_p, c_p$ such that

$$\|r\| \leq (d_u \eta_u c_u + d_p \eta_p c_p)\|\Delta v\|, \tag{4.26}$$

where r is the difference between the true Hessian apply and the perturbed Hessian $r = \Phi''(u)\Delta v - \tilde{\Phi}''(u)\Delta v$.

Proof. We assume that the linear solves for ξ and π are carried out by an iterative solver and stopped early. The stopping criterion is based on the relative error of the residual with respect to the right hand side $g_u \Delta v$, hence in norm less than proportional to the size of the vector Δv. Hence the approximate solution denoted by $\tilde{\xi}$ satisfies

$$g_y(y,u)\tilde{\xi} + g_u(y,u)\Delta v = e_s,$$

and by our assumption we have the following

$$\|e_s\| \leq \eta_u \|g_u \Delta v\| \leq \eta_u \|g_u\|\|\Delta v\| = \eta_u c_u \|\Delta v\|. \tag{4.27}$$

for $\|g_u\| \leq c_u$. The error between ξ and $\tilde{\xi}$ satisfies

$$g_y(y,u)(\tilde{\xi} - \xi) = e_s,$$

and since g_y is assumed to be invertible and using (4.27) we obtain

$$\|\tilde{\xi} - \xi\| \leq \|g_y^{-1}\|\|e_s\| \leq \eta_u \|g_y^{-1}\|\|g_u\|\|\Delta v\|. \tag{4.28}$$

Furthermore for $\tilde{\xi}$ from its definition

$$\|\tilde{\xi}\| \leq \|g_y^{-1}\| \ (\|g_u\|\|\Delta v\| + \|e_s\|) \leq \|g_y^{-1}\| \ [1 + \eta_u] \ \|g_u\|\|\Delta v\|. \tag{4.29}$$

For the approximate solution $\tilde{\pi}$ we have the equation

$$g_y(y,u)^T \tilde{\pi} + (g_{yy}(y,u)\tilde{\xi})^T p + \phi_{yy}(y,u)\tilde{\xi} + (g_{uy}(y,u)\Delta v)^T p + \phi_{uy}(y,u)\Delta v = e_a.$$

By the same assumption as above for $\tilde{\xi}$, and that the error for $\tilde{\pi}$ is relative to the right hand side of the equation for π,

$$\|e_a\| \leq \eta_p \|(g_{yy}(y,u)\tilde{\xi})^T p + \phi_{yy}(y,u)\tilde{\xi} + (g_{uy}(y,u)\Delta v)^T p + \phi_{uy}(y,u)\Delta v\|.$$

By the estimates for $\tilde{\xi}$ from above we obtain

$$\begin{aligned}\|e_a\| \le &\eta_p\|(g_{yy}\tilde{\xi})^T p + \phi_{yy}\tilde{\xi} + (g_{uy}\Delta v)^T p + \phi_{uy}\Delta v\| && (4.30)\\ \le &\eta_p(\|g_{yy}\|\|p\| + \|\phi_{yy}\|)\ \|\tilde{\xi}\| + \eta_p(\|g_{uy}\|\|p\| + \|\phi_{uy}\|)\|\Delta v\| && (4.31)\\ \le &\eta_p\ [\ (\|g_{yy}\|\|p\| + \|\phi_{yy}\|)\ \|g_y^{-1}\|\ [1+\eta_x]\ \|g_u\|]\|\Delta v\| && (4.32)\\ &+\ (\|g_{uy}\|\|p\| + \|\phi_{uy}\|)\ ]\ \|\Delta v\| && (4.33)\\ \le &\eta_p c_p\|\Delta v\|, && (4.34)\end{aligned}$$

where c_p depends on η_x. The error in $\tilde{\pi}$ and π we obtain

$$g_y(y,u)^T(\tilde{\pi}-\pi) + (g_{yy}(y,u)(\tilde{\xi}-\xi))^T p + \phi_{yy}(y,u)(\tilde{\xi}-\xi) = e_a,$$

and by assumption

$$\|e_a\| \le \eta_p\|(g_{yy}(y,u)(\tilde{\xi}-\xi))^T p + \phi_{yy}(y,u)(\tilde{\xi}-\xi)\|. \qquad (4.35)$$

We use the estimates for the error in the ξ's from above to obtain

$$\|e_a\| \le \eta_p\ [\|g_{yy}(y,u)\|\|p\| + \|\phi_{yy}(y,u)\|]\ \eta_x\|g_y^{-1}\|\|g_u\|\|\Delta v\|.$$

By invertibility of g_y^T

$$\|\tilde{\pi}-\pi\| \le \|g_y^{-T}\|\|(g_{yy}(y,u)(\tilde{\xi}-\xi))^T p + \phi_{yy}(y,u)(\tilde{\xi}-\xi) - e_a\|$$

and

$$\|\tilde{\pi}-\pi\| \le \|g_y^{-T}\|\ [\ (\|g_{yy}\|\|p\| + \|\phi_{yy}\|)\ \|\tilde{\xi}-\xi\| + \|e_a\|\ ]$$

and using the previous estimate on the $\xi's$

$$\|\tilde{\pi}-\pi\| \le \|g_y^{-T}\|\ [\ (\|g_{yy}\|\|p\| + \|\phi_{yy}\|)\ \eta_x\|g_y^{-1}\|\|g_u\|\|\Delta v\| + \|e_a\|].$$

If we insert the estimate for the e_a from above, we obtain

$$\begin{aligned}\|\tilde{\pi}-\pi\| \le &\|g_y^{-T}\|\ [\ (\|g_{yy}\|\|p\| + \|\phi_{yy}\|)\ \eta_u\|g_y^{-1}\|\|g_u\|\|\Delta v\| && (4.36)\\ &+ \eta_p\ [\|g_{yy}(y,u)\|\|p\| + \|\phi_{yy}(y,u)\|]\ \eta_x\|g_y^{-1}\|\|g_u\|\|\Delta v\| && (4.37)\\ &\le (\eta_u c_1 + \eta_p c_2)\|\Delta v\|. && (4.38)\end{aligned}$$

Using the approximate quantities, the Hessian-vector product become inexact denoted by $\tilde{\Phi}''(u)\Delta v$ given by

$$\tilde{\Phi}''(u)\Delta v = g_u(y,u)^T\tilde{\pi}+(g_{yu}(y,u)\tilde{\xi})^Tp+\phi_{yu}(y,u)\tilde{\xi}+(g_{uu}(y,u)\Delta v)^Tp+\phi_{uu}(y,u)\Delta v.$$

In order to apply the theory from inexact methods we formulate the exact apply as an inexact

$$\Phi''(u)\Delta v = \tilde{\Phi}''(u)\Delta v + r$$

with perturbation r described by

$$r = g_u(y,u)^T(\pi - \tilde{\pi}) + (g_{yu}(y,u)(\xi - \tilde{\xi}))^Tp + \phi_{yu}(y,u)(\xi - \tilde{\xi}).$$

We would like to estimate $\|r\|$, by employ our estimates (4.38) and (4.35) we can derive

$$\begin{aligned} \|r\| &\leq \|g_u(y,u)^T\|\|\pi - \tilde{\pi}\| + (\|g_{yu}(y,u)\|\|p\| + \|\phi_{yu}(y,u)\|)\ \|\xi - \tilde{\xi}\| \\ &\leq \|g_u(y,u)^T\|\|g_y^{-T}\|\ [\ (\|g_{yy}\|\|p\| + \|\phi_{yy}\|)\ \|g_y^{-1}\|\|e_s\| + \|e_a\|] \\ &\quad +(\|g_{yu}(y,u)\|\|p\| + \|\phi_{yu}(y,u)\|)\ \|g_y^{-1}\|\|e_s\| \\ &\leq d_u\|e_s\| + d_p\|e_a\| \\ &\leq (d_u\eta_u c_u + d_p\eta_p c_p)\|\Delta v\|. \end{aligned}$$

□

Remark 3. *We observe that*

$$(d_u\eta_u c_u + d_p\eta_p c_p)\|\Delta v\| \leq \epsilon$$

holds for some fixed constant ϵ, *if*

$$\eta_u \leq \epsilon/(2d_u c_u\|\Delta v\|), \quad \eta_p \leq \epsilon/(2d_p c_p\|\Delta v\|). \tag{4.39}$$

In practice ϵ is a user supplied constant denoting an upper bound of the error between the true Hessian apply and the perturbation Hessian. To estimate these constants is computationally challenging as these involve computation of matrix norms that at first glance seem expensive to evaluate. However, we note that if one uses the $\|\cdot\|_2$ norms, only the largest eigenvalue will need to be computed and this can be only a rough estimate (as we are interested in the asymptotic behavior of the ratios in (4.39)).

Corollary 3. *Let assumptions 4.2.1 and 4.2.2 hold at* $(y, u) \in Y \times U = I\!R^m \times I\!R^n$ *and equations* (4.22) *and* (4.23) *be solved inexactly as in Theorem 4.4.2. Then we conclude that there are constants* $c_s, c_a > 0$ *such that*

$$\|\tilde{\xi} - \xi\| \leq c_s \|e_s\| \quad \textit{and} \quad \|\tilde{\pi} - \pi\| \leq c_a(\|e_a\| + \|e_s\|). \tag{4.40}$$

Proof. From equations 4.24 we can conclude that

$$g_y(y, u)(\tilde{\xi} - \xi) = e_s$$

and

$$g_y(y, u)^*(\tilde{\pi} - \pi) = e_a - (g_{yy}(y, u)(\tilde{\xi} - \xi))^* p - \phi_{yy}(y, u)(\tilde{\xi} - \xi).$$

By applying assumption 4.2.2 and some inequality work we can derive the result with $c_s = \|g_y(y, u)^{-1}\|$ and $c_a = \max\{c_s, c_s\|(g_{yy}(y, u)\|\|p\| + \|\phi_{yy}(y, u)\|\}$. □

4.5 Inexact Krylov Solvers

There is a vast literature on error estimates for Krylov methods [45, 51, 50, 44, 59]. An important question is “How much error is allowed at each Krylov iteration?”. The stunning result is that towards the end of the iteration, the error does not need to tend to zero, but can stay at a constant prescribed level or in other words the relative error compared to the residual could increase as the iteration progresses. While counter intuitive, such results can be found in the works by Szyld and Simoncini [45] and by Eshof and Sleijpen [55]. Furthermore, the work by Notay [36] and early observation made by Golub and Overton [33, 34, 35] that Krylov methods such as Conjugate Gradient may maintain convergence rate even at loose accuracy. In the book by Kelley [52, 6.2], these issues are also addressed in the context of inexact Newton methods using finite difference approximation. With regard to second order adjoints, the works of Hicken [40] touches on aspects of the convergence.

In the general case of Theorem 4.2.3 the solution y of the equality constraint and p of the adjoint equation have to be computed to high accuracy in order to obtain precise gradient information. But the question arises, if this is also needed in the case of a Hessian-vector product within the inner iteration of a Newton solve like CG or GMRES. Thus, in the next Theorem we examine the difference between the exact Hessian apply and the one obtained from solving the two equations: (second order) sensitivity equation (4.9) and (second order) adjoint equation (4.10).

If we use the inexact solves $\tilde{\xi}$ and $\tilde{\pi}$ in the Hessian vector product according to Theorem 4.2.3, we do not obtain $\Phi''(u)\Delta v$ but rather an approximate quantity which we call $\tilde{H}\Delta v$ defined by

$$\tilde{H}\Delta v = g_u(y,u)^*\tilde{\pi} + (g_{yu}(y,u)\tilde{\xi})^*p + \phi_{yu}(y,u)\tilde{\xi} + (g_{uu}(y,u)\Delta v)^*p + \phi_{uu}(y,u)\Delta v.$$

Theorem 4.5.1. *Let assumptions 4.2.1 and 4.2.2 hold at* $(y,u) \in Y \times U$ *and* $\tilde{H}\Delta v$ *be the Hessian-vector product obtained by solving equations* (4.22) *and* (4.23) *be solved inexactly. Then*

$$\|(\tilde{H} - \Phi''(u))\Delta v\| \leq c_H(\|e_a\| + \|e_s\|). \tag{4.41}$$

Proof. Consider

$$(\tilde{H} - \Phi''(u))\Delta v = g_u(y,u)^*(\tilde{\pi} - \pi) + (g_{yu}(y,u)(\tilde{\xi} - \xi))^*p + \phi_{yu}(y,u)(\tilde{\xi} - \xi)$$

and using the estimates (4.40) and the standard inequality work, we conclude our result, i.e., for some positive constant c_H

$$\|r_i\| \leq c_H(\|e_a\| + \|e_s\|), \tag{4.42}$$

where $r = g_u(y,u)^*(\tilde{\pi} - \pi) + (g_{yu}(y,u)(\tilde{\xi} - \xi))^*p + \phi_{yu}(y,u)(\tilde{\xi} - \xi)$. □

This gives us an estimate for the error in the Hessian-vector product in a rigorous manner for a general Banach space. For the rest of the chapter we will work with the finite dimensional version to be consistent with literature on Krylov methods. Next, the question arises, how does the error, which influences each iteration in a Krylov iteration affect the solution of each Newton step:

$$\Phi''(u)s = -\nabla\Phi(u). \tag{4.43}$$

If the Krylov method solving the inexact system (4.43) terminates with termination criterion ϵ, then we have a Newton system that satisfies

$$\|\tilde{H}s + \nabla\Phi(u)\| \leq \epsilon$$

. We provide analysis in terms of the Richardson iteration as a Krylov solver.

Theorem 4.5.2. *Let r_i and $\tilde{r}_i$ be the residuals of the Richardson iteration (see Equations* (4.12) *and* (4.13)*) solving to the Newton system $\Phi''(u)s_i = -\nabla\Phi(u)$. Furthermore, assume the perturbation g_i from* (4.13) *is bounded*

$$\|g_i\| \leq \eta_i \|\Phi''(u)\| \|s_i\|$$

then the difference of the residuals follows

$$\|r_i - \tilde{r}_i\| \leq \sum_{j=1}^{i} \|I - \alpha\Phi''(u)\|^{j-1} \alpha\eta_{i-j} \|\Phi''(u)\| \|\Phi''(u)s_{i-j} + \nabla\Phi(u)\|, \tag{4.44}$$

where α and η are some constants.

Proof. This can be seen as a direct consequence of Lemma 3, where $A = \Phi''(u)$. Consider the residual

$$\|z_i\| = \|\tilde{r}_i - r_i\| = \| \sum_{j=1}^{i} (I - \alpha A)^{j-1} \alpha g_{i-j} \| \tag{4.45}$$

$$\leq \sum_{j=1}^{i} \|I - \alpha A\|^{j-1} \alpha \|g_{i-j}\|, \tag{4.46}$$

and by assumption the perturbation g_i for the matrix vector product Ar_i is bounded

$$\|g_i\| \leq \eta_i \|A\| \|r_i\|$$

then we can conclude

$$\|r_i - \tilde{r}_i\| \leq \sum_{j=1}^{i} \|I - \alpha A\|^{j-1} \alpha\eta_{i-j} \|A\| \|r_{i-j}\| \tag{4.47}$$

$$= \sum_{j=1}^{i} \|I - \alpha\Phi''(u)\|^{j-1} \alpha\eta_{i-j} \|\Phi''(u)\| \|\Phi''(u)s_{i-j} + \nabla\Phi(u)\|, \tag{4.48}$$

where α was defined in equation (4.11). □

This estimate gives us a good understanding of how the error can propagate in the system. While we were satisfied to do this for Richardson, there is more general theory available for Full Orthogonalization Method (FOM) and the Generalized Minimal Residual Method (GMRES), we point the reader to the works of Szyld and Simoncini [45, 51, 44].

