Particle based sampling and optimization methods for inverse problems

Inauguraldissertation
zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften
der Universität Mannheim

vorgelegt von

Simon Leander Weissmann
aus Eberbach

Mannheim, 2020

Dekan: Dr. Bernd Lübcke, Universität Mannheim
Referent: Prof. Dr. Claudia Schillings, Universität Mannheim
Korreferent: Prof. Dr. Andrew Stuart, California Institute of Technology
Korreferent: Prof. Dr. Dirk Blömker, Universität Augsburg

Tag der mündlichen Prüfung: 18. November 2020

Abstract

In this thesis, we present and analyse several ensemble based methods for inverse problems. The aim is to analyse various particle based methods for sampling as well as optimization in inverse problems.
Firstly we examine the ensemble Kalman inversion, which has been originally introduced as a sampling method for Bayesian inverse problems, but can also be viewed as derivative free optimization method. Furthermore, we present various transformed methods of the ensemble Kalman inversion, which allow to incorporate box-constraints as well as regularization for the underlying optimization problem.
In addition, we also consider a more general class of particle based sampling methods, such as the ensemble Kalman sampler, which is based on an interacting Langevin dynamics, a particle system resulting from an Gaussian approximation, as well as a kernelized Fokker–Planck based particle system.
In the last part of this work, we discuss machine learning applications in inverse problems. Here, we consider data-driven regularization, where the regularization parameter will be chosen by solving a bilevel optimization problem. Moreover, we consider an incorporation of neural networks into inverse problems. For this incorporation the neural network will act as a model-informed surrogate for the complex forward model. The neural network and the unknown parameter will be trained in a one-shot fashion.

Zusammenfassung

In dieser Arbeit präsentieren und analysieren wir verschiedene ensemblebasierte Methoden für inverse Probleme. Das Ziel ist es, verschiedene partikelbasierte Methoden sowohl zur Generierung von Stichproben als auch zur Optimierung für inverse Probleme zu analysieren.
Zunächst behandeln wir die Ensemble Kalman Inversion, die ursprünglich als Stichprobenverfahren für Bayessche inverse Probleme eingeführt wurde, allerdings auch als ableitungsfreies Optimierungsverfahren betrachtet werden kann. Desweiteren stellen wir verschiedene Transformationen der Ensemble Kalman Inversion vor, die es erlauben Box-Einschränkungen sowie Regularisierungsverfahren in das zugrundeliegende Optimierungsproblem einzubauen.
Zusätzlich betrachten wir eine größere Klasse von partikelbasierten Stichprobenverfahren. Diese beinhaltet den Ensemble Kalman Sampler, der auf einer interagierenden Langevin Dynamik basiert, ein aus einer Gaussapproximation resultierendes Partikel-System, sowie ein Partikel-System, das aus einer kernbasierten Fokker–Planck Gleichung entsteht.
In dem letzten Teil dieser Arbeit diskutieren wir Anwendungen des maschinellen Lernens auf inverse Probleme. Wir betrachten daten-getriebene Regularisierungsverfahren, in denen der Regularisierungsparameter durch die Lösung eines Bilevel-Optimierungsproblems gewählt wird. Außerdem betrachten wir die Eingliederung von neuronalen Netzen in inverse Probleme. Für diese spielt das neuronale Netz die Rolle eines Modell-informierten Surrogats für das komplexe Vorwärtsmodell. Das neuronale Netz und der unbekannte Parameter werden parallel trainiert.

Acknowledgements

I would like to express my deep gratitude to Prof. Dr. Claudia Schillings for the intense supervision over the last three years and for giving me the opportunity to write this thesis. I am very thankful for many helpful discussions and advices for my research. Furthermore, I have enjoyed the many opportunities she gave me to visit various conferences and workshops, where I was able to share my research with others.
Moreover, I would like to thank Prof. Dr. Dirk Blömker and Prof. Dr. Andrew Stuart for kindly agreeing to co-referee this thesis.
During my time as PhD student, I had the pleasure to work with numerous people. I would like to thank the "SFB 1294 - Data Assimilation" and in particular Prof. Dr. Sebastian Reich for providing me the five weeks visiting research fellowship at the University of Potsdam. I also thank Dr. Neil Chada and Prof. Dr. Xin Tong for helping me to vistit them at the National University of Singapore. During these research stays my research work progressed greatly. I am also very grateful to Prof. Dr. Dirk Blömker and Dr. Philipp Wacker for having many helpful discussion with me.
I would also like to thank my office mates Niklas, Lukas and Philipp for having interesting discussions on mathematical problems but also for taking a break by talking about non-mathematical topics.
My special thanks goes to my family. In particular, I am deeply thankful to my parents Andrea and Bernhard, my sisters Elena, Larissa and Madita, my brother Philip and specially to my girlfriend Vanessa, for always supporting me and believing in me.
Finally, I am very grateful to the "RTG 1953 - Probability & Statistics group Heidelberg-Mannheim" funded by the Deutsche Forschungsgesellschaft for funding my research and in particular, for the financial support for my research stay in Singapore.

Contents

1 Introduction

The research area of inverse problems has seen a wide range of applications in science. Some examples of this contain medical imaging, which includes tomography and acoustics, or geophysics, such as seismic inversion and Darcy flow. This is just a short extract from a long list of applications. Inverse problems are concerned with the task of recovering some quantity of interest, which cannot be observed directly. Typically, the information of interest can only be observed indirectly through a forward model, which is based on an underlying physical system. These models are often based on partial differential equations.

Inverse problems are typically ill-posed and one aspect in the theory for inverse problems is to provide well-posedness results of the underlying optimization problem. This means, one solves a regularized problem, where a unique solution exists, which is stable with respect to changes in the data. In particular, the stability is an important property in the presence of noise in the data, which could arise from the measurement. Classical approaches for inverse problems are concerned with the choice of the regularization parameter and the corresponding convergence result in the small noise limit.

An alternative approach for the regularization of inverse problems is the Bayesian approach. Here, the unkown parameter as well as the data are modelled as random variables. The task of the Bayesian inverse problem is to quantify the information of the unknown parameter conditioned on the realization of the data. The resulting solution of the Bayesian inverse problem is given by a probability distribution. This distribution might be not accessible in a straightforward way and there has been a wide range of research on sampling methods for Bayesian inverse problems. However, the advantage of this perspective for inverse problems is the possibility of doing statistical analysis for inverse problems.

In this thesis, we are going to analyse several particle based methods for inverse problems. These methods will focus on sampling as well as on optimization for inverse problems. In particular, one major part of this work is about the ensemble Kalman filter applied to inverse problems. The ensemble Kalman filter has been originally introduced for data assimilation problems and more recently has been formulated to solve inverse problems. This method is known as the ensemble Kalman inversion. As it has been introduced originally, the ensemble Kalman inversion can be interpreted as a sampling method for inverse problems in the Bayesian setting. However, by exploiting its gradient flow structure, the ensemble Kalman inversion can also be seen as derivative-free optimization method.

While the ensemble Kalman inversion is the main aspect of the particle based optimization methods, we present a range of alternative particle based sampling methods for inverse problems. Here, we will consider the ensemble Kalman sampler, which is based on an interacting Langevin dynamic. This method can be interpreted as Markov chain Monte Carlo method, where constructed dynamics has the target distribution as stationary distribution in the long-time limit. Furthermore, we provide a particle system which is based on a Gaussian approximation as well as a Fokker–Planck based particle system, which

is constructed by approximating the Fokker–Planck equation of Langevin dynamics in a reproducing kernel Hilbert space.

1.1 Outline

Chapter 2

We start this work with an introduction of several tools which will be needed for the later chapters. This includes an introduction to inverse problems in Section 2.1, where we start with general nonlinear inverse problems. Here, the focus will be on the definition of well-posedness and Tikhonov regularization. In Section 2.2 we consider the Bayesian approach for inverse problems. We discuss well-posedness of the problem, establish the connection to regularization of classical inverse problems via the MAP estimate and introduce basic sampling methods for the posterior distribution. The last Section 2.3 introduces data assimilation problems. We keep the focus on the Kalman filter and its extension to the ensemble Kalman filter.

Chapter 3

In Chapter 3 we consider the so called ensemble Kalman inversion and its formulation as derivativefree particle based optimization method. We firstly introduce the ensemble Kalman inversion as application of the ensemble Kalman filter to inverse problems as it has been originally introduced in [112]. Based on the continuous-time formulation presented in [200], we analyie the optimization scheme in the linear setting. While the existing results were based on unperturbed observations, resulting in a system of ordinary differential equations, we consider the ensemble Kalman inversion with perturbed observations, leading to a system of stochastic differential equations. We show well-posedness of the scheme, which means we verify existence of unique strong solutions of the underlying stochastic differential equation, and quantify the ensemble collapse. Here, the theoretical analysis is based on stochastic Lyapunov functions. Further, we incorporate variance inflation in order to ensure convergence of the data misfit. The presented theoretical analysis will be verified by a numerical example.
This chapter is based on joint work with Dirk Blömker, Claudia Schillings and Philipp Wacker and the corresponding article *Well posedness and convergence analysis of the ensemble Kalman inversion*, Inverse Problems, Volume 35, Number 8, 2019 (doi: 10.1088/1361-6420/ab149c).

Chapter 4

In this chapter, we focus on the analysis of the ensemble Kalman inversion for inverse problems where the unknown parameter satisfies box-constraints. Following the ideas of projected gradient descent method we formulated the projected ensemble Kalman inversion and derive its continuous-time limit resulting again in a system of ordinary differential equations. In order to ensure convergence to a minimizer of the constrained optimization problem, we present a transformed method, which manipulates the preconditioner of the scheme to ensure descent direction of the method. We analyze the introduced method in both linear and nonlinear numerical examples.

This chapter is based on joint work with Neil Chada and Claudia Schillings and contains the results of the publication *On the incorporation of box-constraints for ensemble Kalman inversion*, Foundations of Data Science, Volume 1, Issue 4, 2019 (doi: 10.3934/fods.2019018).

Chapter 5

Chapter 5 is devoted to extend the presented theoretical result of the ensemble Kalman inversion in Chapter 3, which has been based on noisefree data, to the incorporation of noise into the underlying data. Firstly, we introduce the Tikhonov regularized ensemble Kalman inversion presented in [41], and establish well-posedness and convergence results for the resulting system of stochastic differential equations, which is based on a fixed regularization parameter. Secondly, we formulate various ideas of adapting the regularization parameter, which includes data-driven regularization and also the MAP formulation of the Bayesian inverse problem. We conclude this chapter with a numerical analysis of the introduced adaptive schemes, which highlights that these schemes are promising directions to go in.
This chapter is based on joint work with Neil Chada, Claudia Schillings and Xin Tong, which has not been published yet.

Chapter 6

In Chapter 6 we present various particle based sampling methods. Roughly speaking, the considered methods are all designed in order to solve the Bayesian inverse problem by converting a sample from the prior distribution into a sample from the posterior distribution. These methods include the so called ensemble Kalman sampler, which is based on an interacting Langevin dynamics and has been introduced in [83], a particle system resulting from an Gaussian approximation, which aims to minimize the Kullback–Leibler divergence between the Gaussian approximation and the posterior distribution, and lastly a Fokker–Planck based particle system introduced in [175], which approximates the associated Fokker–Planck equation of Langevin dynamics in a reproducing kernel Hilbert space. We connect all of these methods through its gradient structured formulation resulting in a system of ordinary differential equations and stochastic differential equations respectively. Furthermore, we present derivativefree modications of these methods by preconditioning with the sample covariance of the particle system. In order to tune the accuracy of the derivativeapproximation through the sample covariance, we introduce localisation by preconditioning with a localised sample covariance based on weights depending on the distance of the particles. We demonstrate the effectiveness of the presented methods in a row of numerical examples.
This chapter is based on joint work with Sebastian Reich and the correspoding preprint *Fokker–Planck particle systems for Bayesian inference: Computational approaches*, arXiv e-prints, 2019 (arXiv:1911.10832).

Chapter 7

This chapter concerns Machine learning approaches in the context of inverse problems.

The first part of this chapter, Section 7.1, is about data-driven regularization, where we consider bilevel optimization as methodology to learn a regularization parameter for minimization based inverse problems. We view the underlying bilevel optimization problem

as stochastic optimization problem by viewing the unknown parameter and the data as random variables. The upper-level problem is given by some risk-measure between the unknown parameter and the solution of the lower level problem, which is the regularized minimization problem of the corresponding inverse problem depending on the regularization parameter and the data. Assuming to have access to training data, we consider an empirical approximation of the stochastic optimization problem and provide both offline and online consistency results for the size of training data approaching infinity. In the offline setting we analyze the minimization task of the empirical lossfunction, whereas we introduce the stochastic gradient descent method in order to solve the stochastic optimization problem online. In both settings, we firstly provide an abstract consistency result for general nonlinear forward models and general regularization function and secondly verify the presented result for linear forward model under Tikhonov regularization. We provide various numerical examples analyzing the presented consistency results.
This part of the chapter is based on joint work with Neil Chada, Claudia Schillings and Xin Tong, and the preprint *Consistency analysis of bilevel data-driven learning in inverse problems*, arXiv e-prints, 2020 (arXiv:2007.02677).

In the second part of this chapter, Section 7.2-7.2.2, we consider the incorporation of neural networks into inverse problems. We replace the complex forward model by a neural network acting as a physics-informed surrogate model, which will be trained in a one-shot fashion. This means we train the unknown parameter and the neural network at once, i.e. the neural network is only trained for the underlying unknown parameter. We connect the neural network based one-shot formulation to the Bayesian approach for inverse problems and apply the ensemble Kalman inversion in order to solve the optimization problem. We provide numerical experiments to highlight the promising direction of neural network based one-shot formulation together with the application of the ensemble Kalman inversion.
This part of the chapter is based on joint work with Philipp Guth and Claudia Schillings and the preprint *Ensemble Kalman filter for neural network based one-shot inversion*, arXiv e-prints, 2020 (arXiv:2005.02039).

2 Preliminaries

The first chapter is devoted to set up some basic background which will be needed in the rest of this work. In Section 2.1, we start the discussion by introducing inverse problems with a focus on well-posedness and Tikhonov regularization. This means, we state existence and uniqueness results, a stability result and convergence results from the literature. In particular, this part is based on the textbooks [76, 118, 204].
The second part, Section 2.2, concerns the Bayesian approach for inverse problems, where we discuss well-posedness of the problem, discuss prior modelling and connect the Baysian inverse problem to classical regularization via the maximum a-posteriori estimators. Furthermore, we give a short introduction of the Markov chain Monte Carlo method, which is used to build up Monte Carlo estimates for the posterior distribution. This section follows the approaches presented in the textbook [216] and the Acta Numerica article [215].
In Section 2.3, we introduce the research field of data assimilation with focus on the Kalman filter, which is based on the textbooks [147, 186, 216] and the Acta Numerica article [185].

2.1 Introduction to inverse problems

In this section, we introduce a general setting of ill-posed inverse problems and explain basic ideas of regularization. Typically, inverse problems are applied in situations where the quantity of interest is only availiable indirectly through observations. To define an inverse problem, one firstly has to define the corresponding forward problem, which typically based on an underlying physical system. The task of the inverse problem is to quantify information, which cannot be observed directly, but it can be observed indirectly through the forward model. The research field of inverse problems has a wide range of applications such as (medical) imaging [18, 35, 42, 167], geophysics [219], oil industry [114] and many more. From a mathematical perspective, inverse problems are often applied to models based on partial differential equations [51, 178] such as inverse scattering [132] or parameter identifications [13, 141]. The focus will be on the minimization based formulation with regularization of inverse problems [107, 124] . In this work, we will mainly focus on the so called Tikhonov regularization. For more details, we refer the reader to the following textbooks [76, 118, 204].

2.1.1 Inverse problems

In the following we introduce the challenges of inverse problems. Let $\mathcal{X}$ and $\mathcal{Y}$ be Banach spaces and consider a possibly nonlinear operator $H : \mathcal{D}(H) \to \mathcal{Y}$, where $\mathcal{D}(H) \subset \mathcal{X}$ denotes the domain of H. We call the computation of the *data* (sometimes also called

observation)

$$y = H(\theta), \tag{2.1}$$

for given *parameter* $\theta \in \mathcal{X}$, the **forward problem**. The corresponding **inverse problem** is to recover the unknown parameter $\theta \in \mathcal{X}$ for given data $y \in \mathcal{Y}$. Usually, the given data y results from an approximation of a physical model. We include noise models, which will be the additive noise model in this work, i.e. the inverse problem is to recover the parameter θ from noisy observation

$$H(\theta) + \xi = y, \tag{2.2}$$

where ξ denotes observational noise.
So far we have introduced two problems, the forward and inverse problem. But what are the challenges of these introduced problems? Firstly, we will define what we mean by well-posed problems and explain why inverse problems are in general ill-posed.

Definition 2.1.1 (Hadamard 1902, [91])**.** *A problem is called* ***well-posed****, if*

- *there exists a solution (existence),*
- *the solution is unique (uniqueness),*
- *the solution's behavior changes continuously with the initial conditions (stability).*

If one of these properties does not hold, the problem is called ***ill-posed****.*

Remark 2.1.2. *We assume that the forward model H is linear, i.e. we assume that $H(\cdot) = L\cdot$ for some $L \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$, where $\mathcal{L}(\mathcal{X}, \mathcal{Y})$ denotes the space of all linear and bounded maps from $\mathcal{X}$ to $\mathcal{Y}$. While the forward problem* (2.1) *is obviously well-posed, there are certain assumptions necessary to ensure well-posedness of the corresponding inverse problem* (2.2)*. For example sufficient conditions are given in the following:*

- *The assumption $y \in \mathcal{R}(L)$ or surjectivity of L ensures existence of solutions. Through noise it could happen, that the perturbation shifts the noisefree observation outside of the range of the forward map L although there exists a true parameter $\theta^\dagger$ which generates the data $y = L\theta^\dagger + \xi$. This is issue is illustrated in Figure 2.1.*

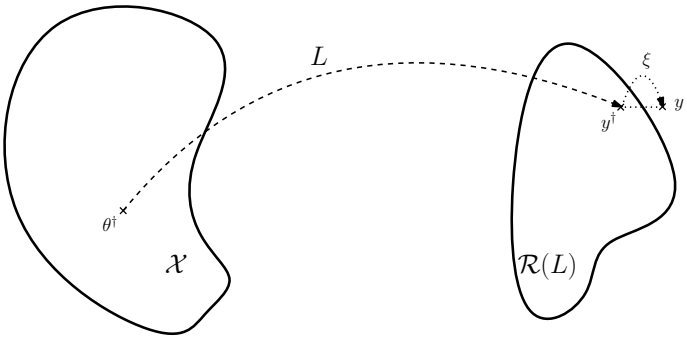

Figure 2.1: Ill-posed through observational noise. Occurance of noise might shift the observed data outside of the range of the forward map L.

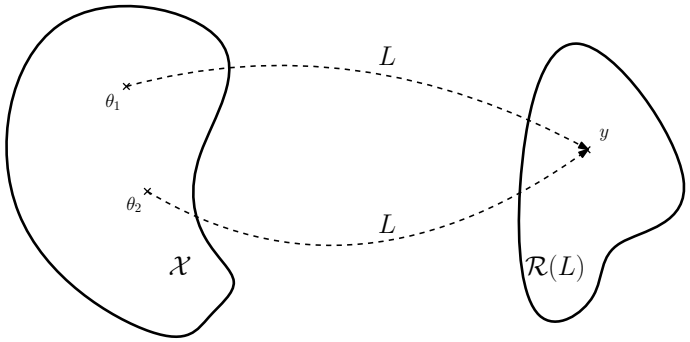

Figure 2.2: Ill-posed through multiple solutions. Two different parameter $\theta_1 \in \mathcal{X}$ and $\theta_2 \in \mathcal{X}$ might map onto the observed data y.

- *Injectivity of L ensures uniqueness of solutions, whereas in other cases problems in distinguishability of the solutions could arise. We demonstrate this problem in Figure 2.2.*
- *If L^{-1} exists and is continuous, the solution is stable. In case of discontinuous inverse L^{-1} noise in the measurement will obviously lead to instability in the solution of the inverse problem, as the following Figure 2.3 demonstrates.*

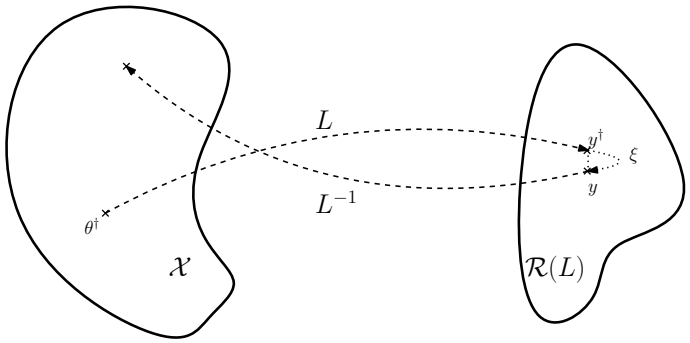

Figure 2.3: Ill-posed through discontinuity. Instability in the solution of the inverse problem resulting from a discontinuous inverse operator L^{-1}.

In the following example we will illustrate the occurrence of ill-posed inverse problems.

Example 2.1.3 ([118, Example 1])**.** *We consider an inverse problem based on the heat-equation*

$$\begin{cases} \frac{\partial^2 u(t,x)}{\partial t^2} - \frac{\partial u(t,x)}{\partial t} = 0, & t > 0, \\ u(t,x) = 0, & x \in \{0,1\}, \\ u(t,x) = \theta_0(x), & t = 0. \end{cases} \tag{2.3}$$

The inverse problem is to recover the unknown initial condition $\theta_0(\cdot)$ given the state $u(T,\cdot)$ at time $T > 0$. As an interpretation, one can think about a stick of length 1 with unit thermal conductivity and the ends of the stick are set with fixed temperature, in this setting 0. The temperature distribution $u(t,x)$ is modelled through (2.3). Hence, we aim to recover

the initial temperature distribution of the stick given the current temperature distribution in $T > 0$.
Applying the Fourier transformation, we can write the solution of the heat-equation as

$$u(t,x) = \sum_{k=1}^{\infty} c_k \exp(-(k\pi)^2 t) \sin(k\pi x),$$

where the coefficients c_k can be found by the Fourier sine coefficients of the initial state

$$\theta_0(x) = \sum_{k=1}^{\infty} c_k \sin(k\pi x). \tag{2.4}$$

Hence, the inverse problem can be broken down to finding the coefficients c_k of the initial state given the data at time $T > 0$. The issue of this task can be seen in the following scenario. Let us consider two initial states $\theta_0^{(1)}(\cdot)$ and $\theta_0^{(2)}(\cdot)$, which are equal in all components of (2.4), but only differ in one high-frequency component. The difference of both states can be written as

$$\theta_0^{(1)}(x) - \theta_0^{(2)}(x) = c_N \sin(N\pi x)$$

for large $N \in \mathbb{N}$. Pushing the initial conditions forward to the solution of the heat-equation, the difference of both solutions at time $T > 0$ is exponentially small

$$u^{(1)}(T,x) - u^{(2)}(T,x) = c_N \exp(-(N\pi)^2 T) \sin(N\pi x),$$

which means that information arising in high-frequency has a low effect on the corresponding solution of the heat-equation and hence, will not be covered in the presence of measurements errors.

The first intuition how to solve the inverse problem (2.2) would be to minimize the data misfit, i.e. to solve

$$\min_{\theta \in \mathcal{D}(H)} \|H(\theta) - y\|^2. \tag{2.5}$$

As we have seen in Remark 2.1.2, solving (2.5) is again not a well-posed problem. In fact, just fitting the data will usually lead to so called *overfitting*, i.e. the optimization method to solve (2.5) fits the noise within our data for the inverse problem (2.2), which will lead to worse resulting estimation for the unknown truth $\theta^\dagger$. This incident can also be seen in Figure 2.3 and will be discussed in more details in subsection 2.1.2. To stabilize the solution of inverse problems, the key idea is to introduce regularization of the optimization problem (2.5) [14, 76, 204].

2.1.2 Tikhonov regularization

To ensure that the inverse problem (2.2) is well-posed, we incorporate Tikhonov regularization into the minimization problem (2.5). The idea behind regularization is to control the data misfit and the norm of the approximate solution simultaneously. This means we can control the bias and variance trade-off. While on the one side we control how accurate the data should be fitted, on the other side we control regularity in our parameters. We define the Tikhonov regularization as follows:

Definition 2.1.4. *For given* ***regularization parameter*** $\kappa > 0$ *and compact, positive and convex nonnegative* ***regularization functional*** $\varphi : \mathcal{X} \to \mathbb{R}_+$ *we define the* ***Tikhonov loss function*** $T_\kappa : \mathcal{X} \to \mathbb{R}_+$ *through*

$$T_\kappa(\theta) = \frac{1}{2}\|H(\theta) - y\|_{\mathcal{Y}}^2 + \kappa\varphi(\theta). \tag{2.6}$$

Every minimizer $\theta_\kappa \in \mathcal{X}$ *of* (2.6) *(provided that it exists) is called* ***Tikhonov regularized solution*** *and we write*

$$\theta_\kappa \in \operatorname*{arg\,min}_{\theta \in \mathcal{D}(H)} T_\kappa(\theta). \tag{2.7}$$

Remark 2.1.5. *Suppose* $H(\cdot) = L\cdot$ *for some* $L \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ *and* $\varphi(\theta) = \|\theta - m\|_C^2$ *for given* $m \in \mathcal{X}$ *and given compact, positive and self-adjoint operator* $C \in \mathcal{L}(\mathcal{Y}, \mathcal{Y})$*, i.e. we consider the Tikhonov loss function*

$$T_\kappa(\theta) = \frac{1}{2}\|L\theta - y\|_{\mathcal{Y}}^2 + \frac{\kappa}{2}\|\theta - m\|_C^2. \tag{2.8}$$

For fixed regularization parameter we compute the first and second order derivatives w.r.t. θ *of the Tikhonov loss function*

$$\nabla_\theta T_\kappa(\theta) = L^*(L\theta - y) + \kappa C^{-1}(\theta - m),$$
$$\nabla_\theta^2 T_\kappa(\theta) = L^*L + \kappa C^{-1}.$$

Here, L^* *denotes the adjoint operator of* L*. Since* C *is a positive definite operator, the Tikhonov regularization acts as shifting the eigenvalues of the Hessian information of the Tikhonov loss function away from* 0*. In particular, in the linear setting this result leads to a strongly convex Tikhonov loss function.*

In the following, we will formulate the existence result for the Tikhonov loss function (2.8) with linear forward operators, where we assume for simplicity $m = 0$. In the linear setting it is possible to compute the Tikhonov regularized solution explicitly.

Theorem 2.1.6 ([118, Theorem 2.5], [76, Theorem 5.1])**.** *Let* $\mathcal{X}$ *and* $\mathcal{Y}$ *be separable Hilbert spaces,* $H(\cdot) = L\cdot$ *for some* $L \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ *and assume that* L *is a compact operator with singular system* (σ_n, v_n, u_n)*. Then the Tikhonov regularized solution exists, is unique, and is given through*

$$\theta_\kappa = (L^*L + \kappa C)^{-1}L^*y.$$

In the following, we will present the generalized existence result to nonlinear forward models. While in the linear setting the Tikhonov loss function was a quadratic one, in general this is not the case for the nonlinear setting. In particular, it is no longer clear whether a unique minimizer exists and how to compute a minimizer if it exists.
However, the existence of Tikhonov regularized solutions for general convex nonnegative regularization functional φ and sufficiently smooth forward map H can be verified. As it is not possible anymore to compute the Tikhonov regularized solution explicitly, there are several assumptions on the Banach spaces $\mathcal{X}$, $\mathcal{Y}$ and the forward model H, including its domain and its derivative, necessary. We summarize these necessary assumptions:

Assumption 2.1.7. *Let* $\mathcal{X}$ *and* $\mathcal{Y}$ *be infinite dimensional reflexive Banach spaces,* $H : \mathcal{D}(H) \to \mathcal{Y}$ *be a nonlinear map with* $\mathcal{D}(H) \subset \mathcal{X}$ *closed and convex and* $\varphi(\cdot) = \frac{1}{q}\|\cdot\|^q$ *for* $1 \le q < \infty$*. Assume that:*

- *H is weak-to-weak sequentially continuous, i.e. for $x_n \rightharpoonup x_0$ in $\mathcal{X}$, with $x_n \in \mathcal{D}(H)$, $n \in \mathbb{N}$ and $x_0 \in \mathcal{D}(H)$ it holds true that $H(x_n) \rightharpoonup H(x_0)$ in $\mathcal{Y}$. Here, "$\rightharpoonup$" denotes weak convergence.*

- *There exists a minimizing norm solution $\theta^\diamond$ of the operator equation $H(\theta) = y$, which means that*

$$H(\theta^\diamond) = y, \quad \text{with} \quad \|\theta^\diamond\|^q = \inf\{\|\theta\|^q \mid \theta \in \mathcal{D}(H),\ H(\theta) = y\}. \tag{2.9}$$

- *There exists a bounded linear operator $H' \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ such that for the one-side derivative at $\theta^\diamond$ and for every $\theta \in \mathcal{D}(H)$ it holds true that*

$$\lim_{t \to 0} \frac{H(\theta^\diamond + t(\theta^\diamond - \theta)) - H(\theta^\diamond)}{t} = H'(\theta^\diamond)(\theta - \theta^\diamond).$$

We note that the above assumptions and the corresponding existence result can be generalized to further regularization functions φ. Howerver, for simplicity we will focus on the above defined regularization and refer to [204] for more details on the generalization. Under Assumption 2.1.7 a Tikhonov regularized solution, i.e. a minimizer of (2.6), exists.

Theorem 2.1.8 ([204, Proposition 4.1])**.** *Let Assumption 2.1.7 hold. Then for all $\kappa > 0$ and $y \in \mathcal{Y}$ there exists a Tikhonov regularized solution θ_κ minimizing* (2.6)*.*

So far, we have stated the existence result for the Tikhonov regularized solution of the inverse problem. To ensure well-posedness, we state the following stability result w.r.t. the data.

Theorem 2.1.9 ([204, Proposition 4.2])**.** *For all $\kappa > 0$ the minimizers of* (2.6) *are stable w.r.t. the data y. This is, for every sequence $(y_n)_{n\in\mathbb{N}}$ in $\mathcal{Y}$ converging to y, i.e. $\lim_{n\to\infty} \|y_n - y\| = 0$, every sequence $(\theta_\kappa^n)_{n\in\mathbb{N}}$ of minimizers to the corresponding Tikhonov lossfunction*

$$T_\kappa^n(\theta) = \frac{1}{2}\|H(\theta) - y_n\|^2 + \kappa\varphi(\theta)$$

has a subsequence $(\theta_\kappa^{n_k})_{k\in\mathbb{N}}$ which converges weakly in $\mathcal{X}$ and the weak limit $\tilde{\theta}$ of each such subsequent is a minimizer of (2.6)*. Further, it holds true that*

$$\lim_{k\to\infty} \varphi(\theta_\kappa^{n_k}) = \varphi(\tilde{\theta}).$$

Another property one is interested in, is what happens with the regularized solution if the assumed noise level tends to zero. In general, the regularization should be chosen such that the regularized solution in the limit of noise going to zero coincides with the noisefree solution, i.e. with the best-approximate solution to

$$H(\theta) = y^\dagger,$$

which is given by the minimizing norm solution $\theta^\diamond$ defined in (2.9).
Some examples of convergence results with respect to the noise level for Tikhonov regularization can be found in [78, 89, 103] and for general convex loss functionals [31, 90, 166, 187].

Remark 2.1.10. *While the regularization function φ is given, the regularization parameter κ is free to choose. For example in the linear setting, i.e. for the Tikhonov loss function* (2.8)*, the operator C and the shift m can be interpreted as prior belief about how the solution should look like and is assumed to be given. The choice of the regularization parameter is a central question in the literature of Tikhonov regularization [77, 126, 168]. In this part of this work we will focus on a priori choices, which are only depending on the noise level ε, and a posteriori choices, which depend on the noise level ε and the data y.*

We assume that the data y^ε is a perturbed image of the underlying true parameter $\theta^\dagger$, i.e. $y^\varepsilon = H(\theta^\dagger)+\xi^\varepsilon$, where we define $y^\dagger = H(\theta^\dagger)$. Further, we assume that we have an estimate of the noise level $\varepsilon > 0$, such that we can ensure

$$\|y^\varepsilon - y^\dagger\| \leq \varepsilon.$$

We cite the following convergence result under a priori choice of regularization parameter.

Theorem 2.1.11 ([204, Corollary 4.6])**.** *Let $(\varepsilon_n)_{n\in\mathbb{N}}$ be a sequence converging to zero and $y^{\varepsilon_n} \in \mathcal{R}(H)$ such that $\|y^{\varepsilon_n} - y^\dagger\| \leq \varepsilon_n$. If the regularization parameter $\kappa_n = \kappa(\varepsilon_n)$ is chosen depending on the noise level, such that*

$$\lim_{n\to\infty} \kappa(\varepsilon_n) = 0, \quad \lim_{n\to\infty} \frac{\varepsilon_n^2}{\kappa(\varepsilon_n)} = 0,$$

then there exists a subsequence such that

$$\lim_{k\to\infty} \theta_{\kappa(\varepsilon_{n_k})} = \theta^\diamond,$$

where $\theta^\diamond$ is some minimizing norm solution defined in (2.9)*.*

For the a posteriori choice of regularization parameter, we focus on the so called *Morozov discrepancy principle*, which has been originally introduced by Morozov [163] and was studied for example in [7, 26]. Alternative heuristic methods for a posteriori choices are for example the L-curve method [95] or the quasi-optimality condition [12].
The discrepancy principle suggests that we cannot expect a more accurate recovering of $\theta^\dagger$ than the measurements accuracy, since otherwise the solution would get be to the noise in our data. Let θ_κ be the Tikhonov regularized solution depending on $\kappa > 0$ and define

$$\psi(\kappa) : \mathbb{R}_+ \to \mathbb{R}_+, \quad \psi(\kappa) = \|H(\theta_\kappa) - y^\varepsilon\|.$$

The Morozov discrepancy principle states that the regularization parameter $\kappa_{discr} = \kappa(\varepsilon, y^\varepsilon)$ should be chosen such that

$$\tau_1\varepsilon \leq \psi(\kappa(\varepsilon, y^\varepsilon)) \leq \tau_2\varepsilon, \tag{2.10}$$

for some $1 \leq \tau_1 \leq \tau_2$. This means that the regularized solution should not try to fit the data more accurately than up to the specified noise level. For general nonlinear forward models it is not clear, whether κ_{discr} satisfying (2.10) exists or not. We refer to [7] for more details on necessary conditions for existence of κ_{discr}. However, for simplicity we assume that κ_{discr} satisfying (2.10) exists.

Theorem 2.1.12 ([7, Theorem 4.5]). *Assume that for all minimizing norm solutions $\theta^\diamond$ it holds true that*

$$\liminf_{t \searrow 0} \frac{\|(1-t)\theta^\diamond - y\|^2}{t} = 0.$$

Let $(\varepsilon_n)_{n\in\mathbb{N}}$ be a sequence converging to zero and $y^{\varepsilon_n} \in \mathcal{R}(H)$ such that $\|y^{\varepsilon_n} - y^\dagger\| \leq \varepsilon_n$. Further, assume that $\kappa^n_{discr} = \kappa(\varepsilon_n, y^{\varepsilon_n})$ satisfy (2.10). Then it holds true that

$$\lim_{n\to\infty} \kappa(\varepsilon_n, y^{\varepsilon_n}) = 0 \quad \textit{and} \quad \lim_{n\to\infty} \frac{\varepsilon_n^2}{\kappa(\varepsilon_n, y^{\varepsilon_n})} = 0.$$

Note that application of Theorem 2.1.11 implies existence of a subsequence converging to a minimizing norm solution $\theta^\diamond$.

Remark 2.1.13. *The discrepancy principle can also be used within a iterative solution procedure of the inverse problem for fixed regularization parameter κ. As for general nonlinear forward maps H it is not possible to compute the minimizer of T_κ explicitly, one often applies gradient based optimization methods in order to compute a minimizer of T_κ. Denoting θ^i the i-th iteration of the optimization method, the aim is to find a sequence decreasing the Tikhonov lossfunction, i.e.*

$$T_\kappa(\theta^i) \geq T_\kappa(\theta^{i+1}).$$

The discrepancy principle states that one cannot expect to fit the data more accurate than the noise level ε, which can then be used as a early stopping criterion. In particular, the optimization method will be stopped in case

$$\|H(\theta^i) - y^\varepsilon\| \leq \varepsilon.$$

Note that this stopping criterion does not change the choice of regularization parameter. However, it can help to prevent overfitting of the noise in the data.

Remark 2.1.14. *While in Assumption 2.1.7 regularization functions of the form $\varphi(\cdot) = \frac{1}{q}\|\cdot\|^q$, for $1 \leq q < \infty$ are covered, and hence, the stated convergence results hold, we will focus on the choice $\varphi(\cdot) = \frac{1}{2}\|\cdot\|^2$ and we will refer this choice to Tikhonov regularization in the rest of this work. In the particular, we consider Tikhonov loss function $T_\kappa : \mathcal{X} \to \mathbb{R}_+$ of the from*

$$T_\kappa(\theta) = \frac{1}{2}\|H(\theta)\|^2 + \frac{\kappa}{2}\|\theta - m\|^2_{C_0},$$

where $m \in \mathcal{X}$ and $C_0 \in \mathcal{L}(\mathcal{X}, \mathcal{X})$, compact, positive and self-adjoint are given. We will state the connection to the Bayesian approach of inverse problems and in particular, to the maximum a-posteriori estimate in Section 2.2.

Other common regularization methods We give a brief literature overview of alternative common regularization methods for inverse problems. In PDE based inverse problems, iterative regularization methods with the combination of early stopping are often applied [11, 77, 156]. Some examples of iterative methods are the Landweber and steepest descent method [93], regularized Newton methods [122] and iterated Tikhonov methods [198]. For a detailed overview we refer the interested reader to [125].
Parallel to the iterative methods in the imaging community, the focus went to regularization methods based on total variation (TV) [44, 194], which is the most common edge

perserving regularization method [43, 188]. The total generalized variation (TGV) [30] makes use of second and higher order information.
There are several other examples for regularization methods, such as truncated singular value decomposition [94], wavelet shrinkage [72], infimal convolutional type regularization [33, 32] or regularization in a reproducing kernel Hilbert space [164, 165]. Furthermore, the Bayesian approach to inverse problems has been invented which opens up the possibilities to analyze inverse problems from a statistical perspective, see Section 2.2 for more details.

2.1.3 Numerical examples

We will demonstrate in simple toy examples the effects of regularization and the importance of the chosen regularization parameter κ. While our first example is a linear PDE based example, the second example is to train a so called neural network, which will be used to approximate a function. In both examples we see that it is a challenging task to balance between fitting the data and regularization.

Example 2.1.15 (Linear example: Partial differential equation)**.** *We consider the following one-dimensional partial differential equation*

$$\begin{cases} -\frac{\mathrm{d}^2 p}{\mathrm{d}x^2}(x) + p(x) = \theta(x), & x \in D := (0, \pi) \\ p(x) = 0, & x \in \partial D, \end{cases} \tag{2.11}$$

and consider the problem of recovering the unknown function $\theta^\dagger$ from noisy observation

$$y = L\theta^\dagger + \xi^\dagger,$$

where $\xi^\dagger$ is a realization of observational noise and the forward operator is defined through

$$L = \mathcal{O} \circ \mathfrak{L}^{-1}, \quad \mathfrak{L} = -\frac{\mathrm{d}^2}{\mathrm{d}x^2} + \mathrm{id} \quad \text{on } \mathcal{D}(\mathfrak{L}) = H^2 \cap H_0^1.$$

Here the operator $\mathfrak{L}^{-1}$ is the solution operator of (2.11) *and $\mathcal{O}$ observes the dynamical system at $K = 2^4 - 1$ equispaced observation points. We solve the PDE* (2.11) *numerically on a uniform mesh with mesh size $h = 2^{-8}$ by a finite element method with continuous, piecewise linear ansatz functions. The reference solution $\theta^\dagger$ will be sampled from a Gaussian distribution*

$$\theta^\dagger \sim \mathcal{N}(0, C_0),$$

where we set $C_0 = 10 \cdot (-\frac{\mathrm{d}^2}{\mathrm{d}x^2})^{-1}$.
In Figure 2.4 we can see the Tikhonov regularized solutions for different choices of regularization parameter κ. While the estimate without regularization clearly overfits the data, the resulting estimate with Tikhonov regularization improves effectively. We can also see that the choice of regularization parameter κ is crucial, as for too big values, the truth cannot be recovered very well and for too small values we can see overfitting again. In Figure 2.5 we can see the corresponding data fitting of our estimates. While we fit the data exactly in the setting without regularization, we allow more tolerance for greater values of κ. If κ is chosen too small, the noise in the data will be fitted again.

Example 2.1.16 (Nonlinear example: Training a neural network)**.** *Our next model problem will be motivated from a machine learning application example. We consider a deep neural network (DNN) to approximate the function $f : [-1, 1] \to \mathbb{R}$ defined as*

$$f(x) = \sin(2\pi x), \quad x \in [0, 2\pi]$$

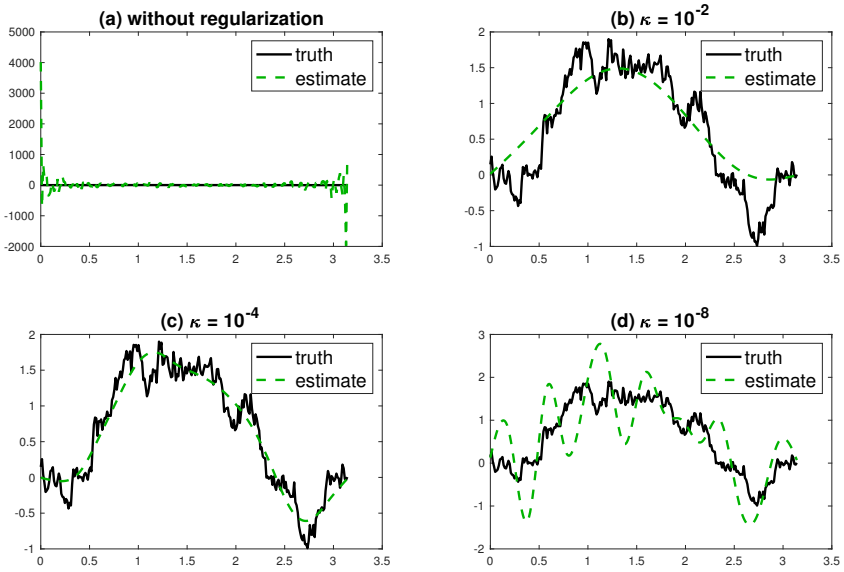

Figure 2.4: Resulting parameter estimates for different choices of regularization parameter: (a) without regularization, (b) with $\kappa = 10^{-2}$, (c) with $\kappa = 10^{-4}$ and (d) with $\kappa = 10^{-8}$

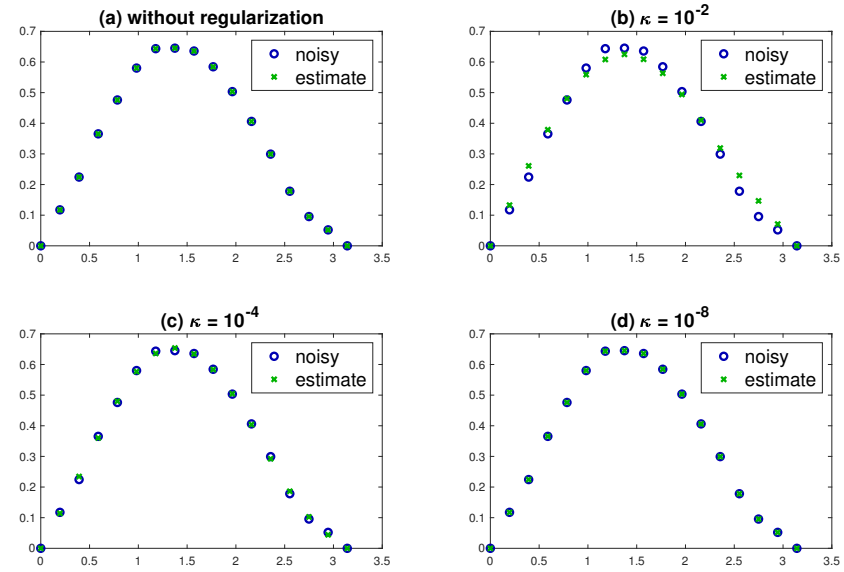

Figure 2.5: Resulting observational estimates for different choices of regularization parameter: (a) without regularization, (b) with $\kappa = 10^{-2}$, (c) with $\kappa = 10^{-4}$ and (d) with $\kappa = 10^{-8}$

given the training data set $\{x^k, f(x^k)\}_{k=1}^K$. *We will consider a DNN which is defined as a function* $f_\theta : \mathbb{R}^d \to \mathbb{R}$, *where the input is defined as* $x \in \mathbb{R}^d$ *and the parameters of the DNN are denoted by* θ. *For more details on the definition of a DNN see Section 7.2.1. By training the NN with respect to the training data* $\{x^k, y_k = u(x^k)\}_{k=1}^K$, *we aim to solve the Tikhonov regularized minimization problem*

$$\min_\theta \frac{1}{2} \sum_{k=1}^{K} \|f_\theta(x_k) - y_k\|^2 + \frac{\kappa}{2}\|\theta\|^2, \tag{2.12}$$

where our training data is perturbed by some noise, i.e.

$$y_k = f(x_k) + \xi_k.$$

Here ξ_k *denotes again the realization of the observational noise. We can translate the minimization problem into the inverse problem of finding* θ *such that*

$$\tilde{y} = f_\theta(\tilde{x}) + \xi, \tag{2.13}$$

given the data set $(\tilde{x}, \tilde{y})$ *with* $\tilde{x} := (x_1, \ldots, x_K)$ *and* $\tilde{y} := (y_1, \ldots, y_K)$, *where* $y_i := f_\theta(x_i)$. *In this setting the inverse problem aims to find the parameter* $\theta \in \mathbb{R}^{N_\theta}$ *such that* $f_\theta(\cdot)$ *fits the data best possible. We will define our NN to approximate the function* $f(x)$ *with* $L = 2$ *hidden layers with* $N_1 = 10,\ N_2 = 10$ *nodes and* $N_3 = 1$ *output node. We choose a logistic function*

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

as activation function. To train our NN we will use the `MATLAB` *function* `fminunc` *to minimize* (2.12) *with and without Tikhonov regularization. The chosen optimization method is a quasi-Newton method.*
Figure 2.6 shows different results of the function approximation through the NN for different choices of regularization parameters. Similarly, as in the example before, in absence of regularization we see the effect of fitting the noise in the data. In the case of too high values of regularization parameter the resulting function approximation acts close to a regression line.

2.2 Bayesian approach for inverse problems

In the next section, we will introduce an alternative viewpoint of inverse problems. Instead of using optimization methods to solve the inverse problem by minimizing deterministic (regularized) loss functions, we will introduce the Bayesian perspective of inverse problems. In the Bayesian approach, we are viewing the unknown parameter as random variable and using the incoming data to do statistical inference. The Bayesian approach can be interpreted as regularization, which makes the inverse problem well-posed. In particular, it is possible to connect the Bayesian approach via the maximum a posteriori estimate to the Tikhonov regularized solution introduced in the previous section for classical inverse problems. We will give more details on the connection in Section 2.2.3. For an example of the application of the Bayesian approach in inverse problems, presented in Section 2.1 and in particular, to inverse problems based on elliptic PDEs similar to the Example 2.1.15, we refer to [59, 224].

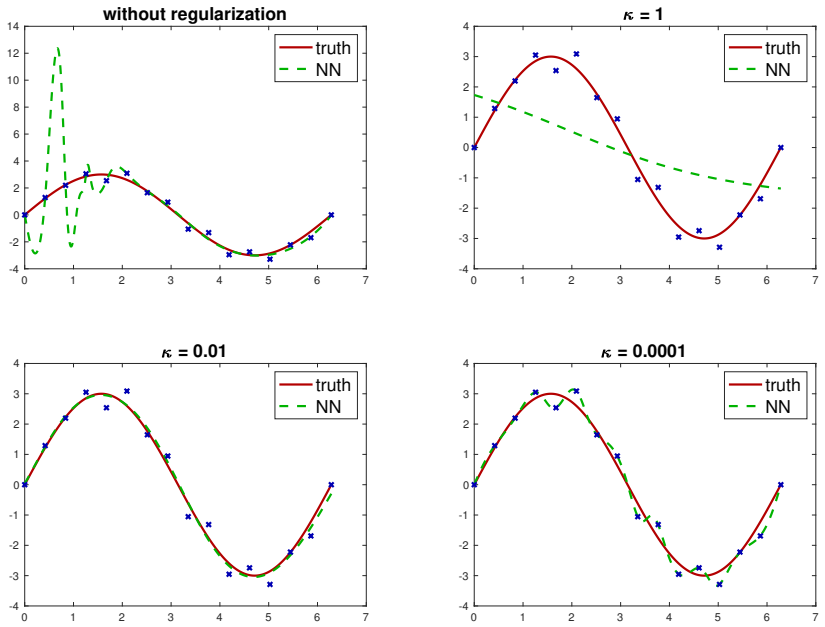

Figure 2.6: Resulting function approximations for different choices of regularization parameter: (a) without regularization, (b) with $\kappa = 1$, (c) with $\kappa = 10^{-2}$ and (d) with $\kappa = 10^{-4}$

The presented introduction is mainly based on the textbooks [118, 216] and the articles [60, 215]. While the finite dimensional setting has been discussed in [118, 215], the focus lies on the general setting of infinite dimensional Banach spaces in [60, 216].
For the rest of this work, we will denote our underlying probability space by $(\Omega, \mathcal{A}, \mathbb{P})$, where Ω is some nonempty set, $\mathcal{A}$ is a σ-algebra over Ω and $\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{A})$. If we talk about a random variable S which is valued on a separable Hilbert space $\mathcal{X}$, we consider a $\mathcal{A}$-$\mathcal{B}(\mathcal{X})$-measurable mapping $S : \Omega \to \mathcal{X}$, where $\mathcal{B}(\mathcal{X})$ denotes the Borel-σ-algebra over $\mathcal{X}$. Further, we sometimes write $S \sim \rho$ for some probability measure ρ on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, which corresponds to the distribution of S, i.e.

$$\mathbb{P}(S \in B) = \rho(B), \quad B \in \mathcal{B}(\mathcal{X}).$$

Usually, we denote random variables by capital letters, while we denote the realizations of these random variables through the corresponding lowercase letters.

2.2.1 Bayesian inverse problems: Well-posedness results

In the following, we will introduce the Bayesian setting for our inverse problem. Let $\mathcal{X}$ be a separable Hilbert space, denoting our (possibly infinite-dimensional) parameter space, and $\mathcal{Y} = \mathbb{R}^K$ denotes our finite-dimensional observation or data space respectively. We consider the stochastic model

$$Y = H(\Theta) + \Xi, \tag{2.14}$$

where

- Θ denotes the unknown parameter modelled as $\mathcal{X}$-valued random variable,
- $H : \mathcal{X} \to \mathbb{R}^K$ is our given forward model, which is some (possibly nonlinear) measurable mapping from the parameter space to the observation space,
- Ξ denotes the additive observational noise modelled as $\mathbb{R}^K$-valued random variable with mean zero and symmetric positive definite covariance matrix Γ,
- Y denotes the noisy observation modelled as $\mathbb{R}^K$-valued random variable.

Within this model we study the joint probability distribution of (Θ, Y), and in particular, we employ the Bayesian approach to inverse problems, where we condition the random variable Θ on the realization of observation $Y = y$. We take the following assumptions:

- We assume that the noise Ξ is Gaussian distributed, i.e. $\Xi \sim \mathcal{N}(0, \Gamma)$, where $\Gamma \in \mathbb{R}^{K\times K}$ is a symmetric positive definite covariance matrix .
- We assume that we have access to some prior information about our unknown parameter Θ, which is given through the marginal distribution of Θ, denoted by $\mathbb{Q}_0$.
- We assume that the random variable Θ and the noise Ξ are independent.

The **Bayesian inverse problem** is the task of conditioning the random variable Θ with **prior distribution** $\mathbb{Q}_0$ on the realization of the observation $Y = y$, i.e. to find

$$\mathbb{Q}_y^*(B) = \mathbb{P}(\Theta \in B \mid Y = y), \quad B \in \mathcal{B}(\mathcal{X}). \tag{2.15}$$

We call the conditioned distribution (2.15) **posterior distribution**.
In general, it is not clear whether the resulting posterior distribution has also Lebesgue density and can be written down in a simple way, since the parameter space is possibly infinite-dimensional.
However, application of Bayes' rule in the finite-dimensional setting suggests, that the Radon–Nikodým derivative of the posterior distribution $\mathbb{Q}_y^*$ with respect to the prior $\mathbb{Q}_0$, exists and is given through

$$\frac{\mathrm{d}\mathbb{Q}_y^*}{\mathrm{d}\mathbb{Q}_0}(\theta) \propto \exp\left(-\frac{1}{2}\|H(\theta) - y\|_\Gamma^2\right).$$

This can be seen in the following finite dimensional example.

Example 2.2.1. *In this example we assume a finite dimensional parameter space $\mathcal{X} = \mathbb{R}^I$ and derive the posterior distribution given by its conditional probability density function. We assume that $\mathbb{Q}_0$ has a Lebesgue probability density function, i.e. $\mathbb{Q}_0(\mathrm{d}\theta) = q_0(\theta)\,\mathrm{d}\theta$, where q_0 denotes the prior density, and we denote the joint probability density function of (Θ, Y) by $q_{(\Theta,Y)}$. Further, we assume that the marginal density function of Y is positive, i.e. $q_Y(y) = \int_{\mathbb{R}^I} q_{(\Theta,Y)}(\theta, y)\,\mathrm{d}\theta > $ for all $y \in \mathbb{R}^K$. Hence, by application of Bayes Theorem, Θ conditioned on $Y = y$ is distributed according to the probability density function defined by*

$$\rho^*(\theta) = \frac{q_{Y|\Theta=\theta}(y)}{q_Y(y)} q_0(\theta),$$

where $q_{Y|\Theta=\theta}$ denotes the conditional probability density function of Y conditioned on $\Theta = \theta$. As we have assumed Gaussian noise, $Y \mid \Theta = \theta$ has Lebesgue density defined by

$$q_{Y|\Theta=\theta}(y) = \frac{1}{\sqrt{\det(2\pi\Gamma)}} \exp(-\frac{1}{2}\|H(\theta) - y\|_\Gamma^2),$$

and the posterior density function ρ^* can be computed by

$$\rho^*(\theta) = \frac{1}{Z}\exp(-\frac{1}{2}\|H(\theta) - y\|_\Gamma^2)q_0(\theta),$$

where $Z = \int_{\mathbb{R}^I}\exp(-\frac{1}{2}\|H(\theta)-y\|_\Gamma^2)q_0(\theta)\,\mathrm{d}\theta$ denotes the normalization constant. Suppressing the normalization constant, we often write

$$\rho^*(\theta) \propto \exp(-\frac{1}{2}\|H(\theta) - y\|_\Gamma^2)q_0(\theta).$$

We note that we have introduced the notation $\|\cdot\|_\Gamma = \|\Gamma^{-1/2}\cdot\|$, where $\|\cdot\|$ denotes the norm of the underlying finite dimensional space, in this particular case $\mathbb{R}^K$. The following theorem extends the result of Example 2.2.1 to the infinite dimensional setting.

Theorem 2.2.2 ([216], Theorem 6.6)**.** *Let $H : \mathcal{X} \to \mathbb{R}^K$ be continuous, $\Xi \sim \mathcal{N}(0, \Gamma)$ for some symmetric positive definite covariance matrix $\Gamma \in \mathbb{R}^{K\times K}$ and $\Theta \sim \mathbb{Q}_0$. Then $\mathbb{Q}_y^*$ is absolutely continuous w.r.t. $\mathbb{Q}_0$, where the Radon–Nikodým derivative is given through*

$$\frac{\mathrm{d}\mathbb{Q}_y^*}{\mathrm{d}\mathbb{Q}_0}(\theta) \propto \exp\left(-\frac{1}{2}\|H(\theta) - y\|_\Gamma^2\right). \tag{2.16}$$

*We call this expression **likelihood** and define $\Phi(\theta, y) := \frac{1}{2}\|H(\theta) - y\|_\Gamma^2$ as the **potential**.*

The next question is if the posterior distribution given through (2.16) is well-defined and if it is stable w.r.t. changes in the data y. We will present existing well-posedness results for the Bayesian inverse problem from the literature. Well-posedness in the sense of consistency w.r.t. changes in the data has been discussed in [1, 52, 139], while consistency w.r.t. numerical approximation of the forward model can be found in [53].
We will state sufficient conditions on the forward map H and the prior distribution $\mathbb{Q}_0$ such that the posterior distribution exists and is stable w.r.t. the data y. We note that all of the presented well-posedness results also hold in the setting of general infinite-dimensional spaces, i.e. for some Banach space $\mathcal{X}$ as parameter space and some Banach space $\mathcal{Y}$ as observation space, see [60]. Nevertheless, we will stick to the case of finite-dimensional observation to stay consistent with the rest of this work.
The following assumptions on the potential Φ ensure that the posterior measure defined through (2.16) is well-defined

Assumption 2.2.3. *Let $\Phi : \mathcal{X} \times \mathbb{R}^K \to \mathbb{R}_+$ be continuous, $\mathcal{X}' \subset \mathcal{X}$ with $\mathbb{Q}_0(\mathcal{X}') = 1$ and assume the following:*

1. *There exists a monotonically nondeacreasing function $b_1 : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ separately in each argument such that for all $\theta \in \mathcal{X}$, $r > 0$ and $y \in B_{\mathbb{R}^K}(0, r)$*

$$\Phi(\theta, y) \geq -b_1(r, \|\theta\|_\mathcal{X}).$$

2. *There exists a monotonically nondeacreasing strictly positive function $b_2 : \mathbb{R}_+\times\mathbb{R}_+ \to \mathbb{R}_+$ such that for all $\theta \in \mathcal{X}'$, $r > 0$ and $y_1, y_2 \in B_{\mathbb{R}^K}(0, r)$*

$$|\Phi(\theta, y_1) - \Phi(\theta, y_2)| \leq b_2(r, \|\theta\|_\mathcal{X})\|y_1 - y_2\|_{\mathbb{R}^K}.$$

3. *For every $r > 0$ it holds true that*

$$\exp(b_1(r, \|\Theta\|_\mathcal{X})) \in L^1(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{Q}_0).$$

4. *We have* $\mathbb{Q}_0(B) > 0$ *for some bounded set* $B \in \mathcal{B}(\mathcal{X})$.

With these assumptions we are now able to ensure that the posterior distribution is well-defined.

Theorem 2.2.4 ([60], Theorem 4.3)**.** *Let Assumption 2.2.3 hold. Then, for every* $y \in \mathbb{R}^K$

$$Z(y) := \int_{\mathcal{X}} \exp(-\Phi(\theta, y))\, \mathbb{Q}_0(\mathrm{d}\theta) \tag{2.17}$$

is positive and finite and the posterior probability measure defined through

$$\frac{\mathrm{d}\mathbb{Q}_y^*}{\mathrm{d}\mathbb{Q}_0}(\theta) = \frac{1}{Z(y)} \exp\left(-\frac{1}{2}\|H(\theta) - y\|_\Gamma^2\right)$$

is well-defined.

In order to state stability results for the posterior distribution, we introduce the Hellinger distance, which is a distance between two probability measures.

Definition 2.2.5. *For two probability measures* $\mu_1, \mu_2 \in \mathcal{P}(\mathcal{X}, \mathcal{B}(\mathcal{X}))$*, where* $\mathcal{P}$ *denotes the space of probability measures on the measurable space* $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$*, we define the* ***Hellinger distance*** *through*

$$d_H(\mu_1, \mu_2) := \left(\int_{\mathcal{X}} \left(\sqrt{\frac{\mathrm{d}\mu_1}{\mathrm{d}\nu}(\theta)} - \sqrt{\frac{\mathrm{d}\mu_2}{\mathrm{d}\nu}(\theta)}\right)^2 \nu(\mathrm{d}\theta)\right)^{1/2},$$

where ν *is a dominating measure of* μ_1 *and* μ_2*, i.e.* $\mu_1 \ll \nu$ *and* $\mu_2 \ll \nu$.

The stability analysis is based on the Hellinger distance, as one can bound distances of moments w.r.t. two measures by the Hellinger distance, i.e. for $f \in L^2(\mathcal{X}, \mathcal{B}(X), \mu_1) \cap L^2(\mathcal{X}, \mathcal{B}(X), \mu_2)$ it holds

$$\|\mathbb{E}_{\mu_1}[f] - \mathbb{E}_{\mu_2}[f]\|_{\mathcal{X}} \leq c d_H(\mu_1, \mu_2),$$

for some constant $c > 0$. Further, one can bound the total variation distance, which is defined as

$$d_{TV}(\mu_1, \mu_2) = \sup_{B \in \mathcal{B}(\mathcal{X})} |\mu_1(B) - \mu_2(B)|,$$

by the Hellinger distance through

$$\frac{1}{2} d_H^2(\mu_1, \mu_2) \leq d_{TV}(\mu_1, \mu_2) \leq d_H(\mu_1, \mu_2).$$

For the well-posedness of the solution to the Bayesian inverse problem, it is left to consider stability w.r.t. changes in the initial conditions. First, we will formulate the stability w.r.t. the data y.

Theorem 2.2.6 ([60], Theorem 4.5)**.** *Let Assumption 2.2.3 hold and assume further, that*

$$\exp(b_1(r, \|\Theta\|_{\mathcal{X}}))(1 + b_2(r, \|\Theta\|)^2) \in L^1(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{Q}_0).$$

Then there exists $C = C(r) > 0$ *such that for all* $y_1, y_2 \in B_{\mathbb{R}^K}(0, r)$

$$d_H(\mathbb{Q}_{y_1}^*, \mathbb{Q}_{y_2}^*) \leq C\|y_1 - y_2\|_{\mathbb{R}^K}.$$

This result states that small changes in the data will result in small changes of the posterior measure, quantified in the Hellinger distance.
Another interesting property to consider is the stability w.r.t. changes in the forward model. Since in most of the examples for inverse problems a numerical approximation is necessary, it is crucial to ensure stability of the posterior in changes of say the accuracy of the forward model. We denote the potential resulting from the approximated forward model through Φ^N, where Φ^N converges in some sense to Φ with N going to infinity. We make the following assumptions.

Assumption 2.2.7. *Let $\Phi : \mathcal{X} \times \mathbb{R}^K \to \mathbb{R}_+$ and $\Phi_N : \mathcal{X} \times \mathbb{R}^K$, $N \in \mathbb{N}$, be continuous, $\mathcal{X}' \subset \mathcal{X}$ with $\mathbb{Q}_0(\mathcal{X}') = 1$ and assume the following:*

1. *There exists a monotonically nondecreasing function $b_1 : \mathbb{R}_+ \to \mathbb{R}_+$ separately in each argument, which is independent of $N \in \mathbb{N}$, such that for all $\theta \in \mathcal{X}'$*
$$\Phi(\theta, y) \geq -b_1(\|\theta\|_{\mathcal{X}}),$$
$$\Phi_N(\theta, y) \geq -b_1(\|\theta\|_{\mathcal{X}}).$$
2. *There exists a monotonically nondeacreasing strictly positive function $b_2 : \mathbb{R}_+ \to \mathbb{R}_+$, which is independent of $N \in \mathbb{N}$ and there exists a sequence $(\delta_N)_{N\in\mathbb{N}}$ converging to zero, such that for all $\theta \in \mathcal{X}'$*
$$|\Phi(\theta, y) - \Phi_N(\theta, y)| \leq b_2(\|\theta\|_{\mathcal{X}})\delta_N.$$
3. *It holds true that*
$$\exp(b_1(\|\Theta\|_{\mathcal{X}})) \in L^1(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{Q}_0).$$
4. *We have $\mathbb{Q}_0(B) > 0$ for some bounded set $B \in \mathcal{B}(\mathcal{X})$.*

The next Theorem establishes the the stability of the posterior with respect to approximation errors in the forward model, i.e. it proves that the approximated posterior distribution defined through

$$\frac{\mathrm{d}\mathbb{Q}_y^N}{\mathrm{d}\mathbb{Q}_0}(\theta) = \frac{1}{Z^N(y)} \exp\left(-\Phi_N(\theta, y)\right), \tag{2.18}$$
$$Z^N(y) := \int_{\mathcal{X}} \exp(-\Phi_N(\theta, y))\, \mathbb{Q}_0(\mathrm{d}\theta) \tag{2.19}$$

is well-defined under Assumption 2.2.7 and converges under further integrability conditions on b_1 and b_2 in Hellinger distance against the posterior distribution $\mathbb{Q}_y^*$.

Theorem 2.2.8 ([60], Theorem 4.8 and Theorem 4.9)**.** *Let Assumption 2.2.7 hold and fix $y \in \mathcal{Y}$ arbitrary. Then* (2.17) *and* (2.19) *are positive and finite, and the (approximate) posterior measures* (2.16) *and* (2.18) *respectively are well-defined. Furthermore, the lower bound on* (2.19) *is independent on N.*
Assume additionally that

$$\exp(b_1(\|\Theta\|_{\mathcal{X}}))\left(1 + b_2(\|\Theta\|)^2\right) \in L^1(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{Q}_0),$$

then there exists a constant $C > 0$ such that for sufficiently large N it holds true that

$$d_H(\mathbb{Q}_y^*, \mathbb{Q}_y^N) \leq C\delta_N.$$

So far, we have seen that the Bayesian inverse problem of finding (2.15) is a well-posed problem. In general it is not possible to write down the posterior distribution in a closed way or to construct straightforward samples from it. Therefore, there are several methods necessary to construct estimates for different kind of quantities of interest.

2.2.2 Gaussian measures

In the previous section, we have introduced the well-posed Bayesian inverse problem, which is the update of the prior distribution by inclusion of the observations, resulting in the conditioned probability distribution - the posterior distribution (2.16). So far, we have not introduced special cases of prior distributions, while we have just considered integrability conditions to ensure the well-posedness.
In this section, we will introduce Gaussian distribution, which will be used as prior. We define a Gaussian random variable on separable Hilbert spaces by requiring that each continuous linear functional of this random variable results in a one-dimensional Gaussian random variable.

Definition 2.2.9. *Let μ be a probability measure on a separable Hilbert space $\mathcal{X}$ and assume that the second moment of μ exists. We call μ* ***Gaussian measure*** *with mean $m \in \mathcal{X}$ and covariance operator $C \in \mathcal{L}(\mathcal{X}, \mathcal{X})$, which is positive, self-adjoint and trace class, if, for $\xi \sim \mu$ it holds true that*

$$\langle \xi, h\rangle_{\mathcal{X}} \sim \mathcal{N}\left(\langle m, h\rangle_{\mathcal{X}}, \langle h, Ch\rangle_{\mathcal{X}}\right)$$

for all $h \in \mathcal{X}$. We call the random variable ξ ***Gaussian distributed****.*

Similar to the finite-dimensional case, Gaussian measures are uniquely determined by its mean and covariance operator. Further, the following Theorem ensures existence of Gaussian measures.

Theorem 2.2.10 ([197])**.** *Let $\mathcal{X}$ be a separable Hilbert space and $C \in \mathcal{L}(\mathcal{X}, \mathcal{X})$ be a positive, self-adjoint and trace class operator. Then there exists a Gaussian measure on $\mathcal{X}$ with covariance operator C.*

One important property of Gaussian distributed random variables is that every affine linear transformation is again Gaussian distributed.

Proposition 2.2.11. *Let $\mathcal{X}_1$ and $\mathcal{X}_2$ be separable Hilbert spaces and $\xi \sim \mathcal{N}(m, C)$ be a Gaussian distributed random variable on $\mathcal{X}_1$. For $s \in \mathcal{X}_2$ and $L \in \mathcal{L}(\mathcal{X}_1, \mathcal{X}_2)$ it holds true that*

$$L\xi + s \sim \mathcal{N}(Lm + s, LCL^*).$$

In the following example, we study the Bayesian inverse problem under a Gaussian prior assumption.

Example 2.2.12 (Linear Gaussian case)**.** *We assume that $\mathcal{X} = \mathbb{R}^I$ and the prior $\mathbb{Q}_0$ is a Gaussian distribution $\mathcal{N}(m_0, C_0)$, and the forward map H is linear, defined through $H(\theta) = L\theta$, where $L \in \mathbb{R}^{K\times n}$. Hence, we can write the prior density through*

$$\mathrm{d}\mathbb{Q}_0(\theta) = \rho_0(\theta)\,\mathrm{d}\theta = \frac{1}{\sqrt{\det(2\pi C_0)}} \exp\left(-\frac{1}{2}\|m_0 - \theta\|^2_{C_0}\right)\,\mathrm{d}\theta,$$

and the likelihood through

$$\frac{\mathrm{d}\mathbb{Q}^*_y}{\mathrm{d}\mathbb{Q}_0}(\theta) = \frac{1}{\sqrt{\det(2\pi\Gamma)}} \exp\left(-\frac{1}{2}\|L\theta - y\|^2_{\Gamma}\right).$$

Using

$$d\mathbb{Q}_y^*(\theta) = \frac{1}{\sqrt{\det(2\pi\Gamma)}} \exp\left(-\frac{1}{2}\|L\theta - y\|_\Gamma^2\right) \frac{1}{\sqrt{\det(2\pi C_0)}} \exp\left(-\frac{1}{2}\|m_0 - \theta\|_{C_0}^2\right) \mathrm{d}\theta$$

gives with some further computations that the posterior distribution is again Gaussian with mean m_y and covariance matrix C_y, given through

$$m_y = m + C_0 L^\top (LC_0 L^\top + \Gamma)^{-1}(y - Lm), \quad C_y = C_0 - C_0 L^\top (LC_0 L^\top + \Gamma)^{-1} LC_0. \tag{2.20}$$

While the mean is getting shifted into direction of the data y, the uncertainty in the distribution, given through the covariances, is getting reduced. Note that this example can be extended to the infinite dimensional setting.

2.2.3 Maximum a-posteriori estimators and connection to regularization

In this section, we want to connect the introduced Bayesian approach to the classical regularization methods based on optimization. Therefore, we introduce the point estimator called *maximum a-posteriori estimator (MAP)*. We will stick to the finite-dimensional setting for simplicity, i.e. we will assume that our parameter space is given by $\mathcal{X} = \mathbb{R}^I$ and, similar as before, our observation space is $\mathcal{Y} = \mathbb{R}^K$.
Before introducing MAP estimators, we will briefly recall the idea of *maximum likelihood* estimator. In our setting the *likelihood function* for given data $y_1, \dots, y_M \in \mathbb{R}^K$ is defined through

$$L_M : \mathbb{R}^I \to \mathbb{R}_+, \quad \theta \mapsto L_M(\theta) = \prod_{i=1}^M q(y_i \mid \theta) \tag{2.21}$$

where in our setting we have denoted the likelihood for arbitrary data $y \in \mathbb{R}^K$ and fixed $\theta \in \mathbb{R}^I$ by

$$\begin{aligned} q(y \mid \theta) &= \frac{1}{C(y)} \exp(-\Phi(\theta, y)), \\ C(y) &= \int_{\mathbb{R}^I} \exp(-\Phi(\theta, y))\, \mathrm{d}\theta. \end{aligned} \tag{2.22}$$

Note that in general, one can take other density functions instead of $q(y \mid \theta)$ to define likelihood estimators. However, we will stick to this class of density functions defined through (2.22).
Before defining the MAP estimator, we define the maximum likelihood estimator arising in nonparametric statistics in order to view the MAP estimator as its modification in the Bayesian perspective.

Definition 2.2.13. *For given data $y_1, \dots, y_M \in \mathbb{R}^K$ the **maximum likelihood estimator** (MLE) is defined by the maximizer of the Likelihood function L_M defined in* (2.21) *provided it exists. We write for a maximum likelihood estimate*

$$\theta_{ML} \in \arg\max_{\theta \in \mathbb{R}^I} L_M(\theta).$$

In our context, we assume to have prior belief on the unknown parameter $\theta \in \mathbb{R}^I$, such that we can use this information to construct in some sense an update of the ML estimator. If

we assume that the prior distribution has Lebesgue density ρ_0, we can write the posterior distribution through Lebesgue density again, i.e.

$$\begin{aligned} \mathrm{d}\mathbb{Q}_y^*(\theta) &= \frac{1}{Z(y)} \exp(-\Phi(\theta, y)) \cdot \rho_0(\theta)\, \mathrm{d}\theta, \\ Z(y) &= \int_{\mathbb{R}^I} \exp(-\Phi(\theta, y)) \cdot \rho_0(\theta)\, \mathrm{d}\theta. \end{aligned}$$

In this context, we will denote the *posterior density* through

$$q_y^*(\theta) = \frac{1}{Z(y)} \exp(-\Phi(\theta, y)) \cdot \rho_0(\theta) \tag{2.23}$$

and define the MAP estimator in the following.

Definition 2.2.14. *Let $\mathcal{X} = \mathbb{R}^I$, $\mathcal{Y} = \mathbb{R}^K$ and consider a prior distribution with Lebesgue density ρ_0, i.e.*

$$\mathrm{d}\mathbb{Q}_0(\theta) = \rho_0(\theta)\, \mathrm{d}\theta.$$

If a maximizer of (2.23) *exists, we denote every maximizer of* (2.23) *by* ***maximum a-posteriori estimator*** *and we write*

$$\theta_{MAP} \in \arg\max_{\theta \in \mathbb{R}^I} q_y^*(\theta).$$

We can interpret the MAP estimator in some sense as regularized maximizer of the likelihood function. This interpretation will be more obvious if we consider a Gaussian prior and connect the MAP estimator to the Tikhonov regularized solution defined in (2.7).
We assume that our prior is Gaussian $\mathcal{N}(m_0, C_0)$ with mean $m_0 \in \mathbb{R}^I$ and symmetric positive definite covariance $C_0 \in \mathbb{R}^{I \times I}$. Hence, we can write the prior density explicitly through

$$\rho_0(\theta) = \frac{1}{\sqrt{\det(2\pi C_0)}} \exp(-\frac{1}{2}\|m_0 - \theta\|_{C_0}^2),$$

such that the MAP estimators are given by maximizing the posterior density

$$q_y^*(\theta) \propto \exp\left(-\frac{1}{2}\|H(\theta) - y\|_\Gamma^2 - \frac{1}{2}\|m_0 - \theta\|_{C_0}^2\right). \tag{2.24}$$

Taking the negative logarithm of (2.24) leads to the equivalent computation of the MAP estimator through the minimization problem

$$\arg\min_{\theta \in \mathbb{R}^I} \|H(\theta) - y\|_\Gamma^2 + \|m_0 - \theta\|_{C_0}^2. \tag{2.25}$$

By this representation of the MAP estimator, we can view the computation as the computation of the Tikhonov regularized solution similar as in (2.7).
We refer the interested reader to [118] in the finite dimensional setting, and for a generalization to the infinite dimensional setting of the MAP estimators and its consistency analysis, we refer to [61].

2.2.4 Karhunen–Loéve expansion: Alternative prior models

In this section, we give a brief introduction to the Karhunen–Loéve (KL) expansion based on [154]. The KL expansion can be interpreted as spectral decomposition of a random field corresponding to its covariance operator. Before introducing different classes of prior choices, we introduce the KL expansion for general $L^2(D)$ valued random fields for some domain $D \subset \mathbb{R}^d$.
When talking about a second-order random field, we mean a random field $\Theta = (\Theta(x))_{x\in D}$, which is a family of real value random variables on the underlying probability space, such that $\Theta(x) \in L^2(\Omega)$ for all $x \in D$. We define the mean function of Θ by

$$m : D \to \mathbb{R}, \quad m(x) = \mathbb{E}[\Theta(x)],$$

and the covariance function by

$$c : D \times C \to \mathbb{R}, \quad c(x, y) = \mathbb{E}[(\Theta(x) - m(x))(\Theta(y) - m(y))].$$

We note that for $\mathcal{X} = L^2(D)$, we can interpret the Gaussian measure $\mathcal{N}(m, C)$ defined in Definition 2.2.9 as random field with mean function m and covariance function c representing the covariance operator C, i.e. for each $h \in L^2(D)$ we can write

$$Ch(x) = \int_D c(x, y)h(y)\, dy.$$

Karhunen–Loéve expansion for random fields

The KL expansion can be interpreted as Fourier series representation of a stochastic process or more general of a random field respectively. We will formulate the Karhunen–Loéve Theorem for $L^2(D)$-valued random field for some $D \subset \mathbb{R}^d$. The KL expansion provides a representation of the random field based on an orthonormal basis of its covariance operator $C \in \mathcal{L}(L^2(D), L^2(D))$, defined by

$$Ch(x) = \int_D c(x, y)h(y)\, \mathrm{d}y,$$

with covariance function $c : D \times D \to \mathbb{R}$.

Theorem 2.2.15 (Karhunen–Loéve Theorem for L^2 random field, [154, Theorem 7.52])**.** *Let $D \subset \mathbb{R}^d$ and consider a second-order random field $\Theta = (\Theta(x))_{x\in D}$. Then for all $\omega \in \Omega$ we can write*

$$\Theta(x, \omega) = m(x) + \sum_{i=1}^{\infty} \sqrt{\nu_i}\varphi_i(x)\zeta_i(\omega), \tag{2.26}$$

where the sum converges in L^2-sense,

$$\zeta_i(\omega) := \frac{1}{\nu_i}\langle\Theta(\cdot, \omega), \varphi_i(\cdot)\rangle_{L^2(D)}$$

and $(\nu_i, \varphi_i)_{i\in\mathbb{N}}$ are the eigenvalues and eigenfunctions of the covariance operator C, i.e.

$$C\varphi_i(x) = \int_D c(x, y)\varphi_i(y)\, \mathrm{d}y = \nu_i\varphi_i(x)$$

with $\nu_1 \geq \nu_2 \geq \cdots \geq 0$.

We note that the random variables (ζ_i) have mean zero and are pairwise uncorrelated. Furthermore, if Θ is a Gaussian random field, the random variables (ζ_i) are i.i.d. $\mathcal{N}(0,1)$ distributed.
The truncation of this KL expansion representation gives the possibility to approximate random fields, i.e. we can approximate the second-order random field Θ by

$$\Theta_M(x,\omega) = m(x) + \sum_{i=1}^{M} \sqrt{\nu_i}\varphi_i(x)\zeta_i(\omega).$$

This results in a random field Θ_M with mean function m and covariance function

$$c_M = \sum_{i=1}^{M} \nu_i \varphi_i(x)\varphi_i(y).$$

This means, in order to use the KL expansion to approximate the random field Θ one has to solve the eigenvalue problem $C\varphi_i = \nu\varphi_i$, for the covariance operator C of Θ.
The truncated random field Θ_M approximates the original random field Θ in the following sense.

Theorem 2.2.16 (uniform convergence of the KL expansion, [154, Theorem 7.53])**.** *Let $D \subset \mathbb{R}^d$ be closed and bounded and consider a second-order random field $\Theta = (\Theta(x))_{x\in D}$ with continuous covariance function c. Then the eigenfunctions $\varphi_i(\cdot)$ of the covariance operator C are continuous and the series expansion of C converges uniformly. In particular,*

$$\sup_{x,y\in D} |c(x,y) - c_M(x,y)| \leq \sup_{x\in D} \sum_{i=M+1}^{\infty} \nu_i\varphi_i(x) \to 0,$$

as $M \to \infty$. Furthermore, the truncated random field Θ_J converges to Θ in the sense that

$$\lim_{M\to\infty} \sup_{x\in D} \mathbb{E}[|\Theta_M(x) - \Theta(x)|^2] = 0.$$

The KL expansion gives now the motivation of choosing a prior model through a series representation

$$\Theta(x,\omega) = m(x) + \sum_{i=1}^{\infty} \sqrt{\nu_i}\varphi_i(x)\zeta_i(\omega) \tag{2.27}$$

by different choices of the system $(\nu_i, \varphi_i)_{i\in\mathbb{N}}$ and different distributions of $(\zeta_i)_{i\in\mathbb{N}}$. We consider models, such that $\mathbb{E}[\zeta_1] = 0$ and the function m acts as mean function of the random field.

Uniform priors

Motivated by the series representation (2.26), the first class of priors we consider is the class of uniform priors. The idea is to start with a series representation of the underlying functional, and randomizing this series representation by introducing random coefficients. We refer to [117] for more details on random functions generated by this way. It is worthwhile to refer to [180] for the construction of probability measures of infinite sequences of i.i.d. random variables, which is fundamental in order to define these kind of prior models. For the introduction of uniform priors, we consider an underlying Banach space $\mathcal{X} = L^\infty(D)$, $D \subset \mathbb{R}^d$ bounded and open with Lipschitz boundary, and let $\sqrt{\nu} = (\sqrt{\nu_i})_{i\in\mathbb{N}}$ be

in ℓ^1. The series of random variables is specified as uniformly distributed, i.e. we consider $\zeta = (\zeta_i)_{i\in\mathbb{N}}$ as i.i.d. sequence of random variables with $\zeta_1 \sim \mathcal{U}([-1,1])$. Further, we assume the following boundary conditions on the mean function m and the sequence $\sqrt{\nu}$:

$$\operatorname*{ess\,inf}_{x\in D} \; m(x) \geq m_l, \quad \operatorname*{ess\,sup}_{x\in D} \; m(x) \leq m_u, \quad \|\sqrt{\nu}\|_{\ell^1} = \frac{\delta}{1+\delta} m_l,$$

where $m_l \leq m_u$ and $\delta > 0$ are constants. Let $\varphi_i : D \to \mathbb{R}$ be real-valued normalized functions, i.e $\varphi_i \in \mathcal{X}$ with $\|\varphi_i\|_{L^\infty} = 1$, and we define $\mathcal{X}'$ as the closure of the linear span of $(\varphi_i)_{i\in\mathbb{N}}$ and m w.r.t. the norm $\|\cdot\|_{L^\infty}$. Hence, we have found a separable Banach space $(\mathcal{X}', \|\cdot\|_{L^\infty})$ and we can state the following result regarding the truncated series representation

$$\Theta_M(x,\omega) = m(x) + \sum_{i=1}^{M} \sqrt{\nu_i}\varphi_i(x)\zeta_i(\omega). \tag{2.28}$$

Theorem 2.2.17 ([60, Theorem 2.1])**.** *For $\mathbb{P}$-almost all $\omega \in \Omega$, the sequence $(\Theta_M(\cdot,\omega))_{M\in\mathbb{N}}$ given by (2.28) is a Cauchy sequence in $\mathcal{X}'$ and the limiting function $\Theta(\cdot,\omega)$ given by (2.27) satisfies for almost every $x \in D$*

$$\frac{1}{1+\delta} m_l \leq \Theta(x,\omega) \leq m_u + \frac{\delta}{1+\delta}.$$

Furthermore, under certain regularity assumptions on ν, φ and m, one can ensure Hölder continuity of the random field Θ.

Theorem 2.2.18 ([60, Theorem 2.3])**.** *Assume that there are constants C, $a > 0$, $\alpha \in (0,1]$, such that for all $i \in \mathbb{N}$*

$$|\varphi_i(x) - \varphi_i(y)| \leq C i^a |x-y|^\alpha, \quad |m(x) - m(y)| \leq C|x-y|^\alpha, \quad x,y \in D.$$

Further, assume that $\sum_{i=1}^{\infty} |\nu_i| i^{a\gamma} < \infty$ for some $\gamma \in (0,2)$. Then for $\mathbb{P}$-almost all $\omega \in \Omega$ it holds true that $\Theta(\cdot,\omega) \in C^{0,\beta}(D)$ for all $\beta < \frac{\alpha\gamma}{2}$.

Random fields constructed in this way have been considered in the context of forward uncertainty quantification [49, 50] where the effect of randomizing the input data on the solution of the model equation has been discussed. For Bayesian inverse problems, uniform prior models have been considered in [206].

Gaussian random field priors

Returning to the setting of $L^2(D)$-valued random fields, we consider the special case of Gaussian random fields and its KL expansion. In many PDE based inverse problems, the unknown parameter is modelled as a Gaussian random field whose covariance function is described by the Whittle–Matérn class. Following [192, 151], we firstly introduce the Whittle–Matérn class covariance and secondly connect this class to a stochastic partial differential equation resulting in a fast approximation approach.
The covariance function of the Whittle–Matérn class is defined by

$$c(x,y) = \frac{\sigma^2}{\Gamma(\alpha)2^{\alpha-1}} \left(\frac{\|x-y\|}{\ell}\right)^\alpha K_\alpha \left(\frac{\|x-y\|}{\ell}\right), \quad x,y \in D = \mathbb{R}^d$$

where K_α denotes the modified Bessel function of second kind and order $\alpha > 0$ and $\Gamma(\cdot)$ denotes the Gamma function. The length parameter $\ell > 0$ is a scaling parameter, σ^2 is

the marginal variance of the random field and the parameter α controlling the smoothness of the random field. The stochastic partial differential equation (SPDE) approach states that samples of the Whittle–Matérn class may be generated by solving the SPDE

$$(\mathrm{Id} - \ell^2 \Delta)^{(\alpha+d/2)/2} \Theta = \ell^{d/2} \sqrt{\gamma} \mathbb{W},$$

where $\mathbb{W}$ is Gaussian white noise on $D = \mathbb{R}^d$, Δ denotes a Laplacian operator on D and the constant γ is defined as

$$\gamma = \sigma^2 \frac{2^d \pi^{d/2} \Gamma(\alpha + d/2)}{\Gamma(\alpha)}.$$

We can formally describe the covariance operator of the Gaussian random field by

$$C_{\mathrm{WM}} = \ell^d \gamma (\mathrm{Id} - \ell^2 \Delta)^{-(\alpha+d/2)}.$$

In our application for our presented numerical examples, we often choose $\sigma > 0$ such that $\ell^d \gamma = \beta > 0$ and we set $\alpha = \alpha + d/2$ as well as $\tau = \ell^{-1}$, such that we can specify the covariance operator of the Whittle–Matérn class as

$$C_{\mathrm{WM}} = \beta \cdot (\tau^2 \cdot \mathrm{Id} - \Delta)^{-\alpha},$$

with parameters $\beta > 0$, $\tau > 0$ and $\alpha > d/2$. In order to generate samples of the Whittle–Matérn class, one can apply the KL expansion using the eigensystem of C_{WM}. For more details of the relation between α and the smoothness of the corresponding random field we refer to [60]. Furthermore, we note that similar to the ideas of uniform prior modelling in (2.28), one can introduce various prior models by choosing different series representations. For example besov priors for Bayesian inverse problems have been considered in [58, 146].

2.2.5 Basic sampling methods for Bayesian inverse problems

In the following section, we will discuss the Markov chain Monte Carlo method, which is a sampling based methods to solve the Bayesian inverse problem. The method is based on either generating samples of the posterior distribution (2.16) or approximating quantities of interest, which are given as expected values w.r.t. the posterior distribution, i.e. for $Z \sim \mathbb{Q}_y^*$:

$$Q_{Int} = \mathbb{E}[F(Z)] = \int_{\mathcal{X}} F(\theta) \, \mathbb{Q}_y^*(\mathrm{d}\theta). \tag{2.29}$$

Assuming to have a finite-dimensional parameter space $\mathcal{X} = \mathbb{R}^I$ and to have access to a Lebesgue density ρ^*, one could use general methods of numerical integration to approximate the integral

$$\int_{\mathbb{R}^I} F(\theta) \rho^*(\theta) \, \mathrm{d}\theta.$$

However, in many cases the parameter space is of high dimension or we even have no access to a Lebesgue density, such that basic numerical integration methods will fail. The basic idea for the following part are Monte Carlo methods, a method which approximates integrals independent of dimension of the state space, but in our case has to assume to have access to samples of the posterior distribution. To generate samples of the posterior distribution or to approximate the quantity of interest (2.29) directly, we will introduce Markov chain Monte Carlo methods.

Markov chain Monte Carlo methods

The basic idea of Markov chain Monte Carlo (MCMC) methods is to construct a Markov chain with stationary distribution given by posterior $\mathbb{Q}^*$ in order to construct a sample of the posterior distribution and hence, construct a Monte Carlo estimate of the quantity of interest (2.29). The idea behind this can be seen from the following: Suppose that the current state of a Markov Chain $(X_k)_{k\in\mathbb{N}}$ with transition kernel P is initialized according to $X_0 \sim \mathbb{Q}^*$, where P is invariant w.r.t. $\mathbb{Q}^*$, it follows that

$$X_1 \sim \mathbb{Q}^* \circ P = \mathbb{Q}^*, \ldots, X_k \sim \mathbb{Q}^* \circ P^k = \mathbb{Q}^*,$$

where we have defined $\mu \circ P(\mathrm{d}x) = \int_{\mathcal{X}} P(z, \mathrm{d}x)\mu(\mathrm{d}z)$ and $P^k = P \circ P \circ \cdots \circ P$. Further, suppose that P^k converges weakly to $\mathbb{Q}^*$ as k tends to infinity, then for large enough $k \in \mathbb{N}$, P^k approximates $\mathbb{Q}^*$ and heuristically it holds

$$X_{k+1} \sim P^{k+1} = P^k \circ P \approx \mathbb{Q}^* \circ P = \mathbb{Q}^*.$$

Hence, one expects the possibility of generating a sample of the target distribution $\mathbb{Q}^*$ by running the Markov chain, which can then be used to approximate the quantity of interest through a Monte Carlo estimate

$$\frac{1}{N}\sum_{i=1}^{N} F(X_{k+i}) \approx \mathbb{E}[F(Z)].$$

While in the case of an i.i.d. sample of $\mathbb{Q}^*$, this approximation can easily be verified through the strong law of large numbers, one has to investigate more work in the setting of Markov chains, as the resulting sample is correlated. To make this idea rigorous, we follow the derivation in [220] which is mainly based on the textbook [169]. We also refer to [221] for Metropolis-Hastings methods on general state spaces. We will focus on time-homogenous Markov chains with transition kernel P, which means for the Markov chain $(X_k)_{k\in\mathbb{N}}$ it holds

$$\mathbb{P}(X_{k+1} \in B \mid X_1 = x_1, \ldots, X_k = x_k) = \mathbb{P}(X_{k+1} \in B \mid X_k = x_k) = P(x_k, B),$$

for all $B \in \mathcal{B}(\mathcal{X})$, $x_1, \ldots, x_k \in \mathcal{X}$ and $k \in \mathbb{N}$. We start this discussion by a formal definition of invariance of a Markov chain.

Definition 2.2.19. *Let P be the transition kernel of a Markov chain. We call a Markov chain **invariant** w.r.t. a probability measure μ, if $\mu \circ P = P$.*

In order to formulate the convergence theorems for MCMC methods, we need to define the properties of irreducibility and periodicity.

Definition 2.2.20. *A Markov chain and its corresponding transition kernel P are called **irreducible** w.r.t. a σ-finite measure φ on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ with $\varphi(\mathcal{X}) > 0$, if for each $x \in \mathcal{X}$ and $A \in \mathcal{B}(\mathcal{X})$ with $\varphi(A) > 0$, there exists an $k \in \mathbb{N}$ depending on x and A, such that $P^k(x, A) > 0$. We will write φ-irreducible Markov chain and φ-irreducible transition kernel respectively.*

Definition 2.2.21. *Let μ be a probability measure on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. A μ-irreducible transition kernel P is called **periodic**, if there exists $d \geq 2$ and $A_0, \ldots, A_{d-1} \in \mathcal{B}(\mathcal{X})$ nonempty and disjoint such that for all $i \in \{0, \ldots, d-1\}$ and all $x \in A_i$*

$$P(x, A_j) = 1 \quad \text{for } j = i + 1 \bmod d.$$

*Otherwise, we call the transition kernel **aperiodic**.*

We note that for a φ-irreducible Markov chain, every set $A \in \mathcal{B}(\mathcal{X})$ with $\varphi(A) > 0$ can be reached with positive probability after finitely many steps. Further, periodicity of a Markov chain means that the chain stays in a loop with probability one.
The next crucial property to be considered is the notion of recurrence, which we define in the following.

Definition 2.2.22. *Let μ be a probability measure on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. A μ-irreducible Markov chain $(X_k)_{k\in\mathbb{N}}$ with invariant distribution is called* ***recurrent****, if*

$$\mathbb{P}_x(\limsup_{k\to\infty} \{X_k \in A\}) > 0, \quad \text{for all } x \in \mathcal{X},$$
$$\mathbb{P}_x(\limsup_{k\to\infty} \{X_k \in A\}) = 1, \quad \text{for } \mu\text{-almost all } x \in \mathcal{X},$$

for all $A \in \mathcal{B}(\mathcal{X})$ with $\mu(A) > 0$. Here, $\mathbb{P}_x$ denotes the measure under which the Markov chain starts a.s. in $x \in \mathcal{X}$. Further, we call the Markov chain ***Harris recurrent****, if $\mathbb{P}_x(\limsup_{k\to\infty} \{X_k \in A\}) = 1$ for all $x \in \mathcal{X}$.*

Recurrence, in particular Harris recurrence, of a Markov chain means that the process returns infinitely many times to each $A \in \mathcal{X}$ with $\mu(A) > 0$ independent of the starting position. We call a Markov chain **ergodic** if it is Harris recurrent and aperiodic.
The following theorem states, that irreducibility and invariance of the Markov chain imply recurrence and if the Markov chain is also aperiodic, the Markov chain converges to its invariant distribution.

Theorem 2.2.23 ([220, Theorem 1])**.** *Let P be a μ-irreducible transition kernel and assume that $\mu \circ P = P$. Then P is recurrent and μ is the unique invariant distribution of P. Further, if P is aperiodic, then for μ-almost all $x \in \mathcal{X}$*

$$\lim_{k\to\infty} d_{TV}(P^k(x, \cdot), \mu) = 0.$$

If P is Harris recurrent, then the convergence holds for all $x \in \mathcal{X}$.

We are now ready to formulate the law of large numbers for Markov chains, which result from the ergodic theorem for Markov chains.

Theorem 2.2.24 ([220, Theorem 3])**.** *Let $(X_k)_{k\in\mathbb{N}}$ be a ergodic Markov chain with invariant distribution μ and assume $F \in L^1(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mu)$, real valued. Then for any initial distribution of X_0 it holds true that*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} F(X_i) \overset{a.s.}{=} \mathbb{E}_\mu[F] = \int_{\mathcal{X}} F(x)\, \mu(dx),$$

where the convergence holds almost surely.

In order to apply Theorem 2.2.24 to generate an estimate for (2.29) one has to construct a Markov chain with transition kernel P satisfying

- P is invariant w.r.t. the posterior $\mathbb{Q}^*$,
- P is a $\mathbb{Q}^*$-irreducible transition kernel,

- P is Harris recurrent.

To ensure the invariance of P, it is often easier to verify the so called detailed balance condition, which is also referred to the reversibility of the Markov chain.

Definition 2.2.25. *The transition kernel P of a Markov chain satisfies the detailed balance condition w.r.t. μ if*

$$\mu(\mathrm{d}x)P(x,\mathrm{d}z) = \mu(\mathrm{d}z)P(z,\mathrm{d}x).$$

The Markov chain will then be called reversible.

Lemma 2.2.26 ([60, Lemma 5.2])**.** *A Markov chain which is reversible w.r.t. μ is also invariant with respect to μ.*

Metropolis–Hastings algorithm

The aim is to construct a transition kernel P which is invariant w.r.t. the posterior $\mathbb{Q}^*$, $\mathbb{Q}^*$-irreducible and Harris recurrent. The Metropolis-Hastings algorithm has been introduced by Nicholas Metropolis et al. in 1953 [161] and has been generalized by Wilfred Keith Hastings in 1970 [97]. The algorithm is based on an acceptance-rejection method. Based on a proposal kernel, the next state will be proposed and then accepted with probability specified by a likelihood ratio depending on the target distribution. For simplicity, we assume that the parameter space is finite dimensional $\mathcal{X} = \mathbb{R}^I$ and the posterior distribution $\mathbb{Q}^*$ has Lebesgue density ρ^*. We formulate the Metropolis-Hastings method in Algorithm 1.

Remark 2.2.27. *Algorithm 1 first proposes the next state regarding the proposal transition kernel, and then accepts the proposed state with probability $\alpha(\theta_k, \theta'_{k+1})$. Heuristically, α can be interpreted in the following way: Assume that the proposal kernel is symmetric, i.e. $q(x,x') = q(x',x)$. Then $\alpha(x,x') = \min(1, \frac{\rho^*(x')}{\rho^*(x)})$ measures the ratio between the posterior density evaluated at x and x'. Firstly, if the density value at the proposed state x' is greater than at the current state, the proposed state will always be accepted. Secondly, if the density at the proposed state is smaller than at the current state, we accept with the ratio $\rho^*(x')/\rho^*(x)$, such that we reject states which are less likely w.r.t. posterior distribution.*

The following proposition verifies the application of Algorithm 1.

Proposition 2.2.28 ([118, Proposition 3.12], [220, Corollary 2])**.** *Let $\rho^* : \mathbb{R}^I \to [0,\infty)$ be a probability density function and Q be the proposal Markov transition kernel with density q. Let $(\Theta_k)_{k\in\mathbb{N}}$ be the Markov chain generated by Algorithm 1 with corresponding transition kernel P. If Q is an aperiodic transition kernel, then the transition kernel P is also aperiodic. Further, if Q is $\mathbb{Q}^*$-irreducible and $\alpha(\theta,\theta') > 0$ for $\mathbb{Q}^*$-almost all $\theta,\theta' \in \mathbb{R}^I$, then the transition kernel P is also $\mathbb{Q}^*$-irreducible and P is Harris recurrent.*

With these properties we can apply Theorem 2.2.23 and Theorem 2.2.24 in order to verify the application of a sample generated by Algorithm 1 as MC estimate for the a quanitity of interest (2.29).
In the following, we will give an example of a proposal transition kernel Q which can be used in Algorithm 1.

Algorithm 1: Metropolis–Hastings algorithm

Input:

- target distribution $\mathbb{Q}^*(\mathrm{d}\theta) \propto \rho^* \,\mathrm{d}\theta$,
- proposal Markov transition kernel $Q : \mathbb{R}^I \times \mathcal{B}(\mathbb{R}^I) \to [0,1]$ with density $q : \mathbb{R}^I \times \mathbb{R}^I \to [0,\infty)$, i.e.

$$Q(x,A) = \int_A q(x,x')\,\mathrm{d}x', \quad A \in \mathcal{B}(\mathcal{X}),$$

- acceptance probability

$$\alpha(x,x') := \begin{cases} \min\left(1, \frac{\rho^*(x')q(x',x)}{\rho^*(x)q(x,x')}\right) & , \ \rho(x)q(x,x') > 0 \\ 1 & , \ \text{else} \end{cases},$$

- initial probability distribution ν_0 on $\mathbb{R}^I$.

Output: Markov chain $(\Theta_k)_{k\in\{1,\dots,N\}}$
Draw $X_1 \sim \nu_0$ and set $\Theta_1 = X_1$.
for *k=1,. . . , N* **do**

- Given the current state $\Theta_k = \theta_k$, propose θ'_{k+1} according to $Q(\theta_k, \cdot)$.
- Draw $U \sim \mathcal{U}([0,1])$ and set

$$\Theta_{k+1} = \begin{cases} \theta'_{k+1} & , \ \text{if } U \le \alpha(\theta_k, \theta'_{k+1}), \\ \theta_k & , \ \text{else.} \end{cases}$$

Example 2.2.29. *We choose a Gaussian random walk kernel, which can be described by $Q : \mathbb{R}^I \times \mathcal{B}(\mathbb{R}^I)$ with $Q_s(\theta,\cdot) = \mathcal{N}(\theta, s^2 C_0)$, where $C_0 \in \mathbb{R}^{I\times I}$ could be the covariance matrix of the prior distribution or some alternative symmetric positive-definite matrix. We note that $s > 0$ is a tuning parameter, which can be optimized. The resulting Metropolis–Hastings algorithm finds the next state by jumping randomly around the current state until one of the proposed states is accepted. A rule of thumb states that the step size parameter s should be chosen such that the acception rate*

$$\int_{\mathbb{R}^I} \alpha(\theta,\theta') Q_s(\theta, \mathrm{d}\theta') \mathbb{Q}^*(\mathrm{d}\theta) \approx \frac{1}{N} \sum_{k=1}^{N} \alpha(\theta_k, \theta'_{k+1})$$

is approximately 25%, see [191].

Remark 2.2.30. *We note that the Metropolis–Hastings algorithm can be generalized to the infinite dimensional setting, where one has to replace the Lebesgue densities in the definition of α by Radon–Nikodým derivatives. In this setting, we can choose a Gaussian random walk kernel, defined by the preconditioned Crank-Nicolson (pCN) proposal, see for example [54]. For Gaussian prior assumption $\mathbb{Q}_0 = \mathcal{N}(0, C_0)$, the proposal can be*

described by $Q : \mathcal{X} \times \mathcal{B}(\mathcal{X})$ *with*

$$Q_s(\theta, \cdot) = \mathcal{N}(\sqrt{1 - s^2}\theta, s^2 C_0),$$

where s *is again a step size parameter, which can be optimized. The acceptance probability* α *is then defined by*

$$\alpha(\theta, \theta') = \min\left(1, \exp(\Phi(\theta) - \Phi(\theta'))\right),$$

where Φ *is the Radon-Nikodým derivative of the posterior distribution* $\mathbb{Q}^*$ *w.r.t. prior* $\mathbb{Q}_0$.

Example 2.2.31. *To give more details on the procedure of the Metropolis–Hastings algorithm, we consider the following* 2*-dimensional toy example. We define the forward map* $H : \mathbb{R}^2 \to \mathbb{R}$ *with* $H(\theta_1, \theta_2) = \log(\theta_1 - (\theta_2^2 + 1))$. *Our prior is assumed to be Gaussian* $\mathbb{Q}_0 = \mathcal{N}(0, C_0)$, *with* $C_0 = 5 \cdot \mathrm{Id} \in \mathbb{R}^{2\times 2}$ *and the noise is assumed to be* $\Sigma \sim \mathcal{N}(0, \Gamma)$ *with* $\Gamma = 0.1 \in \mathbb{R}_+$. *We generate an underlying truth* $\theta^\dagger \sim \mathbb{Q}_0$ *and the corresponding data* $y = H(\theta^\dagger) + \xi^\dagger$, *where* $\xi^\dagger \sim \mathcal{N}(0, \Gamma)$ *is a realization of the noise. We run the MCMC method with pCN proposal, where we have chosen* $s = 0.8$ *in order to ensure an acceptance rate approximately around* 25%. *In Figure 2.7 we can see the first* 10 *and the first* 100 *iterations of Algorithm 1, where the green points correspond to the realized points, while the red points denote the rejected proposed points. In Figure 2.8 we show a sample of size* 1000 *generated through Algorithm 1, which has been run for* $N = 10^5$ *iterations, and we have collected the last* 1000 *iterations to generate the sample.*

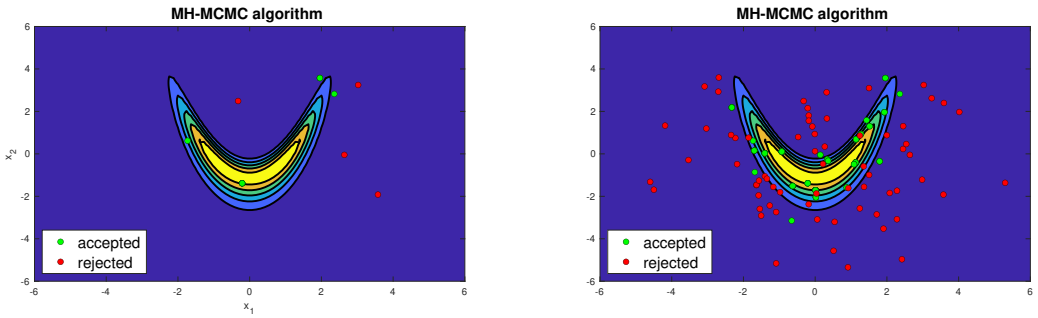

Figure 2.7: Resulting path of the Metropolis–Hastings MCMC method. The first 10 (left) and 100 (right) iterations are shown.

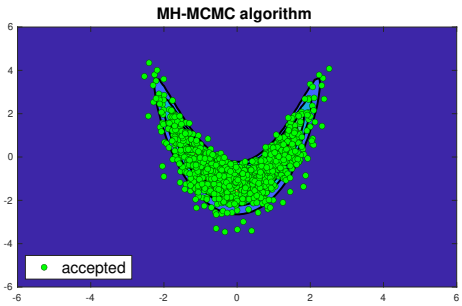

Figure 2.8: Resulting sample of the Metropolis–Hastings MCMC method. The last 1000 iterations are shown.

Beside the Random Walk Metropolis Hastings method, general sampling methods for Bayesian inverse problems and specifically MCMC methods have been studied extensively. Similar to the ideas of the Metropolis Hastings methods, the Gibbs sampling method [86, 85] or Hybrid monte carlo methods [73, 21], which are based on Hamiltonian systems, can be used to construct a Markov chain whose stationary distribution is given by the posterior. There exists also various MCMC methods based on multi-level Monte Carlo methods [87, 48]. Furthermore, MCMC methods can be constructed by applying of the stationary distribution resulting from the Langevin equation [171, 177, 92], which will be discussed in Chapter 6 in more details.
An alternative approach to construct MC estimates is to use particle based sampling methods. The main idea goes back to sequential Monte–Carlo (SMC) methods [162], which have been successfully applied to PDE based inverse problems [129] (Navier–Stokes) and to elliptic inverse problems [22]. The basic idea is to sequentially evolve a particle system from the prior to the posterior distribution by tempering the posterior density function. Here, in each step it is possible to implement resampling through importance sampling [9, 152], which itself can be applied to Bayesian inverse problems [2]. There is much ongoing research in the area of particle based sampling methods for Bayesian inverse problems, such as ensemble Kalman inversion [112] - Chapter 3, ensemble Kalman sampling [83] - Chapter 6 and discretization of the Fokker-Planck equation [175] - Chapter 6. Furthermore, in [153] the promising Stein variational gradient descent method has been introduced. This method is based on minimizing a kernalized Stein discrepancy, which quantifies the observance of the Stein identity. The accuracy of these methods are described in the Kullback–Leibler divergence or Wasserstein distance. Similar ideas are considered in [160] in order to sample via measure transport maps, where the aim is to minimize the Kullback–Leibler divergence.

2.3 Introduction to data assimilation

In the following section, we introduce the field of data assimilation, see for example [186, 216, 147, 185]. The data assimilation problem deals with the combination of two information sources.

1. **Dynamical system:** We consider a time-dependent physical system described through our mathematical model.

2. **Observations:** We assume to have access to a time series of observations of the underlying dynamical system. These observation are usually modelled to be perturbed by some noise.

The field of data assimilation uses both the dynamical system as well as the observations in order to build sequentially more accurate estimates of the state of the dynamical system or to construct predictions of the future state. Typically, the tools are based on Bayesian models. The methodology to combine models with data goes back to Kalman [119, 120]. The research field of data assimilation has a wide range of application such as in weather forecasting [121], oil reservoir simulation [173], turbulence modeling [157] and geophysical sciences [37].
We introduce the mathematical model of the filtering problem and present various versions of the filtering tool called Kalman filter (KF). Throughout this section we will consider a

finite-dimensional state space $\mathbb{R}^I$. We refer the interested reader to [131] for extension to Hilbert space valued state spaces.

2.3.1 The mathematical model

In this work, we will focus on the discrete time formulation of the data assimilation problem. We assume that the state of the dynamical system is given through the Markov chain $Z = (Z_j)_{j\in\mathbb{N}}$ defined by

$$Z_{j+1} = H_j(Z_j) + \xi_j, \quad j \in \mathbb{N}, \tag{2.30}$$

with $Z_0 \sim \pi_0$ for some probability distribution π_0 on $\mathbb{R}^I$, where the dynamics are described through the possibly nonlinear mappings $H_j : \mathbb{R}^I \to \mathbb{R}^I$. We assume that our dynamic is perturbed by noise given through $\xi = (\xi_j)_{j\in\mathbb{N}}$, which is an i.i.d. sequence with $\xi_j \sim \mathcal{N}(0, \Sigma)$ for symmetric and positive definite $\Sigma \in \mathbb{R}^{I\times I}$, where ξ_0 and Z_0 are stochastically independent. We denote the current state Z_j at each time as **signal** and refer to equation (2.30) as the **stochastic dynamical system**.
Further, we assume to have access to a given time series of **data**, or also called **observations**, $Y = (Y_j)_{j\in\mathbb{N}}$ which are described through the observation model (2.31) below. The observations are playing the role of reducing the uncertainty in the stochastic dynamical system.

$$Y_{j+1} = h_{j+1}(Z_{j+1}) + \eta_{j+1}, \quad j \in \mathbb{N}, \tag{2.31}$$

where $h_j : \mathbb{R}^I \to \mathbb{R}^K$ denotes the observation map and $\eta = (\eta_j)_{j\in\mathbb{N}}$ denotes noise, which is given through an i.i.d. sequence with $\eta_1 \sim \mathcal{N}(0, \Gamma)$ for symmetric and positive definite $\Gamma \in \mathbb{R}^{K\times K}$.
We call the task of determining information about the signal Z, given the observation y, **data assimilation problem**.

Remark 2.3.1. *We can easily connect the data assimilation problem to the inverse problem introduced in section 2.1, by considering the* ***deterministic dynamical system***

$$Z_{j+1} = H_j(Z_j), \quad j \in \mathbb{N},$$

with $Z_0 \sim \pi_0$. The task of recovering Z_0 given the whole time series of observation y defined through (2.31) *can be formulated by the model* (2.2)*.*

Since the task is to condition the information about the signal on the incoming data, similar as in the inverse problem context, the basic idea is to use Baysian methods. The idea is to update the information about the state, which is given through a probability distribution, by conditioning to the incoming data sequentially. In this way, one could interpret the data assimilation problem as a time depending sequential inverse problem.

2.3.2 The prediction, filtering and smoothing problem

The prior information about the signal is given through our stochastic dynamical system and the assumptions on the noise model ξ. Using the Chapman–Kolmogorov equation for the Markov chain resulting from the system (2.30) we obtain the marginal distribution density $\pi_j(z)$ of Z_j given through

$$\pi_{Z_{j+1}}(\mathrm{d}z') = \mathbb{P}(Z_{j+1} \in \mathrm{d}z') = \int_{\mathbb{R}^I} \pi_j(\mathrm{d}z' \mid z)\pi_{Z_j}(\mathrm{d}z),$$

where we assume that the distribution of Z_0 has Lebesgue density π_0 and the transition probability densities can be computed through

$$\pi_j(\mathrm{d}z' \mid z) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\left(-\frac{1}{2}\|z' - H_j(z)\|_\Sigma^2\right) \mathrm{d}z'.$$

Similar to the Bayesian approach for inverse problems, we can compute the probability density function (PDF) of the observation Y_j conditioned on the state $Z_j = z$ which is given through

$$\pi_{Y_j}(y \mid z) = \frac{1}{\sqrt{\det(2\pi\Gamma)}} \exp\left(-\frac{1}{2}\|y - h(z)\|_\Gamma^2\right).$$

We define the different problems of prediction, filtering and smoothing as the task of conditioning the PDF of the signal Z_j to the given time series of observations. We assume that a realization $y^{1:N_{obs}} = (y_1, \ldots, y_{N_{obs}})$ of observations $Y^{1:N_{obs}} = (Y_1, \ldots, Y_{N_{obs}})$, $N_{obs} \in \mathbb{N}$, is given. The data assimilation problem is to compute the conditional distribution of the state Z_j given the observations $y^{1:N_{obs}}$, i.e.

$$\pi_{Z_j|y^{1:N_{obs}}}(\mathrm{d}z) = \mathbb{P}(Z_j \in \mathrm{d}z \mid Y^{1:N_{obs}} = y^{1:N_{obs}}). \tag{2.32}$$

We call the task of computing (2.32)

1. **prediction problem** if $j > N_{obs}$,
2. **filtering problem** if $j = N_{obs}$,
3. and **smoothing problem** if $j < N_{obs}$.

We denote (2.32) as **prediction, filtering and smoothing distribution** respectively.
In this work, the focus is on the filtering problem and we introduce the Kalman filter, a method which aims to compute the filtering distribution sequentially for $N_{obs} = 1, 2, \ldots, J$, $J \geq 0$. We note that filtering and smoothing problems are related by the end time of any specified time interval, where, conditioned on the same data, both solutions have to coincide [147, Theorem 2.12].
Filtering methods are typically split into two parts. Given the filtering distribution $\pi_{Z_j|y^{1:j}}(\mathrm{d}z)$, the first part is the **prediction step**, which computes the distribution of the next state through

$$\pi_{Z_{j+1}|y^{1:j}}(\mathrm{d}z) = \mathbb{P}(Z_{j+1} \in \mathrm{d}z \mid Y^{1:j} = y^{1:j}) = \int_{\mathbb{R}^I} \pi_{j+1}(\mathrm{d}z \mid z')\pi_{Z_j|y^{1:j}}(\mathrm{d}z'). \tag{2.33}$$

The second step is the **Bayesian assimilation step**, which uses the distribution resulting from the prediction step as prior distribution to compute the filtering distribution $\pi_{Z_{j+1}|y^{1:j+1}}(\mathrm{d}z)$ with the help of the Bayesian theorem

$$\pi_{Z_{j+1}|y^{1:j+1}}(\mathrm{d}z) = \frac{\pi_{Y_{j+1}}(y_{j+1} \mid z)\pi_{Z_{j+1}|y^{1:j}}(\mathrm{d}z)}{\int_{\mathbb{R}^I} \pi_{Y_{j+1}}(y_{j+1} \mid z)\pi_{Z_{j+1}|y^{1:j}}(\mathrm{d}z)}. \tag{2.34}$$

Similar as in the case of Bayesian inverse problems, methods to compute, approximate or sample from the filtering distribution are necessary.
Given a stochastic dynamical system (2.30) and observations (2.31), we call the recursively computation of the prediction (2.33) followed by the Baysian update step (2.34) **sequential data assimilation**.

2.3.3 Linear Kalman filter

We introduce the Kalman filter, which solves the sequential data assimilation problem in a linear and Gaussian setting exactly. We assume that the signal is described through

$$Z_{j+1} = F_j Z_j + \xi_j, \quad j \in \mathbb{N}, \tag{2.35}$$

for linear forward maps $F_j \in \mathcal{L}(\mathbb{R}^I, \mathbb{R}^I)$ and the observations are given by

$$Y_{j+1} = A_{j+1} Z_{j+1} + \eta_{j+1}, \quad j \in \mathbb{N} \tag{2.36}$$

with linear observation operator $A_j \in \mathcal{L}(\mathbb{R}^I, \mathbb{R}^K)$. For simplicity, we assume that $F = F_j$ and $A = A_j$ are constant for all $j \in \mathbb{N}$. We remind the reader that the noise is described through i.i.d. sequences $(\xi_j)_{j\in\mathbb{N}}$ and $(\eta_j)_{j\in\mathbb{N}}$ with $\xi_j \sim \mathcal{N}(0, \Sigma)$ and $\eta \sim \mathcal{N}(0, \Gamma)$. Since our dynamical model is linear and the noise is assumed to be Gaussian, our filtering distribution $\pi_{Z_j|y^{1:j}}$ remains Gaussian if the initial probability distribution is also Gaussian. We assume that $Z_0 \sim \mathcal{N}(m_0, C_0)$ and compute the filtering distribution recursively.
Through the linear and Gaussian assumptions, we can describe the filtering distribution through

$$\pi_{Z_j|y^{1:j}} = \mathcal{N}(m_j, C_j),$$

and derive the computation of the mean m_j and the covariance C_j.
Given the mean m_j and covariance C_j of iteration j, the first part is to compute the update of the mean and the covariance based on the prediction step (2.33), which is just using the dynamics (2.35). The prediction step is given by

$$\widehat{m}_{j+1} = F m_j, \quad \widehat{C}_{j+1} = F C_j F^\top + \Sigma, \tag{2.37}$$

where we have used that $Z_j \sim \mathcal{N}(m_j, C_j)$ and ξ_j are independent.
The next step is the Bayesian update step (2.34) and corresponds to the Bayesian formulation for inverse problems in the linear Gaussian setting (2.20) where we set our prior distribution to $\mathbb{Q}_0 = \mathcal{N}(\widehat{m}_{j+1}, \widehat{C}_{j+1})$. We update the mean and covariance by application of (2.20)

$$\begin{aligned} m_{j+1} &= \widehat{m}_{j+1} + \widehat{C}_{j+1} A^\top (A \widehat{C}_{j+1} A^\top + \Gamma)^{-1} (y_{j+1} - A \widehat{m}_{j+1}), \\ C_{j+1} &= \widehat{C}_{j+1} - \widehat{C}_{j+1} A^\top (A \widehat{C}_{j+1} A^\top + \Gamma)^{-1} A \widehat{C}_{j+1}. \end{aligned}$$

We define the so called Kalman gain matrix by

$$K_j = \widehat{C}_j A^\top (A \widehat{C}_j A^\top + \Gamma)^{-1}, \tag{2.38}$$

and write the Bayesian update step as

$$\begin{aligned} m_{j+1} &= \widehat{m}_{j+1} + K_{j+1}(y_{j+1} - A \widehat{m}_{j+1}), \\ C_{j+1} &= \widehat{C}_{j+1} - K_{j+1} A \widehat{C}_{j+1}. \end{aligned} \tag{2.39}$$

We will refer m_j as the **state estimator** of the current iteration j. In the following

Algorithm, we summarize the **linear Kalman filter**.

Algorithm 2: Linear Kalman filter

Input: initial mean m_0 and covariance C_0, observations $(y_1, \dots, y_N)$
Output: $(\pi_j)_{j=1,\dots,N}$
for $j = 0, \dots, N-1$ **do**
Prediction step:
Map the mean and covariance through the dynamical system

$$\widehat{m}_{j+1} = Fm_j, \quad \widehat{C}_{j+1} = FC_jF^\top + \Sigma$$

Bayesian assimilation step:
Update the mean and the covariance by

$$\begin{aligned} m_{j+1} &= \widehat{m}_{j+1} + K_{j+1}(y_{j+1} - A\widehat{m}_{j+1}), \\ C_{j+1} &= \widehat{C}_{j+1} - K_{j+1}A\widehat{C}_{j+1}, \\ K_{j+1} &= \widehat{C}_{j+1}A^\top(A\widehat{C}_{j+1}A^\top + \Gamma)^{-1}. \end{aligned}$$

Filtering distribution: $\pi_{j+1} = \mathcal{N}(m_{j+1}, C_{j+1})$.

Further, we note that given $C_j > 0$, we can ensure that also $\widehat{C}_{j+1} > 0$ and hence, $C_{j+1} > 0$. Suppose $C_j > 0$, then

$$\widehat{C}_{j+1} = FC_jF^\top + \Sigma > 0,$$

since $\Sigma > 0$. To ensure that C_{j+1} also stays positive definite, we apply the Woodbury-matrix identy, see for example [147, Lemma 4.4],

$$C_{j+1} = \widehat{C}_{j+1} - \widehat{C}_{j+1}A^\top(A\widehat{C}_{j+1}A^\top + \Gamma)^{-1}A\widehat{C}_{j+1} = (\widehat{C}_{j+1}^{-1} + A^\top\Gamma^{-1}A)^{-1},$$

which implies the update

$$C_{j+1}^{-1} = \widehat{C}_{j+1}^{-1} + A^\top\Gamma^{-1}A.$$

Hence, C_{j+1}^{-1} exists and is positive definite, and we imply that also C_{j+1} stays positive definite.
To give an overview of the described KF method, we present the following Figure 2.9.

2.3.4 Variational perspective of the Kalman filter

Before introducing the ensemble Kalman filter, we will first generalize the ideas of the linear Kalman filter to non-Gaussian models. While in the linear Kalman filter setting the filtering distribution stays always Gaussian, we now introduce another perspective of the Kalman filter, where we view the update steps in the sense of solving minimization problems. We write the update of the mean from equation (2.39) through the minimization problem

$$m_{j+1} = \operatorname*{arg\,min}_{v \in \mathbb{R}^I} \mathcal{I}_{j+1}(v), \tag{2.40}$$

with

$$\mathcal{I}_j(v) := \frac{1}{2}\|y_j - Av\|_\Gamma^2 + \frac{1}{2}\|v - \widehat{m}_j\|_{\widehat{C}_j}^2, \tag{2.41}$$

where $\widehat{m}_j$ and $\widehat{C}_j$ are defined through the prediction step (2.37). The structure of this minimization problem can be seen through the Bayesian perspective of the update with

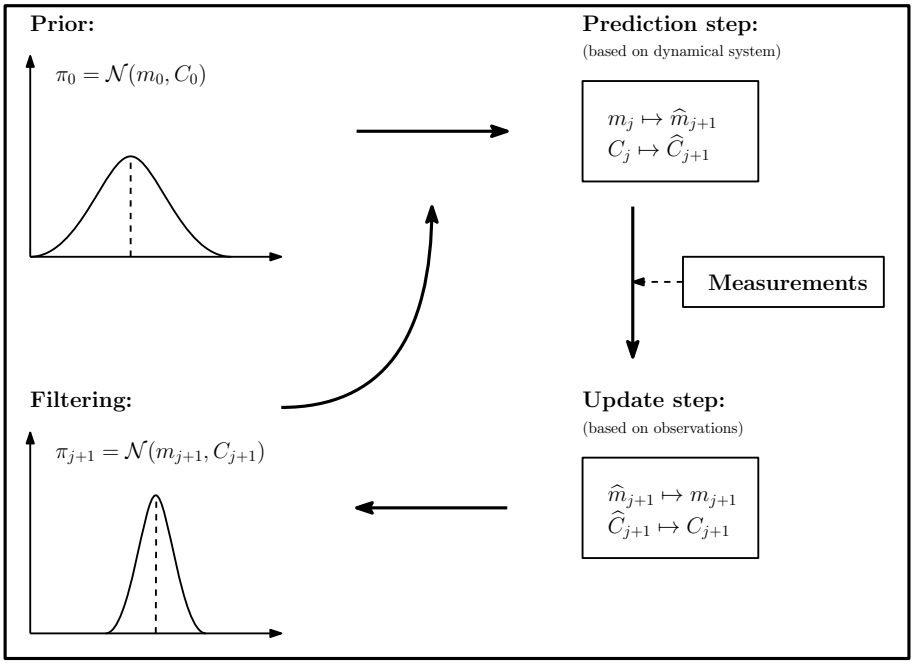

Figure 2.9: Summary of the linear Kalman filter method.

prior $\mathbb{Q}_0 = \mathcal{N}(\widehat{m}_j, \widehat{C}_j)$ and the connection to the MAP estimator introduced in section 2.2.3.
While the prediction step considers only the dynamical system to give a state estimation, the Bayesian update connects the predicted state to the incoming data. This can also be seen in the definition of the loss functional (2.41), where the first term measures the data misfit, and the second part regularizes this data fit to the predicted state, which is interpreted as prior information.
The following Theorem states that the Bayesian update step on m_{j+1} in (2.39) can be derived through solving the minimization problem (2.40).

Theorem 2.3.2. *Let $A \in \mathbb{R}^{K\times n}$ be of rank K and $\widehat{m}_{j+1}$, $\widehat{C}_{j+1}$ be given through* (2.37)*. The solution of the minimization problem* (2.40) *is given through*

$$v^* = \widehat{m}_{j+1} + K_{j+1}(y_{j+1} - A\widehat{m}_{j+1}).$$

Proof. The gradient of $\mathcal{I}$ w.r.t. v is given by

$$\nabla_v \mathcal{I}(v) = -A^\top \Gamma^{-1}(y_{j+1} - Av) + \widehat{C}_{j+1}^{-1}(v - \widehat{m}_{j+1})$$

and the Hessian

$$\nabla_v^2 \mathcal{I}(v) = A^\top \Gamma^{-1} A + \widehat{C}_{j+1}^{-1} > 0,$$

which ensures that $\mathcal{I}$ is strictly convex, and hence it is sufficient to solve

$$\nabla_v \mathcal{I}(v) = 0 \tag{2.42}$$

in order to find the global minimizer of $\mathcal{I}$. Note that $\widehat{C}_{j+1}$ is positive definite and invertible. Solving (2.42) leads to the global minimizer

$$v^* = (\widehat{C}_{j+1}^{-1} + A^\top \Gamma^{-1} A)^{-1}(A^\top y_{j+1} + \widehat{C}_{j+1}^{-1}\widehat{m}_{j+1})$$

$$= (I - K_{j+1}A)\widehat{C}_{j+1}(A^\top \Gamma^{-1} y_{j+1} + \widehat{C}_{j+1}^{-1}\widehat{m}_{j+1}),$$

where we have used the Woodbury-matrix identity to compute

$$(\widehat{C}_{j+1}^{-1} + A^\top \Gamma^{-1} A)^{-1} = \widehat{C}_{j+1} - \widehat{C}_{j+1}A^\top(\Gamma + A\widehat{C}_{j+1}A^\top)^{-1}A\widehat{C}_{j+1} = (I - K_{j+1}A)\widehat{C}_{j+1}.$$

Note that the last equality states

$$\Big((I - K_{j+1}A)\widehat{C}_{j+1}\Big) \cdot \Big(\widehat{C}_{j+1}^{-1} + A^\top \Gamma^{-1} A\Big) = I,$$

which can be reordered such that we have

$$(I - K_{j+1}A)\widehat{C}_{j+1}A^\top \Gamma^{-1} A = K_{j+1}A,$$

and since A is of rank K, we can write

$$\begin{aligned} v^* &= (I - K_{j+1}A)\widehat{C}_{j+1}A^\top \Gamma^{-1} y_{j+1} + (I - K_{j+1}A)\widehat{m}_{j+1} \\ &= \widehat{m}_{j+1} + K_{j+1}(y_{j+1} - A\widehat{m}_{j+1}), \end{aligned}$$

which coincides with the update formula given in (2.39). □

This optimization perspective gives the opportunity to generalize the idea of the Kalman filter to general nonlinear dynamics and observation models.

2.3.5 Extended Kalman filter

We extend the ideas of the Kalman filter to nonlinear dynamical system. Therefore, we consider the stochastic dynamical system (2.30) and the corresponding observations (2.31)

$$Z_{j+1} = H(Z_j) + \xi_j, \quad Y_{j+1} = h(Z_{j+1}) + \eta_{j+1}, \quad j \in \mathbb{N},$$

where $H : \mathbb{R}^I \to \mathbb{R}^I$ and $h : \mathbb{R}^I \to \mathbb{R}^K$ are possibly nonlinear. The idea behind the extension of the Kalman filter, is to linearize in each iteration the nonlinear system around the current state estimate and apply the linear Kalman filter update step introduced in section 2.3.3. The accuracy of this linearization approximation depends crucially on how strong the nonlinearity in the dynamical system is. In particular, for strongly nonlinear forward models, methods based on Gaussian approximation perform poorly, as the resulting distribution from (2.33) and (2.34) are poorly approximated through Gaussian measures. However, the so called extended Kalman filter (ExKF), can be viewed as the best linear unbiased estimator of the linearized dynamical system. This estimator can often be a good approximation of the original nonlinear system.
Given the previous state estimation m_j, we obtain the linearized approximation of the stochastic dynamical system (2.30) through

$$Z_{j+1} = H_j(m_j) + DH_j(m_j)(Z_j - m_j) + \xi_j, \quad j \in \mathbb{N},$$

where DH_j denotes the derivative of the forward map H_j. For simplicity, we assume again the linear observation model (2.36). Note that one can easily extend the ideas to a linearized observation model.
Further, we assume again for simplicity a fixed observational model, i.e. $A = A_j$ for all $j \in \mathbb{N}$. We define the linearized dynamical system through

$$Z_{j+1} = F_j Z_j + b_j + \xi_j, \quad j \in \mathbb{N},$$

for $F_j := DH(m_j)$ and $b_j := H_j(m_j) - F_j m_j$. We assume again Gaussian initial distribution $Z_0 \sim \mathcal{N}(m_0, C_0)$ and compute the prediction step similar to (2.37) by

$$\widehat{m}_{j+1} = F_j m_j + b_j, \quad \widehat{C}_{j+1} = F_j C_j F_j^\top + \Sigma.$$

and the Bayesian update step similar to (2.39)

$$\begin{aligned} m_{j+1} &= \widehat{m}_{j+1} + K_{j+1}(y_{j+1} - A\widehat{m}_{j+1}), \\ C_j &= \widehat{C}_{j+1} - K_{j+1}A\widehat{C}_{j+1}, \end{aligned}$$

where the Kalman gain matrix K_j is defined through (2.38).

2.3.6 Ensemble Kalman filter

The ensemble Kalman filter can be viewed as a Monte Carlo approximation of the Kalman filter. While in the linear setting of the dynamical system (2.35) and observations (2.36) the filtering distribution stays Gaussian, in the nonlinear setting the filtering distribution is in general non-Gaussian. The idea of the ensemble Kalman filter is to use a particle system, i.e. a sample initialized by the prior distribution $Z_0 \sim \pi_0$, which will be updated to approximate the non-Gaussian filtering distribution (2.34). These updates are based on the linear Kalman filter steps (2.37) and (2.39). Since we do not use any Gaussian assumptions and approximate the filtering distribution through a particle system, we are able to apply the ensemble Kalman filter in nonlinear dynamical systems (2.30). For simplicity, we again assume linear observations (2.36).
In the previous presented linear Kalman filter, we have approximated the filtering distribution through $\mathcal{N}(m_j, C_j)$ in each iteration. As in the nonlinear case we are non-Gaussian, we approximate the filtering distribution through

$$\pi_{Z_j|y^{1:j}}(\mathrm{d}v) \approx \widehat{\pi}_j(\mathrm{d}v) = \frac{1}{M}\sum_{m=1}^{M} \delta_{v_j^{(m)}}(\mathrm{d}v), \tag{2.43}$$

where $(v_j^{(m)})_{m=1,\dots,M}$ denotes the particle system of the current iteration, initialized as i.i.d. sample $v_0^{(m)} \sim \pi_0$, $m = 1, \dots, M$. We will denote M as the *ensemble size* of the particle system.
We introduce the **ensemble Kalman filter (EnKF)** with **perturbed observation**. Given the current particle system $(v_j^{(m)})_{m=1,\dots,M}$, the particles are updated in the following **prediction step** and **analysis step**. The prediction step uses the dynamical system (2.30) to predict the system's current state. Therefore, we map the particles within our dynamical system system

$$\widehat{v}_{j+1}^{(m)} = H(v_j^{(m)}) + \xi_j^{(m)}, \quad m = 1, \dots, M, \tag{2.44}$$

where we again assume fixed $H = H_j$ for all $j \in \mathbb{N}$ and $\xi_j^{(m)}$ are i.i.d. samples from $\mathcal{N}(0, \Sigma)$. Further, we compute the empirical mean and the sample covariance of the particle system

$$\widehat{m}_{j+1} = \frac{1}{M}\sum_{m=1}^{M} \widehat{v}_{j+1}^{(m)}, \quad \widehat{C}_{j+1} = \frac{1}{M}\sum_{m=1}^{M} (v_{j+1}^{(m)} - \widehat{m}_{j+1})(v_{j+1}^{(m)} - \widehat{m}_{j+1})^\top. \tag{2.45}$$

We refer to (2.44) and (2.45) as **prediction step**. Using these predictions, we apply the linear Kalman update (2.39) to each particle itself, which corresponds to a Gaussian approximation. This means we update each particle by

$$\begin{aligned} v_{j+1}^{(m)} &= \widehat{v}_{j+1}^{(m)} + K_{j+1}(\widetilde{y}_{j+1}^{(m)} - A\widehat{v}_{j+1}^{(m)}), \\ \widetilde{y}_{j+1}^{(m)} &= y_j + \eta_{j+1}^{(m)}, \quad \eta_{j+1}^{(m)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Gamma). \\ K_j &= \widehat{C}_j A^\top (A\widehat{C}_j A^\top + \Gamma)^{-1}, \end{aligned} \tag{2.46}$$

where we denote $\widetilde{y}_{j+1}^{(m)}$ as **perturbed observation** and K_j is again the introduced Kalman gain. We refer to (2.46) as **analysis step**. One additional advantage of the EnKF is, that we do not have to update the covariance matrix as in (2.37) and (2.39). Instead, we only have to compute the sample covariance in each iteration, which saves computational effort. To give an overview of the described EnKF method, we present the Figure 2.10 and Algorithm 3.

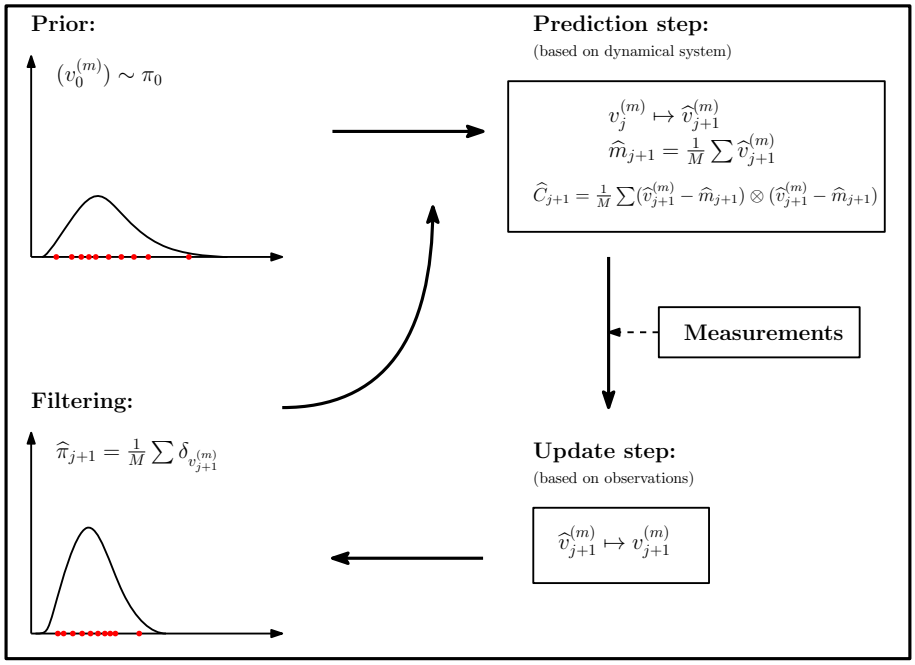

Figure 2.10: Summary of the ensemble Kalman filter method.

Similar to the variational motivation of the KF, one can also view the analysis step of the EnKF in a variational fashion. Given $\widehat{v}_j^m$ and $\widehat{C}_{j+1}$ from the prediction step, the natural choice of loss function would be

$$\mathcal{I}_j^{(m)}(v) = \frac{1}{2}\|y_j^{(m)} - v\|_\Gamma + \frac{1}{2}\|v - \widehat{v}_j^{(m)}\|_{\widehat{C}_j}^2,$$

and update each particle by minimizing the loss function $\mathcal{I}_j^{(m)}$. However, since we have a finite ensemble size, we can not ensure that the sample covariance $\widehat{C}_j$ is positive definite, such that we have to include some auxilary term by defining

$$\widehat{C}_j^\varepsilon := \widehat{C}_j + \varepsilon I.$$

Algorithm 3: Ensemble Kalman filter

Input: initial ensemble $(v_0^{(j)})_{j=1}^J \sim \pi_0$, observations $(y_1, \dots, y_N)$
Output: $(\widehat{\pi}_j)_{j=1,\dots,N}$
for $j = 0, \dots, N-1$ **do**

Prediction step:

- Map the particles through the dynamical system

$$\widehat{v}_{j+1}^{(m)} = H(v_j^{(m)}) + \xi_j^{(m)}, \quad m = 1, \dots, M,$$

- Define sample mean and sample covariance

$$\widehat{m}_{j+1} = \frac{1}{M} \sum_{m=1}^{M} \widehat{v}_{j+1}^{(m)},$$
$$\widehat{C}_{j+1} = \frac{1}{M} \sum_{m=1}^{M} (v_{j+1}^{(m)} - \widehat{m}_{j+1})(v_{j+1}^{(m)} - \widehat{m}_{j+1})^\top.$$

Analysis step:

- Define the Kalman gain

$$K_{j+1} = \widehat{C}_{j+1} A^\top (A \widehat{C}_{j+1} A^\top + \Gamma)^{-1}.$$

- Update each ensemble member by

$$v_{j+1}^{(m)} = \widehat{v}_{j+1}^{(m)} + K_{j+1}(\widetilde{y}_{j+1}^{(m)} - A\widehat{v}_{j+1}^{(m)}),$$

where we consider perturbed observation

$$\widetilde{y}_{j+1}^{(m)} = y_j + \eta_{j+1}^{(m)}, \quad \eta_{j+1}^{(m)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Gamma).$$

Approximate filtering distribution: $\widehat{\pi}_{j+1} = \frac{1}{M} \sum_{m=1}^{M} \delta_{v_{j+1}^{(m)}}$.

We define the loss functions

$$\mathcal{I}_j^{(m),\varepsilon}(v) = \frac{1}{2}\|y_j^{(m)} - v\|_\Gamma + \frac{1}{2}\|v - \widehat{v}_j^{(m)}\|^2_{\widehat{C}_j^\varepsilon} \tag{2.47}$$

and verify that

$$\lim_{\varepsilon \to 0} (v_\varepsilon^*)^{(m)} = v_{j+1}^{(m)},$$

where $(v_\varepsilon^*)^{(m)}$ are the global minimizers of (2.47) and $v_{j+1}^{(m)}$ are the updates computed in (2.46).

Theorem 2.3.3. *Let $A \in \mathbb{R}^{K \times n}$ be of rank K and $\{\widehat{v}_{j+1}^{(m)}\}_{m=1}^M$, $\widehat{C}_{j+1}$ be given through* (2.44) *and* (2.45)*. The solutions of the minimization problems*

$$\min_{v \in \mathbb{R}^I} \mathcal{I}_{j+1}^{(m)}(v)$$

are given through

$$(v_\varepsilon^*)^{(m)} = \widehat{m}_{j+1} + K_{j+1}^\varepsilon (y_{j+1} - A\widehat{m}_{j+1}),$$

where

$$K_j^\varepsilon = \widehat{C}_j^\varepsilon A^\top (A\widehat{C}_j^\varepsilon A^\top + \Gamma)^{-1}.$$

Furthermore, it follows that

$$\lim_{\varepsilon \to 0} (v_\varepsilon^*)^{(m)} = v_{j+1}^{(m)},$$

Proof. The proof follows similar to the proof of Theorem 2.3.2 □

Literature overview

The EnKF has been originally introduced by Evensen [80] and has been reported to produce reliable estimates of the unknown parameters with low computational cost, making the method very appealing for large scale problems. Areas of application include, for example, groundwater flow[172], climate models [203], biological problems [105], image reconstruction [34], building [213] and material sciences [108]. For linear dynamical systems and Gaussian initial conditions, analyses of the large ensemble size limit has been done e.g. in [143, 149], and for nonlinear systems the mean-field Kalman filter has been considered in [148]. In [46, 102] multilevel methods for the EnKF have been proposed. In comparison to the mean-field limit, another interesting perspective to consider is the long-time behaviour of the scheme. In [133, 136, 222] the long-time behaviour and ergodicity of the ensemble Kalman filter with arbitrary ensemble size have been analyzed by establishing time uniform bounds to control the filter divergence with variance inflation techniques and ensuring, in addition, the existence of an invariant measure. Accuracy results for a fixed ensemble size in the linear Gaussian setting can be found in [158, 223], and for the ensemble Kalman-Bucy filters applied to continuous-time filtering problems in [64, 67].

3 A particle based optimization method - Basics of ensemble Kalman inversion

The following chapter is devoted to give a basic introduction to the EnKF applied to inverse problems. Following [112] we will introduce in Section 3.1 the ensemble Kalman inversion by interpreting the inverse problem as artificial dynamical system and applying the EnKF. The resulting algorithm can be interpreted from an optimization perspective as well as from the Bayesian perspective as sequential update of the posterior distribution. We derive a continuous-time limit of the ensemble Kalman inversion in Section 3.2 in form of an coupled system of stochastic differential equations (SDE) and illustrate the resulting gradient flow structure driven by the drift term. In Section 3.3 we present a well-posedness result in form of unique existence of strong solutions of the underlying SDE system. Convergence results of the ensemble Kalman inversion in the linear setting are presented in Section 3.4, where we quantify the ensemble collapse and provide convergence results of the data misfit. The presented results in Section 3.3-3.4 will mainly extend the existing results from [200, 201] to the ensemble Kalman inversion with perturbed observation. We finally present numerical results in Section 3.5 in order to verify the presented theoretical results.

3.1 The ensemble Kalman filter applied to inverse problems

The application of the EnKF to solve inverse problems of the structure (2.2) has been introduced in [112]. In this section, we will follow the presented derivation of the EnKF applied to inverse problems - the so called **ensemble Kalman inversion** (EKI).
The introduction will be based on the general inverse problem

$$q = G(\theta) + \zeta, \tag{3.1}$$

where $G : \mathcal{H} \to \mathbb{R}^p$ is some possible nonlinear map between a separable Hilbert spaces $\mathcal{H}$ and a finite-dimensional space $\mathbb{R}^p$, $q \in \mathbb{R}^p$ is some observational measurement and $\zeta \in \mathbb{R}^p$ denotes Gaussian observational noise with known symmetric positive definite covariance operator $\Gamma \in \mathbb{R}^{p\times p}$, i.e. $\zeta \sim \mathcal{N}(0, \Gamma)$.

Remark 3.1.1. *We note that for different choices of spaces $\mathcal{H}$, $\mathbb{R}^p$ and forward models G, we will be able to include additional regularization. More details on this will be presented in chapter 5 and chapter 7. In the setting of solving general inverse problems*

$$y = H(\theta) + \xi,$$

we will set $\mathcal{H} := \mathcal{X}$, $G := H : \mathcal{X} \to \mathbb{R}^K$, $q := y \in \mathbb{R}^K$, $\zeta = \xi$ and $p = K$. However, we will introduce the EKI for (3.1) *in order to be able to introduce different variants of the EKI by modifying the forward model G.*

While the KF and the EnKF respectively has been introduced for time-dependent dynamical systems, we assume that all of the time-dependence is included in the forward model G, for example by observing the total time interval, and consider a static system modelled through G. As seen in the preliminaries 2, inverse problems modelled through (3.1) are typically ill-posed and we consider regularization through prior knowledge.
We define an artificial dynamical system with state space $\mathcal{Z} = \mathcal{H} \times \mathbb{R}^p$ through the signal

$$Z_{n+1} = \psi_n(Z_n),$$

where we define the mapping

$$\psi_n(z) := \psi(z) := \begin{pmatrix} \theta \\ G(\theta) \end{pmatrix} \quad \text{for } z \in \mathcal{Z}.$$

The corresponding observations are modelled through

$$q_{n+1} = \mathcal{O} Z_{n+1} + \zeta_{n+1},$$

with observation operator $\mathcal{O} = [0, I]$, which projects Z_n to its observation $G(\theta_n)$, and $(\zeta_n)_{n\in\mathbb{N}}$ is an i.i.d. sequence with $\zeta_1 \sim \mathcal{N}(0, \Gamma)$. This leads to the perturbed observation case, where the original data q is used to generate $(q_n)_{n\in\mathbb{N}}$ by perturbation

$$q_n = q + \zeta_n.$$

This time-dependent dynamical system has been introduced as artificial time-dependent system, where the time is completely independent of the forward model G.
In order to solve this artificial system, the authors in [112] propose to apply the data assimilation tool EnKF, which is known as the EKI method.
We consider the interacting particle system $(Z_n^{(j)})_{j=1,\dots,J}$ with $Z_n^{(j)} \in \mathcal{Z}$ from which we can generate the estimate of the unknown parameter θ of the inverse problem (3.1) through

$$\bar{\theta}_n = \frac{1}{J}\sum_{j=1}^{J} \theta_n^{(j)} = \frac{1}{J}\sum_{j=1}^{J} \mathcal{O}^{\perp} Z_n^{(j)}, \quad \mathcal{O}^{\perp} = [I, 0].$$

We assume that there exists a true unknown parameter $\theta^\dagger$ which generates the observation y, i.e.

$$q = G(\theta^\dagger) + \zeta^\dagger.$$

The initialization of this particle system is based on some prior knowledge about the unknown parameter $\theta^\dagger$ available through some probability distribution $\mathbb{Q}_0$. We generate an i.i.d. sample $(\theta_0^{(j)})_{j=1,\dots,J}$ from the distribution $\mathbb{Q}_0$ and set

$$Z_0^{(j)} = \begin{pmatrix} \theta_0^{(j)} \\ G(\theta_0^{(j)}) \end{pmatrix}.$$

We will see that based on this initialization we can ensure that the resulting EKI estimate $\bar{\theta}_n$ will stay in the linear subspace spanned from the initial ensemble, i.e.

$$\bar{\theta}_n \in \mathcal{S} := \operatorname{span}\{\theta_0^{(j)}, j = 1, \dots, J\}.$$

Hence, the choice of the initial ensemble is a design parameter which is related to the choice of a subspace $\mathcal{S} \subset \mathcal{H}$, where the underlying truth $\theta^\dagger$ is expected to be inside. We

note that for second-order random fields it is sometimes usefull to initialize the ensemble through the truncated KL-basis of the covariance function.
Following the EnKF prediction and analysis step, the EKI algorithm is described in the follwing.

Algorithm 4: Ensemble Kalman inversion (original)

Input: initial ensemble $(Z_0^{(j)})_{j=1}^J$, observation q
Output: $\bar{\theta}_N$
for $n = 0, \dots, N-1$ **do**
Prediction step:

- Map the ensemble of particles through dynamics

$$\widehat{Z}_{n+1}^{(j)} = \psi(Z_n^{(j)}).$$

- Define sample mean and sample covariance

$$\bar{Z}_{n+1} = \frac{1}{J}\sum_{j=1}^{J} \widehat{Z}_{n+1}^{(j)},$$
$$C_{n+1} = \frac{1}{J}\sum_{j=1}^{J} (\widehat{Z}_{n+1}^{(j)} - \bar{Z}_{n+1}) \otimes (\widehat{Z}_{n+1}^{(j)} - \bar{Z}_{n+1}).$$

Analysis step:

- Define the Kalman gain

$$K_{n+1} = C_{n+1}\mathcal{O}^*(\mathcal{O}C_{n+1}\mathcal{O}^* + \Gamma)^{-1},$$

where $\mathcal{O}^*$ is the adjoint operator of $\mathcal{O} = [0, I]$.

- Update each ensemble member by

$$Z_{n+1}^{(j)} = \widehat{Z}_{n+1}^{(j)} + K_{n+1}(q_{n+1}^{(j)} - \mathcal{O}\widehat{Z}_{n+1}^{(j)}), \tag{3.2}$$

where we consider perturbed observation

$$q_{n+1}^{(j)} = q + \zeta_{n+1}^{(j)}, \quad \zeta_{n+1}^{(j)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Gamma).$$

Estimate: $\bar{\theta}_N = \frac{1}{J}\sum_{j=1}^{J} \theta_N^{(j)} = \frac{1}{J}\sum_{j=1}^{J} \mathcal{O}^{\perp} Z_N^{(j)}$.

We note that in the standard EnKF method the application of perturbed observations is necessary to capture statistical properties of the distribution conditioned to observations. In the EKI context the perturbed observations are motivated as randomization of the data in order to move around in the initial subspace $\mathcal{S}$ and improve the approximation. Though both methods, i.e. the limit of the EKI with perturbed observations and the later discussed deterministic limit from [200], can be analysed from an optimization perspective. The EKI variant with perturbed observation is shown to be second order accurate, whereas the deterministic limit underestimates the covariance in the linear, Gaussian setting, see

e.g. [80]. In addition, in the nonlinear setting, methods that add noise to data are reported to be more robust to assumptions about linearity and normality, see e.g. [230] and the references therein. The focus in Section 3.3-3.4 will be on the EKI with perturbed observations, as the presented analysis provides valuable insights for the development of methods for the nonlinear, non-Gaussian setting.
While the EKI has been introduced in an artificial state space $\mathcal{Z}$, the analysis of this algorithm will be based on the estimates in the parameter space $\mathcal{H}$. In order to prove the well-known subspace property, we take a deeper look into the update formula (3.2) for

$$Z_n^{(j)} = \begin{pmatrix} \theta_n^{(j)} \\ p_n^{(j)} \end{pmatrix}$$

initialized by $p_0^{(j)} = G(\theta_0^{(j)})$.
We split the prediction step in $\mathcal{Z}$ into the following computations

$$\widehat{Z}_{n+1}^{(j)} = \begin{pmatrix} \widehat{\theta}_{n+1}^{(j)} \\ \widehat{p}_{n+1}^{(j)} \end{pmatrix} = \begin{pmatrix} \theta_n^{(j)} \\ G(\theta_n^{(j)}) \end{pmatrix},$$
$$\bar{Z}_{n+1} = \begin{pmatrix} \bar{\theta}_n \\ \bar{G}_n \end{pmatrix}, C_{n+1} = \begin{pmatrix} C_{n+1}^{\theta\theta} & C_{n+1}^{\theta p} \\ (C_{n+1}^{\theta p})^\top & C_{n+1}^{pp} \end{pmatrix},$$

where we have defined the empirical means and sample covariances

$$\bar{\theta}_n := \frac{1}{J}\sum_{j=1}^{J} \theta_n^{(j)}, \ \bar{G}_n := \frac{1}{J}\sum_{j=1}^{J} G(\theta_n^{(j)}),$$
$$C_{n+1}^{\theta\theta} := \frac{1}{J}\sum_{j=1}^{J} (\theta_n^{(j)} - \bar{\theta}_n) \otimes (\theta_n^{(j)} - \bar{\theta}),$$
$$C_{n+1}^{\theta p} := \frac{1}{J}\sum_{j=1}^{J} (\theta_n^{(j)} - \bar{\theta}_n) \otimes (G(\theta_n^{(j)}) - \bar{G}),$$
$$C_{n+1}^{pp} := \frac{1}{J}\sum_{j=1}^{J} (G(\theta_n^{(j)}) - \bar{G}) \otimes (G(\theta_n^{(j)}) - \bar{G}).$$

Here, the operator $\otimes$ denotes the tensor product (or rank one operator) given by

$$z_1 \otimes z_2 : \mathcal{H}_2 \to \mathcal{H}_1 \text{ with } h \mapsto z_1 \otimes z_2(h) := \langle z_2, h\rangle_{\mathcal{H}_2} \cdot z_1$$

for Hilbert spaces $(\mathcal{H}_1, \langle\cdot,\cdot\rangle_{\mathcal{H}_1}), (\mathcal{H}_2, \langle\cdot,\cdot\rangle_{\mathcal{H}_2})$ and $z_1 \in \mathcal{H}_1, z_2 \in \mathcal{H}_2$. This gives the possibility to split the Kalman gain in

$$K_{n+1} = \begin{pmatrix} C_{n+1}^{\theta p}(C_{n+1}^{pp} + \Gamma)^{-1} \\ C_{n+1}^{pp}(C_{n+1}^{pp} + \Gamma)^{-1}, \end{pmatrix}$$

and write the update of each ensemble member $Z_n^{(j)}$ through

$$Z_{n+1}^{(j)} = \begin{pmatrix} \theta_{n+1}^{(j)} \\ p_{n+1}^{(j)} \end{pmatrix} = \begin{pmatrix} \theta_n^{(j)} + C_{n+1}^{\theta p}(C_{n+1}^{pp} + \Gamma)^{-1}(q_{n+1}^{(j)} - G(\theta_n^{(j)}) \\ G(\theta_n^{(j)}) + C_{n+1}^{pp}(C_{n+1}^{pp} + \Gamma)^{-1}(q_{n+1}^{(j)} - G(\theta_n^{(j)}) \end{pmatrix}.$$

We are now ready to formulate the subspace property for the EKI method.

Lemma 3.1.2 ([112, Theorem 2.1]). *Let $\mathcal{S}$ be the linear span of $\{\theta_0^{(j)}\}_{j=1}^J$, then $\theta_n^{(j)} \in \mathcal{S}$ for all $(n, j) \in \mathbb{N} \times \{1, \dots, J\}$.*

Proof. Using the definition of C_{n+1}^{up}, we obtain the update formula

$$\theta_{n+1}^{(j)} = \theta_n^{(j)} + \frac{1}{J}\sum_{k=1}^{J} \langle G(\theta_n^{(k)}) - \bar{G}_n, (C_{n+1}^{pp} + \Gamma)^{-1}(q_{n+1}^{(j)} - G(\theta_n^{(j)}))\rangle \theta_n^{(k)},$$

which states that the updated $\theta_{n+1}^{(j)}$ is a linear combination of the previous ensemble $\{\theta_n^{(j)}\}_{j=1}^J$. □

In particular, with this result, one can ensure that the EKI estimate $\bar{\theta}_n$ always stays in the subspace spanned by the initial ensemble.

Since our theory will be based on the inverse problem (2.2) with parameter space $\mathcal{X}$, we formulate the updates in the parameter space and set the notation $\mathcal{H} := \mathcal{X}$, $G := H : \mathcal{X} \to \mathbb{R}^K$, $q := y \in \mathbb{R}^K$, $\zeta = \xi$ and $p = K$. Hence, the theory will be based on the EKI applied to (2.2), which is formulated in the following algorithm.

Algorithm 5: Ensemble Kalman inversion

Input: initial ensemble $(\theta_0^{(j)})_{j=1}^J$, observation y
Output: $\bar{\theta}_N$
for $n = 0, \dots, N-1$ **do**
Prediction step:
Define sample mean and sample covariance

$$\begin{aligned} \bar{\theta} &= \frac{1}{J}\sum_{j=1}^{J} \theta^{(j)}, \quad \bar{H} = \frac{1}{J}\sum_{j=1}^{J} H(\theta^{(j)}), \\ C^{pp} &= \frac{1}{J}\sum_{j=1}^{J} (H(\theta^{(j)}) - \bar{H}) \otimes (H(\theta^{(j)}) - \bar{H}), \\ C^{\theta p} &= \frac{1}{J}\sum_{j=1}^{J} (\theta^{(j)} - \bar{\theta}) \otimes (H(\theta^{(j)}) - \bar{H}) \end{aligned} \tag{3.3}$$

Analysis step:
Update each ensemble member by

$$\theta_{n+1}^{(j)} = \theta_n^{(j)} + C_{n+1}^{\theta p}(C_{n+1}^{pp} + \Gamma)^{-1}(y_{n+1}^{(j)} - G(\theta_n^{(j)}), \tag{3.4}$$

where we consider perturbed observation

$$y_{n+1}^{(j)} = y + \xi_{n+1}^{(j)}, \quad \xi_{n+1}^{(j)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Gamma). \tag{3.5}$$

Estimate: $\bar{\theta}_N = \frac{1}{J}\sum_{j=1}^{J} \theta_N^{(j)}$.

Literature overview

While the EKI has been invented in [112], there have been earlier ensemble based methods which aim to solve inverse problems [45, 75]. Although the EKI shows promising success from a practical perspective, the theoretical understanding of the scheme is limited. Most of the existing results in the literature are based on a formally derived continuous-time limit of the algorithm presented in [200], which can be seen as an interacting system of gradient flows described by a coupled system of SDEs. We also refer to [15, 16, 183] for the continuous-time limit of the EnKF in the data assimilation context. In [25], first theoretical verifications of the continuous-time limit can be found, where the authors view the original EKI algorithm as a discrete approximation of the underlying system of coupled SDEs. More recently, a stabilized continuous-time limit has been proposed in [8]. We distinguish between two perspectives of the EKI. The first one is the motivation of the EKI as a sampling method, whereas in the second perspective the focus lies in the long-time behaviour as an optimization method.

EKI as sampling method: The large ensemble size limit of the EKI has been discussed in [79], where the authors show that the EKI method is not consistent with the Bayesian perspective in general nonlinear settings. However, it can be interpreted as a point estimator of the underlying unknown parameter. Furthermore, the mean-field limit has been analyzed based on kinetic methods in [99] and a connection of the mean-field limit to the Fokker–Planck equation has been discussed in [69]. For the original EKI method, in [71] the authors suggest to introduce weights on the particles in order to correct the particle system to obtain a consistent mean-field limit from a Bayesian perspective. A alternative perspective has been proposed in [83] and further discussed in [70, 113], known as the ensemble Kalman sampling method. The basic idea is to shift the pertubation from the observations to the particles itself, resulting in a system of SDEs where the diffusion part takes place in the parameter space itself. Theoretical studies are based on the mean-field limit and the corresponding Fokker–Planck equation.

EKI as optimization method: From an alternative perspective, the EKI can be viewed as an optimization method of the least-squares misfit functional. The gradient flow structure of the EKI has been highlighted in [200, 201] and the viewpoint as derivative-free optimization method has been pointed out in [140], where the authors apply the EKI for the training of neural networks. First convergence results for the EKI as an optimization method have been shown in the linear setting based on the continuous-time limit [200, 201], while convergence for the nonlinear setting for the discrete method has been considered in [39]. Here, the authors suggest non-constant step sizes and covariance inflation in order to tune the method. However, when talking about EKI as an optimization method for inverse problems, a natural question is how to deal with noise in the data. First results were based on the incorporation of the discrepancy principle [201]. In the discrete setting, the connection to deterministic regularisation techniques have been established in [109, 110]. In particular, in [41] the authors incorporate Tikhonov regularization within the algorithm of EKI and present first theoretical results based on the continuous-time limit of the resulting scheme.

3.1.1 Motivation through Optimization

In the following we will build a connection between the EKI estimate and the Tikhonov regularized solution introduced in subsection 2.1.2 and the MAP estimate of the BIP. To

do so, we consider the inverse problem

$$y = L\theta + \xi,$$

for linear forward map $L \in \mathcal{L}(\mathcal{X}, \mathbb{R}^K)$ with the corresponding Tikhonov loss function

$$T_\kappa = \frac{1}{2}\|L\theta - y\|_\Gamma^2 + \frac{1}{2}\|\theta - m\|_C^2, \quad \kappa = 1,$$

for some symmetric and positive definite operators $\Gamma \in \mathbb{R}^{K\times K}$, $C \in \mathcal{L}(\mathcal{X}, \mathcal{X})$ and the regularized solution

$$\theta_* = (L^*\Gamma^{-1}L + C)^{-1}(L^*y + C^{-1}m).$$

Recall, that this solution corresponds to the MAP estimate of the BIP with Gaussian prior $\mathbb{Q}_0 = \mathcal{N}(m, C)$. Similar to example 2.2.12 application of the Woodbury-matrix-identity gives

$$\theta_* = m + CL^*(LCL + \Gamma)^{-1}(y - Lm).$$

Consider an i.i.d. initial ensemble $\{\theta_0^{(j)}\}_{j=1}^J$ with $\theta_0^{(1)} \sim \mathcal{N}(m, C)$ and consider one update step of Algorithm 5, i.e. we set $N = 1$. Define the empirical mean and sample covariance of the initial ensemble by

$$m_J = \frac{1}{J}\sum_{j=1}^J \theta_0^{(j)}, \quad C_J = \frac{1}{J-1}\sum_{j=1}^J (\theta_0^{(j)} - m_J) \otimes (\theta_0^{(j)} - m_J).$$

Since we can write for linear forward models

$$\begin{aligned} C^{\theta p} &= \frac{1}{J}\sum_{j=1}^J (\theta_0^{(j)} - m_J) \otimes (L\theta_0^{(j)} - Lm_J) \\ &= \frac{1}{J}\sum_{j=1}^J (\theta_0^{(j)} - m_J) \otimes (\theta_0^{(j)} - m_J)L^* = \frac{J-1}{J} C_J L^* \end{aligned}$$

and similarly

$$C_1^{pp} = \frac{J-1}{J} L C_J L^*,$$

the EKI estimate can be written through

$$\bar{\theta}_1 = m_J + C_J L(LC_J L^* + \Gamma)^{-1}(y + \frac{1}{J}\sum_{j=1}^J \xi_1^{(j)} - Lm_J).$$

Taking the limit $J \to \infty$ gives that

$$\lim_{J\to\infty} \bar{\theta}_1 = m + CL(LCL^* + \Gamma)^{-1}(y - Lm) = \theta_*$$

almost surely. Thus, we have seen, that for linear forward models and Gaussian prior, the EKI estimate results exactly in the classical Kalman filter update and in the MAP estimate respectively for the large ensemble size limit $J \to \infty$.
For fixed $J \geq 2$, we can interpret the update of the EKI estimate

$$\bar{\theta}_{n+1} = \bar{\theta}_n + C_n^{\theta\theta} L(LC_n^{\theta\theta} L^* + \Gamma)^{-1}(y + \frac{1}{J}\sum_{j=1}^J \xi_{n+1}^{(j)} - L\bar{\theta}_n)$$

as randomized computation of the MAP estimate for an sequential prior distribution $\mathbb{Q}_n = \mathcal{N}(\bar{\theta}_n, C_n^{\theta\theta})$, i.e. $\mathbb{E}[\bar{\theta}_{n+1}]$ is the minimizer of the loss functional

$$I_n(\theta) = \frac{1}{2}\|L\theta - y\|_\Gamma^2 + \frac{1}{2}\|\theta - \bar{\theta}_n\|_{C_n^{\theta\theta}},$$

if one can ensure that $C_n^{\theta\theta}$ is strictly positive definite. This can be guaranteed for a large enough amount of particles or by introduction of socalled covariance inflation, which we introduce in subsection 3.2.2.
Note that without taking expectation on $\bar{\theta}_{n+1}$ this update is not exact, as it is perturbed by the randomization through $\frac{1}{J}\sum_{j=1}^{J} \xi_{n+1}^{(j)}$. This fact gives the rise to interpret the EKI as sequential method in a Bayesian fashion, where the particles represents a sequentially updated distribution $\mathbb{Q}_n$. We will give more details in the next section.

3.1.2 Motivation through Bayesian inverse problems

We motivate the EKI method from a Bayesian perspective of inverse problems. Therefore, we consider the stochastic model (2.14), where we assume that we have some prior information about the unknown parameter given through a probability distribution, i.e. $\Theta \sim \mathbb{Q}_0$, and the noise is assumed to be Gaussian $\Xi \sim \mathcal{N}(0, \Gamma)$, stochastically independent of Θ. Recall, that the posterior is given through

$$\mathbb{Q}_y^*(\mathrm{d}\theta) = \frac{1}{Z}\exp(-\Phi(\theta, y))\,\mathbb{Q}_0(\mathrm{d}\theta),$$
$$Z = \int_{\mathcal{X}} \exp(-\Phi(\theta, y))\,\mathbb{Q}_0(\mathrm{d}\theta),$$

where we have defined the potential $\Phi(\theta, y) = \|H(\theta) - y\|_\Gamma$. We introduce an artificial discrete-time dynamical system, mapping the prior into the posterior distribution, where we define

$$\mathbb{Q}_n(\mathrm{d}\theta) \propto \exp(-\frac{n}{N}\Phi(\theta, y))\,\mathbb{Q}_0(\mathrm{d}\theta), \tag{3.6}$$

which gives the posterior distribution for $n = N$, i.e. $\mathbb{Q}_N = \mathbb{Q}_y^*$. This formulation gives rise to introduce the sequential update formulation

$$\mathbb{Q}_{n+1}(\mathrm{d}\theta) = \frac{1}{Z_n}\exp\big(-h\Phi(\theta, y)\big)\mathbb{Q}_n(\mathrm{d}\theta)\,, \quad n = 0, \dots, N-1,$$
$$Z_n = \int \exp(-h\Phi(\theta, y))\mathbb{Q}_n(\mathrm{d}\theta),$$

with $h = N^{-1}$ denoting the step size of artificial time. This means that the current iteration is absolutely continuous w.r.t. the previous one, i.e. $\mathbb{Q}_{n+1} \ll \mathbb{Q}_n$.
Figure 3.1 demonstrates this idea as sequential update of the prior measure ending up in the posterior at stage N.
Taking a deeper look into the log-likelihood, we see

$$-h\Phi(\theta, y) = -\frac{h}{2}\|H(\theta) - y\|_\Gamma = -\frac{1}{2}\|H(\theta) - y\|_{\frac{1}{h}\Gamma},$$

which states that the introduced artificial discrete-time system coincides with a scaling of the noise covariance

$$\Gamma \mapsto \frac{1}{h}\Gamma. \tag{3.7}$$

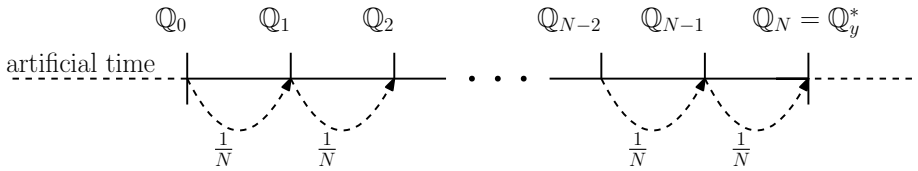

Figure 3.1: Sequential update from the prior distribution to the posterior distribution.

The application of the EKI method can now be viewed as sequential MC-method. Similar to (2.43), we aim to approximate $\mathbb{Q}_n$ in each update step through a an ensemble of particles $(\theta_n^{(j)})$, i.e.

$$\mathbb{Q}_n(\mathrm{d}\theta) \approx \frac{1}{J}\sum_{j=1}^{J} \delta_{\theta_n^{(j)}}(\mathrm{d}\theta).$$

We note that for general method one can introduce weights $w_n^{(j)}$, summing up to 1, which has been set equaly weighted in our setting, i.e. $w_n^{(j)} = \frac{1}{J}$ similar to (2.43). The ensemble of particles at time n can now be sequentially updated to those at time $n+1$ in order to approximate the distribution $\mathbb{Q}_{n+1}$. One hopes that these update steps are exact in the limit $J \to \infty$. We introduce the scaling (3.7) of the noise covariance and apply EKI in order to approximate $\mathbb{Q}_n$ through a Gaussian approximation.