4.6 Model problem setup

To illustrate the effect of using an inexact Hessian-apply using the dynamic tolerance proposed in Section 4.4.1 (see Theorem 4.4.2 and equation (4.39)), we consider the inference of the log-coefficient field in an elliptic partial differential equation. *The forward model* can be mathematically expressed as

$$\begin{aligned} -\nabla \cdot (e^u \nabla y) &= f \quad \text{in } \mathcal{D}, \\ y &= g \quad \text{on } \Gamma_D, \\ e^u \nabla y \cdot \boldsymbol{n} &= h \quad \text{on } \Gamma_N, \end{aligned} \tag{4.49}$$

where $\mathcal{D} \subset \mathbb{R}^d$ $(d = 2, 3)$ is an open bounded domain with sufficiently smooth boundary $\Gamma = \Gamma_D \cup \Gamma_N$, $\Gamma_D \cap \Gamma_N = \emptyset$. Here, y is the state variable, $f \in L^2(\mathcal{D})$ is the source term, and u is an uncertain parameter field in $\mathcal{E} = \mathsf{dom}(\mathcal{A})$, where $\mathcal{A}$ is a Laplacian-like operator, as defined in [76, 75]. To state the weak form of (4.49), we define the space,

$$\mathcal{V}_g = \{v \in H^1(\mathcal{D}) : v\big|_{\Gamma_D} = g\}, \quad \mathcal{V}_0 = \{v \in H^1(\mathcal{D}) : v\big|_{\Gamma_D} = 0\},$$

where $H^1(\mathcal{D})$ is the Sobolev space of functions in $L^2(\mathcal{D})$ with square integrable derivatives. Then, the weak form of (4.49) is as follows: Find $y \in \mathcal{V}_g$ such that

$$\langle e^u \nabla y, \nabla p \rangle = \langle f, p \rangle + \langle h, p \rangle_{\Gamma_N}, \quad \forall p \in \mathcal{V}_0. \tag{4.50}$$

Here $\langle \cdot, \cdot \rangle$ denotes the standard inner products in $L^2(\mathcal{D})$.

For the numerical experiments we choose $\mathcal{D} := [0,1] \times [0,1]$ and the boundaries $\Gamma_N := \{0,1\} \times (0,1)$ and $\Gamma_D := (0,1) \times \{0,1\}$. We also choose no source term, (i.e., $f = 0$) and no normal flux on $\Gamma_N := \{0,1\} \times (0,1)$ (i.e., the homogeneous Neumann condition $e^u \nabla y \cdot \boldsymbol{n} = 0$ on Γ_N) are imposed. Dirichlet conditions are prescribed on the top and bottom boundaries, in particular $y = 1$ on $(0,1) \times \{1\}$ and $y = 0$ on $(0,1) \times \{0\}$. In Figure 4.1, we illustrate the "truth" parameter field used in our numerical tests, and the corresponding state solution.

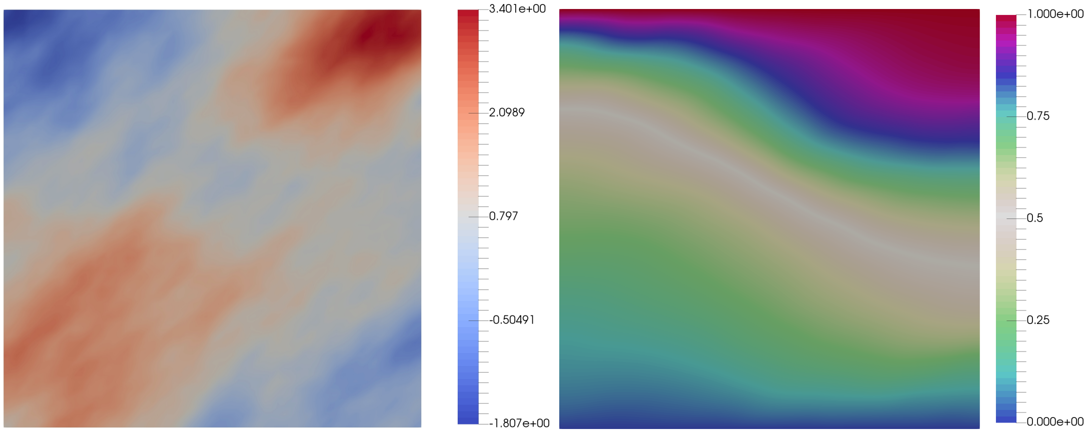


Figure 4.1: Log parameter field u_{true} (a) and state y obtained by solving the forward (state) equation (4.49) with u_{true} . (b).

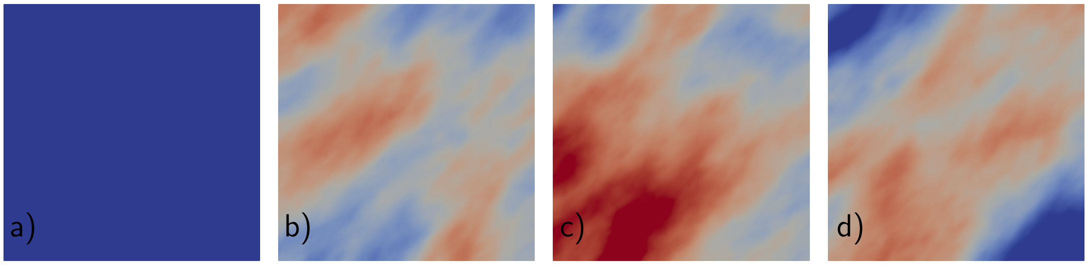


Figure 4.2: Prior mean (a), and samples drawn from the prior distribution (b)–(d).

4.7 Computational results

In this section, we present numerical results for the model problem described in section 4.6. The numerical results presented in this paper are obtained using hIPPYlib (an inverse problem Python library [127, 126]). hIPPYlib implements state-of-the-art scalable adjoint-based algorithms for PDE-based deterministic and Bayesian inverse problems. It builds on FEniCS [128, 129] for the discretization of the PDEs and on PETSc [131] for scalable and efficient linear algebra operations and solvers needed for the solution of the PDEs.

4.7.1 Comparison of the performance of the Newton inverse solver with exact and inexact Hessian-applies

Here, we compute the MAP point discussed in Section 2.2. As can be seen in Table 4.1 (a), the tolerance we derived in Section 4.4.1 is relaxed as we converge to the inverse problem solution. This table also shows that the Newton inverse solver with exact and inexact Hessian-applies perform similarly. Namely, both reduce the optimization objective function to the same value and require the same number of outer Newton iterations to converge (to the given tolerance). We would like to note that while in this table the Newton inverse solver with inexact Hessian-applies requires the same number of outer iterations to converge, we have also observed tests where the number of iterations is slightly larger. This is expected due to the approximate Hessian being used.

Figure 4.3 (left) confirms the results presented in Table 4.1 visually. In particular, from this figure we see that overall the inexact Hessian-apply leads to fewer Krylov iterations than the exact Hessian-apply. The difference is more significant in the middle of the iterations due to the lower dynamic tolerances. We note a high number of inner iterations for the last Newton step even though the dynamic tolerance is much smaller than the fixed tolerance. We attribute this to the fact that the number of Newton-CG iterations was higher for the inexact Hessian-apply case. On the right plot we show the total number of Krylov iterations for each Newton iteration. As can be seen, the number of inner Krylov iterations for the inexact Hessian-apply case is smaller than for the exact Hessian-apply case, and the difference increases as we approach the inverse problem solution.

It is also important to consider the computational cost of the inexact versus exact Hessian-based methods. To compare the two approaches we record the total number of Krylov iterations for the two PDE solves required by the Hessian-apply. As can be seen Table 4.1, the Newton inverse solver with inexact Hessian-applies outperforms the one with exact Hessian-applies. We remind the reader though that to compute the dynamic tolerance, one needs to compute matrix norms. However, as discussed in Section 4.4.1 getting an approximate value for the tolerance has a cost of computing the asymptotic behavior of the leading eigenvalue of ϕ_{uy}, ϕ_{uu},

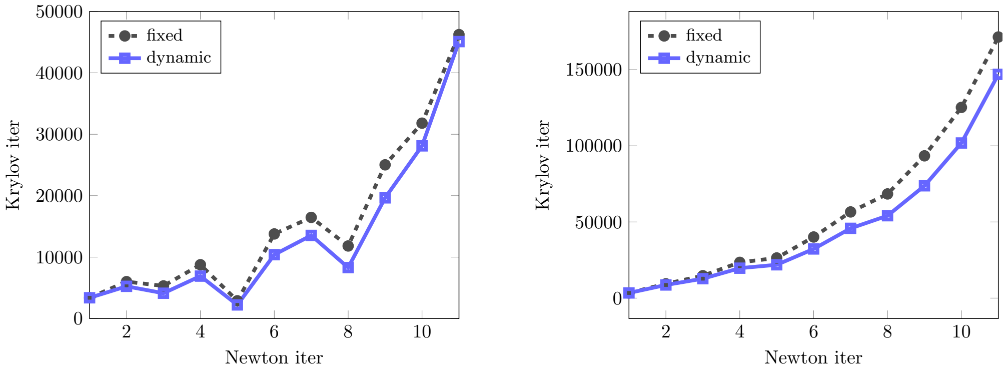


Figure 4.3: The performance of the inverse solver for the dynamic (black) and fixed (blue) tolerances. The left and right figures show the number of Krylov iterations per outer Newton iteration and the total Krylov iterations, respectively.

g_y, g_u [31, 32, 61, 60, 38]. In Table 4.3 we summarize the computational cost for estimating the largest eigenvalues (i.e., l_2 matrix norms) necessary for computing the dynamic tolerance using randomized SVD and the power method. To show the influence of the relative error on the final CG iterations, we compute these estimates using 10% (left) and 5% (right) relative errors. The relative error is defined as $\frac{|\|A_{ref}\|_2 - \|A\|_2|}{\|A_{ref}\|_2}$, where $\|A_{ref}\|_2$ is the accurate l_2 norm of the reference matrix, and $\|A\|_2$ is estimated. These results show that the dynamic tolerance is not very sensitive to the accuracy of the matrix norms. We also note that the power method was slightly more efficient (for this problem) than the randomized SVD. Finally, we would like to point out that in practice we found that the dynamic tolerance can be reused for a number of iterations, which can save some computational effort.

Next we compare the spectrum of the data misfit Hessian evaluated at the MAP point. Figure 4.4 shows a logarithmic plot of the eigenvalues of the generalized symmetric eigenproblem involving the exact and inexact Hessian-applies [126]. This plot shows that the spectrums coincide for the two approaces and that, as expected, they decay rapidly. As shown in [126], an accurate low-rank based approximation of the inverse (exact and inexact) Hessian can be obtained by neglecting eigenvalues that are small compared to 1.

In Figure 4.5 we show the MAP point (a) and samples from the Laplace approximation (20) of the posterior probability density function (b)-(d). These samples

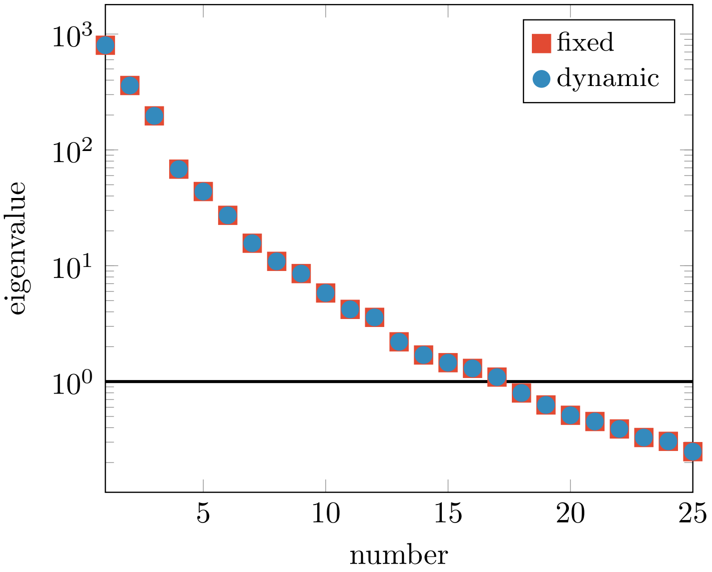


Figure 4.4: Log-linear plot of first 25 eigenvalues of the prior-preconditioned Hessian of the negative log-likelihood for the dynamic (blue) and fixed (red) tolerances. The low-rank based approximation captures the dominant, data-informed portion of the spectrum. The eigenvalues are truncated at $\lambda = 1$.

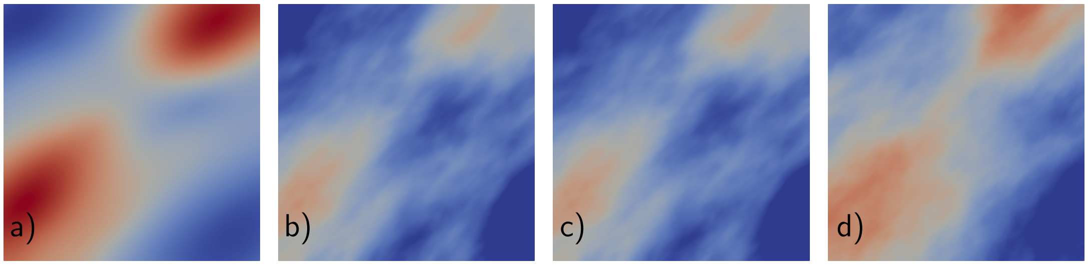


Figure 4.5: The MAP point (a) and samples drawn from the Laplace approximation of posterior distribution (b)–(d).

were obtained using the inexact Hessian-apply. The variance reduction between the posterior samples and prior samples shown in Figure 4.2 reflects the information gained from the data in solving the inverse problem. We note that the MAP point (a) resembles the truth everywhere in the domain due to having observations everywhere.

4.8 Conclusion

In our work we relax the second adjoint and sensitivity equations required for the Hessian-vector apply, we provide general and rigorous derivation on the tolerances used in for the inexact solves. In addition we show the tolerances controlling the inexactness of the adjoint and sensitivity equations are allowed to increase (as the number of Newton iterations in the inverse problem increases), while maintain our convergence properties (see Theorem 4.4.2). Our work provides a basis for practical dynamic strategies for the relaxation of the Hessian-vector products. We employ our dynamical tolerances as a stopping criteria for inner Krylov solver for both the deterministic and Bayesian framework. We provide error bounds on the difference of the inexact Hessian-vector product and the true Hessian-vector product (see Theorem 4.5.1). We demonstrate the computational value of our analytical work in an illustrative example, a statistical inverse problem.

Acknowledgments

This chapter is joint work with Prof. Ekkehard Sachs. We would like to acknowledge his major contributions to the theoretical part of this project. The financial support for this project was from the U.S. National Science Foundation, Division of Mathematical Sciences, Computational Mathematics, grant CAREER-1654311.