3.2 Continuous-time limit of the ensemble Kalman inversion

We present a formal derivation of the continuous-time limit for the EKI, which has been firstly proposed in [200], and preliminary theoretical studied in [25]. In the context of EnKF and the ESRF applied to continuous-time data assimilation problems, the time limit has been studied in [145, 144] as well as for the ensemble Kalman–Bucy filter [17] and its long-time stability [64].
Recall, that for a given artificial step-size $h > 0$ and $J \geq 2$ particles, the EKI iteration for the j-th particle is given by

$$\theta_{n+1}^{(j)} = \theta_n^{(j)} + C^{\theta p}(\theta_n)(C^{pp}(\theta_n) + h^{-1}\Gamma)^{-1}(y_{n+1}^{(j)} - H(\theta_n^{(j)})), \quad j = 1, \ldots, J\,, \tag{3.8}$$

where the initial particles $\theta_0^{(j)}$, $j = 1, \ldots, J$ are draws from the prior distribution, and in each step, we consider artificially perturbed data

$$y_{n+1}^{(j)} = y + \xi_{n+1}^{(j)}\,,$$

where the perturbations $\xi_{n+1}^{(j)}$, w.r.t. both j and n, are i.i.d. random variables distributed according to $\mathcal{N}(0, h^{-1}\Gamma)$.
We can rewrite (3.8) in terms of

$$\begin{aligned}\theta_{n+1}^{(j)} &= \theta_n^{(j)} + C^{\theta p}(\theta_n)(C^{pp}(\theta_n) + h^{-1}\Gamma)^{-1}(y_{n+1}^{(j)} - H(\theta_n^{(j)})) \\ &= \theta_n^{(j)} + hC^{\theta p}(\theta_n)(hC^{pp}(\theta_n) + \Gamma)^{-1}(y - H(\theta_n^{(j)})) \\ &\quad + \sqrt{h}C^{\theta p}(\theta_n)(hC^{pp}(\theta_n) + \Gamma)^{-1}\Gamma^{\frac{1}{2}}\xi_{n+1}^{(j)},\end{aligned}$$

with $\xi_{n+1}^{(j)}$ are i.i.d. distributed according to $\mathcal{N}(0, I)$, again w.r.t. both j and n. Hence, we can view $\Delta W_{n+1}^{(j)} := \sqrt{\Delta t}\, \xi_{n+1}^{(j)}$, $\Delta t = h$, as increments of J independent brownian motions in $\mathbb{R}^K$, and we rewrite the EKI update as

$$\begin{aligned}\theta_{n+1}^{(j)} = \theta_n^{(j)} &+ \Delta t C^{\theta p}(\theta_n)(hC^{pp}(\theta_n) + \Gamma)^{-1}(y - H(\theta_n^{(j)})) \\ &+ C^{\theta p}(\theta_n)(\Delta t C^{pp}(\theta_n) + \Gamma)^{-1}\Gamma^{\frac{1}{2}}\Delta W_{n+1}^{(j)}.\end{aligned}$$

With this point of view, the continuous-time limit of the discrete EKI (3.8) is formally a time discretization of the following SDE:

$$\mathrm{d}\theta_t^{(j)} = C^{\theta p}(\theta_t)\Gamma^{-1}(y - H(\theta_t^{(j)}))\,\mathrm{d}t + C^{\theta p}(\theta_t)\Gamma^{-1/2}\,\mathrm{d}W_t^{(j)}. \tag{3.9}$$

The processes $W^{(j)}$ are independent Brownian motions on $\mathbb{R}^K$. We further denote by $\mathcal{F}_t = \sigma(\theta_s, s \leq t)$ the filtration introduced by the particle dynamics.
The continuous-time limit of the EKI is still an open point, and has to be verified. In particular, there are two open questions

- Does the continuous-time interpolation $\theta_h^{(j)}(t)$ of the EKI converge to the continuous-time limit $\theta^{(j)}(t)$, which is given through a solution of a set of coupled SDEs, and in which sense does it converge?
- Can we say something about the stability error $\|\theta_h^{(j)}(t) - \theta^{(j)}(t)\|$, if the time tends to infinity? Do the limits of $\theta_h^{(j)}(t)$ and $\theta^{(j)}(t)$ coincide with $t \to \infty$?

The second point is an important point, as our analysis is based on the long-time behaviour study of the continuous-time limit (3.9).

3.2.1 Gradient flow structure - Derivative free optimization method

Through the derived continuous-time limit, the authors from [200] proposed to view the EKI method through its gradient flow structure. To see the connection between the EKI and a gradient flow, we will first assume that the forward response operator is linear, i.e. $H(\cdot) = L\cdot$ with $L \in \mathcal{L}(\mathcal{X}, \mathbb{R}^K)$. The continuous-time limit (3.9) can be read as

$$\mathrm{d}\theta_t^{(j)} = \frac{1}{J}\sum_{k=1}^{J}\left\langle L(\theta_t^{(k)} - \bar{\theta}_t), (y - L\theta_t^{(j)})\,\mathrm{d}t + \sqrt{\Gamma}\,\mathrm{d}W_t^{(j)}\right\rangle_{\Gamma}(\theta_t^{(k)} - \bar{\theta}_t). \tag{3.10}$$

By defining the empirical covariance operator

$$C(\theta) = \frac{1}{J}\sum_{k=1}^{J}(\theta^{(k)} - \bar{\theta}) \otimes (\theta^{(k)} - \bar{\theta}),$$

we simplify notation and equation (3.10) can be rewritten in the form

$$\mathrm{d}\theta_t^{(j)} = C(\theta_t)L^*\Gamma^{-1}(y - L\theta_t^{(j)})\,\mathrm{d}t + C(\theta_t)L^*\Gamma^{-1/2}\,\mathrm{d}W_t^{(j)}. \tag{3.11}$$

Since we assume a linear setting, we can compute the derivative of the potential w.r.t. the parameter θ

$$\nabla_\theta \Phi(\theta, y) = \nabla_\theta\left(\frac{1}{2}\|L\theta - y\|_\Gamma^2\right) = L^*\Gamma^{-1}(L\theta - y).$$

By this fact, we can interpret the drift of (3.11) as preconditioned gradient flow, where the preconditioner is given through the sample covariance $C(\theta)$,

$$\mathrm{d}\theta_t^{(j)} = -C(\theta_t)\nabla_\theta \Phi(\theta_t^{(j)}, y)\,\mathrm{d}t + C(\theta_t)L^*\Gamma^{-1/2}\,\mathrm{d}W_t^{(j)}.$$

In [200] the main analysis has been based on this gradient flow structure. While the diffusion part has been suppressed, the convergence analysis is based in a deterministic setting on the system of coupled ODEs

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta^{(j)} = -C(\theta)\nabla_\theta \Phi(\theta^{(j)}, y), \tag{3.12}$$

and its long-time behaviour $t \to \infty$ as the EKI has been viewed as optimization method. This perspective of the EKI as gradient flow, opens the possibility to interpret the EKI as derivative free optimization method and therefore as black-box solver of the corresponding optimization problem. Also for nonlinear forward models H we consider the corresponding system of coupled ODEs, and connect this system to an approximation of its gradient flow. To do so, we consider the drift of (3.9) as system of ODEs

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta^{(j)} = C^{\theta p}(\theta)\Gamma^{-1}(y - H(\theta)). \tag{3.13}$$

We follow the interpretation of (3.13) as approximate gradient flow in [140]. Using the definition of the empirical covariance, (3.13) can be formulated equivalently as

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta^{(j)} = \frac{1}{J}\sum_{k=1}^{J}\left\langle H(\theta^{(k)}) - \bar{H}, \Gamma^{-1}(y - H(\theta^{(j)})\right\rangle(\theta^{(k)} - \bar{\theta}). \tag{3.14}$$

We assume that the forward map H is Fréchet differentiable and consider the linearization

$$H(\theta^{(k)}) = H(\theta^{(j)} + \theta^{(k)} - \theta^{(j)}) \approx H(\theta^{(j)}) + DH(\theta^{(j)})(\theta^{(k)} - \theta^{(j)}),$$

where DH denotes the Fréchet derivative of H. The linearization is getting the more accurate the closer the particles $\theta^{(j)}$ and $\theta^{(k)}$ are lying to each other, i.e. the smaller $\|\theta^{(k)} - \theta^{(j)}\|_{\mathcal{X}}^2$ is. Applying

$$\begin{aligned}
H(\theta^{(k)}) - \bar{H} &= \frac{1}{J}\sum_{l=1}^{J}\Big(H(\theta^{(k)}) - H(\theta^{(l)})\Big) \\
&\approx \frac{1}{J}\sum_{l=1}^{J}\Big(H(\theta^{(j)}) + DH(\theta^{(j)})(\theta^{(k)} - \theta^{(j)}) - \Big(H(\theta^{(j)}) + DH(\theta^{(j)})(\theta^{(l)} - \theta^{(j)})\Big)\Big) \\
&= \frac{1}{J}\sum_{l=1}^{J} DH(\theta^{(j)})(\theta^{(k)} - \theta^{(l)}) \\
&= DH(\theta^{(j)})(\theta^{(k)} - \bar{\theta})
\end{aligned}$$

to (3.14) leads to

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta^{(j)} \approx \frac{1}{J}\sum_{k=1}^{J}\left\langle DH(\theta^{(j)})(\theta^{(k)} - \bar{\theta}), \Gamma^{-1}(y - H(\theta^{(j)})\right\rangle(\theta^{(k)} - \bar{\theta})$$

$$= \frac{1}{J}\sum_{k=1}^{J}\left\langle \theta^{(k)} - \bar{\theta}, DH^*(\theta^{(j)})\Gamma^{-1}(y - H(\theta^{(j)})\right\rangle (\theta^{(k)} - \bar{\theta})$$
$$= C(\theta)DH^*(\theta^{(j)})\Gamma^{-1}(y - H(\theta^{(j)})).$$

By computing

$$\nabla_\theta \Phi(\theta, y) = \nabla_\theta \left(\frac{1}{2}\|H(\theta) - y\|_\Gamma^2\right) = DH^*(\theta)\Gamma^{-1}(H(\theta) - y)$$

we obtain the approximate gradient flow structure

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta^{(j)} \approx -C(\theta)\nabla_\theta \Phi(\theta^{(j)}, y).$$

This approximation leads to the hope of first proving that the particle system is collapsing, i.e. the spread $(\theta^{(j)} - \bar{\theta})$ is converging to zero, and then using the property of the approximate gradient flow structure in order to apply EKI as optimizer. However, this idea itself leads to the conflict, that also the preconditioner $C(\theta)$ will degenerate and might not stay strictly positive definite in time, such that no descent direction can be guaranteed. We will discuss this issue in section 3.4 in more detail.
Another arising problem through the gradient flow persepective and the analysis of the long-time behaviour $t \to \infty$, is the loss of regularization effect from the original introduced EKI method, which would only coincide with the solution until time $t = 1$. We will first study theoretical convergence results based on noisefree data in section 3.4 and in chapter 5 we will discuss possibilities how to incorporate regularization to handle the case of noisy data.
Through the deterministic perspective, one can simplify the analysis of the EKI and apply theory based on ODEs to study the asymptotical behaviour. First theoretical results for EKI has been based on this deterministic version, see for example [200, 201].
However, the original derivation of the EKI has been based on perturbed observations, which imply a nontrivial diffusion of the continuous-time limit (3.9). Therefore, our theoretical results will be mainly based on analysis of the set of coupled SDEs (3.9).

Remark 3.2.1. *We note that the presented methods in [200] and also the results for the EKI method we are going to present in the rest of this work can be straightforwardly extended to the ensemble square root filter (ESRF), a filtering method which itself can be viewed as a deterministic variant of the EnKF. The basic idea in the filtering context is to update the particles deterministically in a way such that the empirical covariance exactly satisfies the Kalman identity* (2.39)*. To do so, in the linear setting an explicit transformation of the particles, also known as the ensemble transform Kalman filter, can be found [186, 147]. Based on [16, 15] the continuous-time limit of the ESRF applied to inverse problems can be formulated as*

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta^{(j)} = \frac{1}{J}\sum_{k=1}^{J}\left\langle H(\theta^{(k)}) - \bar{H}, \Gamma^{-1}(y - \frac{1}{2}H(\theta^{(j)} - \frac{1}{2}\bar{H})\right\rangle (\theta^{(k)} - \bar{\theta}).$$

In the linear setting, a gradient flow structure can be achieved in the sense of

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta^{(j)} = -C(\theta)\left(\frac{1}{2}\nabla_\theta \Phi(\theta^{(j)}, y) + \frac{1}{2}\nabla_\theta \Phi(\bar{\theta}, y)\right).$$

This approach can be seen as a deterministic variant of the EKI method.

3.2.2 Covariance inflation

In the high-dimensional setting, the EnKF is known to have certain difficulties, in particular in the case where the dimension of the parameter space is larger than the ensemble size. The main issue is that the particle system resulting from the EnKF underestimates the uncertainty stored through the sample covariance operator. One common way to alleviate this issue is through the incorporation of covariance inflation [5, 6, 135]. The idea is to inflate the sample covariance by addition of another positive definite operator, for example by the prior covariance operator, i.e.

$$C(\theta) \mapsto C(\theta) + \kappa C_0,$$

where $\kappa > 0$ is a parameter to choose, which scales the effect of the covariance inflation. In the previous section, we have seen that for EKI the sample covariance is acting as a preconditioner of a gradient flow structured dynamic. If the ensemble size is less than the size of parameter space, which is in fact a problem in high dimensions, the preconditioned gradient flow structure can not ensure a strictly decreasing direction w.r.t. the potential Φ, since the sample covariance $C(\theta)$ might not be strictly positive definit. Following the ideas in [200], we propose to inflate the preconditioning effect through the sample covariance artificially. In the linear deterministic EKI setting, i.e. the system of coupled ODEs (3.12), we can inflate the effect of the covariance by additive covariance inflation which leads to the modified gradient flow structure

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta^{(j)} = -(C(\theta) + \kappa B)\nabla_\theta \Phi(\theta^{(j)}, y),$$

for some self-adjoint, positive definite operator $B : \mathcal{X} \to \mathcal{X}$.
While in the linear setting, due to the gradient flow structure, it is straightforward to introduce the covariance inflation into EKI, it is not the case for nonlinear forward problems. In the previous section we have seen, that the EKI can be viewed as gradient flow approximatively. Assuming that H is Fréchet differentiable, we consider the following approximation based on the Taylor expansion

$$H(\theta^{(j)}) - H(\bar{\theta}) \approx DH(\bar{\theta})(\theta^{(j)} - \bar{\theta}),$$

where $DH(\bar{\theta})$ denotes the Fréchet derivative of H at $\bar{\theta}$. Hence, we will approximate the mixed sample covariance $C^{\theta p}(\theta)$ by

$$C^{\theta p}(\theta) \approx \frac{1}{J}\sum_{j=1}^{J}(\theta^{(j)} - \bar{\theta}) \otimes (DH(\bar{\theta})(\theta^{(j)} - \bar{\theta})) = C(u)DH^*(\bar{\theta}).$$

Here, we have introduced a further approximation $\bar{H} \approx H(\bar{\theta})$. We will incorporate covariance inflation in sense of

$$C^{\theta p}(\theta) \approx C(\theta)DH^*(\bar{\theta}) \mapsto (C(\theta) + \kappa B)DH^*(\bar{\theta}),$$

and we will write the EKI with covariance inflation through the following approximation

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta^{(j)} = \Big(C^{\theta p}(\theta) + \kappa B\ DH^*(\bar{\theta})\Big)\,\Gamma^{-1}(y - H(\theta^{(j)})). \tag{3.15}$$

By application of this method of variance inflation one should mention that there exists the disadvantage of computing the derivative of the nonlinear map H.

3.3 Well-posedness of the ensemble Kalman inversion

In the previous section, we have introduced the continuous-time limit of the EKI, which is given through a system of coupled SDEs (3.9), which takes values in an infinite-dimensional space $\mathcal{X}$. As we want to study the behaviour of solutions to these infinite-dimensional SDEs, our aim is to break everything down w.l.o.g. to the finite-dimensional case. Therefore, we will first extend the subspace property from the discrete version Lemma 3.1.2 to a continuous version of it. With the help of this subspace property we will be able to work in a coordinate system of this initialized subspace and consider w.l.o.g. a finite-dimensional parameter space $\mathcal{X}$.
Using the definition of the empirical covariance, (3.9) can be formulated equivalently as

$$\mathrm{d}\theta_t^{(j)} = \frac{1}{J}\sum_{k=1}^{J}\left\langle H(\theta_t^{(k)}) - \bar{H}_t, (y - H(\theta_t^{(j)}))\,\mathrm{d}t + \sqrt{\Gamma}\,\mathrm{d}W_t^{(j)}\right\rangle_{\Gamma}(\theta_t^{(k)} - \bar{\theta}_t). \tag{3.16}$$

The formulation (3.16) reveals that solutions satisfy a generalization of the subspace property of Lemma 3.1.2 to continuous-time.

Lemma 3.3.1. *Assume that H is locally Lipschitz and let $\mathcal{S}$ be the linear span of $\{\theta_0^{(j)}\}_{j=1}^J$, then $\theta_t^{(j)} \in \mathcal{S}$ for all $(t,j) \in [0,\infty)\times\{1,\dots,J\}$ almost surely.*

We rely on the subspace property of Lemma 3.3.1, and first show that we can reduce the infinite-dimensional $\mathcal{X}$-valued setting without loss of generality to a finite-dimensional setting.

Lemma 3.3.2. *Without loss of generality we assume that the initial ensemble $(\theta_0^{(j)})_{j\in\{1,\dots,J\}}$ is linearly independent almost surely and spans a J-dimensional vector space $\mathcal{S}$. Furthermore, we assume that $H(\cdot) = L\cdot$ for some $L \in \mathcal{L}(\mathcal{X}, \mathbb{R}^K)$.*
Then there exists a linear operator $\tilde{L}: \mathbb{R}^J \to \mathbb{R}^K$ such that equation (3.11) restricted to $\mathcal{S}$ is equivalent to

$$\mathrm{d}v_t^{(j)} = \frac{1}{J}\sum_{k=1}^{J}\left\langle \tilde{L}v_t^{(k)} - \tilde{L}\bar{v}_t, (y - \tilde{L}v_t^{(j)})\,\mathrm{d}t + \Gamma^{\frac{1}{2}}\,\mathrm{d}W_t^{(j)}\right\rangle_{\Gamma}(v_t^{(k)} - \bar{v}_t) \tag{3.17}$$

for $v_t^{(j)} \in \mathbb{R}^J$, $\bar{v}_t := \frac{1}{J}\sum_{k=1}^{J} v_t^{(k)}$, in the following sense: For $\theta_t^{(j)} = \sum_{k=1}^{J}(v_t^{(j)})_k \cdot \theta_0^{(k)}$ one has that θ_t is a $\mathcal{S}$-valued solution of (3.11) if and only if v_t is a solution of (3.17).

Proof. By Lemma 3.3.1, any $\mathcal{S}$-valued process $u(t)$ can be uniquely expanded as a linear combination $\theta^{(j)}(t) = \sum_{l=1}^{J} v_l^{(j)}(t)\cdot\theta^{(l)}(0)$ for every $j \in \{1,\dots,J\}$, $t \geq 0$ and coordinates $v_l^{(j)}(t) \in \mathbb{R}$. Let $\Psi^{-1}: \mathbb{R}^J \to \mathcal{S}$ denote the basis isomorphism, i.e. $\Psi: \mathcal{S} \to \mathbb{R}^J$ with $u = \sum_{l=1}^{J} v_l\theta^{(l)}(0) \overset{\Psi}{\mapsto} (v_1,\dots,v_J)^\top$. Since Ψ is a linear isomorphism, (3.11) can be equivalently transformed to

$$\begin{aligned}\mathrm{d}\Psi(\theta^{(j)}(t)) &= \Psi(\mathrm{d}\theta^{(j)}(t))\\ &= \frac{1}{J}\sum_{k=1}^{J}\langle L(\theta^{(k)}(t) - \bar{\theta}(t)), (y - L\theta^{(j)}(t))\,\mathrm{d}t + \Gamma^{\frac{1}{2}}\,\mathrm{d}W_t^{(j)}\rangle_{\Gamma}(\Psi(\theta^{(k)}(t)) - \Psi(\bar{\theta}(t)))\,.\end{aligned}$$

Thus, with $\tilde{L} = L\Psi^{-1}$, we obtain

$$\begin{aligned} \mathrm{d}\Psi(\theta^{(j)}(t)) =& \frac{1}{J}\sum_{k=1}^{J}\langle \tilde{L}\Psi(\theta^{(k)}(t) - \bar{\theta}(t)), (y - \tilde{L}\Psi(\theta^{(j)}(t)))\,\mathrm{d}t + \Gamma^{\frac{1}{2}}\,\mathrm{d}W_t^{(j)}\rangle_\Gamma \\ & \cdot (\Psi(\theta^{(k)}(t)) - \Psi(\bar{\theta}(t))) \end{aligned}$$

The assertion follows with $v^{(j)} := \Psi(\theta^{(j)})$. □

Remark 3.3.3. *We note that Lemma 3.3.2 can be generalized to the nonlinear setting, by introduction of of $\tilde{H}(\cdot) = (H \circ \Psi^{-1})(\cdot)$, which is again nonlinear. Hence, also in the nonlinear setting, the original system of infinite dimensional SDEs can be breaked down to a system of finite dimensional SDEs with a similar structure as* (3.17).

3.3.1 Well-posedness result - Linear setting

The following section is devoted to prove existence and uniqueness of global solutions of the set of coupled SDEs (3.11). The local existence and uniqueness of $\mathcal{X}$-valued local solutions to (3.11) is straightforward by the local Lipschitz-property of the drift and diffusion on the right-hand side.
In this part, we will assume that the forward response operator is linear, i.e. $H(\cdot) = L\cdot$ with $L \in \mathcal{L}(\mathcal{X}, \mathbb{R}^K)$. Then recall, that the continuous-time limit (3.16) reads as

$$\mathrm{d}\theta_t^{(j)} = \frac{1}{J}\sum_{k=1}^{J}\left\langle L(\theta_t^{(k)} - \bar{\theta}_t), (y - L\theta_t^{(j)})\,\mathrm{d}t + \sqrt{\Gamma}\,\mathrm{d}W_t^{(j)}\right\rangle_\Gamma (\theta_t^{(k)} - \bar{\theta}_t).$$

Recall that the empirical covariance operator is defined by

$$C(\theta) = \frac{1}{J}\sum_{k=1}^{J}(\theta^{(k)} - \bar{\theta}) \otimes (\theta^{(k)} - \bar{\theta}).$$

and equation (3.10) can be rewritten in the form

$$\mathrm{d}\theta_t^{(j)} = C(\theta_t)L^*\Gamma^{-1}(y - L\theta_t^{(j)})\,\mathrm{d}t + C(\theta_t)L^*\Gamma^{-1/2}\,\mathrm{d}W_t^{(j)}.$$

Remark 3.3.4. *By Lemma 3.3.2 we can imply that solving equation* (3.11) *is equivalent to solving the finite-dimensional equation* (3.17). *Thus, to simplify notation we will assume w.l.o.g. that $\mathcal{X} = \mathbb{R}^I$, $I \in \mathbb{N}$, $I \leq J$. Further, in the case of linearly independent initial ensemble we can assume $I = J$.*
When studying the dynamical behavior of the ensemble, we will sometimes require the following assumption for the extension of our results to the parameters space:

$$\textit{The linear operator } \tilde{L} \textit{ defined above is one-to-one.} \tag{3.18}$$

However, we note that Assumption (3.18) *seems to be a rather strict assumption: It requires that the forward operator "sees everything" and secondly, this means that*

$$\{\tilde{L}\Psi(u_0^{(j)})\}_{j=1}^{J} \subset \mathbb{R}^K$$

is linearly independent. This implies the restriction on the number of particles $J \leq K$.

However, note that this assumption is on the operator $\tilde{L}$, i.e. we do not assume that L is one-to-one. The discretization of the parameter space via the ensemble of particles acts as a regularization of the inverse problem in this setting. We will need assumption (3.18) *only when we want to prove dynamical properties in the parameter space. This makes sense as we cannot hope for convergence to the true parameter if the forward operator is indifferent with respect to some components of this parameter value. Our convergence results in the observation space hold without assumption* (3.18)*.*

In order to prove the existence and uniqueness of global solutions we rewrite the set of coupled SDEs (3.11) as a single SDE of the following form:

$$d\theta_t = F(\theta_t)\,dt + G(\theta_t)\,dW_t,$$

with $\theta_t = (\theta_t^{(j)})_{j\in\{1,\dots,J\}} \in \mathbb{R}^{IJ\times 1}, W_t = (W_t^{(j)})_{j\in\{1,\dots,J\}} \in \mathbb{R}^{J^2\times 1}$ and

$$\begin{aligned} F(x) &= (C(x)L^*\Gamma^{-1}(y - Lx^{(j)}))_{j\in\{1,\dots,J\}} \in \mathbb{R}^{IJ\times 1},\\ G(x) &= \mathrm{diag}(C(x)L^*\Gamma^{-\frac{1}{2}})_{j\in\{1,\dots,J\}} \in \mathbb{R}^{IJ\times J^2}, \end{aligned}$$

where $x = (x^{(j)})_{j\in\{1,\dots,J\}} \in \mathbb{R}^{IJ\times 1}$ and $\mathrm{diag}(B_j)_{j\in\{1,\dots,J\}}$ is a diagonal block matrix with matrices $(B_j)_{j\in\{1,\dots,J\}}$ on the diagonal.
Before formulating the well-posedness result of the EKI, we will prove the following auxiliary result, which we will need several times when studying the system of SDEs.

Lemma 3.3.5. *Let M be a symmetric and nonnegative $d\times d$-matrix, then for all choices of vectors $(z^{(k)})_{k=1,\dots,J}$ in $\mathbb{R}^d$ we have*

$$\sum_{k,l=1}^{J} \langle z^{(k)}, z^{(l)}\rangle \langle z^{(k)}, Mz^{(l)}\rangle \geq 0.$$

Proof. Let $(v^{(m)})_{m=1,\dots,d}$ be an orthonormal basis of eigenvectors such that $Mv^{(m)} = \lambda_m v^{(m)}$ with $\lambda_m \geq 0$. Then $z^{(l)} = \sum_{m=1}^{d} z_m^{(l)} v^{(m)}$ and thus

$$\sum_{k,l=1}^{J} \langle z^{(k)}, z^{(l)}\rangle \langle z^{(k)}, Mz^{(l)}\rangle = \sum_{k,l=1}^{J}\sum_{m,n=1}^{d} z_n^{(k)} z_n^{(l)} z_m^{(k)} z_m^{(l)} \lambda_m = \sum_{n,m=1}^{d} \lambda_m \Big(\sum_{k=1}^{J} z_n^{(k)} z_m^{(k)}\Big)^2 \geq 0.$$

□

We will now formulate and prove the main result of this section on the well-posedness of the EKI.

Theorem 3.3.6. *Let $\theta_0 = (\theta_0^{(j)})_{j\in\{1,\dots,J\}}$ be $\mathcal{F}_0$-measurable maps $\theta_0^{(j)} : \Omega \to \mathcal{X}$ which are linearly independent almost surely. Then for all $T \geq 0$ there exists a unique strong solution $(\theta_t)_{t\in[0,T]}$ (up to $\mathbb{P}$-indistinguishability) of the set of coupled SDEs* (3.11)*.*

Proof. For the proof we will assume without loss of generality that $\mathcal{X} = \mathbb{R}^I$ as discussed before. Due to the local Lipschitz property of the drift F and the diffusion G, we can ensure through standard arguments existence and uniqueness of local strong solutions for (3.11) (up to a stopping time). Therefore, we note that both F and G are polynomials.

The global existence of a strong solution is based on stochastic Lyapunov theory. See for example Theorem 3.5 of [137]. We just need to construct a function $V \in C^2(\mathcal{X}; \mathbb{R}_+)$ such that for some constant $c > 0$

$$\mathfrak{L}V(x) := \nabla V(x) \cdot F(x) + \frac{1}{2}\operatorname{trace}(G^T(x)\mathrm{Hess}[V](x)G(x)) \leq cV(x) \tag{3.19}$$

and

$$\inf_{|x|>R} V(x) \to \infty \text{ as } \mathbb{R} \to \infty \tag{3.20}$$

hold true.
We can uniquely decompose $y \in \mathbb{R}^K$ as $y = y_1 + y_2$, with $y_1 \in \mathcal{R}(\Gamma^{-\frac{1}{2}}L)$ and $y_2 \in \mathcal{R}(\Gamma^{-\frac{1}{2}}L)^\perp$, where $\mathcal{R}(\Gamma^{-\frac{1}{2}}L)$ denotes the image of $\Gamma^{-\frac{1}{2}}L$. We fix $\tilde{\theta} \in \mathbb{R}^J$ such that $\Gamma^{-\frac{1}{2}}L\tilde{\theta} = y_1$ and define the Lyapunov function

$$V(\theta) := 2V_1(\theta) + V_2(\theta) + \|\Gamma^{-1/2}L\|_{\mathcal{F}}^2 = \frac{2}{J}\sum_{j=1}^{J}\|\theta^{(j)} - \bar{\theta}\|^2 + \varphi(\|\bar{\theta} - \tilde{\theta}\|^2) + \|\Gamma^{-1/2}L\|_{\mathcal{F}}^2,$$

where we define $\varphi(z) = \log(1+z)$. First note that for $z > 0$ it holds true that $\varphi'(z) \leq 1$, $z\varphi'(z) \leq 1$ and $|z\varphi''(z)| \leq 1$. Obviously, (3.20) is satisfied. The generator $\mathfrak{L}$ applied to V is given by $\mathfrak{L}V = \mathfrak{L}V_1 + \mathfrak{L}V_2$ with

$$\begin{aligned}
\mathfrak{L}V_1(\theta) =& -\frac{J+1}{J^3}\sum_{j,l=1}^{J}\langle\theta^{(j)} - \bar{\theta}, \theta^{(l)} - \bar{\theta}\rangle\langle\Gamma^{-\frac{1}{2}}L(\theta^{(l)} - \bar{\theta}), \Gamma^{-\frac{1}{2}}L(\theta^{(j)} - \bar{\theta})\rangle \\
\mathfrak{L}V_2(\theta) =& -\varphi'(\|\bar{\theta} - \tilde{\theta}\|^2)\frac{2}{J}\sum_{l=1}^{J}\langle\bar{\theta} - \tilde{\theta}, \theta^{(l)} - \bar{\theta}\rangle\langle\Gamma^{-\frac{1}{2}}L(\theta^{(l)} - \bar{\theta}), \Gamma^{-\frac{1}{2}}L(\bar{\theta} - \tilde{\theta}) - y_2\rangle \\
&+ \varphi'(\|\bar{\theta} - \tilde{\theta}\|^2)\frac{1}{J^3}\sum_{j,l=1}^{J}\langle\theta^{(j)} - \bar{\theta}, \theta^{(l)} - \bar{\theta}\rangle\langle\Gamma^{-\frac{1}{2}}L(\theta^{(l)} - \bar{\theta}), \Gamma^{-\frac{1}{2}}L(\theta^{(j)} - \bar{\theta})\rangle \\
&+ \varphi''(\|\bar{\theta} - \tilde{\theta}\|^2)\langle\bar{\theta} - \tilde{\theta}, C(\theta)L^\top LC(\theta)(\bar{\theta} - \tilde{\theta})\rangle \\
=& -\varphi'(\|\bar{\theta} - \tilde{\theta}\|^2)\frac{2}{J}\sum_{l=1}^{J}\langle\bar{\theta} - \tilde{\theta}, \theta^{(l)} - \bar{\theta}\rangle\langle\Gamma^{-\frac{1}{2}}L(\theta^{(l)} - \bar{\theta}), \Gamma^{-\frac{1}{2}}L(\bar{\theta} - \tilde{\theta})\rangle \\
&+ \varphi'(\|\bar{\theta} - \tilde{\theta}\|^2)\frac{1}{J^3}\sum_{j,l=1}^{J}\langle\theta^{(j)} - \bar{\theta}, \theta^{(l)} - \bar{\theta}\rangle\langle\Gamma^{-\frac{1}{2}}L(\theta^{(l)} - \bar{\theta}), \Gamma^{-\frac{1}{2}}L(\theta^{(j)} - \bar{\theta})\rangle \\
&+ \varphi''(\|\bar{\theta} - \tilde{\theta}\|^2)\langle\bar{\theta} - \tilde{\theta}, C(\theta)L^\top LC(\theta)(\bar{\theta} - \tilde{\theta})\rangle,
\end{aligned}$$

where we used $\langle\Gamma^{-\frac{1}{2}}L(\theta^{(l)} - \bar{\theta}), y_2\rangle = 0$ for all $l \in \{1, \ldots, J\}$ which is true by construction. We can bound the generator by application of Cauchy-Schwarz inequality and Young's inequality

$$\begin{aligned}
&\varphi'(\|\bar{\theta} - \tilde{\theta}\|^2)\frac{2}{J}\sum_{l=1}^{J}\langle\bar{\theta} - \tilde{\theta}, \theta^{(l)} - \bar{\theta}\rangle\langle\Gamma^{-\frac{1}{2}}L(\theta^{(l)} - \bar{\theta}), \Gamma^{-\frac{1}{2}}L(\bar{\theta} - \tilde{\theta})\rangle \\
&\qquad = 2\varphi'(\|\bar{\theta} - \tilde{\theta}\|^2)\langle\bar{\theta} - \tilde{\theta}, C(\theta)L^\top\Gamma^{-1}L(\bar{\theta} - \tilde{\theta})\rangle
\end{aligned}$$

$$\leq \left(\varphi'(\|\bar{\theta}-\tilde{\theta}\|^2)\|\bar{\theta}-\tilde{\theta}\|^2\|\Gamma^{-1/2}L\|_{\mathcal{F}}\right)^2 + \|C(\theta)L^\top\Gamma^{-\frac{1}{2}}\|_{\mathcal{F}}^2$$
$$\leq \|\Gamma^{-1/2}L\|_{\mathcal{F}}^2 + \|C(\theta)L^\top\Gamma^{-\frac{1}{2}}\|_{\mathcal{F}}^2,$$

and

$$\varphi''(\|\bar{\theta}-\tilde{\theta}\|^2)\langle\bar{\theta}-\tilde{\theta}, C(\theta)L^\top LC(\theta)(\bar{\theta}-\tilde{\theta})\rangle$$
$$\leq |\varphi''(\|\bar{\theta}-\tilde{\theta}\|^2)|\|\bar{\theta}-\tilde{\theta}\|^2\|C(\theta)L^\top\Gamma^{-\frac{1}{2}}\|_{\mathcal{F}}^2$$
$$\leq \|C(\theta)L^\top\Gamma^{-\frac{1}{2}}\|_{\mathcal{F}}^2.$$

Further, we note that

$$\|C(\theta)L^\top\Gamma^{-\frac{1}{2}}\|_{\mathcal{F}}^2 = \frac{1}{J^2}\sum_{j,l=1}^{J}\langle\theta^{(j)}-\bar{\theta}, \theta^{(l)}-\bar{\theta}\rangle\langle\Gamma^{-\frac{1}{2}}L(\theta^{(l)}-\bar{\theta}), \Gamma^{-\frac{1}{2}}L(\theta^{(j)}-\bar{\theta})\rangle.$$

Thus, as $L^\top\Gamma^{-1}L$ is a symmetric non-negative matrix, we can show that the generator satisfies (3.19)

$$\mathfrak{L}V(\theta) = 2\mathfrak{L}V_1(\theta) + \mathfrak{L}V_2(\theta)$$
$$\leq -\frac{2(J+1)}{J}\|C(\theta)L^\top\Gamma^{-\frac{1}{2}}\|_{\mathcal{F}}^2 + \|\Gamma^{-1/2}L\|_{\mathcal{F}}^2 + 2\|C(\theta)L^\top\Gamma^{-\frac{1}{2}}\|_{\mathcal{F}}^2$$
$$\leq V(\theta).$$

□

3.4 Convergence analysis of the ensemble Kalman inversion - Linear setting

In this section, we study the so called ensemble collapse and the convergence of the residuals. The idea is to split the convergence results into two parts. The first part is to prove that the spread of the ensemble of particles stays bounded and that the particles are converging to their joint mean. The second part is to study the behaviour of the mean of the ensemble itself, whether it converges to a good estimate of our underlying inverse problem.
Recall, that we consider a true underlying parameter $\theta^\dagger \in \mathcal{X}$ which constructs our observations, i.e.

$$y = L\theta^\dagger + \xi^\dagger,$$

where $\xi^\dagger \sim \mathcal{N}(0,\Gamma)$ is a realization of the measurements noise. We introduce the following quantities we are going to analyse:

	parameter space	**observation space**
spread	$e^{(j)} = \theta^{(j)} - \bar{\theta}$	$\mathfrak{e}^{(j)} = \Gamma^{1/2}(L\theta^{(j)} - L\bar{\theta})$
residual	$r^{(j)} = \theta^{(j)} - \theta^\dagger$	$\mathfrak{r}^{(j)} = \Gamma^{1/2}(L\theta^{(j)} - y)$

The **spread** e describes the difference of each particle to the ensembles mean and the **residual** r describes the difference of each particle to the underlying truth $\theta^\dagger$. As our forward model L is in general mapping to a lower dimensional space, we can not expect to prove convergence results in the parameter space. We aim to study the quantities mapped by our forward model L, i.e. we consider the spread of the mapped particles $\mathfrak{e}$ and the data misfit $\mathfrak{r}$. For simplicity in our computations, we scale both quantities $\mathfrak{e}$ and $\mathfrak{r}$ by the symmetric and positive definite matrix $\Gamma^{1/2}$.
We can describe the introduced quantities by the SDEs

$$\begin{aligned} \mathrm{d}e_t^{(j)} &= -C(e_t)L^*\Gamma^{-1}Le_t^{(j)}\,\mathrm{d}t + C(e_t)L^*\Gamma^{-\frac{1}{2}}\,\mathrm{d}(W_t^{(j)} - \bar{W}_t), \\ \mathrm{d}r_t^{(j)} &= \mathrm{d}\theta_t^{(j)} = C(\theta_t)L^*\Gamma^{-1}(y - L\theta_t^{(j)})\,\mathrm{d}t + C(\theta_t)L^*\Gamma^{-1/2}\,\mathrm{d}W_t^{(j)}, \end{aligned} \tag{3.21}$$

with $\bar{W}_t := \frac{1}{J}\sum_{j=1}^{J} W^{(j)}$.
The dynamical behavior of the empirical mean is given by

$$\mathrm{d}\bar{\theta}_t = \frac{1}{J}\sum_{k=1}^{J}(\theta_t^{(k)} - \bar{\theta}_t)\langle L(\theta_t^{(k)} - \bar{u}_t), (y - L\bar{\theta}_t)\,\mathrm{d}t + \Gamma^{\frac{1}{2}}\,\mathrm{d}\bar{W}_t\rangle_\Gamma.$$

3.4.1 Quantification of the ensemble collapse

We quantify the ensemble collapse, which means the convergence of the ensemble spread towards zero, in the parameter as well as in the observation space in L^p and almost surely sense.
For the computation of the dynamics describing the processes $\mathfrak{e}$ and $\mathfrak{r}$, we will apply several times Itô's formula. In order to use Itô's formula we have to calculate the following quadratic covariation in many cases:

Lemma 3.4.1. *Let $(W^{(j)})_{j=1,\dots,J}$ be independent Brownian motions in $\mathbb{R}^K$, $u, v \in \mathbb{R}^K$ and let $l \neq j \in \{1, \dots, J\}$. Then with $\bar{W} = \frac{1}{J}\sum_{k=1}^{J} W^{(k)}$,*

$$\langle u, d(W^{(j)} - \bar{W})\rangle\langle v, d(W^{(j)} - \bar{W})\rangle = \frac{J-1}{J}\langle u, v\rangle\, dt,$$

$$\langle u, d(W^{(j)} - \bar{W})\rangle\langle v, d(W^{(l)} - \bar{W})\rangle = -\frac{1}{J}\langle u, v\rangle\, dt.$$

Proof. We can write

$$W^{(j)} - \bar{W} = -\frac{1}{J}\sum_{k=1, k\neq j}^{J} W^{(k)} + \frac{J-1}{J}W^{(j)},$$

and, since $W^{(k)}$ are independent Brownian motions, it follows

$$\begin{aligned} \langle u, d(W^{(j)} - \bar{W})\rangle\langle v, d(W^{(j)} - \bar{W})\rangle &= \frac{1}{J^2}\sum_{k=1, k\neq j}^{J}\langle u, \mathrm{d}W^{(k)}\rangle\langle v, \mathrm{d}W^{(k)}\rangle \\ &\quad + \frac{(J-1)^2}{J^2}\langle u, \mathrm{d}W^{(j)}\rangle\langle v, \mathrm{d}W^{(j)}\rangle \\ &= \frac{J-1}{J}\langle u, v\rangle\,\mathrm{d}t. \end{aligned}$$

Similarly, we obtain

$$\begin{aligned}\langle u, d(W^{(j)} - \bar{W})\rangle\langle v, d(W^{(l)} - \bar{W})\rangle &= -\frac{1}{J}\sum_{k=1}^{J}(\langle u, \mathrm{d}W^{(j)}\rangle\langle v, \mathrm{d}W^{(k)}\rangle + \langle u, \mathrm{d}W^{(k)}\rangle\langle v, \mathrm{d}W^{(l)}\rangle) \\ &\quad + \frac{1}{J^2}\sum_{i,k=1}^{J}\langle u, \mathrm{d}W^{(i)}\rangle\langle v, \mathrm{d}W^{(k)}\rangle \\ &= -\frac{1}{J}\langle u, v\rangle\,\mathrm{d}t.\end{aligned}$$

□

Ensemble collapse in the observation space - Auxiliary results:

We begin with our first auxiliary result, which states that the L^p norm stays bounded in time for p depending on the ensemble size J.

Lemma 3.4.2. *Let $p \in [2, J+3)$ and $\theta_0 = (\theta_0^{(j)})_{j\in\{1,\dots,J\}}$ be $\mathcal{F}_0$-measurable maps $\theta_0^{(j)} : \Omega \to \mathcal{X}$ such that $\mathbb{E}[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_0^{(j)}|^p] < \infty$. Then*

$$t \in [0,\infty) \mapsto \|\mathfrak{e}_t\|_{\mathcal{L}_p(\Omega,\mathbb{R}^K)} := \mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_t^{(j)}|^p\right]^{\frac{1}{p}}$$

is monotonically decreasing in t. Furthermore there exists a constant $C > 0$ such that for all $t \geq 0$

$$\int_0^t \mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_s^{(j)}|^{p+2}\right]\mathrm{d}s < C.$$

Proof for $p = 2$. We will prove the assertion in the case $p = 2$, in order to give the key ideas. The case $p > 2$ is very similar, but much more technical.
Applying $\Gamma^{-\frac{1}{2}}L$ to $e^{(j)}$ implies that the quantity $\mathfrak{e}^{(j)}$ satisfies (see (3.21) and (3.3))

$$\begin{aligned}\mathrm{d}\mathfrak{e}_t^{(j)} &= -C(\mathfrak{e}_t)\mathfrak{e}_t^{(j)}\,\mathrm{d}t + C(\mathfrak{e}_t)\,\mathrm{d}(W_t^{(j)} - \bar{W}_t) \\ &= -\frac{1}{J}\sum_{k=1}^{J}\mathfrak{e}_t^{(k)}\langle\mathfrak{e}_t^{(k)}, \mathfrak{e}_t^{(j)}\rangle\,\mathrm{d}t + \frac{1}{J}\sum_{k=1}^{J}\mathfrak{e}_t^{(k)}\langle\mathfrak{e}_t^{(k)}, \mathrm{d}(W_t^{(j)} - \bar{W}_t)\rangle.\end{aligned}$$

Itô's formula gives

$$\begin{aligned}\mathrm{d}|\mathfrak{e}_t^{(j)}|^2 &= 2\langle\mathfrak{e}_t^{(j)}, \mathrm{d}\mathfrak{e}_t^{(j)}\rangle + \langle\mathrm{d}\mathfrak{e}_t^{(j)}, \mathrm{d}\mathfrak{e}_t^{(j)}\rangle \\ &= -\frac{2}{J}\sum_{k=1}^{J}\left\langle\mathfrak{e}_t^{(j)}, \mathfrak{e}_t^{(k)}\right\rangle^2\mathrm{d}t + 2\mathfrak{e}_t^{(j)T}C(\mathfrak{e}_t)\,\mathrm{d}(W_t^{(j)} - \bar{W}_t) \\ &\quad + \frac{1}{J^2}\sum_{k,l=1}^{J}\left\langle\mathfrak{e}_t^{(k)}, \mathfrak{e}_t^{(l)}\right\rangle\left\langle\mathfrak{e}_t^{(k)}, \mathrm{d}(W_t^{(j)} - \bar{W}_t)\right\rangle\left\langle\mathfrak{e}_t^{(l)}, \mathrm{d}(W_t^{(j)} - \bar{W}_t)\right\rangle\end{aligned}$$

and with Lemma 3.4.1 to evaluate the Itô correction we get

$$\mathrm{d}|\mathfrak{e}_t^{(j)}|^2 = -\frac{2}{J}\sum_{k=1}^{J}\langle \mathfrak{e}_t^{(j)}, \mathfrak{e}_t^{(k)}\rangle^2 \,\mathrm{d}t + 2\mathfrak{e}_t^{(j)T} C(\mathfrak{e}_t)\,\mathrm{d}(W_t^{(j)} - \bar{W}_t) + \frac{J-1}{J^3}\sum_{k,l=1}^{J}\langle \mathfrak{e}_t^{(k)}, \mathfrak{e}_t^{(l)}\rangle^2 \,\mathrm{d}t\,.$$

Taking the sum over all particles leads to

$$\begin{aligned}
\mathrm{d}\left(\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_t^{(j)}|^2\right) &= -\frac{2}{J^2}\sum_{j,k=1}^{J}\langle \mathfrak{e}_t^{(j)}, \mathfrak{e}_t^{(k)}\rangle^2 \,\mathrm{d}t + \frac{2}{J}\sum_{j=1}^{J}\mathfrak{e}_t^{(j)\top} C(\mathfrak{e}_t)\,\mathrm{d}(W_t^{(j)} - \bar{W}_t) \\
&\quad + \frac{J-1}{J^3}\sum_{j,k=1}^{J}\langle \mathfrak{e}_t^{(j)}, \mathfrak{e}_t^{(k)}\rangle^2 \,\mathrm{d}t \\
&= -\frac{J+1}{J^3}\sum_{j,k=1}^{J}\left\langle \mathfrak{e}_t^{(j)}, \mathfrak{e}_t^{(k)}\right\rangle^2 \,\mathrm{d}t + \frac{2}{J}\sum_{j=1}^{J}\mathfrak{e}_t^{(j)\top} C(\mathfrak{e}_t)\,\mathrm{d}(W_t^{(j)} - \bar{W}_t) \\
&= -\frac{J+1}{J^3}\sum_{j,k=1}^{J}\left\langle \mathfrak{e}_t^{(j)}, \mathfrak{e}_t^{(k)}\right\rangle^2 \,\mathrm{d}t + \frac{2}{J}\sum_{j=1}^{J}\mathfrak{e}_t^{(j)\top} C(\mathfrak{e}_t)\,\mathrm{d}W_t^{(j)}.
\end{aligned}$$

The last step follows from $\sum_j \mathfrak{e}^{(j)} = 0$. This yields

$$\begin{aligned}
&\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_t^{(j)}|^2 - \frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_0^{(j)}|^2 \\
&\qquad = -\frac{J+1}{J^3}\int_0^t\sum_{j,k=1}^{J}\langle \mathfrak{e}_t^{(j)}, \mathfrak{e}_t^{(k)}\rangle^2 \,\mathrm{d}t + \frac{2}{J}\int_0^t\sum_{j=1}^{J}\mathfrak{e}_t^{(j)\top} C(\mathfrak{e}_t)\,\mathrm{d}W_t^{(j)}.
\end{aligned} \tag{3.22}$$

Note that we cannot simply take the expectation, as we do not know if the stochastic integral is a martingale. We introduce a localization, where we set $t, s \geq 0$ and let $(\tau_n)_{n\in\mathbb{N}}$ with $\tau_n \overset{n}{\to} \infty$ a.s. be a sequence of deterministically bounded stopping times, such that

$$\int_s^{s+(t\wedge\tau_n)} \mathfrak{e}_s^{(j)T} C(\mathfrak{e}_s)\,\mathrm{d}W_s^{(j)}$$

is a martingale for every $j \in \{1, \cdots, J\}$. This is possible by definition of local martingales, with any stochastic integral being one. For example we can take for τ_n the minimum of n and the first exit time of $\mathfrak{e}_s$ at radius n. Then, for all $n \in \mathbb{N}$, from (3.22) (after rebasing the integration interval from $[0,t]$ to $[s, s+t]$) we obtain

$$\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_{s+(t\wedge\tau_n)}^{(j)}|^2\right] - \mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_s^{(j)}|^2\right] = -\mathbb{E}\left[\int_s^{s+(t\wedge\tau_n)}\frac{J+1}{J^3}\sum_{j,k=1}^{J}\langle \mathfrak{e}_r^{(j)}, \mathfrak{e}_r^{(k)}\rangle^2 \,\mathrm{d}r\right]$$

As $\tau_n \to \infty$, applying Fatou's lemma on the left hand side and applying the monotone convergence theorem on the right hand side gives

$$\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_{s+t}^{(j)}|^2\right] - \mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_s^{(j)}|^2\right] \leq -\mathbb{E}\left[\int_s^{s+t}\frac{J+1}{J^3}\sum_{j,k=1}^{J}\langle \mathfrak{e}_r^{(j)}, \mathfrak{e}_r^{(k)}\rangle^2 \,\mathrm{d}r\right] \leq 0, \tag{3.23}$$

which implies that $\mathbb{E}[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_t^{(j)}|^2]$ is monotonically decreasing in t.
Finally,

$$\int_0^t \mathbb{E}\left[\frac{J+1}{J^3}\sum_{j=1}^{J}|\mathfrak{e}_s^{(j)}|^4\right]\mathrm{d}s \le \int_0^t \mathbb{E}\left[\frac{J+1}{J^3}\sum_{j,k=1}^{J}\langle\mathfrak{e}_s^{(j)},\mathfrak{e}_s^{(k)}\rangle^2\right]\mathrm{d}s \le \mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_0^{(j)}|^2\right],$$

where the first inequality is trivial by inserting non-negative terms in the sum and the second inequality is (3.23) with $s=0$. This proves the second claim.
Let us finally remark that $\tau_n \to \infty$ necessarily holds. If we assume that $\tau_n \to \tau_*$ then the previous argument with $s=0$ and arbitrary $T>0$ gives $\mathbb{E}[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_{t\wedge\tau_*}^{(j)}|^2] < \infty$. Thus, it would follow $t < \tau_*$ for our choice of stopping time. □

Proof for $p>2$. Since $\mathfrak{e}^{(j)} \in \mathbb{R}^K$, note that componentwise

$$\mathrm{d}\mathfrak{e}_m^{(j)} = -\frac{1}{J}\sum_{l=1}^{J}\mathfrak{e}_m^{(l)}\langle\mathfrak{e}^{(l)},\mathfrak{e}^{(j)}\rangle\mathrm{d}t + \frac{1}{J}\sum_{l=1}^{J}\mathfrak{e}_m^{(l)}\langle\mathfrak{e}^{(l)},\mathrm{d}(W^{(j)}-\bar{W})\rangle.$$

We define the Lyapunov function (for equivalent notions of "p-norms" of the ensemble, see Lemma 3.4.3)

$$V_p(\mathfrak{e}) = \frac{1}{J}\sum_{m=1}^{K}(\sum_{j=1}^{J}|\mathfrak{e}_m^{(j)}|^2)^{\frac{p}{2}}$$

and according to Ito's lemma it holds that

$$\mathrm{d}V_p(\mathfrak{e}) = \sum_{m=1}^{K}\sum_{j=1}^{J}\frac{\partial V_p}{\partial\mathfrak{e}_m^{(j)}}\mathrm{d}\mathfrak{e}_m^{(j)} + \frac{1}{2}\sum_{m,m'=1}^{K}\sum_{j,j'=1}^{J}\mathrm{d}\mathfrak{e}_m^{(j)}\frac{\partial^2 V_p}{\partial\mathfrak{e}_m^{(j)}\partial\mathfrak{e}_{m'}^{(j')}}\mathrm{d}\mathfrak{e}_{m'}^{(j')}$$

Analogously to the case $p=2$ the expectation is given by

$$\begin{aligned}\mathbb{E}[V_p(\mathfrak{e}_{s+t})] &= \mathbb{E}[V_p(\mathfrak{e}_s)]\\ &\quad - C(p,J)\mathbb{E}[\int_s^{s+t}\sum_{m=1}^{K}[\{(\sum_{k=1}^{J}|\mathfrak{e}_m^{(k)}|^2)^{\frac{p}{2}-1}\}[\sum_{n=1}^{K}(\sum_{l=1}^{J}\mathfrak{e}_m^{(l)}\mathfrak{e}_n^{(l)})^2]]\,\mathrm{d}r]\\ &\quad + \mathbb{E}[\int_0^t\frac{p}{J^2}\sum_{m=1}^{K}((\sum_{k=1}^{J}|\mathfrak{e}_m^{(k)}|^2)^{\frac{p}{2}-1}\sum_{j,l=1}^{J}\mathfrak{e}_m^{(l)}\mathfrak{e}_m^{(j)}\langle\mathfrak{e}^{(l)},\mathrm{d}(W^{(j)}-\frac{1}{J}\sum_{r=1}^{J}W^{(r)})\rangle)]\end{aligned} \tag{3.24}$$

by defining $C(p,J) := \frac{p}{J^2}(1-\frac{(p-2+J)\cdot(J-1)}{2J^2}-\frac{p-2}{2J^2}) > 0$.
Thus, similarly to Lemma 3.4.2 we obtain by introducing stopping times and using Fatou's Lemma

$$\begin{aligned}&\mathbb{E}[V_p(\mathfrak{e}_{s+t})] - \mathbb{E}[V_p(\mathfrak{e}_s)]\\ &\qquad \le -C(p,J)\mathbb{E}[\int_s^{s+t}\sum_{m=1}^{K}[\{(\sum_{k=1}^{J}|\mathfrak{e}_m^{(k)}|^2)^{\frac{p}{2}-1}\}[\sum_{n=1}^{K}(\sum_{l=1}^{J}\mathfrak{e}_m^{(l)}\mathfrak{e}_n^{(l)})^2]]\,\mathrm{d}r] \le 0,\end{aligned}$$

and setting $s = 0$ leads to

$$\mathbb{E}[V_p(\mathfrak{e}_0)] \geq C(p,J)\mathbb{E}[\int_0^t \sum_{m=1}^{K}[\{(\sum_{k=1}^{J}|\mathfrak{e}_m^{(k)}|^2)^{\frac{p}{2}-1}\}[\sum_{n=1}^{K}(\sum_{l=1}^{J}\mathfrak{e}_m^{(l)}\mathfrak{e}_n^{(l)})^2]]\,\mathrm{d}s].$$

Note that

$$\mathbb{E}[\int_0^t \sum_{m=1}^{K}[\{(\sum_{k=1}^{J}|\mathfrak{e}_m^{(k)}|^2)^{\frac{p}{2}-1}\}[\sum_{n=1}^{K}(\sum_{l=1}^{J}\mathfrak{e}_m^{(l)}\mathfrak{e}_n^{(l)})^2]]\,\mathrm{d}s] < C$$

Now we bound the integrand by below by:

$$\sum_{m=1}^{K}((\sum_{k=1}^{J}|\mathfrak{e}_m^{(k)}|^2)^{\frac{p}{2}-1})(\sum_{n=1}^{K}(\sum_{l=1}^{J}\mathfrak{e}_m^{(l)}\mathfrak{e}_n^{(l)})^2) \geq \sum_{m=1}^{K}(\sum_{k=1}^{J}|\mathfrak{e}_m^{(k)}|^2)^{\frac{p}{2}+1} = JV_{p+2}(\mathfrak{e}),$$

Thus, we also have

$$\mathbb{E}[\int_0^t V_{p+2}(\mathfrak{e}_s)\,\mathrm{d}s] < C \tag{3.25}$$

for all $p < J + 3$.

□

The following result states the equivalent notions of the previous introduced Lyapunov function V_p and the "p-norms" of the ensemble of particles.

Lemma 3.4.3. *For $a_{m,j} \in \mathbb{R}$, $m = 1,\dots,d$, $j = 1,\dots,J$ and $p \in \mathbb{N}$,*

$$\sum_{j=1}^{J}(\sum_{m=1}^{d}|a_{m,j}|^2)^{\frac{p}{2}} \leq d^{(p-1)/2}\sum_{m=1}^{d}\sum_{j=1}^{J}|a_{m,j}|^p$$

and

$$\sum_{m=1}^{d}\sum_{j=1}^{J}|a_{m,j}|^p \leq J^{p/2}\sum_{m=1}^{d}(\sum_{j=1}^{J}|a_{m,j}|^2)^{\frac{p}{2}}.$$

By symmetry we also have

$$\sum_{m=1}^{d}(\sum_{j=1}^{J}|a_{m,j}|^2)^{\frac{p}{2}} \leq J^{p/2}\sum_{m=1}^{d}\sum_{j=1}^{J}|a_{m,j}|^p \quad \text{and} \quad \sum_{m=1}^{d}\sum_{j=1}^{J}|a_{m,j}|^p \leq d^{(p-1)/2}\sum_{j=1}^{J}(\sum_{m=1}^{d}|a_{m,j}|^2)^{\frac{p}{2}}.$$

Proof. We start with the first claim and write

$$\sum_{j=1}^{J}(\sum_{m=1}^{d}|a_{m,j}|^2)^{\frac{p}{2}} = \sum_{j=1}^{J}T_j$$

with $T_j^2 = (\sum_{m=1}^{d}|a_{m,j}|^2)^p$. We continue by expressing T_j^2 using the multinomial theorem and Young's inequality

$$T_j^2 = \sum_{k_1+\cdots+k_d=p}\binom{p}{k_1,\dots,k_d}\prod_{m=1}^{d}|a_{m,j}|^{2k_m}$$

$$= \sum_{k_1+\cdots+k_d=p} \binom{p}{k_1,\ldots,k_d} \prod_{m=1,k_m\neq 0}^{d} |a_{m,j}|^{2k_m}$$
$$\leq \sum_{m=1}^{d} |a_{m,j}|^{2p} \sum_{l_1+\cdots+l_d=p-1} \binom{p-1}{l_1,\ldots,l_d} = \sum_{m=1}^{d} |a_{m,j}|^{2p} d^{p-1}.$$

This means that

$$\sum_{j=1}^{J} (\sum_{m=1}^{d} |a_{m,j}|^2)^{\frac{p}{2}} \leq d^{\frac{p-1}{2}} \sum_{j=1}^{J} \sqrt{\sum_{m=1}^{d} |a_{m,j}|^{2p}} \leq d^{\frac{p-1}{2}} \sum_{j=1}^{J} \sum_{m=1}^{d} |a_{m,j}|^p,$$

which proves the first statement. For the second claim we can write by concavity of the square root

$$\sum_{m=1}^{d} (\sum_{j=1}^{J} |a_{m,j}|^2)^{\frac{p}{2}} = \sum_{m=1}^{d} (\sqrt{J} \sqrt{\sum_{j=1}^{J} \frac{|a_{m,j}|^2}{J}})^p \geq J^{-\frac{p}{2}} \sum_{m=1}^{d} \sum_{j=1}^{J} |a_{m,j}|^p,$$

i.e.

$$\sum_{m=1}^{d} \sum_{j=1}^{J} |a_{m,j}|^p \leq J^{\frac{p}{2}} \sum_{m=1}^{d} (\sum_{j=1}^{J} |a_{m,j}|^2)^{\frac{p}{2}}.$$

□

Before proving the L^p convergence of the spread in the observation space, we state that the stochastic integral in (3.22) is indeed a martingale. This property gives the possibility to take directly the expectation in (3.22) and prove convergence in L^p.

Lemma 3.4.4. *For all $j \in \{1,\ldots,J\}$ the process*

$$(M(t))_{t\geq 0} := \Big(\int_0^t \mathfrak{e}_s^{(j)T} C(\mathfrak{e}_s) \, \mathrm{d}W_s^{(j)} \Big)_{t\geq 0}$$

is a (global) martingale.

Proof. The local martingale given by the stochastic integral is a true martingale by Itô-isometry if we show that following second moment is finite (cp.[84, Theorem 2.4])

$$\|\mathfrak{e}_\cdot^{(j)T} C(\mathfrak{e}_\cdot)\|_{\Lambda_2;T} := \mathbb{E}[\int_0^T \|\mathfrak{e}_s^{(j)T} C(\mathfrak{e}_s)\|_F^2 \, \mathrm{d}s] = \int_0^T \mathbb{E}[\|\mathfrak{e}_s^{(j)T} C(\mathfrak{e}_s)\|_F^2] \, \mathrm{d}s < \infty$$

for all $T \geq 0$. For this, we first estimate the Frobenius norm, denoted by $\|\cdot\|_F$, by

$$\|\mathfrak{e}_s^{(j)T} C(\mathfrak{e}_s)\|_F^2 := \operatorname{trace} \mathfrak{e}^{(j)T} C(\mathfrak{e})(\mathfrak{e}^{(j)T} C(\mathfrak{e}))^T = \frac{1}{J^2} \sum_{k,l=1}^{J} \langle \mathfrak{e}^{(l)}, \mathfrak{e}^{(k)} \rangle \langle \mathfrak{e}^{(j)}, \mathfrak{e}^{(k)} \rangle \langle \mathfrak{e}^{(l)}, \mathfrak{e}^{(j)} \rangle$$
$$\leq \frac{1}{J^2} \sum_{k,l=1}^{J} |\mathfrak{e}^{(l)}|^2 |\mathfrak{e}^{(j)}|^2 |\mathfrak{e}^{(k)}|^2$$

Thus, it holds true that

$$\frac{1}{J}\sum_{j=1}^{J}\|\mathfrak{e}_s^{(j)T}C(\mathfrak{e}_s)\|_F^2 \leq \frac{1}{J^3}\sum_{j,k,l=1}^{J}|\mathfrak{e}^{(l)}|^2|\mathfrak{e}^{(j)}|^2|\mathfrak{e}^{(k)}|^2 = (\frac{1}{J}\sum_{j=1}^{J}|e^{(j)}|^2)^3 \leq \frac{1}{J}\sum_{j=1}^{J}|e^{(j)}|^6$$

and with Lemma 3.4.2 it follows

$$\frac{1}{J}\sum_{j=1}^{J}\|\mathfrak{e}_\cdot^{(j)T}C(\mathfrak{e}_\cdot)\|_{\Lambda_2;T} \leq \int_0^T \mathbb{E}[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}^{(j)}|^6]\,\mathrm{d}s \leq C,$$

since $p+2 := 6 \leq J+4$. □

Analogously, we prove the martingale property for the following stochastic integral in the setting of $p > 2$ which arises in (3.24). The idea of the proof is similar to the setting for $p = 2$, but is getting more technically again.

Lemma 3.4.5. *For all $k \in \{1,\dots,J\}$ and $p \in (2, \frac{J+3}{2})$ the process*

$$(M(t))_{t\geq 0} := \left(\int_0^t \frac{p}{J^2}\sum_{m=1}^{K}((\sum_{k=1}^{J}|\mathfrak{e}_m^{(k)}|^2)^{\frac{p}{2}-1}\sum_{j,l=1}^{J}\mathfrak{e}_m^{(l)}\mathfrak{e}_m^{(j)})\mathfrak{e}^{(l)\top}\mathrm{d}W^{(k)}\right)$$

is a (global) martingale.

Proof. Similarly to the proof of Lemma 3.4.4 we estimate the Frobenius norm of the integrand by

$$\begin{aligned}\|\sum_{m=1}^{K}((\sum_{k=1}^{J}|\mathfrak{e}_m^{(k)}|^2)^{\frac{p}{2}-1}\sum_{j,l=1}^{J}\mathfrak{e}_m^{(l)}\mathfrak{e}_m^{(j)})\mathfrak{e}^{(l)\top}\|_F^2 &\leq C_1(J)\sum_{m=1}^{K}(\sum_{k=1}^{J}|\mathfrak{e}_m^{(k)}|^2)^{p-2}\sum_{j,l=1}^{J}(\mathfrak{e}_m^{(l)})^2(\mathfrak{e}_m^{(j)})^2|\mathfrak{e}^{(l)}|^2\\ &\leq C_2(J,K)\sum_{k=1}^{J}|\mathfrak{e}^{(k)}|^{2(p-2)}\sum_{j,l=1}^{J}|\mathfrak{e}^{(l)}|^4|\mathfrak{e}^{(j)}|^2\\ &\leq C_3(J,K)\sum_{l=1}^{J}|\mathfrak{e}^{(l)}|^{2p+2},\end{aligned}$$

where we have used Jensen's inequality and the fact $|\mathfrak{e}_m^{(j)}|^2 \leq \sum_{n=1}^{K}|\mathfrak{e}_n^{(j)}|^2 = |\mathfrak{e}^{(j)}|^2$. The assertion follows by the bound (3.25) in the proof of Lemma 3.4.2, which we obtained by localization and Fatou's Lemma without martingale property. □

Ensemble collapse in the observation space - Main results for L^2 convergence:

We are now ready to formulate our main result for the quantification of the ensemble collapse in L^p sense. We first formulate the result in the case for $p = 2$ to give more details on the main idea, and second give the more general result for $p > 2$, which is again more technically.

Theorem 3.4.6. *Let $\theta_0 = (\theta_0^{(j)})_{j\in\{1,\dots,J\}}$ be $\mathcal{F}_0$-measurable random variables $\theta_0^{(j)} : \Omega \to \mathcal{X}$ such that $C_0 := \mathbb{E}[\frac{1}{J}\sum_{j=1}^{J} |\mathfrak{e}_0^{(j)}|^2] < \infty$. Then, the ensemble collapse is quantified by*

$$\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J} |\mathfrak{e}_t^{(j)}|^2\right] \leq \frac{1}{\frac{J+1}{J^2}t + \frac{1}{C_0}}. \tag{3.26}$$

Proof. By Lemma 3.4.4 we can directly take expectations in (3.22) to obtain

$$\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J} |\mathfrak{e}_t^{(j)}|^2\right] = \mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J} |\mathfrak{e}_0^{(j)}|^2\right] - \frac{J+1}{J^3}\int_0^t \mathbb{E}\left[\sum_{j,k=1}^{J} \langle \mathfrak{e}_s^{(j)}, \mathfrak{e}_s^{(k)}\rangle^2\right] \mathrm{d}s.$$

Note that by dropping the non-negative mixed terms $j \neq k$ and by using Jensen's and Young's inequality

$$\frac{J+1}{J^3}\mathbb{E}\left[\sum_{j,k=1}^{J} \langle \mathfrak{e}_s^{(j)}, \mathfrak{e}_s^{(k)}\rangle^2\right] \geq \frac{J+1}{J^2}\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J} |\mathfrak{e}_s^{(j)}|^2\right]^2.$$

Thus setting $t \mapsto h(t) := \mathbb{E}[\frac{1}{J}\sum_{j=1}^{J} |\mathfrak{e}_t^{(j)}|^2]$ we can write

$$h(t) = h(0) - \frac{J+1}{J^2}\int_0^t h^2(s)\,\mathrm{d}s - \int_0^t p(s)ds$$

for a non-negative function $p \geq 0$. Hence, we can differentiate to obtain the differential inequality

$$h' \leq -\frac{J+1}{J^2}h^2\,,$$

from which by a comparison argument for scalar ODE it follows that

$$h(t) = \mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J} |\mathfrak{e}_t^{(j)}|^2\right] \leq \frac{1}{\frac{J+1}{J^2}t + \frac{1}{h(0)}}\,.$$

□

Corollary 3.4.7. *Under the same assumptions as in Theorem 3.4.6 and under Assumption* (3.18) *it holds true that*

$$\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J} |e_t^{(j)}|^2\right] \leq \frac{1}{\sigma_{\min}}\frac{1}{\frac{J+1}{J^2}t + \frac{1}{C_0}},$$

where $\sigma_{\min}$ is the smallest eigenvalue of the positive definite operator $L^*\Gamma^{-1}L$.

Proof. The assertion follows directly from the inequality

$$|\mathfrak{e}^{(j)}|^2 = |\Gamma^{-\frac{1}{2}}Le^{(j)}|^2 = \langle e^{(j)}, L^*\Gamma^{-1}Le^{(j)}\rangle \geq \sigma_{\min}|e^{(j)}|^2,$$

since $L^*\Gamma^{-1}L$ is positive definite. □

Remark 3.4.8. *We note that the bound in* (3.26) *deteriorates with growing number of particles J, i.e. the result does not quantify the ensemble collapse in the large ensemble size limit. However, the presented analysis is tailored for fixed ensemble size and we will demonstrate in the numerical experiments that the derived bound* (3.26) *can be efficiently used to quantify the collapse in this setting.*

Ensemble collapse in the observation space - Main results for higher-order moments:

We state our next main result for the ensemble collapse in the sense of higher-order moments convergence, where the proofs are very similar to but technically more involved than the case $p = 2$.

Theorem 3.4.9. *Let $p \in (2, \frac{J+3}{2})$ and let $\theta_0 = (\theta_0^{(j)})_{j\in\{1,\dots,J\}}$ be $\mathcal{F}_0$-measurable maps $\theta_0^{(j)} : \Omega \to \mathcal{X}$ such that $\mathbb{E}[\frac{1}{J}\sum_{j=1}^{J} |\mathfrak{e}_0^{(j)}|^p] < \infty$. Then it holds true that*

$$\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_t^{(j)}|^p\right] \leq \frac{J^{\frac{p}{2}}}{\left(\frac{2}{p}C(p,J)K^{-\frac{2}{p}}J^{1-\frac{2}{p}}t + \left(K^{\frac{p-1}{2}}\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_0^{(j)}|^p\right]\right)^{-\frac{2}{p}}\right)^{\frac{p}{2}}}$$

with $C(p,J) := \frac{p}{J^2}(1 - \frac{(p-2+J)\cdot(J-1)}{2J^2} - \frac{p-2}{2J^2}) > 0$.

Proof. The proof based on Itô's formula and a comparison principle for ODEs is very similar to the case $p = 2$. By (3.24) we get that $\mathbb{E}[V_p(\mathfrak{e}_t)]$ is monotonically decreasing and it follows

$$\mathbb{E}[V_p(\mathfrak{e}_t)] \leq \mathbb{E}[V_p(\mathfrak{e}_0)] - C(p,J)J\int_0^t \mathbb{E}[V_{p+2}(\mathfrak{e}_s)]\,ds.$$

By Jensen's inequality it follows

$$V_{p+2}(\mathfrak{e}) = \frac{1}{J}\sum_{m=1}^{K}(\sum_{j=1}^{J}|\mathfrak{e}_m^{(j)}|^2)^{\frac{p}{2}\frac{p+2}{p}} \geq K^{-\frac{2}{p}}J^{-\frac{2}{p}}(V_p(\mathfrak{e}))^{\frac{p+2}{p}}$$

and we obtain

$$\mathbb{E}[V_p(\mathfrak{e}_t)] \leq \mathbb{E}[V_p(\mathfrak{e}_0)] - C(p,J)J^{1-\frac{2}{p}}K^{-\frac{2}{p}}\int_0^t \mathbb{E}[V_p(\mathfrak{e}_s)]^{\frac{p+2}{p}}\,ds.$$

Similarly to the proof of Theorem 3.4.6 we get

$$h' \leq -C(p,J)J^{1-\frac{2}{p}}K^{-\frac{2}{p}}h^{\frac{p+2}{p}},$$

by defining $h(t) := \mathbb{E}[V_p(\mathfrak{e}_t)]$, from which it follows that

$$h(t) \leq (\frac{2}{p}C(p,J)K^{-\frac{2}{p}}J^{1-\frac{2}{p}}t + (h(0))^{-\frac{2}{p}})^{-\frac{p}{2}}.$$

Finally, we conclude with

$$\mathbb{E}[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_t^{(j)}|^p] \leq J^{\frac{p}{2}}(\frac{2}{p}C(p,J)K^{-\frac{2}{p}}J^{1-\frac{2}{p}}t + (K^{\frac{p-1}{2}}\mathbb{E}[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_0^{(j)}|^p])^{-\frac{2}{p}})^{-\frac{p}{2}}$$

by using Lemma 3.4.3.

□

Remark 3.4.10. *Note that a larger ensemble seems to regularize the dynamics. The higher the ensemble number J, the larger is the highest moment of ensemble collapse we can bound.*
The restriction $2p < J + 3$ comes from the fact that we need the martingale property of the stochastic integral, which we obtain from the bounds in Lemma 3.4.2.

Corollary 3.4.11. *Under the same assumptions as in Theorem* (3.4.9) *and under Assumption* (3.18) *it holds true that*

$$\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J}|e_t^{(j)}|^p\right] \leq \frac{J^{\frac{p}{2}}}{\left(\sigma_{\min}\cdot\frac{2}{p}C(p,J)K^{-\frac{2}{p}}J^{1-\frac{2}{p}}t + \left(K^{\frac{p-1}{2}}\mathbb{E}\left[\frac{1}{J}\sum\limits_{j=1}^{J}|\mathfrak{e}_0^{(j)}|^p\right]\right)^{-\frac{2}{p}}\right)^{\frac{p}{2}}},$$

where $\sigma_{\min}$ is the smallest eigenvalue of the positive definite operator $L^*\Gamma^{-1}L$ and $C(p,J)$ is defined in Theorem 3.4.9.

Ensemble collapse in the observation space - Main results for almost sure convergence:

While we have considered the L^p convergence in the previous sections, we now focus on the almost sure setting. This means we aim to prove that the particle members are converging to its mean almost surely. The idea of the proofs for almost sure convergence is based on the application of stochastic Lyapunov theory.

Theorem 3.4.12. *Let $\theta_0 = (\theta_0^{(j)})_{j\in\{1,\dots,J\}}$ be $\mathcal{F}_0$-measurable maps $\theta_0^{(j)} : \Omega \to \mathcal{X}$ and $\gamma : \mathbb{R}_+ \to \mathbb{R}_+$ a positive, monotonically increasing and differentiable function such that $\int_0^\infty \frac{\gamma'(s)^2}{\gamma(s)}\,\mathrm{d}s < \infty$. Then the trivial solution of*

$$\mathrm{d}\mathfrak{e}_t^{(j)} = -C(\mathfrak{e}_t)\mathfrak{e}_t^{(j)}\mathrm{d}t + C(\mathfrak{e}_t)\mathrm{d}(W_t^{(j)} - \bar{W}_t) \tag{3.27}$$

is almost surely asymptotically stable with rate function $\rho(t) = (\gamma(t))^{-\frac{1}{2}}$. In particular, $(\mathfrak{e}_t^{(j)})_{j=1,\dots,J}$ converges to zero almost surely as $t \to \infty$.

For examples of γ see the remark below.

Proof. The idea of this proof is based on Theorem 4.6.2 in [159]. We define the stochastic Lyapunov function

$$V(\mathfrak{e},t) = \gamma(t)\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}^{(j)}|^2.$$

The generator applied to V fulfills

$$\mathfrak{L}V(\mathfrak{e},t) = \frac{\gamma'(t)}{J}\sum_{j=1}^{J}|\mathfrak{e}^{(j)}|^2 - \gamma(t)\frac{J+1}{J^3}\sum_{j,k=1}^{J}\langle\mathfrak{e}^{(k)},\mathfrak{e}^{(j)}\rangle^2$$

$$\leq \frac{\gamma'(t)}{J}\sum_{j=1}^{J}|\mathfrak{e}^{(j)}|^2 - \gamma(t)\frac{J+1}{J^3}\sum_{j=1}^{J}|\mathfrak{e}^{(j)}|^4.$$

We can maximize this expression w.r.t. $(|\mathfrak{e}^{(1)}|^2,\dots,|\mathfrak{e}^{(J)}|^2)$ and get the following bound for $\mathfrak{L}V$.

$$\mathfrak{L}V(\mathfrak{e},t) \leq \frac{\gamma'(t)^2}{\gamma(t)}\frac{1}{4}\frac{J^2}{J+1} =: \eta(t)$$

Since $\int_0^\infty \eta(t)\,\mathrm{d}t < \infty$, with Theorem 4.6.2 of [159] the trivial solution of (3.27) is almost surely asymptotically stable with rate function $\rho(t) = (\gamma(t))^{-\frac{1}{2}}$. □

Corollary 3.4.13. *Under the same assumptions as in Theorem 3.4.12 and assumption* (3.18) *it holds true that* $(e_t^{(j)})_{j=1,\dots,J}$ *converges to zero almost surely as* $t\to\infty$ *with rate function* $\rho(t) = (\gamma(t))^{-\frac{1}{2}}$.

Remark 3.4.14. *We give two examples of admissible* $\gamma(t)$*:*

- $\gamma(t) = (t+\varepsilon)^\alpha$ *for* $\alpha\in(0,1)$ *and* $\varepsilon>0$ *sufficiently small to obtain the rate function* $\rho(t) = \frac{1}{(t+\varepsilon)^{\frac{\alpha}{2}}}$.
- $\gamma(t) = (t+\varepsilon)\log(t+\varepsilon)^{-\alpha}$ *for arbitrarily small* $\alpha > \frac{1}{2}$ *and* $\varepsilon>0$ *to obtain the rate function* $\rho(t) = \frac{\log(t+\varepsilon)^{\frac{\alpha}{2}}}{(t+\varepsilon)^{\frac{1}{2}}}$

Ensemble collapse in the parameter space:

We can also prove some theoretical result for the ensemble spread without the strong assumption (3.18). The result states a monotone decreasing spread over time, but not the collapse. For the ensemble collapse we need (3.18), see also Corollary 3.4.11 and 3.4.13.

Proposition 3.4.15. *Let* $\theta_0 = (\theta_0^{(j)})_{j\in\{1,\dots,J\}}$ *be* $\mathcal{F}_0$*-measurable maps* $\theta_0^{(j)}:\Omega\to\mathcal{X}$ *such that* $\mathbb{E}[\frac{1}{J}\sum_{j=1}^{J}|e_0^{(j)}|^2] < \infty$. *Then it holds true that* $t\mapsto \mathbb{E}[\frac{1}{J}\sum_{j=1}^{J}|e_t^{(j)}|^2]^{\frac{1}{2}}$ *is monotonically decreasing for* $t\geq 0$.

Proof. Itô's formula leads to

$$\begin{aligned}
\mathrm{d}|e_t^{(j)}|^2 &= 2\langle e_t^{(j)}, \mathrm{d}e_t^{(j)}\rangle + \langle \mathrm{d}e_t^{(j)}, \mathrm{d}e_t^{(j)}\rangle \\
&= -\frac{2}{J}\sum_{k=1}^{J}\langle e_t^{(j)}, e_t^{(k)}\rangle\langle \Gamma^{-\frac{1}{2}}Le_t^{(k)}, \Gamma^{-\frac{1}{2}}Le_t^{(j)}\rangle\,\mathrm{d}t \\
&\quad + \frac{2}{J}\sum_{k=1}^{J}\langle e_t^{(j)}, e_t^{(k)}\rangle\langle \Gamma^{-\frac{1}{2}}Le_t^{(k)}, \mathrm{d}(W_t^{(j)} - \bar{W}_t)\rangle \\
&\quad + \frac{1}{J^2}\sum_{k,l=1}^{J}\frac{J-1}{J}\langle e_t^{(k)}, e_t^{(l)}\rangle\langle \Gamma^{-\frac{1}{2}}Le_t^{(k)}, \Gamma^{-\frac{1}{2}}Le_t^{(l)}\rangle\,\mathrm{d}t
\end{aligned}$$

and taking the mean over all particles $j\in\{1,\dots,J\}$ gives

$$\mathrm{d}(\frac{1}{J}\sum_{j=1}^{J}|e_t^{(j)}|^2) = -\frac{J+1}{J^3}\sum_{j,k=1}^{J}\langle e_t^{(k)}, e_t^{(j)}\rangle\langle \Gamma^{-\frac{1}{2}}Le_t^{(k)}, \Gamma^{-\frac{1}{2}}Le_t^{(j)}\rangle\,\mathrm{d}t$$

$$+\frac{2}{J^2}\sum_{k,j=1}^{J}\langle e_t^{(k)},e_t^{(j)}\rangle\langle\Gamma^{-\frac{1}{2}}Le_t^{(k)},\mathrm{d}(W_t^{(j)}-\bar{W}_t)\rangle.$$

Again, we do not know, whether the stochastic integral is a martingale, and we need again a localization. Consider as in Lemma 3.4.2 a sequence of stopping times $(\tau_n)_{n\in\mathbb{N}}$ with $\tau_n\to\infty$ a.s., such that

$$\int_0^{t\wedge\tau_n}\frac{2}{J^2}\sum_{k,j=1}^{J}\langle e_s^{(k)},e_s^{(j)}\rangle\langle\Gamma^{-\frac{1}{2}}Le_s^{(k)},\mathrm{d}(W_s^{(j)}-\bar{W}_s)\rangle$$

is a martingale. We obtain for all $n\in\mathbb{N}$

$$\mathbb{E}[\frac{1}{J}\sum_{j=1}^{J}|e_{t\wedge\tau_n}^{(j)}|^2]=\mathbb{E}[\frac{1}{J}\sum_{j=1}^{J}|e_0^{(j)}|^2]-\mathbb{E}[\int_0^{t\wedge\tau_n}\frac{J+1}{J^3}\sum_{j,k=1}^{J}\langle e_s^{(k)},e_s^{(j)}\rangle\langle\Gamma^{-\frac{1}{2}}Le_s^{(k)},\Gamma^{-\frac{1}{2}}Le_s^{(j)}\rangle\,\mathrm{d}s]$$

and hence, as we have the positivity of the integrand by Lemma 3.3.5, we obtain that $\mathbb{E}[\frac{1}{J}\sum_{j=1}^{J}|e_{t\wedge\tau_n}^{(j)}|^2]$ is monotonically decreasing and bounded. Analogously to the proof of Lemma 3.4.2, we can pass to the limit $n\to\infty$ by Fatou's lemma and the monotone convergence theorem. This implies for $t>s\geq 0$

$$\mathbb{E}[\frac{1}{J}\sum_{j=1}^{J}|e_{t+s}^{(j)}|^2]\leq\mathbb{E}[\frac{1}{J}\sum_{j=1}^{J}|e_s^{(j)}|^2]-\mathbb{E}[\int_s^{s+t}\frac{J+1}{J^3}\sum_{j,k=1}^{J}\langle e_r^{(k)},e_r^{(j)}\rangle\langle\Gamma^{-\frac{1}{2}}Le_r^{(k)},\Gamma^{-\frac{1}{2}}Le_r^{(j)}\rangle\,\mathrm{d}r]$$

In particular, it follows that $\mathbb{E}[\frac{1}{J}\sum_{j=1}^{J}|e_t^{(j)}|^2]$ is monotonically decreasing. □

3.4.2 Convergence to ground truth

While we have seen in the motivation of EKI as optimizer, that to overcome the issue of fitting the noise one has to include some regularization or stopping criterion in the case that the data y is perturbed by some measurements noise. However, we consider the assumption that y is the image of a truth $\theta^\dagger\in\mathcal{X}$ under L without noise and we are interested now in the analysis of the convergence to the truth. This means, we analyse the EKI as optimization method without care of the ill-posedness of the inverse problem itself. More details on the possibilities for consideration of noisy observations will be given in the later part of this work, see chapter 5.
Recall the equation for the residuals $\mathfrak{r}$

$$\mathrm{d}\mathfrak{r}_t^{(j)}=-C(\mathfrak{r}_t)\mathfrak{r}_t^{(j)}\,\mathrm{d}t+C(\mathfrak{r}_t)\,\mathrm{d}W_t^{(j)}\,. \tag{3.28}$$

The following properties can be shown for the residuals.

Proposition 3.4.16. *Let y be the image of a truth $\theta^\dagger\in\mathcal{X}$ under L and $\theta_0=(\theta_0^{(j)})_{j\in\{1,\dots,J\}}$ be $\mathcal{F}_0$-measurable maps $\theta_0^{(j)}:\Omega\to\mathcal{X}$ such that $\mathbb{E}[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{r}_0^{(j)}|^2]<\infty$. Then $\mathbb{E}[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{r}_t^{(j)}|^2]^{\frac{1}{2}}$ is monotonically decreasing.*

Proof. The assertion follows by similar arguments to the proof of Proposition 3.4.15. □

The aim is now to prove the convergence of the residuals towards zero. This means, we want that the particles fit the data. However, by looking at the dynamics of the residuals described through (3.28) we can see the problems of proving this result.
First we not that

$$\begin{aligned} C(\mathfrak{r}) &= \frac{1}{J}\sum_{k=1}^{J}\Big((L\theta^{(k)}-y)-(L\bar{\theta}-y)\Big)\Big((L\theta^{(k)}-y)-(L\bar{\theta}-y)\Big)^{\top} \\ &= \frac{1}{J}\sum_{k=1}^{J}(L\theta^{(k)}-L\bar{\theta})(L\theta^{(k)}-L\bar{\theta})^{\top} = C(\mathfrak{e}). \end{aligned}$$

In the previous part we have proved the convergence of the spread $\mathfrak{e}$ in several senses, including almost sure convergence and L^p convergence. If this convergence, now happens to fast, we can not expect that the residuals $\mathfrak{r}$ are still moving through the dynamics

$$\mathrm{d}\mathfrak{r}_t^{(j)} = -\underbrace{C(\mathfrak{e}_t)}_{\to 0}\mathfrak{r}_t^{(j)}\,\mathrm{d}t + \underbrace{C(\mathfrak{e}_t)}_{\to 0}\,\mathrm{d}W_t^{(j)}\,.$$

For example we could initialize the particle system with an ensemble of zero spread and will stay in the initial state for all of the time.
To give some light into this issue, we consider the following toy example of ODE

$$z'(t) = -t^{\alpha}z(t),$$

where the solution will only converge to zero if the rate of increase of t^{α} is low enough, in particular if $\alpha \leq 1$. We can view t^{α} as playing the role of ensemble collapse, described through $C(\mathfrak{e})$. This means on the one side we have to control the lower bound of the ensemble spread, such that the collapse does not happen to fast, but on the other side, we need the ensemble to collapse as every particle shell converge to the same true observation y.
To avoid this issue of "too fast" or "too slow" ensemble collapse, we introduce the so called variance inflation, which increases the preconditioning effect of the covariance operator artificially. The variance inflation can help to stabilize the convergence of the system and is often used in practice for this reason, see for example [80, 134]. See also Section 3.2.2 for more details.

Variance Inflation

In order to correct rank deficiencies of the empirical covariance operator $C(\mathfrak{r})$, we will use variance inflation in the following sense. Let $B \in \mathcal{L}(\mathbb{R}^K, \mathbb{R}^K)$ be some positive definite operator (for example the identity) and consider the equation

$$\mathrm{d}\mathfrak{r}_t^{(j)} = -\left(C(\mathfrak{r}_t) + \frac{1}{t^{\alpha}+R}B\right)\mathfrak{r}_t^{(j)}\,\mathrm{d}t + C(\mathfrak{r}_t)\,\mathrm{d}W_t^{(j)}, \quad \alpha \in (0,1), R > 0. \tag{3.29}$$

This modification gives convergence of the mapped residuals. For sufficiently small $\mathfrak{r}_t$, the new term will dominate, and for $\alpha \in (0,1)$ we then expect convergence to 0 at a rate faster than any polynomial. The question is now whether and when this asymptotic for small $\mathfrak{r}_t$ sets in.

Theorem 3.4.17. *Assume that y is the image of a truth $\theta^\dagger \in \mathcal{X}$ under L and let $\mathfrak{r}_0 = (\mathfrak{r}_0^{(j)})_{j\in\{1,\dots,J\}}$ be $\mathcal{F}_0$-measurable maps $\mathfrak{r}_0^{(j)} : \Omega \to \mathbb{R}^K$ such that $\mathbb{E}[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{r}_0^{(j)}|^2] < \infty$, $B \in \mathcal{L}(\mathbb{R}^K, \mathbb{R}^K)$ a positive definite operator and $(\mathfrak{r}_t^{(j)})_{t\geq 0, j=1,\dots,J}$ the solution of (3.29). Then for all $\beta > 0$ it holds true that $\mathbb{E}[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{r}_t^{(j)}|^2] \in \mathcal{O}(t^{-\beta})$ and $\mathbb{E}[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{r}_t^{(j)}|^2]$ is monotonically decreasing.*

Proof. Let $B \in \mathcal{L}(\mathbb{R}^K, \mathbb{R}^K)$ be a positive definite operator, $\alpha \in (0,1)$, $R > 0$ and assume, that the smallest eigenvalue of B is $\lambda_{\min} = c > 0$.
We derive an equation for $\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{r}_t^{(j)}|^2$ by using Itô's formula:

$$\begin{aligned}\mathrm{d}|\mathfrak{r}_t^{(j)}|^2 = &-2\left\langle \mathfrak{r}_t^{(j)}, \left(C(\mathfrak{r}_t) + \frac{1}{t^\alpha + R}B\right)\mathfrak{r}_t^{(j)}\right\rangle \mathrm{d}t + 2\langle \mathfrak{r}_t^{(j)}, C(\mathfrak{r}_t)\mathrm{d}W_t^{(j)}\rangle \\ &+ \frac{1}{J}\sum_{j=1}^{J}\left\langle \mathfrak{r}_t^{(k)} - \bar{\mathfrak{r}}_t, C(\mathfrak{r}_t)(\mathfrak{r}_t^{(k)} - \bar{\mathfrak{r}}_t)\right\rangle \mathrm{d}t.\end{aligned}$$

Taking the empirical mean over all particles yields

$$\begin{aligned}\mathrm{d}\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{r}_t^{(j)}|^2 = &-\frac{2}{J}\sum_{j=1}^{J}\left\langle \mathfrak{r}_t^{(j)}, \left(C(\mathfrak{r}_t) + \frac{1}{t^\alpha + R}B\right)\mathfrak{r}_t^{(j)}\right\rangle \mathrm{d}t + \frac{2}{J}\sum_{j=1}^{J}\langle \mathfrak{r}_t^{(j)}, C(\mathfrak{r}_t)\mathrm{d}W^{(j)}\rangle \\ &+ \frac{1}{J}\sum_{k=1}^{J}\langle \mathfrak{r}_t^{(k)} - \bar{\mathfrak{r}}_t, C(\mathfrak{r}_t)(\mathfrak{r}_t^{(k)} - \bar{\mathfrak{r}}_t)\rangle \,\mathrm{d}t\,.\end{aligned}$$

Thus, for all $t, s \geq 0$, it follows similarly to the proof of Lemma 3.4.2 that

$$\begin{aligned}\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{r}_{t+s}^{(j)}|^2\right] \leq\ &\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{r}_s^{(j)}|^2\right] - \frac{2}{J}\int_s^{s+t}\mathbb{E}\left[\sum_{j=1}^{J}\langle \mathfrak{r}_r^{(j)}, C(\mathfrak{r}_r)\mathfrak{r}_r^{(j)}\rangle\right]\mathrm{d}r \\ &- \frac{2}{J}\int_s^{s+t}\frac{1}{r^\alpha + R}\mathbb{E}\left[\sum_{j=1}^{J}\langle \mathfrak{r}_r^{(j)}, B\mathfrak{r}_r^{(j)}\rangle\right]\mathrm{d}r \\ &+ \frac{1}{J}\int_s^{s+t}\mathbb{E}\left[\sum_{j=1}^{J}\langle \mathfrak{r}_r^{(j)} - \bar{\mathfrak{r}}_r, C(\mathfrak{r}_r)(\mathfrak{r}_r^{(j)} - \bar{\mathfrak{r}}_r)\rangle\right]\mathrm{d}r \\ \leq\ &\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{r}_s^{(j)}|^2\right] - \frac{1}{J}\int_s^{s+t}\mathbb{E}\left[\sum_{j=1}^{J}\langle \mathfrak{r}_r^{(j)}, \left(C(\mathfrak{r}_r) + \frac{1}{r^\alpha + R}B\right)\mathfrak{r}_r^{(j)}\rangle\right]\mathrm{d}r\,,\end{aligned}$$

where we have used Lemma 3.3.5 and the non-negativity of B. This gives the monotonicity, as both the covariance $C(\mathfrak{r}_r)$ as well as B are non-negative matrices.
We can even improve the estimate to obtain the asymptotic rate. Consider $S(t) = \frac{1}{J}\sum_{j=1}^{J}|\mathfrak{r}_t^{(j)}|^2$, then

$$\mathrm{d}(t^\beta S(t)) = \beta t^{\beta-1}S(t)dt + t^\beta \mathrm{d}S(t)\,.$$

We use all the previous estimates for the terms in $\mathrm{d}S$ together with the non-negativity of the covariance matrix $C(\mathfrak{r}_t)$ and $B \geq \lambda_{\min} > 0$ to obtain

$$\begin{aligned} t^\beta \mathbb{E}S(t) &\leq \beta \int_0^t \tau^{\beta-1} \mathbb{E}S(\tau) d\tau - \frac{2}{J} \int_0^t \tau^\beta \frac{\lambda_{\min}}{\tau^\alpha + R} \mathbb{E}S(\tau) d\tau \\ &\leq \int_0^t \tau^{\beta-1} \Big[\beta - \frac{2\lambda_{\min}}{J} \frac{\tau}{\tau^\alpha + R}\Big] \mathbb{E}S(\tau) d\tau \, . \end{aligned}$$

There is a time $T > 0$ such that the integrand in the equation above is negative for all $t > T$ and thus using the monotonicity of $\mathbb{E}S(\tau)$ we obtain for all $t > T$

$$t^\beta \mathbb{E}S(t) \leq \int_0^T \tau^{\beta-1} \Big[\beta - \frac{2\lambda_{\min}}{J} \frac{\tau}{\tau^\alpha + R}\Big] d\tau \mathbb{E}S(0),$$

which gives the asymptotic rate $t^{-\beta}$ for $\mathbb{E}S(t)$. □

Remark 3.4.18. *In case of a positive semidefinite matrix B, the convergence of the residuals will then take place in the image space of the matrix B. We can generalize the proof straightforwardly to this setting by projections of the quantities to the corresponding subspace.*

We can also verify almost sure convergence faster than any polynomial rate.

Theorem 3.4.19. *Assume that y is the image of a truth $\theta^\dagger \in \mathcal{X}$ under L and let $\mathfrak{r}_0 = (\mathfrak{r}_0^{(j)})_{j\in\{1,\dots,J\}}$ be $\mathcal{F}_0$-measurable maps $\mathfrak{r}_0^{(j)} : \Omega \to \mathbb{R}^K$ and $B \in \mathcal{L}(\mathbb{R}^K, \mathbb{R}^K)$ a positive definite operator. Then the solution of (3.29) is almost surely asymptotically stable with rate function $\rho(t) = t^{-\frac{\beta}{2}}$ for all $\beta > 0$. In particular, $(\mathfrak{r}_t^{(j)})_{j=1,\dots,J}$ converges to zero almost surely as $t \to \infty$.*

Proof. We define the Lyapunov function

$$V(\mathfrak{r}, t) = t^\beta \frac{1}{J} \sum_{j=1}^J |\mathfrak{r}^{(j)}|^2$$

and obtain

$$\mathfrak{L}V(\mathfrak{r}, t) \leq \frac{\beta t^{\beta-1}}{J} \sum_{j=1}^J |\mathfrak{r}^{(j)}|^2 - t^\beta \frac{1}{J} \sum_{j=1}^J \langle \mathfrak{r}^{(j)}, \left(C(\mathfrak{r}) + \frac{1}{t^\alpha + R} B\right) \mathfrak{r}^{(j)} \rangle.$$

Thus,

$$\mathfrak{L}V(\mathfrak{r}, t) \leq \frac{1}{J} \sum_{j=1}^J |\mathfrak{r}^{(j)}|^2 \left(\beta - \frac{\lambda_{\min} t}{t^\alpha + R}\right) t^{\beta-1}.$$

There exists a $T > 0$ such that the bracket above is non-positive for all $t \geq T$ and we obtain

$$\int_0^\infty \mathfrak{L}V(\mathfrak{r}, t) \, \mathrm{d}t \leq \int_0^T LV(\mathfrak{r}, t) \, \mathrm{d}t.$$

Moreover, by neglecting the negative term in the bracket for $t \leq T$ we obtain

$$\mathbb{E}[\int_0^T \mathfrak{L}V(\mathfrak{r}_t, t) \, \mathrm{d}t] \leq \mathbb{E}[\int_0^T \beta s^{\beta-1} \frac{1}{J} \sum_{j=1}^J |\mathfrak{r}_s^{(j)}|^2 \, \mathrm{d}s] \leq \frac{T^\beta}{J} \mathbb{E}[\sum_{j=1}^J |\mathfrak{r}_0^{(j)}|^2] < \infty,$$

by using the monotonicity of the sum. Hence, $\int_0^\infty \mathfrak{L}V(\mathfrak{r}_t, t) \, \mathrm{d}t < \infty$ and we conclude with $\mathfrak{r}_t$ is almost surely asymptotically stable with rate function $\rho(t) = t^{-\frac{\beta}{2}}$. □

Remark 3.4.20. *Note that the convergence rate is faster than any polynomial rate. However, the proof reveals that the constant in the convergence result will grow w.r.t. the rate β and $\alpha \in (0,1)$, which is consistent with the numerical experiments presented in subsection 3.5.*

Our aim is to use variance inflation in the parameter space, such that we can apply Theorem 3.4.17. We will use variance inflation in the finite-dimensional system of SDEs of the coordinates in the parameter space.
Let $y \in L\mathcal{S}$ where $L\mathcal{S}$ is the linear span of $\{L\theta_0^{(1)}, \dots, L\theta^{(J)}\}$ and consider the equation

$$\mathrm{d}\theta_t^{(j)} = (C(\theta_t) + \frac{1}{t^\alpha + R} B) L^* \Gamma^{-1} (y - L\theta_t^{(j)}) \,\mathrm{d}t + C(\theta_t) L^* \Gamma^{-\frac{1}{2}} \,\mathrm{d}W_t^{(j)}, \tag{3.30}$$

$j = 1, \dots, J$, for B positive definite, $R > 0$ and $\alpha \in (0,1)$. Since $y \in L\mathcal{S}$, the subspace property still holds, i.e. $\theta_t^{(j)} \in \mathcal{S}$ for all $(t,j) \in [0,\infty) \times \{1, \dots, J\}$. The following result transfers the results of Theorem 3.4.17 to the parameter space:

Corollary 3.4.21. *Let $y \in L\mathcal{S}$ and assume that y is the image of a truth $\theta^\dagger \in \mathcal{X}$ under L, L^* is assumed to be one-to-one and let $(\theta_t^{(j)})_{t\geq 0, j=1,\dots,J}$ be the solution of* (3.30)*. Then*

1. $\lim_{t\to\infty} \mathbb{E}[\frac{1}{J} \sum_{j=1}^{J} |\mathfrak{e}_t^{(j)}|^2] = 0.$

2. $\lim_{t\to\infty} \mathbb{E}[\frac{1}{J} \sum_{j=1}^{J} |\mathfrak{r}_t^{(j)}|^2] = 0.$

3. $(\mathfrak{r}_t^{(j)})_{t\geq 0}$ *converges almost surely to zero with rate function $\rho(t) = t^{-\frac{\beta}{2}}$ for all $\beta > 0$.*

Proof. Let $R > 0$ and $\alpha \in (0,1)$ and observe

$$\mathrm{d}\mathfrak{r}_t^{(j)} = -(C(\mathfrak{r}_t) + \frac{1}{t^\alpha + R} \Gamma^{-\frac{1}{2}} LB(\Gamma^{-\frac{1}{2}} L)^*) \mathfrak{r}_t^{(j)} \,\mathrm{d}t + C(\mathfrak{r}_t) \,\mathrm{d}W_t^{(j)}.$$

Since $\Gamma^{-\frac{1}{2}} LB(\Gamma^{-\frac{1}{2}} L)^*$ is positive definite the second and third assertion follow directly from Theorem 3.4.17 and 3.4.19. The proof of the first assertion is similar to the proof of Theorem 3.4.6. □

3.5 Numerical results

We consider the problem of recovering the unknown data $\theta^\dagger$ from noisefree observations

$$y^\dagger = L(\theta^\dagger),$$

where $p = \mathcal{L}^{-1}(\theta)$ is again the solution of the one-dimensional elliptic equation

$$\begin{aligned} -\frac{\mathrm{d}^2 p}{\mathrm{d}x^2} + p &= \theta \quad \text{in } D := (0,\pi), \\ p &= 0 \quad \text{on } \partial D, \end{aligned} \tag{3.31}$$

see also (2.11) from subsection 2.1.3.

Recall that the forward response operator is defined by

$$L = \mathcal{O} \circ \mathcal{L}^{-1} \quad \text{with} \quad \mathcal{L} = -\frac{\mathrm{d}^2}{\mathrm{d}x^2} + \mathrm{id} \quad \text{on} \quad \mathcal{D}(\mathcal{L}) = H^2 \cap H_0^1,$$

with the operator $\mathcal{O}$ observing the dynamical system at $K = 2^4 - 1$ equispaced observation points $x_k = \frac{k}{2^4}$, $k = 1, \dots, K$. We approximate the forward problem (3.31) numerically on a uniform mesh with meshwidth $h = 2^{-8}$ by a finite element method with continuous, piecewise linear ansatz functions.
We choose the initial ensemble of particles based on the eigenvalue and eigenfunctions $\{\lambda_j, z_j\}_{j \in \mathbb{N}}$ of the covariance operator C_0, defined by $C_0 = \beta(\mathcal{L} - \mathrm{id})^{-1}$ for $\beta = 10$.
From the Bayesian perspective we may interpret this as prior distributed by $\mu_0 = \mathcal{N}(0, C_0)$. We set our j^{th} initial particle to $\theta^{(j)}(0) = \sqrt{\lambda_j}\zeta_j z_j$ with $\zeta_j \sim \mathcal{N}(0,1)$, i.e. we use the Karhunen-Loève expansion to generate draws from μ_0.
We use equation (3.8) as discretization of the EKI continuous-time limit

$$\mathrm{d}\theta_t^{(j)} = C(\theta_t)L^*\Gamma^{-1}(y - L\theta_t^{(j)})\,\mathrm{d}t + C(\theta_t)L^*\Gamma^{-\frac{1}{2}}\,\mathrm{d}W_t^{(j)}$$

for the following simulations.

Ensemble collapse We illustrate the results from section 3.4, in particular we verify the bounds on the ensemble collapse derived in Theorem 3.4.6 and in Theorem 3.4.9.

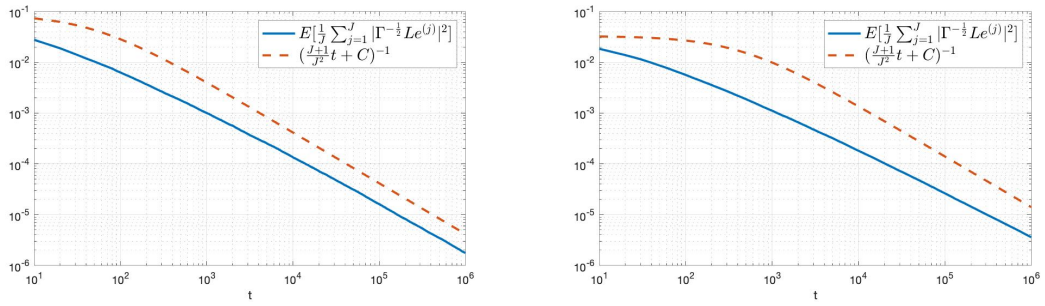

Figure 3.2: $\hat{E}(\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}^{(j)}(t)|^2)$ with w.r. of time. $Q = 1000$ paths with $J = 5$ (left) and $J = 15$ (right) particles has been simulated.

Figure 3.2 shows that the Monte Carlo approximation of the expected value $\hat{E}[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_t^{(j)}|^2]$ is bounded from above by $(\frac{J+1}{J^2}t + C)^{-1}$ with $C = (\hat{E}[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_0^{(j)}|^2])^{-1}$, as expected from Theorem 3.4.6.
Similarly Figure 3.3 demonstrates that the approximated higher moments $\hat{E}[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_t^{(j)}|^p]^{-\frac{1}{p}}$ are bounded by $J^{\frac{1}{2}}(\frac{2}{p}C(p,J)J^{1-\frac{2}{p}}K^{-\frac{2}{p}}t + C)^{-\frac{1}{2}}$ with $C = (K^{\frac{p-1}{2}}\hat{E}[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_0^{(j)}|^p])^{\frac{2}{p}}$, compare Theorem 3.4.9.
In order to verify the almost sure ensemble collapse numerically, we have simulated $Q = 10$ paths.

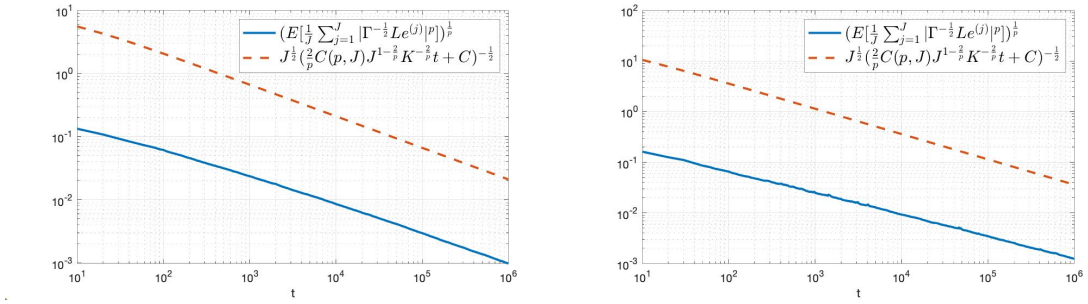

Figure 3.3: $\hat{E}(\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}^{(j)}(t)|^p)^{-\frac{1}{p}}$, $p=\lfloor\frac{J+3}{2}\rfloor-1$, w.r. of time. $Q=1000$ paths with $J=5$ (left) and $J=15$ (right) particles has been simulated.

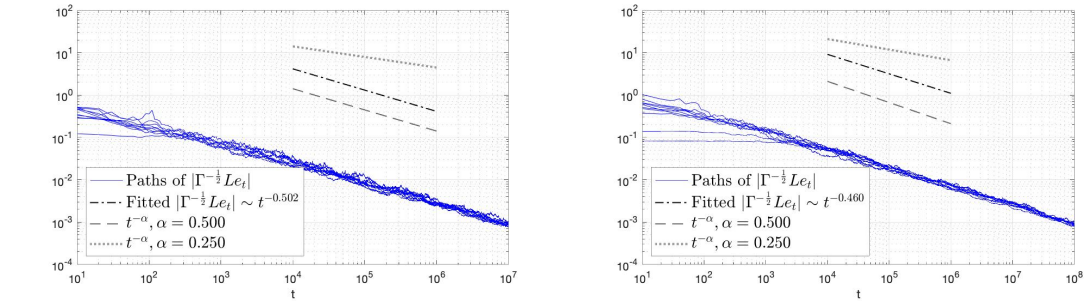

Figure 3.4: Paths of $|\mathfrak{e}(t)|_2$ w.r. of time. Q=10 paths with $J=5$ (left) and $J=15$ (right) particles has been simulated.

From Theorem 3.4.12 we know, that $\mathfrak{e}(t)$ converges almost surely to zero with rate function $\rho(t)=t^{-\frac{\alpha}{2}}$ for every $\alpha\in(0,1)$. Figure 3.4 illustrates this behavior, the expected convergence rates can be observed in this example.

Convergence to ground truth We compare simulations of the ensemble Kalman inversion without variance inflation with simulations of the ensemble Kalman inversion with variance inflation. The variance inflation is used in the following setting: We set $\alpha\in\{\frac{1}{2},\frac{3}{4}\}$ and $R=1$ in equation (3.30). The number of particles is $J=15$, i.e. the forward response operator is bijective as a mapping from the subspace spanned by the initial ensemble to the data space.

Figure 3.5 shows the differences of the EKI estimation in the parameter space as well as in the observation space. We observe that the simulations with variance inflation giving a better estimation in the observation space as well as in the parameter space. If we reduce the variance inflation in time faster, i.e. we increase the parameter α from $\frac{1}{2}$ to $\frac{3}{4}$, the effect of the variance inflation decreases. The following figures demonstrate the effect on the ensemble collapse and the residuals.

The idea of the variance inflation was to slow down the convergence of the particles to the ensemble mean, i.e. to control the rate of the ensemble collapse, in order to ensure the convergence of the residuals in the observation space. Figure 3.6 illustrates that we can ensure a higher spread of the ensemble in the simulations with variance inflation in comparison to the simulations without variance inflation in the observation space.

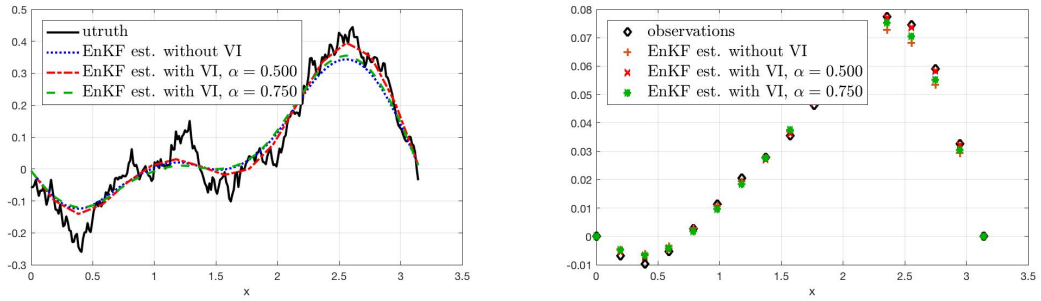

Figure 3.5: EKI estimation without VI vs. EKI estimation with VI. J=15 particles and Q=1000 paths has been simulated.

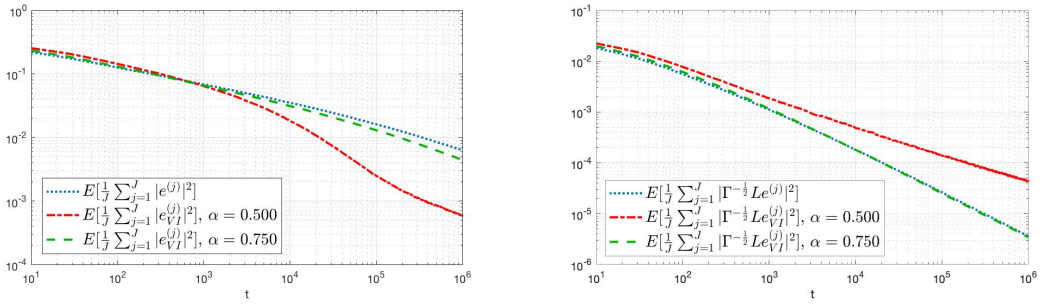

Figure 3.6: Comparison of the spread of the ensemble w.r. to time with VI and without VI.

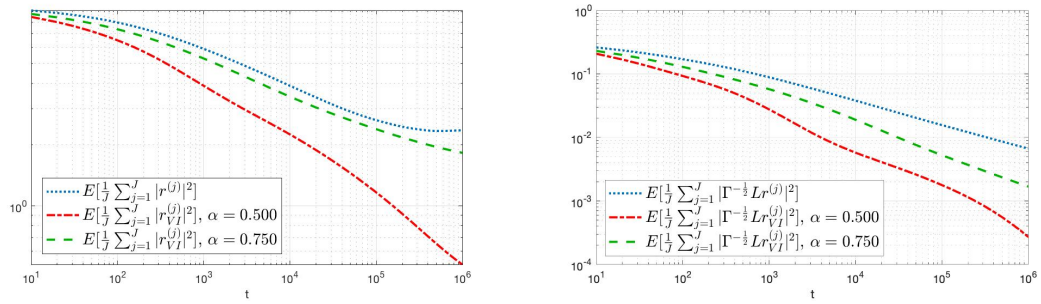

Figure 3.7: Comparison of the residuals w.r. to time with VI and without VI.

Figure 3.7 points out that we end up with convergence of the residuals in the observation and parameter space in case of variance inflation. Without variance inflation the simulations show a slight increase of the residuals in the parameter space, suggesting that the convergence of the residuals will slow down in the observation space as well.

To emphasize this result, we reduce the dimension of the example and we set $h = 2^4$ with $K = 3$ equispaced observation points. Furthermore, we set again $R = 1$ and $\alpha = \frac{1}{2}$ and we use $J = 3$ particles, such that the forward response operator is again bijective as mapping from the subspace spanned by the initial ensemble to the observation space.

Figure 3.8 shows again the difference of the EKI estimation with and without variance inflation. Figure 3.9 points out the effect of the variance inflation. While the residuals in the observation space without variance inflation diverge, we obtain convergence of the

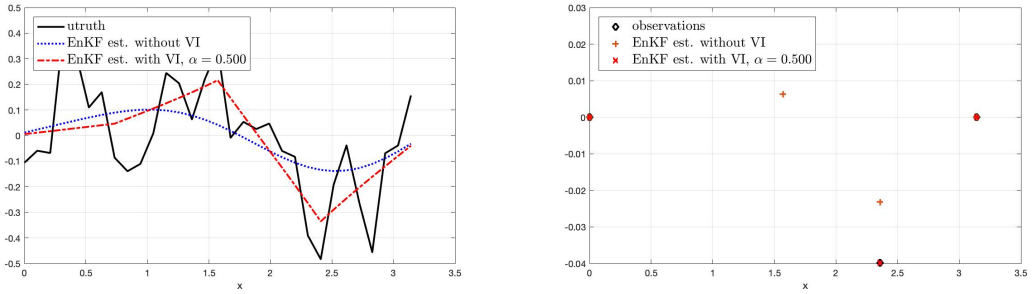

Figure 3.8: EnKF estimation without VI vs. EnKF estimation with VI. J=3 particles and Q=10000 paths has been simulated.

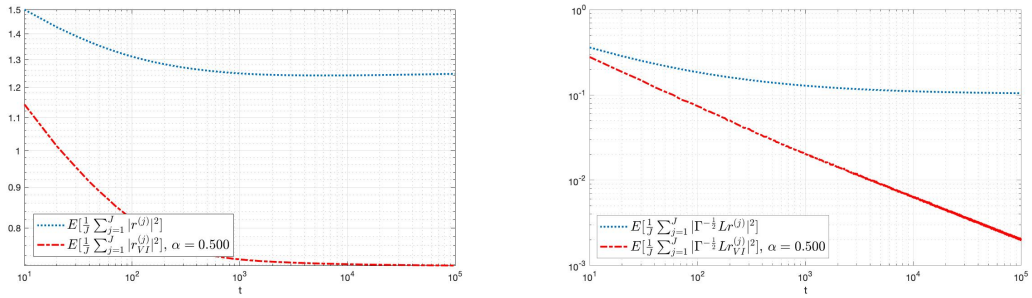

Figure 3.9: Comparison of the residuals w.r. to time with VI and without VI.

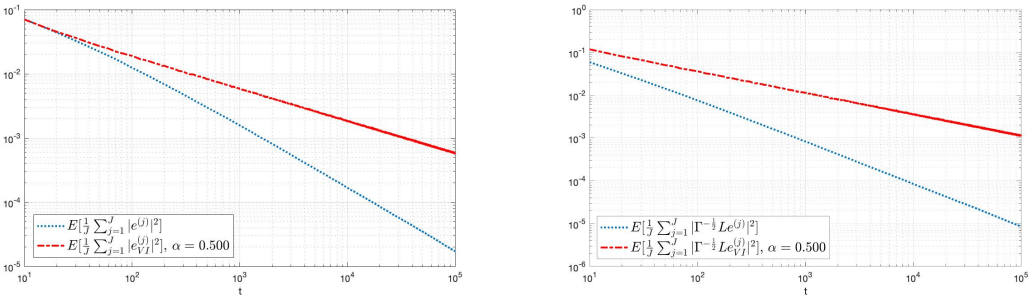

Figure 3.10: Comparison of the ensemble spread w.r. to time with VI and without VI.

residuals in the observation space using variance inflation. In addition, in Figure 3.10 we can see that the ensemble of particles still collapse in the parameter space as well as in the observation space.

4 Constrained ensemble Kalman inversion

In this chapter, we consider the incorporation of box constraints to the EKI method introduced in the previous section. The original EKI method does not allow for additional constraints on the parameter, arising for example from applications due to additional knowledge about the underlying system. Our focus lies in the incorporation of box-constraints, which can, for example, be applied in the context of hierarchical methods for EKI [40]. For these methods, the hyperparameters are often modelled through a uniform distribution. It is well-known that the application of EKI may lead to unfeasible estimates of the unknown parameters, i.e. does not satisfy the box constraints, which are obtained from the uniform distribution on the hierarchical parameters. As a result, we will provide the incorporation of constraints for the EKI. We refer to [4, 212] for a literature overview of existing methods for the treatment of linear and nonlinear constraints for Kalman-based methods. Common approaches are based on projection to the feasible set [128, 225], where a generalization to nonlinear constraints can be obtained through linearization ideas. Many variants are motivated by interpreting the Kalman-based updates as a solution of a corresponding optimization problem [4], which allows to contain straightforward constraints on the state and parameters. In the context of EKI, there has been a new approach to handle linear equality and inequality constraints for the EnKF and EKI by reparameterizing the solution of the optimization problem in the range of the covariance, i.e. by seeking the solution of the optimization problem in a subspace defined by the initial ensemble [3]. In [100], the authors introduce the EKI in order to solve equality constrained inverse problems, which are incorporated in the original least-squares loss functional. The update step results from the stationary point of the corresponding Lagrangian function. For our setting, we focus on the optimization viewpoint of the EKI and we introduce a projection-based method to the feasible set to incorporate box constraints. We modify the original EKI method in order to ensure that the estimate of the unknown parameter remains within a box of the parameter domain. The presented ideas build up on the theory of Bertsekas [19, 20] and others [202, 211]. We note that box constraints for inverse problems have been considered in a different context and for different applications [106].

We start with the formulation of the underlying box-constrained optimization problem in Section 4.1. In Section 4.2, we recall the projected gradient descent method and provide a convergence analysis based on its continuous-time limit. To do so, we make use of ideas inspired from barrier methods, which help to avoid arising discontinuities. In Section 4.2.1, we provide an example showing that preconditioning does not lead to a descent direction in general. Section 4.3 provides the formulation of the projected EKI method, which will be theoretically studied by its continuous-time formulation. We close the chapter with numerical results in Section 4.4

4.1 Formulation of the box-constrained optimization problem

Throughout this section we assume that the parameter space is finite-dimensional, i.e. $\mathcal{X} = \mathbb{R}^I$, the forward model is linear, i.e. let $H(\cdot) = L\cdot$, for $L \in \mathcal{L}(\mathcal{X}, \mathbb{R}^K)$ and that we have linear constraints for the underlying unknown parameter $\theta \in \mathbb{R}^I$, which are defined through the set

$$\mathcal{B}_c = \{\theta \in \mathbb{R}^I \mid \langle c_j, \theta\rangle + \delta_j \leq 0\},$$

for $c_j \in \mathbb{R}^I$ and $\delta_j \in \mathbb{R}$ for $j = 1, \ldots, m$. We define $f_j : \mathbb{R}^I \to \mathbb{R}$ with $f(\theta) = \langle c_j, \theta\rangle + \delta_j$ and write the set of constraints through

$$\mathcal{B}_c = \{\theta \in \mathbb{R}^I \mid f_j(\theta) \leq 0\}. \tag{4.1}$$

For given noisefree observation y, we aim to solve the inverse problem under constraints on the unknown parameter θ

$$y = L\theta + \xi, \quad \text{s.t. } \theta \in \mathcal{B}_c,$$

where $\xi \in \mathbb{R}^K$ is the assumed measurements noise with symmetric and positive definit covariance Γ.
In order to solve this problem, we solve the constrained optimization problem

$$\min_{\theta \in \mathcal{B}_c} \Phi(\theta), \quad \text{with } \Phi(\theta) = \|L\theta - y\|_\Gamma^2. \tag{4.2}$$

We note that the constrained optimization problem (4.2) is convex, which implies that necessary optimality conditions are also sufficient [20]. It is sufficient to assume that θ^* is a Karush Kuhn Tucker (KKT)-point for (4.2), i.e. there exists $\lambda^* \in \mathbb{R}^m$ such that

- $\theta^* \in \mathcal{B}_c$,
- $\lambda_j^* \geq 0$ for all $j \in \{1, \ldots, m\}$,
- $\lambda_j^*(\langle c_j, \theta^*\rangle - \delta_j) = 0$ for all $j \in \{1, \ldots, m\}$,
- $\nabla\Phi(\theta^*) + \sum_{j=1}^{m} \lambda_j^* c_j = 0$,

in order to ensure that θ^* is a global minimizer of (4.2). We can view **box-constraints** as special case of the set (4.1) with $f_j(\theta) = \pm\theta_j + \delta_j$ by setting $c_j = \pm e_j$, where e_j is the j-th unit vector in $\mathbb{R}^I$ and δ_j corresponds to the lower or upper bound on $\theta_j \in \mathbb{R}$.

Remark 4.1.1. *We note that $L^\top\Gamma^{-1}L$ is always symmetric and positive semidefinite. In case that $L^\top\Gamma^{-1}L$ is strictly positive definite, the optimization problem* (4.2) *is strictly convex. In particular, there exists at most one stationary point θ^* which is the global minimum for the problem* (4.2) *(cp. [20], Proposition 2.1.1). However, please note that the assumption on the regularity of $L^\top\Gamma^{-1}L$ is in general not satisfied for inverse problems. Typically, the number of unknown parameters is much larger than the number of observations, i.e. $n \gg K$, in the applications of interest.*

Remark 4.1.2. *The condition $L^\top\Gamma^{-1}L > 0$ is well-known in the context of data assimilation and is related to the strong observability condition. As a result this condition for the case of ensemble Kalman filter has been well-documented, see e.g. [24**?**].*

The idea behind the incorporation of box-constraints in EKI is based on the projected gradient descent method and the theory behind. Therefore, we will introduce the projected gradient descent method, before we will study the box-constraints within EKI.

4.2 Projected gradient descent method

We will give a brief overview of projected gradient descent method in order to solve the constrained optimization problem (4.2). We refer to the work of Bertsekas [20] for more details on this method in the discrete-time setting.
Recall, that we denote the underlying box with $\mathcal{B}_c$ which is now considered to be

$$\mathcal{B}_c = \{\theta \in \mathbb{R}^I \mid a_i \leq \theta_i \leq b_i,\ i = 1, \dots, m\}$$

and define the projection onto the box as $\mathcal{P} : \mathbb{R}^I \to \mathcal{B}_c$ by

$$\begin{aligned}(\mathcal{P}(\theta))_i &= \begin{cases} a_i, & \text{if } \theta_i < a_i, \\ \theta_i, & \text{if } \theta_i \in [a_i, b_i], \\ b_i, & \text{if } \theta_i > b_i, \end{cases} \qquad i = 1, \dots, m, \\ (\mathcal{P}(\theta))_i &= \theta_i, \qquad i = m+1, \dots, n. \end{aligned}$$

The projected gradient method with step size $\alpha_k > 0$ is based on the iteration

$$\theta^{k+1}(\beta_k) = \mathcal{P}(\theta^k - \beta_k \nabla \Phi(\theta^k)). \tag{4.3}$$

By considering β_k going to zero and using directional derivatives one can derive a continuous-time limit.

$$\begin{aligned}\left(\frac{d\theta}{dt}\right)_i &= \begin{cases} -\nabla_i \Phi(\theta), & \theta_i \in (a_i, b_i), \\ -\nabla_i \Phi(\theta) \mathbb{1}_{[0,\infty)}(-\nabla_i \Phi(\theta)), & \theta_i = a_i, \\ -\nabla_i \Phi(\theta) \mathbb{1}_{(-\infty,0]}(-\nabla_i \Phi(\theta)), & \theta_i = b_i, \end{cases} \qquad i = 1, \dots, m, \\ \left(\frac{d\theta}{dt}\right)_i &= -\nabla_i \Phi(\theta), \qquad i = m+1, \dots, n. \end{aligned} \tag{4.4}$$

More details can be found for the continuous-time limit of the projected EKI in Section 4.3.

Remark 4.2.1. *Since the right hand side (RHS) of* (4.4) *is discontinuous, it is not obvious that a solution to this system exists. To ensure unique existence we consider a smoothed version of* (4.4) *by approximating the limit by ideas inspired from barrier methods. To do so, we introduce the following parametrized, convex optimization problems*

$$\min_{\theta \in \mathcal{B}_c} \Phi(\theta) - \frac{1}{\iota} \sum_{i=1}^{2m} log(-f_i(\theta)).$$

with parameter $\iota > 0$ *and inequality constraints* $f_i(\theta) = a_i - \theta_i,\ i = 1, \dots, m$ *and* $f_{i+m}(\theta) = \theta_i - b_i,\ i = 1, \dots, m$. *For* $\iota \to \infty$, *the log barrier functions get closer to the indicator function of the feasible set of the original problem. In the following, we consider the equivalent problems*

$$\min_{\theta \in \mathcal{B}_c} \iota \Phi(\theta) - \sum_{i=1}^{2m} log(-f_i(\theta)), \tag{4.5}$$

where we define $\tilde{\Phi}(\theta) = \iota\Phi(\theta) - \sum_{i=1}^{2m} log(-f_i(\theta))$. *We approximate* (4.4) *for* $i \in \{1, \dots, m\}$ *by*

$$\frac{\mathrm{d}\theta}{dt} = -\iota\nabla\Phi(\theta) + \sum_{i=1}^{2m} \frac{1}{f_i(\theta)} \nabla f_i(\theta) = -\nabla\tilde{\Phi}(\theta)\,. \tag{4.6}$$

For our theoretical results we will always consider the smoothed initial value problem (4.5). In the following theorem we present the convergence result for the smoothed initial value problem.

Theorem 4.2.2. *Let* $\theta_0 \in \mathcal{B}$ *and* $\theta(t)$ *denote the solution of the smoothed initial value problem* (4.6) *with* $\theta(0) = \theta_0$. *Further assume that* $L^\top \Gamma^{-1} L$ *is positive definite, and there exists a (unique global) minimizer* θ_ι^* *of* (4.5). *Then for each* $\iota > 0$ *it holds true that*

$$\lim_{t\to\infty} \theta(t) = \theta_\iota^*,$$

i.e. the solution $\theta(t)$ *converges to the (global) minimizer of* (4.5).

Proof. We define $V(\theta) = \frac{1}{2}\|\theta - \theta_\iota^*\|^2$ and prove that V is a strict Lyapunov-function by the strict convexity of the optimization problem. The flow of V satisfies

$$\frac{\mathrm{d}V(\theta)}{dt} = \langle \frac{\mathrm{d}\theta}{\mathrm{d}t}, \theta - \theta_\iota^* \rangle = \langle -\nabla\tilde{\Phi}(\theta), \theta - \theta_\iota^* \rangle < 0\,,$$

thus, the claim follows. □

Remark 4.2.3. *By duality arguments (see [29, 11.2] for more details), the accuracy of the approximation can be bounded by*

$$\Phi(\theta_\iota^*) - \Phi(\theta^*) \leq \frac{2m}{\iota}\,,$$

where θ^* *denotes the minimizer of the original problem* (4.2). *In particular,* $\Phi(\theta_\iota^*) \to \Phi(\theta^*)$ *for* $\iota \to \infty$ *and thus* $\theta_\iota^* \to \theta^*$.

Corollary 4.2.4. *Let* $\theta_0 \in \mathcal{B}_c$ *and* $\theta(t)$ *denote the solution of the smoothed initial value problem* (4.6) *with* $\theta(0) = \theta_0$. *Further assume that there exists a (global) minimizer of* (4.5). *Then it holds true that*

$$\lim_{t\to\infty} \Phi(\theta(t)) = \Phi(\theta_\iota^*),$$

where θ_ι^* *is a KKT-point of* (4.5).

Proof. Let θ_ι^* be an arbitrary KKT-point of (4.5). The flow in the observation space is given by

$$\frac{\mathrm{d}L\theta}{\mathrm{d}t} = -\iota L\nabla\Phi(\theta) + \sum_{i=1}^{2m} \frac{1}{f_i(\theta)} L\nabla f_i(\theta)\,,$$

which corresponds to the gradient flow of a strictly convex optimization problem in the observation space. Thus, the claim follows by the same arguments as before in Theorem 4.2.2. □

4.2.1 The preconditioned projected gradient method

We extend the previous results to the preconditioned version of the iteration (4.3), which is in discrete-time given by

$$\theta^{k+1}(\beta) = \mathcal{P}(\theta^k - \beta D_k \nabla\Phi(u^k)),$$

for a symmetric, positive definite matrix D_k. It is well known that arbitrary choice of D_k gives no descent for any choice of $\beta > 0$. We discuss this issue briefly in an example, which demonstrates that the preconditioning of the gradient flow does not lead to a descent direction in general (cp. [19]). This can also be seen in Figure 4.1.

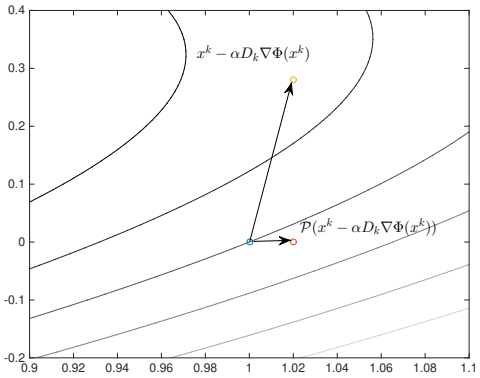

Figure 4.1: Varying contour lines of the function $\Phi(x)$ defined in Example 4.2.5, with both the preconditioned descent direction in the unconstrained case and the projected preconditioned descent direction.

Example 4.2.5. *We consider the 2-dimensional quadratic example, where we define the quadratic function*

$$\Phi(x) = x^\top Q x + x_1 = (x_1 - x_2)^2 + 2x_2^2 + x_1, \quad \textit{with} \quad Q = \begin{pmatrix} 1 & -1 \\ -1 & 3 \end{pmatrix},$$

and consider the minimization problem of Φ with the constraints $x_1 \in \mathbb{R}$, $x_2 \leq 0$. The gradient of Φ is given by

$$\nabla\Phi(x) = \begin{pmatrix} 2x_1 - 2x_2 + 1 \\ -2x_1 + 6x_2 \end{pmatrix}.$$

We assume that the current iterate is given by $x^k = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and we choose the symmetric, positive definite symmetric matrix

$$D_k = \begin{pmatrix} 1 & 2 \\ 2 & 10 \end{pmatrix}$$

as a preconditioner. For arbitrary $\beta > 0$, the next iteration is given by

$$x^{k+1} = \mathcal{P}(x^k - \alpha D_k \nabla\Phi(x^k)) = \begin{pmatrix} 1 + \alpha \\ 0 \end{pmatrix},$$

where $\mathcal{P}$ is the projection onto $\mathbb{R} \times \mathbb{R}_{\leq 0}$. Then,

$$\Phi(x^{k+1}) = (1+\alpha)^2 + 1 + \alpha > 2 = \Phi(x^k)\,,$$

i.e. for all $\beta > 0$ the function value of the objective function increases.

Example 4.2.5 shows that a simple projection strategy for the EKI, which is a preconditioned gradient flow in the linear case, does not lead to a convergent descent method in general.
To ensure a descent direction for the preconditioned projected gradient method, we follow the approach of [19, Proposition 1], which suggests to use matrices D_k diagonal with respect to the subset of indices containing

$$\mathcal{I}^+(\theta^k) = \left\{ i \in \{1,\dots,m\} \mid \theta_i^k = a_i, \frac{\partial \Phi(\theta^k)}{\partial \theta^i} > 0 \vee \theta_i^k = b_i, \frac{\partial \Phi(\theta^k)}{\partial \theta^i} < 0 \right\}. \tag{4.7}$$

A matrix $B = (b_{i,j})_{i,j=1\dots,m} \in \mathbb{R}^{m\times m}$ is called diagonal with respect to a subset $\mathcal{I} \subset \{1,\cdots,m\}$ if

$$b_{i,j} = 0, \quad \text{for all } i \in \mathcal{I},\ j \in \{1,\dots,m\} \setminus \{i\}.$$

We will give more details on the modification of the preconditioner in the context of the EKI in the following. In particular, we will make use of variance inflation introduced in Section 3.2.2 to ensure a descent.

4.3 Projected ensemble Kalman inversion

In this section, we introduce the incorporation of projection into EKI and derive the corresponding continuous limit of the algorithm. We provide a complete convergence analysis of the proposed modification in the linear setting. This analysis includes the analysis of the ensemble collapse and the convergence to the truth.
To incorporate the projection into our scheme we modify the iteration (3.8) described in Section 3.1. For the derivation we assume a possibly nonlinear forward map $H : \mathbb{R}^I \to \mathbb{R}^K$ and we define the prediction step in its projected form as

$$\theta_{n,\mathcal{P}}^{(j)} = \mathcal{P}(\theta_n^{(j)}), \quad \bar{\theta}_{\mathcal{P}} = \frac{1}{J}\sum_{j=1}^{J}(\theta_{n,\mathcal{P}}^{(j)}), \quad \bar{H}_{\mathcal{P}} = \frac{1}{J}\sum_{j=1}^{J} H(u_{n,\mathcal{P}}^{(j)}),$$

and similarly for the covariances

$$C_{n,\mathcal{P}}^{\theta p} = \frac{1}{J}\sum_{j=1}^{J}(\theta_{n,\mathcal{P}}^{(j)} - \bar{\theta}_{\mathcal{P}}) \otimes (H(\theta_{n,\mathcal{P}}^{(j)}) - \bar{H}_{\mathcal{P}}),$$

$$C_{n,\mathcal{P}}^{pp} = \frac{1}{J}\sum_{j=1}^{J}(H(\theta_{n,\mathcal{P}}^{(j)}) - \bar{H}_{\mathcal{P}}) \otimes (H(\theta_{n,\mathcal{P}}^{(j)}) - \bar{H}_{\mathcal{P}}).$$

We construct our update formula in its closed form by

$$\begin{cases} \tilde{\theta}_{n+1,\mathcal{P}}^{(j)} = \theta_{n,\mathcal{P}}^{(j)} + C_{n,\mathcal{P}}^{\theta p}(C_{n,\mathcal{P}}^{pp} + h^{-1}\Gamma)^{-1}(y - H(\theta_{n.\mathcal{P}}^{(j)})), \\ \theta_{n+1,\mathcal{P}}^{(j)} = \mathcal{P}(\tilde{\theta}_{n+1,\mathcal{P}}^{(j)}), \end{cases}$$

where we consider the case of unperturbed observation, i.e. we consider $y_{n+1}^{(j)} = y$ for all $n \in \mathbb{N}$ and all $j = 1,\dots,J$.

4.3.1 Continuous-time limit

We now derive the continuous-time limit for the projected EnKF for inverse problems. Recall the equations in its closed form, given by the componentwise increments

$$\begin{aligned}(\theta_{n+1}^{(j)})_i - (\theta_n^{(j)})_i = \mathcal{P}\Big([\mathcal{P}(\tilde{\theta}_n^{(j)})]_i + \Big[C_{n,\mathcal{P}}^{\theta p}(C_{n,\mathcal{P}}^{pp} + \frac{1}{h}\Gamma)^{-1}(y - H(\mathcal{P}(\theta_n^{(j)}))\Big]_i\Big) \\ - \mathcal{P}\Big([\mathcal{P}(\tilde{\theta}_n^{(j)})]_i\Big).\end{aligned}$$

By using the Neumann expansion for part of the Kalman gain, we observe for positive-semidefinite $C \in \mathbb{R}^{K\times K}$ that

$$\begin{aligned}\left(\frac{1}{h}\Gamma + C\right)^{-1} &= \left(\frac{1}{h}\Gamma(I + \Gamma^{-1}C)\right)^{-1} \\ &= \left(\frac{1}{h}\Gamma\right)^{-1} + \sum_{k=1}^{\infty}\left(-\left(\frac{1}{h}\Gamma\right)^{-1}C\right)^k\left(\frac{1}{h}\Gamma\right)^{-1} \\ &= h\Gamma^{-1} + \sum_{k=1}^{\infty} h^{k+1}(\Gamma^{-1}C)^k\Gamma^{-1}. \end{aligned} \tag{4.8}$$

By defining

$$v_i := \left[C_{n,\mathcal{P}}^{\theta p}\Gamma^{-1}(y - H(\mathcal{P}(\theta_n^{(j)}))\right]_i,$$

using (4.8) and the definition of directional derivatives, we obtain

$$\begin{aligned}\lim_{h\to 0}\frac{(\theta_{n+1}^{(j)})_i - (\theta_n^{(j)})_i}{h} &= \begin{cases} v_i, & (\tilde{\theta}_n^{(j)})_i \in (a_i, b_i), \\ \mathbb{1}_{[0,\infty)}(v_i)v_i, & (\tilde{\theta}_n^{(j)})_i \le a_i, \\ \mathbb{1}_{(-\infty,0]}(v_i)v_i, & (\tilde{\theta}_n^{(j)})_i \ge b_i, \end{cases} \qquad i = 1,\dots,m \\ \lim_{h\to 0}\frac{(\theta_{n+1}^{(j)})_i - (\theta_n^{(j)})_i}{h} &= v_i(\theta_t^{(j)}), \quad i = m+1,\dots,n.\end{aligned}$$

Finally, we obtain the continuous-time limit for the EKI by

$$\begin{aligned}\left(\frac{\mathrm{d}\theta^{(j)}}{\mathrm{d}t}\right)_i &= \begin{cases} v_i(\theta_t^{(j)}), & (\theta_t^{(j)})_i \in (a_i, b_i), \\ \mathbb{1}_{[0,\infty)}(v_i(\theta_t^{(j)}))v_i(\theta_t^{(j)}), & (\theta_t^{(j)})_i = a_i, \\ \mathbb{1}_{(-\infty,0]}(v_i(\theta_t^{(j)}))v_i(\theta_t^{(j)}), & (\theta_t^{(j)})_i = b_i, \end{cases} \qquad i = 1,\dots,m \\ \left(\frac{\mathrm{d}\theta^{(j)}}{\mathrm{d}t}\right)_i &= v_i(\theta_t^{(j)}) \quad i = m+1,\dots,n,\end{aligned} \tag{4.9}$$

where $v_i(\theta^{(j)})$ is given by

$$v_i(\theta^{(j)}) = \left[C^{\theta p}(\theta)\Gamma^{-1}(y - H(\theta^{(j)}))\right]_i.$$

We remind the reader that in the linear case we can write $v(\theta^{(j)})$ as

$$v(\theta^{(j)}) = -C(\theta)\nabla\Phi(\theta^{(j)}), \tag{4.10}$$

with $\Phi(\theta) = \frac{1}{2}\|L\theta - y\|_\Gamma^2$ from the minimization problem (4.2), which shows the interpretation of the projected EKI as preconditioned projected gradient method, where the preconditioning matrix $C(\theta)$ is adapted over time.

Remark 4.3.1. *Due to the projection onto the feasible set, the RHS of* (4.9) *is discontinuous similar as in the case of the projected gradient descent method. Based on the ideas of Remark 4.2.1 we introduce again a smoothed system*

$$\frac{\mathrm{d}\theta^{(j)}}{\mathrm{d}t} = -\iota C(\theta)\nabla\Phi(\theta^{(j)}) + \sum_{i=1}^{2m}\frac{1}{f_i(\theta)}\nabla f_i(\theta), \quad j = 1, \dots, J,$$

where $f_i(\theta) = a_i - \theta_i,\ f_{i+m}(\theta) = \theta_i - b_i,\ i = 1, \dots, m.$

4.3.2 Transformed method for the EnKF

In Example 4.2.5 we have seen that in the case of preconditioned projected gradient methods, it is not possible to ensure a descent direction for an arbitrary choice of a positive definite matrix D_k. Based on the results of Bertsekas [19], we will transform (4.9) in a way such that the preconditioner is diagonal with respect to an index set $\mathcal{I}^+(u)$ which is built similar to the set from (4.7). Since we consider a system of particles we will use a preconditioner which is diagonal with respect to an index set which depends on the whole ensemble of particles $\theta = (\theta^{(j)})_{j=1,\dots,J}$, in particular we set

$$\mathcal{I}^+(\theta) := \left\{ i \in \{1, \dots, m\} \mid \bar\theta_i = a_i,\ \frac{\partial\Phi(\bar\theta)}{\partial x^i} > 0 \quad \vee \quad \bar\theta_i = b_i,\ \frac{\partial\Phi(\bar\theta)}{\partial x^i} < 0 \right\}.$$

Similarly, we could also choose the index set to be

$$\hat{\mathcal{I}}^+(\theta) := \bigcup_{j\in\{1,\dots,J\}} \left\{ i \in \{1, \dots, m\} \mid \theta_i^{(j)} = a_i,\ \frac{\partial\Phi(\theta^{(j)})}{\partial x^i} > 0 \quad \vee \right.$$
$$\left. \theta_i^{(j)} = b_i,\ \frac{\partial\Phi(\theta^{(j)})}{\partial x^i} < 0 \right\}.$$

In this work, we will focus on the choice of $\mathcal{I}^+(\theta)$. Therefore, we consider the preconditioned gradient flow given by

$$\left(\frac{\mathrm{d}\theta^{(j)}}{\mathrm{d}t}\right)_i = \begin{cases} p_i(\theta_t^{(j)}), & (\theta_t^{(j)})_i \in (a_i, b_i) \\ \mathbb{1}_{[0,\infty)}(p_i(\theta_t^{(j)}))(p_i(\theta_t^{(j)})), & (\theta_t^{(j)})_i = a_i \\ \mathbb{1}_{(-\infty,0]}(p_i(\theta_t^{(j)}))(p_i(\theta_t^{(j)})), & (\theta_t^{(j)})_i = b_i, \end{cases} \quad i = 1, \dots, m,$$
$$\left(\frac{\mathrm{d}\theta^{(j)}}{\mathrm{d}t}\right)_i = p_i(\theta_t^{(j)}), \quad i = m+1, \dots, n, \tag{4.11}$$

where $p(\theta_t^{(j)}) = -D(\theta_t)\nabla\Phi(\theta_t^{(j)})$ and the preconditioner is given by

$$(D(\theta_t))_{i,j} = \begin{cases} (C(\theta_t) + \varepsilon I)_{i,j}, & i,j \in \{1, \dots, n\} \setminus \mathcal{I}^+(\theta_t) \\ 0, & i \in \mathcal{I}^+(\theta_t),\ j \neq i \vee j \in \mathcal{I}^+(\theta_t),\ j \neq i \\ \varepsilon, & i = j \in \mathcal{I}^+(\theta_t). \end{cases} \tag{4.12}$$

Let $\{x^{(j)}\}_{j=1}^J$ be a system of particles in $\mathbb{R}^I$. Without loss of generality we assume $\mathcal{I}^+(x) = \{1,\dots,r\}$ for $r \le m$, such that the the preconditioner $D(x)$ can be written as

$$\begin{pmatrix} \varepsilon \operatorname{Id}_r & 0 \\ 0 & \hat{C}(x) + \varepsilon \operatorname{Id}_{n-r} \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & \hat{C}(x) \end{pmatrix} + \varepsilon \operatorname*{Id}_n,$$

where $(\hat{C}(x))_{i,j\in\{1,\dots,n-r\}} = (C(x))_{i,j\in\{r+1,\dots,n\}}$.

Remark 4.3.2. *Consider the continuous-time limit of the projected EKI in equation* (4.9)*. When we introduce variance inflation for* (4.10) *with the help of a diagonal matrix, e.g. by*

$$v(\theta^{(j)}) = -(C(\theta) + \varepsilon I)\nabla\Phi(\theta^{(j)}),$$

then the preconditioner $D(\theta) = C(\theta) + \varepsilon I$ is diagonal with respect to the index set $\mathcal{I}^+(\theta)$ and can be written in the form of (4.12)*.*
This can be seen as follows: For simplicity, we assume $m = n$. Since $\theta^{(j)}$ evolves by (4.11)*, we have that $\theta_i^{(j)} \in [a_i, b_i]$ for all i and j. Let $i \in \mathcal{I}^+(\theta)$, then it follows*

$$\bar{\theta}_i = a_i \qquad \vee \qquad \bar{\theta}_i = b_i.$$

Since the particles are feasible for all $t \ge 0$, we obtain

$$\theta_i^{(j)} = a_i \quad \forall j \qquad \vee \qquad \theta_i^{(j)} = b_i \quad j = 1,\dots,J,$$

and finally $\theta_i^{(j)} - \bar{\theta}_i = 0$ for all j. This leads to

$$(C(\theta) + \varepsilon I)_{ik} = \frac{1}{J}\sum_{j=1}^{J}(\theta_i^{(j)} - \bar{\theta}_i)(\theta_k^{(j)} - \bar{\theta}_k) = 0,$$

for $i \in \mathcal{I}^+(\theta)$, $k \in \{1,\dots,n\}$ and $i \ne k$.
Hence, we can view the projected EnKF with variance inflation as special case of the presented transformed method.

4.3.3 Convergence Results

For the next part we will study the convergence of the solution $(\theta_t^{(j)})$ of the smoothed system of (4.11)

$$\frac{du_t^{(j)}}{dt} = -\iota D(u_t)\nabla\Phi(u_t^{(j)}) + \sum_{i=1}^{2m}\frac{1}{f_i(u)}\nabla f_i(u), \tag{4.13}$$

to possible KKT-points for (4.5). Note that we have defined f in Remark 4.2.1.
We consider a KKT-point θ_ι^* of (4.5) and aim to prove the convergence of the residuals

$$\tilde{r}_t^{(j)} := \theta_t^{(j)} - \theta_\iota^*.$$

In order to prove convergence of the residuals we have to control the empirical covariance $C(\theta_t)$, in particular we have to control the ensemble spread

$$e_t^{(j)} = \theta_t^{(j)} - \bar{\theta}_t,$$

which we have already studied for the unconstrained EKI in Section 3.4.

4.3.4 Ensemble Collapse

The following proposition states that we can bound the ensemble spread e_t over time.

Proposition 4.3.3. *Let $\theta_0 = (\theta_0^{(j)})_{j\in\{1,\dots,J\}} \in \mathcal{B}_c$, $j = 1,\dots,J$ be the initial ensemble and $\theta(t)$ denotes the solution for the smoothed flow* (4.13)*. Then, it holds true that*

$$\frac{1}{J}\sum_{j=1}^{J}|e_t^{(j)}|^2 \leq \frac{1}{J}\sum_{j=1}^{J}|e_0^{(j)}|^2,$$

for all $t \geq 0$.

Proof. We define the function $V(\theta) = \frac{1}{J}\sum\limits_{j=1}^{J}\frac{1}{2}|\theta^{(j)} - \bar{\theta}|^2$ and show that $V(u) \leq 0$ in order to imply the monotonicity of the quantity e. We have

$$\begin{aligned}\frac{\mathrm{d}V(\theta_t)}{\mathrm{d}t} =& -\frac{1}{J}\sum_{j=1}^{J}\langle\theta_t^{(j)} - \bar{\theta}_t, \iota D(\theta_t)L^\top\Gamma^{-1}L(\theta_t^{(j)} - \bar{\theta}_t)\rangle \\ &-\frac{1}{J}\sum_{j=1}^{J}\langle\theta_t^{(j)} - \bar{\theta}_t, \nabla\tilde{f}(\theta_t^{(j)}) - \overline{\nabla\tilde{f}(\theta_t)})\rangle,\end{aligned}$$

with $\tilde{f}(\theta) = \frac{1}{\iota}\sum_{i=1}^{2m}\log(-f_i(\theta))$ and $\overline{\nabla\tilde{f}(\theta_t)} = \frac{1}{J}\sum_{j=1}^{J}\tilde{f}(\theta^{(j)})$. The first inner product can be straightforwardly shown to be smaller or equal than zero by exploiting the linearity of $\nabla\Phi$. Using the convexity of $\tilde{f}$ gives

$$-\frac{1}{J}\sum_{j=1}^{J}\langle\theta_t^{(j)} - \bar{\theta}_t, \nabla\tilde{f}(\theta_t^{(j)}) - \overline{\nabla\tilde{f}(\theta_t)})\rangle \leq 0\,,$$

which implies the monotonicity of the quantity e and proves the assertion.

□

4.3.5 Convergence of the residuals

We are now ready to formulate the following theorem regarding the residual in the parameter space $\tilde{r}_t^{(j)}$.

Theorem 4.3.4. *Let $\theta_0 = (\theta_0^{(j)})_{j\in\{1,\dots,J\}} \in \mathcal{B}_c$ be the initial ensemble and $\theta(t)$ denotes the solution of* (4.13)*, $L^\top\Gamma^{-1}L$ be positive definite, $\varepsilon > 0$, $\mathcal{B}_c \neq \emptyset$ and θ_ι^* be the KKT-point for* (4.5)*. Then it holds true that*

$$\lim_{t\to\infty}\frac{1}{J}\sum_{j=1}^{J}|\tilde{r}_t^{(j)}|^2 = \lim_{t\to\infty}\frac{1}{J}\sum_{j=1}^{J}|\theta_t^{(j)} - \theta_\iota^*|^2 = 0.$$

Proof. By assumption, we have that $L^\top\Gamma^{-1}L$ is positive definite and $\mathcal{B}_c \neq \emptyset$, i.e. there exists a unique global minimizer θ_ι^* of (4.5).
We define the function $V(\theta) = \frac{1}{J}\sum\limits_{j=1}^{J}\frac{1}{2}|\theta^{(j)} - \theta_\iota^*|^2$ and prove

$$\frac{\mathrm{d}V(\theta_t)}{\mathrm{d}t} < 0,$$

for $\theta_t = (\theta_t^{(j)})_{j=1,\dots,J}$.
The variance inflation breaks the subspace property of the EKI and gives the convergence to the KKT-point:

$$\begin{aligned}
\frac{1}{J}\sum_{j=1}^{J}\frac{\mathrm{d}\frac{1}{2}|\theta_t^{(j)}-\theta_\iota^*|^2}{\mathrm{d}t} =& \frac{1}{J}\sum_{j=1}^{J}\langle\theta_t^{(j)}-\theta_\iota^*, -D(\theta)\nabla\Phi(\theta_t^{(j)})-\nabla\tilde{f}(\theta_t^{(j)})\rangle \\
=& -\frac{1}{J}\sum_{j=1}^{J}\langle\theta_t^{(j)}-\theta_\iota^*, C(\theta)L^\top\Gamma^{-1}L(\theta_t^{(j)}-\theta_\iota^*)\rangle \\
& -\frac{1}{J}\sum_{j=1}^{J}\epsilon\langle\theta_t^{(j)}-\theta_\iota^*, \nabla\Phi(\theta_t^{(j)})+\nabla\tilde{f}(\theta_t^{(j)})\rangle \\
=& -\frac{1}{J^2}\sum_{j,k=1}^{J}\langle\theta_t^{(j)}-\theta_\iota^*), \theta_t^{(k)}-\theta_\iota^*\rangle\langle\theta_t^{(k)}-\theta_\iota^*, L^\top\Gamma^{-1}L(\theta_t^{(j)}-\theta_\iota^*)\rangle \\
& -\frac{1}{J}\sum_{j=1}^{J}\epsilon\langle\theta_t^{(j)}-\theta_\iota^*, \nabla\Phi(\theta_t^{(j)})+\nabla\tilde{f}(\theta_t^{(j)})\rangle < 0,
\end{aligned}$$

where we have used the convexity of Φ and $\tilde{f}$ and Lemma 3.3.5. □

Corollary 4.3.5. *Let $\theta_0 = (\theta_0^{(j)}) \in \Omega$ be the initial ensemble, $\varepsilon > 0$ and assume that there exists a (global) minimizer for* (4.5)*. Then it holds true that*

$$\lim_{t\to\infty}\frac{1}{J}\sum_{j=1}^{J}|\Phi(\theta_t^{(j)})-\Phi(\theta_\iota^*)|^2 = 0,$$

where θ_ι^* is a KKT-point of (4.5)*.*

Proof. The claim follows similarly to the proof of Corollary 4.2.4. □

Time-dependent variance inflation Instead of using variance inflation with constant multiplier ε, we will reduce the variance inflation over time similar to the variance inflation in the unconstrained case (3.30). We set $\varepsilon(t) = \frac{1}{t^\alpha+R}$, with $\alpha \in (0,1)$ and $R > 0$.
We will quantify the ensemble collapse with given rate in the following Proposition.

Proposition 4.3.6. *Let $\theta_0 = (\theta_0^{(j)})_{j\in\{1,\dots,J\}} \in \mathcal{B}_c$ be the initial ensemble and $\theta(t)$ denotes the solution of* (4.13)*, $L^\top\Gamma^{-1}L$ be positive definite and $\varepsilon(t) = \frac{1}{t^\alpha+R}$, with $\alpha \in (0,1)$ and $R > 0$. Then it holds true that*

$$\frac{1}{J}\sum_{j=1}^{J}|e_t^{(j)}|^2 \in \mathcal{O}(t^{-(1-\alpha)}).$$

Proof. We consider again the function $V(\theta) = \frac{1}{J}\sum_{j=1}^{J}\frac{1}{2}|\theta^{(j)}-\bar{\theta}|^2$ and use

$$\frac{\mathrm{d}V(\theta_t)}{\mathrm{d}t} = -\frac{1}{J}\sum_{j=1}^{J}\langle\theta_t^{(j)}-\bar{\theta}_t, D(\theta_t)L^\top\Gamma^{-1}L(\theta_t^{(j)}-\bar{\theta}_t)\rangle$$

$$
\begin{aligned}
&- \frac{1}{J}\sum_{j=1}^{J}\langle \theta_t^{(j)} - \bar{\theta}_t, \nabla \tilde{f}(\theta_t^{(j)}) - \bar{\nabla}\tilde{f}(\theta_t))\rangle \\
&\leq -\frac{1}{J}\sum_{j=1}^{J}\langle \theta_t^{(j)} - \bar{\theta}_t, \varepsilon(t) L^\top \Gamma^{-1} L(\theta_t^{(j)} - \bar{\theta}_t)\rangle,
\end{aligned}
$$

similarly to the proof of 4.3.3. Thus, it follows that

$$
\frac{\mathrm{d}V(\theta_t)}{\mathrm{d}t} \leq -\varepsilon(t)\sigma_{\min}(L^\top \Gamma^{-1} L) V(\theta_t).
$$

By using the bound

$$
V(\theta_0) \geq \int_0^t \sigma_{\min}(L^\top \Gamma^{-1} L)\varepsilon(s)\,\mathrm{d}s \cdot V(\theta_t),
$$

for all $t \geq 0$ we conclude with

$$
\frac{1}{J}\sum_{j=1}^{J} |e_t^{(j)}|^2 \in \mathcal{O}(t^{-(1-\alpha)}).
$$

□

As shown before we can also prove the convergence of the residuals in the parameter space when we reduce the variance inflation over time.

Corollary 4.3.7. *Let $\theta_0 = (\theta_0^{(j)})_{j\in\{1,\dots,J\}} \in \mathcal{B}_c$ be the initial ensemble and $\theta(t)$ denotes the solution of* (4.13)*, $L^\top\Gamma^{-1}L$ be positive definite and $\varepsilon(t) = \frac{1}{t^\alpha + R}$, with $\alpha \in (\frac{1}{2}, 1)$ and $R > 0$. Furthermore, let θ_ι^* be the KKT-point of* (4.5)*. Then it holds true that*

$$
\frac{1}{J}\sum_{j=1}^{J} |\tilde{r}_t^{(j)}|^2 \in \mathcal{O}(t^{-(1-\alpha)}).
$$

Proof. We define the function $V(\theta) = \frac{1}{J}\sum_{j=1}^{J} \frac{1}{2}|\theta^{(j)} - \theta_\iota^*|^2$ and prove

$$
\frac{\mathrm{d}V(\theta_t)}{\mathrm{d}t} < 0,
$$

for $\theta_t = \{\theta_t^{(j)}\}_{j=1}^J$. Similarly to the proof of Theorem 4.3.4 we obtain

$$
\begin{aligned}
\frac{\mathrm{d}V(\theta_t)}{\mathrm{d}t} =& -\frac{1}{J^2}\sum_{j,k=1}^{J} \langle \theta_t^{(j)} - \theta_\iota^*, \theta_t^{(k)} - \theta_\iota^*\rangle \langle L^\top\Gamma^{-1}L(\theta_t^{(k)} - \theta_\iota^*), \theta_t^{(j)} - \theta_\iota^*\rangle \\
&- \frac{1}{J}\sum_{j=1}^{J}(\varepsilon(t)\langle \theta_t^{(j)} - \theta_\iota^*, \nabla\Phi(\theta_t^{(j)})\rangle + \langle \theta_t^{(j)} - \theta_\iota^*, \nabla\tilde{f}(\theta_t^{(j)})\rangle).
\end{aligned}
$$

and we conclude with

$$
\frac{\mathrm{d}V(\theta_t)}{\mathrm{d}t} \leq -\varepsilon(t)\sigma_{\min}(L^\top\Gamma^{-1}L)V(\theta_t).
$$

□

4.4 Numerical results

We seek to verify that the BC optimization introduced in Section 4.3 works in practice. In the following section, we consider two PDE based examples in order to illustrate the effect of the incorporation projection. The first example is based on a linear 1D elliptic PDE, whereas the second example is based on a nonlinear 2D PDE. While our theory is only provided in the linear setting, we observe promising results in the nonlinear example, too. For the continuum limit we solve the ODE through the `MATLAB` solver `ode45`.

4.4.1 Linear PDE

Our forward model is again the linear 1D elliptic PDE from subsection 3.5, where we seek a solution $p \in \mathcal{U} := H_0^1(D)$ from

$$\begin{aligned} -\frac{\mathrm{d}^2 p}{\mathrm{d}x^2} + p &= \theta \quad x \in D = (0, \pi), \\ p &= 0 \quad x \in \partial D. \end{aligned}$$

We specify the covariance of the noise as $\Gamma = \gamma^2 I$ where $\gamma = 0.01$, and we choose $T = 10^6$. The initial ensemble θ_0 is again chosen through a Fourier basis representation of $\mathcal{N}(0, C_0)$ with $C_0 = \beta(\mathfrak{L} - \mathrm{id})$, and the inflation parameters are taken as $\alpha = 0.75$, $R = 1$ and an ensemble-size $J = 5$.
As stated in Remark 4.3.2, the projected EKI with variance inflation can be viewed as a special case of the transformed version provided in (4.11). This suggests that both methods should perform similarly and outperform the original EKI with no constraint. For the numerics we now specify the projected EKI as the projected EKI without variance inflation and the transformed EKI as the projected EKI with variance inflation. For the linear case we compared the projected EKI, the transformed version and the original method without projecting onto the box. Note that the numerical solution of both the projected EKI and the transformed EKI are based on a smoothed version of the indicator function $1_y(x)$ by a linear function $\tilde{1}_y(x)$ with $\tilde{1}_y(y) = 1$ and $\tilde{1}_y(y \pm \iota) = 0$ for the upper and lower bound respectively.
Our numerics will consist of two different cases:

1. The truth $\theta^\dagger$ lies outside the box and $\mathcal{O}$ gives full observations.

2. The truth $\theta^\dagger$ lies outside the box, and $\mathcal{O}$ gives low-dimensional observations.

To understand the effect of the different forms of observations, if we have full observations then there exists a unique KKT-point to (4.2), implying the true parameter is a KKT point. If we have low-dimensional observations then $L^\top \Gamma^{-1} L$ is only positives semi-definite. Thus, we can only expect convergence of the cost functions. For the first case we choose full observations and for the latter we choose 15 observations.
Our first set of experiments for the linear PDE are shown in Figures 4.2 - 4.4, where we assume that we have full observations. The left hand side image of Figure 4.2 compares the performance of the different methods at reconstructing the truth. As known with EKI, it can exhibit a smooth reconstruction which we can see. For the projected EKI we notice a similar performance, however it takes into consideration the constraints as part of its reconstruction is on the boundary. However when analyzing the transformed EKI, we see

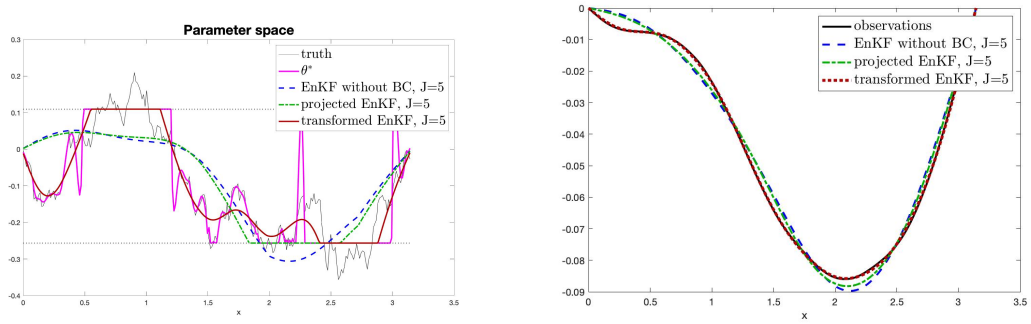

Figure 4.2: Transformed EnKF estimation in comparison to the EnKF estimation and the projected EnKF estimation. $J = 5$ particles have been simulated.

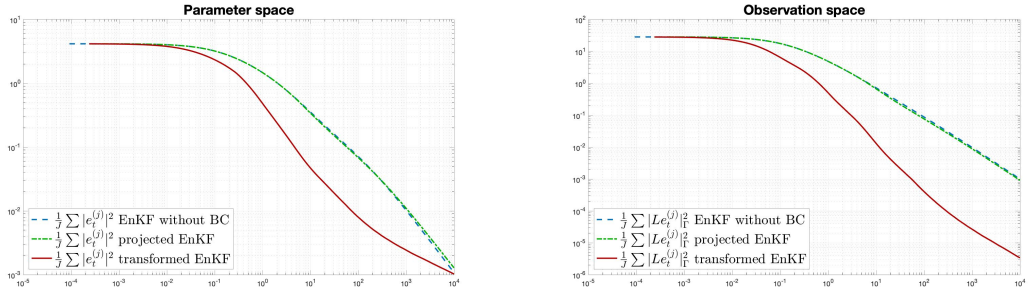

Figure 4.3: Ensemble spread in the transformed EnKF in comparison to the EnKF and the projected EnKF. $J = 5$ particles have been simulated.

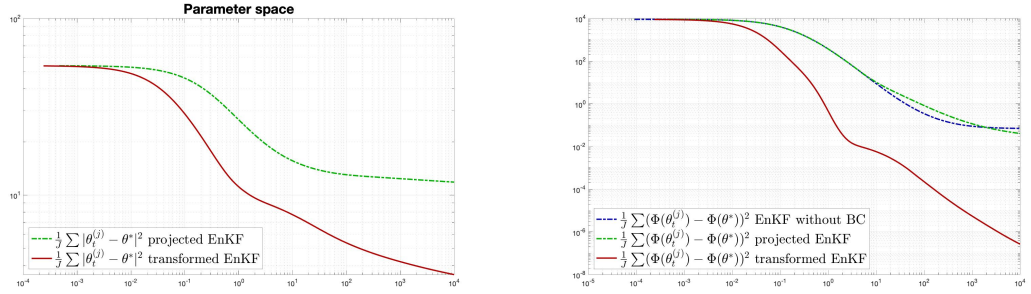

Figure 4.4: KKT-Residuals and difference of the misfit functional and the global minimum in the transformed EnKF in comparison to the projected EnKF. $J = 5$ particles have been simulated.

an improvement over previous methods, which is also evident from the comparison of the observations from the right hand image.

Figure 4.3 shows the expected ensemble collapse of each method. We observe the spread behaves almost at an identical rate for the EKI and projected EKI. To highlight further the benefit of using the transformed version, Figure 4.4 demonstrates this by showing a sharper decrease in both the KKT residuals as well as the difference of the misfit functional to global minimizer. As the projected method eventually levels flattens at 10^{-1} for both, the transformed version continues to achieve smaller differences.

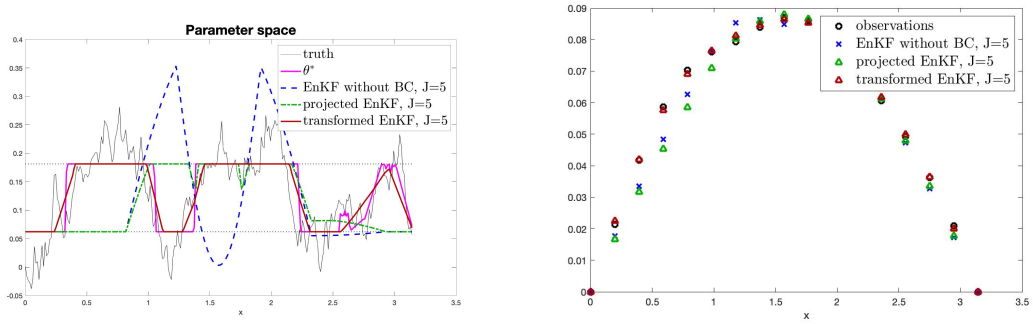

Figure 4.5: Transformed EKI estimation in comparison to the EKI estimation and the projected EKI estimation. $J = 5$ particles have been simulated.

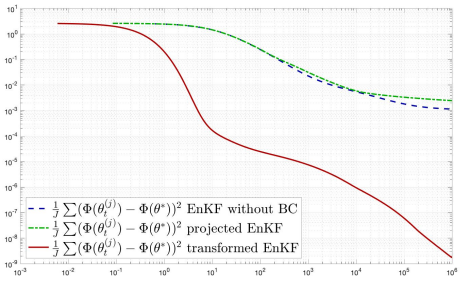

Figure 4.6: Difference of the misfit functional and the global minimum in the transformed EKI in comparison to the EKI and the projected EKI. $J = 5$ particles have been simulated.

The second set of experiments in the linear setting are shown in Figures 4.5 - 4.6, where we assume that we have only 15 observation points. The obtained results are analogous to the full observation case, in which the transformed version outperforms the other methods again.