Nit	Cost	$\|\boldsymbol{g}\|$	CGit	SensIt	AdjIt	PDE Tol
(a) inexact Hessian-apply						
1	1.10E+03	1.59E+04	2	1580	1800	1.00E-09
2	2.90E+02	5.97E+03	2	2756	2490	9.06E-14
3	1.72E+02	2.57E+03	2	2164	1962	5.37E-13
4	6.00E+01	1.63E+03	4	3548	3362	1.64E-12
5	5.50E+01	1.05E+03	1	1053	1166	2.96E-12
6	4.20E+01	5.46E+02	6	5139	5258	8.45E-12
7	3.91E+01	3.17E+02	8	6679	6854	3.34E-11
8	3.85E+01	2.27E+02	5	4092	4205	4.72E-11
9	3.78E+01	1.56E+02	13	8712	10936	1.56E-10
10	3.78E+01	2.97E+01	20	11930	16183	4.15E-10
11	3.78E+01	1.28E+00	41	15742	29346	2.13E-09
Total iterations			104	146957		
(b) exact Hessian-apply						
1	1.10E+03	1.59E+04	2	1580	1800	1.00E-09
2	2.90E+02	5.97E+03	2	3327	2681	1.00E-14
3	1.72E+02	2.57E+03	2	3013	2302	1.00E-14
4	6.00E+01	1.63E+03	4	4640	4125	1.00E-14
5	5.50E+01	1.05E+03	1	1490	1401	1.00E-14
6	4.20E+01	5.46E+02	6	6917	6857	1.00E-14
7	3.91E+01	3.17E+02	7	8186	8275	1.00E-14
8	3.85E+01	2.24E+02	5	5619	6197	1.00E-14
9	3.78E+01	1.56E+02	11	12095	12928	1.00E-14
10	3.78E+01	2.98E+01	15	14263	17545	1.00E-14
11	3.78E+01	1.35E+00	23	18959	27270	1.00E-14
Total iterations			78	171470		

Table 4.1: The number of total Krylov iterations for dynamic (a) and fixed (b) tolerances. The **Nit** lists the number of outer Newton iterations needed to reduce the norm of the gradient $\|g\|$ five orders of magnitude. The **Cost** lists the value of the optimization objective (i.e., the negative log posterior); the **CGit** lists the number of CG iterations needed to solve the Newton system; the **SensIt** and **AdjIt** list the number of Krylov iterations needed to solve the sensitivity and second order adjoint problems to a tolerance given by **PDE Tol**.

Test 1					**Test 3** (Hicken*)				
Nit	**CG**	**SensIt**	**AdjIt**	**Tol**	**Nit**	**CG**	**SensIt**	**AdjIt**	**Tol**
1	1	19	18	10^{-6}	1	1	15	14	10^{-5}
2	1	19	17	10^{-6}	2	1	19	17	10^{-6}
3	1	20	12	10^{-6}	3	1	24	17	10^{-8}
4	21	366	157	10^{-6}	4	11	250	141	10^{-8}
5	1	19	11	10^{-6}	5	4	100	65	10^{-10}
6	147	2001	563	10^{-6}	6	16	350	205	10^{-10}
7	2	23	8	10^{-6}	7	18	422	277	10^{-13}
Total	174	3253			**Total**	52	1927		

Test 2					**Test 4** (Our approach)				
1	1	47	46	10^{-15}	1	1	19	18	10^{-7}
2	1	47	45	10^{-15}	2	1	34	32	10^{-12}
3	1	47	39	10^{-15}	3	1	31	26	10^{-11}
4	10	452	327	10^{-15}	4	10	291	215	10^{-11}
5	4	171	122	10^{-15}	5	5	126	82	10^{-10}
6	14	563	416	10^{-15}	6	16	344	190	10^{-10}
7	20	621	470	10^{-15}	7	6	49	9	10^{-8}
Total	51	3413			**Total**	51	1426		

Table 4.2: Solving the deterministic inverse problem 3.30 with fixed and adaptive tolerances; **Test 1** and **Test 2** used fixed tolerances, **Test 3** we used an adaptive tolerance suggested by Hicken and Alonso in [40], **Test 4** uses the dynamic tolerances from equation (4.39) where we fixed $\epsilon = 10^{-12}$. The **Nit** column lists the number of outer Newton iterations needed to reduce the norm of the gradient $\|g\|$ five orders of magnitude; the **CGit** column lists the number of CG iterations needed to solve the Newton system; the **SensIt** and **AdjIt** columns list the number of Krylov iterations needed to solve the sensitivity and second order adjoint problems to a tolerance given by **PDE Tol**, listed in the last column; **Total** is the total number of Krylov iterations for the **CG**, **SensIt** and **AdjIt**.

RelError<0.1	**Ref**	**RSVD**	**PM**
$\|g_y\|_2$	N/A	88	51
$\|g_u\|_2$	N/A	22	6
$\|\phi_{uy}\|_2$	N/A	28	7
$\|\phi_{uu}\|_2$	N/A	66	6
Total	N/A	204	66
CG	1461	1461	1463

RelError<0.05	**Ref**	**RSVD**	**PM**
$\|g_y\|_2$	N/A	399	180
$\|g_u\|_2$	N/A	28	28
$\|\phi_{uy}\|_2$	N/A	30	20
$\|\phi_{uu}\|_2$	N/A	74	9
Total	N/A	527	237
CG	1461	1461	1461

Table 4.3: A numerical study of the accuracy required for estimating the matrix norms for the dynamic tolerances applied in an inexact Newton-CG solver. In the columns we show the reference norm (**Ref**) computed within machine precision using a direct method, the relative error (**RelError**), and the number of mat-vecs for randomized SVD (**RSVD**) and the number of power method iterations (**PM**) necessary to reach the accuracy prescribed by the relative error.

Chapter 5

Variance reduction for the Bayesian approximation error (BAE) with application to the Stokes ice sheet model under uncertain thermal distribution

5.1 Introduction

Inverse problems governed by physics-based models (*e.g.*, expressed via PDEs) typically contain multiple parameters which are uncertain. A common approach is to formulate and solve the inverse problem to infer the unknown/uncertain parameters simultaneously. However, this approach will result in a highly ill-posed and potentially computationally intractable problem. Often times, these additional uncertainties are not taken into account, which leads to infeasible estimates, as shown in [23, 21]. To overcome these difficulties, one approach is to premarginalize over the *not important* or so-called secondary or auxiliary parameters, and then invert for the *important* or so-called primary parameters. This can be done via the Bayesian approximation error (BAE) approach [71, 23, 21].

Bayesian inversion combined with BAE will lead to a modified likelihood term that will bring in the additional uncertainties via a modified (non-diagonal) correlation matrix. It is common practice to approximate the model error with a Gaussian distribution whose mean and covariance can be estimated with Monte Carlo sampling. The goal of this work is to reduce the computational cost, i.e., reduce the number of samples needed to accurately estimate the statistics (i.e., the mean and variance) of the model error. The key idea is to use a control variate approach [136, 137] using a Taylor linear approximation of the model discrepancy (i.e., the difference of the accurate and approximate models). This approach will allow us to compute parts of the mean and covariance via analytical formulas, as in [138, 139] and the remainder via Monte Carlo sampling. We illustrate our approach with an ice sheet inverse problem. The primary parameter is the so-called basal sliding parameter field (a parameter field that has been the focus of ice sheet inversion in the last decade [13, 11, 12, 10, 8, 7, 6, 5, 4, 114, 72, 9]). The secondary parameters are parameters that go into the computation of the temperature field that affects the viscosity of the ice.

The main contributions of this chapter of the thesis is threefold. First, we show that simply ignoring the temperature distribution within the ice sheet flow model by setting it to a (possibly well justified) nominal value can lead to significantly overly confident and biased estimates for the basal sliding coefficient field if the additional uncertainty is not accounted for. Secondly, we show that the BAE approach can be used to take into account these additional uncertainties at essentially no additional computational costs at the *online stage*, as all computations are carried out prior to the acquisition of data. Thirdly, we show that employing a linear Taylor expansion as a control variate can reduce the *offline stage* costs associated with using the BAE approach.

This chapter is organized as follows. In Section 5.2 we discuss the conventional Bayesian inverse problem formulation and provide necessary background on the Bayesian approximation error approach. In Section 5.3, we describe the forward ice sheet flow problem that is used for the inference of the basal sliding coefficient field under uncertain rheology, and the mathematical model guiding the thermal distribution of the ice. In Section 5.4 we show how to estimate the mean and variance for the model error, and in Section 5.5 we discuss the control variate approach to reduce the variance of the model error. Finally, in Section 5.6 we show numerical results. Section 5.7 provides concluding remarks.

5.2 Background

5.2.1 Conventional Bayesian Inverse Ice Sheet Problem

Here we summarize the Bayesian inverse ice sheet problem, for a more in-depth discussion see [72, 114, 113, 116]. The goal of the inverse ice sheet problem is to estimate the basal sliding parameter based on noisy surface velocity measurements. It is well understood, however, that the surface velocity can also be significantly influenced by the temperature dependent rheology of the ice (among other things). As such, we write the accurate representation of the relationship between the surface observations, $\boldsymbol{d}_{\mathrm{o}}$, and parameters, i.e., the noise model, as

$$\boldsymbol{d}_{\mathrm{o}} = \mathcal{F}(\beta, z) + \boldsymbol{\epsilon}, \tag{5.1}$$

where $\boldsymbol{\epsilon}$ denotes measurement errors in the data, β the basal sliding parameter field, and z any other unknown parameters, i.e., the geothermal heat flux and thermal conductivity of the ice.

We pose the inverse problem in the Bayesian framework [76, 126, 67] as it allows for systematic incorporation of uncertainty, including model uncertainties [69, 70]. As discussed in Chapter 2, within the Bayesian framework all unknowns are treated as random variables, and are assigned prior probability densities which encode any prior beliefs on the parameters. Here we model these prior densities as Gaussians, i.e., $\beta \sim \mathcal{N}(\beta_0, \mathcal{C}_\beta)$ and $z \sim \mathcal{N}(z_0, \mathcal{C}_z)$, where β_0 and z_0 denote the means, and $\mathcal{C}_\beta$ and $\mathcal{C}_z$ the respective covariance operators. The covariance operators are define using Laplacian-like PDE operators, namely $\mathcal{C}_\beta = \mathcal{A}_\beta^{-1}$ and $\mathcal{C}_z = \mathcal{A}_z^{-1}$, with

$$\mathcal{A}_\beta = -\nabla \cdot (\Theta_\beta)\nabla + \delta_\beta \mathcal{I}, \quad \mathcal{A}_z = -\nabla \cdot (\Theta_z)\nabla + \delta_z \mathcal{I}, \tag{5.2}$$

where $(\Theta_\beta,\ \Theta_z)$ and $(\delta_\beta,\ \delta_z)$ control the correlation lengths and the pointwise variance of the prior operator, respectively [76, 126].

Discretization: We discretize each of the parameters using continuous linear Lagrange basis functions, leading to the approximations $\beta_h(\boldsymbol{s}) = \sum_{i=1}^{p} \beta_i \phi_i(\boldsymbol{s})$ and $z_h(\boldsymbol{x}) = \sum_{k=1}^{q} z_k \psi_k(\boldsymbol{x})$, where the basis functions ϕ and ψ may or may not coincide. The unknowns are then $\boldsymbol{\beta} = [\beta_1, \beta_2, \dots \beta_p]$ and $\boldsymbol{z} = [z_1, z_2, \dots z_q]$ which are normally distributed with means $\boldsymbol{\beta}_0$ and $\boldsymbol{z}_0$ and covariance matrices

$$\left[\boldsymbol{\Gamma}_\beta^{-1}\right]_{ij} = \int_{\mathcal{D}_\beta} \phi_i(\boldsymbol{s}) \mathcal{A}_\beta^2 \phi_j(\boldsymbol{s}) d\boldsymbol{s} \quad \left[\boldsymbol{\Gamma}_z^{-1}\right]_{kl} = \int_{\mathcal{D}_z} \psi_k(\boldsymbol{x}) \mathcal{A}_z^2 \psi_l(\boldsymbol{x}) d\boldsymbol{x}, \tag{5.3}$$

for $i, j \in \{1, 2, \dots, p\}$ and $k, l \in \{1, 2, \dots, q\}$, see [126] for details. The basal sliding parameter is assumed to be (a priori) independent of the thermal parameters of the ice. Thus the discrete joint prior distribution can be written as

$$\pi_{\text{prior}}(\boldsymbol{\beta}, \boldsymbol{z}) \propto \exp\left\{-\frac{1}{2}\left\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\right\|^2_{\boldsymbol{\Gamma}_\beta^{-1}} - \frac{1}{2}\left\|\boldsymbol{z} - \boldsymbol{z}_0\right\|^2_{\boldsymbol{\Gamma}_z^{-1}}\right\}, \tag{5.4}$$

where $\|\cdot\|_{\boldsymbol{\Gamma}_\beta^{-1}}$ and $\|\cdot\|_{\boldsymbol{\Gamma}_z^{-1}}$ denote the $\boldsymbol{\Gamma}_\beta^{-1}$ and $\boldsymbol{\Gamma}_z^{-1}$ weighted l_2 norms, respectively.

The solution of the Bayesian inverse problem is the parameter posterior probability distribution, i.e., the distribution of the parameter conditioned on the observations. Bayes' theorem [66, 24] allows us to write the posterior in terms of the prior density and the likelihood,

$$\pi(\boldsymbol{\beta}, \boldsymbol{z} | \boldsymbol{d}_{\text{o}}) \propto \pi_{\text{like}}(\boldsymbol{d}_{\text{o}} | \boldsymbol{\beta}, \boldsymbol{z}) \pi_{\text{prior}}(\boldsymbol{\beta}, \boldsymbol{z}).$$

Assuming the measurement errors are independent of all parameters and are normally distributed with mean $\boldsymbol{0}$, i.e., $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Gamma}_e)$, the discrete likelihood is of the form [126]

$$\pi_{\text{like}}(\boldsymbol{d}_{\text{o}} | \boldsymbol{\beta}, \boldsymbol{z}) \propto \exp\left\{-\frac{1}{2}\left\|\boldsymbol{d}_{\text{o}} - \mathcal{F}(\boldsymbol{\beta}, \boldsymbol{z})\right\|^2_{\boldsymbol{\Gamma}_e^{-1}}\right\}. \tag{5.5}$$

Plugging this into Bayes' theorem we have

$$\pi_{\text{post}}(\boldsymbol{\beta}, \boldsymbol{z} | \boldsymbol{d}_{\text{o}}) \propto \exp\left\{-\frac{1}{2}\left(\left\|\boldsymbol{d}_{\text{o}} - \mathcal{F}(\boldsymbol{\beta}, \boldsymbol{z})\right\|^2_{\boldsymbol{\Gamma}_e^{-1}} + \left\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\right\|^2_{\boldsymbol{\Gamma}_m^{-1}} + \left\|\boldsymbol{z} - \boldsymbol{z}_0\right\|^2_{\boldsymbol{\Gamma}_z^{-1}}\right)\right\}.$$

When the parameter-to-observable map $\mathcal{F}$ is nonlinear, the posterior is not Gaussian. Full characterization of the posterior would then necessitate the use of sampling-based approaches such as Markov chain Monte Carlo (MCMC). However, for large-scale problems with computationally expensive forward problems, such as the problem at hand, sampling based approaches are computationally infeasible. As a computationally practical alternative, the maximum a posteriori (MAP) estimate, *i.e.*, the parameters which maximize the posterior density, and the approximate posterior covariance are computed. More specifically, letting $\boldsymbol{\omega} = [\boldsymbol{\beta}^T, \boldsymbol{z}^T]^T \in \mathbb{R}^{p+q}$, the approximation $\pi_{\text{post}}(\boldsymbol{\omega}|\boldsymbol{d}_{\text{o}}) \approx \mathcal{N}(\boldsymbol{\omega}_{\text{MAP}}, \boldsymbol{\Gamma}_{\text{post}})$ is made, where the (joint) MAP estimate is defined

$$\boldsymbol{\omega}_{\text{MAP}} := \arg \min_{\boldsymbol{\omega} \in \mathbb{R}^{p+q}} \pi_{\text{post}}(\boldsymbol{\theta}|\boldsymbol{d}_{\text{o}}), \tag{5.6}$$

and the approximate (joint) posterior covariance matrix is

$$\boldsymbol{\Gamma}_{\text{post}} := \left(\boldsymbol{F}^T \boldsymbol{\Gamma}_e^{-1} \boldsymbol{F} + \boldsymbol{\Gamma}_{\text{prior}}^{-1}\right)^{-1}. \tag{5.7}$$

Here $\boldsymbol{F} = [\boldsymbol{F}_\beta \, \boldsymbol{F}_z] \in \mathbb{R}^{d\times(p+q)}$ with $\boldsymbol{F}_\beta \in \mathbb{R}^{d\times p}$ and $\boldsymbol{F}_z \in \mathbb{R}^{d\times q}$ denoting the Jacobian matrices of the parameter-to-observable map with respect to $\boldsymbol{\beta}$ and $\boldsymbol{z}$, respectively, and $\boldsymbol{\Gamma}_{\text{prior}} \in \mathbb{R}^{(p+q)\times(p+q)}$ the joint prior (block diagonal) covariance matrix, *i.e.*, $\boldsymbol{\Gamma}_{\text{prior}} = \text{diag}(\boldsymbol{\Gamma}_\beta, \boldsymbol{\Gamma}_z)$. Here we are only concerned with the estimation of the basal sliding parameter, i.e., the uncertain (secondary or auxiliary) thermal parameters of the ice are of little or no interest. That is, we only wish to find the marginal posterior $\pi(\boldsymbol{\beta}|\boldsymbol{d}_{\text{o}})$. We follow the Bayesian approximation error approach [71, 23, 21], which provides a means to approximately *premarginalize* over the auxiliary parameters, *i.e.*, marginalize over the auxiliary parameters prior to collecting data. In what follows, we give a brief overview of this approach.