4.4.2 Darcy flow

To demonstrate the effectiveness of the transformed projection, we consider a second example of a 2D nonlinear PDE. We will use an analogous nonlinear version of our linear elliptic PDE, which arises in subsurface flow. The forward problem associated with the Darcy flow model is using the permeability $a \in L^\infty(D)$ to solve

$$\begin{aligned} -\nabla \cdot (a \nabla p) &= f, \quad x \in D = (0, \pi)^2, \\ p &= 0, \quad x \in \partial D, \end{aligned} \tag{4.14}$$

where $\nabla\cdot$ denotes the divergence and we have imposed zero boundary conditions. Our forward solver is based on a second-order centred finite difference method with mesh size $h = 2^{-4}$. We take a constant source term of (4.14) as $f = 1$. The inverse problem associated with (4.14) , where we specify $a = \exp(\theta)$, is to reconstruct θ from noisy measurement

$$y = \mathcal{O}(p) + \xi, \quad \xi \sim N(0, \Gamma),$$

where $\mathcal{O} : H_0^1(D) \cap H^2(D) \to \mathbb{R}^K$ denotes the linear observation map, which takes measurements at K equidistant chosen points in D, i.e. $\mathcal{O}(p) = (p(x_1), \dots, p(x_K))^\top$, for $p \in \mathcal{V}$, $x_1, \dots, x_K \in D$.
We stick to the same setting as in subsection 4.4.1, but with the modifications of only having 16 observation points, and testing a system with $n = 64$. Our prior $\theta \sim N(0, C_0)$ is simulated through a KL expansion of the form

$$\theta = \sum_j^{\mathcal{J}} \sqrt{\nu_j}\zeta_j\varphi_j, \quad \zeta \sim N(0, I),$$

where (ν_j, φ_j) is the eigenbasis of the covariance operator C_0, expressed as

$$C_0 := \beta(\mathrm{Id} - \Delta)^{-\alpha},$$

where the hyperparameters are specified as $\beta = 1$ and $\alpha = 2$. See Section 2.2.4 for more details on the KL expansion.
Further, we have modified (4.9) by incorporation of variance inflation following the ideas introduced in (3.15) in section 3.2.2

$$v_i(\theta^{(j)}) = \left[\left(C^{\theta p}(\theta) + \kappa B\ DH^*(\bar{\theta})\right)\Gamma^{-1}(y - H(\theta^{(j)}))\right]_i,$$

to which we again refer as transformed method.

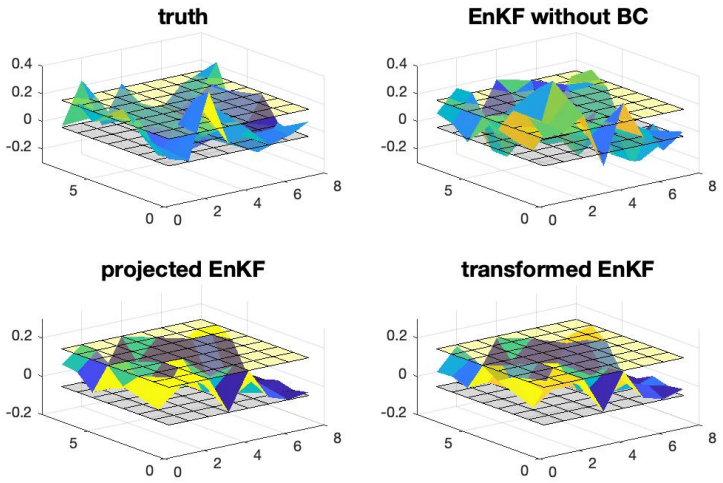

Figure 4.7: Transformed EKI parameter estimation in comparison to the EKI estimation and the projected EnKF estimation. $J = 5$ particles have been simulated.

For the nonlinear experiments Figure 4.7 shows the performance of each method w.r.t. the truth. It can be seen that the EKI without constraints does not remain within the feasible set unlike the other two methods, despite Figure 4.8 looking identical across all methods. To see a more indepth representation of the performance we analyze the ensemble spread seen in Figure 4.9 where they all seem to converge to zero and the rates look similar. However, by looking at the difference of the misfit functional with the minimizer in Figure 4.10, we see the difference of the transformed method continues to decrease while for the projected EKI it starts to flatten, likewise with the original EKI. This highlights further the benefit of using the transformed EKI.

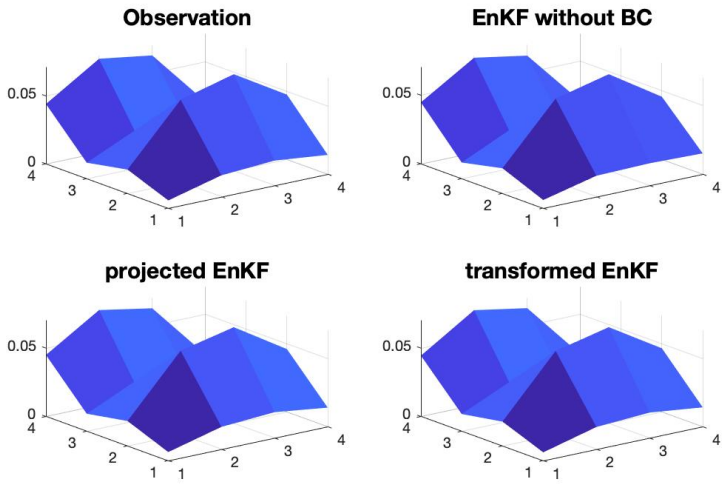

Figure 4.8: Transformed EKI observation estimation in comparison to the EKI estimation and the projected EKI estimation. $J = 5$ particles have been simulated.

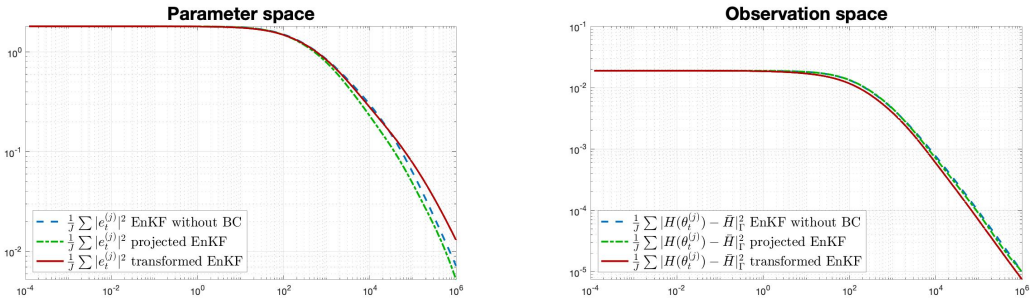

Figure 4.9: Ensemble spread in the transformed EKI in comparison to the EKI and the projected EKI. $J = 5$ particles have been simulated.

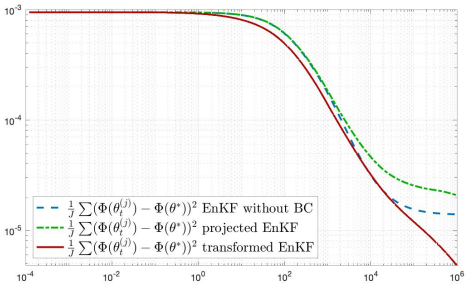

Figure 4.10: Difference of the misfit functional and the global minimum in the transformed EKI in comparison to the projected EKI. $J = 5$ particles have been simulated.

5 Tikhonov regularization for ensemble Kalman inversion

In the previous chapter, we have seen that the EKI can be interpreted as an optimizer to fit the observed data. While we have proved the convergence of the residual in the case of noisefree observations, the natural question then is how to avoid fitting the noise within the data if there is pertubation within the measurements. In [201], the authors suggest to use a stopping criterion based on the Morozov discrepancy principle. Another recent approach is based on the viewpoint of the EKI as an iterative regularization method [110] and incorporates regularization within the EKI algorithm. In particular, in [41] the authors incorporate Tikhonov regularization within the EKI method. In this thesis, we will keep the focus on the Tikhonov regularized EKI. Throughout this section, we will work in the finite-dimensional setting, i.e. we assume $\mathcal{X} = \mathbb{R}^I$. However, using again the subspace property in Lemma 5.1.2, the presented results can be extended to the infinite-dimensional setting.
We start by formulating the Tikhonov regularized EKI in Section 5.1. In Section 5.2, based on the continuous-time limit of the method, we provide a well-posedness result for fixed regularization parameter choice, which is based on the unique existence of strong solutions from the underlying system of coupled SDEs. Furthermore, we formulate convergence results in the long-time limit. The theoretical results are again strongly based on stochastic Lyapunov functions. Section 5.3 is devoted to present ideas of adaptive choices of the regularization parameter. We again close the chapter by verifying of the presented methods through numerical experiments in Section 5.4.

5.1 Introduction of the Tikhonov regularized ensemble Kalman inversion

In order to incorporate Tikhonov regularization into EKI, following the derivation in [41], we extend the underlying inverse problem by prior information about the unknown paramater, which leads to the following equations

$$\begin{aligned} y &= H(\theta) + \xi \\ 0 &= \theta + \eta, \end{aligned} \tag{5.1}$$

where $\xi \sim \mathcal{N}(0, \Gamma)$ is the noise of the original inverse problem and $\eta \sim \mathcal{N}(0, \kappa^{-1} C_0)$ corresponds to the prior belief on θ. Here $C_0 \in \mathbb{R}^{I \times I}$ is a positive definite matrix, which stores the covariance structure of θ and $\kappa > 0$ corresponds to the regularization parameter. We will see the connection to the regularization parameter defined in Definition 2.1.4.

To derive the Tikhonov regularized ensemble Kalman inversion (TEKI), we set

$$\mathcal{H} = \mathbb{R}^I,\ p = K + I,\ q = \begin{pmatrix} y \\ 0 \end{pmatrix},\ G(\theta) = \begin{pmatrix} H(\theta) \\ \theta \end{pmatrix},\ \zeta = \begin{pmatrix} \xi \\ \eta \end{pmatrix},$$

in (3.1), resulting in the extended inverse problem

$$q = G(\theta) + \zeta,$$

with

$$\zeta \sim \mathcal{N}(0, \Sigma), \quad \Sigma = \begin{pmatrix} \Gamma & 0 \\ 0 & \kappa^{-1} C_0 \end{pmatrix}.$$

Application of the EKI to this inverse problem leads to the following algorithm.

Algorithm 6: Tikhonov regularized ensemble Kalman inversion

Input: initial ensemble $(\theta_0^{(j)})_{j=1}^J$, extended observation q
Output: $\bar{\theta}_N$
for $n = 0, \dots, N-1$ **do**
Prediction step:
Define sample mean and sample covariance

$$\bar{\theta} = \frac{1}{J}\sum_{j=1}^{J} \theta^{(j)}, \quad \bar{H} = \frac{1}{J}\sum_{j=1}^{J} H(\theta^{(j)}), \quad \bar{G} = \begin{pmatrix} \bar{H} \\ \bar{\theta} \end{pmatrix}$$

$$B^{pp} = \frac{1}{J}\sum_{j=1}^{J} (G(\theta^{(j)}) - \bar{G}) \otimes (G(\theta^{(j)}) - \bar{G}),$$

$$B^{\theta p} = \frac{1}{J}\sum_{j=1}^{J} (\theta^{(j)} - \bar{\theta}) \otimes (G(\theta^{(j)}) - \bar{G})$$

Analysis step:
Update each ensemble member by

$$\theta_{n+1}^{(j)} = \theta_n^{(j)} + B_{n+1}^{\theta p}(B_{n+1}^{pp} + \Sigma)^{-1}(q_{n+1}^{(j)} - G(\theta_n^{(j)})), \tag{5.2}$$

where we consider perturbed observation

$$q_{n+1}^{(j)} = q + \zeta_{n+1}^{(j)}, \quad \xi_{n+1}^{(j)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma'). \tag{5.3}$$

Estimate: $\bar{\theta}_N = \frac{1}{J}\sum_{j=1}^{J} \theta_N^{(j)}$.

Note that the usage of Σ' in the perturbed observations gives us the possibility to consider different cases. By setting $\Sigma' = 0$, we end up in the unperturbed observations case, which results in the ODE setting and has been studied for EKI in [200]. In [41] the authors also focused mainly on the case of $\Sigma' = 0$. In this thesis, we will extend the results to the setting of

$$\Sigma' = \begin{pmatrix} \Gamma & 0 \\ 0 & C_{0,} \end{pmatrix}$$

resulting in an SDE driven by a Brownian motion in $\mathbb{R}^K$ and a Brownian motion in $\mathbb{R}^I$ independently.

Following the continuous-time interpretation of the EKI, the continuous-time limit of the TEKI is formally given by

$$\mathrm{d}\theta_t^{(j)} = B^{\theta p}(\theta_t)\Sigma^{-1}(q - G(\theta_t^{(j)})\,\mathrm{d}t + B^{\theta p}(\Sigma')^{-\frac{1}{2}}\,\mathrm{d}\tilde{W}_t^{(j)}, \tag{5.4}$$

where $\tilde{W}^{(j)}$ are independent Brownian motions on $R^K \times \mathbb{R}^I$. Note that we can write the sample covariances as

$$B^{pp}(\theta) = \begin{pmatrix} C^{pp}(\theta) & C^{p\theta}(\theta) \\ C^{\theta p}(\theta) & C^{\theta\theta}(\theta) \end{pmatrix}, \quad B^{\theta p}(\theta) = \begin{pmatrix} C^{\theta p}(\theta) \\ C^{\theta\theta}(\theta) \end{pmatrix},$$

with

$$\begin{aligned} C^{pp}(\theta) &= \frac{1}{J}\sum_{j=1}^{J}(H(\theta^{(j)}) - \bar{H}) \otimes (H(\theta^{(j)}) - \bar{H}), \\ C^{\theta p}(\theta) &= \frac{1}{J}\sum_{j=1}^{J}(\theta^{(j)} - \bar{\theta}) \otimes (H(\theta^{(j)}) - \bar{H}), \\ C^{\theta\theta}(\theta) &= \frac{1}{J}\sum_{j=1}^{J}(\theta^{(j)} - \bar{\theta}) \otimes (\theta^{(j)} - \bar{\theta}) \end{aligned}$$

Assuming

$$\Sigma' = \Sigma = \begin{pmatrix} \Gamma & 0 \\ 0 & \kappa^{-1}C_0 \end{pmatrix}$$

we can write (5.4) as

$$\begin{aligned} \mathrm{d}\theta_t^{(j)} = {} & C^{\theta p}(\theta_t)\Gamma^{-1}(y - H(\theta_t^{(j)}))\,\mathrm{d}t + \kappa C^{\theta\theta}(\theta_t)C_0^{-1}\theta_t^{(j)}\,\mathrm{d}t \\ & + C^{\theta p}(\theta_t)\Gamma^{-\frac{1}{2}}\,\mathrm{d}W_t^{(j)} + \kappa^{\frac{1}{2}}C^{\theta\theta}(\theta_t)C_0^{-\frac{1}{2}}\,\mathrm{d}\hat{W}_t^{(j)}, \end{aligned} \tag{5.5}$$

where $W^{(j)}$ are Brownian motions on $\mathbb{R}^K$ and $\hat{W}^{(j)}$ are Brownian motions on $\mathbb{R}^I$, which are all independent from each other.

Remark 5.1.1. *We note that Σ' might also be chosen as*

$$\Sigma' = \begin{pmatrix} \Gamma & 0 \\ 0 & 0 \end{pmatrix} \tag{5.6}$$

resulting formally in the system of SDEs

$$\begin{aligned} \mathrm{d}\theta_t^{(j)} = {} & C^{\theta p}(\theta_t)\Gamma^{-1}(y - H(\theta_t^{(j)}))\,\mathrm{d}t + \kappa C^{\theta\theta}(\theta_t)C_0^{-1}\theta_t^{(j)}\,\mathrm{d}t \\ & + C^{\theta p}(\theta_t)\Gamma^{-\frac{1}{2}}\,\mathrm{d}W_t^{(j)}, \end{aligned} \tag{5.7}$$

where the diffusion only takes part in the observation space through the Brownian motions $W^{(j)}$ in $\mathbb{R}^K$. This method corresponds to the scheme, where we consider the perturbation (5.3) *only in the observations and no pertubations in the particles itself.*
The results based on (5.5) *presented in the following part, can be straightforwardly obtained similar for* (5.7) *by dropping the corresponding part resulting from the diffusion driven by $\hat{W}^{(j)}$. The adaptive choice of the regularization parameters will also be based on* (5.7).

Following the ideas of approximate preconditioned gradient structure of the EKI, in the long time behavior the TEKI aims to minimize the Tikhonov regularized loss functional

$$T_\kappa(\theta) = \frac{1}{2}\|H(\theta) - y\|^2 + \frac{\kappa}{2}\|\theta\|^2_{C_0}. \tag{5.8}$$

Note that through the scaling by C_0 one can enforce regularity on θ. We use the definition of the sample covariances and write (5.5) as

$$\begin{aligned}
\mathrm{d}\theta_t^{(j)} =& \frac{1}{J}\sum_{k=1}^{J}\langle H(\theta^{(k)}) - \bar{H}, \Gamma^{-1}(y - H(\theta_t^{(j)}))\,\mathrm{d}t\rangle(\theta^{(k)} - \bar{\theta}) \\
&+\kappa\frac{1}{J}\sum_{k=1}^{J}\langle \theta^{(k)} - \bar{\theta}, C_0^{-1}\theta_t^{(j)}\,\mathrm{d}t\rangle(\theta^{(k)} - \bar{\theta}) \\
&+\frac{1}{J}\sum_{k=1}^{J}\langle H(\theta^{(k)}) - \bar{H}, \Gamma^{-\frac{1}{2}}\,\mathrm{d}W_t^{(j)}\rangle(\theta^{(k)} - \bar{\theta}) \\
&+\kappa^{\frac{1}{2}}\frac{1}{J}\sum_{k=1}^{J}\langle \theta^{(k)} - \bar{\theta}, C_0^{-\frac{1}{2}}\,\mathrm{d}\hat{W}_t^{(j)}\rangle(\theta^{(k)} - \bar{\theta}).
\end{aligned}$$

This formulation again reveals that solutions satisfy the well known subspace property of Lemma 3.3.1.

Lemma 5.1.2. *Assume that H is locally Lipschitz and let $\mathcal{S}$ be the linear span of $\{\theta_0^{(j)}\}_{j=1}^J$, then $\theta_t^{(j)} \in \mathcal{S}$ for all $(t, j) \in [0, \infty) \times \{1, \dots, J\}$ almost surely.*

Remark 5.1.3. *Considering* (5.5) *with a linear and bounded forward map $G(\cdot) = A\cdot$, we observe that we can mainly transfer the results of section 3.4 to the resulting SDE. Hence, we can on the one side ensure existence of unique strong solutions for* (5.5) *and on the other side we can omit the results for the ensemble collapse. Furthermore, by Lemma 3.3.1 it follows again, that one can assume without lose of generality a finite-dimensional parameter space X, see Lemma 3.3.2.*

5.2 Convergence results for fixed regularization parameter

For the convergence analysis of the TEKI, we will assume a linear forward model, i.e. we assume $H(\cdot) = L\cdot$ for some $L \in \mathcal{L}(\mathcal{X}, \mathbb{R}^K)$. Hence, denoting $C(\theta) = C^{\theta\theta}(\theta)$, we can write (5.5) as

$$\begin{aligned}
\mathrm{d}\theta_t^{(j)} &= C(\theta_t)L^*\Gamma^{-1}(y - L\theta_t^{(j)})\,\mathrm{d}t + \kappa C(\theta_t)C_0^{-1}\theta_t^{(j)}\,\mathrm{d}t \\
&\quad + C(\theta_t)\left(L^\top\Gamma^{-\frac{1}{2}}\,\mathrm{d}W_t^{(j)} + \kappa^{\frac{1}{2}}C_0^{-\frac{1}{2}}\,\mathrm{d}\hat{W}_t^{(j)}\right),
\end{aligned} \tag{5.9}$$

By taking the limits in (5.9), we now establish the existence and uniqueness of strong solutions.

Corollary 5.2.1. *Let $\theta_0 = (\theta_0^{(j)})_{j\in\{1,\dots,J\}}$ be $\mathcal{F}_0$-measurable maps $\theta_0^{(j)} : \Omega \to X$ which are linearly independent almost surely. Then for all $T \geq 0$ there exists a unique strong solution $(\theta_t)_{t\in[0,T]}$ (up to $\mathbb{P}$-indstinguishability) of the set of coupled SDEs* (5.9).

Proof. Since the set of coupled SDEs (5.9) can be viewed by (5.4) with $G : \mathbb{R}^I \to \mathbb{R}^K \times \mathbb{R}^I$ with

$$G(\theta) = \begin{bmatrix} L \\ I \end{bmatrix} \theta,$$

which is again a bounded and linear map, the result follows directly by application of Theorem 3.3.6. □

5.2.1 Quantification of the ensemble collapse

In comparison to the EKI without regularization, we are now able to prove the ensemble collapse in the parameter space.

Corollary 5.2.2. *Let $(\theta_0^{(j)})_{j\in\{1,\dots,J\}}$ be $\mathcal{F}_0$-measurable maps $\theta_0^{(j)} : \Omega \to X$ such that $K_0 = \mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J} |\Sigma^{-\frac{1}{2}} G(e_0^{(j)})|^2\right] < \infty$, then it holds true that*

$$\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J} |\Sigma^{-\frac{1}{2}} G(e_t^{(j)})|^2\right] \leq \frac{1}{C_0^{-1} + \frac{J+1}{J^2}t}.$$

Furthermore, it follows

$$\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J} |e_t^{(j)}|^2\right] \in \mathcal{O}\left(\frac{1}{t}\right).$$

Proof. The first assertion follows by Theorem 3.4.6 and the second assertion follows by the definition of G. □

Similarly, we can also ensure the almost sure ensemble collapse in the observation space as well as in the parameter space, see Theorem 3.4.12.

5.2.2 Convergence of the regularized loss function

In the following we will present convergence results regarding the regularized loss function given by (5.8). First of all, by application of Proposition 3.4.16 we can ensure that

$$\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J} T_\kappa(\theta_t^{(j)})\right],$$

is monotonically decreasing in time. Furthermore, we will analyse

$$\widetilde{r}_t^{(j)} := \theta_t^{(j)} - \theta^*,$$

where θ^* is the global minimizer of T_κ, i.e. θ^* is given by the Tikhonov regularized solution

$$\theta^* := \left(L^\top \Gamma^{-1} L + C_0^{-1}\right)^{-1} L^\top \Gamma^{-1} y,$$

and satisfies

$$\nabla T_\kappa(\theta^*) = 0.$$

Theorem 5.2.3. *Assume that y are noisy measurements of the true parameter $\theta^\dagger$ under L, i.e. $y = L\theta^\dagger + \eta^\dagger$, where $\eta^\dagger \in \mathbb{R}^K$ denotes a realization of noise and let $\theta_0 = (\theta_0^{(j)})_{j\in\{1,\dots,J\}}$ be $\mathcal{F}_0$-measurable maps $\theta_0^{(j)} : \Omega \to \mathbb{R}^K$ such that we have bounded moments $\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J} |\widetilde{r}_0^{(j)}|^2\right] < \infty$. Then $\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J} |\widetilde{r}_t^{(j)}|^2\right]$ is strictly monotonically decreasing in time.*

Proof. We assume w.l.o.g. $\kappa = 1$, as we can extend the result to arbitrary $\kappa > 0$ by transforming $C_0 \mapsto \frac{1}{\kappa}C_0$. By application of Itô's formula to (5.9) we can derive the dynamics of $\|\widetilde{r}_t^{(j)}\|^2$ for fixed $j \in \{1,\dots,J\}$. The dynamics are given by

$$
\begin{aligned}
\mathrm{d}\|\widetilde{r}_t^{(j)}\|^2 =& 2\langle \theta_t^{(j)} - \theta^*, \mathrm{d}\theta_t^{(j)}\rangle + \langle \mathrm{d}\theta_t^{(j)}, \mathrm{d}\theta_t^{(j)}\rangle \\
=& \frac{2}{J}\sum_{k=1}^{J}\left(\langle L(\theta^{(k)} - \bar\theta), y - L\theta^{(j)}\rangle_\Gamma - \langle \theta^{(k)} - \bar\theta, \theta^{(j)}\rangle_{C_0}\right)\langle \theta^{(j)} - \theta^*, \theta^{(k)} - \bar\theta\rangle\,\mathrm{d}t \\
&+ \frac{1}{J^2}\sum_{k,l=1}^{J}\langle L(\theta^{(k)} - \bar\theta), \Gamma^{\frac12}\mathrm{d}W^{(j)}\rangle_\Gamma\langle L(\theta^{(l)} - \bar\theta), \Gamma^{\frac12}\mathrm{d}W^{(j)})\rangle_\Gamma\langle \theta^{(k)} - \bar\theta, \theta^{(l)} - \bar\theta\rangle\,\mathrm{d}t \\
&+ \frac{1}{J^2}\sum_{k,l=1}^{J}\langle \theta^{(k)} - \bar\theta, C_0^{\frac12}\mathrm{d}\hat W^{(j)}\rangle_{C_0}\langle \theta^{(l)} - \bar\theta, C_0^{\frac12}\mathrm{d}\hat W^{(j)})\rangle_{C_0}\langle \theta^{(k)} - \bar\theta, \theta^{(l)} - \bar\theta\rangle\,\mathrm{d}t \\
&+ \frac{1}{J}\sum_{k=1}^{J}\left(\langle L(\theta^{(k)} - \bar\theta), \Gamma^{\frac12}\mathrm{d}W^{(j)}\rangle_\Gamma\right)\langle \theta^{(k)} - \bar\theta, \theta^{(j)} - \theta^*\rangle \\
&+ \frac{1}{J}\sum_{k=1}^{J}\left(\langle \theta^{(k)} - \bar\theta, C_0^{\frac12}\mathrm{d}\hat W^{(j)}\rangle_{C_0}\right)\langle \theta^{(k)} - \bar\theta, \theta^{(j)} - \theta^*\rangle,
\end{aligned}
$$

which leads to

$$
\begin{aligned}
=& \frac{2}{J}\sum_{k=1}^{J}\left(\langle L(\theta^{(k)} - \bar\theta), y - L\theta^{(j)}\rangle_\Gamma - \langle \theta^{(k)} - \bar\theta, \theta^{(j)}\rangle_{C_0}\right)\langle \theta^{(j)} - \theta^*, \theta^{(k)} - \bar\theta\rangle\,\mathrm{d}t \\
&+ \frac{1}{J^2}\sum_{k,l=1}^{J}\langle L(\theta^{(k)} - \bar\theta), L(\theta^{(l)} - \bar\theta)\rangle_\Gamma\langle \theta^{(k)} - \bar\theta, \theta^{(l)} - \bar\theta\rangle\,\mathrm{d}t \\
&+ \frac{1}{J^2}\sum_{k,l=1}^{J}\langle \theta^{(k)} - \bar\theta, \theta^{(l)} - \bar\theta\rangle_\Gamma\langle \theta^{(k)} - \bar\theta, \theta^{(l)} - \bar\theta\rangle\,\mathrm{d}t \\
&+ \frac{1}{J}\sum_{k=1}^{J}\left(\langle L(\theta^{(k)} - \bar\theta), \Gamma^{\frac12}\mathrm{d}W^{(j)}\rangle_\Gamma\right)\langle \theta^{(k)} - \bar\theta, \theta^{(j)} - \theta^*\rangle \\
&+ \frac{1}{J}\sum_{k=1}^{J}\left(\langle \theta^{(k)} - \bar\theta, C_0^{\frac12}\mathrm{d}\hat W^{(j)}\rangle_{C_0}\right)\langle \theta^{(k)} - \bar\theta, \theta^{(j)} - \theta^*\rangle \\
=:& LV(\theta_t^{(j)})\,\mathrm{d}t + \mathrm{d}M_t.
\end{aligned}
$$

Note that integration over the stochastic part of this dynamic, denoted by dM_t, describes a local martingale. Hence, we will focus on $LV(\theta^{(j)})$. Using $\nabla T_\kappa(\theta^*) = 0$ and taking the

mean over all $j \in \{1, \dots, J\}$, we obtain

$$\begin{aligned}\frac{1}{J}\sum_{j=1}^{J} LV(\theta^{(j)}) = \Bigg(\frac{2}{J^2} \sum_{j,k=1}^{J} \Big(&\langle \theta^{(k)} - \bar{\theta}, L\Gamma^{-1}(y - L\theta^{(j)}) - C_0^{-1}\theta^{(j)} \\ &+ L^\top \Gamma^{-1}(L\theta^* - y) + C_0^{-1}\theta^* \rangle \langle \theta^{(j)} - \theta^*, \theta^{(k)} - \bar{\theta} \rangle \Big) \, \mathrm{d}t \\ &+ \frac{1}{J^2} \sum_{k,l=1}^{J} \langle \theta^{(k)} - \bar{\theta}, \left(L^\top \Gamma^{-1} L + C_0^{-1} \right) (\theta^{(l)} - \bar{\theta}) \rangle \langle \theta^{(k)} - \bar{\theta}, \theta^{(l)} - \bar{\theta} \rangle \, \mathrm{d}t \\ = \Bigg(&- \frac{2}{J^2} \sum_{j,k=1}^{J} \langle \theta^{(k)} - \bar{\theta}, \left(L\Gamma^{-1}L^\top + C_0^{-1} \right) (\theta^{(j)} - \theta^*) \rangle \cdot \langle \theta^{(j)} - \theta^*, \theta^{(k)} - \bar{\theta} \rangle \, \mathrm{d}t \Bigg) \\ &+ \frac{1}{J^2} \sum_{k,l=1}^{J} \langle \theta^{(k)} - \bar{\theta}, \left(L^\top \Gamma^{-1} L + C_0^{-1} \right) (\theta^{(l)} - \bar{\theta}) \rangle \langle \theta^{(k)} - \bar{\theta}, \theta^{(l)} - \bar{\theta} \rangle \, \mathrm{d}t,\end{aligned}$$

and by $\theta^{(l)} - \bar{\theta} = (\theta^{(l)} - \theta^*) - (\bar{\theta} - \theta^*)$ we obtain

$$\begin{aligned} = &- \frac{1}{J^2} \sum_{j,k=1}^{J} \langle \theta^{(k)} - \bar{\theta}, \left(L^\top \Gamma^{-1} L + C_0^{-1} \right) (\theta^{(j)} - \theta^*) \rangle \langle \theta^{(j)} - \theta^*, \theta^{(k)} - \bar{\theta} \rangle \, \mathrm{d}t \\ &- \frac{1}{J} \sum_{k=1}^{J} \langle \theta^{(k)} - \bar{\theta}, \left(L^\top \Gamma^{-1} L + C_0^{-1} \right) (\bar{\theta} - \theta^*) \rangle \langle \bar{\theta} - \theta^*, \theta^{(k)} - \bar{\theta} \rangle \, \mathrm{d}t. \end{aligned}$$

Similarly to the proof of Lemma 3.4.2, we consider a sequence of stopping times $(\tau_n)_{n\in\mathbb{N}}$ with $\tau_n \to \infty$ a.s., such that

$$M_{t\wedge\tau_n} = \int_0^{t\wedge\tau_n} \frac{1}{J} \sum_{k=1}^{J} \left(\langle L(\theta^{(k)} - \bar{\theta}), \Gamma^{\frac{1}{2}} \mathrm{d}W^{(j)} \rangle_\Gamma + \langle \theta^{(k)} - \bar{\theta}), C_0^{\frac{1}{2}} \mathrm{d}\hat{W}^{(j)} \rangle_{C_0} \right) \langle \theta^{(k)} - \bar{\theta}, \theta^{(j)} - \theta^* \rangle$$

is a martingale. We obtain for all for all $n \in \mathbb{N}$

$$\mathbb{E}\left[\frac{1}{J} \sum_{j=1}^{J} |\widetilde{r}_{t\wedge\tau_n}^{(j)}|^2 \right] = \mathbb{E}\left[\frac{1}{J} \sum_{j=1}^{J} |\widetilde{r}_0^{(j)}|^2 \right] + \mathbb{E}[\int_0^{t\wedge\tau_n} \frac{1}{J} \sum_{j=1}^{J} LV(\theta_s^{(j)}) \, ds],$$

Since $(L^\top \Gamma L + C_0^{-1})$ is strictly positive definite, by application of Lemma 3.3.5 we have that $\frac{1}{J}\sum_{j=1}^{J} LV(\theta^{(j)}) < 0$ and we obtain that φ is monotonically decreasing and bounded. Similar to the proof of Lemma 3.4.2, by Fatou's lemma and the monotone convergence theorem, we obtain in the limit

$$\mathbb{E}\left[\frac{1}{J} \sum_{j=1}^{J} |\widetilde{r}_{t+s}^{(j)}|^2 \right] \leq \mathbb{E}\left[\frac{1}{J} \sum_{j=1}^{J} |\widetilde{r}_s^{(j)}|^2 \right] + \int_s^t \mathbb{E}[\frac{1}{J} \sum_{j=1}^{J} LV(\theta_r^{(j)})] \, dr,$$

which implies that $\mathbb{E}\left[\frac{1}{J} \sum_{j=1}^{J} |\widetilde{r}_t^{(j)}|^2 \right]$ is strictly monotonically decreasing in time. □

Variance inflation Recalling the perspective of the EKI as preconditioned gradient flow, we will incorporate variance inflation into the algorithm in order to ensure convergence of the Tikhonov regularized loss function. In the case of the EKI without Tikhonov regularization, variance inflation lead to convergence of the residuals, see Theorem 3.4.17 and 3.4.19. By translating this to the setting with Tikhonov regularization, this would mean

$$\lim_{t\to\infty} \mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J} \mathcal{I}(u_t^{(j)}, y)\right] \to 0.$$

This above result is not desirable as it would suggest that $\lim_{t\to\infty} \theta_t = 0$, since G is mapping into a space which dimension is higher than $\mathbb{R}^I$ itself. The aim will be to prove, that the scheme will converge to the minimizer of $T_\kappa(\theta)$.
Under the assumption of the forward problem being linear and C_0 being strictly positive definite, it follows that the loss function T_κ is strongly convex, since

$$\nabla^2 T_\kappa = L^\top \Gamma^{-1} L + C_0^{-1} > 0.$$

We will denote the smallest eigenvalue of $\nabla^2 T_\kappa$ by $\lambda_{\min} > 0$. In order to ensure convergence to the unique minimizer of T_κ, we will incorporate variance inflation into the system of SDEs (5.9) in the following way

$$\begin{aligned} \mathrm{d}\theta_t^{(j)} = &\left(C(\theta_t) + \frac{1}{t^\alpha + R} B\right)\left(L^\top \Gamma^{-1}(y - L\theta_t^{(j)}) + C_0^{-1}\theta^{(j)}\right)\mathrm{d}t \\ &+ C(\theta)\left(L^\top \Gamma^{-1/2}\,\mathrm{d}W_t^{(j)} + \kappa^{\frac{1}{2}} C_0^{-\frac{1}{2}}\,\mathrm{d}\hat{W}_t^{(j)}\right), \end{aligned} \tag{5.10}$$

where $\alpha \in (0,1)$, $R > 0$ and B denotes a strictly positive definite matrix. Now that we have introduced variance inflation into scheme, we state the result of convergence towards the minimizer of the functional $T_\kappa(\theta)$. The intuition behind this result can be seen in the gradient flow structure of (5.10)

$$\mathrm{d}\theta_t^{(j)} = \left(C(\theta_t) + \frac{1}{t^\alpha + R} B\right)\nabla T_\kappa(\theta^{(j)})\,\mathrm{d}t + C(\theta)\left(L^\top \Gamma^{-1/2}\,\mathrm{d}W_t^{(j)} + \kappa^{\frac{1}{2}} C_0^{-\frac{1}{2}}\,\mathrm{d}\hat{W}_t^{(j)}\right).$$

Theorem 5.2.4. *Assume that y are noisy measurements of the true parameter $\theta^\dagger$ under L, i.e. $y = L\theta^\dagger + \eta^\dagger$, where $\eta^\dagger \in \mathbb{R}^K$ denotes a realization of noise and let $\theta_0 = (\theta_0^{(j)})_{j\in\{1,\dots,J\}}$ be $\mathcal{F}_0$-measurable maps $\theta_0^{(j)} : \Omega \to \mathbb{R}^K$ such that we have bounded moments $\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J} |\widetilde{r}_0^{(j)}|^2\right] < \infty$. Furthermore let $B \in \mathcal{L}(\mathbb{R}^I, \mathbb{R}^I)$ be a strictly positive definite operator. Then for all $\alpha > 0$ it holds true that*

$$\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J} |\widetilde{r}_t^{(j)}|^2\right] \in \mathcal{O}\left(t^{-(1-\alpha)}\right).$$

Proof. We again assume w.l.o.g. $\kappa = 1$. Using the results presented in the proof of Theorem 5.2.3, the SDE for $\frac{1}{J}\sum_{j=1}^{J} |\widetilde{r}_t^{(j)}|^2$ corresponding to (5.10) is given by

$$\mathrm{d}\frac{1}{J}\sum_{j=1}^{J} \|\widetilde{r}_t^{(j)}\|^2 = -\frac{1}{J^2}\sum_{j,k=1}^{J} \langle \theta^{(k)} - \bar{\theta}, \left(L^\top \Gamma^{-1} L + C_0^{-1}\right)(\theta^{(j)} - \theta^*)\rangle \langle \theta^{(j)} - \theta^*, \theta^{(k)} - \bar{\theta}\rangle\,\mathrm{d}t$$

$$
\begin{aligned}
&- \frac{1}{J}\sum_{k=1}^{J}\langle \theta^{(k)} - \bar{\theta}, \left(L^\top \Gamma^{-1} L + C_0^{-1}\right)(\bar{\theta} - \theta^*)\rangle \langle \bar{\theta} - \theta^*, \theta^{(k)} - \bar{\theta}\rangle \,\mathrm{d}t \\
&- \frac{1}{J}\frac{1}{t^\alpha + R}\sum_{j=1}^{J}\langle \theta^{(j)} - \theta^*, \left(L^\top \Gamma^{-1} L + C_0^{-1}\right)(\theta^{(j)} - \theta^*)\rangle \,\mathrm{d}t \\
&+ \frac{1}{J}\sum_{k=1}^{J}\left(\langle L(\theta^{(k)} - \bar{\theta}), \Gamma^{\frac{1}{2}}\mathrm{d}W^{(j)}\rangle_\Gamma\right)\langle \theta^{(k)} - \bar{\theta}, \theta^{(j)} - \theta^*\rangle \\
&+ \frac{1}{J}\sum_{k=1}^{J}\left(\langle \theta^{(k)} - \bar{\theta}, C_0^{\frac{1}{2}}\mathrm{d}\hat{W}^{(j)}\rangle_{C_0}\right)\langle \theta^{(k)} - \bar{\theta}, \theta^{(j)} - \theta^*\rangle.
\end{aligned}
$$

Taking the integral and expectation and similar localization arguments as in Lemma 3.4.2 leads to

$$h(t+s) \leq h(s) - \lambda_{\min}\int_s^t \frac{1}{r^\alpha + c} h(r)\, dr,$$

where we have set $t \mapsto h(t) := \mathbb{E}[\frac{1}{J}\sum_{j=1}^{J}\|\widetilde{r}_t^{(j)}\|^2]$ and $\lambda_{\min}$ denotes the smallest eigenvalue of $(L^\top \Gamma^{-1} L + C_0^{-1})$. It follows

$$\int_0^t \frac{1}{r^\alpha + c} h(r)\, dr \leq h(0),$$

for all $t \geq 0$. By using the monotonicity of $h(r)$, we obtain

$$\int_1^t \frac{1}{r^\alpha + c}\, dr h(t) \leq h(0),$$

and by using

$$\int_1^t \frac{1}{r^\alpha + c}\, dr \geq \frac{1}{(1+R)(1-\alpha)}(t^{1-\alpha} - 1)$$

we conclude with

$$h(r) \in \mathcal{O}(t^{-(1-\alpha)}).$$

□

5.3 Adaptive regularization parameter choice

An important point to consider, in the theory of regularization for inverse problems, is the choice of the regularization parameter. The parameter itself can depend largely on the problem itself and the specific form of regularization. In this section, we describe a specific way to find a good choice of the Tikhonov parameter κ based on the bilevel data-driven regularization introduced in Section 7.1. Figure 5.1 describes the task of adapting the regularization parameter withing the algorithm of EKI. While in the algorithm presented in section 5.1 we kept the regularization parameter κ fixed, we now consider approaches where we adapt the regularization parameter between the prediction step and the update step.

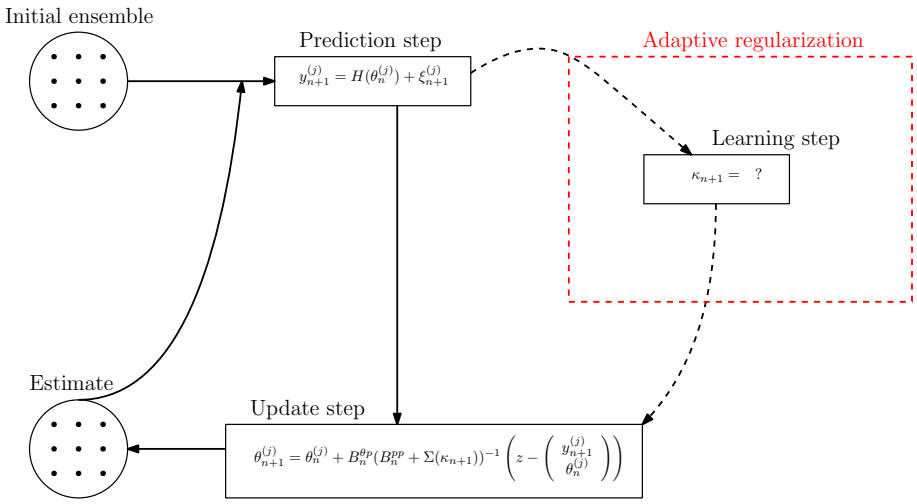

Figure 5.1: Description of the task of choosing the regularization parameter adaptively.

- The first method is based on a bilevel optimization problem (7.1). We will mainly present some heuristic idea of incorporation of the bilevel optimization approach to learn the regularization parameter within TEKI. We will give theoretical verification for the regularization based on bilevel optimization in general in section 7.1. While in section 7.1 we will assume to have access to training data, we will use our prediction step to generate artificial training data which will then be used to adapt the regularization parameter.

- The second method is based on the MAP formulation in the Bayesian framework of inverse problems.

We note that the presented methods were build up independently of the recently proposed adaptive regularization methods for EKI provided in [111], which is motivated from the viewpoint of the EKI as Gaussian approximation in a Bayesian tempering setting and further, incorporates early stopping criteria.

5.3.1 Data-driven regularization within EKI

Recall that through the Bayesian setting of the inverse problem

$$Y = H(\Theta) + \Xi,$$

we view Θ and Ξ as independent random variables distributed by $\mathcal{N}(0, \kappa^{-1}C_0) \otimes \mathcal{N}(0, \Gamma)$. To get access to training data, we can draw $(\theta^{(j)})_{j=1}^J$ samples of the prior distribution, $(\xi^{(j)})_{j=1}^J$ realizations of the assumed noise and compute

$$y^{(j)} = H(\theta^{(j)}) + \xi^{(j)}.$$

To incorporate those ideas of learning the regularization parameter from *training data*, we will give an alternative view point of EKI with perturbed observations. Instead of considering perturbations directly to the true observation as in (3.5), we will now view the perturbation as producing training data in each iteration.

Furthermore, instead of computing an optimal regularization parameter only at the beginning of the methods, we assume in each iteration that our current ensemble of particles represents current prior information in the form of an empirical distribution

$$\mathbb{Q}_n = \frac{1}{J}\sum_{j=1}^{J} \delta_{\theta_n^{(j)}}.$$

We will first introduce the basic idea of the adaptive regularization scheme in the linear setting, i.e. we assume $H(\cdot) = L\cdot$ for $L \in \mathcal{L}(\mathcal{X}, \mathbb{R}^K)$. Afterwards we will extend this idea to the general nonlinear setting.
We view $(\Theta, \Xi) \sim \mathbb{Q}_n \otimes \mathcal{N}(0, \Gamma)$ and compute

$$\widehat{\kappa}_{n+1}^J = \arg\min_{\kappa} \mathbb{E}_{(\Theta,\Xi)}[\|(L^\top\Gamma^{-1}L + \kappa C_0^{-1})^{-1}L^\top\Gamma^{-1}(L\Theta + \Xi) - \Theta\|^2],$$

where $(L^\top\Gamma^{-1}L + \kappa C_0^{-1})^{-1}L^\top\Gamma^{-1}(L\theta + \xi)$ denotes the Tikhonov regularized solution for realized θ and ξ, i.e. for the generated data $y = L\theta + \xi$. Using the artificial data we approximate the expected value empirically, i.e.

$$\begin{aligned}&\mathbb{E}_{(\Theta,\Xi)}[\|(L^\top\Gamma^{-1}L + \kappa C_0^{-1})^{-1}L^\top\Gamma^{-1}(L\Theta + \Xi) - \Theta\|^2] \\ &\approx \frac{1}{J}\sum_{j=1}^{J} \|(L^\top\Gamma^{-1}L + \kappa C_0^{-1})^{-1}L^\top\Gamma^{-1}y_{n+1}^{(j)} - \theta_n^{(j)}\|^2.\end{aligned} \tag{5.11}$$

Our training data is produced by perturbing the particles mapped by the forward operator,

$$y_{n+1}^{(j)} = L\theta_n^{(j)} + \xi_{n+1}^{(j)}.$$

Following the ideas of [10, 47] we now employ a way to update the regularization parameter κ_n in each iteration by seeking to decrease (5.11) adaptively. To do so, we will do in each update step a gradient descent step of

$$f_{n+1}(\kappa) := \frac{1}{J}\sum_{j=1}^{J} \|(L^\top\Gamma^{-1}L + \kappa C_0^{-1})^{-1}L^\top\Gamma^{-1}y_{n+1}^{(j)} - \theta_n^{(j)}\|^2,$$

with respect to κ. We define

$$v_{n+1}^{(j)}(\kappa) := \theta_\kappa(y_{n+1}^{(j)}) - \theta_n^{(j)}, \tag{5.12}$$

which represents the difference of the current particle $\theta_n^{(j)}$ to the minimizer $\theta_\kappa(y_{n+1}^{(j)}) = (L^\top\Gamma^{-1}L + \kappa C_0^{-1})^{-1}L^\top\Gamma^{-1}y_{n+1}^{(j)}$ of the Tikhonov regularized loss function. From (5.12) we can write the loss function

$$f_{n+1}(\kappa) := \frac{1}{J}\sum_{j=1}^{J} \frac{1}{2}\|v_{n+1}^{(j)}(\kappa)\|^2, \tag{5.13}$$

where for simplicity we now drop the dependance of n and j. To do a gradient descent with respect to (5.13) we need to compute its derivative which means we also need to compute the derivative of (5.12). To proceed we compute both $f'(\kappa)$ and $v'(\kappa)$. To aid we use the chain rule,

$$\frac{d\|v_{n+1}^{(j)}(\kappa)\|^2}{d\kappa} = (v_{n+1}^{(j)})^\top(\kappa) \cdot (v_{n+1}^{(j)})'(\kappa), \tag{5.14}$$

and compute

$$\begin{aligned}(v_{n+1}^{(j)})'(\kappa) = \frac{d\theta_{n+1}^{(j)}(\kappa)}{d\kappa} &= \frac{d(L^\top\Gamma^{-1}L + \kappa C_0^{-1})^{-1}L^\top\Gamma^{-1}y_{n+1}^{(j)}}{d\kappa}, \\ &= \frac{d(L^\top\Gamma^{-1}L + \kappa C_0^{-1})^{-1}}{d\kappa}L^\top\Gamma^{-1}y_{n+1}^{(j)}, \\ &= -(L^\top\Gamma^{-1}L + \kappa C_0^{-1})^{-1}C_0^{-1}(L^\top\Gamma^{-1}L + \kappa C_0^{-1})^{-1} \\ &\cdot L^\top\Gamma^{-1}y_{n+1}^{(j)}.\end{aligned}$$

Therefore, using the expression for the derivative of v, we can now express the derivative of (5.14) as

$$\begin{aligned}\frac{d\|v_{n+1}^{(j)}(\kappa)\|^2}{d\kappa} = &-\left((L^\top\Gamma^{-1}L + \kappa C_0^{-1})^{-1}L^\top\Gamma^{-1}y_{n+1}^{(j)} - \theta_n^{(j)}\right)(L^\top\Gamma^{-1}L + \kappa C_0^{-1})^{-1} \\ &\cdot C_0^{-1}(L^\top\Gamma^{-1}L + \kappa C_0^{-1})^{-1}L^\top\Gamma^{-1}y_{n+1}^{(j)}.\end{aligned}$$

To simplify the derivation described above we present Tikhonov EKI with adaptive regularization in Algorithm 7. Note that we only consider perturbation in the observation y described by the choice of Σ' defined in (5.6).

Algorithm 7: Linear TEKI: adaptive learning regularization

Input: initial ensemble $(\theta_0^{(j)})_{j=1}^J$, $\kappa_0 = 1$, observation y

Output: $\bar{\theta}_N$

for $n = 0, \dots, N-1$ **do**

Learning step:

Construct training data by

$$y_{n+1}^{(j)} = L\theta_n^{(j)} + \xi_{n+1}^{(j)}$$

and adaptively learn κ_n by updating

$$\kappa_{n+1} = \kappa_n - \gamma_n \cdot f'_{n+1}(\kappa_n),$$

Update step:

Update the ensemble of particle by using the TEKI update formula (5.2)

$$\theta_{n+1}^{(j)} = \theta_n^{(j)} + B_n^{\theta p}\big(B_n^{pp} + \Sigma(\kappa_{n+1})\big)^{-1}\left(q - \begin{pmatrix} y_{n+1}^{(j)} \\ \theta_n^{(j)} \end{pmatrix}\right).$$

Estimate: $\bar{\theta}_N = \frac{1}{J}\sum_{j=1}^{J}\theta_N^{(j)}$.

Remark 5.3.1. *An important question to ask from Algorithm 7 is how to choose the step size γ_n. It is well known in optimization that it can be beneficial to choose a non-fixed step size for maximum learning. Our choice for γ_n will be based on the Armijo rule to ensure that we have a correct descent direction at every iteration.*

With the introduced method of learning the regularization parameter κ, we can now fill in the gap of Figure 5.1 by the following Figure 5.2, which explains the idea of learning the regularization parameter by the use of artificial training data.

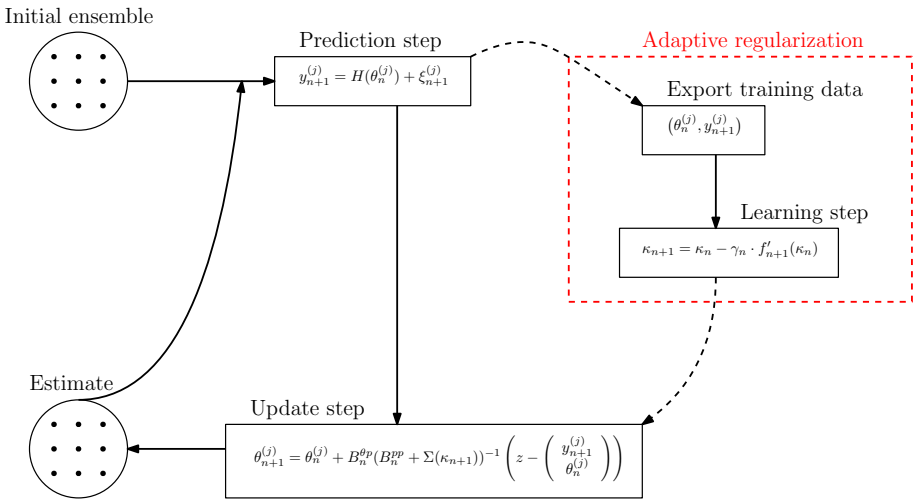

Figure 5.2: Representation of adaptive regularized ensemble Kalman inversion with the inclusion of data-driven learning.

5.3.2 Generalization to nonlinear setting

While for linear forward models we have used the closed expression of the Tikhonov regularized solution, we are not able to use this expression in the nonlinear setting. To avoid this issue, we will present another way of choosing the regularization parameter adaptively.

For our first method, we will make use of the data-driven regularization approach, which will be studied in more detail in section 7.1. In particular, we consider the following bilevel optimization problem in a general nonlinear setting with Tikhonov regularization, i.e.

$$\widehat{\kappa} \in \arg\min_{\kappa>0} \mathbb{E}_{\mathbb{Q}(\Theta,Y)}[\|R_\kappa(Y) - \Theta\|^2]$$
$$R_\kappa(y) := \arg\min_{\theta\in\mathbb{R}^I} \frac{1}{2}\|H(\theta) - y\|_\Gamma^2 + \frac{\kappa}{2}\|\theta\|_{C_0}^2,$$

where we view $(\Theta, Y) \sim \mathbb{Q}(\Theta, Y)$ as random variables. We assume that we have given the current ensemble of particles $(\theta_n^{(j)})_{j=1}^J$, which represent current information about the unknown true parameter $\theta^\dagger$. Furthermore, assume that we have given a current regularization parameter κ_n. The method is similarly to the previous one based on learning the regularization parameter over time with the help of artificial training data $(\theta_n^{(j)}, y_{n+1}^{(j)})$, constructed in the prediction step. The update step (5.2) pushes the current ensemble to $(\theta_{n+1}^{(j)})_{j=1}^J$ in order to get closer to the minimizer of the Tikhonov functional, i.e. into direction of $R_{\kappa_{n+1}}(y_{n+1}^{(j)})$. In the linear setting we have chosen κ_{n+1} minimizing the difference of the Tikhonov regularized solution to the particles itself. Using an empirical approximation we aim to choose κ_{n+1} minimizing the difference

$$v_{n+1}(\kappa) = \frac{1}{J}\sum_{j=1}^{J} \frac{1}{2}\|R_\kappa(y_{n+1}^{(j)}) - \theta_n^{(j)}\|^2.$$

We will use the updated ensemble $(\theta_{n+1}^{(j)})_{j=1}^J$ to approximate the minimizer of the Tikhonov functional $(\theta_{n+1}^{(j)}(\kappa_n))_{j=1}^J$. If our update step becomes closer to be stationary, this approxi-

mation will get more accurate, since the Tikhonov functional is strictly convex. This leads to choosing κ_{n+1} by minimizing

$$g_{n+1}(\kappa) := \frac{1}{J}\sum_{j=1}^{J}\frac{1}{2}\|\theta_n^{(j)} - \theta_{n+1}^{(j)}\|^2 = \frac{1}{J}\sum_{j=1}^{J}\frac{1}{2}\|B_n^{\theta p}\big(B_n^{pp} + \Sigma(\kappa)\big)^{-1}(z - z_{n+1}^{(j)})\|^2,$$

where we have defined $z_{n+1}^{(j)} = \begin{pmatrix} y_{n+1}^{(j)} \\ \theta_n^{(j)} \end{pmatrix}$ and

$$\Sigma(\kappa) = \begin{bmatrix} \Gamma & 0 \\ 0 & \kappa^{-1}C_0 \end{bmatrix},$$

to emphasize the dependence of Σ from κ. To update κ we will use again a gradient descent step in each iteration. Therefore, we will compute the derivative of $g_{n+1}^{(j)}(\kappa) := \frac{1}{2}\|\theta_n^{(j)} - \theta_{n+1}^{(j)}\|^2$ by

$$(g_{n+1}^{(j)})'(\kappa) = (z - z_{n+1}^{(j)})^\top (B_n^{pp} + \Sigma(\kappa))^{-1} B_n^{p\theta} B_n^{\theta p} \frac{\mathrm{d}(B_n^{pp} + \Sigma(\kappa))^{-1}}{\mathrm{d}\kappa}(z - z_{n+1}^{(j)}).$$

Defining $\widehat{C}_0 = \begin{pmatrix} 0 & 0 \\ 0 & C_0 \end{pmatrix}$ gives

$$\begin{aligned}(g_{n+1}^{(j)})'(\kappa) = \frac{1}{\kappa^2}(z - z_{n+1}^{(j)})^\top (B_n^{pp} + \Sigma(\kappa))^{-1} B_n^{p\theta} B_n^{\theta p} (B_n^{pp} + \Sigma(\kappa))^{-1} \widehat{C}_0 (B_n^{pp} + \Sigma(\kappa))^{-1} \\ \cdot (z - z_{n+1}^{(j)}),\end{aligned}$$

and we conclude with the update step

$$\kappa_{n+1} = \kappa_n - \gamma_n \cdot g'_{n+1}(\kappa_n),$$

where $g'_{n+1}(\kappa) = \frac{1}{J}\sum_{j=1}^{J}(g_{n+1}^{(j)})'(\kappa)$ and γ_n denotes a step size. To simplify the derivation described above we present in algorithmic form the TEKI with adaptive regularization in general nonlinear setting through Algorithm 8. We again consider pertubation only in the observation y described by the choice of Σ' defined in (5.6).

5.3.3 MAP formulation

We describe another way to find the parameter, which is based on a Hierarchical Bayesian framework. Given some initial guess on κ which we can define through a prior of the form $\kappa \sim \mathcal{U}[\kappa_l, \kappa_u]$, we define the parameter estimation as MAP estimate

$$\underset{\theta\in\mathbb{R}^I, \kappa\in[\kappa_l,\kappa_u]}{\arg\max} \frac{1}{\sqrt{\det(2\pi\Gamma)}}\exp(-\frac{1}{2}\|H(\theta) - y\|_\Gamma^2)\frac{1}{\sqrt{\det(2\pi\kappa^{-1}C_0)}}\exp(-\frac{1}{2}\|\theta\|_{\kappa^{-1}C_0}^2).$$

Or alternatively by taking the logarithm and ignoring the constants

$$\underset{\theta\in\mathbb{R}^I, \kappa\in[\kappa_l,\kappa_u]}{\arg\min} \frac{1}{2}\|H(\theta) - y\|_\Gamma^2 + \frac{\kappa}{2}\|\theta\|_{C_0}^2 - \frac{d_\theta}{2}\log\kappa,$$

Algorithm 8: Nonlinear TEKI: adaptive learning regularization

Input: initial ensemble $(\theta_0^{(j)})_{j=1}^J$, $\kappa_0 = 1$

Output: $\bar{\theta}_N$

for $n = 0, \dots, N-1$ **do**

Learning step:

Construct training data by

$$y_{n+1}^{(j)} = H(\theta_n^{(j)}) + \xi_{n+1}^{(j)}$$

and adaptively learn κ_n by updating

$$\kappa_{n+1} = \kappa_n - \gamma_n \cdot g'_{n+1}(\kappa_n),$$

Update step:

Update the ensemble of particle by using the TEKI update formula (5.2)

$$\theta_{n+1}^{(j)} = \theta_n^{(j)} + B_n^{\theta p}\big(B_n^{pp} + \Sigma(\kappa_{n+1})\big)^{-1}\left(q - \begin{pmatrix} y_{n+1}^{(j)} \\ \theta_n^{(j)} \end{pmatrix}\right).$$

Estimate: $\bar{\theta}_N = \frac{1}{J}\sum_{j=1}^{J} \theta_N^{(j)}$.

where d_θ denotes the dimension of θ. Notice that when θ is given, the minimizer of κ is explicitly found using critical point

$$\kappa_* = \left(\frac{1}{d_\theta}\|\theta\|_{C_0}^2\right)^{-1}.$$

Viewing each update step of the Tikhonov EKI as step into direction of the MAP estimator, leads to the following update

$$\kappa_{n+1} = \left(\frac{1}{d_\theta}\|\bar{\theta}_n\|_{C_0}^2\right)^{-1}, \quad \text{or} \quad \kappa_{n+1} = \left(\frac{1}{Jd_\theta}\sum_{j=1}^{J}\|\theta_n^{(j)}\|_{C_0}^2\right)^{-1}.$$

5.4 Numerical results

In this section, we numerically test and implement the adaptive strategies discussed in Section 5.3. As our analysis for fixed regularization parameter is based on the linear case, we will test our algorithms on the linear PDE introduced in subsection 2.1.15 in equation (2.11). We also test our algorithms on two nonlinear problems, the first one will be an analogous version of PDE based nonlinear example introduced in subsection 4.4.2 with one-dimensional domain and the second nonlinear example will be based on training a DNN, as introduced in example 2.1.16.

Algorithm 9: TEKI: adaptive learning regularization using the MAP

Input: initial ensemble $(\theta_0^{(j)})_{j=1}^J$, $\kappa_0 = 1$

Output: $\bar{\theta}_N$

for $n = 0, \dots, N-1$ **do**

Learning step:

Compute

$$\kappa_{n+1} = \left(\frac{1}{Jd_\theta} \sum_{j=1}^{J} \|\theta_n^{(j)}\|_{C_0}^2 \right)^{-1}.$$

Update step:

Update the ensemble of particle by using the TEKI update formula (5.2)

$$\theta_{n+1}^{(j)} = \theta_n^{(j)} + B_n^{\theta p} \big(B_n^{pp} + \Sigma(\kappa_{n+1}) \big)^{-1} \left(q - \begin{pmatrix} y_{n+1}^{(j)} \\ \theta_n^{(j)} \end{pmatrix} \right).$$

Estimate: $\bar{\theta}_N = \frac{1}{J} \sum_{j=1}^{J} \theta_N^{(j)}$.

5.4.1 Linear PDE

In this section, we run numerical experiments to highlight the effect of Tikhonov regularization within EKI with noisy observations. Our motivation again to emphasize will be to see if the overfitting of data can be prevented, and also to see the effect of the learned regularization parameter κ. Our first set of experiments is to show, with the help of inverse elliptic PDE, that overfitting can be reduced in the noisy SDE case given through (5.9). Throughout our experiments we are interested in assessing the performance of the TEKI in the noisy case through:

1. Ensemble spread: $\mathbb{E}\big[\frac{1}{J} \sum_{j=1}^{J} |e^{(j)}|_2^2\big]$.
2. Data misfit: $\mathbb{E}\big[\frac{1}{J} \sum_{j=1}^{J} |L\theta^{(j)} - y|_\Gamma^2\big]$.
3. Tikhonov loss function: $\mathbb{E}\big[\frac{1}{J} \sum_{j=1}^{J} T_\kappa(\theta^{(j)})\big]$.
4. Residual: $\mathbb{E}\big[\frac{1}{J} \sum_{j=1}^{J} |\widetilde{r}^{(j)}|^2\big]$.

For the continuum limit (5.7) of the algorithm with Σ' defined in (5.6), we will use (5.2) as discretization method, where we only perturb y through (5.3). Here we have incorporated a stepsize by setting $\Sigma \mapsto h^{-1}\Sigma$ and $\Sigma' \mapsto h^{-1}\Sigma'$ respectively. In both cases we use an ensemble size of $J = 15$. Our forward model will be the linear 1D elliptic PDE introduced in (2.11), where we set a mesh size $h = 2^{-4}$ and $K = 2^3 - 1$ equispaced observation points. We specify the covariance of the noise as $\Gamma = \gamma^2 \cdot I$ where $\gamma = 0.1$ and consider the prior assumption

$$\theta_0 \sim \mathcal{N}(0, C_0),$$

where $C_0 := \kappa^\dagger \cdot 10 \cdot (-\frac{\mathrm{d}^2}{\mathrm{d}x^2})^{-0.5}$. For our numerical examples we will consider the true unknown parameter

$$\theta^\dagger \sim \mathcal{N}(0, (\kappa^\dagger)^{-1} \cdot C_0),$$

such that our aim will be to find regularization parameters close to $\kappa^\dagger$. For the variance inflation in all of our numerical results we choose the inflation factor to be $\alpha = 1/2$ and $R = 1$. In our linear numerical example since we can compute the difference of the Tikhonov minimizer $\theta_\kappa(y)$ to the known $\theta^\dagger$, we will compare all the results to the best possible approximation to be expected by

$$\kappa_{\text{best}} = \arg\min_{\kappa} \|\theta_\kappa(y) - \theta^\dagger\|^2.$$

For the fixed regularization comparison, we will choose $\kappa = 1$, i.e. in this case we "trust" the prior assumption. For each regularization algorithm we test two different examples which correspond to different values of $\kappa^\dagger$. These will be chosen as $\kappa^\dagger = 50,\ 0.04$. We will keep the number of paths and particles consistent for each example and algorithm, specified as $Q = 1000$ and $J = 15$.

Example 1:

We set $\kappa^\dagger = 50$ and run $Q = 1000$ paths and $J = 15$ particles.

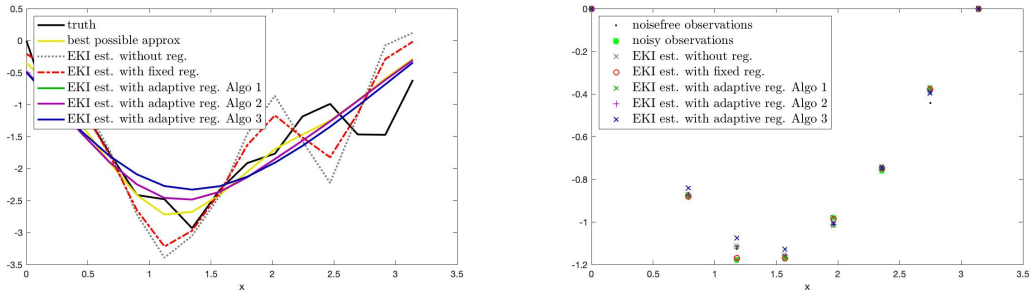

Figure 5.3: (T)EKI estimation for the different presented algorithms. $J = 15$ Particles and $Q = 1000$ paths has been simulated.

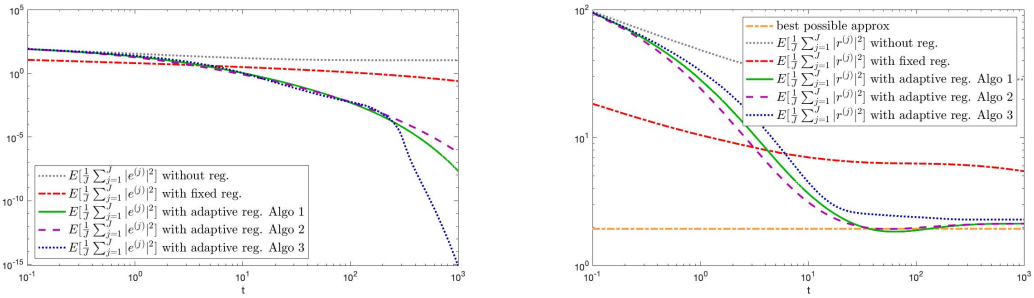

Figure 5.4: Spread of the particles and residuals for the different presented algorithms. $J = 15$ Particles and $Q = 1000$ paths has been simulated.

Our first results from the numerics constitute to the example where we take $\kappa^\dagger = 50$. The reconstruction of each algorithm compared to fixed and no regularization is shown

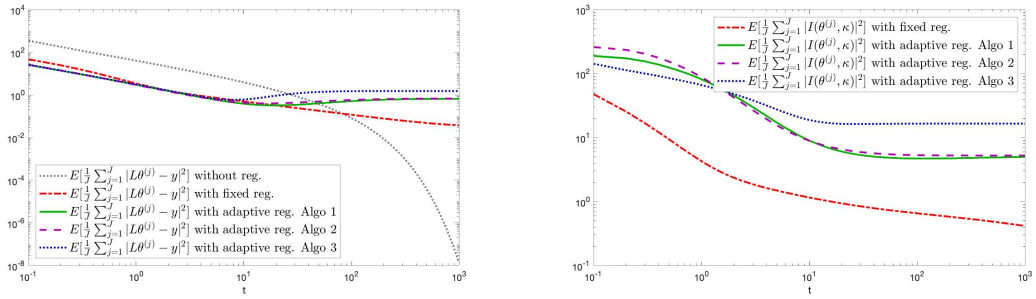

Figure 5.5: Data misfit and Tikhonov regularized loss for the different presented algorithms. $J = 15$ Particles and $Q = 1000$ paths has been simulated.

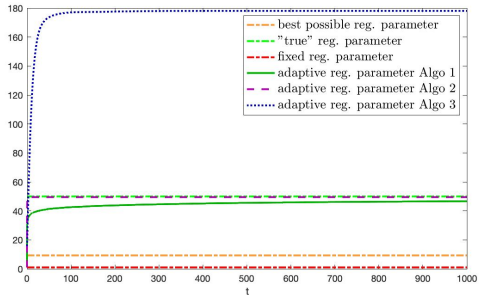

Figure 5.6: Learned regularization parameter for the different presented algorithms. J=15 Particles and $Q = 1000$ paths has been simulated.

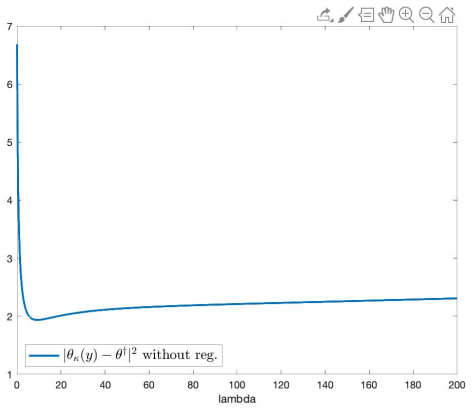

Figure 5.7: Plotted the difference of the Tikhonov minimizer $\theta_\kappa(y)$ to the "known" unknown true parameter $\theta^\dagger$.

in Figure 5.3. As we can see EKI with no regularization performs the worst with most variation followed by fixed. As expected the 3 algorithms based on adaptively learning the regularization parameter seem to perform similarly. However when we analyze Figures 5.4 - 5.6, we start to see a further discrepancies. With respect to the ensemble spread and the residuals we see Algorithm 9 perform betters. For the data misfit and regularized loss function we observe all adaptive algorithms do not overfit the data and perform better than using fixed regularization.

Example 2:

We set $\kappa^\dagger = 0.04$ and run $Q = 1000$ paths and $J = 15$ particles.

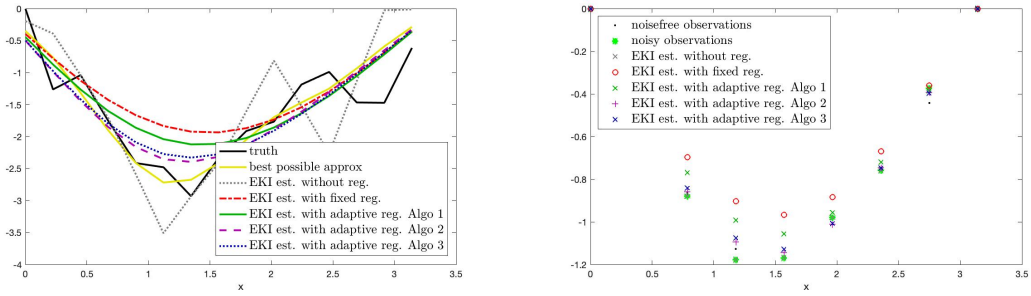

Figure 5.8: (T)EKI estimation for the different presented algorithms. $J = 15$ Particles and $Q = 1000$ paths has been simulated.

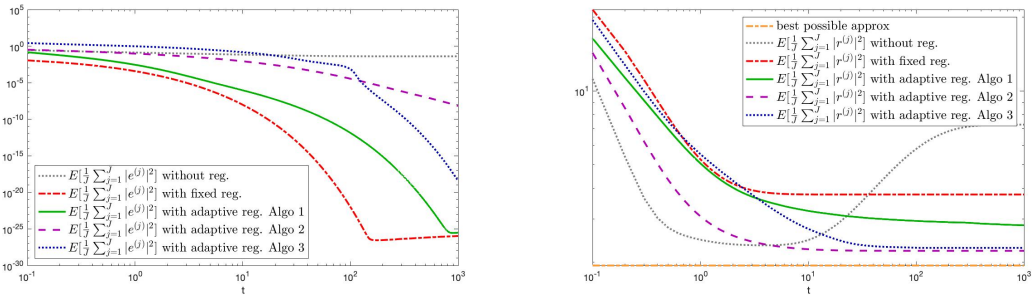

Figure 5.9: Spread of the particles and residuals for the different presented algorithms. $J = 15$ Particles and $Q = 1000$ paths has been simulated.

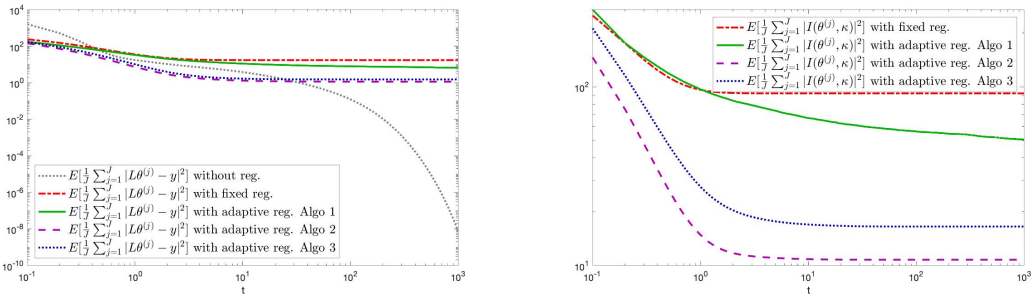

Figure 5.10: Data misfit and Tikhonov regularized loss for the different presented algorithms. $J = 15$ Particles and $Q = 1000$ paths has been simulated.

For the second example we modify the true regularization parameter to $\kappa^\dagger = 0.04$. We interestingly observe a different trend which is that Algorithm 8 outperforms TEKI with fixed regularization and Algorithms 7 and 9. This is seen through the reconstruction in Figure 5.8, and is verified further through Figures 5.9 - 5.11.

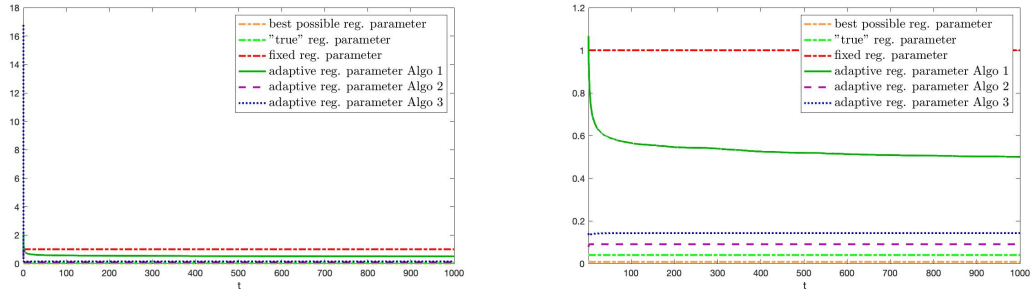

Figure 5.11: Learned regularization parameter (right: zoomed) for the different presented algorithms. $J = 15$ Particles and $Q = 1000$ paths has been simulated.

5.4.2 Darcy flow

As second model problem to test our regularization schemes we choose Darcys flow, which has already been considered in subsection 4.4.2. Given a source term f and permeability $a = \exp(\theta) \in L^\infty(D)$, then the forward problems is to solve the PDE

$$\begin{aligned} -\nabla \cdot (a \nabla p) &= f, \ x \in D = (0,1), \\ p &= 0, \ x \in \partial D \end{aligned}$$

for $p \in H_0^1(D)$ which is subject to zero Dirichlet boundary conditions. We choose a prior to $\theta \sim \mathcal{N}(0, C_0)$ where we define

$$C_0 := \kappa^\dagger \sigma^2 (I - \Delta)^{-\nu},$$

where σ^2 is a scaling constant, $\nu > 1/2$ is the smoothness of the prior and Δ is the Laplace operator in 1D and we will again simulate our prior through a KL expansion. We set $\kappa^\dagger = 5$, $\sigma = 2$, $\nu = 0.5$ and run $Q = 1000$ paths of the TEKI with $J = 100$ particles. The measurements noise is set to $N(0, 0.01 \cdot I)$ with $K = 16$ observation points. The physical domain $D = [0,1]$ has been discretizised by the equidistant grid $\{\frac{i}{2^5},\ i = 0, \dots, 2^5\}$. As before for the fixed regularization we choose $\kappa = 1$.
To incorporate variance inflation in a nonlinear setting we use the technique introduced in (3.15). We approximate the mixed sample covariance $B^{\theta p}(\theta)$ by

$$\mathcal{B}^{\theta p}(\theta) \approx \frac{1}{J} \sum_{j=1}^{J} (\theta^{(j)} - \bar{\theta})(\theta^{(j)} - \bar{\theta})^\top DG(\bar{\theta})^\top = C(\theta) DG(\bar{\theta})^\top,$$

where $G(\theta^{(j)}) - G(\bar{\theta}) \approx DG(\bar{\theta})(\theta^{(j)} - \bar{\theta})$ with denoting the derivative of G with respect to θ by $DG(\theta)$. The variance inflation now becomes

$$B^{\theta p}(\theta_t) \approx C(\theta_t) DG(\bar{\theta}_t)^\top \mapsto (C(\theta_t) + \vartheta(t) C_0) DG(\bar{\theta}_t)^\top, \ t \geq 0$$

such that we will write the SDE resulting from the TEKI update formula as

$$\mathrm{d}\theta^{(j)} = \left(\mathcal{B}^{\theta p}(\theta) - \frac{1}{t^\alpha + 1} C_0 DG(\bar{\theta})^\top \right) \Sigma^{-1} (q - G(\theta^{(j)})) \mathrm{d}t - C^{\theta p}(u) \Gamma^{-1/2} dW_t^{(j)},$$

where $\alpha > 0$ is a free parameter to choose for the variance inflation which scales the reduction of the variance inflation in time. For our numerical results we have set $\alpha = 0.5$.

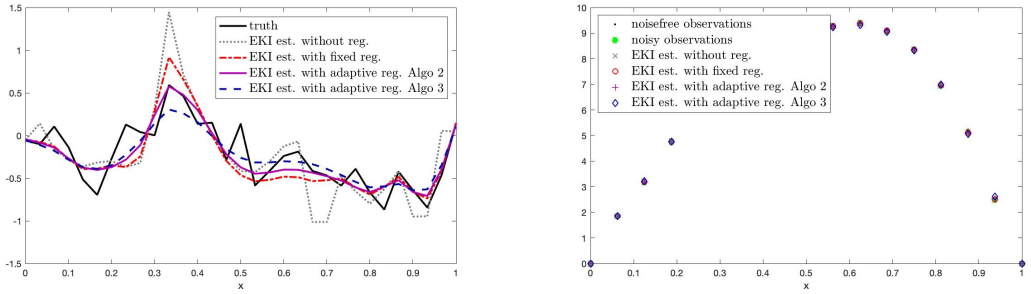

Figure 5.12: (T)EKI estimation for the different presented algorithms. $J = 100$ Particles and $Q = 1000$ paths has been simulated.

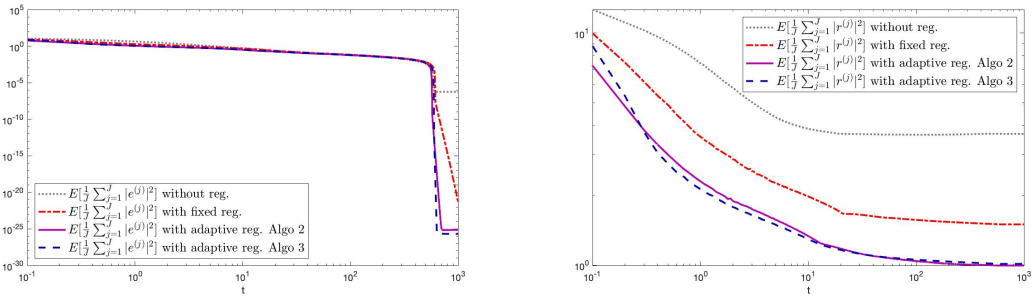

Figure 5.13: Spread of the particles and residuals for the different presented algorithms. $J = 100$ Particles and $Q = 1000$ paths has been simulated.

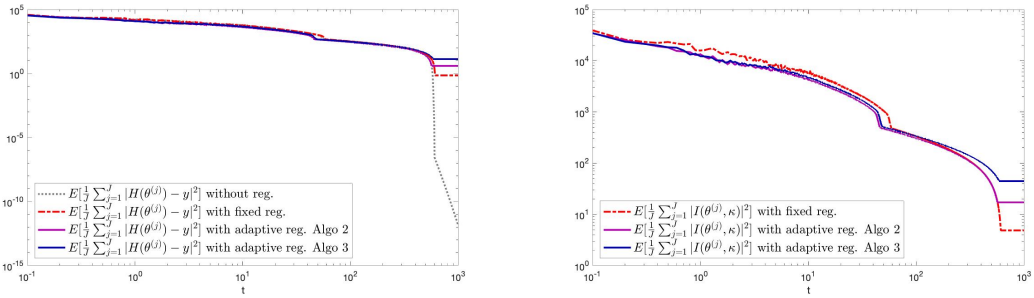

Figure 5.14: Data misfit and tikhonov regularized loss for the different presented algorithms. $J = 100$ Particles and $Q = 1000$ paths has been simulated.

5.4.3 Training of neural networks

Our final model problem will be motivated from machine learning where we consider a DNN. With the help of the NN we will try to learn the function $f : [-1, 1] \to \mathbb{R}$ defined as

$$f(x) = 5 \cdot \exp(-x^2) - 0.3,$$

given the training data set $\{x^k, f(x^k)\}_{k=1}^K$, see Example 2.1.16 for more details. By training the DNN with respect to the training data $\{x^k, y_k = \theta(x^k)\}_{k=1}^K$, we aim to solve the minimization problem (2.12) with the help of TEKI, where our training data is perturbed by some noise, i.e.

$$y_k = f(x_k) + \xi_k, \quad \xi_k \sim \mathcal{N}(0, \Gamma).$$

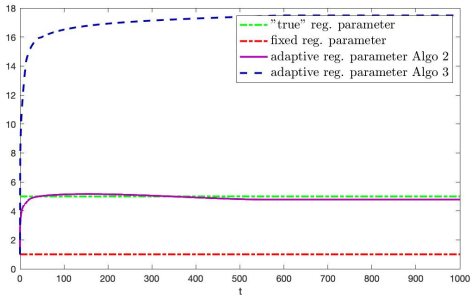

Figure 5.15: Learned regularization parameter for the different presented algorithms. $J = 100$ Particles and $Q = 1000$ paths has been simulated.

We will define our NN to approximate the function $f(x)$ with $L = 2$ hidden layers with $N_1 = 10$, $N_2 = 5$ hidden nodes and $N_3 = 1$ output node. Our choice of the activation function will be

$$\sigma(x) = \frac{1}{1 + e^{-x}},$$

which is the standard logistic function. To train our NN we will use the EKI to solve the inverse problem (2.13) to minimize (2.12) without regularization. In comparison to this we will use TEKI with fixed regularization and adaptively chosen regularization with $C_0 = 5 \cdot I$, where I denotes the $N_\theta \times N_\theta$ identity matrix. The measurement noise covariance has been chosen to be $\Gamma = 0.01 \cdot \mathrm{Id}_K$ with $K = 16$ observation points, while the true observation has been perturbed with noise $\xi \sim N(0, \mathrm{Id}_K)$. To measure the accuracy of our NN we will consider the quantity

$$r(\theta) = \frac{1}{K_{thin}} \sum_{i=1}^{K_{thin}} \|p_\theta(x_i) - f(x_i)\|^2,$$

with x_i chosen from a finer grid of $[-1, 1]$ with size $K_{thin} = 2^{10}$. For our numerical results we have set the parameter for variance inflation for $\alpha = 0.75$. Finally the fixed regularization parameter is again chosen as $\kappa = 1$.

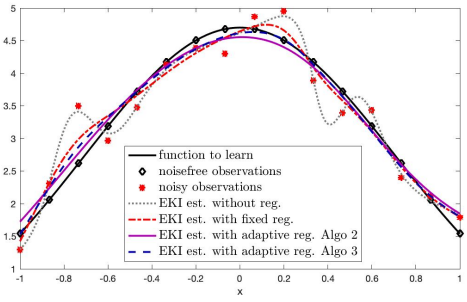

Figure 5.16: (T)EKI estimation for the different presented algorithms. $J = 100$ Particles and $Q = 1000$ paths has been simulated.