5.2.2 The Bayesian Approximation Error Approach

The key idea of the Bayesian approximation error approach is to write the noise model given in Equation 5.1 as follows

$$\boldsymbol{d}_{\text{o}} = \mathcal{F}(\beta, z) + \boldsymbol{\epsilon} = \mathcal{F}(\beta, z) - \mathcal{G}(\beta) + \mathcal{G}(\beta) + \boldsymbol{\epsilon} = \mathcal{G}(\beta) + \boldsymbol{\eta}(\beta, z), \tag{5.8}$$

where $\mathcal{F}(\beta, z)$ is the accurate (sometimes referred to as high-fidelity) model, $\mathcal{G}(\beta) = \mathcal{F}(\beta, z_0)$ is the accurate model (also referred to as the low fidelity model) evaluated at a nominal value for the additional uncertainty), $\boldsymbol{\eta}(\beta, z)$ is the total error subsuming the model error (or discrepancy) $\boldsymbol{r}(\beta, z) = \mathcal{F}(\beta, z) - \mathcal{G}(\beta)$ and the noise in measurements $\boldsymbol{\epsilon}$. Both the primary and auxiliary parameters are random variables, and thus so is $\boldsymbol{r}$. The next step in the BAE approach is to approximate $\boldsymbol{r}|(\beta, z)$ as Gaussian, *i.e.*, $\boldsymbol{r}|(\beta, z) \sim \mathcal{N}(\boldsymbol{\varepsilon}_*, \boldsymbol{\Gamma}_r)$. The mean, $\boldsymbol{\varepsilon}_*$, and covariance, $\boldsymbol{\Gamma}_r$, can in general not be computed analytically, and thus we build these by sampling. Note that all the sampling however can be carried out before we collect the data, and is thus often referred to as being carried out *offline*.

It is common practice to approximate $\boldsymbol{r}$ as uncorrelated with the parameter of interest, β, meaning that $\boldsymbol{r}|(\beta, z) = \boldsymbol{r}$. As a result, the noise model (5.1) can be rewritten as

$$\boldsymbol{d}_{\mathrm{o}} = \mathcal{G}(\beta) + \boldsymbol{\eta}, \tag{5.9}$$

with $\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{\eta}_*, \boldsymbol{\Gamma}_\eta) = \mathcal{N}(\boldsymbol{\varepsilon}_*, \boldsymbol{\Gamma}_e + \boldsymbol{\Gamma}_r)$. We note that using BAE leads to an updated likelihood. The MAP estimate then becomes

$$\beta_{\mathrm{MAP}} := \arg\min_{\beta} \; \|\boldsymbol{d}_{\mathrm{o}} - \mathcal{G}(\beta) - \boldsymbol{\varepsilon}_*\|^2_{\boldsymbol{\Gamma}_\eta^{-1}} + \|\beta - \beta_*\|^2_{\boldsymbol{\Gamma}_\beta^{-1}}, \tag{5.10}$$

and the posterior covariance changes to

$$\boldsymbol{\Gamma}_{\mathrm{post}} = (\mathcal{F}^T \boldsymbol{\Gamma}_\eta^{-1} \mathcal{F} + \boldsymbol{\Gamma}_\beta^{-1})^{-1}. \tag{5.11}$$

5.3 Forward Ice Sheet Flow Model

In this section, we describe the accurate (*i.e.*, high fidelity) and approximate (*i.e.*, low fidelity) forward ice sheet models described by the nonlinear Stokes equations and the model for the temperature distribution that influences the viscosity term in the accurate Stokes problem. For the approximate forward model, the temperature is fixed, *i.e.*, this model is unaware of the changes in the temperature.

The accurate (high fidelity) ice sheet Stokes model. We model the flow of ice as a nonisothermal, viscous, shear-thinning, incompressible fluid via the balance of mass and linear momentum [Hutter, 1983, Marshall, 2005, Paterson, 1994]

$$-\nabla \cdot \boldsymbol{\sigma}_{\boldsymbol{u}} = \rho \boldsymbol{g} \qquad \text{in } \Omega, \tag{5.12a}$$

$$\nabla \cdot \boldsymbol{u} = 0 \qquad \text{in } \Omega, \tag{5.12b}$$

where $\boldsymbol{u}$ denotes the velocity field, $\boldsymbol{\sigma}_{\boldsymbol{u}}$ the stress tensor, ρ the density of the ice, and $\boldsymbol{g}$ gravity. The stress, $\boldsymbol{\sigma}_{\boldsymbol{u}}$, can be decomposed as $\boldsymbol{\sigma}_{\boldsymbol{u}} = \boldsymbol{\tau}_{\boldsymbol{u}} - \boldsymbol{I}p$, where $\boldsymbol{\tau}_{\boldsymbol{u}}$ is the deviatoric stress tensor, p the pressure, and $\boldsymbol{I}$ the identity tensor. We employ Glen's flow law [119, 27], which relates the stress and strain rate tensors by

$$\boldsymbol{\tau}_{\boldsymbol{u}} = 2\eta(\boldsymbol{u},\theta)\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}} \quad \text{with} \quad \eta(\boldsymbol{u},\theta) = \frac{1}{2}A(\theta)^{-\frac{1}{n}}\dot{\boldsymbol{\varepsilon}}_{\mathrm{II}}^{\frac{1-n}{2n}}, \tag{5.12c}$$

where n is the Glen's flow law exponent parameter (here taken 3 as in most of ice sheet models), η is the effective viscosity, the Arrhenius $A(\theta) = A_0 \exp\left(-\frac{Q}{R\theta}\right)$ is the temperature dependent flow rate factor with Q the activation energy, R the Boltzmann constant, and A_0 is a pre-exponential constant. Finally, $\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}} = \frac{1}{2}(\nabla\boldsymbol{u} + \nabla\boldsymbol{u}^T)$ and $\dot{\boldsymbol{\varepsilon}}_{\mathrm{II}} = \frac{1}{2}\mathrm{tr}(\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}^2)$ are the strain rate tensor and its second invariant, respectively. The top boundary, Γ_{t}, is equipped with a traction-free boundary condition, while on the basal boundary Γ_{b} we apply a no flow condition for the normal component of the velocity along with a linear sliding law for the tangential components (as shown in Figure 5.1). That is, the boundary conditions can be summarized as follows

$$\boldsymbol{\sigma}_{\boldsymbol{u}}\boldsymbol{n} = \boldsymbol{0} \qquad \text{on } \Gamma_{\mathrm{t}}, \tag{5.12d}$$

$$\boldsymbol{u}\cdot\boldsymbol{n} = 0 \qquad \text{on } \Gamma_{\mathrm{b}}, \tag{5.12e}$$

$$\boldsymbol{T}\boldsymbol{\sigma}_{\boldsymbol{u}}\boldsymbol{n} + \exp(\beta)\boldsymbol{T}\boldsymbol{u} = \boldsymbol{0} \qquad \text{on } \Gamma_{\mathrm{b}}, \tag{5.12f}$$

where $\beta(\boldsymbol{x})$ is the log basal sliding coefficient field[1], $\boldsymbol{n}$ is the outward normal unit vector, and $\boldsymbol{T} := \boldsymbol{I} - \boldsymbol{n}\boldsymbol{n}^T$ is the projection onto the tangential plane. The observational data is comprised of (noisy) point-wise measurements of the velocities, $\boldsymbol{u}$. The noise model with the accurate parameter-to-observable mapping can be written as

$$\boldsymbol{d}_{\mathrm{o}} = \mathcal{F}(\beta, z) + \boldsymbol{\epsilon} = \mathcal{B}\boldsymbol{u} + \boldsymbol{\epsilon}, \tag{5.12g}$$

where $\mathcal{B}$ denotes the observation operator. We note that this formulation is in line with [9, 8, 23, 72, 113, 114, 115].

[1] The 'exp' parametrization is used to ensure the basal sliding coefficient remains positive. Therefore β is the log basal sliding coefficient field, however for simplicity, in what follows, we refer to β as the basal sliding coefficient.

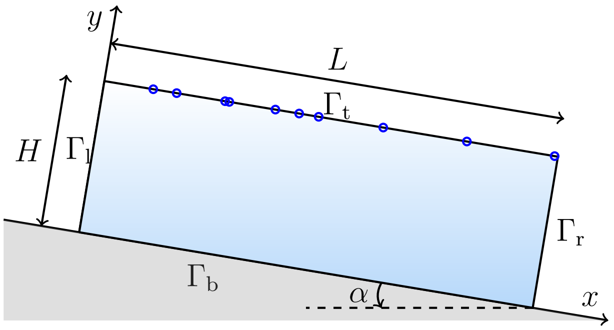


Figure 5.1: Schematic of a two-dimensional slab of ice, as used in the computational experiments. The blue circles show representative (random) measurement locations. This figure is a modification of Figure 2 from [114].

The model for the thermal distribution of ice: To model the temperature distribution within the ice, we use a simple steady state heat equation. We note that more complex coupled thermal Stokes model can be used, see for instance [115]. The temperature $\theta = \theta(\boldsymbol{x})$ is computed by solving the following boundary value problem

$$
\begin{aligned}
-\nabla \cdot (\exp(K)\nabla\theta) &= 0 && \text{in } \Omega, && (5.13a)\\
\theta &= \theta_s && \text{on } \Gamma_{\mathrm{t}}, && (5.13b)\\
\exp(K)\nabla\theta \cdot n &= \exp(G) && \text{on } \Gamma_{\mathrm{b}}, && (5.13c)
\end{aligned}
$$

where θ_s is the prescribed temperature at the top surface (chosen $\theta_s = 230$ K to be consistent with the surface temperature of a glacier [22, 25]), $\exp(G)$ is the (distributed) (log-)geothermal heat flux, and $\exp(K)$ is the (distributed) (-log) thermal conductivity of the ice. A key insight at this point is that θ is a random variable which depends (only) on G and K, assuming θ_s is known. As such the additional (auxiliary) uncertain parameters for this example are the (log-)geothermal heat flux and thermal conductivity fields. That is z in this case is $z = (G, K)$.

The approximate (low fidelity) ice sheet Stokes model model: As an approximate, but often used model in the literature, we consider the isothermal version of the Stokes problem (5.12). This model is identical to (5.12), except here we use the isothermal version of Glen's flow law, namely

$$\boldsymbol{\tau}_{\boldsymbol{u}} = 2\eta(\boldsymbol{u})\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}} \quad \text{with} \quad \eta(\boldsymbol{u}) = \frac{1}{2}A^{-\frac{1}{n}}\dot{\boldsymbol{\varepsilon}}_{\mathrm{II}}^{\frac{1-n}{2n}}, \tag{5.14}$$

where η is the effective viscosity, A is the isothermal (*i.e.*, temperature independent) flow rate factor given by $A = A_0 \exp\left(-\frac{Q}{R\theta_0}\right)$, where θ_0 is the (fixed) temperature obtained by solving the heat problem (5.13) with K and G defined as the means of their distributions.

5.4 Computing the Statistics of the Approximation Error

As discussed above, we model the error model via a Gaussian distribution. It's mean and variance (*i.e.*, it's statistics) are estimated via Monte Carlo sampling. The basic idea is the following: first we define Gaussian distributions for the additional uncertain parameters $z = (G, K)$, where $K \sim \mathcal{N}(K_*, \boldsymbol{\Gamma}_K)$ and $G \sim \mathcal{N}(G_*, \boldsymbol{\Gamma}_G)$, draw samples $(\beta^{(\ell)}, z^{(\ell)})$, $\ell = 1, 2, \ldots, N$, evaluate the model discrepancy $\boldsymbol{r}^{(\ell)} = \mathcal{F}(\beta^{(\ell)}, z^{(\ell)}) - \mathcal{G}(\beta^{(\ell)})$, and compute the mean and covariance as follows

$$\hat{\boldsymbol{r}} := \mathbb{E}[\boldsymbol{r}] \approx \boldsymbol{r}_* = \frac{1}{N}\sum_{\ell=1}^{N} \boldsymbol{r}^{(\ell)}, \quad \boldsymbol{\Gamma}_r := \mathbf{Var}[\boldsymbol{r}] \approx \frac{1}{N-1}\boldsymbol{R}\boldsymbol{R}^T, \tag{5.15}$$

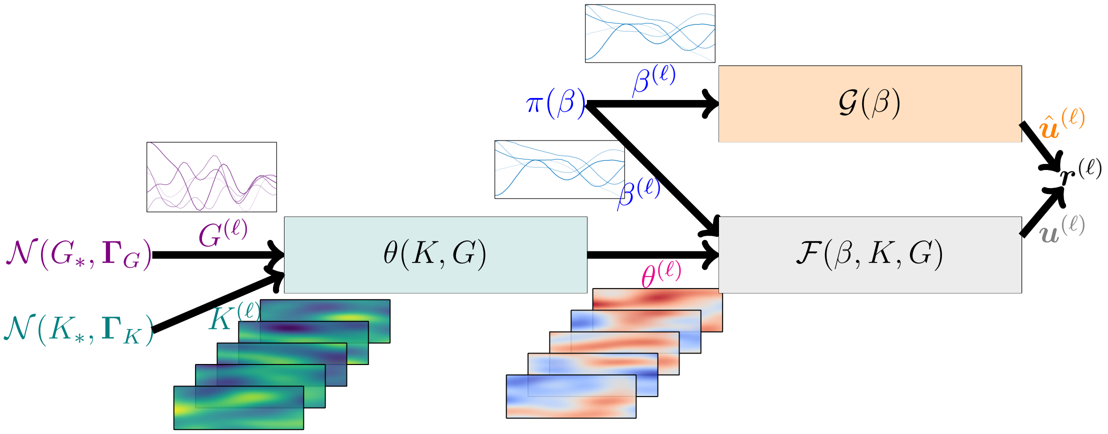
where $\boldsymbol{R} = [\boldsymbol{r}^{(1)} - \boldsymbol{r}_*, \boldsymbol{r}^{(2)} - \boldsymbol{r}_*, \ldots, \boldsymbol{r}^{(N)} - \boldsymbol{r}_*]$. The sampling procedure is summarized in **Algorithm 1**. The entire process is also illustrated in Figure 5.2.