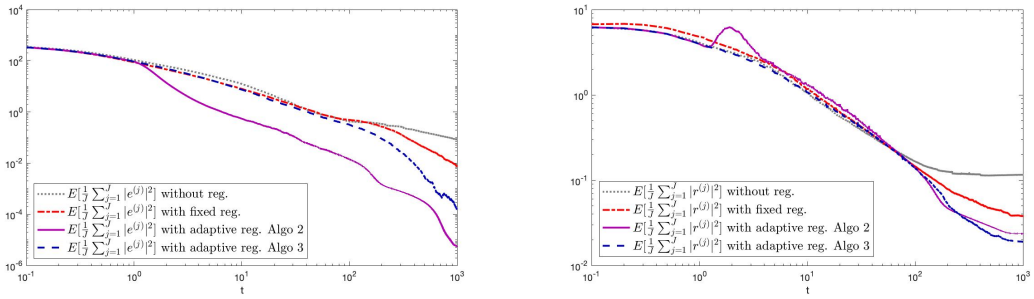

Figure 5.17: Spread of the particles and residuals for the different presented algorithms. $J = 100$ Particles and $Q = 1000$ paths has been simulated.

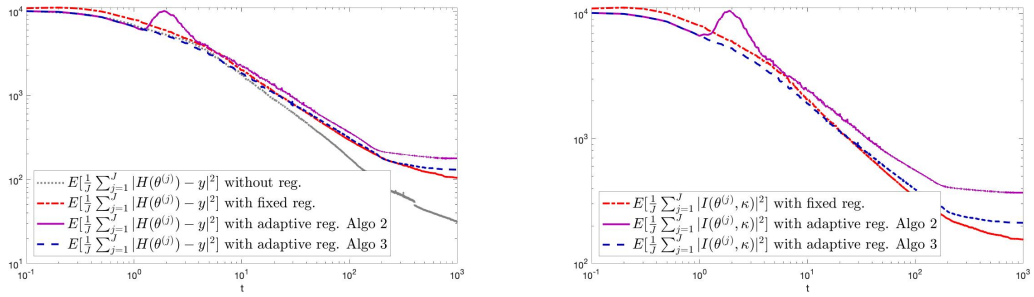

Figure 5.18: Data misfit and tikhonov regularized loss for the different presented algorithms. $J = 100$ Particles and $Q = 1000$ paths has been simulated.

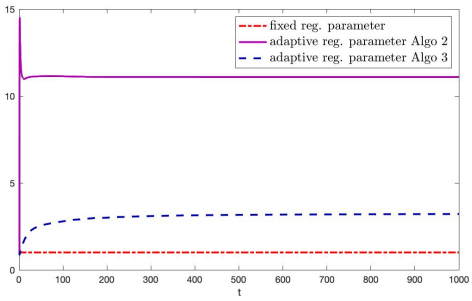

Figure 5.19: Learned regularization parameter for the different presented algorithms. $J = 100$ Particles and $Q = 1000$ paths has been simulated.

By analyzing both sets of nonlinear experiments, we see consistent results which show that Algorithm 8 slightly outperforms Algorithm 9, despite both working well. For the Darcy flow experiments we see the ensemble spread in Figure 5.13 is similar for both adaptive algorithms. However from Figures 5.14 - 5.15 we see that the Algorithm 8 correctly recovers the true regularization parameter and has a higher data misfit. From the NN experiments, we have no true regularization parameter $\kappa^\dagger$ but from Figures 5.17 - 5.18 we see as before that both algorithms perform well, with Algorithm 8 performing slightly better. Again as expected working adaptively overall achieves a better level of accuracy than fixed regularization.

6 Computational aspects for particle based sampling methods

While in the previous chapters the ensemble Kalman inversion as particle based optimization method for inverse problems delivered a point estimate, we will now consider different particle based sampling methods in order to solve the Bayesian inverse problem introduced in section 2.2. The presented ideas are closely related to the ideas of coupling of measures [62, 184, 160]. In particular, we will introduce 4 different particle based method, which are linked in the sense that all of the methods are seeking to convert a sample from the prior distribution into a sample from the posterior distribution.

- Langevin dynamic: Based on SDEs and their underlying Fokker–Planck equation, the idea is to find an equation such that the posterior is invariant under the driving stochastic process. In particular, Brownian dynamics can be viewed as a gradient flow in the space of probability measures [116], which minimises the Kullback–Leibler divergence between the current distribution of the underlying stochastic process and the posterior distribution. We will give a brief introduction to MCMC methods based on Langevin dynamics in Section 6.1.

- Ensemble Kalman sampler: The key idea of this approach is to run an interacting particles system through a preconditioned Langevin dynamics driven by Brownian motions, where the mean field limit corresponds to an SDE whose invariant distribution of the underlying Fokker–Planck equation is given by the posterior distribution. From an alternative perspective, this method can also be seen as a modification of the EKI method, where the noise arising in the particle system is shifted from pertubation of the observation to a pertubation of the particles itself. We will briefly discuss the approach introduced in [83] in Section 6.2.

- Gaussian approximation: In Section 6.3, we consider a particle system describing the mean and covariance of a Gaussian approximation to the posterior distribution. The evolution of the particle system can be found by minimizing the Kullback-Leibler divergence between the Gaussian approximation and the posterior distribution.

- Fokker–Planck particle systems: In this approach, we approximate the Fokker–Planck equation which arises from a Langevin dynamic in a reproducing Kernel Hilbert space. The key idea is again to minimize the Kullback–Leibler divergence between the underlying distribution of the dynamics and the posterior distribution. This time, the Kullback–Leibler divergence will be regularized in a reproducing Kernel Hilbert space and the distribution described by the Fokker–Planck equation will be approximated by the empirical measure of a particle system. The particles dynamic evolves in order to minimize the Kullback–Leibler divergence in the

reproducing Kernel Hilbert space. This approach will be discussed in Section 6.4.

We will identify these methods in a gradient flow structure and extend them in order to avoid the computation of the derivative of the forward model. This will lead to derivative-free sampling methods similar to the EKI as derivative-free optimization method.
Throughout this section, we assume that the posterior distribution $\mathbb{Q}_y^*$ defined in (2.16) can be represented by a probability density function ρ^* w.r.t. the Lebesgue measure on the finite-dimensional parameter space $\mathcal{X} = \mathbb{R}^I$, i.e. we can write

$$\mathbb{Q}_y^*(\mathrm{d}\theta) = \rho^*(\theta)\,\mathrm{d}\theta.$$

The aim of the methods presented in the following part will be to construct a time-dependent interacting particle system $(\theta_t^{(j)})_{j\in\{1,\dots,J\}}$, initialized by the prior distribution, with the property that this particle system approximates ρ^* as $t \to \infty$. We collect the particle system in one matrix $Z_t \in \mathcal{Z} = \mathbb{R}^{J\times I}$, i.e.

$$Z_t = \left((\theta_t^{(1)})^\top, \quad \dots, \quad (\theta_t^{(J)})^\top\right)^\top \tag{6.1}$$

and consider the gradient-based evolution equations of the form

$$\mathrm{d}Z_t = \mathcal{A}(Z_t)\nabla_z \mathcal{V}(Z_t)\,\mathrm{d}t + \Sigma(Z_t)\,\mathrm{d}W_t, \tag{6.2}$$

for some positive semi-definit matrix-valued preconditioner $\mathcal{A}(z) \in \mathcal{L}(\mathcal{Z},\mathcal{Z})$ and $\Sigma(z) \in \mathcal{L}(\mathcal{Z},\mathcal{Z})$, $z \in \mathcal{Z}$. By W we denote a Brownian motion on $\mathcal{Z}$ and $\mathcal{V}$ denotes a potential we are aiming to minimize in order to ensure the posterior approximation. The preconditioner $\mathcal{A}$ can be chosen in a way, such that the particles of the ensemble are interacting with each other. A common choice, which we have considered in the context of EKI, is a preconditioning with the sample covariance of the particle system. The potential $\mathcal{V}$ is chosen in a way to push the particle system into the "right" direction. For the Langevin dynamic, $\mathcal{V}$ will be a crucial part of the stationary distribution of the underlying SDE. For the Gaussian approximation or the Fokker–Planck particle system the potential $\mathcal{V}$ is designed such that the resulting flow aims to minimize a loss function, which is described through the Kullback–Leibler divergence. We will specify $\mathcal{V}$ for our presented algorithms. We start the discussion by introducing the 4 different particle based sampling methods in Section 6.1-6.4, and provide various numerical experiments in Section 6.6.

6.1 Langevin dynamics: A Markov chain Monte Carlo method

In this section, we introduce a Markov chain Monte Carlo method based on the Fokker-Planck equation. This method is based on the first-order Langevin dynamics (also called Brownian dynamics) defined as the following $\mathbb{R}^I$-valued SDE

$$\mathrm{d}\Theta_t = -\nabla_x \mathcal{V}(\Theta_t)\,\mathrm{d}t + \sqrt{2\beta^{-1}}\,\mathrm{d}W_t, \tag{6.3}$$

where W_t denotes a $\mathbb{R}^I$-dimensional Brownian motion, i.e $\mathcal{Z} = \mathbb{R}^I$ and $\mathcal{V}(\theta) := \frac{1}{2}\|\theta\|^2$ denotes the underlying potential, where the choice of $\mathcal{V}$ will be clear from equation (6.9). Following [177], we will introduce the MCMC method based on (6.3).
The generator of the process (Θ_t) driven by (6.3) is given by

$$\mathcal{L} = -\nabla_\theta \mathcal{V}(\theta)\cdot\nabla + \beta^{-1}\Delta. \tag{6.4}$$

Further, we can compute the time dependent probability density function ρ_t of Θ_t by the Fokker–Planck equation

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot (\nabla \mathcal{V} \rho_t) + \beta^{-1} \Delta \rho_t \tag{6.5}$$

initialized by the probability density function ρ_0 of Θ_0. The Fokker–Planck equation (6.5) in this particular form is often refered to Smoluchowski equation.

Remark 6.1.1. *In general the Fokker–Planck equation can formally be derived through the Kolmogorov equation. In the following we briefly discuss the connection between SDEs and PDEs, for more details we refer to [130]. We consider a SDE of general form*

$$dX_t = b(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}W_t, \tag{6.6}$$

where X takes values in $\mathbb{R}^d$, W is Brownian motion in $\mathbb{R}^K$, $b : \mathbb{R}^d \to \mathbb{R}^d$, and $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times K}$. The generator $\mathcal{L}$ of the SDE is defined by

$$\mathcal{L}f(x) = (b \cdot \nabla f)(x) + \frac{1}{2}\operatorname{trace}(\sigma^\top D^2 f \sigma)(x), \quad f \in C^2.$$

Under certain assumptions we obtain by Itô's formula

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[f(X_t)] = \mathbb{E}[\mathcal{L}f(X_t)]$$

for $f \in C_b^2$, which is also known as the Kolmogorov forward equation. If we assume that the distribution of X_t can be described by a probability density $\rho(t,\cdot)$, $t \geq 0$, we compute

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^d} f(x)\rho(t,x)\,\mathrm{d}x = \int_{\mathbb{R}^d} \mathcal{L}f(x)\rho(t,x)\,\mathrm{d}x = \int_{\mathbb{R}^d} f(x)\mathcal{L}^*\rho(t,x)\,\mathrm{d}x,$$

where $\mathcal{L}^*$ is the adjoint operator of $\mathcal{L}$. Note that we can compute

$$\mathcal{L}^* f(x) = -\nabla(f \cdot b)(x) + \operatorname{trace}(\sigma^\top D^2 f \sigma)(x).$$

From this computation we formally obtain the Fokker–Planck equation as

$$\partial_t \rho(t,x) = \mathcal{L}^*\rho(t,x), \quad \rho(0,x) = \rho_0(x). \tag{6.7}$$

With the help of the Fokker–Planck equation, one can solve the SDE (6.6) *by solving the PDE described by the Fokker–Planck equation* (6.7)*. It is also possible to extend this idea by solving a PDE through solving a SDE, where the connection is based on the Kolmogorov backward equation. We define $u(t,x) = \mathbb{E}_x[\varphi(X_t)]$, where $\mathbb{E}_x$ denotes the expected value under $X_0 = x$ a.s. and formulate the corresponding PDE for u. Under certain assumptions by the Markov property and application of Itô's formula it holds true that*

$$u(t+h,x) = \mathbb{E}_x[\mathbb{E}_x[f(X_{t+h}) \mid \mathcal{F}_h]] = \mathbb{E}[u(t,X_h)] = \mathbb{E}_x[u(t,X_0) + \int_0^h \mathcal{L}u(t,X_s)\,\mathrm{d}s],$$

where $\mathcal{F} = (\mathcal{F}_t)_{t\geq 0}$ denotes the filtration introduced by the SDE (6.6)*. Taking the limit $h \to 0$ formally leads to the Kolmogorov backward equation*

$$\partial_t u(t,x) = \mathcal{L}u(t,x), \quad u(0,x) = f(x). \tag{6.8}$$

Hence, instead of solving the PDE (6.8) *one can also solve the SDE* (6.6) *and compute $u(t,x) := \mathbb{E}_x[\varphi(X_t)]$. This connection between SDEs and PDEs will be the basic for the following particle based sampling methods.*

We want to apply the Fokker–Planck equation in order to produce samples of the posterior distribution in the Bayesian setting for inverse problems. The following Proposition states, that under suitable assumptions on $\mathcal{V}$ it is possible to use the invariant distribution of the process following the dynamics (6.3) to construct samples of probability densities of the form

$$\rho_\beta(\theta) = \frac{1}{Z}\exp(-\beta\mathcal{V}(\theta)), \tag{6.9}$$

where $Z = \int_{\mathbb{R}^I} \exp(-\beta\mathcal{V}(\theta))\,\mathrm{d}\theta$ is a normalization constant.

Proposition 6.1.2 ([177, Proposition 4.6])**.** *Assume that* $\exp(-\beta\mathcal{V}(\cdot))$ *is integrable for all* $\beta > 0$*. Then the Markov process with generator* (6.4) *is ergodic and the unique invariant distribution is given by* ρ_β*.*

Further, under sufficient conditions on $\mathcal{V}$ the solution of the Fokker–Planck equation (6.5) converges exponentially fast to its equilibrium.

Theorem 6.1.3 ([177, Theorem 4.9])**.** *Suppose that* $\mathcal{V} \in C^2(\mathbb{R}^I)$ *with*

$$\lim_{\|\theta\|\to+\infty}\left(\frac{\|\nabla\mathcal{V}(\theta)\|^2}{2} - \Delta\mathcal{V}(\theta)\right) = +\infty \tag{6.10}$$

and let ρ_t *denote the solution of the Fokker–Planck equation* (6.5) *with* $\rho_0 \in L^2(\mathbb{R}^I;\rho_\beta^{-1})$*. Then there exists* $\lambda > 0$ *such that* ρ_t *converges exponentially fast to* ρ_β *defined in* (6.9)*, i.e.*

$$\|\rho_t - \rho_\beta\|^2_{L_2(\mathbb{R}^I;\rho_\beta^{-1})} \leq \exp(-\lambda\beta^{-1}t)\|\rho_0 - \rho_\beta\|^2_{L^2(\mathbb{R}^I;\rho_\beta^{-1})}.$$

We note that the property (6.10) is a sufficient condition to ensure that ρ_β satisfies the Poincaré inequality with constant $\lambda > 0$, which is

$$\lambda\|f\|^2_{L^2(\mathbb{R}^I;\rho_\beta)} \leq \|\nabla f\|^2_{L^2(\mathbb{R}^I;\rho_\beta)} \tag{6.11}$$

for all $f \in C^1(\mathbb{R}^I) \cap L^2(\mathbb{R}^I;\rho_\beta)$ with $\int f\rho_\beta\,\mathrm{d}\theta = 0$, which is a necessary assumption in the original stated Theorem 4.9 in[177], see Theorem 4.8 in[177]. It turns out that the socalled Bakry–Emery criterion states that a convexity condition on $\mathcal{V}$

$$D^2\mathcal{V} \geq \lambda I$$

ensures that ρ_β satisfies the Poincaré inequality (6.11).
We will choose $\mathcal{V}$ in order to use (6.3) to construct an MC estimate for some quantity of interest

$$Q_{int} = \int_{\mathbb{R}^I} F(\theta)\rho^*(\theta)\,\mathrm{d}\theta,$$

which will be an MCMC method. The idea is to find a drift b and diffusion Σ of the SDE

$$\mathrm{d}\Theta_t = b(\Theta_t)\,\mathrm{d}t + \Sigma(\Theta_t)\,\mathrm{d}W_t,$$

such that the resulting invariant distribution is given by ρ^*, i.e. ρ^* is a stationary point of the Fokker–Planck equation, this is

$$\nabla\cdot\left(-b\rho^* + \frac{1}{2}\nabla\cdot(\Sigma\rho^*)\right) = 0. \tag{6.12}$$

We can reduce this problem of finding b and Σ to the sufficient conditon

$$-b\rho^* + \frac{1}{2}\nabla \cdot (\Sigma\rho^*) = 0. \tag{6.13}$$

We can set the diffusion to $\Sigma = 2I$ and the drift term to

$$b = (\rho^*)^{-1}\nabla\rho^* = \nabla \log \rho^*,$$

such that (6.13) hold. We note that b and Σ are not uniquely determined in order to ensure the stationary condition (6.12). However, our choice for b and Σ leads to the Langevin dynamics

$$\mathrm{d}\Theta_t = \nabla \log \rho^*(\Theta_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t,$$

whose stationary distribution can be used to construct approximation of $Q_{\mathrm{i}nt}$ thanks to the ergodic theorem for Markov chains, which states that

$$\lim_{T\to\infty} \int_0^T f(\Theta_t)\,\mathrm{d}t = \int_{\mathbb{R}^I} f(\theta)\rho^*(\theta)\,\mathrm{d}\theta.$$

Provided that the potential $\mathcal{V}(\theta) = -\log \rho^*$ satisfies the Poincaré inequality (6.11), Theorem 6.1.3 ensures exponential convergence to the target density ρ^*. If we set the target distribution to be the posterior density function ρ^* defined in (2.23) with Gaussian prior $\mathcal{N}(m_0, C_0)$ this means we have to ensure that

$$D^2\Phi_R \geq \lambda I,$$

for some $\lambda > 0$, where we have defined $\Phi_R(\theta, y) = \Phi(\theta, y) + R(\theta)$ with $R(\theta) = \frac{1}{2}\|m_0 - \theta\|^2_{C_0}$. In order to view this method as particle based sampling method in the form of (6.2), we will modify the dynamics in the following way

$$\mathrm{d}\theta^{(j)} = -\mathcal{C}\nabla_\theta \Phi_R(\theta^{(j)}, y)\,\mathrm{d}t + \sqrt{2\mathcal{C}}\,\mathrm{d}W_t^{(j)}, \quad j = 1, \ldots, J, \tag{6.14}$$

where $W^{(j)}$ are independent Brownian motions on $\mathbb{R}^I$ and $\mathcal{C} \in \mathcal{L}(\mathbb{R}^I, \mathbb{R}^I)$ is a constant symmetric positive-definite matrix. We note that combined with an acceptance/rejection step this method is also known as metropolis adjusted langevin algorithm [190].
Equation (6.14) leads to the particle system of the form (6.2) with $\mathcal{A} = \mathrm{kron}(\mathrm{Id}_J, \mathcal{C})$ and $\Sigma = \sqrt{2}\mathrm{kron}(\mathrm{Id}_J, \sqrt{\mathcal{C}})$, where kron denotes the Kronecker product of two matrices. Further the potential $\mathcal{V}$ is given by

$$\mathcal{V}(z) = \frac{1}{J}\sum_{j=1}^{J} \Phi_R(\theta^{(j)}, y).$$

The corresponding probability density function $\rho_t^{(j)}$ of each particle $\theta^{(j)}$ of (6.14), $j = 1, \ldots, J$ satisfies the Fokker–Planck equation

$$\partial_t \rho_t = \nabla_\theta \cdot \left(\rho_t \mathcal{C}\nabla_\theta \frac{\delta \mathrm{KL}(\rho_t \mid \rho^*)}{\delta \rho_t}\right) \tag{6.15}$$

with $\rho_t = \rho_t^i$ and the Kullback–Leibler divergence defined by

$$\mathrm{KL}(\rho \mid \rho^*) = \left\langle \rho, \log\left(\frac{\rho}{\rho^*}\right)\right\rangle.$$

Note that its variational derivative is given by

$$\frac{\delta \mathrm{KL}(\rho \mid \rho^*)}{\delta \rho} = \log\left(\frac{\rho}{\rho^*}\right).$$

We will provide more details on the computation in the proof of Lemma 6.4.1. This is the starting point of the recently introduced interacting Langevin dynamics, where the basic idea is to choose a preconditioner $\mathcal{C}$ depending on time as sample covariance of the current particle system, which leads to an interaction between the different particles in the dynamical system.
We note that the Fokker–Planck equation (6.15) can be formally viewed as the Liouville equation corresponding to the mean-field ordinary differential equation (ODE)

$$\frac{\mathrm{d}}{\mathrm{d}t}\Theta_t = \mathcal{F}(\Theta_t, \rho_t) = -\mathcal{C}\nabla_\theta \log\left(\frac{\rho_t}{\rho^*}\right)(\Theta_t). \tag{6.16}$$

This reformulation provides the starting point for the deterministic interacting particle formulations proposed in [185, 175] for BIPs and for the blob method for diffusion in [38]. We will discuss these formulations in Section 6.4 in more details.

6.2 Interacting Langevin dynamics: Ensemble Kalman sampler

In [83], based on the dynamics (6.14) the authors propose the interacting Langevin dynamics

$$\mathrm{d}\theta_t^{(j)} = -C(\theta_t)\nabla_\theta \Phi_R(\theta_t^{(j)})\,\mathrm{d}t + \sqrt{2C(\theta_t)}\,\mathrm{d}W_t^{(j)}, \quad j = 1, \dots, J, \tag{6.17}$$

where $W^{(j)}$ denote independent R^I-dimensional Brownian motions. Taking $J \to \infty$ leads formally to the mean-field equation

$$\mathrm{d}\Theta_t = -\mathcal{C}(\rho_t)\nabla_\theta \Phi_R(\Theta_t) + \sqrt{2\mathcal{C}}(\rho_t)\,\mathrm{d}W_t,$$

with $\mathcal{C}(\rho_t)$ defined as

$$\mathcal{C}(\rho_t) = \mathbb{E}\left[(\Theta_t - \bar{\Theta}_t)(\Theta_t - \bar{\Theta}_t)^\top\right], \qquad \bar{\Theta}_t = \mathbb{E}\left[\Theta_t\right],$$

where the corresponding marginal densities $\rho_t,\ t \geq 0$, evolve according to the nonlinear Fokker–Planck equation

$$\partial_t \rho_t = \nabla_\theta \cdot \left(\rho_t \mathcal{C}(\rho_t)\nabla_\theta \frac{\delta \mathrm{KL}(\rho_t \mid \rho^*)}{\delta \rho_t}\right). \tag{6.18}$$

In [83] the authors have studied the mathematical properties of (6.18). In particular, it is possible to extend Theorem 6.1.3 to the Fokker–Planck equation in form of (6.18), see Proposition 2 in [83].
Furthermore, in [170], the authors propose a corrected finite-size particle system in order to obtain the correct long-time behaviour even under finite ensemble sizes. More specifically, the corrected particle system evolves according to the system of SDEs

$$\mathrm{d}\theta_t^{(j)} = -C(\theta_t)\nabla_\theta \Phi_R(\theta_t^{(j)})\,\mathrm{d}t + \frac{I+1}{J}(\theta_t^{(j)} - \bar{\theta}_t)\,\mathrm{d}t + \sqrt{2C(\theta_t)}\,\mathrm{d}W_t^j, \quad j = 1, \dots, J, \tag{6.19}$$

where the gradient descent direction in the drift term has been corrected through the expression

$$\nabla_{\theta^{(j)}} \cdot C(\theta) = \frac{I+1}{J}(\theta_t^{(j)} - \bar{\theta}_t) \in \mathbb{R}^I. \tag{6.20}$$

This correction term helps the particle system to stay spread in time in order to cover the support of the underlying target distribution. This effect will also be seen in the numerical example in Section 6.6.4. We note that (6.19) fits within the general framework of (6.2) with preconditioning $\mathcal{A}(Z_t) = \mathrm{kron}(\mathrm{Id}_J, C(\theta_t))$, diffusion $\Sigma(Z_t) = \sqrt{2}\,\mathrm{kron}(\mathrm{Id}_J, \sqrt{C(\theta_t)})$ and potential

$$\mathcal{V}(Z_t) = \sum_{j=1}^{J} \Phi_R(\theta_t^{(j)}, y) - \frac{I+1}{2} \log|C(\theta_t)|.$$

Here, we have assumed that $C(\theta_t)$ has full rank, i.e. $J > I$, and used that

$$\frac{\partial}{\partial (C(\theta_t))_{ij}} \log|C(\theta_t)| = \left((C(\theta_t))^{-1}\right)_{ij},$$

and, hence,

$$\frac{1}{2}\nabla_{\theta^{(j)}} \log|C(\theta_t)| = \frac{1}{J}(C(\theta_t))^{-1}(\theta_t^{(j)} - \bar{\theta}_t).$$

Note that for generalizations on $J \leq I$ one has to consider the dynamic of (6.17) in the underyling subspace spanned by the initial ensemble. In particular, also in the case $J > I$ it is not clear whether $C(\theta)$ has full rank, which is a crucial assumptions in the theoretical results. From computational aspects, a possibility to avoid this issue might be incorporation of variance inflation as it has been discussed in Section 3.2.2.
As demonstrated in [170], the corrected dynamic (6.19) leads to the Fokker–Planck equation

$$\partial_t \varphi_t = \nabla_z \cdot \left(\varphi_t \mathcal{A} \nabla_z \frac{\delta \mathrm{KL}(\varphi_t \mid \varphi^*)}{\delta \varphi_t} \right),$$

for the marginal PDF $\varphi_t(z)$ in the state variable (6.1) and the asymptotic behaviour

$$\lim_{t\to\infty} \varphi_t = \varphi^*$$

follows under appropriate conditions on the potential Φ_R. Here we consider the product density $\varphi^*(z) := \prod_{j=1}^{J} \rho^*(\theta^{(j)})$. Hence, formulation (6.19) leads to an generalisation of (6.3) under the state-dependent diffusion matrix $\Sigma(Z_t)$ and finite ensemble sizes J. While the original method proposed in [83] needed to be considered in the large data limit to produce suitable samples for the posterior distribution, in the corrected method the particle system can be viewed as a sample of the posterior distribution itself.
A detailed theoretical analysis of this method, including its affine invariance, of the formulation (6.19) as well as efficient numerical implementations in order to avoid the computation of $\sqrt{C(\theta_t)}$ can be found in [113].

6.3 Gaussian approximation

In this section, we present a particle based sampling method coming from a Gaussian approximation. Suppose we have given an initial ensemble of J particles $\theta_0^j \in \mathbb{R}^I$ drawn from the prior distribution ρ_0 given by a Gaussian measure $\mathcal{N}(m_0, C_0)$ with mean $m_0 \in \mathbb{R}^I$ and covariance matrix $C_0 \in \mathbb{R}^{I\times I}$. Further, we need the assumption that the ensemble size

is larger than the space, i.e. $J \geq I + 1$, as we have to compute the inverse of the sample covariance $C(\theta)$ in the derivation. However, if we precondition the resulting gradient flow by the sample covariance itself, it is possible to drop this assumption.
We employ a Gaussian approximation for the time-evolved distributions ρ_t based on the particle induced mean $\bar{\theta}_t$ and sample covariance matrix $C(\theta_t)$, i.e. we approximate $\rho_t \approx \mathcal{N}(\bar{\theta}_t, C(\theta_t))$. We still need to define evolution equations for the particle locations θ_t^j. Therefore, we will make use of the Kullback–Leibler divergence between ρ_t and posterior distribution ρ^* which is given by

$$\begin{aligned}\mathrm{KL}(\rho_t \mid \rho^*) &= \int_{\mathbb{R}^I} \rho_t(\theta) \log \rho_t(\theta) \,\mathrm{d}\theta - \int_{\mathbb{R}^I} \rho_t(\theta) \log \rho^*(\theta) \,\mathrm{d}\theta \\ &= -\frac{1}{2} \log(2\pi \mathrm{e} \,|C(\theta_t)|) - \int_{\mathbb{R}^I} \rho_t(\theta) \log \rho^*(\theta) \,\mathrm{d}\theta,\end{aligned}$$

where we have used, that the entropy of a Gaussian with covariance Σ is given by $\frac{1}{2} \log(2\pi \mathrm{e} \,|\Sigma|)$.
We can approximate the Kullback–Leibler divergence empirically as follows

$$\mathcal{V}(\{\theta^j\}) = \frac{1}{J} \sum_{j=1}^{J} - \log \rho^*(\theta^{(j)}) - \frac{1}{2} \log |C(\theta_t)|.$$

where we suppress constants. Here we have used the empirical approximation of the Gaussian measure $\tilde{\rho}_t$ by the particles $\{\theta^{(j)}\}_{j=1}^J$ defined by

$$\hat{\rho}_t(\theta) = \frac{1}{J} \sum_{j=1}^{J} \delta_{\theta^{(j)}}(x).$$

The gradient of the potential $\mathcal{V}$ is given by

$$\nabla_{\theta^{(j)}} \mathcal{V}(\{\theta^{(l)}\}) = \frac{1}{J} \left(-\nabla_{\theta^{(j)}} \log \rho^*(\theta^{(j)}) - (C(\theta))^{-1} (\theta^{(j)} - \bar{\theta}_t) \right)$$

and the deterministic particle dynamics by

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} \theta_t^{(j)} &= -J \nabla_{\theta^{(j)}} \mathcal{V}(\{\theta_t^{(l)}\}) \\ &= \nabla_{\theta^{(j)}} \log \rho^*(\theta_t^{(j)}) + (C(\theta_t))^{-1} (\theta_t^{(j)} - \bar{\theta}_t).\end{aligned}$$

Preconditioning by the sample covariance gives the preconditioned particle dynamics

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} \theta_t^{(j)} &= -JC(\theta_t) \nabla_{\theta^{(j)}} \mathcal{V}(\{\theta_t^{(l)}\}) \\ &= C(\theta) \nabla_{\theta^{(j)}} \log \rho^*(\theta_t^{(j)}) + \theta_t^{(j)} - \bar{\theta}_t,\end{aligned}$$

where we avoid the computation of the inverse of the sample covariance. Note, that we have used that

$$\nabla_P \log |P| = P^{-1}$$

for symmetric matrices P and, hence,

$$\frac{1}{2} \nabla_{\theta^{(j)}} \log |C(\theta_t)| = \frac{1}{J} (C(\theta_t))^{-1} (\theta_t^{(j)} - \bar{\theta}_t).$$

We plug in the definition of ρ^* under a Gaussian prior assumption and consider the the deterministic interacting particle system

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta_t^{(j)} = -\nabla_\theta \Phi_R(\theta_t^{(j)}, y) + (C(\theta))^{-1}\,(\theta_t^{(j)} - \bar{\theta}_t). \tag{6.21}$$

Assume that the posterior distribution is Gaussian $\mathcal{N}(m^*, C^*)$, i.e. we assume that the forward model $H(\cdot) = L\cdot$ for some $L \in \mathcal{L}(\mathbb{R}^I, \mathbb{R}^I)$. If one choses the initial particle positions X_0^i such that the associated empirical covariance matrix $C(\theta_0)$ is non-singular, then the particle system (6.21) satisfies

$$\lim_{t\to\infty} \bar{\theta}_t = m^*, \qquad \lim_{t\to\infty} C(\theta_t) = C^*, \tag{6.22}$$

where

$$\Phi_R(\theta, y) = \frac{1}{2}\|\theta - m^*\|_{C^*}.$$

Indeed, it is easily verified that (6.21) implies

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{\theta}_t = -(C^*)^{-1}(\bar{\theta}_t - m^*)$$

as well as

$$\frac{\mathrm{d}}{\mathrm{d}t}C(\theta_t) = -(C^*)^{-1}C(\theta_t) - C(\theta_t)(C^*)^{-1} + 2\,\underset{I}{\mathrm{Id}}\,.$$

Also note that (6.21) fits into the framework (6.2) with $\mathcal{A} = \mathrm{Id}_{J\cdot I}$, $\Sigma = 0_{J\cdot I}$, and potential

$$\mathcal{V}(Z_t) = \sum_{i=1}^{J} \Phi_R(\theta_t^{(i)}, y) - \frac{J}{2}\log|C(\theta_t)|. \tag{6.23}$$

Alternatively, one could set $\mathcal{A} = J\,\mathrm{Id}_{J\cdot I}$ and scale the potential (6.23) by J^{-1}. This formulation has the advantage that the resulting potential can be interpreted as an approximation to an expectation value. However, we will stick to unnormalised potentials of the form (6.23).

6.4 Fokker–Planck based particle systems

In this section, we introduce a particle based approximation of the Fokker–Planck equation (6.15). To do so, we will work in a reproducing kernel Hilbert space (RKHS) and follow the approach presented in [175]. We refer the reader to [176] for a detailed introduction to RKHS.

The basic idea is to consider the Kullback–Leibler divergence between to pdf ρ and the target distribution ρ^* and to regularize this divergence in a RKHS. In combination with the resulting kernelized Fokker–Planck equation it is possible to construct a particle approximation which follows a gradient flow structure (6.2).

For simplicity, we start with the case $\mathcal{C} = \mathrm{Id}_I$. Given a RKHS $\mathcal{H}$ with symmetric kernel function $k(\theta, \theta')$ and inner product $\langle g, f\rangle_{\mathcal{H}}$ we consider the RKHS Kullback–Leibler divergence

$$\mathrm{KL}_{\mathcal{H}}(\rho \mid \rho^*) := \left\langle \widetilde{\rho}, \log\left(\frac{\widetilde{\rho}}{\rho^*}\right)\right\rangle_{\mathcal{H}}, \tag{6.24}$$

with RKHS PDF

$$\widetilde{\rho}(\theta) = \int_{\mathbb{R}^I} k(\theta, \theta')\rho(\theta')\,\mathrm{d}\theta' = \langle k(\theta, \cdot), \rho(\cdot)\rangle.$$

Here, we have assumed that the kernel satisfies for each $\theta' \in \mathbb{R}^I$

$$\int_{\mathbb{R}^I} k(\theta, \theta')\mathrm{d}\theta = 1. \tag{6.25}$$

In the variational derivative of (6.24) any normalisation constant vanishes and it turns out that one can work with unnormalised kernel functions. Hence, w.l.o.g. we drop condition (6.25) in the following.
The associated RKHS Fokker–Planck equation in ρ_t is now defined by

$$\partial_t \rho_t = -\nabla_\theta \cdot (\rho_t \mathcal{F}) \tag{6.26}$$

with the vector field $\mathcal{F}$ given by

$$\mathcal{F}(\theta, \rho_t) = -\nabla_\theta \frac{\delta \mathrm{KL}_{\mathcal{H}}(\rho_t \mid \rho^*)}{\delta \rho_t}(\theta). \tag{6.27}$$

Lemma 6.4.1. *Assuming that* $\log \rho^* \in \mathcal{H}$*, the variational derivative of the RKHS Kullback–Leibler divergence is given by*

$$\frac{\delta \mathrm{KL}_{\mathcal{H}}(\rho_t \mid \rho^*)}{\delta \rho_t} = \log \widetilde{\rho}_t - \log \rho^* + \int_{\mathbb{R}^I} k(\cdot, \theta')\frac{\rho_t(\theta')}{\widetilde{\rho}_t(\theta')}\mathrm{d}\theta'. \tag{6.28}$$

Proof. We first note that the reproducing kernel property $f(\theta') = \langle f(\cdot), k(\cdot, \theta')\rangle_{\mathcal{H}}$ implies that

$$\mathrm{KL}_{\mathcal{H}}(\rho \mid \rho^*) = \left\langle \rho, \log\left(\frac{\widetilde{\rho}}{\rho^*}\right)\right\rangle. \tag{6.29}$$

This can be seen as follows

$$\begin{aligned}
\mathrm{KL}_{\mathcal{H}}(\rho \mid \rho^*) = \left\langle \widetilde{\rho}, \log\left(\frac{\widetilde{\rho}}{\rho^*}\right)\right\rangle_{\mathcal{H}} &= \left\langle \int_{\mathbb{R}^I} k(\theta', \cdot)\rho(\theta')\,\mathrm{d}\theta', \log\left(\frac{\widetilde{\rho}}{\rho^*}\right)\right\rangle_{\mathcal{H}} \\
&= \int_{\mathbb{R}^I} \rho(\theta')\left\langle k(\theta', \cdot), \log\left(\frac{\widetilde{\rho}}{\rho^*}\right)\right\rangle_{\mathcal{H}} \mathrm{d}\theta' \\
&= \int_{\mathbb{R}^I} \rho(\theta') \log\left(\frac{\widetilde{\rho}}{\rho^*}\right)(\theta')\,\mathrm{d}\theta' \\
&= \left\langle \rho, \log\left(\frac{\widetilde{\rho}}{\rho^*}\right)\right\rangle.
\end{aligned}$$

The assertion follows from the definition of the variational derivative

$$\left\langle \frac{\delta \mathrm{KL}_{\mathcal{H}}(\rho \mid \rho^*)}{\delta \rho}, h\right\rangle = \lim_{\epsilon \to 0} \frac{\mathrm{KL}_{\mathcal{H}}(\rho + \epsilon h \mid \rho^*) - \mathrm{KL}_{\mathcal{H}}(\rho \mid \rho^*)}{\epsilon}$$

for $h \in \mathcal{H}$. We have that

$$\begin{aligned}
\mathrm{KL}_{\mathcal{H}}(\rho + \epsilon h \mid \rho^*) &= \left\langle \widetilde{\rho + \varepsilon h}, \log\left(\frac{\widetilde{\rho + \varepsilon h}}{\rho^*}\right)\right\rangle_{\mathcal{H}} \\
&= \langle \widetilde{\rho}, \log(\widetilde{\rho} + \widetilde{\varepsilon h})\rangle_{\mathcal{H}} - \langle \widetilde{\rho}, \log \rho^*\rangle_{\mathcal{H}} + \langle \widetilde{\varepsilon h}, \log(\widetilde{\rho} + \widetilde{\varepsilon h})\rangle_{\mathcal{H}} - \langle \widetilde{\varepsilon h}, \log \rho^*\rangle_{\mathcal{H}},
\end{aligned}$$

and in particular, we can write

$$\frac{\mathrm{KL}_{\mathcal{H}}(\rho+\epsilon h \mid \rho^*)-\mathrm{KL}_{\mathcal{H}}(\rho \mid \rho^*)}{\epsilon} = \frac{1}{\varepsilon}\Big\langle \rho, \log \frac{\int_{\mathbb{R}^I} k(\cdot,\theta')(\rho(\theta')+\epsilon h(\theta'))\mathrm{d}\theta'}{\int_{\mathbb{R}^I} k(\cdot,\theta')\rho(\theta')\mathrm{d}\theta'}\Big\rangle_{\mathcal{H}} + \langle \widetilde{\varepsilon h}, \log \widetilde{\rho}\rangle - \langle \widetilde{\varepsilon h}, \log \rho^*\rangle.$$

Hence, using a similar argument as in equation (6.29), the assertion follows from the computation

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon}\Big\langle \rho, \log \frac{\int_{\mathbb{R}^I} k(\cdot,\theta')(\rho(\theta')+\epsilon h(\theta'))\mathrm{d}\theta'}{\int_{\mathbb{R}^I} k(\cdot,\theta')\rho(\theta')\mathrm{d}\theta'}\Big\rangle = \Big\langle \frac{\rho}{\widetilde{\rho}}, \int_{\mathbb{R}^I} k(\cdot,\theta')h(\theta')\mathrm{d}\theta'\Big\rangle = \Big\langle \int_{\mathbb{R}^I} k(\cdot,\theta')\frac{\rho(\theta')}{\widetilde{\rho}(\theta')}\mathrm{d}\theta', h\Big\rangle.$$

□

Remark 6.4.2. *We note that for a choice $k(\theta,\theta') = \delta(\theta-\theta')$ the proposed approach leads formally back to the standard definition of the Kullback–Leibler divergence, where the second term in* (6.28) *vanishes.*
Furthermore, we note that the blob method proposed in [38], which relies on a regularised Kullback–Leibler divergence

$$\mathrm{KL}_\epsilon(\rho \mid \rho^*) = \Big\langle \rho, \log\Big(\frac{\rho_\epsilon}{\rho^*}\Big)\Big\rangle$$

with mollified PDF

$$\rho_\epsilon(\theta) = \int_{\mathbb{R}^I} \phi_\epsilon(\theta-\theta')\rho(\theta)\mathrm{d}\theta,$$

and the regularisation parameter $\epsilon > 0$ becomes identical to (6.26) *for mollification und kernel functions satisfying $\phi_\epsilon(\theta-\theta') = k(\theta,\theta')$. This follows from the equivalence of* (6.24) *and* (6.29). *We note that the Gaussian kernel* (6.37) *satisfies this property while the data-driven kernel* (6.38) *does not.*

Using (6.26) and (6.27) we can ensure a decreasing Kullback–Leibler divergence in time

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}_{\mathcal{H}}(\rho_t \mid \rho^*) &= \Big\langle \frac{\delta \mathrm{KL}_{\mathcal{H}}(\rho_t \mid \rho^*)}{\delta \rho_t}, \partial_t \rho_t\Big\rangle \\ &= -\int_{\mathbb{R}^I} \Big\| \nabla_\theta \frac{\delta \mathrm{KL}_{\mathcal{H}}(\rho_t \mid \rho^*)}{\delta \rho_t}\Big\|^2 \rho_t \mathrm{d}\theta = -\int_{\mathbb{R}^I} \big\|\mathcal{F}\big\|^2 \rho_t \mathrm{d}\theta \\ &\le 0\end{aligned}$$

and, according to (6.28), we can quantify critical points ρ_c of $\mathrm{KL}_{\mathcal{H}}$ by

$$0 = \log \widetilde{\rho}_\mathrm{c} - \log \rho^* + \int_{\mathbb{R}^I} k(\cdot,\theta')\frac{\rho_\mathrm{c}(\theta')}{\widetilde{\rho}_\mathrm{c}(\theta')}\mathrm{d}\theta' + c,$$

where c is a normalisation constant. This means it holds true that

$$\rho^*(x) \propto \widetilde{\rho}_\mathrm{c}(x) e^{l(x)} \tag{6.30}$$

with log-likelihood function

$$l(\theta) := \int_{\mathbb{R}^I} k(\cdot, \theta') \frac{\rho_{\mathrm{c}}(\theta')}{\widetilde{\rho}_{\mathrm{c}}(\theta')} \mathrm{d}\theta',$$

which means that samples $\theta_{\mathrm{c}}^{(j)}$, $j = 1, \ldots, J$, from $\widetilde{\rho}_{\mathrm{c}}$ can be used to approximate expectation values with respect to the target measure ρ^* by assigning them importance weights

$$W_{\mathrm{c}}^{(j)} \propto e^{l(\theta_{\mathrm{c}}^{(j)})}.$$

The idea is that these weights are more uniform than the importance weights arising from the prior particles $\theta_0^{(j)}$, $j = 1, \ldots, J$, and the log-likelihood function Φ_R.
We construct a discrete Fokker–Planck particle dynamics by replacing ρ_t with the empirical measure

$$\rho_t(x) = \frac{1}{J} \sum_{j=1}^{J} \delta(\theta - \theta_t^{(j)}),$$

which leads from (6.26) to

$$\frac{\mathrm{d}}{\mathrm{d}t} \theta_t^{(j)} = F_t(\theta_t^{(j)}) \tag{6.31}$$

with drift term $F_t(\theta)$ given by

$$F_t(\theta) = -\nabla_\theta \left\{ \log \left(\frac{1}{J} \sum_{i=1}^{J} k(\theta, \theta_t^{(i)}) \right) - \log \rho^*(\theta) + \sum_{i=1}^{J} \frac{k(\theta, \theta_t^{(i)})}{\sum_{l=1}^{J} k(\theta_t^{(l)}, \theta_t^{(i)})} \right\}. \tag{6.32}$$

The resulting particle dynamics is equivalent to the one derived in [185, 175] starting from a discrete approximation of the regularised Kullback–Leibler divergence. These equations are of gradient flow structure (6.2) with potential

$$\mathcal{V}(z) = \sum_{j=1}^{J} \left\{ \log \left(\frac{1}{J} \sum_{i=1}^{J} k(\theta^{(j)}, \theta^{(i)}) \right) - \log \rho^*(\theta^{(j)}) \right\}, \tag{6.33}$$

$\mathcal{A} = \mathrm{Id}_{J.I}$, and $\Gamma \equiv 0_{J.I}$, that is, $F_t(\theta_t^{(j)}) = -\nabla_{\theta^{(j)}} \mathcal{V}(Z_t)$. Also we recall that the log-likelihood function is given by

$$-\log \rho^*(x) = \Phi_R(x) + C.$$

with appropriate normalisation constant $C > 0$ which is irrelevant for the particle dynamics (6.31).

Remark 6.4.3. *We note that we have assumed* $\log \rho^* \in \mathcal{H}$ *for the computation of the variational derivative in Lemma 6.4.1. In general this assumption might be strong, such that one has to generalize the presented computation, where the posterior should also be described as element of the RKHS through* $\log \widetilde{\rho}$*. This would lead to the variational derivative*

$$\frac{\delta \mathrm{KL}_{\mathcal{H}}(\rho_t \mid \rho^*)}{\delta \rho_t} = \log \widetilde{\rho}_t - \log \widetilde{\rho}^* + \int_{\mathbb{R}^I} k(\cdot, \theta') \frac{\rho_t(\theta')}{\widetilde{\rho}_t(\theta')} \mathrm{d}\theta'$$

and the generalization of the drift term (6.32) *in form of*

$$F_t(\theta) = -\nabla_\theta \Bigg\{ \log\left(\frac{1}{J}\sum_{i=1}^{J} k(\theta, \theta_t^{(i)})\right) - \frac{1}{M}\sum_{i=1}^{J} k(\theta, \theta^{(i)}) \log \rho^*(\theta^{(i)}) + \sum_{i=1}^{J} \frac{k(\theta, \theta_t^{(i)})}{\sum_{l=1}^{J} k(\theta_t^{(l)}, \theta_t^{(i)})} \Bigg\}, \tag{6.34}$$

which is of a similar form as the gradient flow resulting from the Stein variational gradient descent method [153]. While the particles are pushed into the direction of the data with the same drift in form of $\frac{1}{M}\sum_{i=1}^{J} k(\theta, \theta^{(i)}) \log \rho^*(\theta^{(i)})$*, the spreading of the particles in* (6.34) *are in a different form as the ones from Stein variational gradient descent. We also note that the resulting gradient flow with* (6.34) *is not in particular of the form* (6.2)*, as the drift for each particle into the direction of the posterior is averaged over the whole particle system. We refer to [74] for a detailed analysis of the Stein variational gradient descent method as a gradient flow in a space of probability measures.*

Remark 6.4.4. *For the number of particles J approaching infinity, one can expect* (6.31) *converging to the associated RKHS Fokker–Planck dynamics* (6.26)*. In [38] the authors present a rigorous theoretical analysis of the closely related blob method for diffusion. See also [195] and [66] for earlier work on deterministic numerical methods for approximating diffusion processes. Furthermore, in [217] a diffusion map approach has been suggested to approximate $\nabla_\theta \log \rho_t$ in the Fokker–Planck equation* (6.16) *without using an explicit kernel density estimate for ρ_t. While this method is computationally attractive, it is not clear whether such an approximation leads to a particle system fitting the gradient flow structure* (6.2)*.*

Equation (6.30) suggests that a particle system in the equilibrium

$$Z_{\mathrm{c}} = \{\theta_{\mathrm{c}}^{(j)}\}_{j=1}^{J} = \operatorname*{arg\,inf}_{z \in \mathbb{R}^{J \cdot I}} \mathcal{V}(z)$$

can be used to approximate expected values with respect to ρ^* for J sufficiently large using the approximation

$$\widehat{\rho}^*(\theta) \propto \left(\frac{1}{J}\sum_{j=1}^{J} k(\theta, \theta_{\mathrm{c}}^{(j)})\right) \exp\left(\sum_{i=1}^{J} \frac{k(\theta, \theta_{\mathrm{c}}^{(i)})}{\sum_{l=1}^{J} k(\theta_{\mathrm{c}}^{(l)}, \theta_{\mathrm{c}}^{(i)})}\right). \tag{6.35}$$

More specifically, one first collects a desired number of realisations $\theta^{(j),k}$, $k = 1, \ldots, K$, from each of the PDFs $k(\cdot, \theta_{\mathrm{c}}^{(j)})$, $j = 1, \ldots, J$. These $N = J \cdot K$ realisations are then assigned with importance weights

$$W^{(j),k} \propto \exp\left(\sum_{i=1}^{J} \frac{k(\theta^{(i),k}, \theta_{\mathrm{c}}^{(i)})}{\sum_{l=1}^{J} k(\theta_{\mathrm{c}}^{(l)}, \theta_{\mathrm{c}}^{(i)})}\right).$$

The choice of the kernel functions $k(\theta, \theta')$ constitutes an important aspect for the computational implementation of (6.31). We have already mentioned the Gaussian kernel (6.37) which requires the specification of an appropriate covariance matrix B. Given J samples $\theta_0^{(j)}$, $j = 1, \dots, J$, from the prior PDF ρ_0 with corresponding empirical covariance matrix $C(\theta_0)$, we set

$$B = \alpha C(\theta_0) \tag{6.36}$$

for $\alpha > 0$ appropriately chosen. The choice of the bandwidth α is itself a difficult task arising in general in kernel density estimation. We also use the $\theta_0^{(j)}$'s as initial conditions for (6.31) in our numerical experiments. Alternatively, the data-driven kernel (6.38) can be implemented with $\theta^{(j)} = \theta_0^{(j)}$ and B given by (6.36).

Remark 6.4.5. *In our numerical examples we mainly choose the class of kernel functions which are provided by the Gaussian kernels*

$$k(\theta, \theta') = \psi(\|\theta - \theta'\|_B^2) \tag{6.37}$$

where $\psi(r) = \exp(-r^2/2)$ for some appropriate symmetric positive-definite matrix $B \in \mathbb{R}^{I \times I}$. We will see that the choice of B is a crucial task in order to tune the approximation results for the posterior distribution. In particular, the choice of the bandwidth of the underlying kernel will be the key for tuning this method.
One may also consider data-driven kernel functions such as

$$k(\theta, \theta') = \frac{\psi(\|\theta - \theta'\|_B^2)}{\sqrt{\sum_{j=1}^J \psi(\|\theta^{(j)} - \theta'\|_B^2)} \sqrt{\sum_{j=1}^J \psi(\|x - \theta^{(j)}\|_B^2)}}, \tag{6.38}$$

which arise from a diffusion map approximation to the semigroup generated by a reversible diffusion process with invariant measure ρ provided that $\theta^{(j)} \sim \rho$ and $B = 2\epsilon \, \mathrm{Id}_I$ for $\epsilon > 0$ sufficiently small [96, 218].

Remark 6.4.6. *In high dimensional setting the statistical curse of dimensionality arises, see for example [226]. If the parameters have dimension I, then we need an ensemble size J growing exponentially fast with I. The kernel density estimator approximates the target distribution in a local neighbourhood of the members in our particle system of size J and in high dimensions those particle locations will be sparse in the parameter space. This problem can be counteracted partially by adaptive kernel functions. Adaptive choices of the kernel functions such as*

$$B = \alpha C(\theta_t) \tag{6.39}$$

in the Gaussian kernel (6.37) *can, for examples, be considered in the definition of the potential* (6.33)*) and can help to capture the structure of the target distribution. However, the computation of the corresponding drift $F_t(\theta_t^{(j)}) = -\nabla_{\theta^{(j)}} \mathcal{V}(Z_t)$ becomes, more involved. Therefore, in our high dimensional numerical example we will consider the adaptive choice* (6.39) *but for simplicity suppress the dependence of B on θ in the computation of the drift* (6.32)*. We note, that this method will no longer be of gradient flow structure* (6.2)*, but still leads to an effective improvement of the numerical results.*

6.5 Derivative free modification - preconditioning and localisation

We have introduced most of the methods from the previous section with a trivial choice of preconditioning matrix. We now move to the evolution equations (6.2) with a non-trivial choice of the matrix $\mathcal{A}(z)$ and $\Sigma(z) = 0_{N_z}$. More specifically, we choose

$$\mathcal{A}(Z_t) = \mathrm{Id}_J \otimes C(\theta_t). \tag{6.40}$$

which is motivated by the gradient flow structure of the EnKF [140] and ensemble Kalman–Bucy filter [15]. Further, we refer to [175] for a more general discussion of preconditioned gradient flows in the context of BIPs. The same preconditioning has recently been considered for stochastic interacting particle systems with $\Sigma(Z_t) = \sqrt{2}\mathcal{A}(Z_t)^{1/2}$ in [83], which we have discussed in section 6.2. We also note that related ideas have previously appeared in the Markov chain Monte Carlo literature. See, for example, [150] and references therein. Application of (6.40) to the Fokker–Planck particle dynamics (6.31) introduced in Section 6.4, leads to

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta_t^{(j)} = C(\theta_t)F_t(\theta_t^{(j)}). \tag{6.41}$$

Example 6.5.1. *We consider the Gaussian approximation* (6.21) *introduced in Section 6.3 and transform this particle system into the form of* (6.41)*, i.e. we replace* (6.21) *by*

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\theta_t^{(j)} &= -C(\theta_t)\nabla_\theta\Phi_R(\theta_t^{(j)}) + \theta_t^{(j)} - \bar{\theta}_t \\ &= -C(\theta_t)(C^*)^{-1}(\theta_t^{(j)} - \bar{\theta}^*) + \theta_t^{(j)} - \bar{\theta}_t.\end{aligned}$$

From (6.22)*, we obtain*

$$C(\theta_t)(C^*)^{-1}(\theta_t^{(j)} - \bar{\theta}^*) \approx \theta_t^{(j)} - \bar{\theta}^*$$

and, hence,

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta_t^{(j)} \approx -(\bar{\theta}_t - \bar{\theta}^*)$$

for $t \gg 1$ sufficiently large. This suggests that the preconditioning (6.40) *is asymptotically optimal for linear BIPs, that is, all directions of state space $\mathbb{R}^I$ and all particles are treated equally as $t \to \infty$.*

6.5.1 Localisation

While (6.41) is asymptotically optimal for Gaussian posterior PDFs ρ^*, this is not necessarily the case for multimodal posterior PDFs. In this case, one may use a localised covariance matrix, as first considered in [150]. In particular, we first define distance-dependent weights

$$w_t^{ji} = \frac{\exp\left(-\frac{1}{2\gamma}\|\theta_t^{(j)} - \theta_t^{(i)}\|_D^2\right)}{\sum_{l=1}^J \exp\left(-\frac{1}{2\gamma}\|\theta_t^{(j)} - \theta_t^{(l)}\|_D^2\right)}, \tag{6.42}$$

where $\gamma > 0$ is an appropriate scaling parameter and $D \in \mathbb{R}^{I\times I}$ an appropriate symmetric positive-definite matrix. Each particle $\theta_t^{(j)}$ is assigned the weighted covariance matrix

$$C(\theta_t^{(j)}) = \sum_{i=1}^J w_t^{ji}\left(\theta_t^{(i)} - \bar{\theta}_t^{(j)}\right)\left(\theta_t^{(i)} - \bar{\theta}_t^{(j)}\right)^\top = \sum_{i=1}^J w_t^{ji}\theta_t^{(i)}\left(\theta_t^{(i)}\right)^\top - \bar{\theta}_t^{(j)}\left(\bar{\theta}_t^{(j)}\right)^\top \tag{6.43}$$

with localised mean

$$\bar{\theta}_t^{(j)} = \sum_{i=1}^{J} w_t^{ji} \theta_t^{(i)}. \tag{6.44}$$

Finally, we replace the evolution equations (6.41) by

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta_t^{(j)} = C(\theta_t^{(j)}) F_t(\theta_t^{(j)}) \tag{6.45}$$

for $j = 1, \dots, J$. We note that (6.45) also fits into the framework of (6.2), where $\mathcal{A}(Z_t) \in \mathcal{L}(\mathbb{R}^{J \cdot I}, \mathbb{R}^{J \cdot I})$ is block-diagonal with its j-th block entry given by $C(\theta_t^{(j)})$. Furthermore, it holds true that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{V}(Z_t) \leq 0$$

along solutions of (6.45). Finally, we mention that considering $\gamma \to \infty$ leads to $w_t^{ij} = 1/J$ in (6.42) and the covariance matrix $C(\theta_t)$ is recovered from (6.43). In the sequel we always assume that

$$D = C(\theta_0). \tag{6.46}$$

6.5.2 Invariance under affine transformations

We now demonstrate that both of the preconditioned formulations (6.41) and (6.45) are invariant under affine transformations. For the importance of affine invariant sampling methods for BIP, we refer [88]. In the context of the general framework (6.2) we define affine invariance as follows.

Definition 6.5.2. *We call an SDE* (6.2) ***invariant under an affine transformation***

$$Z_t = MV_t + c,$$

$M \in \mathcal{L}(\mathbb{R}^{J \cdot I}, \mathbb{R}^{J \cdot I})$ *invertible, if the associated equations of* V_t *are of the form*

$$\mathrm{d}V_t = -\mathcal{A}(V_t)\nabla_v \mathcal{U}(V_t) + \Gamma(V_t)\,\mathrm{d}W_t$$

with the potential $\mathcal{U}$ *defined by*

$$\mathcal{U}(V_t) = \mathcal{V}(MV_t + c). \tag{6.47}$$

We consider only affine transformations defined component-wise, i.e.

$$\theta_t^{(j)} = Av_t^{(j)} + b, \qquad V_t = \left((v_t^{(1)})^\top, \dots, (v_t^{(J)})^\top\right)^\top, \tag{6.48}$$

with $A \in \mathbb{R}^{I \times I}$ being invertible.

Lemma 6.5.3. *The preconditioned formulations* (6.41) *and* (6.45)*, respectively, are invariant under affine transformations of the form* (6.48)*.*

Proof. First, we note that (6.46) implies that the weights (6.42) are invariant under (6.48). Furthermore, we obtain

$$C(\theta_t^{(j)}) = C(Av_t^{(j)} + b) = AC(v_t^{(j)})A^\top$$

and

$$\nabla_{v^{(i)}}\mathcal{U}(V_t) = A^\top \nabla_{\theta^{(j)}}\mathcal{V}(Z_t)$$

with $\mathcal{U}$ defined by (6.47). The invariance property follows now from

$$A\,\frac{\mathrm{d}}{\mathrm{d}t}v_t^{(j)} = \frac{\mathrm{d}}{\mathrm{d}t}\theta_t^{(j)} = -C(\theta_t^{(j)})\nabla_{\theta^{(i)}}\mathcal{V}(Z_t) = -A\left(C(v_t^{(j)})\nabla_{v^{(j)}}\mathcal{U}(V_t)\right)$$

and thus,

$$\frac{\mathrm{d}}{\mathrm{d}t}V_t = \mathcal{A}(V_t)\nabla\mathcal{U}(V_t).$$

□

Remark 6.5.4. *Choosing* (6.36) *for the Gaussian kernels (6.37) leads to the transformed potential* (6.47) *given by*

$$\mathcal{U}(v) = \sum_{j=1}^{J}\left\{\log\left(\frac{1}{J}\sum_{i=1}^{J}k(v^{(j)},v^{(i)})\right) - \log\rho^*(Av^{(j)}+b)\right\}.$$

6.5.3 Derivative free formulation

As we have seen in Subsection 3.2.1, one of the attractive features of the EnKF is its gradientfree formulation. In order to extend this approach to our preconditioned gradient flow formulations, we recall that

$$\begin{aligned} C(\theta_t)\nabla_\theta \log\rho^*(\theta_t^{(j)}) &= -C(\theta_t)\nabla_\theta\Phi_R(\theta_t^{(j)}) \\ &= -C(\theta_t)DH(\theta)^\top\Gamma^{-1}(H(\theta_t^{(j)})-y) - C(\theta_t)C_0^{-1}(\theta_t^{(j)}-\bar\theta_0)). \end{aligned}$$

The key idea of the gradientfree formulations discussed in Subsection 3.2.1 is to replace $C(\theta_t)DH(\theta)^\top$ with the covariance matrix

$$C^{\theta p}(\theta_t) = \frac{1}{J}\sum_{i=1}^{J}(\theta_t^{(i)}-\bar\theta_t)(H(\theta_t^{(i)})-\bar H_t)^\top, \qquad \overline{H}_t = \frac{1}{J}\sum_{i=1}^{J}H(\theta_t^{(i)}).$$

Recall that

$$C(\theta_t)DH(\theta)^\top = C^{\theta p}(\theta_t) \tag{6.49}$$

for linear forward maps H. For nonlinear forward maps H this approximation will get more accurate if the particles are close to each other. Since the particles are representing a distribution, a localised version of the covariance matrix suggests to improve the accuracy of the derivative free formulation. Increasing the ensemble size will also improve the accuracy of the derivative free formulation, and in particular in the localised formulation. Following the corresponding continuous-time formulations of the ensemble Kalman–Bucy filter [17, 175], we obtain the following derivative free reformulation of (6.41):

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\theta_t^{(j)} = -C(\theta_t)\nabla_{\theta^{(j)}}&\left\{\log\left(\frac{1}{J}\sum_{i=1}^{J}k(\theta_t^{(j)},\theta_t^{(i)})\right) + \sum_{i=1}^{J}\frac{k(\theta_t^{(j)},\theta_t^{(i)})}{\sum_{l=1}^{J}k(\theta_t^{(l)},\theta_t^{(i)})}\right\} \\ &- \left\{C^{\theta p}(\theta_t)\Gamma^{-1}(H(\theta_t^{(j)})-y) + C(\theta_t)C_0^{-1}(\theta_t^{(j)}-m_0)\right\}. \end{aligned} \tag{6.50}$$

While the derivative free formulation is no longer of gradient flow structure (6.2), the potential (6.33) still allows us to monitor the behaviour of (6.50) in the large time limit. We note that (6.49) also holds for our localised covariance matrices $C(\theta_t^{(j)})$ with $C^{\theta p}(\theta_t^{(j)})$ defined by

$$C^{\theta p}(\theta_t^{(j)}) = \sum_{i=1}^{J} w_t^{ji}(\theta_t^{(i)} - \bar{\theta}_t^{(j)})(H(\theta_t^{(i)}) - \bar{H}_t^{(j)})^\top, \qquad \bar{H}_t^{(j)} = \sum_{i=1}^{J} w_t^{ji} H(\theta_t^{(i)}).$$

Hence, the localised formulation (6.45) gives rise to the derivative free formulation

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\theta_t^{(j)} = -C(\theta_t^{(j)})\nabla_{\theta^{(j)}} \left\{ \log\left(\frac{1}{J}\sum_{i=1}^{J} k(\theta_t^{(j)}, \theta_t^{(i)}) \right) + \sum_{i=1}^{J} \frac{k(\theta_t^{(j)}, \theta_t^{(i)})}{\sum_{l=1}^{J} k(\theta_t^{(l)}, \theta_t^{(i)})} \right\} \\ - \left\{ C^{\theta p}(\theta_t^{(j)})\Gamma^{-1}(H(\theta_t^{(j)}) - y) + C(\theta_t^{(j)})C_0^{-1}(\theta_t^{(j)} - m_0) \right\}. \end{aligned} \tag{6.51}$$

We will demonstrate in Section 6.6 that the localised formulation (6.51) is beneficial in case of multimodal posterior PDFs ρ^*.

Remark 6.5.5. *We emphasise that the localisation strategy proposed here is different from standard B-localisation employed for ensemble Kalman filters, where the empirical covariance matrix $C(\theta_t)$ is tempered by a second matrix C such that the preconditioning matrix becomes $C \circ C(\theta_t)$ where $\circ$ denotes the Schur product of two matrices. See, for example, [80, 186]. Furthermore, the proposed localised strategy can also be applied in the context of EKI discussed in Chapter 3.*

6.5.4 Localised interacting Langevin dynamics

We now derive a correction term for the particle evolution with $C(\theta_t)$ in (6.17) replaced by the localised empirical covariance matrix (6.43). Following [170], the correction term is given by $\nabla_{\theta^{(j)}} \cdot C(\theta_t^{(j)})$ and an explicit expression is provided in the following lemma.

Lemma 6.5.6. *It holds that*

$$\begin{aligned} \nabla_{\theta^{(j)}} \cdot C(\theta_t^{(j)}) = w_t^{jj}(I+1)(\theta_t^{(j)} - \bar{\theta}_t^{(j)}) + \sum_{i=1}^{J} \theta_t^{(i)} \left(\theta_t^{(i)}\right)^\top \nabla_{\theta^{(j)}} w_t^{ji} \\ - \sum_{i=1}^{J} \bar{\theta}_t^{(j)} \left(\theta_t^{(i)}\right)^\top \nabla_{\theta^{(j)}} w_t^{ji} - \sum_{i=1}^{J} \theta_t^{(i)} \left(\bar{\theta}_t^{(j)}\right)^\top \nabla_{\theta^{(j)}} w_t^{ji}, \end{aligned} \tag{6.52}$$

where $\bar{\theta}_t^{(j)} \in \mathbb{R}^I$ denotes the localised mean defined in (6.44) *and $w_t^{ji} \in [0,1]$ denote the localisation weights* (6.42)*.*

Proof. In order to prove the expression of the correction term, we will make use of the following identities for vectors $(x^{(j)})_{j=1,\dots,J}$ in $\mathbb{R}^I$:

$$\begin{aligned} \nabla_{x^{(j)}} \cdot (w_t^{jj} x^{(j)}(x^{(j)})^\top) &= w_t^{jj}(I+1)x^{(j)} + x^{(j)}(x^{(j)})^\top \nabla_{x^{(j)}} w_t^{jj} \\ \nabla_{x^{(j)}} \cdot (w_t^{ji} x^{(j)}(x^{(i)})^\top) &= w_t^{ji} x^{(i)} + x^{(j)}(x^{(i)})^\top \nabla_{x^{(j)}} w_t^{ji} \\ \nabla_{x^{(j)}} \cdot (w_t^{ji} x^{(i)}(x^{(j)})^\top) &= w_t^{ji} I x^{(i)} + x^{(i)}(x^{(j)})^\top \nabla_{x^{(j)}} w_t^{ji} \\ \nabla_{x^{(j)}} \cdot (w_t^{ji} x^{(i)}(x^{(i)})^\top) &= x^{(i)}(x^{(i)})^\top \nabla_{x^{(j)}} w_t^{ji}, \end{aligned}$$

where we assume that $j \neq i$. We apply these identities to the localised covariance matrix (6.43) yielding

$$\nabla_{\theta^{(j)}} \cdot \left(\sum_{i=1}^{J} w_t^{ji} \theta_t^{(i)} \left(\theta_t^{(i)} \right)^\top \right) = w_t^{jj} (I+1) \theta_t^{(j)} + \sum_{i=1}^{J} \theta_t^{(i)} \left(\theta_t^{(i)} \right)^\top \nabla_{x^{(j)}} w_t^{ji}$$

as well as

$$\nabla_{x^{(j)}} \cdot \left(\bar{\theta}_t^{(j)} \left(\bar{\theta}_t^{(j)} \right)^\top \right) = w_t^{jj} (I+1) \bar{\theta}_t^{(j)} + \sum_{i=1}^{J} \left\{ \bar{\theta}_t^{(j)} \left(\theta_t^{(i)} \right)^\top + \theta_t^{(i)} \left(\bar{\theta}_t^{(j)} \right)^\top \right\} \nabla_{x^{(j)}} w_t^{ji},$$

which implies (6.52). □

In order to compute the correction term (6.52) we have to compute $\nabla_{\theta^{(j)}} w_t^{ji}$ for the weights given by (6.42):

$$\begin{aligned}
\nabla_{x^{(j)}} w_t^{ji} &= \frac{w_t^{ji}}{\gamma} D^{-1} \left\{ -(\theta_t^{(j)} - \theta_t^{(i)}) + \frac{\sum_{l=1}^{J} \exp\left(-\frac{1}{2\gamma} \|\theta_t^{(j)} - \theta_t^{(l)}\|_D^2 \right) (\theta_t^{(j)} - \theta_t^{(l)})}{\sum_{l=1}^{J} \exp\left(-\frac{1}{2\gamma} \|\theta_t^{(j)} - \theta_t^{(l)}\|_D^2 \right)} \right\} \\
&= \frac{w_t^{ji}}{\gamma} D^{-1} \left\{ -(\theta_t^{(j)} - \theta_t^{(i)}) + \sum_{l=1}^{J} w_t^{jl} (\theta_t^{(j)} - \theta_t^{(l)}) \right\} \\
&= \frac{w_t^{ji}}{\gamma} D^{-1} \left\{ \theta_t^{(i)} - \bar{\theta}_t^{(j)} \right\}.
\end{aligned}$$

We are now ready to formulate corrected evolution system for the interacting Langevin diffusion model with localised covariance matrix by adding the drift $\nabla_{\theta^{(j)}} \cdot C(\theta_t^{(j)})$

$$\mathrm{d}\theta_t^{(j)} = -C(\theta_t^{(j)}) \nabla \Phi_R(\theta_t^{(j)}) \, \mathrm{d}t + \nabla_{\theta^{(j)}} \cdot C(\theta_t^{(j)}) \, \mathrm{d}t + \sqrt{2C(\theta_t^{(j)})} \, \mathrm{d}W_t^{(j)}. \tag{6.53}$$

Similar as for the particle system (6.19), one can derive the joint density of the particle system evolving through (6.53), see [170]

Lemma 6.5.7. *The joint density corresponding to the particle system evolving by* (6.53) *satisfies the Fokker–Planck equation given by*

$$\partial_t \varphi_t = \nabla_z \cdot \left(\varphi_t \mathcal{A} \nabla_z \frac{\delta \mathrm{KL}(\varphi_t \mid \varphi^*)}{\delta \varphi_t} \right)$$

with $\mathcal{A} \in \mathcal{L}(\mathcal{Z}, \mathcal{Z})$ a block diagonal matrix with its j-th block entry given by $C(\theta_t^{(j)})$ and $\varphi^*(z) := \prod_{j=1}^{J} \rho^*(\theta^{(j)})$.

Hence by the Kalman–Wasserstein gradient flow structure [83, 170] in the joint distribution φ_t, the finite-size particle system (6.53) can be used in the long-time limit to approximate i.i.d. samples from ρ^*.

We can further formulate derivative free variants of both formulations (6.19) and (6.53), respectively, by straightforward replacement of $C(\theta_t) DH(\theta_t^{(j)})^\top$ by $C^{\theta p}(\theta_t)$, see [83], or $C^{\theta p}(\theta_t^{(j)})$ respectively. We obtain the following derivative free formulation

$$\begin{aligned}
\mathrm{d}\theta_t^{(j)} = &- \left\{ C^{\theta p}(\theta_t) \Gamma^{-1} (H(\theta_t^{(j)}) - y) + C(\theta_t) C_0^{-1} (\theta_t^{(j)} - m_0) \right\} \mathrm{d}t \\
&+ \nabla_{\theta^{(j)}} \cdot C(\theta_t) \, \mathrm{d}t + \sqrt{2C(\theta_t)} \, \mathrm{d}W_t^{(j)}.
\end{aligned} \tag{6.54}$$

and in its localised formulation

$$\begin{aligned} \mathrm{d}\theta_t^{(j)} = &- \Big\{ C^{\theta p}(\theta_t^{(j)})\Gamma^{-1}(H(\theta_t^{(j)}) - y) + C(\theta_t^{(j)})C_0^{-1}(\theta_t^{(j)} - m_0) \Big\}\,\mathrm{d}t \\ &+ \nabla_{\theta^{(j)}} \cdot C(\theta_t^{(j)})\,\mathrm{d}t + \sqrt{2C(\theta_t^{(j)})}\,\mathrm{d}W_t^{(j)}. \end{aligned} \tag{6.55}$$

We note that the presented derivative free and localised formulations are only based on heuristic derivations and we did not provide any theoretical verification.

6.6 Numerical results

We will apply the proposed methods to a sequence of numerical examples with increasing challenge ranging from low to high dimensional and unimodal to multimodal examples.

6.6.1 2-dimensional unimodal example

Our first example is nearly Gaussian, which was originally presented in [79] and later also used in [83, 99].
More specifically, we consider the following one-dimensional elliptic boundary-value problem

$$-\frac{\mathrm{d}}{\mathrm{d}s}\left(\exp(\theta_1)\frac{\mathrm{d}}{\mathrm{d}s}p(s)\right) = 1, \quad s \in [0,1],$$

with boundary condition $p(0) = 0$ and $p(1) = \theta_2$. An explicit solution of this boundary-value problem in the parameters $\theta = (\theta_1, \theta_2)^{\mathbb{T}} \in \mathbb{R}^2$ is given by

$$p(s,\theta) = \theta_2 s + \exp(-\theta_1)\left(-\frac{s^2}{2} + \frac{s}{2}\right)$$

and we define the forward map in (2.2) by

$$H(\theta) = (p(s_1,\theta), p(s_2,\theta))^\top,$$

this means, we assume that noisy measurements of $p(\cdot,\theta)$ are given at locations $s_1 = 0.25$ and $s_2 = 0.75$.
Furthermore, we assume Gaussian measurement errors $\Xi \sim \mathcal{N}(0,\Gamma)$ with $\Gamma = 0.01 \cdot \mathrm{Id}_2$, $\mathrm{Id}_2 \in \mathbb{R}^{2\times 2}$ the identity matrix, and a Gaussian prior $\rho_0 = \mathcal{N}(0, C_0)$ with $C_0 = 100 \cdot \mathrm{Id}_2$. We draw a reference parameter $\theta^\dagger \sim \rho_0$ in order to construct the data and set the observation to $y = H(\theta^\dagger) + \xi^\dagger$, where $\xi^\dagger$ is a realisation of the measurement error. Our numerical results are based on the realisations $\theta^\dagger = (0.0865, -0.8157)^\top$ and $y = (-0.0173, -0.573)^\top$.
To solve the ODE representing the deterministic Fokker–Planck dynamics (6.50) we apply the `MATLAB` solver `ode45` and we have implemented an Euler–Maruyama scheme with step-size $\Delta t = 0.0001$ for the interacting Langevin sampler (6.55). The preconditioned Fokker–Planck dynamics is implemented using Gaussian kernels (6.37) with $B = \alpha C_0$ and $\alpha = 0.05$.
In Figure 6.1 we show the posterior approximation through the resulting particle systems with $J = 200$ particles, which were generated from the deterministic Fokker–Planck dynamics as well as the interacting Langevin sampler. One finds that both methods approximate the posterior distribution well.

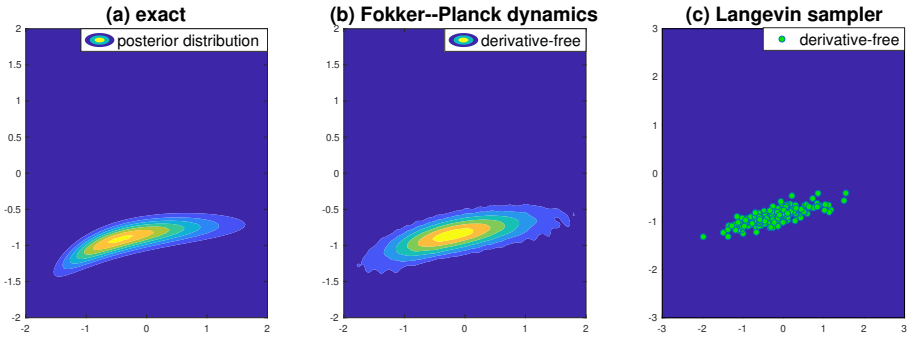

Figure 6.1: Approximations to the posterior PDF from interacting Fokker–Planck dynamics: b) kernel density estimate from deterministic Fokker–Planck dynamics, c) particle system from interacting Langevin sampler. The exact posterior PDF is displayed in panel a).

Kernel density estimates for the target distribution for different choices of α are displayed in Figure 6.2 from which it can be seen that the performance of the deterministic Fokker–Planck dynamics depends crucially on the kernel parameter α in (6.36). The effect of the different choices of α is also demonstrated in the time evolution of the potential $\mathcal{V}$ from (6.33), which can be found in Figure 6.3. For too small choices of α the resulting density underestimates the spread, whereas too large α lead to an overestimated spread. Overall, the choice of the scaling parameter α is crucial and quite sensitive, which is a general challenge in kernel density estimation.
We conclude that both the deterministic and stochastic interacting particle formulations work well for this simple 2-dimensional example. While the interacting Langevin dynamics is easier to implement, the Fokker–Planck dynamics immediately results in a kernel density estimate for the posterior distribution and its performance can be monitored through the time evolution of the potential energy (6.33).

6.6.2 2-dimensional bimodal example

To numerically test the effect of the localisation introduced in Subsection 6.5.1, we next consider a 2-dimensional bimodal example resulting from the nonlinear forward map

$$H : \mathbb{R}^2 \to \mathbb{R}, \quad H(\theta) = (\theta_1 - \theta_2)^2.$$

We assume a Gaussian prior with mean zero and covariance $C_0 = \mathrm{Id}_2 \in \mathbb{R}^{2\times 2}$ and Gaussian measurements errors $\Xi \sim \mathcal{N}(0, \Gamma)$ with $\Gamma = \mathrm{Id}_2$. We again draw a reference parameter $\theta^\dagger \sim \rho_0$ and construct observation $y = H(\theta^\dagger) + \xi^\dagger$, where $\xi^\dagger$ is a realisation of the random measurement errors Ξ. Our numerical results are based on the realisation $\theta^\dagger = (-1.5621, -0.0021)^\top$ and $y = 4.2297$.
We implement the preconditioned version of the Fokker–Planck based particle system (6.31). Furthermore, we also consider its derivative free formulation (6.50) as well as the localised formulation (6.51). We compare the results to those from the corresponding interacting Langevin sampler (6.19), its derivative free formulation (6.54), and its localised formulation (6.55), respectively.
We again use the `MATLAB` solver `ode45` to time-step the deterministic Fokker–Planck dynamics (6.50) and the Euler–Maruyama scheme with step-size $\Delta t = 0.0001$ for the inter-

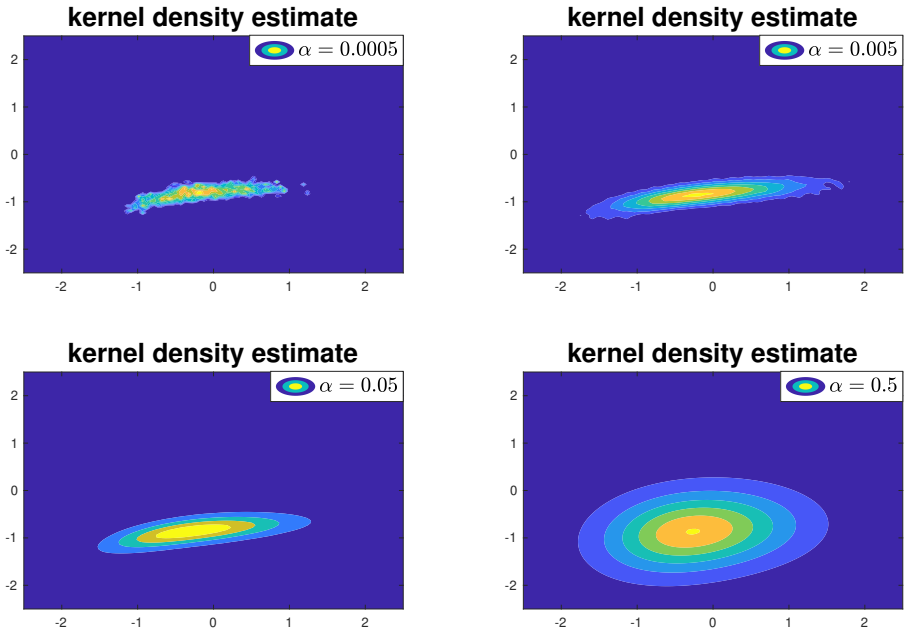

Figure 6.2: Approximations to the posterior PDF from deterministic Fokker–Planck dynamics for different values of the kernel parameter α: a) $\alpha = 0.0005$, b) $\alpha = 0.005$, c) $\alpha = 0.05$, d) $\alpha = 0.5$.

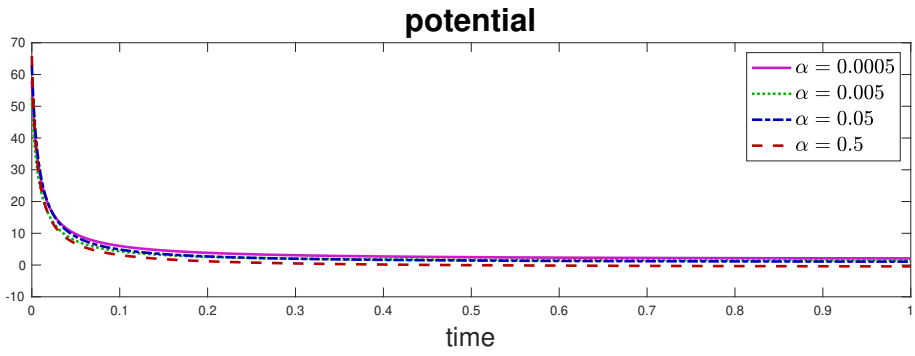

Figure 6.3: Time evolution of the potential function $\mathcal{V}$ for different choices of α.

acting Langevin sampler (6.55). We set our Gaussian kernel (6.37) with $B = \alpha P_0$ and $\alpha = 0.01$ for the preconditioned Fokker–Planck dynamics. The weights for the localised formulation (6.42) are computed with $\gamma = 0.5$ and $D = C_0 \in \mathbb{R}^{2\times 2}$. All simulations use $J = 200$ particles.

In Figure 6.4 we present the kernel density estimates resulting from the preconditioned Fokker–Planck dynamics, which demonstrates that the particle system is representing the target density. While the derivative free formulation (6.50) is getting pushed to one of the peaks, the localised formulation (6.51) leads to an effectively improved approximation of the posterior density. The time evolution of the potential (6.33) along the three different interacting particle approximations is displayed in Figure 6.5.

Further, we present the variational derivative of the RKHS Kullback–Leibler divergence

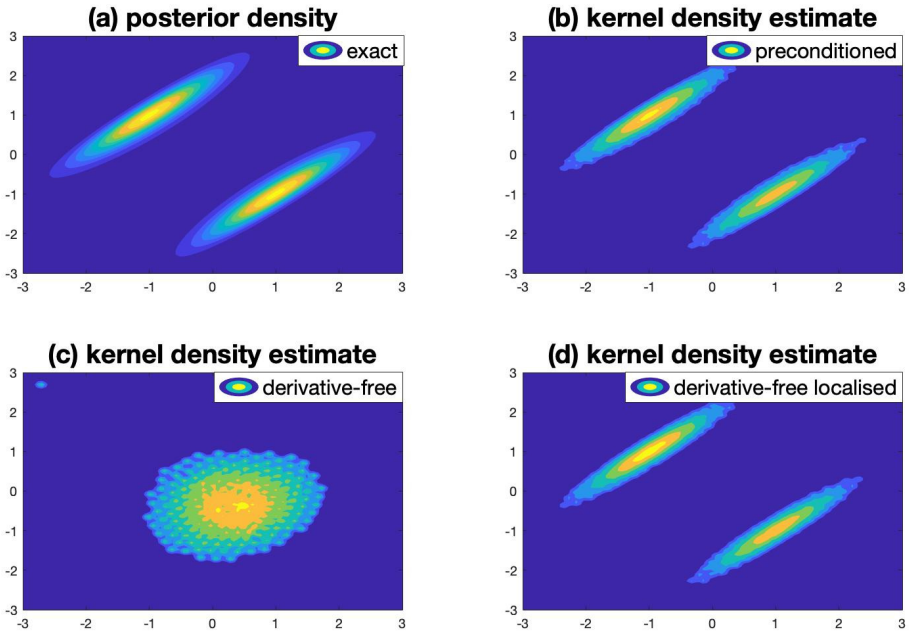

Figure 6.4: Approximations to the posterior PDF from interacting Fokker–Planck dynamics: a) exact posterior PDF, b) fitted PDF from preconditioned dynamics using exact gradients, c) fitted PDF from preconditioned gradient-free dynamics, d) fitted PDF from localised gradient-free dynamics.

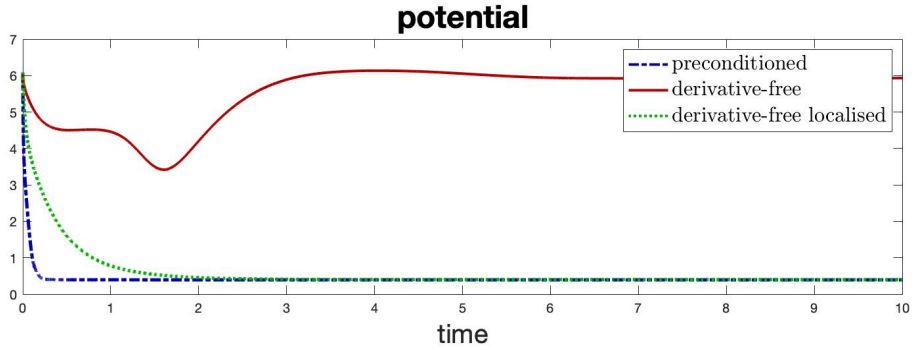

Figure 6.5: Time evolution of potential function $\mathcal{V}$ for different implementations of Fokker–Planck particle dynamics: (blue) preconditioned dynamics using exact gradients, (red) preconditioned gradient-free dynamics, (green) localised gradient-free dynamics.

(6.28) resulting from the equilibrium particle positions $\{X_c^i\}$ as well as the weight function

$$W(\theta) = \exp\left(\sum_{i=1}^{J} \frac{k(\theta, \theta_{\mathrm{c}}^{(i)})}{\sum_{l=1}^{J} k(\theta_{\mathrm{c}}^{(l)}, \theta_{\mathrm{c}}^{(i)})}\right)$$

(compare (6.35)) corresponding to the equilibrium particle positions in Figure 6.6. We find that both the exact gradient method as well as the localised derivative free formulation are leading to a variational derivative, which is nearly constant in the region of state space covered by the equilibrium particle positions $\{\theta_c^{(j)}\}$.

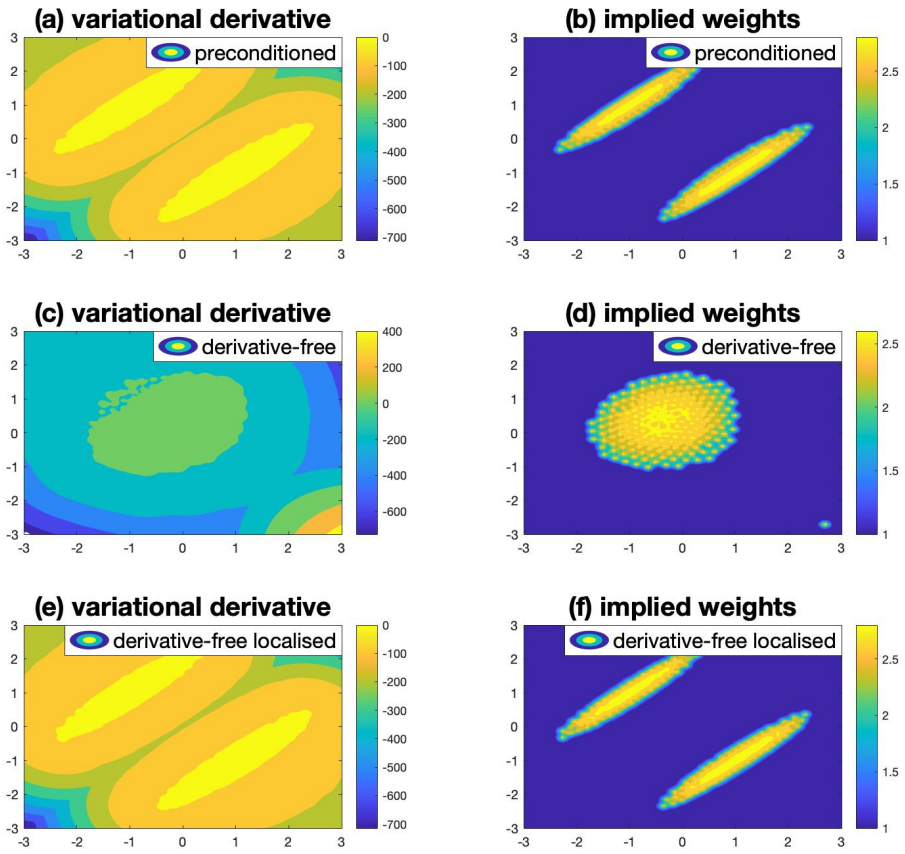

Figure 6.6: Variational derivative (left) and implied weights (right) of the RKHS Kullback–Leibler divergence for different implementations of Fokker–Planck particle dynamics: (a)-(b) preconditioned dynamics using exact gradients, (c)-(d) preconditioned gradient-free dynamics, (e)-(f) localised gradient-free dynamics.

We obtain results, which are qualitatively similar, from the corresponding implementations of the interacting Langevin dynamics. The estimate can be seen in Figure 6.7. We find again a focus of the derivative free formulation into the direction of one of the peaks. However, the localisation highly improves the approximation of the posterior distribution. In Figure 6.8 we present the kernel density estimate resulting from the localised deterministic Fokker–Planck dynamics (6.51), where we consider different choices of the localisation scaling $\gamma > 0$. For small choices of γ we can see a too strong effect of the localisation such that the particles are moving slowly, as the particles do not find nearby particles to interact with. On the other hand, for too large scaling parameter γ, the particles start to concentrating on one of the two peaks, as the localisation effect is too weak and the

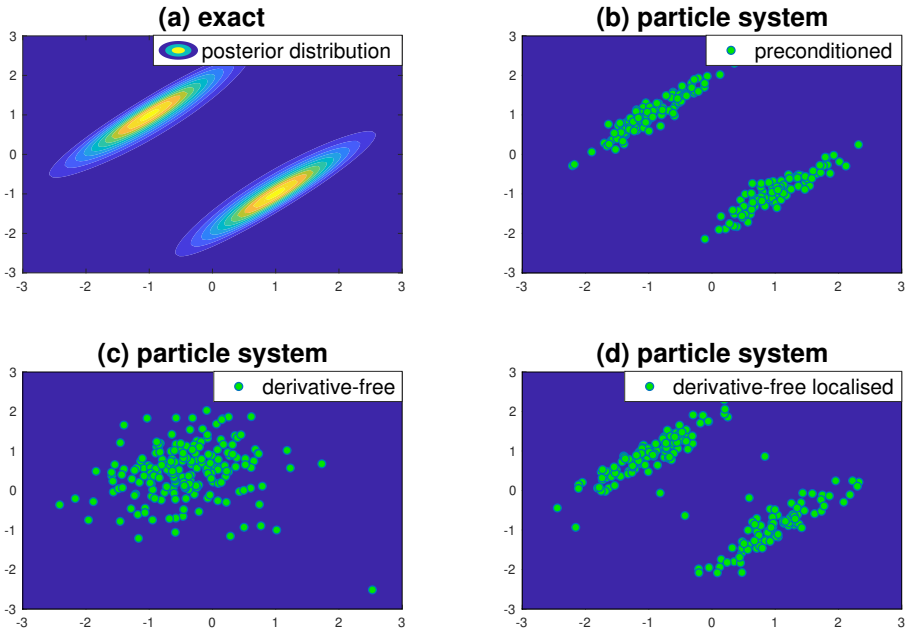

Figure 6.7: Approximations to the posterior PDF from interacting Langevin dynamics: a) exact posterior PDF, b) particle system from preconditioned dynamics using exact gradients, c) particle system from preconditioned gradient-free dynamics, d) particle system from localised gradient-free dynamics.

gradient approximation of the forward model is inaccurate.
We compare the time evolution of the potential (6.33) for different scaling localisation parameter γ in Figure 6.9, where we can see that estimate for $\gamma = 10$ is leading to a lower value of the potential compared to the estimate for $\gamma = 0.1$ with $\gamma = 1$ being optimal. The reason for this is that while the estimate for $\gamma = 10$ represents one of the peaks nearly perfectly, but ignoring the second peak, the two other estimates approximates both peaks fairly well.
To conclude, we have seen the effectiveness of the localised derivative free formulations of both the deterministic and the stochastic methods. However, one has to choose carefully the parameter γ as smaller values of γ require larger ensemble sizes J in order to provide faithful gradient approximations.

6.6.3 Scalability in high dimensions

In the following we consider a simple linear Gaussian toy example in order to study the behavior of the deterministic Fokker–Planck particle system (6.50) based on the RKHS approach. To do so, we consider a Gaussian process $GP(0, (-\Delta)^{\tau})$ represented by the KL expansion

$$u(s, x) = \sum_{k=1}^{\infty} \theta_k \varphi_k(s), \tag{6.56}$$

where Δ denotes the Laplacian operator over $\mathcal{D} = [0, 1]$ equipped with Dirichlet boundary conditions, $\theta = (\theta_k)_{k \in \mathbb{N}}$ is a sequence of independent Gaussian distributed random vari-

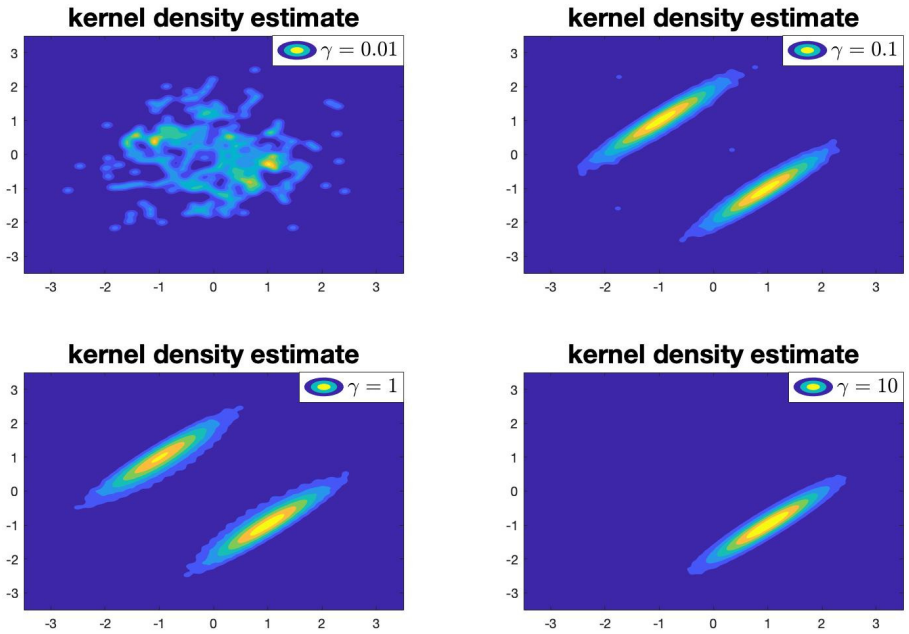

Figure 6.8: Approximations to the posterior PDF from interacting Fokker–Planck dynamics: a) $\gamma = 0.1$, b) $\gamma = 1$, c) $\gamma = 4$, d) $\gamma = 10$.

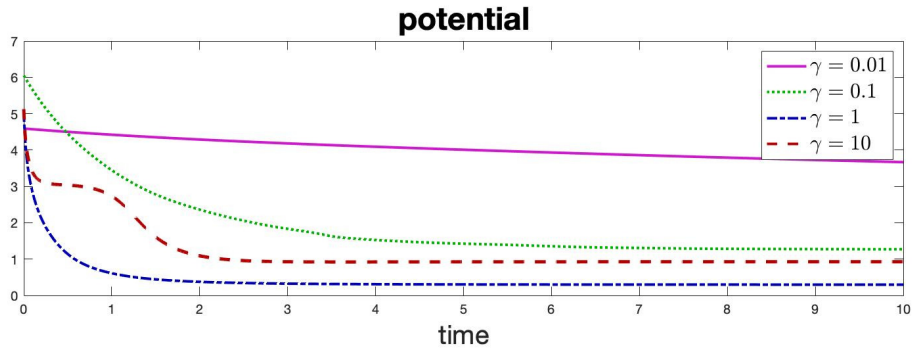

Figure 6.9: Time evolution of the potential function $\mathcal{V}$ for different choices of γ.

ables $\mathcal{N}(0, \nu_k)$ with $\nu_k = k^{-2\tau}$ and $\varphi_k(s) = \sqrt{2\pi}\sin(2\pi s)$, see Section 2.2.4 for more details. We consider the inverse problem of recovering the coefficients $\theta = (\theta_1, \dots, \theta_I)^\top \in \mathbb{R}^I$ of one observed sampled path of the Gaussian process. The KL expansion will be truncated at index I, and the state space of the Gaussian process will be discretized on a uniform grid $\mathcal{D}_l \subset [0, 1]$ with mesh size $h = 2^{-l}$. We will analyze the performance of the deterministic Fokker–Planck dynamics (6.50) for increasing the dimension in both I and l. We set a prior to $X_0 \sim \mathcal{N}(0, C_0)$ with $C_0 \in \mathbb{R}^{I\times I}$ being a diagonal matrix with entries ν_k, $k = 1, \dots, I$ and $\tau = 1$. For fixed I and fixed l, we can write the problem as linear inverse problem in the form of (2.2) where the forward model is defined by $h(\cdot) = L\cdot$, where the $k - th$ column of $L = L_{I,l} \in \mathbb{R}^{2^l\times I}$ consists of $(\varphi_k(s_1), \dots, \varphi(s_{2^l})^\top$, $s_i = i \cdot h$, $i = 1, \dots 2^l$. The reference data y will be constructed by drawing $\theta_k^\dagger \sim \mathcal{N}(0, \nu_k)$ and computing $y_{I,l} = L_{I,l}\theta^\dagger$

with $\theta^\dagger = (\theta_1^\dagger, \dots, \theta_I^\dagger)^\top$. We consider two settings:

- In the first setting we keep the truncation of the KL expansion fixed to $I = 4$ and increase the discretization of the state by choosing $l \in \{4, 6, 8\}$.
- In the second setting we increase the truncation of the KL expansion by choosing the index $I \in \{4, 6, 8\}$ and keeping the discretization of the state fixed to $l = 6$.

In our numerical results, we initialize the particle system by an i.i.d. sample of $N(0, C_0)$ and solve

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta_t^{(i)} = P_t \cdot F_t(\theta_t^{(i)})$$

for F defined by (6.32), where we consider Gaussian kernels (6.37) with different choices of B. In particular, we will choose B such that it scales with the dimension I and the ensemble size J in order to obtain good approximation results, i.e. we test

$$B_1 = \frac{c_\delta}{J^\delta}\,\mathrm{diag}((C_0)_{ii}),\ B_2 = \frac{c_\delta}{J^\delta}\,\mathrm{diag}((C_*)_{ii}) \text{ and } B_3 = \frac{c_\delta}{M^\delta}\,\mathrm{diag}((C(\theta_t))_{ii}),$$

where we define

$$\delta = \frac{1}{I+4} \quad \text{and} \quad c_\delta = \left(\frac{4}{I+2}\right)^\delta. \tag{6.57}$$

These choices of kernels correspond to the product of univariate kernels in each dimensions with optimal bandwidth for gaussian kernels minimizing the asymptotic mean integrated squared error [208, Section 6.3.1]. While in the choice of B_2 we assume to have access to the theoretical variance $\sigma_i^2 = (C_*)_{ii}$ of each component, we are using approximation $\widehat{\sigma}_i^2 = (C_0)_{ii}^2$ for the choice B_1 and $\widehat{\sigma}_i^2 = (C(\theta_t)^{xx})_{ii}$.
Testing these choices of kernels, we observe that for B_1 the approximation results are getting worse if the prior covariance C_0 is far away from the target covariance C_* as $\widehat{\sigma}_i^2 = (C_0)_{ii}^2$ is overestimating the variance of the posterior. For the choice B_2 we obtain high accurate approximation results, however, in practical situations it is an infeasible choice as the true covariance is typically unknown. The third choice B_3 corresponds to the adaptive kernel choice introduced in Remark 6.4.6. This choice helps by approximating C_* through the particle system and updating the underlying RKHS adaptively. In our numerical results, we observe for increasing dimension I but fixed l that the variance σ_i in each component gets underestimated as the particle system collapses, such that the resulting approximation fails for increasing I. To encounter this problem, we introduce a time-depending variance inflation for B_3, i.e. we consider

$$B_3^t = \frac{d_\delta}{J^\delta}\,\mathrm{diag}((C(\theta_t))_{ii} + \frac{1}{1+t}\cdot(C_0)_{ii}).$$

In the limit for t approaching infinity, we obtain again the optimal bandwidth for the product kernel, but we also obtain a better approximation result of $\widehat{\sigma}_i$. However, we still observe, that we have to increase the ensemble size J as I increases.
To illustrate our numerical results, we have created for both settings a table where we compare the trace of the estimated covariance for the posterior distribution. In Table 6.1 we keep $I = 4$ fixed and increase l from 4 to 8, whereas in Table 6.2 we keep $l = 6$ fixed and increase I from 4 to 8. The covariances for the Tables 6.1-6.2 have been estimated by solving (6.50) up to a fixed time of $T = 1000$ and producing samples according to (6.35) using the final particle system. The ODE has been solved again with the `MATLAB`

$l \setminus J$		50	100	200	theoretical
4		0.754	0.707	0.668	0.396
6	B_1	0.703	0.653	0.606	0.151
8		0.728	0.657	0.605	0.046
4		0.3873	0.404	0.399	0.396
6	B_2	0.151	0.152	0.151	0.151
8		0.046	0.046	0.046	0.046
4		0.333	0.371	0.386	0.396
6	B_3^t	0.130	0.142	0.148	0.151
8		0.041	0.045	0.045	0.046

Table 6.1: Trace of the estimated covariance in comparison for the three different choices of $B \in \{B_1, B_2, B_3^t\}$ and $l \in \{4, 6, 8\}$ and fixed $I = 4$.

$I \setminus J$		50	100	200	theoretical
4		0.723	0.658	0.604	0.151
6	B_1	0.869	0.795	0.750	0.191
8		0.994	0.931	0.873	0.217
4		0.151	0.151	0.152	0.151
6	B_2	0.180	0.186	0.188	0.191
8		0.206	0.207	0.210	0.217
4		0.143	0.154	0.156	0.151
6	B_3^t	0.093	0.126	0.156	0.191
8		0.078	0.086	0.099	0.217

Table 6.2: Trace of the estimated covariance in comparison for the three different choices of $B \in \{B_1, B_2, B_3^t\}$, $I \in \{4, 6, 8\}$ and fixed $l = 6$.

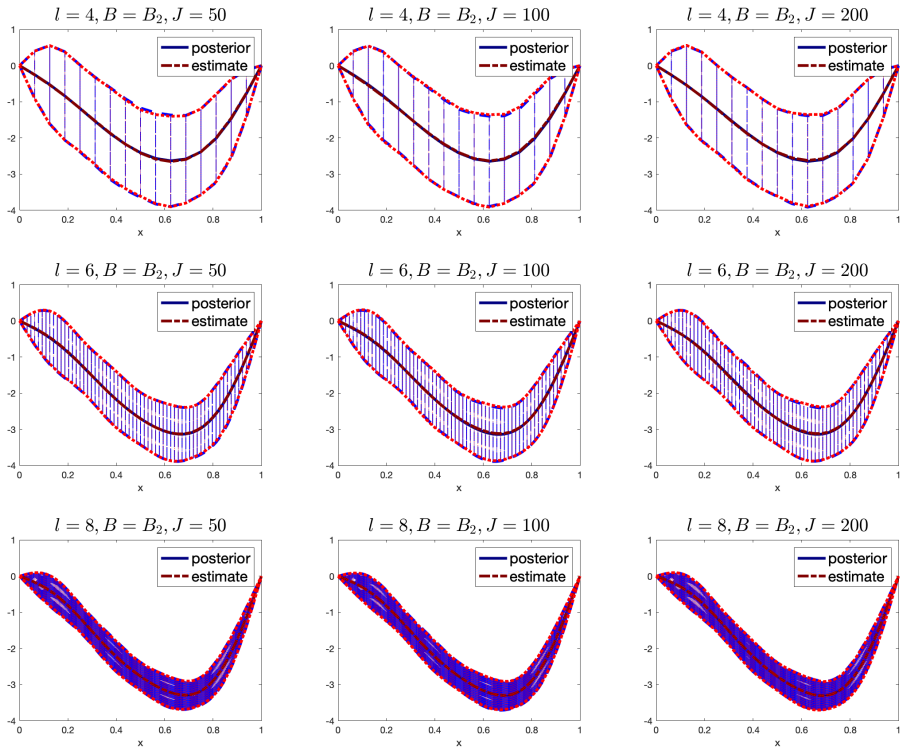

Figure 6.10: Approximations to the unknown parameter from the deterministic Fokker–Planck dynamics with $B = B_2$ for different choices of $l \in \{4, 6, 8\}$, $J \in \{50, 100, 200\}$ and fixed $I = 4$.

solver `ode45`. For keeping the size I of the parameter fixed, we observe, that we estimate the trace of the posterior covariance closely exact for B_2 and B_3^t independently of the choices of $l \in \{4, 6, 8\}$, while B_1 overestimates for each choice. In contrast, we find that for fixed discretization $h = 2^{-l}$, $l = 6$ the choice of B_3^t needs to include a larger ensemble size for increasing parameter dimension in order to estimate the trace of the posterior covariance correctly. The choice B_2 performs again very well, whereas the choice B_1 fails. These results can also be observed in Figure 6.10-Figure 6.11 for the scaling in l and in Figure 6.12-Figure 6.13 for the scaling in I, where we compare the evaluation of the KL expansion (6.56) of the kernel based methods to the evaluation of the posterior distribution. We can see again that for the choice B_3^t, the sample size has to be increased for increaing I, while for fixed I and increasing l the obtained approximation results are stable.

In summary, we observe that the deterministic Fokker–Planck particle system is a promising method as long as it is possible to choose RKHS representing the posterior distribution well. If there is no information available about the covariance structure of the posterior, it is a challenging task to choose a satisfying kernel.

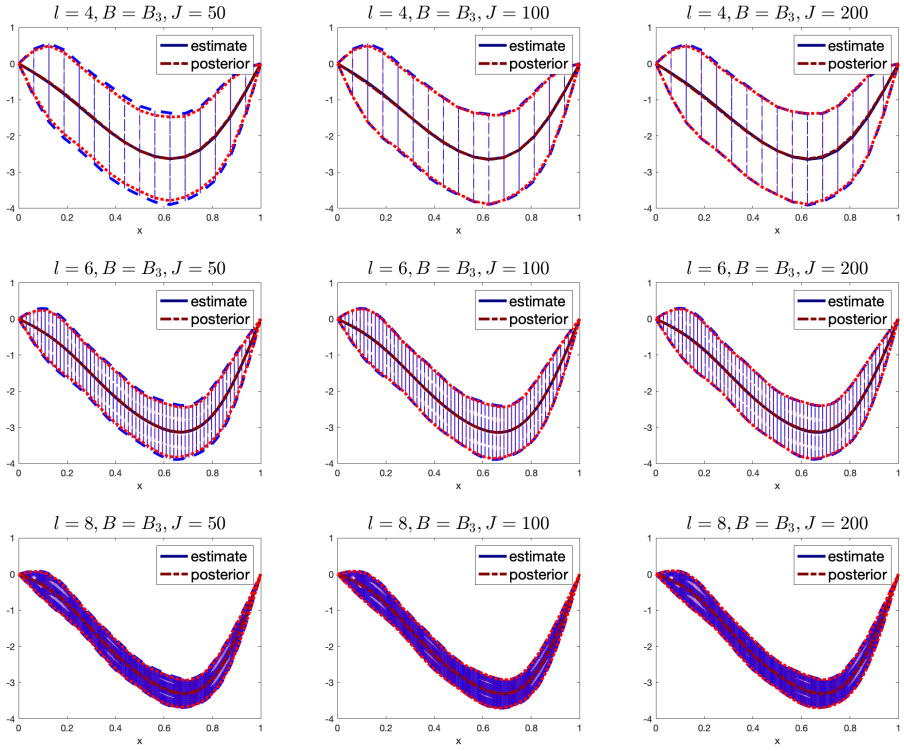

Figure 6.11: Approximations to the unknown parameter from the deterministic Fokker–Planck dynamics with $B = B_3^t$ for different choices of $l \in \{4, 6, 8\}$, $J \in \{50, 100, 200\}$ and fixed $I = 4$.

6.6.4 High dimensional example

We again consider the one-dimensional elliptic boundary-value problem from Subection 4.4.2 and 5.4.2, which is

$$\begin{aligned} -\nabla \cdot (\exp(u)\nabla p) &= f, \quad x \in D = (0,1), \\ p &= 0, \quad x \in \partial D \end{aligned} \tag{6.58}$$

for given source term $f = 1$ and unknown permeability $a = \exp(u) \in L^\infty(D)$. Recall that $p \in H_0^1(D)$ subjected to zero Dirichlet boundary conditions. Inspired by [83] we infer the coefficients of KL expansion of u, where we assume a Gaussian prior $\mathcal{N}(0, (-\Delta)^{-\tau})$. Thus, we can write $u \sim \mathcal{N}(0, (-\Delta)^{-\tau})$ a.s. through the KL expansion (6.56)
where $\theta = (\theta_k)_{k\in\mathbb{N}}$ is again a sequence of independent Gaussian distributed random variables $\mathcal{N}(0, \nu_k)$ with $\nu_k = k^{-2\tau}$ and $\varphi_k(s) = \sqrt{2\pi}\sin(k\pi s)$.
The inverse problem is to recover the coefficients $(\theta_k)_{k\in\mathbb{N}}$ corresponding to the KL expansion (6.56) given discrete noisy observations y of (6.58), $y = (\mathcal{O} \circ S)(u(s,\theta)) + \xi$, where $S : L^\infty([0,1]) \to H_0^1([0,1];\mathbb{R})$ denotes the solution operator of (6.58) and $\mathcal{O} : H_0^1([0,1];\mathbb{R}) \to \mathbb{R}^K$ denotes the observation operator, $z(\cdot) \in H_0^1([0,1];\mathbb{R}) \mapsto \mathcal{O}(z(\cdot)) = (z(s_1), \dots, z(s_K))^\top$, $s_i = \frac{i}{K}$, $i = 1, \dots, K$.
We will truncate the KL expansion (6.56) at index $I = 32$, such that the unknown param-

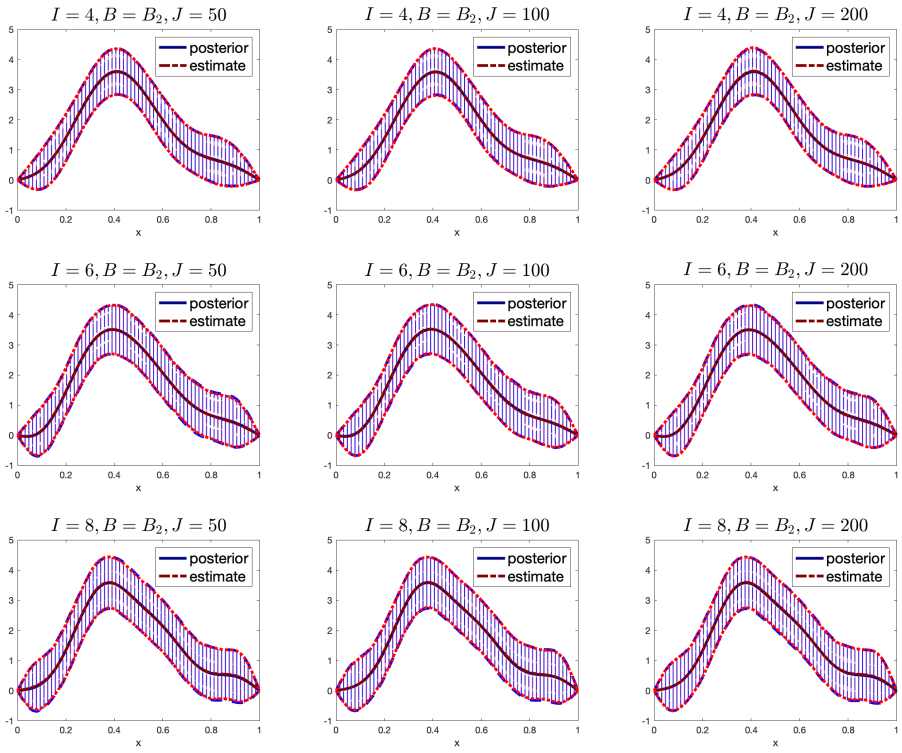

Figure 6.12: Approximations to the unknown parameter from the deterministic Fokker–Planck dynamics with $B = B_2$ for different choices of $I \in \{4, 6, 8\}$, $J \in \{50, 100, 200\}$ and fixed $l = 6$.

eters are given by $\theta = (\theta_1, \dots, \theta_I)^\top \in \mathbb{R}^I$ and we define the forward map H by

$$H : \mathbb{R}^I \to \mathbb{R}^K, \quad \text{with} \quad \theta \mapsto u(\cdot, \theta) \mapsto (\mathcal{O} \circ S)(u(\cdot, \theta)).$$

For our numerical results we replace S by an numerical solution operator for (6.58) on the grid $\mathcal{D} \subset [0, 1]$ with mesh size $h = 2^{-8}$ and restrict $u(\cdot, \theta)$ to the computational grid $s_l = l\,h$, $l = 1, \dots, 2^8 - 1$.
We set the prior to $\Theta_0 \sim \mathcal{N}(0, C_0)$, with $C_0 \in \mathbb{R}^{I \times I}$ being a diagonal matrix with entries ν_k, $k = 1, \dots, I$ and $\tau = 1.5$. We assume the measurement errors to be zero Gaussian, $\Xi \sim \mathcal{N}(0, \Gamma)$ with $\Gamma = 0.0001 \cdot \mathrm{Id}_K$ and the resulting inverse problem is of the form (2.2). The reference data y has been constructed by drawing $\theta_k^\dagger \sim \mathcal{N}(0, \nu_k)$ for $k \leq 4$ and set $\theta_k^\dagger = 0$ for $k > 4$ and realized measurement error $\xi^\dagger \sim \mathcal{N}(0, \Gamma)$ in order to compute $y = H(\theta^\dagger) + \xi^\dagger$. In our numerical experiments, we truncate the KL expansion (6.56) at $I = 32$ and take $K = 16$ equidistant observation points of the solution $p(s)$ of (6.58).
We again use the `MATLAB` solver `ode45` to solve the deterministic Fokker–Planck dynamics (6.50) and implement a Euler–Maruyama scheme this time with an adaptive step-size $\Delta t_k \leq 0.1/\beta_k$ for the interacting Langevin sampler (6.55). Here, β_k is chosen such that

$$\beta_k = \max\{\|P_{t_k}^{xh} R^{-1}(h(X_{t_k}^{(i)}) - y) + P_{t_k}^{xx} P_0^{-1}(X_{t_k}^{(i)} - \bar{x}_0)\|, \quad i = 1, \dots, M\}.$$

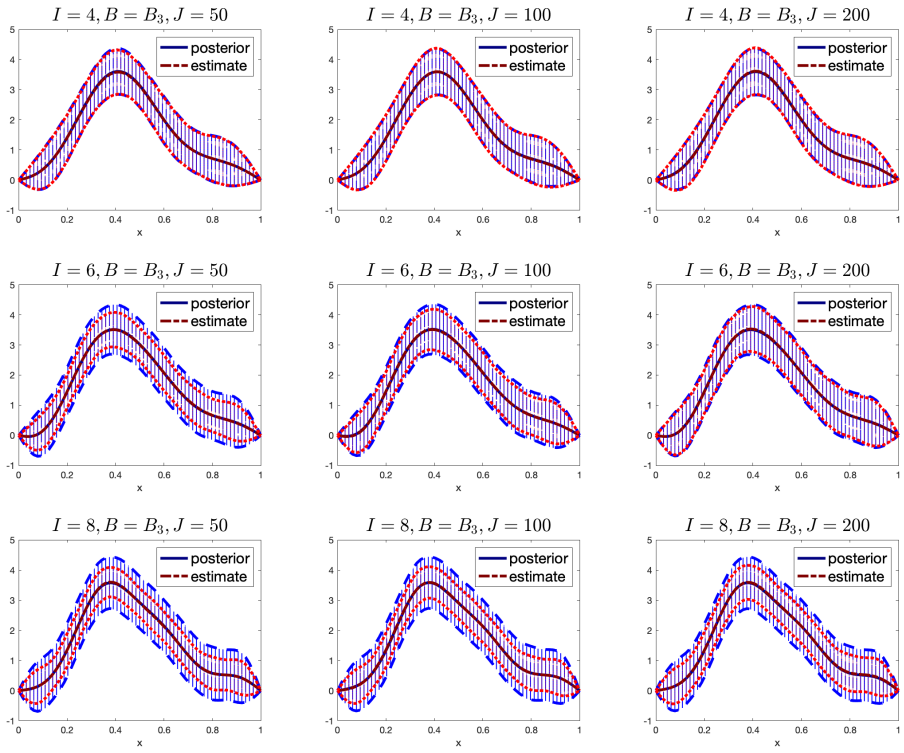

Figure 6.13: Approximations to the unknown parameter from the deterministic Fokker–Planck dynamics with $B = B_3^t$ for different choices of $I \in \{4, 6, 8\}$, $J \in \{50, 100, 200\}$ and fixed $l = 6$.

The Gaussian kernel (6.37) now depends on the current empirical covariance matrix, that is, $B_t = c_\delta / J^\delta \operatorname{diag}((C(\theta_t)^{xx})_{ii} + \frac{1}{1+t}(C_0)_{ii})$, with δ and c_δ defined in (6.57), for the preconditioned Fokker–Planck dynamics. We again inflate the B_t of order $1/t$.
We will consider similar summary statistics to the high dimensional example in [83], as our setup is quite similar.
We run both the deterministic Fokker–Planck dynamics (6.50) as well as the interacting Langevin dynamic (6.54) up to a fixed time of $T = 100$ and construct an empirical approximation to the posterior $X \mid y$ by a sample of size 512 for both methods, in order to analyse the numerical results. We use the particle system of the deterministic Fokker–Planck dynamics at final time and produce samples according to (6.35). For the Langevin sampler, we collect the required total number samples using the temporal evolution paths of the particles. Therefore, we add the current ensemble of particles every 1000 time steps to the already collected samples once the dynamics can be considered as equilibrated. As comparison we approximate the posterior by a Random Walk Metropolis Hastings algorithm with pCN proposal [54], i.e. we propose for given state θ_k

$$\hat{\theta}_{k+1} \sim \mathcal{N}(\sqrt{1 - s^2}\theta_k, s^2 C_0),$$

where s is a step size parameter. We set the step size to $s = 0.07$, which results in an acceptance rate of approximately 25% as discussed in Subsection 2.2.5.

As additional experiment we test a sequential Monte Carlo method (SMC) [129], which we combine with the interacting Langevin dynamic (6.54). In particular, we initialize by J particles $(\theta_0^{(i)})$ drawn from ρ_0, compute weights $W_0^{(i)} = 1/J,\ i = 1, \dots, J$ and proceed as follows for $n = 1, \dots, N$:

- Importance weights: update weights by $W_n^{(i)} \propto W_{n-1}^{(i)} \cdot \rho_n$, with $\sum_{i=1}^{J} W_n^{(i)} = 1$ and $\rho_n(x) = \frac{n}{N}\|y - H(x)\|_\Gamma^2$.
- If the effective sample size $(\sum_{i=1}^{J}(W_n^{(i)})^2)^{-1} < J_{\text{tol}}$, we resample according to the weights $(W_n^{(i)})$.
- we update the particles $\theta_{n-1}^{(i)} \mapsto \theta_n^{(i)}$ by running the interacting Langevin dynamic with scaled drift into direction of the data initialized by $\theta_{n-1}^{(i)}$ and up to time t_n:

$$\begin{aligned} \mathrm{d}\theta_t^{(j)} = &- \Big\{ \frac{n}{N} C^{\theta p}(\theta_t)\Gamma^{-1}(H(\theta_t^{(j)}) - y) + C(\theta_t)C_0^{-1}(\theta_t^{(j)} - m_0) \Big\}\,\mathrm{d}t \\ &+ \nabla_{\theta^{(j)}} \cdot C(\theta_t)\,\mathrm{d}t + \sqrt{2C(\theta_t)}\,\mathrm{d}W_t^{(j)}. \end{aligned}$$

We compare the SMC with the interacting Langevin dynamic (6.54) by running both methods up to a final time $T = 0.5$ with $J = 512$ particles. The SMC method will be simulated for $N = 250$, with $t_n = 0.002$, $n = 1, \dots, N$ and $J_{\text{tol}} = 256$, such that it runs up to final time $T = 0.5$, too.

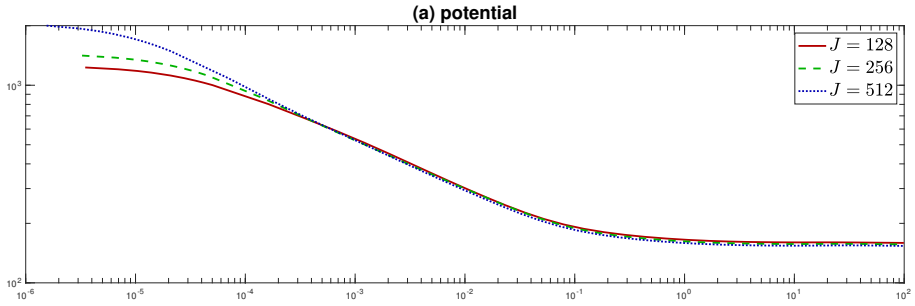

Figure 6.14: Time evolution of the potential function $\mathcal{V}$ for different choices of M (a).

Figure 6.14 shows the time evolution of the potential function (6.33) for different choices of ensemble sizes J. There is no crucial sensitivity on those parameters detectable as all parameter choices result in quite a similar behaviour. In contrast to the temporal behaviour of the potential, we can see significant differences in the particle distributions by viewing scatter plots.
Figure 6.15 shows samples resulting from different choices of ensemble size J, which shows that a large ensemble size is necessary to produce good approximations to the posterior distribution.
For the Langevin sampler we can see the effect of the introduced correction term (6.20) for small ensemble sizes M. Without correction term the resulting sample without correction concentrates in a small area, which can be seen from Figure 6.16, whereas Figure 6.17 shows that the resulting sample with corrected dynamics approximates the posterior distribution quite well already for an ensemble size $M = 16$.

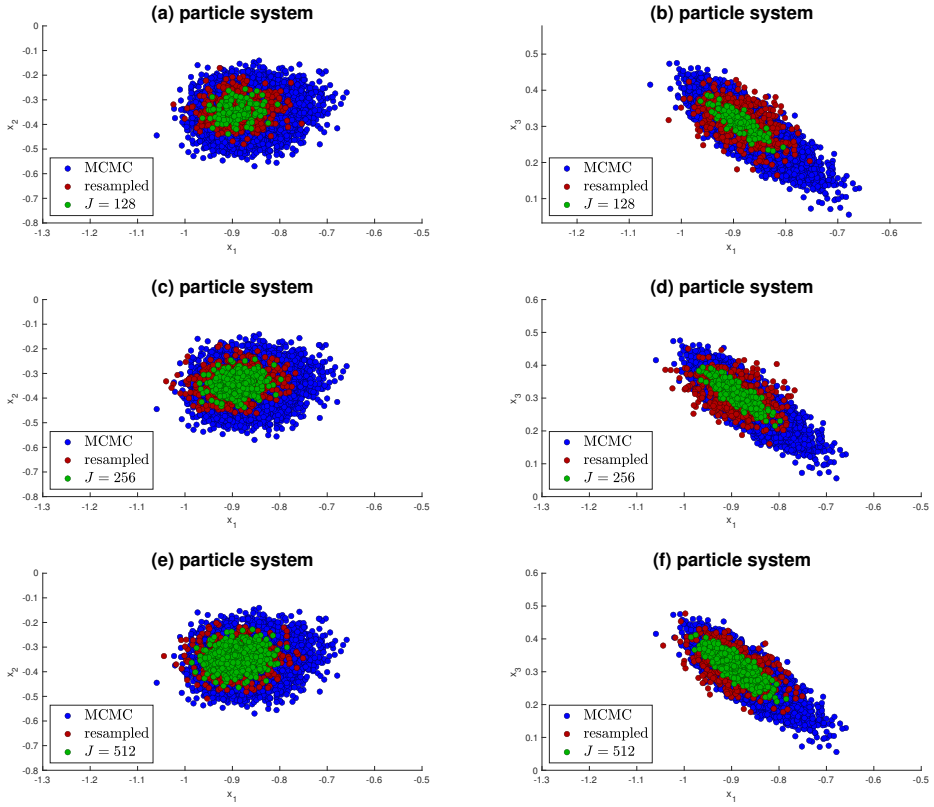

Figure 6.15: Approximations to two different marginals of the posterior PDF from deterministic Fokker–Planck dynamics for different values of the ensemble size J: a) - b) $J = 128$, c) - d) $J = 256$, e) - f) $J = 512$.

This difference also occurs in the time evolution of the spread of the ensemble

$$e_t := \frac{1}{M} \sum_{j=1}^{J} |\theta_t^{(j)} - \bar{\theta}_t|^2,$$

which can be seen in Figure 6.18. The particles of the ensemble stay spread for different choices of ensemble sizes under the corrected dynamics, while we can see again the concentration effect for the uncorrected case. Further, we compare the resulting parameter estimation for both the deterministic as well as the stochastic method. The plot on the left in Figure 6.19 shows the resulting estimate for the deterministic version, that is, the evaluation of the truncated KL expansion (6.56)

$$u^{(j)} = u(\cdot, \theta^{(j)}) = \sum_{k=1}^{I} \theta_k^{(j)} \varphi_k(\cdot).$$

Similarly, we can see the estimates resulting from the stochastic method in the plot on the right. While the stochastic method fits the mean corresponding to the Random Walk

Metropolis Hastings algorithms fairly well, the deterministic method seems to need a higher ensemble size to find the posterior distribution.
We close the discussion by analyzing the effect of the incorporation of the SMC method in the interacting Langevin dynamics. We find that the inclusion of the resampling can accelerate the convergence to the posterior distribution. This can firstly be seen in the spread over time, Figure 6.20, where we see that the occuring resampling is shifting the particle system into direction of the posterior distribution. In Figure 6.21 we see that already after final time $T = 0.5$ the SMC method produces very good approximations of the posterior distribution, while the interacting Langevin dynamics still spreads too much. This result can also be seen in the parameter estimation in Figure 6.22.
We conclude by a summary of the numerical experiment. For the deterministic Fokker–Planck dynamics we have seen the crucial dependency of the choice of the kernel. Here it turned out that it is beneficial to introduce a timedependent kernel B_t, while in the low dimensional examples it was enough to tune the parameter α for the initial kernel covariance $B = \alpha C(\theta_0)$. Further, we have demonstrated the improvement of the interacting Langevin dynamics through the correction term (6.20). While it is enough to chose a small ensemble size for the corrected sampler, we have to increase the ensemble size to $J \gg I$ in the method without correction.

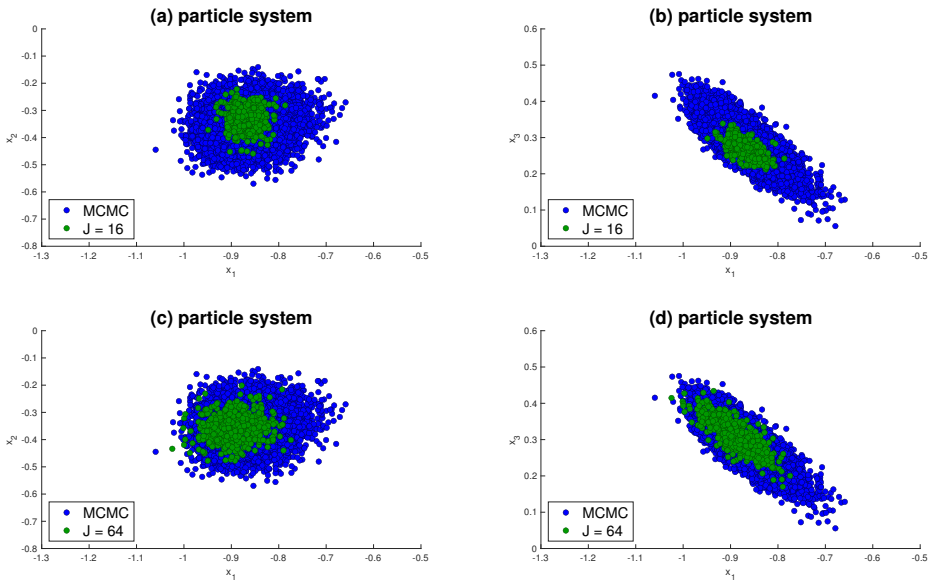

Figure 6.16: Approximations to two different marginals of the posterior PDF from interacting Langevin dynamics without correction for different values of the ensemble size J: a) - b) $J = 16$, c) - d) $J = 64$.

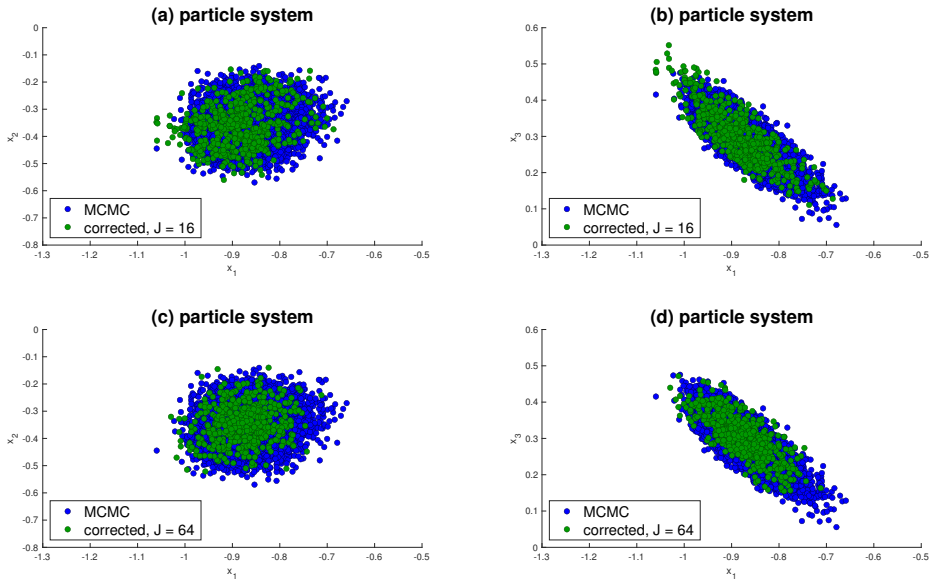

Figure 6.17: Approximations to two different marginals of the posterior PDF from interacting Langevin dynamics with correction for different values of the ensemble size J: a) - b) $J = 16$, c) - d) $J = 64$.

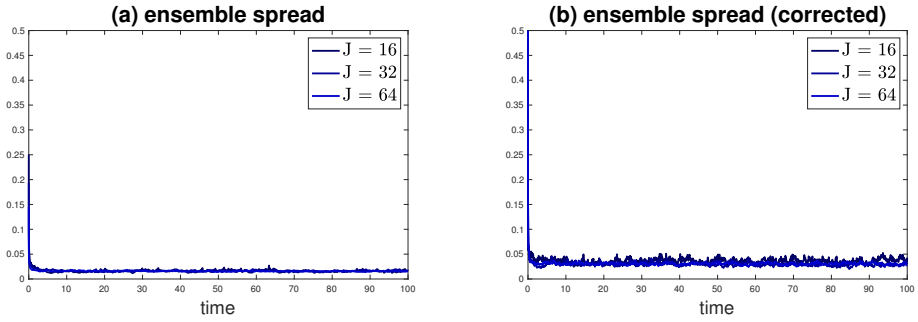

Figure 6.18: Comparison of the spread of the ensemble from interacting Langevin dynamics with and without correction over time: a) without correction, b) with correction.

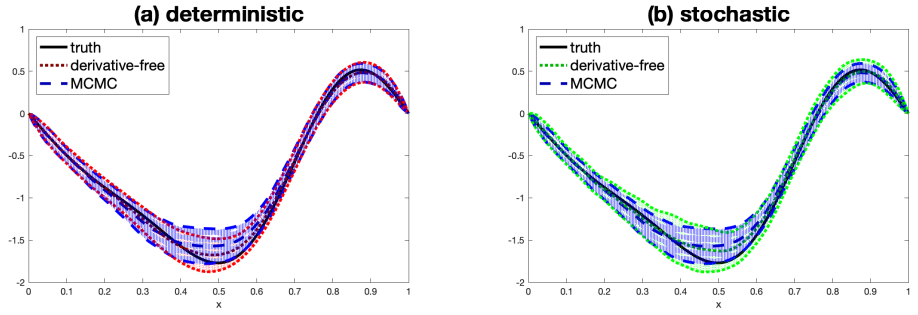

Figure 6.19: Approximations to the unknown parameter from: a) deterministic Fokker–Planck dynamics, b) interacting Langevin dynamics.

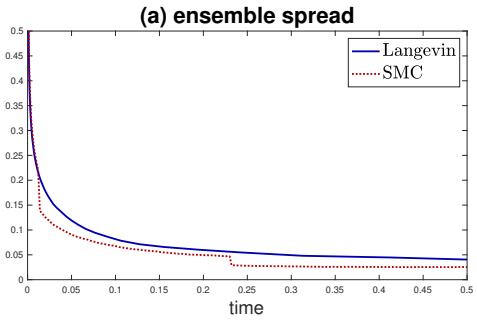

Figure 6.20: Comparison of the spread of the ensemble from the sequential Monte Carlo method and the interacting Langevin dynamics.

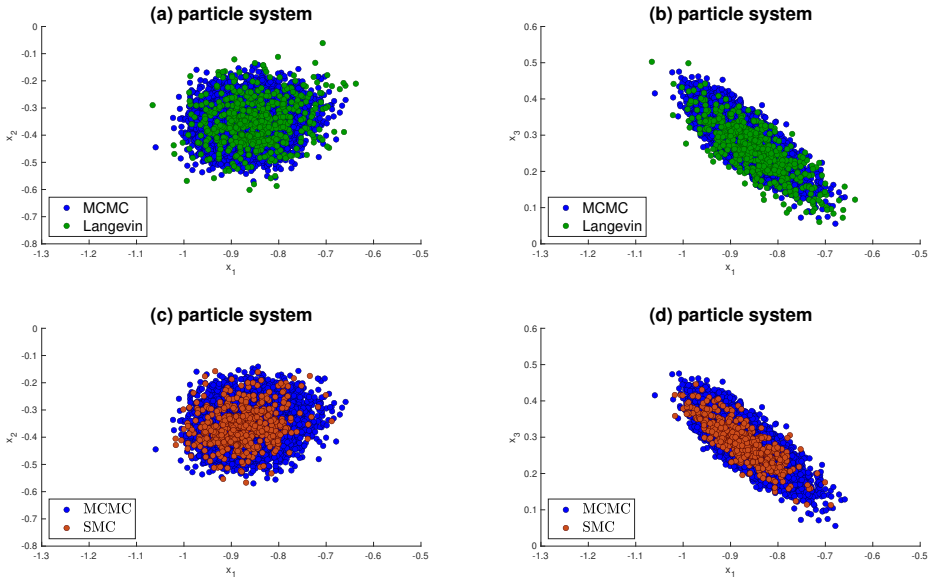

Figure 6.21: Approximations to two different marginals of the posterior PDF from sequential Monte Carlo method fixed ensemble size $J = 512$: a)-b) interacting Langevin dynamics, c)-d) sequential Monte Carlo method

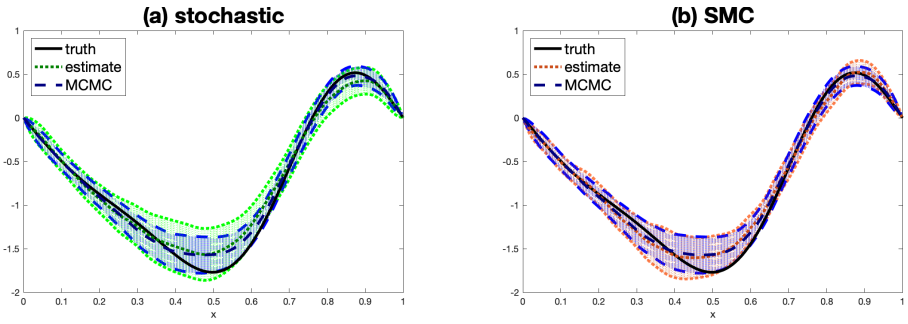

Figure 6.22: Approximations to the unknown parameter from: a) interacting Langevin dynamics, b) sequential Monte Carlo method.

7 Machine learning application in inverse problems

In this chapter, we consider different applications of machine learning in inverse problems. The first application presented in Section 7.1 is about a regularization parameter choice based on data-driven learning. The basic idea is to consider a bilevel optimization problem, where the upper-level problem is given by a risk measure between the unknown parameter and the regularized solution, whereas the lower-level problem is the corresponding regularized optimization problem for a given regularization parameter. We formulate the method as an empirical risk minimization problem and provide both offline and online consistency results for the large training data limit. In the second Section 7.2, we incorporate neural networks into the setting of inverse problems. The complex forward model will be replaced by a neural network surrogate model, which will be informed through the underlying physics in the model. We then train the neural network and the unknown parameter in an one-shot fashion. Furthermore, we provide the connection to the Bayesian approach for inverse problems and are able to apply the EKI method.

7.1 Data driven regularization

This section is devoted to provide some theoretical verification for the introduced data-driven regularization within EKI in Section 5.3.1. We formulate the task of finding the best possible regularization parameter for general inverse problems of the form (2.2) as empirical risk minimization problem which we analyze in the large data behavior. We provide both offline and online consistency results for increasing size of training data. In the offline setting we analyze the empirical risk minimization problem itself and quantify the accuracy of the corresponding optimal solution, while in the online setting we formulate the stochastic gradient descent (SGD) method in order to reduce the associated computational costs. In both settings, we firstly provide an abstract but general consistency result, which can be applied to general nonlinear inverse problems with general regularization function, and secondly we verify the presented results for linear problems under Tikhonov regularization.
We recall, that for general inverse problems of the form (2.2), following Definition 2.1.4 we aim to minimize the Tikhonov regularized loss function

$$T_\kappa(\theta) = \frac{1}{2}\|H(\theta) - y\|^2 + \varphi_\kappa(\theta),$$

where $\kappa > 0$ is the regularization parameter and $\varphi_\kappa : \mathcal{X} \to \mathbb{R}_+$ is the regularization function. While our general results, can be extended to general regularization function φ_κ, we provide verification of the presented results for Tikhonov regularization.

As we have seen in the discussion in Section 2.1 and in Chapter 5, the choice of the regularization parameter $\kappa > 0$ crucially effects the resulting solution of the underlying inverse problem. In order to choose the regularization parameter within inverse problems, a recently proposed method relies on bilevel optimization [36, 47, 68, 63, 104].

It seeks to learn the regularization parameter in a variational manner, and it can be viewed as a data-driven regularization [10]. To formulate this approach, we view unknown parameter $\Theta \in \mathcal{X}$ and the data $Y \in \mathcal{Y}$ in the model (2.2) as a jointly distributed random variable with distribution $\mathbb{Q}_{(\Theta,Y)}$. To find the best possible regularization parameter of the model (2.2), the bilevel minimization seeks to solve

$$\begin{aligned} \kappa_* &\in \arg\min_{\kappa>0} F(\kappa), \quad F(\kappa) = \mathbb{E}_{\mathbb{Q}_{(\Theta,Y)}}[\mathcal{L}_{\mathcal{X}}(\theta_\kappa(Y), \Theta)], \quad \text{(upper level)} \\ \theta_\kappa(Y) &:= \arg\min_{\theta\in\mathcal{X}} \mathcal{L}_{\mathcal{Y}}(H(\theta), Y) + \varphi_\kappa(\theta), \qquad \text{(lower level)} \end{aligned} \tag{7.1}$$

where $\mathcal{L}_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ is some metric in the parameter space $\mathcal{X}$ and $\mathcal{L}_{\mathcal{Y}} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ some metric in the observation space. The upper level problem seeks to minimize the distance between the unknown parameter Θ and the regularized solution corresponding to its data Y, which is computed through $\theta_\kappa(Y)$ in the lower level problem. We have assumed here, that for the lower level problem there exists unique solutions. To solve this (stochastic) bilevel optimization problem, we assume that we have access to training data, given through samples of $(\Theta_i, Y_i) \sim \mu_{(\Theta,Y)}$, and the function F in (7.1) can be approximated by its empirical Monte–Carlo approximation.

Holler et. al [104] consider bilevel optimization for PDE based inverse problems with Tikhonov regularization. They provide theory which suggests existence of solutions and formulate their problem as an optimal control problem. Learning of regularization parameters of Tikhonov regularization have also been discussed in [47, 214]. The work of De los Reyes, Schönlieb [36, 68, 155, 63] and coauthors considered the application of bilevel optimization to denoising and deblurring, where non-smooth regularization is used such as total variation and Bregman regularization. The latter forms of regularization are useful in imaging as they preserve non-smooth features, such as edges and straight lines.

We start our discussion in Section 7.1.1 by formulating the empirical risk minimization problem and provide offline consistency results for recovery of the regularization parameter. In section 7.1.2 we formulate the stochastic gradient descent method for the bilevel optimization approach in order to quantify the regularization parameter online. The discussion on bilevel learning will be closed with several numerical examples in Section 7.1.3.

7.1.1 Regularization parameter offline recovery

In this section, we discuss how to use offline bilevel optimization to recover regularization parameters. We also show the solution is statistically consistent under suitable conditions.

Offline bilevel optimization

As stated before, in regularization parameter learning by bilevel optimization we view the unknown parameter Θ and the data Y as a jointly distributed random variable with distribution $\mathbb{Q}_{(\Theta,Y)}$, see e.g. [10] for more details. Recall the bilevel optimization problem

is given by

$$\kappa_* = \arg\min_{\kappa\in\Lambda} F(\kappa), \quad F(\kappa) = \mathbb{E}_{\mathbb{Q}_{(\Theta,Y)}}[\mathcal{L}_{\mathcal{X}}(\theta_\kappa(Y),\Theta)], \qquad \text{(upper level)}$$
$$\theta_\kappa(Y) := \arg\min_{\theta\in\mathcal{X}} \Psi(\kappa,\theta,Y), \quad \Psi(\kappa,\theta,y) := \mathcal{L}_{\mathcal{Y}}(H(\theta),y) + \varphi_\kappa(\theta), \text{ (lower level)}$$

where $\mathcal{L}_{\mathcal{X}}$ denotes a discrepancy function in the parameter space $\mathcal{X} := \mathbb{R}^I$ and $\mathcal{L}_{\mathcal{Y}}$ denotes a discrepancy function in the observation space $\mathcal{Y} := \mathbb{R}^K$. The function $\varphi_\kappa(\theta)$ represents the regularization with parameter $\kappa \in \Lambda$. Here, Λ represents the range of regularization parameters which often come from physical constraints. For simplicity, we assume all the functions here are continuous and integrable, and so are their first and second order derivatives with respect to κ.
In general, we do not know the exact distribution $\mathbb{Q}$ in the upper level of (7.1). However, we assume to have access to training data $(\theta^{(j)}, y^{(j)})_{j=1}^n$, arising as i.i.d. samples from $\mathbb{Q}_{(\Theta,Y)}$. Using these data, we approximate F in (7.1) by its empirical average:

$$\widehat{F}_n = \frac{1}{n}\sum_{j=1}^{n} \mathcal{L}_{\mathcal{X}}(\theta_\kappa(y^{(j)}),\theta^{(j)}). \tag{7.2}$$

Solving this problem leads to a data-driven estimator of the regularization parameter,

$$\begin{aligned} \widehat{\kappa}_n &= \arg\min_{\kappa\in\Lambda} \widehat{F}_n, \\ \theta_\kappa(y^{(j)}) &= \arg\min_{\theta\in\mathcal{X}} \mathcal{L}_{\mathcal{Y}}(H(\theta),y^{(j)}) + \varphi_\kappa(\theta). \end{aligned} \tag{7.3}$$

This method of estimation is often known as empirical risk minimization in machine learning [210]. We refer to this is as "offline" since minimizing $\widehat{F}_n$ involves all n data points at each algorithmic iteration. With $\widehat{\kappa}_n$ being formulated, it is of natural interest to investigate its convergence to the true parameter κ_*, when the sample size increases. Consistency analysis is of central interest in the study of statistics. In particular, if $\widehat{\kappa}_n$ is the global minimum of $\widehat{F}_n$, we formulate the following theorem 5.2.3 [23] from Bickel and Doksum in our notation

Theorem 7.1.1. *Suppose for any $\epsilon > 0$*

$$\mathbb{P}(\sup\{\kappa\in\Lambda, |\widehat{F}_n(\kappa) - F(\kappa)|\} > \epsilon) \to 0,$$

as $n\to\infty$, $\widehat{\kappa}_n$ is the global minimizer of $\widehat{F}_n$, and κ_* is the unique minimizer of F, then $\widehat{\kappa}_n$ is a consistent estimator.

In more practical scenarios, the finding of $\widehat{\kappa}_n$ relies on the choice of optimization algorithms. If we are using gradient based algorithms, such as gradient descent, $\widehat{\kappa}_n$ can be the global minimum of $\widehat{F}_n$ if $\widehat{F}_n$ is convex. More generally, we can only assume $\widehat{\kappa}_n$ to be a stationary point of $\widehat{F}_n$, i.e. $\nabla\widehat{F}_n(\widehat{\kappa}_n) = 0$. In such situations, we provide the following alternative tool replacing Theorem 7.1.1:

Proposition 7.1.2. *Suppose F is C^2, κ_* is a local minimum of F, and $\widehat{\kappa}_n$ is a local minimum of $\widehat{F}_n$. Let $\mathcal{D}$ be an open convex neighborhood of κ_* in the parameter space and c_0 be a positive constant. We denote $\mathcal{A}_n$ as the event*

$$\mathcal{A}_n = \{\widehat{\kappa}_n \in \mathcal{D}, \nabla^2_\kappa \widehat{F}_n(\kappa) \succeq c_0\,\mathrm{Id} \ \textit{ for all } \kappa\in\mathcal{D}\}.$$

When $\mathcal{A}_n$ takes place, the following holds:

$$\|\widehat{\kappa}_n - \kappa_*\| \leq \frac{\|\nabla_\kappa \widehat{F}_n(\kappa_*) - \nabla_\kappa F(\kappa_*)\|}{c_0}.$$

In particular, we have

$$\mathbb{E}\mathbb{1}_{\mathcal{A}_n}\|\widehat{\kappa}_n - \kappa_*\| \leq \frac{\sqrt{\operatorname{trace}(\operatorname{Var}(\nabla_\kappa f(\kappa_*, \Theta, Y)))}}{c_0\sqrt{n}}.$$

Proof. We denote the data couple (θ, y) by z and denote the data loss function by

$$f(\kappa, z) = \mathcal{L}_{\mathcal{X}}(\theta_\kappa(y), \theta).$$

in order to simplify notation.
When $\widehat{\kappa}_n \in \mathcal{D}$, we apply the fundamental theorem of calculus on $\nabla_\kappa \widehat{F}_n$, and find

$$\nabla_\kappa \widehat{F}_n(\kappa_*) = \nabla_\kappa \widehat{F}_n(\widehat{\kappa}_n) + \int_0^1 \nabla_\kappa^2 \widehat{F}_n(s\kappa_* + (1-s)\widehat{\kappa}_n)(\kappa_* - \widehat{\kappa}_n)ds = A_F(\widehat{\kappa}_n - \kappa_*),$$

where

$$A_F := \int_0^1 \nabla_\kappa^2 \widehat{F}_n((1-s)\widehat{\kappa}_n + s\kappa_*)ds \succeq c_0 \operatorname{Id}.$$

We note that

$$\begin{aligned} 0 = \nabla_\kappa F(\kappa_*) &= \nabla_\kappa F(\kappa_*) - \nabla_\kappa \widehat{F}_n(\kappa_*) + \nabla_\kappa \widehat{F}_n(\kappa_*) \\ &= A_F(\widehat{\kappa}_n - \kappa_*) + \nabla_\kappa F(\kappa_*) - \nabla_\kappa \widehat{F}_n(\kappa_*). \end{aligned}$$

Hence,

$$-\left(\nabla_\kappa \widehat{F}_n(\kappa_*) - \nabla_\kappa F(\kappa_*)\right) = A_F(\widehat{\kappa}_n - \kappa_*),$$

which implies a formula for the error $\kappa_* - \widehat{\kappa}_n$

$$\|\kappa_* - \widehat{\kappa}_n\| = \left\| A_F^{-1}\left(\nabla_\kappa \widehat{F}_n(\kappa_*) - \nabla_\kappa F(\kappa_*)\right)\right\| \leq c_0^{-1}\left\|\nabla_\kappa \widehat{F}_n(\kappa_*) - \nabla_\kappa F(\kappa_*)\right\|.$$

We use $\nabla_\kappa \mathbb{E} f(\kappa, Z) = \mathbb{E}\nabla_\kappa f(\kappa, Z)$, see [199, Theorem 12.5], and obtain

$$\nabla_\kappa \widehat{F}_n(\kappa_*) - \nabla_\kappa F(\kappa_*) = \frac{1}{n}\sum_{i=1}^n \nabla_\kappa f(\kappa_*, z_i) - \mathbb{E}\nabla_\kappa f(\kappa_*, Z).$$

It follows

$$\mathbb{E}\|\nabla_\kappa \widehat{F}_n(\kappa_*) - \nabla_\kappa F(\kappa_*)\|^2 = \frac{1}{n}\operatorname{trace}(\operatorname{Var}(\nabla_\kappa f(\kappa_*, Z))).$$

by Cauchy-Schwarz we imply the second assertion

$$\mathbb{E}\mathbb{1}_{\mathcal{A}_n}\|\nabla_\kappa \widehat{F}_n(\kappa_*) - \nabla_\kappa F(\kappa_*)\| \leq \sqrt{\mathbb{E}\|\nabla_\kappa \widehat{F}_n(\kappa_*) - \nabla_\kappa F(\kappa_*)\|^2}.$$

□

The stated Proposition 7.1.2 provides two claims: The first claim suggests, that we can have more accurate estimates on large or medium deviations. And secondly, we can see that κ_n converges to κ_* with rate of $1/\sqrt{n}$. While we did not specify assumptions on the forward model, the regularization function or the underlying distribution of (Θ, Y), we are going to show how to apply this result to linear forward models with Tikhonov regularization.
Furthermore, for Proposition 7.1.2 we need to assume that $\widehat{F}_n$ is locally Lipschitz in a domain where $\widehat{\kappa}_n$ and κ_* are in it. This is cruicial as there might be multiple local minimums, and we will have issues to distinguish different local minimums.
In order to apply Proposition 7.1.2, one needs to find $\mathcal{D}$ and bound the probability of outlier cases $\mathcal{A}_n^{\complement}$. This procedure can be nontrivial, and requires some advanced tools from probability. We demonstrate how to do so for the linear inverse problem.

Offline consistency analysis with linear observation models

In this section, we demonstrate how to apply Proposition 7.1.2 for linear forward models with Tikhonov regularization. In particular, we assume $\theta \in \mathbb{R}^d$ and the data y is observed through a matrix $L \in \mathbb{R}^{K\times I}$

$$y = L\theta + \xi,$$

with Gaussian prior information $\theta \sim \mathcal{N}(0, \frac{1}{\kappa_*}C_0)$ and Gaussian noise $\xi \sim \mathcal{N}(0, \Gamma)$. The common choice of discrepancy functions in the lower level are the corresponding negative log-likelihoods

$$\mathcal{L}_{\mathcal{Y}}(H(\theta), y) = \frac{1}{2}\|L\theta - y\|_{\Gamma}^2, \quad \varphi_\kappa(\theta) = \frac{\kappa}{2}\|\theta\|_{C_0}^2.$$

Since both of these functions are quadratic in θ, the lower level optimization problem has an explicit solution

$$\theta_\kappa(y) = (L^\top\Gamma^{-1}L + \kappa C_0)^{-1}L^\top\Gamma^{-1}y_i,$$

see Theorem 2.1.6. If we use the root-mean-square error in the upper level to learn κ, the discrepancy function is given by

$$f(\kappa, \theta, y) = \|\theta_\kappa(y) - \theta\|^2.$$

and the empirical loss function is defined by

$$\widehat{F}_n(\kappa) = \frac{1}{n}\sum_{i=1}^{n}\|\theta_\kappa(y_i) - \theta_i\|^2.$$

It is worth mentioning that $F(\kappa)$ is not convex on the real line despite that H is linear. The detailed calculation can be found in Remark 7.1.7. Hence, in Proposition 7.1.2 it is necessary to introduce the local region $\mathcal{D}$ such that F is convex inside.
Since the formulation of θ_κ involves the inversion of matrix $L^\top\Gamma^{-1}L + \kappa C_0$, such an operation may be unstable for κ approaching ∞. When κ approaches ∞, the gradient of $\widehat{F}_n$ approaches zero, so ∞ can be a stationary point that an optimization algorithm tries to converge to. To avoid these issues, we assume that there are lower and upper bounds such that

$$0 < \kappa_l < \frac{1}{2}\kappa_* < \frac{3}{2}\kappa_* < \kappa_u,$$

where κ_l can be chosen as a very small number and κ_u can be very large. Their values often can be obtained from physical restrictions from the underlying inverse problem. By

assuming their existence, we can restrict $\widehat{\kappa}_n$ to be in the interval $\Lambda = (\kappa_l, \kappa_u)$. We are now ready to formulate our main result in the offline recovery setting. In particular, we show $\widehat{\kappa}_n$ converges to κ_* with high probability.

Theorem 7.1.3. *Suppose $\widehat{\kappa}_n \in (\kappa_l, \kappa_u)$ is a local minimum of $\widehat{F}_n$. Then there exist constants $C_*, c_* > 0$ such that for any $\varepsilon \in (0,1)$ and $n \in \mathbb{N}$,*

$$\mathbb{P}(|\widehat{\kappa}_n - \kappa_*| > \varepsilon, \kappa_l < \widehat{\kappa}_n < \kappa_u) \leq C_* \exp(-c_* n \min(\varepsilon, \varepsilon^2)).$$

The values of $C_*, c_* > 0$ depend on $\kappa_l, \kappa_u, \kappa_*, C_0$ but not on n.

Since we can obtain consistency assuming that $\widehat{\kappa}_n$ is a local minimum, we do not demonstrate how to implement Theorem 7.1.1 for the more restrictive scenario where $\widehat{\kappa}_n$ is a global minimum.

Remark 7.1.4. *We note that in the Gaussian setting with Tikhonov regularization one can also estimate κ_* empirically by using the maximum likelihood estimator*

$$\widehat{\kappa}_n = I \cdot \left(\frac{1}{n} \sum_{j=1}^{n} (\theta^{(j)})^\top C_0^{-1} \theta^{(j)} \right)^{-1},$$

where I denotes the dimension of $\mathbb{R}^I$. However, only in the Gaussian setting with Tikhonov regularization the estimate will lead to the optimal solution of (7.1). When considering alternative regularization, or dropping the Gaussian assumption on θ, it is not clear whether this approach still leads to a good estimate of κ_*.

Before presenting the proof of Theorem 7.1.3, in the following we will formulate various auxiliary results. We denote $D := C_0^{1/2} L^\top \Gamma^{-1/2}, \Omega_0 = C_0^{-1}, v_i = \Omega_0^{1/2} \theta_i$, and $\xi_i = -\Gamma^{-1/2}(L\theta_i - y_i) \sim \mathcal{N}(0, \mathrm{Id})$ and we note that

$$\begin{aligned}
(L^\top \Gamma^{-1} L + \kappa C_0^{-1})^{-1} L^\top \Gamma^{-1} y_i - \theta_i &= (L^\top \Gamma^{-1} L + \kappa C_0^{-1})^{-1} L^\top \Gamma^{-1} (L\theta_i + \xi_i) - \theta_i \\
&= (L^\top \Gamma^{-1} L + \kappa C_0^{-1})^{-1} (\kappa C_0^{-1} \theta_i + L^\top \Gamma^{-1/2} \xi_i) \\
&= C_0^{1/2} (C_0^{1/2} L^\top \Gamma^{-1} L C_0^{1/2} + \kappa \,\mathrm{Id})^{-1} C_0^{1/2} (\kappa C_0^{-1} \theta_i + L^\top \Gamma^{-1/2} \xi_i) \\
&= C_0^{1/2} (DD^\top + \kappa \,\mathrm{Id})^{-1} (\kappa v_i + D\xi_i).
\end{aligned}$$

Therefore we define

$$Q_\kappa = (DD^\top + \kappa \,\mathrm{Id})^{-1},$$

and we can express f by the introduced notation

$$\begin{aligned}
f(\kappa, z) &= \mathrm{trace}(Q_\kappa C_0 Q_\kappa (\kappa v + D\xi)(\kappa v + D\xi)^\top) \\
&= \mathrm{trace}(Q_\kappa C_0 Q_\kappa (\kappa^2 v v^\top + 2\kappa D\xi v^\top + D\xi\xi^\top D^\top)).
\end{aligned}$$

We further introduce the following notation

$$P_1 = Q_\kappa C_0 Q_\kappa, \quad P_2 = \frac{\partial P_1}{\partial \kappa} = -(Q_\kappa^2 C_0 Q_\kappa + Q_\kappa C_0 Q_\kappa^2),$$

$$P_3 = \frac{\partial P_2}{\partial \kappa} = 2(Q_\kappa^3 C_0 Q_\kappa + Q_\kappa^2 C_0 Q_\kappa^2 + Q_\kappa C_0 Q_\kappa^3),$$

$$P_4 = \frac{\partial P_3}{\partial \kappa} = -6(Q_\kappa^4 C_0 Q_\kappa + Q_\kappa^3 C_0 Q_\kappa^2 + Q_\kappa^2 C_0 Q_\kappa^3 + Q_\kappa C_0 Q_\kappa^4).$$

Note that $\|Q_\kappa\| \leq \kappa^{-1}$ and it follows

$$|\operatorname{trace}(Q_\kappa^k C_0 Q_\kappa^j)| = |\operatorname{trace}(C_0 Q_\kappa^{j+k})| \leq \frac{1}{\kappa^{j+k}} \operatorname{trace}(C_0),$$

$$\|Q_\kappa^k C_0 Q_\kappa^j\| \leq \|Q_\kappa\|^{j+k}\|C_0\| \leq \frac{1}{\kappa^{j+k}}\|C_0\|,$$

$$\|Q_\kappa^k C_0 Q_\kappa^j\|_{\mathcal{F}} \leq \|Q_\kappa^k\|\|C_0 Q_\kappa^j\|_{\mathcal{F}} \leq \|Q_\kappa\|^{j+k}\|C_0\|_{\mathcal{F}} \leq \frac{1}{\kappa^{j+k}}\|C_0\|_{\mathcal{F}}.$$

Finally, for any choice of T being $T(A) = |\operatorname{trace}(A)|$ or $T(A) = \|A\|$ or $T(A) = \|A\|_{\mathcal{F}}$, we all have

$$T(P_k) \leq (\frac{2}{\kappa})^{k+1} T(C_0).$$

Using these notations, we can compute derivatives of f

$$\partial_\kappa f(\kappa, z) = \operatorname{trace}\left(P_2(\kappa^2 vv^\top + 2\kappa D\xi v^\top + D\xi\xi^\top D^\top) + P_1(2\kappa vv^\top + 2D\xi v^\top)\right),$$
$$\partial_\kappa^2 f(\kappa, z) = \operatorname{trace}\left(P_3(\kappa^2 vv^\top + 2\kappa D\xi v^\top + D\xi\xi^\top D^\top) + 4P_2(\kappa vv^\top + D\xi v^\top) + 2P_1 vv^\top\right),$$
$$\partial_\kappa^3 f(\kappa, z) = \operatorname{trace}\left(P_4(\kappa^2 vv^\top + 2\kappa D\xi v^\top + D\xi\xi^\top D^\top) + 6P_3(\kappa vv^\top + D\xi v^\top) + 6P_2 vv^\top\right).$$

Pointwise consistency analysis In order to apply Proposition 7.1.2, we have to show that the gradient of $\widehat{F}_n(\kappa)$ is a good approximation of $\nabla F(\kappa)$ at $\kappa = \kappa_*$ with high probability. This is actually true for general κ. We start by proving consistency of the sample covariance.

Lemma 7.1.5. *Let $X^{(i)} \in \mathbb{R}^d$ and $Y^{(i)} \in \mathbb{R}^{d_y}$ be i.i.d. $\mathcal{N}(0, \mathrm{Id}_d)$ and $\mathcal{N}(0, \mathrm{Id}_{d_y})$ respectively, $i \in \mathbb{N}$, and let $\Sigma \in \mathbb{R}^{d\times d}$ be fixed. There exists a universal constant $c > 0$, such that for any n and*

$$C_n = \frac{1}{n}\sum_{i=1}^n X^{(i)}(X^{(i)})^\top, \quad B_n = \frac{1}{n}\sum_{i=1}^n X^{(i)}(Y^{(i)})^\top,$$

the following holds for all $t > 0$

$$\mathbb{P}(|\operatorname{trace}(\Sigma C_n) - \operatorname{trace}(\Sigma)| > t) \leq 2\exp\left(-cn\min\left(\frac{t^2}{\|\Sigma\|_{\mathcal{F}}^2}, \frac{t}{\|\Sigma\|}\right)\right),$$

$$\mathbb{P}(|\operatorname{trace}(\Sigma B_n)| > t) \leq 2\exp\left(-cn\min\left(\frac{t^2}{\|\Sigma\|_{\mathcal{F}}^2}, \frac{t}{\|\Sigma\|}\right)\right).$$

Proof. We first write

$$\operatorname{trace}(\Sigma X^{(i)}(X^{(i)})^\top) = (X^{(i)})^\top \Sigma X^{(i)},$$

and we define the block-diagonal matrix $D_\Sigma \in \mathbb{R}^{nd\times nd}$ consisting of n blocks of Σ, and $Z = [X^{(1)}; X^{(2)}; \cdots; X^{(n)}] \in \mathbb{R}^{nd}$. Note that

$$\operatorname{trace}(\Sigma C_X) = Z^\top(\frac{1}{n}D_\Sigma)Z.$$

By application of the Hanson–Wright inequality [193, Theorem 1.1], we obtain for some constants c_0 and K_0,

$$\mathbb{P}(|\operatorname{trace}(\Sigma C_X) - \operatorname{trace}(\Sigma)| > t) \leq 2\exp\left(-c_0 \min\left(\frac{t^2}{K_0^4 \|\frac{1}{n} D_\Sigma\|_{\mathcal{F}}^2}, \frac{t}{K_0^2 \|\frac{1}{n} D_\Sigma\|}\right)\right).$$

Note that

$$\|\frac{1}{n} D_\Sigma\|_{\mathcal{F}}^2 = \frac{1}{n^2}\|D_\Sigma\|_{\mathcal{F}}^2 = \frac{1}{n}\|\Sigma\|_{\mathcal{F}}^2,$$
$$\|\frac{1}{n} D_\Sigma\| = \frac{1}{n}\|D_\Sigma\| = \frac{1}{n}\|\Sigma\|,$$

which implies the first assertion. For the second claim we first note that

$$\operatorname{trace}(\Sigma Y^{(i)}(X^{(i)})^\top) = (X^{(i)})^\top \Sigma Y^{(i)} = \left[(X^{(i)})^\top, (Y^{(i)})^\top\right] Q \begin{bmatrix} X^{(i)} \\ Y^{(i)} \end{bmatrix},$$
$$Q = \begin{bmatrix} 0 & \Sigma \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{(d+d_y)\times(d+d_y)}.$$

We then consider a block-diagonal matrix $D_Q \in \mathbb{R}^{n(d+d_y)\times n(d+d_y)}$ consisting of n blocks of Q, and $Z = [X^{(1)}; Y^{(1)}; X^{(2)}; Y^{(2)}; \cdots; X^{(n)}; Y^{(n)}] \in \mathbb{R}^{n(d+d_y)}$. Then we can verify that

$$\operatorname{trace}(\Sigma B) = Z^\top (\tfrac{1}{n} D_Q) Z.$$

Application of the Hanson–Wright inequality [193, Theorem 1.1] leads to

$$\mathbb{P}(|\operatorname{trace}(\Sigma B)| > t) \leq 2\exp\left(-c \min\left(\frac{t^2}{K_0^4 \|\frac{1}{n} D_Q\|_{\mathcal{F}}^2}, \frac{t}{K_0^2 \|\frac{1}{n} D_Q\|}\right)\right).$$

We note again that

$$\|\frac{1}{n} D_Q\|_{\mathcal{F}}^2 = \frac{1}{n^2}\|D_Q\|_{\mathcal{F}}^2 = \frac{1}{n}\|\Sigma\|_{\mathcal{F}}^2,$$
$$\|\frac{1}{n} D_Q\| = \frac{1}{n}\|D_Q\| = \frac{1}{n}\|Q\|.$$

and finally end up with

$$\mathbb{P}(|\operatorname{trace}(\Sigma B)| > t) \leq 2\exp\left(-cn \min\left(\frac{t^2}{K_0^4 \|\Sigma\|_{\mathcal{F}}^2}, \frac{t}{K_0^2 \|\Sigma\|}\right)\right).$$

□

By the previous consistency result we obtain the following convergence results.