5.5 Variance Reduction for the Bayesian Approximation Error Approach

Outlined above is the *standard version* of the BAE approach. Although all sampling is carried out *offline*, for large-scale problems, such as the continental ice sheet problem, this still poses a significant computational burden. One of the

Algorithm 1 Calculate Statistics of Approximation Error

```
procedure $(\beta_*, \boldsymbol{\Gamma}_\beta, G_*, \boldsymbol{\Gamma}_G, K_*, \boldsymbol{\Gamma}_K, N)$
    for $i \leq N$ do
        Sample $(\beta^{(i)}, G^{(i)}, K^{(i)})$ from associated priors
        Solve for $\theta^{(i)}(G^{(i)}, K^{(i)})$
        Compute $\boldsymbol{r}^{(i)} = \mathcal{F}(\beta^{(i)}, G^{(i)}, K^{(i)}) - \mathcal{G}(\beta^{(i)})$
    end for
    Compute $\boldsymbol{r}_* = \frac{1}{N}\sum_{\ell=1}^{N} \boldsymbol{r}^{(\ell)}$
    Compute $\boldsymbol{R} = [\boldsymbol{r}^{(1)} - \boldsymbol{r}_*, \boldsymbol{r}^{(2)} - \boldsymbol{r}_*, \ldots, \boldsymbol{r}^{(N)} - \boldsymbol{r}_*]$.
    Compute $\boldsymbol{\Gamma}_r = \frac{1}{N-1}\boldsymbol{R}\boldsymbol{R}^T$
end procedure
```



$\pi(\beta)$
$\beta^{(\ell)}$
$\mathcal{G}(\beta)$
$\hat{\boldsymbol{u}}^{(\ell)}$
$\boldsymbol{r}^{(\ell)}$
$\beta^{(\ell)}$
$\boldsymbol{u}^{(\ell)}$
$\mathcal{N}(G_*, \boldsymbol{\Gamma}_G)$
$G^{(\ell)}$
$\theta(K, G)$
$\theta^{(\ell)}$
$\mathcal{F}(\beta, K, G)$
$\mathcal{N}(K_*, \boldsymbol{\Gamma}_K)$
$K^{(\ell)}$

Figure 5.2: A schematic for the approximation errors sampling scheme.

major contributions of this work is to show that the number of samples required to calculate $\boldsymbol{r}_*$ and $\boldsymbol{\Gamma}_r$ can be reduced. In particular we propose the use of the first order (i.e., linear) Taylor approximation of $\boldsymbol{r}(\beta, z)$ as a control variate to reduce the variance of the Monte Carlo estimator in (5.15) for both the mean and the covariance of the model error. By reducing the number of samples we alleviate the major computational burden of the BAE approach, but still benefit from the premarginalization over the parameter and auxiliary (thermal) parameters.

Recall, $\boldsymbol{r} = \mathcal{F}(\beta, z) - \mathcal{G}(\beta)$, where β and $z = (G, K)$ are random variables.

As such, our problem reduces to a forward uncertainty propagation. There are a variety of approaches aimed at reducing the number of samples required to carry out uncertainty propagation. These methods are typically referred to as *variance reduction techniques.* Two of the more noteworthy approaches are the quasi-Monte Carlo (qMC) method [18] and the control variates method [16, 17, 14]. The draw back of qMC is that it typically does not provide any reduction in the number of samples for large dimensional cases [18]. We next outline the application of the control variates method to the BAE approach.

5.5.1 Control Variates for the Bayesian Approximation Error Approach

In what follows, let us consider the linear Taylor expansion of $\boldsymbol{r}$ in terms of all the unknowns (β, z), i.e.,

$$\boldsymbol{r}(\beta, z) \approx \boldsymbol{r}_{\text{lin}}(\beta, z) = \boldsymbol{r}(\beta_0, z) + \mathsf{D}_\beta \boldsymbol{r}(\beta_0, z)(\beta - \beta_0) - \boldsymbol{r}(\beta_0, z_0),$$

where $\mathsf{D}_\beta \boldsymbol{r}(\beta_0, z)$ denotes the Fréchet derivative of $\boldsymbol{r}$ in the direction β evaluated at (β_0, z). Consider introducing the following,

$$\begin{aligned}
&\overline{\boldsymbol{r}} = \boldsymbol{r} + \gamma\left(\boldsymbol{r}_{\text{L}} - \mathbb{E}_\beta\left[\boldsymbol{r}_{\text{L}}\right]\right), \\
&\boldsymbol{S} = (\boldsymbol{r} - \mathbb{E}_\beta[\boldsymbol{r}]) \otimes (\boldsymbol{r} - \mathbb{E}_\beta[\boldsymbol{r}]) + \gamma\left(\left(\boldsymbol{r}_{\text{L}} - \mathbb{E}_\beta\left[\boldsymbol{r}_{\text{L}}\right]\right) \otimes \left(\boldsymbol{r}_{\text{L}} - \mathbb{E}_\beta\left[\boldsymbol{r}_{\text{L}}\right]\right) - \boldsymbol{\Gamma}^{\beta}_{r_{\text{L}}}\right),
\end{aligned}$$

where at this point $\boldsymbol{r}_{\text{L}}$ is not yet chosen, but one seeks to choose a random variable strongly correlated with $\boldsymbol{r}$ (such as a linear Taylor approximation), γ is a weight constant we choose $\gamma = 1$, but in general an optimal choice can be computed see [14, 15], $\mathbb{E}_\beta[X]$ is the expectation of of a random variable X, with respect to β, and $\otimes$ is the outer product defined in Definition 7. Then clearly $\mathbb{E}_\beta\left[\overline{\boldsymbol{r}}\right] = \mathbb{E}_\beta\left[\boldsymbol{r}\right]$

and $\mathbb{E}_\beta\left[\boldsymbol{S}\right] = \boldsymbol{\Gamma}_r^\beta$ for any[2] γ, and thus

$$\mathbb{E}_\beta\left[\boldsymbol{r}\right] \approx \frac{1}{N_{\mathrm{cv}}} \sum_{\ell=1}^{N_{\mathrm{cv}}} \left(\boldsymbol{r}^{(\ell)} + \gamma\left(\boldsymbol{r}_{\mathrm{L}}^{(\ell)} - \mathbb{E}_\beta\left[\boldsymbol{r}_{\mathrm{L}}\right]\right)\right) \tag{5.16}$$

$$\boldsymbol{\Gamma}_r^\beta \approx \frac{1}{N_{\mathrm{cv}}-1} \sum_{\ell=1}^{N_{\mathrm{cv}}} \left(\boldsymbol{r}^{(\ell)} - \mathbb{E}_\beta[\boldsymbol{r}]\right) \otimes \left(\boldsymbol{r}^{(\ell)} - \mathbb{E}_\beta[\boldsymbol{r}]\right) \tag{5.17}$$

$$+ \gamma\left(\left(\boldsymbol{r}_{\mathrm{L}}^{(\ell)} - \mathbb{E}_\beta\left[\boldsymbol{r}_{\mathrm{L}}^{(\ell)}\right]\right) \otimes \left(\boldsymbol{r}_{\mathrm{L}}^{(\ell)} - \mathbb{E}_\beta\left[\boldsymbol{r}_{\mathrm{L}}^{(\ell)}\right]\right) - \boldsymbol{\Gamma}_{r_{\mathrm{L}}}^\beta\right), \tag{5.18}$$

where N_{cv} denotes the number of samples taken.

While this would appear arbitrary choice, we clarify our reasoning shortly. First, if we take $\beta_0 = \beta_*$ (while $\beta_0 \neq \beta_*$ is not required it provides some cancellations), then

$$\begin{aligned} \mathbb{E}_\beta\left[\boldsymbol{r}_{\mathrm{L}}\right] &= \boldsymbol{r}(\beta_*, z), \\ \boldsymbol{\Gamma}_{r_{\mathrm{L}}}^\beta &= \mathsf{D}_\beta \boldsymbol{r}(\beta_*, z) \boldsymbol{\Gamma}_\beta \mathsf{D}_\beta \boldsymbol{r}(\beta_*, z) \end{aligned}$$

can be computed using (relatively) few forward runs. Second, and this is the entire justification for using the control variates approach, is that the higher the correlation between $\boldsymbol{r}$ and $\boldsymbol{r}_{\mathrm{L}}$, the better the variance reduction, *i.e.*, the less samples are required to calculate $\mathbb{E}_\beta\left[\boldsymbol{r}\right]$ and $\boldsymbol{\Gamma}_r^\beta$ (using (5.16) and (5.18), respectively). For the auxiliary parameters, we use a linearization in z and use a similar approach as before. That is, we linearize around (β_0, z_0),

$$\boldsymbol{r}(\beta, z) \approx \overline{\boldsymbol{r}}_{\mathrm{L}}(\beta, z) = \boldsymbol{r}(\beta_0, z_0) + \mathsf{D}_\beta \boldsymbol{r}(\beta_0, z_0)(\beta - \beta_0) + \mathsf{D}_z \boldsymbol{r}(\beta_0, z_0)(z - z_0),$$

where $\mathsf{D}_z \boldsymbol{r}(\beta_0, z_0)$ is computed using chain rule.

For our case, *i.e.*, $z = (G, K)$, the full (Taylor) linearization is then of the form

$$\begin{aligned} \boldsymbol{r} \approx \overline{\boldsymbol{r}}_{\mathrm{L}} = \boldsymbol{r}_0 &+ \mathsf{D}_\beta \boldsymbol{r}(\beta_0, G_0, K_0)(\beta - \beta_0) + \mathsf{D}_G \boldsymbol{r}(\beta_0, G_0, K_0)(G - G_0) \\ &+ \mathsf{D}_K \boldsymbol{r}(\beta_0, G_0, K_0)(K - K_0), \end{aligned}$$

while expressing the derivatives using the *chain rule* gives

$$\mathsf{D}_K \boldsymbol{u} = \mathsf{D}_A \boldsymbol{u} \mathsf{D}_\theta A \mathsf{D}_K \theta, \quad \mathsf{D}_G \boldsymbol{u} = \mathsf{D}_A \boldsymbol{u} \mathsf{D}_\theta A \mathsf{D}_G \theta,$$

[2] An optimal γ does typically exists and is related to the correlation between $\boldsymbol{r}$ and $\boldsymbol{r}_{\mathrm{L}}$. However, it is unknown a priori in our case.

where $\mathsf{D}_K\boldsymbol{u}$ is the Fréchet partial derivative of $\boldsymbol{u}$ with respect to K, $\mathsf{D}_A\boldsymbol{u}$ is the partial with respect to the flow rate factor $A(\theta)$ (see equation (5.12c)), $\mathsf{D}_\theta A$ is the partial of A with respect to θ, $\mathsf{D}_K\theta$ is the partial of θ with respect to K, and similarly when taking the derivative with respect to G. Or more concisely,

$$[\mathsf{D}_G\boldsymbol{u},\, \mathsf{D}_K\boldsymbol{u}] = \mathsf{D}_A\boldsymbol{u}\mathsf{D}_\theta A\,[\mathsf{D}_G\theta,\, \mathsf{D}_G\theta]\,.$$

The computation of the linear Taylor approximation and in particular the partial derivatives required was carried out in the FEniCS environment. FEniCS employs Unified Form Language (UFL), which is a domain-specific language for representing weak formulations of PDEs. In UFL, the derivative of a form is based on the Gateaux derivative [128]. The computation of $\mathsf{D}_\beta\boldsymbol{r}(\beta_0,z)(\beta-\beta_0)$ will be carried out by using the UFL library to compute the derivative of the variational form encoded in FEniCS in conjunction with the Implicit Function Theorem 4.2.1, specifically we compute

$$\begin{aligned}\mathsf{D}_\beta\boldsymbol{r}(\beta_0,z)(\beta-\beta_0) &= \left(\mathsf{D}_\beta\mathcal{F}(\beta_0,z) - \mathsf{D}_\beta\mathcal{G}(\beta_0)\right)(\beta-\beta_0)\\ &= -\big([\mathsf{D}_u\boldsymbol{w}_{HF}(\beta_0,z)]^{-1}\,\mathsf{D}_\beta\boldsymbol{w}_{HF}(\beta_0,z)\big)\,(\beta-\beta_0)\\ &\quad -\big(-\,[\mathsf{D}_u\boldsymbol{w}_{LF}(\beta_0,z)]^{-1}\,\mathsf{D}_\beta\boldsymbol{w}_{LF}(\beta_0,z)\big)\,(\beta-\beta_0),\end{aligned}$$

where $\boldsymbol{w}_{HF}$ and $\boldsymbol{w}_{LF}$ are the variational representations of the PDEs associated with $\mathcal{F}$ and $\mathcal{G}$ supplied to FEniCS. Similarly we can compute the remaining derivatives in the Taylor approximation.The sampling procedure for BAE with control variate (linear Taylor approximation) can be executed using the pseudocode outlined in **Algorithm 2**.

5.6 Numerical Results

In this section, we outline the numerical example to assess the applicability, performance, and robustness of the BAE control variate approach to account for uncertain parameters in the thermal distribution of the ice. The primary uncertainty parameter is the basal sliding coefficient β and the secondary (auxiliary) parameters are $z = (G, K)$, where G is the (log-)geothermal heat flux, and K is the

Algorithm 2 Calculate Statistics of Approximation Error

```
procedure ($\beta_*, \mathbf{\Gamma}_\beta, G_*, \mathbf{\Gamma}_G, K_*, \mathbf{\Gamma}_K, N$)
    for $i \leq N$ do                                                    ▷ Parallelisable
        Sample $(\beta^{(i)}, G^{(i)}, K^{(i)})$ from associated priors
        Solve for $\theta^{(i)}(G^{(i)}, K^{(i)})$
        Compute $\boldsymbol{r}^{(i)} = \mathcal{F}(\beta^{(i)}, \theta^{(i)}) - \mathcal{G}(\beta^{(i)})$
        Compute $\boldsymbol{c}^{(i)} = \boldsymbol{r}^{(mean)} + \boldsymbol{r}_\beta^{(mean)}(\beta_i - \beta_{mean}) + \boldsymbol{r}_z^{(mean)}(z_i - z_{mean})$
    end for
    Compute $\boldsymbol{r}_* = \frac{1}{N}\sum_{\ell=1}^{N} \boldsymbol{r}^{(\ell)}$
    Compute $\boldsymbol{R} = [\boldsymbol{r}^{(1)} - \boldsymbol{r}_*, \boldsymbol{r}^{(2)} - \boldsymbol{r}_*, \ldots, \boldsymbol{r}^{(N)} - \boldsymbol{r}_*]$.
    Compute $\mathbf{\Gamma}_{r,cv} = \frac{1}{N-1}\boldsymbol{R}\boldsymbol{R}^T - \frac{1}{N-1}\boldsymbol{C}\boldsymbol{C}^T + \mathbf{\Gamma}_{\boldsymbol{c}}$
end procedure
```

(log-)thermal conductivity. The forward problems considered here are inspired by the models used in the Ice Sheet Model Intercomparison Project for Higher-Order Models (ISMIP-HOM) benchmark study carried out in [29, 27].

To set up the problem and geometry, we follow [114]. In particular, we consider a box-like geometry with $\Omega = [0, L] \times [0, H]$, where $L = 10$km and $H = 250$m inclined with slope $\theta = 0.1$ degrees. The density of the ice is fixed $\rho = 910$kg/m, and the standard gravitational constant $g = 9.81\text{s}^{-2}$. The true basal sliding coefficient field is defined as

$$\beta(s) = 7 + \sin(ws), \quad \forall s \in \Gamma_b.$$

For the inversion, we use synthetic measurements. These are randomly placed noisy pointwise measurements of each component of the velocity on the top surface of the domain Γ_{top}. The noise model $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \Gamma_e)$ with covariance matrix $\Gamma_e = \delta_e^2\mathbf{I}$. In particular we choose $\delta_e = 0.01$, *i.e.*, 1% of the range of the true synthetic measurements. The mean on the geothermal heat flux distribution (shown in Figure 5.3) was chosen such that $\exp(G)$ matches the one used in [72]. Similarly, the values for $\exp(K)$ are chosen to be within an acceptable range of a plausible thermal conductivity for ice of temperature $-40°$C on top and $-10°$C on the bottom, as suggested in [73].

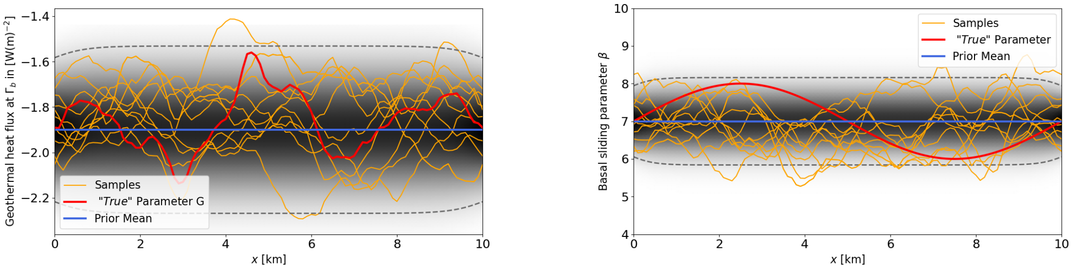


Figure 5.3: Samples (orange) from the distributions of the geothermal heat flux G (left) and of the basal sliding coefficient β (right). Red denotes the true parameter fields and blue denotes the mean of the distributions.

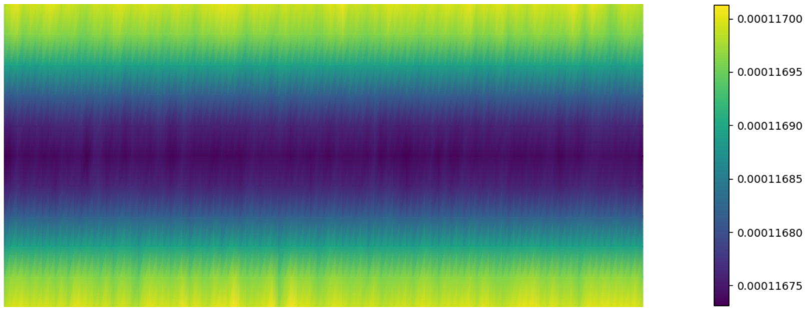
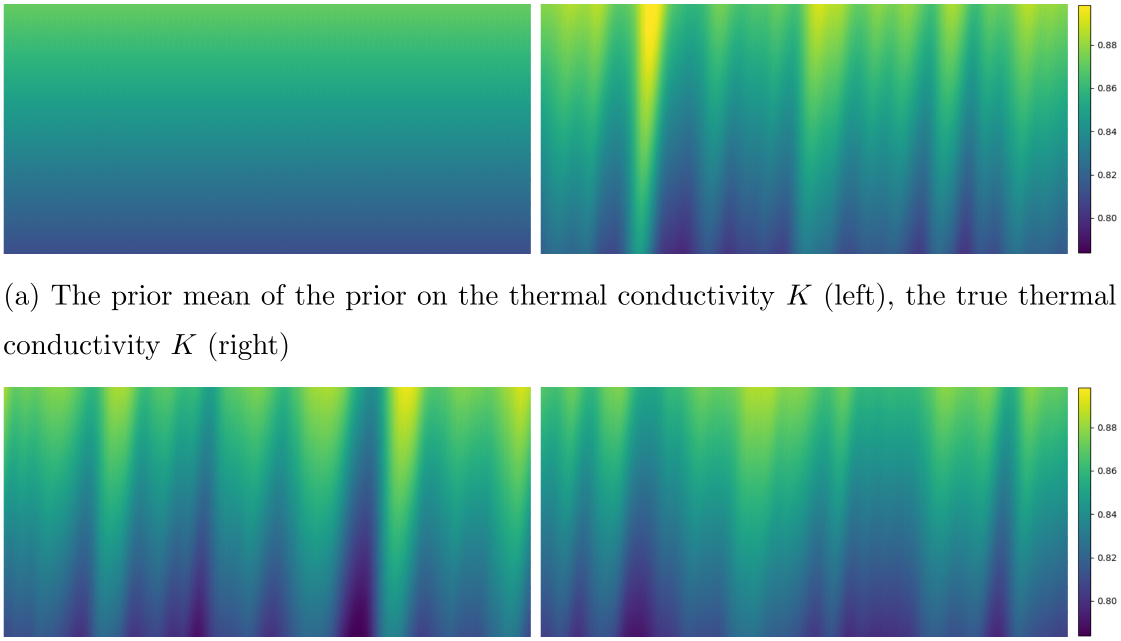


Figure 5.4: Prior variance for the thermal conductivity K.

In Figure 5.6 we show that the surface velocity can be significantly influenced by the temperature dependent rheology of the ice. Neglecting the uncertainty in the auxiliary parameters $z = (K, G)$ leads to a posterior that fails to capture the true basal sliding parameter field using the standard conventional error model (CEM). Furthermore, in Figure 5.7, one can see how the auxiliary parameters affect the temperature and in turn the basal velocity. On the other hand, if we had the truth for K and G we note that the posterior fully captures the true basal sliding coefficient (this is the so-called accurate or reference (REF) case). Finally, the Bayesian approximation error approach, which we presented in section 5.2.2, leads to a MAP point that is close to the true basal sliding coefficient, and with this approach the posterior is clearly feasible in the sense that the true basal sliding coefficient is well supported by the Gaussian approximation of the posterior.

To validate the control-variate based BAE approach, in Figure 5.9 we show the MAP estimates of the basal sliding parameter for the example problem described



(a) The prior mean of the prior on the thermal conductivity K (left), the true thermal conductivity K (right)

(b) Two samples from the thermal conductivity prior K.

Figure 5.5

above using the two approaches, the standard (left) and the control variate-based (right) BAE. In these plots we also show the true basal sliding parameter (red), samples from the respective distributions (orange), and the marginal distribution (shaded) with darker shading indicating higher probability, and the ±2 (approximate) standard deviation intervals (dashed black line). These results show that the two approaches converge to the same result. The difference between the two approaches is the number of samples required to converge, as detailed below.

In what follows we study the computational gain of the proposed control variate-based BAE approach. In Table 5.1 we report on the mean square error of the statistics estimated via the standard ($\text{MSE}(R)$) and the control variate-based ($\text{MSE}(D)$) BAE approaches. This error is defined as

$$\text{MSE}(Q) = \frac{\text{Var}[Q]}{N}, \tag{5.19}$$

where $\text{Var}[Q]$ is the variance of Q at N random samples. In the last column in

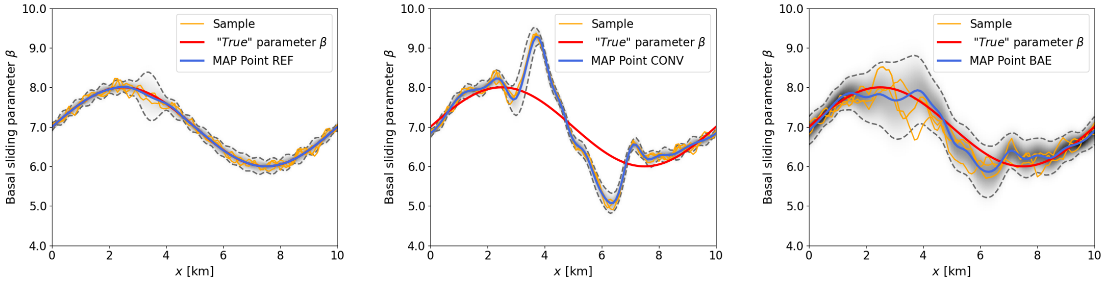


Figure 5.6: Example problem, accurate/reference (REF) case (left), conventional error model (CEM) case (center), and Bayesian approximation error (BAE) case (right), red denotes the true parameter, orange denotes samples, and blue is the inversion. The axes have been stretched in the y-direction for ease of visualization.

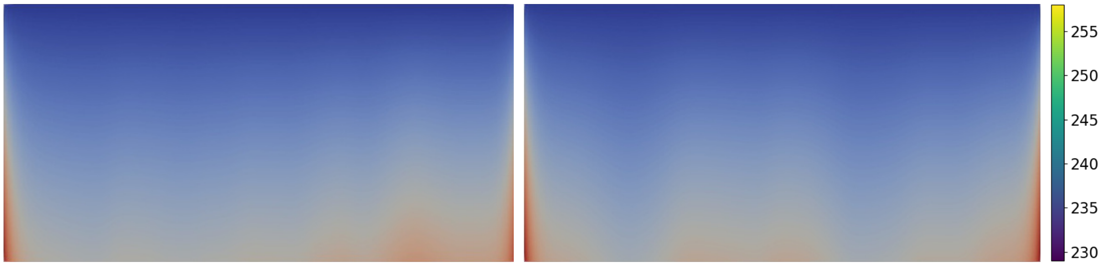


Figure 5.7: Temperature θ realized over samples of K and G.

this table, we report on the so-called speed up factor, defined as

$$\text{SpeedUpFactor} = \frac{\text{MSE}(R)}{\text{MSE}(D)},$$

which measures the computational gain of the control-variate based BAE. The results show that a 8.60 speed up can be achieved when we take a large number of samples. This means that the number of samples needed to achieve a target MSE is about eight time smaller for the control variate BAE when compared to the standard BAE.

To further study and compare the convergence properties of the two approaches, in Figure 5.10 we show the error between the “true” (*i.e.*, reference) and estimated error means defined in (5.15). The reference error mean, $\boldsymbol{r}_*$, was obtained by using the standard BAE approach with 10,000 samples. More concretely, we computed and plot the errors $\|\boldsymbol{r}_* - \boldsymbol{r}_N\|$ and $\|\boldsymbol{r}_* - \boldsymbol{r}_{N,CV}\|$, with $N = 2^k$, $k = 5, \ldots, 10$. In this plot we also check the convergence rate against the well-known Monte

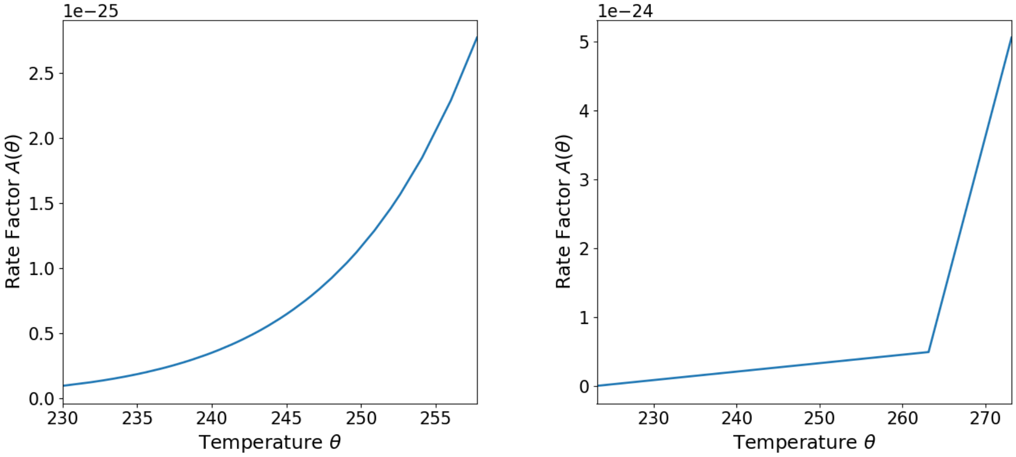


Figure 5.8: Rate factor at the mean of the prior distributions $A(\theta_{mean})$ where for the temperature range from 230K to 259K. Note (relative to the pressure melting point) according to the Arrhenius law (263K) was not reached [27, 25] . As such we do not see a kink at 263.15K is due to the piecewise definition of the pre-exponential constant A_0 and the activation energy Q.

Carlo convergence rate. The results show the control variate-based BAE approach arrives to a smaller error with a smaller number of samples when compared to the standard BAE approach.

5.7 Conclusions and Outlook

In this work, we have considered the inversion for the basal sliding coefficient field for ice sheet flow problems under uncertain rheology stemming from uncertainty in the thermal distribution of the ice. To account for the resulting model error/uncertainties, we employed the Bayesian approximation error approach. This approach shifts all uncertainty into a single additive total error term, which is approximated as Gaussian, and can be premarginalized over.

We quantified the uncertainty in the estimated basal sliding coefficient via Bayesian inversion (under Gaussian approximation of the posterior) and showed that fixing the additional or auxiliary uncertain parameters to some nominal values

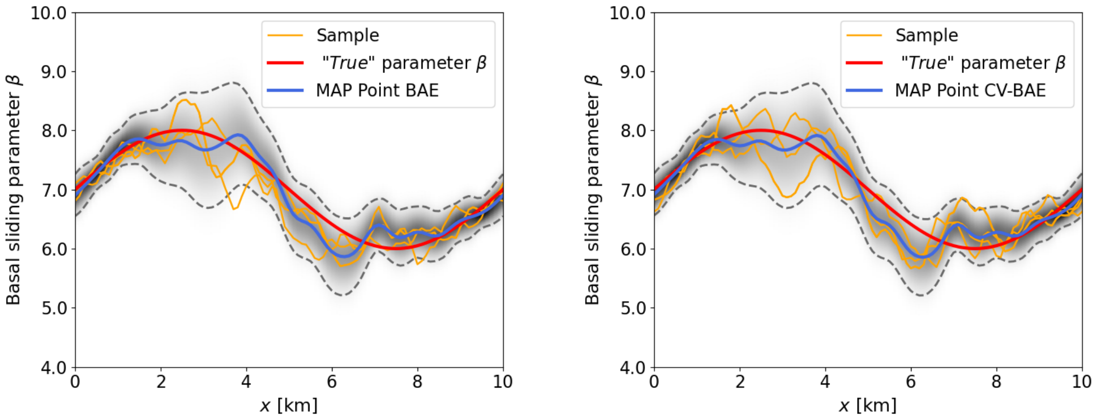


Figure 5.9: The MAP estimates of the basal sliding parameter obtained via the standard (left) and the control variate-based (right) BAE approaches. Shown are also the true basal sliding parameter (red), samples from the respective distributions (orange), the marginal distribution (shaded) with darker shading indicating higher probability, and the ±2 (approximate) standard deviation intervals (dashed black line). We note that the axes have been stretched in the y-direction for ease of visualization.

can lead to overconfident and heavily biased results.

Furthermore, we proposed a computational framework to reduce the *offline* cost of the BAE approach. Specifically, we advocate for the use of linear Taylor expansion as control variates to reduce the variance of the Monte Carlo estimator. Preliminary results suggest that the computational cost of the offline sampling stage can be reduced via this approach.

Ns	MSE(R)	MSE(D)	SpeedUpFactor
32	2.87E+00	3.22E-01	8.92
64	1.27E+00	1.20E-01	10.6
128	6.27E-01	7.03E-02	8.91
256	3.54E-01	4.11E-02	8.63
512	1.79E-01	2.02E-02	8.83
1024	8.68E-02	1.03E-02	8.41

Table 5.1: Comparison of the variance of the estimators for the standard BAE approach and our proposed control variate BAE for problem (5.10). Here **Ns** is the number of samples used for the estimator, **MSE(R)** is the mean square error of the standard BAE estimator, **MSE(R)** is the MSE of the control variate BAE estimator, and **SpeedUpFactor** is the computational gain.

Ns	MSE($\boldsymbol{\Gamma}_{r}$)	MSE($\boldsymbol{\Gamma}_{r,cv}$)	SpeedUpFactor
32	2.65E+01	9.97E+00	2.66
64	4.19E+00	1.52E+00	2.76
128	1.00E+00	4.49E-01	2.24
256	3.34E-01	1.83E-01	1.83
512	7.30E-02	3.95E-02	1.85
1024	1.60E-02	8.52E-03	1.88

Table 5.2: Comparison of the variance of the covariance matrix for the standard and control variate-based BAE approaches. Here **Ns** is the number of samples used for the estimator, **MSE($\boldsymbol{\Gamma}_{r}$)** is the mean square error of the standard BAE covariance matrix, **MSE($\boldsymbol{\Gamma}_{r,cv}$)** is the MSE of the control variate BAE covariance matrix, **SpeedUpFactor** is the computational gain.

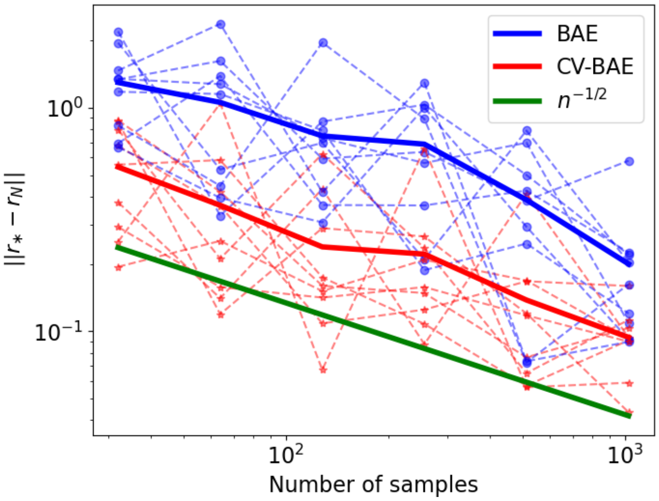


Figure 5.10: The error between the “true” (*i.e.*, reference) and estimated error means defined in (5.15) obtained via the standard (blue) and control variate-based (red) BAE approaches. The reference error mean, $\boldsymbol{r}_*$, was obtained by using the standard BAE approach with 10,000 samples. Shown are the errors obtained using various realizations (dashed blue), the mean of these errors (solid blue), and the expected $N^{-1/2}$ Monte Carlo convergence rate (green) versus number of samples.

Chapter 6

Conclusion

In this thesis, we focused on computational methods to solve large-scale inverse problems governed by PDEs. In particular, we adopted a derivative-based with line search optimization approach and computed the derivatives via adjoint methods. The first part of the thesis (Chapter 3) focused on quasi-Newton methods in infinite dimensions. In this context we derived the well-known quasi-Newton formulas in an infinite-dimensional Hilbert space setting. The second part of the thesis (Chapter 4) focused on Newton's method. Here we focused on reducing the computational cost when solving the Newton system by introducing inexactness into the underlying PDE solves. We applied these approaches for a coefficient field inverse problem governed by an elliptic PDE. The third part of the thesis (Chapter 5) was devoted to computational methods for inverse problems governed by uncertain PDEs, where the uncertainty stems from additional unknown or uncertain parameters in the forward PDE model. Such inverse problems build on deterministic inverse problems, therefore all the developments in the first parts of the thesis can be applied to reduce the computational cost of the inverse solver. In this part we build on Bayesian inversion and approximation error to account for the model error (stemming from additional unknown/uncertain parameters) when solving inverse problems governed by PDEs. In what follows, we outline possible future research avenues.

- As discussed above, in Chapter 4, we presented a rigorous framework for inexact Hessian-vector products for Newton's method. In particular we derived bounds for the tolerances to control the error of each Hessian apply, *i.e.*, PDE solve. The computation of the tolerances required the computation of the leading eigenvalue of the sub-blocks of the Hessian. As future work, we plan to explore more efficient means, such as preconditioners and warm starts to estimate the tolerances. Another idea is to explore the possibility to use reduced order models for the Hessian-applies and the effect on the convergence properties of Newton's method.

- In the context of the Bayesian Approximation Error (BAE) combined with control variates, the first goal is to investigate second-order Taylor expansion as a control variate to obtain further reduction in the computational cost of the sampling stage of BAE. Finally, we hope to apply this framework to a more realistic ice sheet inverse problem.

- A tangential project to the BAE-related research I hope to work on is sensitivity analysis for inverse problems governed by uncertain PDEs. The idea is to make the primary and secondary uncertain parameters more systematically via this analysis.

- The simulation and analysis of high-dimensional problems is often infeasible due to the curse of dimensionality. To mitigate this limitation for inverse problems, I plan to build on my internship experience with using tensor train decompositions. Specifically, I plan to apply tensor train numerical schemes for solving the forward model within inverse problems.

Bibliography

[1] G.M. Murphy, *Ordinary differential equations and their solutions.* Courier Corporation, 2011.

[2] P.J. Olver, *The calculus of variations.* Applied Mathematics Lecture Notes. Sec, 21, 4, 2012.

[3] R.G. Vuchkov, C.G. Petra, and N. Petra, *On the derivation of quasi-newton formulas for optimization in function spaces.* Numerical Functional Analysis and Optimization, 41(13), 1564-1587, 2020

[4] M. Giudici, F. Baratelli, A. Comunian, C. Vassena, and L. Cattaneo, *Model calibration for ice sheets and glaciers dynamics: a general theory of inverse problems in glaciology*, The Cryosphere Discussions, 8(5), 5511-5537, 2014.

[5] C. Zhao, R.M. Gladstone, R.C. Warner, M.A. King, T.. Zwinger, and M. Morlighem, *Basal friction of Fleming Glacier, Antarctica–Part 2: Evolution from 2008 to 2015.* The Cryosphere, 12(8), 2653-2666, 2018

[6] C. Zhao, R.M. Gladstone, R.C. Warner, M.A. King, T.. Zwinger, and M. Morlighem, *Basal friction of Fleming Glacier, Antarctica–Part 1: Sensitivity of inversion to temperature and bedrock uncertainty*, The Cryosphere, 12(8), 2637-2652, 2018.

[7] M. Morlighem, H. Seroussi, E. Larour, and E. Rignot, E, *Inversion of basal friction in Antarctica using exact and incomplete adjoints of a higher-order model.* Journal of Geophysical Research: Earth Surface, 118(3), 1746-1753, 2013

[8] T. Isaac, G. Stadler, and O. Ghattas, *Solution of nonlinear Stokes equations discretized by high-order finite elements on nonconforming and anisotropic meshes, with application to ice sheet dynamics.* SIAM Journal on Scientific Computing, 37(6), B804-B833, 2012.

[9] T. Isaac, N. Petra, G. Stadler, and O. Ghattas, *Scalable and efficient algorithms for the propagation of uncertainty from data through inference to*

prediction for large-scale problems, with application to flow of the Antarctic ice sheet. Journal of Computational Physics, 296, 348-368, 2015

[10] D. Pollard, and R.M. DeConto, *A simple inverse method for the distribution of basal sliding coefficients under ice sheets, applied to Antarctica.* The Cryosphere, 6(5), 953-971, 2012.

[11] M.J Raymond, and G.H. Gudmundsson, *Estimating basal properties of glaciers from surface measurements: a non-linear Bayesian inversion approach.* Cryosphere Discuss 3.1 (2009): 181-222, 2009.

[12] D. Farinotti, H. Corr, and G.H. Gudmundsson, *The ice thickness distribution of Flask Glacier, Antarctic Peninsula, determined by combining radio-echo soundings, surface velocity data and flow modelling.* Annals of Glaciology, 54(63), 18-24, 2013.

[13] M. Truffer, *The basal speed of valley glaciers: an inverse approach.* Journal of Glaciology, 50(169), 236-242, 2004.

[14] C.P. Robert, and G. Casella, *Monte Carlo statistical methods (Vol. 2).* New York: Springer, 1999.

[15] S.P. Brooks, and A. Gelman, *General methods for monitoring convergence of iterative simulations. Journal of computational and graphical statistics*, 7(4), 434-455, 1998.

[16] B. Peherstorfer, K. Willcox, and M. Gunzburger, *Survey of multifidelity methods in uncertainty propagation, inference, and optimization*, Siam Review, 60(3), 550-591, 2018.

[17] P.W. Glynn, and R. Szechtman, *Some new perspectives on the method of control variates in Monte Carlo and Quasi-Monte Carlo Methods 2000* (pp. 27-49), Springer, Berlin, Heidelberg, 2002.

[18] R.E. Caflisch, *Monte Carlo and quasi-Monte Carlo methods.* Acta numerica, 7, 1-49, 1998.

[19] J.P. Kaipio, T. Huttunen, T. Luostari, T. Lähivaara, and P.B. Monk, *A Bayesian approach to improving the Born approximation for inverse scattering with high-contrast materials*, Inverse Problems, 35(8), 084001, 2019.

[20] V. Rimpiläinen, A. Koulouri, F. Lucka, J.P. Kaipio, and C.H. Wolters, *Improved EEG source localization with Bayesian uncertainty modelling of unknown skull conductivity*, NeuroImage, 188, 252-260, 2019.

[21] R. Nicholson, N. Petra, and J.P. Kaipio, *Estimation of the Robin coefficient field in a Poisson problem with uncertain conductivity field*, Inverse Problems, 34(11), 115005, 2018.

[22] Y.C. Yen, *Review of thermal properties of snow, ice, and sea ice*, (Vol. 81, No. 10). US Army, Corps of Engineers, Cold Regions Research and Engineering Laboratory.

[23] O. Babaniyi, R. Nicholson, U. Villa, and N. Petra, *Inferring the basal sliding coefficient field for the Stokes ice sheet model under rheological uncertainty.* The Cryosphere, 15(4), 1731-1750, 2021

[24] L. Wasserman, *Bayesian inference.* In All of Statistics (pp. 175-192). Springer, New York, NY, 2007.

[25] K.M. Cuffey, and W.S.B Paterson, *The physics of glaciers*, Academic Press, 2010.

[26] S.I. Kabanikhin, Sergey I. *Inverse and ill-posed problems*, Inverse and Ill-posed Problems. de Gruyter, 2011.

[27] R. Greve, and H. Blatter. *Dynamics of ice sheets and glaciers*, Springer Science & Business Media, 2009.

[28] W. S. B. Paterson, *The Physics of Glaciers. Pergamon, 3rd edition*, 1994.

[29] F. Pattyn, L. Perichon,A. Aschwanden, B. Breuer, B. De Smedt, O. Gagliardini, G.H. Gudmundsson, R. Hindmarsh, A. Hubbard, J.V. Johnson. *Benchmark experiments for higher-order and full Stokes ice sheet models (ISMIP-HOM)*, The Cryosphere Discussions, 2, 111–151, 2008.

[30] A.K. Saibaba, J. Lee, and P.K. Kitanidis. *Randomized algorithms for generalized Hermitian eigenvalue problems with application to computing Karhunen–Loève expansion*, Numerical Linear Algebra with Applications, 2016.

[31] G.H. ,Golub, and C.F. Van Loan. *Matrix computations.* JHU press, 2013.

[32] H. A. Van der Vorst. *Iterative Krylov methods for large linear systems* , Cambridge University Press, 2003.

[33] G.H. Golub and M. Overton. *The convergence of inexact Chebyshev and Richardson iterative methods for solving linear systems*, Numer. Math., 53 , pp. 571–593. 1988.

[34] G.H Golub and M. Overton. *Convergence of two-stage Richardson iterative procedure for solving systems of linear equations*, Numerical Analysis, G. Watson, ed., Lectures Notes in Math. 912, Springer-Verlag, Berlin Heidelberg, New York,, pp. 128–139, 1983

[35] E. Giladi, G.H. Golub, and J. B. Keller, *Inner and outer iterations for the Chebyshev algorithm*, SIAM J. Numer. Anal., 35, pp. 300–319. 1998.

[36] Y. Notay. *Flexible conjugate gradients*. SIAM Journal on Scientific Computing, 22(4), 1444-1460, 2000.

[37] J. C. Gilbert, and C. Lemarechal. *Some numerical experiments with variable-storage quasi-Newton algorithms*, Mathematical programming, 45(1), 407-435, 1989.

[38] N. Halko, P. Martinsson, and J.A. Tropp. *Randomized methods for computing low-rank approximations of matrices*, SIAM Review, 2011.

[39] N. Halko, and P. Nathan. *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions*, Doctoral dissertation University of Colorado at Boulder, 2012.

[40] J.E. Hicken, and J. Alonso. *Comparison of reduced-and full-space algorithms for PDE-constrained optimization*, 51st AIAA Aerospace Sciences Meeting including the New Horizons Forum and Aerospace Exposition, 2013.

[41] A. Dener, G.K. Kenway, J.E. Hicken, and J. Martins. *Comparison of inexact- and quasi-Newton algorithms for aerodynamic shape optimization*, 53rd AIAA Aerospace Sciences Meeting, 2015.

[42] E. Haber. *Quasi-Newton methods for large-scale electromagnetic inverse problems*, Inverse problems, 2004.

[43] I. Ambartsumyan, W. Boukaram, T. Bui-Thanh, O. Ghattas, D. Keyes, G. Stadler, G. Turkiyyah, S. Zampini. *Hierarchical Matrix Approximations of Hessians Arising in Inverse Problems Governed by PDEs*, SIAM Journal on Scientific Computing, 2022.

[44] V. Simoncini, and D.B. Szyld. *Theory of inexact Krylov subspace methods and applications to scientific computing*, SIAM Journal on Scientific Computing, 2003.

[45] V. Simoncini, and D.B. Szyld. *Recent computational developments in Krylov subspace methods for linear systems*, Numerical Linear Algebra with Applications, 2007.

[46] A. Bouras, and V. Frayssé. *Inexact matrix-vector products in Krylov methods for solving linear systems: a relaxation strategy*, SIAM Journal on Matrix Analysis and Applications, 2005.

[47] A. Bouras, V. Frayssé, and L. Giraud. *A relaxation strategy for inner-outer linear solvers in domain decomposition methods*, Technical Report CERFACS TR/PA/17, European Centre for Research and Advanced Training in Scientific Computation, 2000.

[48] A. Bouras, and V. Frayssé. *A relaxation strategy for inexact matrix-vector products for Krylov methods*, Technical Report CERFACS TR0PA000015, European Centre for Research and Advanced Training in Scientific Computation, 2000.

[49] A. Bouras, and V. Frayssé. *A relaxation strategy for the Arnoldi method in eigenproblems*, Technical Report CERFACS TR/PA/16, European Centre for Research and Advanced Training in Scientific Computation, 2000.

[50] A. Frommer, and D.B. Szyld . *H-Splittings and two-stage iterative methodsn*, Numerische Mathematik, 1992.

[51] V. Simoncini, and D.B. Szyld. *Relaxed Krylov subspace approximation*, PAMM: Proceedings in Applied Mathematics and Mechanics, 2005.

[52] C.T Kelley. *Iterative methods for linear and nonlinear equations*, SIAM, 1995.

[53] S.C. Eisenstat, and H.F. Walker. *Globally convergent inexact Newton methods*, SIAM Journal on Optimization, 1994.

[54] S.C. Eisenstat, S. C., and H.F. Walker, *Choosing the forcing terms in an inexact Newton method.* SIAM Journal on Scientific Computing, 17(1), 16-32, 1996.

[55] J. Van Den Eshof, and G. Sleijpen. *Inexact Krylov subspace methods for linear systems*, SIAM Journal on Matrix Analysis and Applications, 2002.

[56] J. Van Den Eshof, and G. Sleijpen. *Inexact Krylov subspace methods for linear systems*, SIAM Journal on Matrix Analysis and Applications, 2004.

[57] R.S. Dembo S.C. Eisenstat, and T. Steihaug. *Inexact Newton methods*, SIAM Journal on Numerical analysis, 1982.

[58] T. Steihaug, *The conjugate gradient method and trust regions in large scale optimization.* SIAM Journal on Numerical Analysis, 20(3), 626-637, 1983.

[59] Y. Saad, *Krylov subspace methods for solving large unsymmetric linear systems*, Mathematics of computation, 37(155), 105-126, 1981.

[60] Y. Saad, and M.H. Schultz. *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM Journal on scientific and statistical computing, 7(3), 856-869, 1986.

[61] Y. Saad, *Iterative methods for sparse linear systems*, Society for Industrial and Applied Mathematics, 2003.

[62] M.J. Powell.*On search directions for minimization algorithms*, Mathematical programming, 4(1), 193-201, 1973.

[63] L. Armijo. *Minimization of functions having Lipschitz continuous first partial derivatives*, Pacific Journal of mathematics, 16(1), 1-3, 1966.

[64] P. E. Gill, W. Murray, and M.H. Wright. *Practical optimization*, Society for Industrial and Applied Mathematics, 2019.

[65] M. Evans and T. Swartz. *Approximating Integrals Via Monte Carlo and Deterministic Methods*, Vol. 20. OUP Oxford, 2000.

[66] S. J. Press. *Subjective and Objective Bayesian Statistics: Principles, Methods and Applications*, Wiley, New York, 2003.

[67] T. Bui-Thanh, O. Ghattas, J. Martin, and G. Stadler. *A computational framework for infinite-dimensional Bayesian inverse problems Part I: The linearized case, with application to global seismic inversion.* SIAM Journal on Scientific Computing 35, 6 (2013), A2494–A2523, 2013.

[68] T. Bui-Thanh, C. Burstedde, O. Ghattas, J. Martin, G. Stadler, and L.C. Wilcox. *Extreme-scale UQ for Bayesian inverse problems governed by PDEs.* SC'12: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis, 2012.

[69] J. Kaipio, and E. Somersalo, *Statistical and computational inverse problems*, (Vol. 160), Springer Science & Business Media, 2006.

[70] D. Calvetti, E. Somersalo, *An introduction to Bayesian scientific computing: ten lectures on subjective computing*, (Vol. 2). Springer Science & Business Media, 2007.

[71] A. Nissinen, L.M. Heikkinen, and J.P. Kaipio, *The Bayesian approximation error approach for electrical impedance tomography—experimental results*, Measurement Science and Technology, 19(1), 015501, 2007.

[72] N. Petra, J. Martin, G. Stadler, and O. Ghattas, *A computational framework for infinite-dimensional Bayesian inverse problems, Part II: Stochastic Newton MCMC with application to ice sheet flow inverse problems.*, SIAM Journal on Scientific Computing, 36(4), A1525-A1555, 2014.

[73] *Engineering ToolBox Ice - Thermal Properties*, Available at: https://www.engineeringtoolbox.com/ice-thermal-properties-d_576.html.

[74] G. Biros, O. Ghattas, M. Heinkenschloss, D. Keyes, B. Mallick, L. Tenorio, B. van Bloemen Waanders, K. Willcox, Y. Marzouk, and L. Biegler, *Large-scale inverse problems and quantification of uncertainty*, John Wiley & Sons, 2011.

[75] A. Alexanderian, N. Petra, G. Stadler, and O. Ghattas, *A Fast and Scalable Method for A-Optimal Design of Experiments for Infinite-dimensional Bayesian Nonlinear Inverse Problems.*, SIAM Journal on Scientific Computing, 2016.

[76] A.M. Stuart, *Inverse problems: a Bayesian perspective*, Acta numerica, 19, 451-559, 2010.

[77] N. Petra, E.W. Sachs. *Second Order Adjoints in Optimization*, In: Al-Baali, M., Purnama, A., Grandinetti, L. (eds) Numerical Analysis and Optimization. NAO 2020. Springer Proceedings in Mathematics & Statistics, vol 354, Springer, 2020.

[78] D. Williams, *Probability with martingales*, Cambridge university press., 1991.

[79] F. Troltzsch, *Optimal control of partial differential equations: theory, methods, and applications*,(Vol. 112). American Mathematical Soc., 2010.

[80] H. W. Engl, M. Hanke, A. Neubauer, *Regularization of inverse problems*,(Vol. 375). Springer Science and Business Media, 1996.

[81] A. Tikhonov, N. Arsenin, *Solutions of ill-posed problems*,Vh Winston, 1977.

[82] L. Tenorio, *An introduction to data analysis and uncertainty quantification for inverse problems*, Society for Industrial and Applied Mathematics, 2017.

[83] M. Evans and T. Swartz, *Approximating integrals via Monte Carlo and deterministic methods*, OUP Oxford, 2000.

[84] L.C. Evans, *Partial differential equations (Vol. 19)*, American Mathematical Soc., 2010

[85] Y. Pinchover, J.Rubinstein, *An introduction to partial differential equations (Vol. 10)*, Cambridge university press, 2010.

[86] J. Barzilai and J. M. Borwein, *Two-point step size gradient methods*, IMA Journal of Numerical Analysis, 8 (1988), pp. 141–148.

[87] B. Benahmed, H. Mokhtar-Kharroubi, B. De Malafosse, and A. Yassine, *Quasi-Newton methods in infinite-dimensional spaces and application to matrix equations*, Journal of Global Optimization, 49 (2011), pp. 365–379.

[88] A. Borzì and V. Schulz, *Computational optimization of systems governed by partial differential equations*, SIAM, 2012.

[89] T. Bui-Thanh and O. Ghattas, *Analysis of the Hessian for inverse scattering problems: Ii. inverse medium scattering of acoustic waves*, Inverse Problems, 28 (2012).

[90] G. Biros and O. Ghattas, *Parallel Newton-Krylov Algorithms For PDE–Constrained Optimization*, Proceedings of ACM/IEEE SC99, 1999.

[91] G. Biros and O. Ghattas, *PDE-constrained Optimization*, Springer Science & Business Media. 2012.

[92] R. H. Byrd, J. Nocedal, and R. B. Schnabel, *Representations of quasi-Newton matrices and their use in limited memory methods*, Mathematical Programming, 63 (1994), pp. 129–156.

[93] J. B. Conway, *A course in operator theory*, American Mathematical Soc., 2000.

[94] ——, *A course in functional analysis*, vol. 96, Springer Science & Business Media, 2013.

[95] C. Y. Deng, *A generalization of the Sherman–Morrison–Woodbury formula*, Applied Mathematics Letters, 24 (2011), pp. 1561–1564.

[96] J. E. Dennis, Jr and J. J. Moré, *Quasi-Newton methods, motivation and theory*, SIAM review, 19 (1977), pp. 46–89.

[97] J. Diestel, *Sequences and series in Banach spaces*, vol. 92, Springer Science & Business Media, 2012.

[98] H. W. Engl and C. W. Groetsch, *Inverse and ill-posed problems*, Academic Press, Boston, USA, 1987.

[99] H. P. Flath, L. C. Wilcox, V. Akçelik, J. Hill, B. van Bloemen Waanders, and O. Ghattas, *Fast algorithms for Bayesian uncertainty quantification in large-scale linear inverse problems based on low-rank partial Hessian approximations*, SIAM Journal on Scientific Computing, 33 (2011), pp. 407–432.

[100] M. S. Gockenbach, *Understanding and implementing the finite element method*, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2006.

[101] M. S. Gockenbach, *Partial differential equations: analytical and numerical methods (Vol. 122).* Siam, 2005

[102] W. A. Gruver and E. Sachs, *Algorithmic methods in optimal control*, vol. 47, Pitman Publishing, 1981.

[103] O. Güler, F. Gürtuna, and O. Shevchenko, *Duality in quasi-Newton methods and new variational characterizations of the DFP and BFGS updates*, Optimization Methods & Software, 24 (2009), pp. 45–62.

[104] O. Güler, *Foundations of optimization*, Vol. 258. Springer Science & Business Media, 2010.

[105] L. B. Horwitz and P. E. Sarachik, *Davidon's method in Hilbert space*, SIAM Journal on Applied Mathematics, 16 (1968), pp. 676–695.

[106] J. K. Hunter and B. Nachtergaele, *Applied analysis*, World Scientific Publishing Company, 2001.

[107] R. C. Kirby, *From functional analysis to iterative methods*, SIAM review, 52 (2010), pp. 269–293.

[108] D. G. Luenberger, *Optimization by vector space methods*, John Wiley & Sons, 1997.

[109] R. Mayorga and V. Quintana, *A family of variable metric methods in function space, without exact line searches*, Journal of Optimization Theory and Applications, 31 (1980), pp. 303–329.

[110] G. J. Murphy, *C*-algebras and operator theory*, Academic press, 2014.

[111] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, New York, 2nd ed., 2006.

[112] C. G. Petra, M. S. D. Troya, N. Petra, Y. Choi, G. M. Oxberry, and D. Tortorelli, *A quasi-Newton interior-point method for optimization in Hilbert spaces*, Submitted., (2019).

[113] N. Petra and G. Stadler, *Model variational inverse problems governed by partial differential equations*, Tech. Rep. 11-05, ICES, The University of Texas at Austin, 2011.

[114] N. Petra, H. Zhu, G. Stadler, T.J. Hughes, and O. Ghattas, *An inexact Gauss-Newton method for inversion of basal sliding and rheology parameters in a nonlinear Stokes ice sheet model*, Journal of Glaciology, 58(211), 889-903, 2012

[115] N. Petra, H. Zhu, G. Stadler, T.J. Hughes, and O. Ghattas, *Inversion of geothermal heat flux in a thermomechanically coupled nonlinear Stokes ice sheet model.* The Cryosphere, 10(4), 1477-1494, 2016.

[116] K. Wang, T. Bui-Thanh, O. Ghattas, *A randomized maximum a posteriori method for posterior sampling of high dimensional nonlinear Bayesian inverse problems.* SIAM Journal on Scientific Computing, 40(1), A142-A171.

[117] J. N. Reddy, *An introduction to the finite element method*, McGraw-Hill Higher Education, 1994.

[118] F. Riesz and S. Nagy, *B. Functional analysis*, Dover Publications, Inc., New York. First published in, 3 (1955), p. 35.

[119] J.W. Glen, *The creep of polycrystalline ice*, Proceedings of the Royal Society of London, Series A, Mathematical and Physical Sciences, 228, 519–538, 1955.

[120] W. Rudin, *Functional analysis. 1991*, Internat. Ser. Pure Appl. Math, (1991).

[121] ——, *Real and complex analysis*, Tata McGraw-Hill Education, 2006.

[122] E. W. Sachs, *Broyden's method in Hilbert space*, Mathematical Programming, 35 (1986), pp. 71–82.

[123] T. Schwedes, S. W. Funke, and D. A. Ham, *An iteration count estimate for a mesh-dependent steepest descent method based on finite elements and Riesz inner product representation*, arXiv preprint arXiv:1606.08069, (2016).

[124] T. Schwedes, D. A. Ham, S. W. Funke, and M. D. Piggott, *Mesh dependence in PDE-constrained optimisation. An application in tidal turbine array layouts*, SpringerBriefs in Mathematics of Planet Earth, Springer, 2017.

[125] N. Petra, and G. Stadler. *Model variational inverse problems governed by partial differential equations*, TEXAS UNIV AT AUSTIN INST FOR COMPUTATIONAL ENGINEERING AND SCIENCES, 2011.

[126] U. Villa, N. Petra, and O. Ghattas, *hIPPYlib: An extensible software framework for large-scale inverse problems governed by PDEs; Part I: Deterministic inversion and linearized Bayesian inference*, ACM Transactions on Mathematical Software, (2021).

[127] U. Villa, N. Petra, and O. Ghattas, *hIPPYlib: An Extensible Software Framework for Large-Scale Inverse Problems*, Journal of Open Source Software, 2018.

[128] T. Dupont,J. Hoffman,C. Johnson,R.C. Kirby, M.G. Larson, A. Logg, L.R. Scott, *The FEniCS project*, Chalmers Finite Element Centre, Chalmers University of Technology, 2003.

[129] A. Logg, Kent-Andre Mardal, G.N. Wells, *Automated solution of differential equations by the finite element method*, Lecture Notes in Computational Science and Engineering Vol 84, 2012.

[130] S. Balay, K. Buschelman, W.D. Gropp, D. Kaushik, M. Knepley, L.C. McInnes, B.F. Smith, and H. Zhang, *PETSc home page*, http://www.mcs.anl.gov/petsc, 2001.

[131] S. Balay, K. Buschelman, W.D. Gropp, D. Kaushik, M. Knepley, L.C. McInnes, B.F. Smith, and H. Zhang, *PETSc home page*, http://www.mcs.anl.gov/petsc, 2009.

[132] C. R. Vogel, *Computational methods for inverse problems*, Frontiers in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002.

[133] A. Wächter and L. T. Biegler, *Line search filter methods for nonlinear programming: Local convergence*, SIAM Journal on Optimization, 16 (2005), pp. 32–48.

[134] ——, *Line search filter methods for nonlinear programming: Motivation and global convergence*, SIAM Journal on Optimization, 16 (2005), pp. 1–31.

[135] A. Wächter and L. T. Biegler, *On the implementation of an interior-point filter line-search algorithm for nonlinear programming*, Mathematical Programming, 106 (2006), pp. 25–57.

[136] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods (Springer Texts in Statistics)*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

[137] B. Peherstorfer, K. Willcox, and M. Gunzburger, *Survey of multifidelity methods in uncertainty propagation, inference, and optimization*, Siam Review, 60 (2018), pp. 550–591.

[138] A. Alexanderian, N. Petra, G. Stadler, and O. Ghattas, *Mean-variance risk-averse optimal control of systems governed by PDEs with random parameter fields using quadratic approximations*, SIAM/ASA Journal on Uncertainty Quantification, 5 (2017), pp. 1166–1192.

[139] P. Chen, U. Villa, and O. Ghattas, *Taylor approximation and variance reduction for PDE-constrained optimal control under uncertainty*, Journal of Computational Physics, 385 (2019), pp. 163–186.

[Hutter, 1983] Hutter, Kolumban, 1983. Theoretical Glaciology, Mathematical Approaches to Geophysics, D. Reidel Publishing Company.

[Paterson, 1994] Paterson, W. Stanley B., 1994. The Physics of Glaciers, Butterworth Heinemann, third ed.

[Marshall, 2005] Marshall, Shawn J., 2005. Recent advances in understanding ice sheet dynamics, *Earth and Planetary Science Letters*, **240**, 191–204.