Lemma 7.1.6. *The empirical loss function $\widehat{F}_n$ is C^3 in κ, and for any $\kappa \in (\kappa_l, \kappa_r)$, there exist constants $C,\ c > 0$ such that for all $\varepsilon > 0$*

$$\mathbb{P}(|\partial_\kappa \widehat{F}_n(\kappa) - (\kappa/\kappa_* - 1)\operatorname{trace}(P_2 D D^\top)| > \varepsilon) \leq C\exp(-nc\min\{\varepsilon, \varepsilon^2\}),$$

and

$$\mathbb{P}(|\partial_\kappa^2 \widehat{F}_n(\kappa) - \operatorname{trace}\left((\frac{\kappa^2}{\kappa_*} - \kappa)P_3 + (\frac{4\kappa}{\kappa_*} - 3\kappa)P_2 + \frac{2}{\kappa_*}P_1\right)| > \varepsilon) \leq C\exp(-cn\min\{\varepsilon, \varepsilon^2\}).$$

Proof. Since $\widehat{F}_n(\kappa) = \frac{1}{n}\sum_{i=1}^n f(\kappa, z_i)$, if we let

$$C_\theta = \frac{1}{n}\sum_{i=1}^n \theta_i\theta_i^\top, \quad B = \frac{1}{n}\sum_{i=1}^n \xi_i\theta_i^\top, \quad C_\xi = \frac{1}{n}\sum_{i=1}^n \xi_i\xi_i^\top,$$

then

$$\begin{aligned}\partial_\kappa \widehat{F}_n(\kappa) &= \operatorname{trace}\left(P_2(\kappa^2 C_v + 2\kappa DB + DC_\xi D^\top) + P_1(2\kappa C_v + 2DB)\right) \\ &= \operatorname{trace}\left((P_2\kappa^2 + 2\kappa P_1)C_v + (2P_1 + 2\kappa P_2)DB + D^\top P_2 D C_\xi\right).\end{aligned} \tag{7.4}$$

We note that

$$\begin{aligned}\mathbb{E}\partial_\kappa \widehat{F}_n(\kappa) &= \operatorname{trace}\left(P_2(\kappa^2/\kappa_* \operatorname{Id} + DD^\top) + 2\kappa P_1/\kappa_*\right) \\ &= \operatorname{trace}\left(-(Q_\kappa^2 C_0 Q_\kappa + Q_\kappa C_0 Q_\kappa^2)(\frac{\kappa^2}{\kappa_*}I + DD^\top) + 2\frac{\kappa}{\kappa_*}Q_\kappa C_0 Q_\kappa\right) \\ &= \operatorname{trace}\left((\frac{\kappa}{\kappa_*} - 1)P_2 DD^\top\right).\end{aligned}$$

Moreover, in (7.4), $\partial_\kappa \widehat{F}_n$ can be written as sum of $\operatorname{trace}(\Sigma_1 C_v)$, $\operatorname{trace}(\Sigma_2 B)$ and $\operatorname{trace}(\Sigma_3 C_\xi)$ for certain matrices Σ such as

$$\Sigma_1 = (P_2\kappa^2 + 2\kappa P_1), \quad \Sigma_2 = (2P_1 + 2\kappa P_2)D, \quad \Sigma_3 = D^\top P_2 D.$$

Note that for any random variables A_k

$$\mathbb{P}\left(|\sum_{k=1}^m (A_k - \mathbb{E}A_k)| > \varepsilon\right) \le \sum_{k=1}^m \mathbb{P}(|A_k - \mathbb{E}A_k| > \varepsilon/m).$$

Therefore we can apply Lemma 7.1.5 at each trace term, and bound its probability of deviating from its mean. Therefore, we can find constants C_1, c such that

$$\mathbb{P}(|\partial_\kappa \widehat{F}_n(\kappa) - (\kappa/\kappa_* - 1)\operatorname{trace}(P_2 DD^\top)| > \varepsilon) \le C_1 \exp\left(-cn\min(\varepsilon^2, \varepsilon)\right).$$

For the second claim,

$$\partial_\kappa^2 f(\kappa, z) = \operatorname{trace}\left(P_3(\kappa^2 vv^\top + 2\kappa D\xi v^\top + D\xi\xi^\top D^\top) + 4P_2(\kappa vv^\top + D\xi v^\top) + 2P_1 vv^\top\right).$$

and

$$\partial_\kappa^2 \widehat{F}_n(\kappa) = \operatorname{trace}\left((\kappa^2 P_3 + 4\kappa P_2 + 2P_1)C_v + (2\kappa P_3 + 4P_2)DB + D^\top P_3 D C_\xi\right). \tag{7.5}$$

Therefore,

$$\mathbb{E}\partial_\kappa^2 \widehat{F}_n(\kappa) = \operatorname{trace}\left((\kappa^2 P_3 + 4\kappa P_2 + 2P_1)/\kappa_* + DD^\top P_3\right).$$

The deviation probability can also be obtained by analyzing matrices

$$\Sigma_1' = (\kappa^2 P_3 + 4\kappa P_2 + 2P_1), \quad \Sigma_2' = (2\kappa P_3 + 4P_2)D, \quad \Sigma_3' = D^\top P_3 D.$$

Note that

$$\operatorname{trace}(Q_\kappa^{-1}P_3) = \operatorname{trace}(2Q_\kappa^2 C_0 Q_\kappa + Q_\kappa^2 C_0 Q_\kappa + Q_\kappa C_0 Q_\kappa^2 + 2Q_\kappa C_0 Q_\kappa^2) = -3P_2.$$

$\operatorname{trace}(\kappa P_2 + 2P_1) = \operatorname{trace}((Q_\kappa^{-1} - \kappa I)Q_\kappa^2 C_0 Q_\kappa + Q_\kappa C_0 Q_\kappa^2 (Q_\kappa^{-1} - \kappa \operatorname{Id})) = -\operatorname{trace}(DD^\top P_2)$.
Thus, the average can also be written as

$$\begin{aligned}
\mathbb{E}\partial_\kappa^2 \widehat{F}_n(\kappa) &= \operatorname{trace}\left((\kappa(\kappa \operatorname{Id} + DD^\top)P_3 + 4\kappa P_2 + 2P_1)/\kappa_* + (1-\kappa/\kappa_*)DD^\top P_3\right) \\
&= \operatorname{trace}\left((-3\kappa P_2 + 4\kappa P_2 + 2P_1)/\kappa_* + (1-\kappa/\kappa_*)DD^\top P_3\right) \\
&= \operatorname{trace}\left(-DD^\top P_2/\kappa_* + (1-\kappa/\kappa_*)DD^\top P_3\right) \\
&= \operatorname{trace}\left(DD^\top((1-\kappa/\kappa_*)P_3 - P_2/\kappa_*)\right).
\end{aligned}$$

Similar, we can obtain

$$\partial_\kappa^3 \widehat{F}_n(\kappa) = \operatorname{trace}\left((\kappa^2 P_4 + 6\kappa P_3 + 6P_2)C_v + (2\kappa P_4 + 6P_2)DB + D^\top P_4 D C_\xi\right) \quad (7.6)$$

and

$$\partial_\kappa^3 \widehat{F}_n(\kappa) = \operatorname{trace}\left((\kappa^2 P_4 + 6\kappa P_3 + 6P_2)/\kappa_* + D^\top P_4 D\right).$$

□

Remark 7.1.7. *It is worthwhile to note that*

$$\partial_\kappa^2 F(\kappa) = \mathbb{E}\partial_\kappa^2 \widehat{F}_n(\kappa) = \operatorname{trace}\left(DD^\top((1-\kappa/\kappa_*)P_3 - P_2/\kappa_*)\right),$$

is not always positive, and it can be negative if κ is very large. In other words, F is not convex on the real line. Therefore, it is necessary to introduce a local parameter domain where F is convex inside.

Remark 7.1.8. *We note that through the definition of f, we can ensure that*

$$\mathbb{E}[\partial_\kappa f(\kappa, z)] = \partial_\kappa \mathbb{E}[f(\kappa, z)].$$

This can be seen, by the following computation of $\mathbb{E}[f(\kappa, z)]$. We can write

$$\mathbb{E}[f(\kappa, z)] = \operatorname{trace}\left(\operatorname{Cov}[C_0^{1/2} Q_\kappa(\kappa v + D\xi), C_0^{1/2} Q_\kappa(\kappa v + D\xi)]\right),$$

and with

$$C_0^{1/2} Q_\kappa(\kappa v + D\xi) \sim \mathcal{N}(0, C_0^{1/2} Q_\kappa(\kappa^2 \operatorname{Id} + DD^\top) Q_\kappa C_0^{1/2})$$

we obtain

$$\begin{aligned}
\mathbb{E}[f(\kappa, z)] &= \operatorname{trace}(C_0^{1/2} Q_\kappa(\kappa^2 \operatorname{Id} + DD^\top) Q_\kappa C_0^{1/2}) \\
&= \operatorname{trace}(P_1(\kappa^2 \operatorname{Id} + DD^\top)).
\end{aligned}$$

Hence, we imply

$$\partial_\kappa \mathbb{E}[f(\kappa, z)] = \operatorname{trace}((\frac{\kappa}{\kappa_*} - 1)P_2 DD^\top) = \mathbb{E}[\partial_\kappa f(\kappa, z)].$$

Consistency analysis within an interval In order to apply Proposition 7.1.2, we also have to show that there exists a local region/interval in which $\widehat{F}_n(\kappa)$ is strongly convex. To do so, we use a chaining argument.
First, we show that the empirical loss function has bounded derivatives with high probability.

Lemma 7.1.9. *There exists an $S > 0$ such that the following holds true*

$$\mathbb{P}\left(\max_{\kappa_l \leq \kappa \leq \kappa_u} |\partial_\kappa^k \widehat{F}_n(\kappa)| > S, k = 1, 2, 3\right) \leq 6 \exp(-nc).$$

Proof. Recall that $\|Q_\kappa\| \leq \frac{1}{\kappa_l}$ and the formulae (7.4), (7.5) and (7.6).
We have

$$\mathbb{P}(\max_{\kappa_l \leq \kappa \leq \kappa_u} |\partial_\kappa^k \widehat{F}_n(\kappa) - \mathbb{E}\widehat{F}_n(\kappa)| > t) \leq 2 \exp\left(-cn \min\left(\frac{t^2}{\|\Sigma_k\|_{\mathcal{F}}^2}, \frac{t}{\|\Sigma_k\|}\right)\right),$$

for each $k = 1, 2, 3$. Here each Σ_k consists of matrices of form $P_j S$ or $S P_j$ where $j = 1, 2, 3, 4$ and $S = \mathrm{Id}, D$ or $DD^\top$. It follows

$$\|P_j S\| \leq \|P_j\| \|S\| \leq \frac{\|C_0\| \|S\|}{\kappa_l^{j+1}}, \quad \|P_j S\|_{\mathcal{F}} \leq \frac{\|C_0\|_{\mathcal{F}} \|S\|}{\kappa_l^{j+1}}.$$

We see that c can depend on $\|C_0\| \leq \mathrm{trace}(C_0)$, $\|C_0\|_{\mathcal{F}}$ and $\|D\|$. Furthermore, $\mathbb{E}\partial_\kappa^k \widehat{F}_n$ is a linear sum of some $\mathrm{trace}(P_j)$ and

$$|\,\mathrm{trace}(P_j)| \leq (\frac{2}{\kappa_l})^{j+1} \,\mathrm{trace}(C_0).$$

Hence, L can also be taken as a constant that depends only on $\mathrm{trace}(C_0)$, $\|C_0\|_{\mathcal{F}}$ and $\|D\|$. This concludes our proof.

□

The next result states, that if a function is bounded at each fixed point with high probability, it implies to be bounded on a fixed intervall with high probability if the function is Lipschitz.

Lemma 7.1.10. *Let $f_n(\kappa)$ be function of κ and assume that the following is true for some interval $[\kappa_l, \kappa_u]$*

$$\mathbb{P}(f_n(\kappa) > a) \leq C \exp(-nc_a) \quad \forall \kappa_l \leq \kappa \leq \kappa_u.$$

Then

$$\mathbb{P}\left(\max_{\kappa \in [\kappa_l, \kappa_u]} f_n(\kappa) > 2a, \max_{\kappa \in [\kappa_l, \kappa_u]} |\partial f_n(\kappa)| \leq M\right) \leq a^{-1} |\kappa_u - \kappa_l| MC \exp(-nc_a).$$

Let $f_n(\kappa)$ be function of κ and the following is true for some interval $[\kappa_l, \kappa_u]$

$$\mathbb{P}(f_n(\kappa) < a) \leq \exp(-nc_a) \quad \forall \kappa_l \leq \kappa \leq \kappa_u.$$

Then

$$\mathbb{P}\left(\min_{\kappa \in [\kappa_l, \kappa_u]} f_n(\kappa) < a/2, \max_{\kappa \in [\kappa_l, \kappa_u]} |\partial f_n(\kappa)| \leq M\right) \leq 2a^{-1} |\kappa_u - \kappa_l| MC \exp(-nc_a).$$

Proof. We pick $\kappa_i = \kappa_l + \frac{2a}{|M|}i$ for $i = 0, \dots, \lfloor \frac{|\kappa_u - \kappa_l|M}{2a} \rfloor$, such that $\kappa_l \le \kappa_i \le \kappa_u$, and for any $\kappa_l \le \kappa \le \kappa_u$, $|\kappa - \kappa_i| \le \frac{a}{M}$ for some κ_i. We note that for $|\partial f_n(\kappa)| \le M$, and $f_n(\kappa_i) \le a$, for all i, it follows for any $\kappa_l \le \kappa \le \kappa_u$,

$$f_n(\kappa) \le f_n(\kappa_i) + (\kappa_i - \kappa)\partial_\kappa f_n(\kappa) \le a + \frac{a}{M}M = 2a.$$

Consequentially, by union bound we obtain

$$\begin{aligned}\mathbb{P}\bigg(\min_{\kappa\in[\kappa_l,\kappa_u]} f_n(\kappa) > 2a, \max_{\kappa\in[\kappa_l,\kappa_u]} |\partial f_n(\kappa)| \le M \bigg) &\le \mathbb{P}\bigg(f_n(\kappa_i) > a \text{ for some } i \bigg) \\ &\le a^{-1}|\kappa_u - \kappa_l| MC \exp(-nc_a).\end{aligned}$$

The same argument can be applied to show the second claim, except that we choose $\kappa_i = c + \frac{a}{|M|}$. □

By the next lemma, we indicate that the loss function is strongly convex within $\mathcal{D}$ with high probability.

Lemma 7.1.11. *Assume that the largest eigenvalue of $DD^\top$ is κ_D and let $\kappa \in \mathcal{D} := [\frac{5}{6}\kappa_*, \frac{7}{6}\kappa_*]$. Then for some constants $c, C > 0$,*

$$\mathbb{P}(\min_{\kappa\in\mathcal{D}} \partial_\kappa^2 \widehat{F}_n(\kappa) < H_*/4) \le \frac{C}{\min\{H_*, 1\}} \exp\bigg(- cn \min(H_*^2, H_*, 1)\bigg),$$

with

$$H_* = H_*(\kappa_D, \kappa_*) = \frac{\kappa_D^2}{(\kappa_D + 2\kappa_*)^2\kappa_* \|L^\top \Gamma^{-1} L\|} > 0.$$

Proof. We denote

$$A = DD^\top((1 - \kappa/\kappa_*)P_3 - P_2/\kappa_*),$$

and let v_i be the eigenvector of $DD^\top$ corresponds to eigenvalue σ_i. Note that v_i is also the eigenvector Q_κ with eigenvalue $(\sigma_i + \kappa)^{-1}$, then

$$v_i^\top A v_i = \bigg(\frac{6(1 - \kappa/\kappa_*)}{(\sigma_i + \kappa)^3} + \frac{2}{(\sigma_i + \kappa)^2\kappa_*} \bigg) \sigma_i v_i^\top C_0 v_i.$$

When $\kappa \in \mathcal{D}$, if $7/6\kappa_* \ge \kappa > \kappa_*$, we have that

$$\frac{6(1 - \kappa/\kappa_*)}{(\sigma_i + \kappa)^3} + \frac{2}{(\sigma_i + \kappa)^2\kappa_*} \ge -\frac{1}{(\sigma_i + \kappa)^2\kappa} + \frac{2}{(\sigma_i + \kappa)^2\kappa_*} \ge \frac{1}{(\sigma_i + \kappa)^2\kappa_*}.$$

If $\kappa \le \kappa_*$, the same relation also holds. Then note that if v_i are all the eigenvectors of $DD^\top$ with eigenvalues σ_i, assuming that σ_i are decreasing,

$$\operatorname{trace} A = \sum_{i=1}^d v_i^\top A v_i \ge \sum_{i=1}^d \frac{\sigma_i v_i^\top C_0 v_i}{(\sigma_i + 2\kappa_*)^2\kappa_*} = \frac{\sigma_1 v_1^\top C_0 v_1}{(\sigma_1 + 2\kappa_*)^2\kappa_*}.$$

Finally, note that

$$\kappa_D = \sigma_1 = v_1^\top DD^\top v_1 = v_1^\top C_0^{1/2} L^\top \Gamma^{-1} L C_0^{1/2} v_1 \le \|L^\top \Gamma^{-1} L\| \|C_0^{1/2} v_1\|^2.$$

It follows

$$\operatorname{trace} A \geq \frac{\kappa_D^2}{(\kappa_D + 2\kappa_*)^2 \kappa_* \|L^\top \Gamma^{-1} L\|} = H_*$$

for $\kappa \in \mathcal{D}$ and we set $\varepsilon = H_*/2 > 0$ to apply Corollary 7.1.6. We obtain some C_1, c

$$\begin{aligned}
&C_1 \exp\left(-cn \min(H_*^2, H_*)\right) \\
&\geq \mathbb{P}\left(|\partial_\kappa^2 \widehat{F}_n(\kappa) - \frac{2}{\kappa_*} \operatorname{trace}\left((3\kappa_* \operatorname{Id} - 2\kappa \operatorname{Id} + DD^\top) DD^\top Q_\kappa^4 \right) | > H_*/2 \right) \\
&\geq \mathbb{P}\left(\partial_\kappa^2 \widehat{F}_n(\kappa) < \frac{2}{\kappa_*} \operatorname{trace}\left((3\kappa_* I - 2\kappa \operatorname{Id} + DD^\top) DD^\top Q_\kappa^4 \right) - H_*/2 \right) \\
&\geq \mathbb{P}(\partial_\kappa^2 \widehat{F}_n(\kappa) < H_*/2).
\end{aligned}$$

By Lemma 7.1.9 there exists an $L > 0$ and c_1 such that

$$\mathbb{P}\left(\max_{\kappa \in \mathcal{D}} |\partial_\kappa^3 \widehat{F}_n(\kappa)| > L \right) \leq 6 \exp(-nc_1),$$

and by Lemma 7.1.10 it holds true that

$$\begin{aligned}
&\frac{C_2}{\min(H_*, 1)} \exp\left(- cn \min(H_*^2, H_*, 1) \right) \\
&\qquad \geq \mathbb{P}\left(\min_{\kappa \in \mathcal{D}} \partial_\kappa^2 \widehat{F}_n(\kappa) < H_*/4, \max_{\kappa \in \mathcal{D}} |\partial_\kappa^3 \widehat{F}_n(\kappa)| \leq M \right),
\end{aligned}$$

for some $C_2 > 0$. We define the sets $A_n := \{\min_{\kappa \in \mathcal{D}} \partial_\kappa^2 \widehat{F}_n(\kappa) < H_*/4\}$ and $B_n := \{\max_{\kappa \in \mathcal{D}} |\partial_\kappa^3 \widehat{F}_n(\kappa)| \leq L\}$, and we obtain

$$\begin{aligned}
\mathbb{P}(\min_{\kappa \in \mathcal{D}} \partial_\kappa^2 \widehat{F}_n(\kappa) < H_*/4) &= \mathbb{P}(A_n \mid B_n)\mathbb{P}(B_n) + \mathbb{P}(A_n \mid B_n^{\complement})\mathbb{P}(B_n^{\complement}) \\
&\leq \mathbb{P}(A_n \cap B_n) + \mathbb{P}(B_n^{\complement}) \\
&\leq \frac{C}{\min(H_*, 1)} \exp\left(- cn \min(H_*^2, H_*, 1) \right).
\end{aligned}$$

□

We state the last lemma before proving Theorem 7.1.3, which indicates that the empirical loss function is unlikely to have local minimums outside $[\frac{2}{3}\kappa_*, \frac{4}{3}\kappa_*]$.

Lemma 7.1.12. *Let κ_D be the largest eigenvalue of $DD^\top$ and let*

$$S_* = \frac{2\kappa_D^2}{3(\kappa_D + \kappa_u)^3 \|L^\top \Gamma^{-1} L\|}.$$

Then there exist constants $c, C > 0$ such that

$$\mathbb{P}\left(\min_{\kappa_u \geq \kappa > \frac{4}{3}\kappa_*} \partial_\kappa \widehat{F}_n(\kappa) < S_*/4 \right) \leq \frac{C}{\min\{S_*, 1\}} \exp\left(-cn \min\left(S_*^2, S_*, 1\right) \right),$$

and

$$\mathbb{P}\left(\min_{\frac{2}{3}\kappa_* \geq \kappa > \kappa_l} \partial_\kappa \widehat{F}_n(\kappa) > -S_*/4 \right) \leq \frac{C}{\min\{S_*, 1\}} \exp\left(-cn \min\left(S_*^2, S_*, 1\right) \right).$$

Proof. Let v be again the leading eigenvector of $DD^\top$, i.e. we have

$$\begin{aligned} -\operatorname{trace}(P_2 DD^\top) \geq v^\top (Q_\kappa^2 C_0 Q_\kappa + Q_\kappa C_0 Q_\kappa^2) v &\geq 2\kappa_D \frac{v^\top C_0 v}{(\kappa_D + \kappa_u)^3} \\ &\geq \frac{2\kappa_D^2}{(\kappa_D + \kappa_u)^3 \|L^\top \Gamma^{-1} L\|} =: 3S_*. \end{aligned}$$

For $\kappa > \frac{4}{3}\kappa_*$ we have

$$(1 - \kappa/\kappa_*) \operatorname{trace}(P_2 DD^\top) \geq S_*.$$

We set $\varepsilon = S_*/2$ and apply Lemma 7.1.6 to obtain

$$\begin{aligned} C \exp\left(-nc \min(S_*^2, S_*)\right) &\geq \mathbb{P}(|\partial_\kappa \widehat{F}_n(\kappa) - (1 - \kappa/\kappa_*) \operatorname{trace}(Q_\kappa^3)| > S_*/2) \\ &\geq \mathbb{P}(\partial_\kappa \widehat{F}_n(\kappa) < (1 - \kappa/\kappa_*) \operatorname{trace}(Q_\kappa^3) - S_*/2) \\ &\geq \mathbb{P}(\partial_\kappa \widehat{F}_n(\kappa) < L_*/2). \end{aligned}$$

Similarly as in Lemma 7.1.11, we apply Lemma 7.1.9 and Lemma 7.1.10 to obtain the first assertion. The second assertion follows by using

$$(1 - \kappa/\kappa_*) \operatorname{trace}(P_2 DD^\top) \leq -L_* < 0,$$

for $\kappa < \frac{2}{3}\kappa_*$. □

Final step We are now ready to prove Theorem 7.1.3.

Proof of Theorem 7.1.3. We denote $\mathcal{D} = [\frac{2}{3}\kappa_*, \frac{4}{3}\kappa_*], H_* = \frac{\kappa_D^2}{(\kappa_D + 2\kappa_*)^2 \kappa_* \|L^\top \Gamma^{-1} L\|} > 0$ and define the events

$$B = \{\kappa_l < \widehat{\kappa}_n < \kappa_u\}, \quad \mathcal{A}_n = \{\widehat{\kappa}_n \in \mathcal{D}, \partial_\kappa^2 \widehat{F}_n(\kappa) \geq \frac{1}{4} H_* \text{ for all } \kappa \in \mathcal{D}\}.$$

First, we decompose

$$\begin{aligned} \mathbb{P}(|\widehat{\kappa}_n - \kappa_*| > \varepsilon, B) &= \mathbb{P}(|\widehat{\kappa}_n - \kappa_*| > \varepsilon, B \mid \mathcal{A}_n) \cdot \mathbb{P}(\mathcal{A}_n) \\ &\quad + \mathbb{P}(|\widehat{\kappa}_n - \kappa_*| > \varepsilon, B \mid \mathcal{A}_n^\complement) \cdot \mathbb{P}(\mathcal{A}_n^\complement) \\ &\leq \mathbb{P}(|\widehat{\kappa}_n - \kappa_*| > \varepsilon, B \mid \mathcal{A}_n) + \mathbb{P}(B \cap \mathcal{A}_n^\complement). \end{aligned}$$

In the last step we have used $\mathbb{P}(\widehat{\kappa}_n \leq \kappa_u) = 1$. By Proposition 7.1.2

$$\begin{aligned} \mathbb{P}(|\widehat{\kappa}_n - \kappa_*| > \varepsilon, B \mid \mathcal{A}_n) &\leq \mathbb{P}\left(|\partial_\kappa \widehat{F}_n(\kappa_*) - \partial_\kappa F(\kappa_*)| > \frac{1}{4} H_* \epsilon, B\right) \\ &= \mathbb{P}\left(|\partial_\kappa \widehat{F}_n(\kappa_*)| > \frac{1}{4} H_* \epsilon, B\right), \end{aligned}$$

which we can bound by Lemma 7.1.6 and Lemma 7.1.10

$$\mathbb{P}(|\widehat{\kappa}_n - \kappa_*| > \varepsilon \mid \mathcal{A}_n) \leq C_1 \exp(-nc_1 \min\{\epsilon, \epsilon^2\}),$$

for some $C_1, c_1 > 0$.

We bound the probability $\mathbb{P}(\mathcal{A}_n^{\complement})$ by

$$\mathbb{P}(B \cap \mathcal{A}_n^{\complement}) \leq \mathbb{P}(B, \widehat{\kappa}_n \notin \mathcal{D}) + \mathbb{P}(\{\partial_\kappa^2 \widehat{F}_n(\kappa) \geq H_*/4 \text{ for all } \kappa \in \mathcal{D}\}^{\complement}),$$

and study both terms separately. We first note, by Lemma 7.1.12, for some constants $C_2, c_2 > 0$ the following holds

$$\begin{aligned} \mathbb{P}(B, \widehat{\kappa}_n \notin \mathcal{D}) \leq & \ \mathbb{P}(\partial_\kappa \widehat{F}_n(\kappa) = 0 \text{ for some } \kappa \in (\kappa_l, \kappa_u) \setminus \mathcal{D}) \\ \leq & \ C_2 \exp(-c_2 n). \end{aligned}$$

Second, by Lemma 7.1.11, we imply that for some constants $C_3, c_3 > 0$ we obtain

$$\mathbb{P}(\{\partial_\kappa^2 \widehat{F}_n(\kappa) \geq H_*/4 \text{ for all } \kappa \in \mathcal{D}\}^{\complement}) \leq C_3 \exp(-\, c_3 n).$$

Finally, we conclude that there exist some constants $C_*, c_* > 0$ such that

$$\mathbb{P}(|\widehat{\kappa}_n - \kappa_*| > \epsilon) \leq C_* \exp(-c_* n \min(\epsilon, \epsilon^2)).$$

□

7.1.2 Regularization parameter online recovery

In this part, we consider the implementation of the stochastic gradient descent (SGD) method in order to solve the bilevel optimization online. We formulate the SGD method for general nonlinear inverse problems and state certain assumptions on the forward model and the corresponding regularization function to ensure convergence of the proposed method.

Bilevel stochastic gradient descent method

Solving the bilevel optimization probelm (7.3) online one needs to compute the empirical loss function $\widehat{F}_n$ and its gradient in (7.2). To do so, one has to solve the lower level problem for each training data point $(\theta^{(j)}, y^{(j)})$, $j = 1, \ldots, n$. For very large n this can be computationally very demanding. One promising way to alleviate this is the application of the SGD method. In the context of traditional bilevel optimization various convergence results has been shown [57]. As a result this has been applied to problems in machine learning, most notably support vector machines [55, 56], but also in a more general context without the use of SGD [82, 115]. In our setting, we propose a SGD method to solve the bilevel optimization problem (7.1) online.
We first note that the general gradient descent method for a function $F(\kappa)$ generates iterates κ_{k+1} based on the following update rules

$$\kappa_{k+1} = \kappa_k - \beta_k \nabla_\kappa F(\kappa_k),$$

where β_k is a sequence of stepsizes.
As mentioned above, the population gradient $\nabla_\kappa F$ is often computationally inaccessible, and its empirical approximation $\nabla_\kappa \widehat{F}_n$ is often expensive to compute. One general solution to this issue is using a stochastic approximation of $\nabla_\kappa F$. Here we choose $\nabla_\kappa f(\kappa_k, Z^{(k)})$, as it is an unbiased estimator of $\nabla_\kappa F$:

$$\nabla_\kappa F(\kappa_k) = \mathbb{E}_Z \nabla_\kappa f(\kappa_k, Z).$$

The identity above holds by Fubini's theorem, since we assume f and its second order derivatives are all continuous and differentiable. Comparing with $\nabla_\kappa \widehat{F}_n$, $\nabla_\kappa f$ involves only one data point $Z^{(k)}$, so it has a significantly smaller computation cost. We refer to this method as "online", since it does not require all n data points available at each algorithmic iteration.
In the following we formulate the stochastic gradient descent method to solve (7.1) as Algorithm 10.

Algorithm 10: Bilevel Stochastic Gradient Descent

Input: κ_0, m, $\beta = (\beta_k)_{k=1}^n$, $\beta_k > 0$, i.i.d. sample $(Z^{(k)})_{k\in\{1,\dots,n\}} \sim \mu_{(\Theta,Y)}$.
for $k = 0, \dots, n-1$ **do**

$$\kappa_{k+1} = \chi(\kappa_k - \beta_k \nabla_\kappa f(\kappa_k, Z^{(k)})), \tag{7.7}$$

Output: the average $\bar{\kappa}_n = \frac{1}{m}\sum_{k=n-m+1}^n \kappa_k$

In Algorithm 10, the step size β_k is a sequence which decreases to zero, but not too fast, such that the Robbins–Monro conditions [189] apply:

$$\sum_{k=1}^\infty \beta_k = \infty, \quad \sum_{k=1}^\infty \beta_k^2 < \infty. \tag{7.8}$$

One standard choice is to take a decreasing step size $\beta_k = \beta_0 k^{-\alpha}$ with $\alpha \in (1/2, 1]$. We note that the output of our bilevel SGD method is given by the average over the last iterations $\bar{\kappa}_n$, which has been shown to accelerate the scheme for standard SGD methods, see [179]. The projection map χ ([210], Section 14.4.1) is defined as

$$\chi(\kappa) = \arg\min_{x\in\Lambda}\{\|x - \kappa\|\}.$$

This means, the projection maps κ to itself in case $\kappa \in \Lambda$, and otherwise it outputs the closest point in Λ to κ. This projection ensures that κ_{k+1} stays in the range of the regularization parameter if Λ is closed. This operation in general only shorten the distance between κ_{k+1} and κ_* when Λ is convex, as the following Lemma states.

Lemma 7.1.13 (Lemma 14.9 of [210])**.** *If Λ is convex, then for any κ*

$$\|\chi(\kappa) - \kappa_*\| \le \|\kappa - \kappa_*\|.$$

In particular, the stochastic gradient $\nabla_\kappa f(\kappa_k, Z^{(k)})$ is given by the following lemma, which states sufficient conditions on Ψ to ensure both θ_κ and f are continuously differentiable w.r.t. κ.

Lemma 7.1.14. *Suppose the lower level loss function $\Psi(\kappa, \theta, y)$ is C^2 and strictly convex for (θ, κ) in a neighborhood of $(\theta_{\kappa_0}, \kappa_0)$, then the function $\kappa \mapsto \theta_\kappa(y)$ is continuously differentiable w.r.t. κ near κ_0 and the derivative is given by*

$$\nabla_\kappa \theta_\kappa(y) = -\left(\nabla_\theta^2 [\Psi(\kappa, \theta_\kappa(y), y)]\right)^{-1} \nabla_{\kappa\theta}^2 [\Psi(\kappa, \theta_\kappa(y), y)]$$

and

$$\nabla_\kappa f(\kappa, y, \theta) = \nabla_w \mathcal{L}_\mathcal{X}(\theta_\kappa(y), \theta)^\top \nabla_\kappa \theta_\kappa(y). \tag{7.9}$$

Proof. The proof of this statement is based on the implicit function theorem. For fixed $y \in \mathbb{R}^K$, we define the function

$$\varphi(\kappa, \theta) := \nabla_\theta \Psi(\kappa, \theta, y).$$

Since $(\kappa, \theta) \mapsto \Psi(\kappa, \theta, y)$ is strictly convex, we have that for all (κ, θ) near $(\kappa_0, \theta_{\kappa_0})$ the Jacobian of φ w.r.t. θ is invertible, i.e.

$$D_\theta \varphi(\kappa, \theta) = \nabla_\theta^2 \Psi(\kappa, \theta, y) > 0.$$

Set $\bar{\kappa} \in \mathbb{R}^d$ arbitrary, then for $\bar{\theta} = \theta_{\bar{\kappa}}(y)$ it holds true that

$$\varphi(\bar{\kappa}, \bar{\theta}) = 0$$

and by the implicit function theorem there exists an open neighborhood $\mathcal{D} \subset \mathbb{R}^d$ of κ_0 with $\bar{\kappa} \in \mathcal{D}$ such that there exists a unique continuously differentiable function $\bar{\Theta} : \mathcal{D} \to \mathbb{R}^d$ with $\bar{\Theta}(\bar{\kappa}) = \bar{\theta}$ and

$$\varphi(\kappa, \bar{\Theta}(\kappa)) = 0,$$

for all $\kappa \in \Lambda$, i.e. $\bar{\Theta}$ maps all $\kappa \in \Lambda$ to the corresponding regularized solution $\bar{\Theta}(\kappa) = \theta_\kappa(y)$. Further, the partial derivatives of $\bar{\Theta}$ are given by

$$\frac{\partial \bar{\Theta}}{\partial \kappa_i}(\kappa) = -\left[D_\theta \varphi(\kappa, \bar{\Theta}(\kappa))\right]^{-1} \left[\frac{\partial \varphi}{\partial \kappa_i}(\kappa, \bar{\Theta}(\kappa))\right].$$

Since the choice of $\bar{\kappa} \in \mathbb{R}^d$ is arbitrary, it follows that $\kappa \mapsto \theta_\kappa(y)$ is continuously differentiable with derivative given by

$$\nabla_\kappa \theta_\kappa(y) = -\left(\nabla_\theta^2 \left[\Psi(\kappa, \theta_\kappa(y), y)\right]\right)^{-1} \nabla_{\kappa\theta}^2 \left[\Psi(\kappa, \theta_\kappa(y), y)\right].$$

The computation of $\nabla_\kappa f$ can be obtained by the chain rule.

□

Approximate stochastic gradient method

For the implementation of Algorithm 10, it is necessary to evaluate the gradient $\nabla_\kappa f$. While Lemma 7.1.14 provides a formula to compute the gradient, its evaluation can be expensive for complicated PDE forward models. In these scenarios, it is more reasonable to implement approximate SGD.
We consider the approximate gradient by central finite difference schemes. This involves perturbing certain coordinates in opposite direction, and use the value difference to approximate the gradient:

$$(\widetilde{\nabla}_\kappa f(\kappa_k, z))_i \approx \frac{f(\kappa_k + h_k e_i, z) - f(\kappa_k - h_k e_i, z)}{2h_k}, \tag{7.10}$$

where e_i is the i-th Euclidean basis vector and h_k is a step size. h_k can either be fixed as a small constant, or it can be decaying as k increases, so that higher accuracy gradients are used when the iterates are converging.
In many cases, the higher level optimization uses a L_2 loss function

$$\mathcal{L}_{\mathcal{X}}(y, \theta) = \|y - \theta\|^2.$$

In this case, the exact SGD update step (7.7) can be written as

$$\begin{aligned}\kappa_{k+1} &= \kappa_k - \beta_k \nabla_\kappa \|\theta_{\kappa_k}(y^{(k)}) - \theta^{(k)}\|^2 \\ &= \kappa_k - \beta_k \left(\nabla_\kappa \theta_{\kappa_k}(y^{(k)})\right)^\top (\theta_{\kappa_k}(y^{(k)}) - \theta^{(k)}).\end{aligned}$$

In this case, it makes more sense to apply central difference scheme only on the $\nabla\theta_\kappa$ part:

$$(\nabla_\kappa \theta_\kappa(y^{(k)}))_i = \frac{\partial \theta_\kappa(y^{(k)})}{\partial \kappa_i} \approx \frac{\theta_{\kappa+h_k e_i}(y^{(k)}) - \theta_{\kappa-h_k e_i}(y^{(k)})}{2h_k} =: (\widetilde{\nabla}_\kappa \theta_\kappa(y^{(k)}))_i, \tag{7.11}$$

where $(e_i)_{i=1,\dots,d}$ denote the i–th unit vectors in $\mathbb{R}^d$. Using this approximation, we formulate the approximate SGD method in the following algorithm, where we replace the exact gradient $\nabla_\kappa \theta_\kappa(y^{(k)})$ by the numerical approximation $\widetilde{\nabla}_\kappa \theta_\kappa(y^{(k)})$ defined in (7.11).
Here we have defined the numerical approximation of $\nabla_\kappa f$ by

$$\widetilde{\nabla}_\kappa f(\kappa, (y, \theta)) := \left(\widetilde{\nabla}_\kappa \theta_\kappa(y)\right)^\top (\theta_\kappa(y) - \theta). \tag{7.12}$$

In most finite difference approximation schemes, the approximation error involved is often controlled by h_k. In particular, we assume the centred forward difference scheme used in either (7.10) or (7.12) yields an error of order

$$\|\mathbb{E}\widetilde{\nabla}_\kappa(f(\kappa, Y, \Theta)) - \nabla_\kappa F(\kappa)\| =: \alpha_k = O(h_k^2).$$

Replacing the stochastic gradient in Algorithm 10 with its approximation, we obtain the algorithm below:

Algorithm 11: Approximate Bilevel Stochastic Gradient Descent

Input: κ_0, m, $\beta = (\beta_k)_{k=1}^n$, $\beta_k > 0$, i.i.d. sample $(Z^{(k)})_{k\in\{1,\dots,n\}} \sim \mu_{(\Theta,Y)}$.
for $k = 0, \dots, n-1$ **do**

$$\kappa_{k+1} = \chi(\kappa_k - \beta_k \widetilde{\nabla}_\kappa f(\kappa_k, Z^{(k)})),$$

Output: the average $\bar{\kappa}_n = \frac{1}{m}\sum_{k=n-m+1}^n \kappa_k$

Consistency analysis for online estimators

In the following part we present sufficient conditions which ensure that κ_k converges in L^2 to the optimal solution κ_* of (7.1).

Proposition 7.1.15. *Suppose that there is a convex region $\mathcal{D} \subset \Lambda$ and a constant $c > 0$ such that*

$$\inf_{\kappa\in\mathcal{D}} (\kappa - \kappa_*)^\top \nabla_\kappa F(\kappa) > c\|\kappa - \kappa_*\|^2. \tag{7.13}$$

and there are constants $a,\ b > 0$ such that for all $\kappa \in \mathcal{D}$ it holds true that

$$\mathbb{E}[|\widetilde{\nabla}_\kappa f(\kappa, Z)|^2] < a + b\|\kappa - \kappa_*\|^2. \tag{7.14}$$

Also the bias in the approximated SGD is bounded by

$$\|\mathbb{E}\widetilde{\nabla}_\kappa f(\kappa_k, Z_k) - \nabla_\kappa F(\kappa_k)\|^2 \le \alpha_k. \tag{7.15}$$

Let $\mathcal{A}_k$ be the event that $\kappa_k \in \mathcal{D}$. Suppose $\beta_0 \leq \frac{c}{b}$. Then if the approximation error is bounded by a small constant $\alpha_k \leq \alpha_0$, there is a constant C_n such that

$$\mathbb{E}\mathbb{1}_{\mathcal{A}_n}\|\kappa_n - \kappa_*\|^2 \leq \left(\mathbb{E}Q_0 + 2a\sum_{j=1}^{\infty}\beta_j^2\right)C_n + \frac{\alpha_0}{c^2}.$$

Here

$$C_n = \min_{k\leq n}\max\left\{\prod_{j=k+1}^{n}(1-c\beta_j), a\beta_k/c\right\}$$

is a sequence converging to zero.
If the approximation error is decaying so that $\alpha_k \leq D\beta_k$, then we have the estimation error

$$\mathbb{E}\mathbb{1}_{\mathcal{A}_n}\|\kappa_n - \kappa_*\|^2 \leq \left(\mathbb{E}Q_0 + 2(a + D/c)\sum_{j=1}^{\infty}\beta_j^2\right)C_n.$$

Remark 7.1.16. *We note that the above result also leads to similar convergence of the average estimator $\bar{\kappa}_n$ since by Jensen's inequality*

$$\|\bar{\kappa}_n - \kappa_*\|^2 \leq \frac{1}{m}\sum_{k=n-m+1}^{n}\|\kappa_k - \kappa_*\|^2.$$

Further, for standard SGD methods the averaging step has been shown to lead to the highest possible convergence rate under suitable assumptions. We refer interested readers to [179] for more details.

Proof of Proposition 7.1.15. We note that

$$\Delta_{k+1} = \chi(\kappa_k - \beta_k\widetilde{\nabla}_\kappa f(\kappa_k, Z_k)) - \kappa_*,$$

and apply Lemma 7.1.13

$$\begin{aligned}\|\Delta_{k+1}\|^2 = \|\chi\left(\kappa_k - \beta_k\widetilde{\nabla}_\kappa f(\kappa_k, Z_k)\right) - \kappa_*\|^2 &\leq \|\kappa_k - \beta_k\widetilde{\nabla}_\kappa f(\kappa_k, Z_k) - \kappa_*\|^2 \\ &= \|\Delta_k - \beta_k\widetilde{\nabla}_\kappa f(\kappa_k, Z_k)\|^2 \\ &= \|\Delta_k - \beta_k\nabla_\kappa F(\kappa_k, Z_k) - \beta_k\delta_k - \beta_k\xi_k\|^2,\end{aligned}$$

where we have defined the bias and noise in the stochastic gradient by

$$\delta_k = \mathbb{E}_k\widetilde{\nabla}f(\kappa_k, Z) - \nabla F(\kappa_k), \quad \xi_k = \widetilde{\nabla}_\kappa f(\kappa_k, Z_k) - \mathbb{E}_k\widetilde{\nabla}_\kappa f(\kappa_k, Z).$$

Further, we denote the expectation conditioned on information available at step k as $\mathbb{E}_k$ and define the first exit time of $\mathcal{D}$ by with $\tau = \inf\{k \geq 0 \mid \kappa_k \in \mathcal{D}\}$. Next, we note that

$$\mathbb{E}_k\|\nabla f(\kappa_k, Z_k)\|^2 = \|\nabla_\kappa F(\kappa_k) + \delta_k\|^2 + \mathbb{E}_k\|\xi_k\|^2.$$

So, conditioned on $\tau \geq k$, we have

$$\begin{aligned}\mathbb{E}_k\|\Delta_{k+1}\|^2 &\leq \|\Delta_k\|^2 - 2\beta_k\Delta_k^T(\nabla_\kappa F(\kappa_k) + \delta_k) + \beta_k^2\|\nabla_\kappa F(\kappa_k) + \delta_k\|^2 + \mathbb{E}_k\|\xi_k\|^2 \\ &\leq \|\Delta_k\|^2 - 2\beta_k\Delta_k^T\nabla_\kappa F(\kappa_k) + 2\beta_k\|\Delta_k\|\|\delta_k\| + \beta_k^2(a + b\|\Delta_k\|^2)\end{aligned}$$

$$\leq \|\Delta_k\|^2 - 2c\beta_k\|\Delta_k\|^2 + \frac{1}{2}c\beta_k\|\Delta_k\|^2 + \frac{2}{c}\beta_k\|\delta_k\|^2 + \beta_k^2 a + b\beta_k^2\|\Delta_k\|^2$$
$$\leq (1 - 1.5c\beta_k + b\beta_k^2)\|\Delta_k\|^2 + (a\beta_k + 2\alpha_k/c)\beta_k.$$

Since $\beta_k < c/2b$, we have

$$\mathbb{E}_k \mathbb{1}_{\tau \geq k+1}\|\Delta_{k+1}\|^2 \leq \mathbb{E}_k \mathbb{1}_{\tau \geq k}\|\Delta_{k+1}\|^2 \leq \mathbb{1}_{\tau \geq k}(1 - c\beta_k)\|\Delta_k\|^2 + (a\beta_k + 2\alpha_k/c)\beta_k.$$

Let $Q_k = \mathbb{1}_{\tau_k \geq k}\|\Delta_k\|^2$, then we have just derived that

$$\mathbb{E}Q_{k+1} \leq (1 - c\beta_k)\mathbb{E}Q_k + (a\beta_k + 2\alpha_k/c)\beta_k.$$

Therefore by application of Gronwall's inequality

$$\mathbb{E}Q_n \leq a\sum_{k=1}^{n}\left(\prod_{j=k+1}^{n}(1 - c\beta_j)\beta_k^2\right) + \frac{2}{c}\sum_{k=1}^{n}\left(\prod_{j=k+1}^{n}(1 - c\beta_j)\beta_k\alpha_k\right) + \exp\left(-c\sum_{j=1}^{n}\beta_j\right)\mathbb{E}Q_0. \tag{7.16}$$

Next we look at the 2nd term of (7.16). Note that when $\alpha_k \leq \alpha_0$, then

$$\frac{2}{c}\sum_{k=1}^{n}\prod_{j=k+1}^{n}(1 - c\beta_j)\beta_k\alpha_k \leq \frac{\alpha_0}{c^2}\sum_{k=1}^{n}\prod_{j=k+1}^{n}(1 - c\beta_j)c\beta_k$$
$$\leq \frac{\alpha_0}{c^2}\sum_{k=1}^{n}\left(\prod_{j=k+1}^{n}(1 - c\beta_j) - \prod_{j=k}^{n}(1 - c\beta_j)\right) \leq \frac{\alpha_0}{c^2}.$$

In this case, (7.16) becomes

$$\mathbb{E}Q_n \leq a\sum_{k=1}^{n}\left(\prod_{j=k+1}^{n}(1 - c\beta_j)\beta_k^2\right) + \frac{\alpha_0}{c^2} + \exp\left(-c\sum_{j=1}^{n}\beta_j\right)\mathbb{E}Q_0.$$

And if $\alpha_k \leq D\beta_k$, then (7.16) can be simplified as

$$\mathbb{E}Q_n \leq (a + D/c)\sum_{k=1}^{n}\left(\prod_{j=k+1}^{n}(1 - c\beta_j)\beta_k^2\right) + \exp\left(-c\sum_{j=1}^{n}\beta_j\right)\mathbb{E}Q_0.$$

In both cases, to show our claim, we just need to show

$$\sum_{k=1}^{n}\left(\prod_{j=k+1}^{n}(1 - c\beta_j)\beta_k^2\right) \leq 2C_n, \quad \exp\left(-c\sum_{j=1}^{n}\beta_j\right) \leq C_n.$$

Let k_0 be the minimizer of

$$k_0 = \arg\min_{k \leq n}\max\{\prod_{j=k+1}^{n}(1 - c\beta_j), a\beta_k/c\}$$

Then note that,

$$\sum_{k=1}^{k_0}\prod_{j=k+1}^{n}(1-c\beta_j)\beta_k^2 \le \sum_{k=1}^{k_0}\prod_{j=k_0+1}^{n}(1-c\beta_j)\beta_k^2 \le \prod_{j=k_0+1}^{n}(1-c\beta_j)\sum_{k=1}^{\infty}\beta_k^2 \le C_n.$$

and also

$$\begin{aligned}\sum_{k=k_0+1}^{n}\prod_{j=k+1}^{n}(1-c\beta_j)\beta_k^2 &\le \frac{1}{c}\beta_{k_0}\sum_{k=1}^{k_0}\prod_{j=k+1}^{n}(1-c\beta_j)c\beta_k \\ &\le \frac{1}{c}\beta_{k_0}\sum_{k=1}^{k_0}\left(\prod_{j=k+1}^{n}(1-c\beta_j)-\prod_{j=k}^{n}(1-c\beta_j)\right) \\ &\le \frac{1}{c}\beta_{k_0} = C_n.\end{aligned}$$

The sum of the previous two inequalities leads to

$$\sum_{k=1}^{n}\left(\prod_{j=k+1}^{n}(1-c\beta_j)\beta_k^2\right) \le 2C_n.$$

Finally, we obtain

$$\exp(-c\sum_{j=1}^{n}\beta_j)\mathbb{E}Q_0 \le \exp(-c\sum_{j=k_0+1}^{n}\beta_j)\mathbb{E}Q_0 \le C_n.$$

To see that C_n converges to zero, simply let

$$k_n = \max_k\left\{\prod_{j=1}^{k}(1-c\beta_j) > \sqrt{\prod_{j=1}^{n}(1-c\beta_j)}\right\}$$

Because $\prod_{j=1}^{n}(1-c\beta_j)$ decays to zero when $n\to\infty$, so k_n will increases to ∞, and β_{k_n} will decay to zero. Meanwhile,

$$C_n \le \min\left\{\prod_{j=k+1}^{n}(1-c\beta_j), \beta_{k_n}\right\} \le \min\left\{\sqrt{\prod_{j=1}^{n}(1-c\beta_j)}, \beta_{k_n}\right\},$$

which will decay to zero when $n\to\infty$. □

Consistency analysis with linear inverse problem

We consider again the linear inverse problem from Section 7.1.1

$$y = L\theta + \xi,$$

with Gaussian prior information $\theta\sim\mathcal{N}(0,\frac{1}{\kappa_*}C_0)$ and Gaussian noise $\xi\sim\mathcal{N}(0,\Gamma)$, and the corresponding bilevel optimization with least squares data misfit and Tikhonov regularization, i.e.

$$\mathcal{L}_{\mathcal{Y}}(L\theta,y) = \frac{1}{2}\|L\theta-y\|_{\Gamma}^2, \quad \kappa\varphi(\theta) = \frac{\kappa}{2}\|\theta\|_{C_0}^2.$$

Theorem 7.1.17. *Let $\beta = (\beta_k)_{k\in\mathbb{N}}$ be a sequence of step sizes with $\beta_k > 0$, $\sum_{k=1}^{\infty} \beta_k = \infty$, and $\sum_{k=1}^{\infty} \beta_k^2 < \infty$. Then for some constant B and a sequence C_n converging to zero, the following hold*

1. *the iterates generated from the exact SGD, Algorithm 10, converge to κ_* in the sense*

$$\mathbb{E}\|\kappa_n - \kappa_*\|^2 \leq BC_n,$$

2. *the iterates generated from the aproximate SGD, Algorithm 11 with formula* (7.12) *and $h_k = h$, converge to κ_* up to an error of order $\mathcal{O}(h^4)$, i.e.*

$$\mathbb{E}\|\kappa_n - \kappa_*\|^2 \leq B(C_n + h^4).$$

If we use decaying finite difference stepsize $h_k \leq h\beta_k^{1/4}$, then the error can be further bounded by

$$\mathbb{E}\|\kappa_n - \kappa_*\|^2 \leq BC_n.$$

Remark 7.1.18. *While in the offline setting the proof of the consistency result for the linear Gaussian setting was heavily relying on the Gaussian assumption on θ and ξ, in the online setting we are able to extend the result to non Gaussian distributions of θ and ξ. For our proof of Theorem 7.1.17 we only need to assume that $\mathbb{E}[|\theta|^4] < \infty$ and $\mathbb{E}[|L^\top\Gamma^{-1}\xi|^4] < \infty$. Hence, it can also be applied to general linear inverse problems without Gaussian assumption on the unknown parameter or Gaussian assumption on the noise.*

Proof of Theorem 7.1.17. We denote $\mathcal{D} = \Lambda = [\kappa_l, \kappa_u]$ and observe that, because χ always bring κ_k back into $\mathcal{D}$, the event $\mathcal{A}$ always happen.
Recall that

$$\partial_\kappa f(\kappa, z) = \operatorname{trace}\Big(P_2(\kappa^2 vv^\top + 2\kappa D\xi v^\top + D\xi\xi^\top D^\top) + P_1(2\kappa vv^\top + 2D\xi v^\top)\Big).$$

$$\partial_\kappa^2 f(\kappa, z) = \operatorname{trace}\Big(P_3(\kappa^2 vv^\top + 2\kappa D\xi v^\top + D\xi\xi^\top D^\top) + 4P_2(\kappa vv^\top + D\xi v^\top) + 2P_1 vv^\top\Big).$$

$$\partial_\kappa^3 f(\kappa, z) = \operatorname{trace}\Big(P_4(\kappa^2 vv^\top + 2\kappa D\xi v^\top + D\xi\xi^\top D^\top) + 6P_3(\kappa vv^\top + D\xi v^\top) + 6P_2 vv^\top\Big).$$

We observed in the proof of Lemma 7.1.12, that

$$-\nabla_\kappa F(\kappa) = (\kappa/\kappa_* - 1)\operatorname{trace}(P_2 DD^\top).$$

Next, we multiply with $(\kappa - \kappa_*)$, which gives

$$-(\kappa - \kappa_*)\partial_\kappa F(\kappa) = (\kappa - \kappa_*)^2 \operatorname{trace}(P_2 D^T D/\kappa_*).$$

For v being the eigenvector of $DD^\top$ with maximum eigenvalue κ_D we have that

$$\operatorname{trace}(P_2 DD^\top/\kappa_*) \geq v^\top P_2 DD^\top v/\kappa_* = \frac{\kappa_D v^\top C_0 v}{\kappa_*(\kappa_D + \kappa)^3} \geq \frac{\kappa_D^2}{\kappa_*(\kappa_D + \kappa_u)^3\|L^\top\Gamma^{-1}L\|}.$$

So we define c by

$$c = \frac{2\kappa_D^2}{\kappa_*(\kappa_D + \kappa_u)^3 \|L^\top \Gamma^{-1} L\|},$$

and (7.13) is verified.
By Taylor's theorem, there are some w_k, w_k' between $\kappa_k - h_k$ and $\kappa_k + h_k$ such that

$$|\widetilde{\partial}_\kappa f(\kappa_k, Z) - \partial_\kappa f(\kappa_k, Z)| = \frac{1}{6} h_k^2 |\partial_{\kappa^3}^3 f(w_k, Z) + \partial_{\kappa^3}^3 f(w_k', Z)|.$$

Taking the expectation gives

$$\mathbb{E}|\widetilde{\partial}_\kappa f(\kappa_k, Z) - \partial_\kappa f(\kappa_k, Z)|^2 = \frac{1}{18} h_k^4 (\mathbb{E}|\partial_{\kappa^3}^3 f(w_k, Z)|^2 + \mathbb{E}|\partial_{\kappa^3}^3 f(w_k', Z)|^2).$$

We show that there is a constant B_κ that may depend on κ such that

$$\mathbb{E}|\partial_{\kappa^3}^3 f(w_k, Z)|^2 \leq B_\kappa, \quad \text{and} \quad \mathbb{E}|\partial_{\kappa^3}^3 f(w_k', Z)|^2 \leq B_\kappa. \tag{7.17}$$

This comes from the fact that each component of $\partial_{\kappa^3}^3 f(w_k, Z)$ can be written as $\operatorname{trace}(\Sigma C_v)$ or $\operatorname{trace}(\Psi B)$ or $\operatorname{trace}(\Sigma C_\xi)$, with some Σ and Ψ. Here, we introduce

$$C_v = vv^\top, \quad B = \xi v^\top, \quad C_\xi = \xi\xi^\top.$$

We apply Lemma 7.1.10 with $n = 1$, which shows that for some universal constant C

$$\mathbb{E}(\Sigma C_v) \leq 2|\operatorname{trace}(\Sigma)|^2 + C(\|\Sigma\|_{\mathcal{F}}^3 + \|\Sigma\|^3).$$
$$\mathbb{E}(\Sigma C_\xi) \leq 2|\operatorname{trace}(\Sigma)|^2 + C(\|\Sigma\|_{\mathcal{F}}^3 + \|\Sigma\|^3).$$
$$\mathbb{E}(\Psi B) \leq C(\|\Psi\|_{\mathcal{F}}^3 + \|\Psi\|^3).$$

The matrices Σ in $\partial_{\kappa^3}^3 f(w_k, Z)$ are of form P_j or $D^\top P_j D$, which we have seen to have bounded operator, trace and Frobenius norms. This comes from the proof of Proposition 7.1.15. Furthermore, the matrix Ψ is of form $P_j D$ and we have $\|P_j D\|_{\mathcal{F}} \leq \|P_j\|_{\mathcal{F}} \|D\|$, $\|P_j D\| \leq \|P_j\| \|D\|$. We can conclude that there is a constant B, such that (7.17) holds.
Further, we verify (7.15) by Jensen's inequality

$$\begin{aligned}|\mathbb{E}\widetilde{\partial}_\kappa f(\kappa_k, Z) - \partial_\kappa F(\kappa_k)|^2 &= |\mathbb{E}\widetilde{\partial}_\kappa f(\kappa_k, Z) - \mathbb{E}\partial_\kappa f(\kappa_k, Z)|^2 \\ &\leq \mathbb{E}|\widetilde{\partial}_\kappa f(\kappa_k, Z) - \partial_\kappa f(\kappa_k, Z)|^2 \leq \frac{1}{9} h_k^4 B_\kappa.\end{aligned}$$

Finally, because $\kappa \in [\kappa_l, \kappa_u]$, so B_κ can be bounded as well.
In order to prove that (7.14) is satisfied, we note that by Young's inequality

$$\mathbb{E}|\widetilde{\partial}_\kappa f(\kappa_k, Z)|^2 \leq 2\mathbb{E}|\widetilde{\partial}_\kappa f(\kappa_k, Z) - \partial_\kappa f(\kappa_k, Z)|^2 + 2\mathbb{E}|\partial_\kappa f(\kappa_k, Z)|^2.$$

As we can bound $\mathbb{E}|\widetilde{\partial}_\kappa f(\kappa_k, Z) - \partial_\kappa f(\kappa_k, Z)|^2$, it is sufficient to bound $\mathbb{E}|\partial_\kappa f(\kappa_k, Z)|^2$ by a constant A_κ. Again each component of $\partial_\kappa f(w_k, Z)$ can be written as $\operatorname{trace}(\Sigma C_v)$ or $\operatorname{trace}(\Psi B)$ or $\operatorname{trace}(\Sigma C_\xi)$, with some Σ and Ψ. We apply Lemma 7.1.10 again with $n = 1$ to show that for some universal constant C

$$\mathbb{E}(\Sigma C_v) \leq 2|\operatorname{trace}(\Sigma)|^2 + C(\|\Sigma\|_{\mathcal{F}}^3 + \|\Sigma\|^3).$$
$$\mathbb{E}(\Sigma C_\xi) \leq 2|\operatorname{trace}(\Sigma)|^2 + C(\|\Sigma\|_{\mathcal{F}}^3 + \|\Sigma\|^3).$$
$$\mathbb{E}(\Psi B) \leq C(\|\Psi\|_{\mathcal{F}}^3 + \|\Psi\|^3).$$

Similar as before, we can conclude that there is a dimension free constant A_κ, such that $\mathbb{E}|\partial_\kappa f(\kappa_k, Z)|^2 \leq A_\kappa$. Finally, we conclude the proof as $\kappa \in [\kappa_l, \kappa_u]$. □

7.1.3 Numerical results

In the following part we will test the presented theoretical results in numerical examples. We consider various inverse problems which can be formulated in the form of (7.1) in order to learn the regularization parameter. These example will be based on partial differential equations, both linear and nonlinear which includes again a linear 2D Laplace equation, a 2D Darcy flow from geophysical sciences and in addition a 2D eikonal equation which arises in wave propagation. As a final numerical experiment we test our theory on an image denoising problem.
For the linear example, we have access to the exact derivative of the Tikhonov solution for the bilevel optimization. In particular, we can implement both offline and online bilevel optimization methodologies. In contrast, finding the exact derivatives for nonlinear inverse problems is difficult both in derivation and computation, so we will only use online methods with approximated gradient. For online methods, we implement the following variants:

- bSGD: Application of the bilevel SGD, Algorithm 10 with exact derivative (7.9).
- bSGD$_a$: Application of the bilevel SGD, Algorithm 11 with derivative approximation (7.12) for fixed $h_k = h_0$ in (7.11).

For our first model we have tested both bSGD and bSGD$_a$, while for the nonlinear models we have used bSGD$_a$. It is worth mentioning that we have also tested, as a side experiment, using the adaptive derivative $h_k = h_0/k^{1/4}$. For these experiments it was shown that the adaptive derivative scheme does not show any major difference to the case of fixed $h_k = h_0$. In fact, Theorem 7.1.17 has already implied this, since the difference between the two scheme is of order h_0^{-4}, which is often smaller than the error from the numerical forward map solver or the use of $\bar{\kappa}_n$. For this reason, we do not present this scheme in our numerics.

Linear example: 2D Laplace equation

As first example, we consider the following forward model

$$\begin{cases} -\Delta p(x) &= \theta(x), \quad x \in D := [0,1]^2, \\ p(x) &= 0, \qquad x \in \partial D, \end{cases} \tag{7.18}$$

with Lipschitz domain D and consider the corresponding inverse problem of recovering the unknown $\theta^\dagger$ from observation of (7.18), which are described through

$$y = \mathcal{O}(p) + \xi,$$

where $\xi \sim \mathcal{N}(0, \Gamma)$ is measurements noise and p solves (7.18). The PDE has been solved in weak form, where we denote the solution operator for (7.18) by $\mathfrak{L}^{-1} : \mathcal{X} \to \mathcal{V}$, with $\mathcal{X} = L^\infty(D)$ and $\mathcal{V} = H_0^1(D) \cap H^2(D)$, and $\mathcal{O} : \mathcal{V} \to \mathbb{R}^K$ denotes again the observation map taking measurements at K randomly chosen points in D, i.e. $\mathcal{O}(p) = (p(x_1), \dots, p(x_K))^\top$, for $p \in \mathcal{V}$, $x_1, \dots, x_K \in D$. In our experiments we observe $K = 250$ points, which are illustrated in Figure 7.1. We can express this problem as a linear inverse problem in the reduced form (2.2) by

$$y^\dagger = L\theta^\dagger + \eta \in \mathbb{R}^K,$$

where $L = \mathcal{O} \circ \mathfrak{L}^{-1}$ is the forward operator taking measurements of (7.18). We again solve the forward model (7.18) numerically on a uniform mesh with 1024 grid points in D by a finite element method with continuous, piecewise linear finite element basis functions.
Our unknown parameter $\theta^\dagger$ is assumed to be Gaussian distributed $\mathcal{N}(0, \frac{1}{\kappa_*}C_0)$, where the covariance is defined as

$$C_0 = \beta \cdot (\tau^2 \cdot \mathrm{Id} - \Delta)^{-\alpha}, \tag{7.19}$$

with Laplacian operator Δ equipped with Dirichlet boundary conditions, known $\beta,\ \tau > 0$, $\alpha > 1$ and unknown $\kappa^* > 0$. To sample from the Gaussian distribution, we consider the truncated KL expansion, see section 2.2.4, which is a series representation for $\theta \sim \mathcal{N}(0, C_0)$, i.e.

$$\theta(x) = \sum_{i=1}^{\infty} \zeta_i \sqrt{\frac{1}{\kappa_*}\nu_i}\varphi_i(x), \tag{7.20}$$

where $(\nu_i, \varphi_i)_{i\in\mathbb{N}}$ are the eigenvalues and eigenfunction of the covariance operator C_0 and $\zeta = (\zeta_i)_{i\in\mathbb{N}}$ is an i.i.d. sequence with $\zeta_1 \sim \mathcal{N}(0,1)$ i.i.d.. Here, we have sampled from the KL expansion for the discretized C_0 on the uniform mesh. Furthermore, we assume to have access to training data $(\theta^{(j)}, y^{(j)} = L\theta^{(j)} + \xi^{(j)})_{j=1,\dots,n}$, $n \in \mathbb{N}$, which we will use to learn the unknown scaling parameter κ_* before solving the inverse problem. For the numerical experiment we set $\beta = 100,\ \tau = 0.1,\ \alpha = 2$ and $\kappa_* = 0.1$. After learning the regularization parameter, we will compare the estimated parameter through the different results of the Tikhonov minimum

$$\theta_{\kappa_i}(y^\dagger) = (L^\top \Gamma^{-1} L + \kappa_i \cdot C_0)^{-1} L^\top y^\dagger,$$

for $\kappa_1 = \widehat{\kappa}$ learned from the training data, $\kappa_2 = \kappa_*$ and fixed $\kappa_3 = 1$. We have used the `MATLAB` function `fmincon` to recover the the regularization parameter offline by solving the empirical optimization problem

$$\widehat{\kappa}_n \in \arg\min_{\kappa>0} \frac{1}{n}\sum_{j=1}^{n} |\theta_\kappa(y^{(j)}) - \theta^{(j)}|^2.$$

We use $M = 1000$ samples of the training data to construct Monte–Carlo estimates of $\mathbb{E}[|\widehat{\kappa}_n - \kappa^\dagger|^2]$. Further, we also compare the results to the proposed online recovery in form of the SGD method in order to learn the regularization parameter κ. Here, we run Algorithm 10 with chosen step size $\beta_k = 200/k$, range of regularization parameter $\Lambda = [0.0001, 10]$ and initial value $\kappa_0 = 1$. The resulting iterate κ_k can be seen in Figure 7.2 on the right side.
Summarizing the numerical experiments for the linear example, we observe that the numerics match our derived theory. For the offline recovery setting, we compare the MSE with theoretical rate, see Figure 7.2 on the left side, which decay at the same rate. In the same Figure we see on the right side the convergence of the online recovery towards κ_* as iteration progress. Further, we show the result of the approximate bSGD method Algorithm 11 for fixed chosen $h_k = 0.01$ in (7.11). As the derivative approximation (7.11) is closely exact, we see very similar good performance of the approximate bSGD method. Finally, we show the recovery of the underlying unknown through different choices of κ in Figure 7.3. We verify that the adaptive choice of κ outperforms that of fixed choice of the regularization parameter $\kappa = 1$.

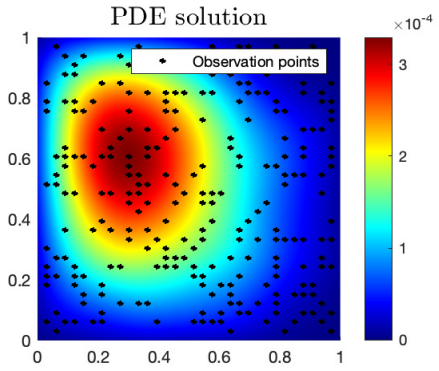

Figure 7.1: Reference PDE solution for the Laplace equation of the underlying unknown parameter $\theta^\dagger$, and the corresponding randomized observation points $x_1, \dots, x_K \in D$.

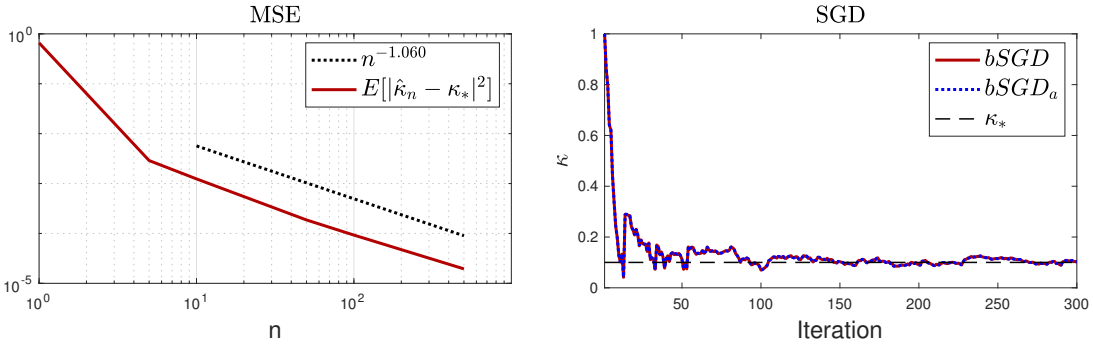

Figure 7.2: MSE (left) resulting from the offline recovery depending on training data size. Learned regularization parameter κ_k (right) resulting from the online recovery, Algorithm 10 for the Laplace equation.

Nonlinear example: 2D Darcy flow

We now consider again the elliptic PDE introduced in Section 4.4.2. The forward model is again concerned using the log-permeability $\log \theta \in L^\infty(D) =: \mathcal{X}$ to solve for the pressure $p \in H_0^1(D) \cap H^2(D) =: \mathcal{V}$ from

$$\begin{cases} -\nabla \cdot (\exp(u(x)) \nabla p(x)) = f(x), & x \in D := [0,1]^2 \\ p(x) = 0, & x \in \partial D \end{cases}$$

with known scalar field $f \in \mathbb{R}$, see also (4.14). The corresponding inverse problem is the recovery of the unknown parameter $\theta^\dagger$ from noisy observation of (4.14), described through

$$y = \mathcal{O}(p) + \xi,$$

where $\mathcal{O} : \mathcal{V} \to \mathbb{R}^K$ denotes the linear observation map, which takes again measurements at K randomly chosen points in D, i.e. $\mathcal{O}(p) = (p(x_1), \dots, p(x_K))^\top$, for $p \in \mathcal{V}$, $x_1, \dots, x_K \in D$. For our numerical setting we choose $K = 125$ observational points, which can again be seen in Figure 7.4. The measurements noise is denoted by $\xi \in \mathcal{N}(0, \Gamma)$, for $\Gamma \in \mathbb{R}^{K \times K}$ symmetric and positive definite.

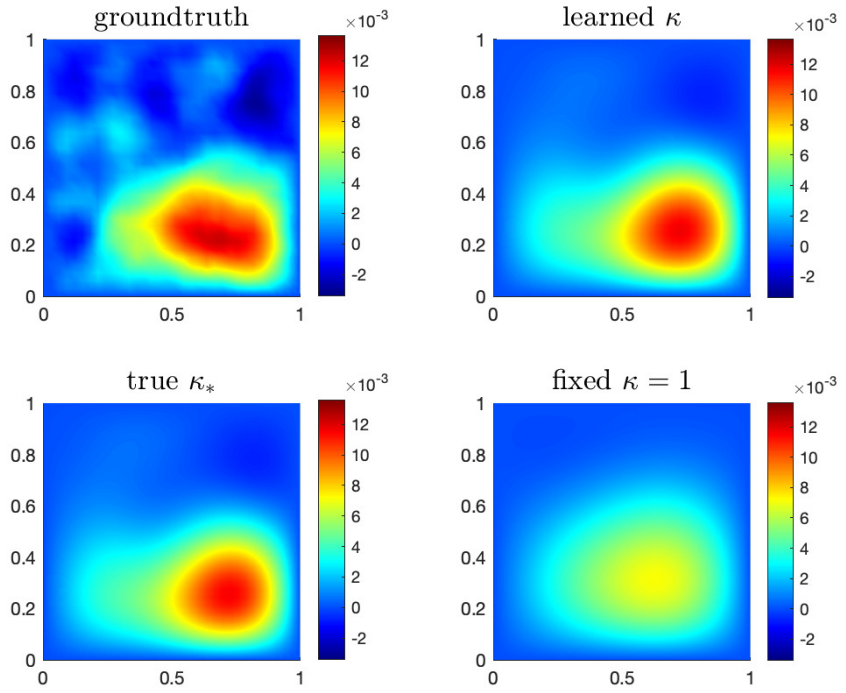

Figure 7.3: Comparison of different Tikhonov solutions for choices of regularization parameter κ_i. The learned Tikhonov regularized solution corresponds to the resulting one of the SGD method Algorithm 10 for the Laplace equation.

To apply the presented methods, we formulate the inverse problem through

$$y^\dagger = H(\theta^\dagger) + \xi,$$

with $H = \mathcal{O} \circ S$, where $S : \mathcal{X} \to \mathcal{V}$ denotes the solution operator of (4.14), solving the PDE (4.14) in weak form. The forward problem (4.14) has been solved again by a second-order centered finite difference method on a uniform mesh with 256 grid points.
We assume that $\theta^\dagger$ follows the Gaussian distribution $\mathcal{N}(0, \frac{1}{\kappa_*} C_0)$ with a covariance operator (7.19) prescribed with Neumann boundary condition. Similar as before, $\beta, \tau > 0$ and $\alpha > 1$ are known, while $\kappa_* > 0$ is unknown. We again apply the KL expansion and this time, we do estimation of the coefficients ζ, see Section 6.6.4 for more details. Therefore, we truncate (7.20) up to I and consider the nonlinear map $H : \mathbb{R}^I \to \mathbb{R}^K$, with $H(\zeta) = \mathcal{O} \circ S(\theta^\zeta(\cdot))$ and

$$\theta^\zeta(\cdot) = \sum_{i=1}^{d} \zeta_i \sqrt{\frac{1}{\kappa_*} \nu_i} \varphi_i(\cdot).$$

Hence, our unknown parameter is given by $\zeta \in \mathbb{R}^I$ and we set a Gaussian prior on ζ with $\mathcal{N}(0, \frac{1}{\kappa_*} \mathrm{Id})$, where $\kappa_* > 0$ is unknown.
Furthermore, we again assume to have access to training data $(\zeta^{(j)}, y^{(j)})_{j=1,\dots,n}$, $n \in \mathbb{N}$,

where $\zeta^{(j)} \sim Z \sim \mathcal{N}(0, \frac{1}{\kappa_*}\,\mathrm{Id})$ and we aim to solve the original bilevel optimization problem

$$\widehat{\kappa} \in \arg\min_{\kappa>0} \mathbb{E}[\|\theta_\kappa(Y) - Z\|^2], \quad \theta_\kappa(Y) = \arg\min_{\zeta\in\mathbb{R}^I} \frac{1}{2}\|H(\zeta) - Y\|_\Gamma^2 + \frac{\kappa}{2}\|\zeta\|_{\mathrm{Id}}^2.$$

The corresponding empirical optimization problem is given by

$$\widehat{\kappa}^n \in \arg\min_{\kappa>0} \frac{1}{n}\sum_{j=1}^n \|\theta_\kappa(y^{(j)}) - \zeta^{(j)}\|^2, \quad \theta_\kappa(y^{(j)}) = \arg\min_{\zeta\in\mathbb{R}^I} \frac{1}{2}\|H(\zeta) - y^{(j)}\|_\Gamma^2 + \frac{\kappa}{2}\|\zeta\|_{\mathrm{Id}}^2, \tag{7.21}$$

for a given size of the training data n. We are now not able to compute the Tikhonov minimum analytically for each observation $y^{(j)}$, and it requires more computational power to solve (7.21). Hence, we solve (7.21) online by application of Algorithm 11, where we approximate the derivative of the forward model by centered different method (7.11). We keep the accuracy of the numerical approximation fixed to $h_k = 0.01$.
For our numerical results we choose $I = 25$ coefficients in the KL expansion and the noise covariance $\Gamma = \gamma^2\,\mathrm{Id}$ with $\gamma = 0.001$. For the prior model set $\beta = 10$, $\alpha = 2$, $\tau = 3$ and the true scaling parameter $\kappa_* = 0.1$.
For the SGD method we have chosen a step size $\beta_k = 0.001k^{-1}$. We observe that the learned parameter moves fast into direction of the true scaling parameter κ_*, where it oscillates around this value. The variance reduces with the iterations, as seen in Figure 7.6.
Finally, Figure 7.5 highlights again the importance and improvements of choosing the right regularization parameters.

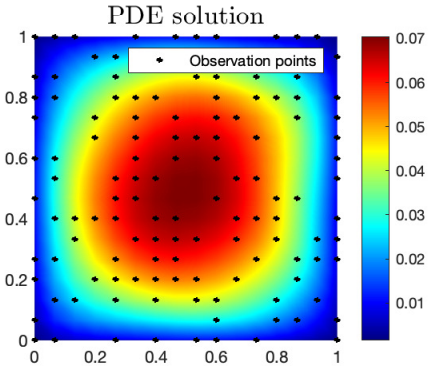

Figure 7.4: Reference PDE solution for Darcy flow of the underlying unknown parameter $\theta^\dagger$ and the corresponding randomized observation points $x_1, \ldots, x_K \in D$.

Nonlinear example: Eikonal equation

As third example we seek to test our theory on the eikonal equation, which concerns with wave propagation. Given a slowness or inverse velocity function $s(x) \in C^0(\bar{D}) =: \mathcal{X}$, characterizing the medium, and a source location $x_0 \in D$, the forward eikonal equation is to solve for travel time $T(x) \in C^0(\bar{D}) =: \mathcal{V}$ satisfying

$$\begin{cases} |\nabla T(x)| &= s(x), \quad x \in D \setminus \{x_0\}, \\ T(x_0) &= 0, \\ \nabla T(x) \cdot \tau(x) &\geq 0, \quad x \in \partial D. \end{cases} \tag{7.22}$$

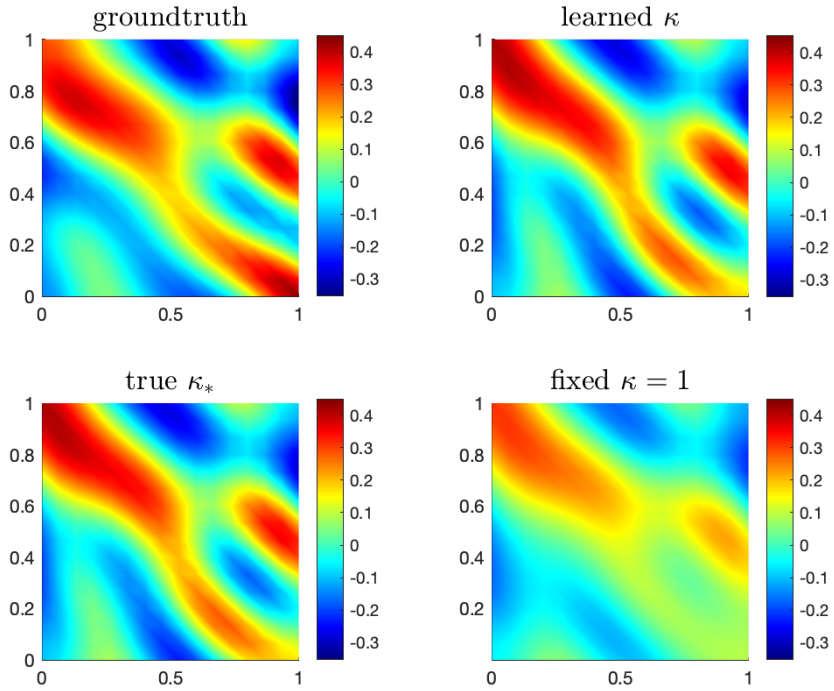

Figure 7.5: Comparison of different Tikhonov solutions for choices of the regularization parameter κ. The learned Tikhonov regularized solution corresponds to the resulting one of the SGD method Algorithm 11 for Darcy flow.

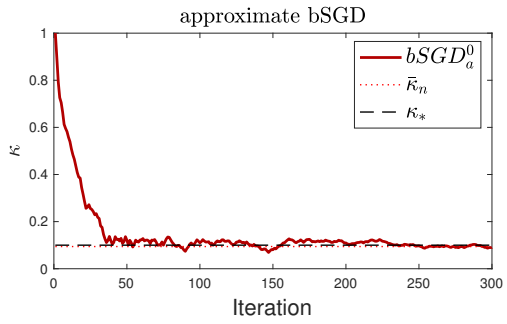

Figure 7.6: Learned regularization parameter κ_k, for Darcy flow, resulting from the approximate bilevel SGD method Algorithm 11 with fixed derivative accuracy $h = h_0$ and the corresponding mean over the last 50 iterations $\bar{\kappa}_n$. We obtain an error $|\kappa_* - \bar{\kappa}_n|^2 = 3.3640\text{e}{-05}$.

The forward solution $T(x)$ represents the shortest travel time from x_0 to a point in the domain D. The Soner boundary condition imposes that the wave propagates along the unit outward normal $\tau(x)$ on the boundary of the domain.

We formulate the inverse problem for (7.22) as the recovery of the speed function $s = \exp(\theta(x))$ from measurements of the shortest travel time $T(x)$. The data is assumed to

take the form

$$y = \mathcal{O}(T) + \xi,$$

where $\mathcal{O} : \mathcal{V} \to \mathbb{R}^K$ denotes again the linear observation map, taking measurements at $K = 125$ randomly chosen grid points in D, i.e. $\mathcal{O}(p) = (T(x_1), \dots, T(x_K))^\top$, for $T \in \mathcal{V}$, $x_1, \dots, x_K \in D$. The observed points can be seen in Figure 7.7. The measurements noise is again denoted by $\xi \in \mathcal{N}(0, \Gamma)$, for $\Gamma \in \mathbb{R}^{K \times K}$ symmetric and positive definite. We formulate the inverse problem through

$$y^\dagger = H(\theta^\dagger) + \xi,$$

with $H = \mathcal{O} \circ S$, where $S : \mathcal{X} \to \mathcal{V}$ denotes the solution operator of (7.22). The unknown parameter $\theta^\dagger$ is again assumed to be distributed according to a Gaussian measure with mean zero and covariance structure (7.19). We set $\beta = 1,\ \tau = 0.1,\ \alpha = 2$ and $\kappa_* = 0.1$ and truncate the KL expansion such that the unknown parameters $\zeta \in \mathbb{R}^I$ with $I = 25$. We apply a similar approach as in the previous example, that is we apply the SGD described through Algorithm 11. We choose an adaptive step size

$$\beta_k = \min\left(0.002, \frac{\kappa_0}{|\nabla_\kappa f(\kappa_k, Z^{(k)})|}\right) k^{-1}.$$

Here, the chosen step size β_k provides a bound on the maximal moved step in each SGD step, i.e.

$$|\beta_k \cdot \nabla_\kappa f(\kappa_k, Z^{(k)})| \leq \kappa_0 / k.$$

This helps to avoid instability arising through the high variance of the stochastic gradient, but the step size will be mainly of order $0.002/k$. However, from theoretical side it is not clear whether assumption of (7.8) is still satisfied. Therefore, we will also show the resulting $\sum_{k=1}^n \beta_k$ and the realisation of the stochastic gradient $\nabla f(\kappa_k, Z^{(k)})$ in Figure 7.10.
Our setting for the parameter choices of our prior and for the bilevel-optimization problem remain the same. To discretize (7.22) on a uniform mesh with 256 grid points we use a fast marching method, described by the work of Sethian [65, 209].
We highlight in Figure 7.8 that using the learned κ_n provides almost identical recoveries to that of using the true scaling κ_*. As we have expected for both cases we see an improvement over the case $\kappa = 1$. The convergence of the online recovery is verified through Figure 7.9, where we again see oscillations of the learned κ_k around the true scaling κ_*. Finally from Figure 7.10 we see that the summation of our choice β_k diverges, but not as quickly as the summation of the deterministic step size $0.002/k$ does, which is the implication of the introduced adaptive upper bound based on the size of the stochastic gradient $\nabla_\kappa f(\kappa_k, Z^{(k)})$. Figure 7.10 also shows the histrogram of the stochastic gradient and its rare realized large values.

Signal denoising example

The next example is devoted to implement our methods on a image denoising example. We are interested in denoising a 1D compound Poisson process of the form

$$\Theta_t = \sum_{i=1}^{N_t} X_i, \tag{7.23}$$

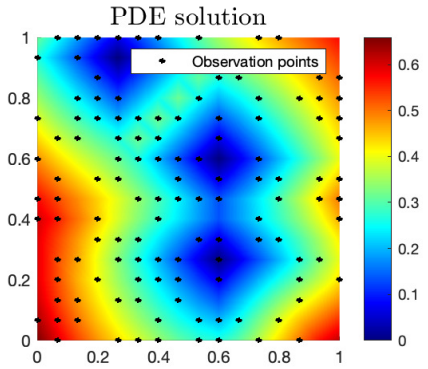

Figure 7.7: Reference PDE solution for the eikonal equation of the underlying unknown parameter $\theta^\dagger$, and the corresponding randomized observation points $x_1, \dots, x_K \in D$.

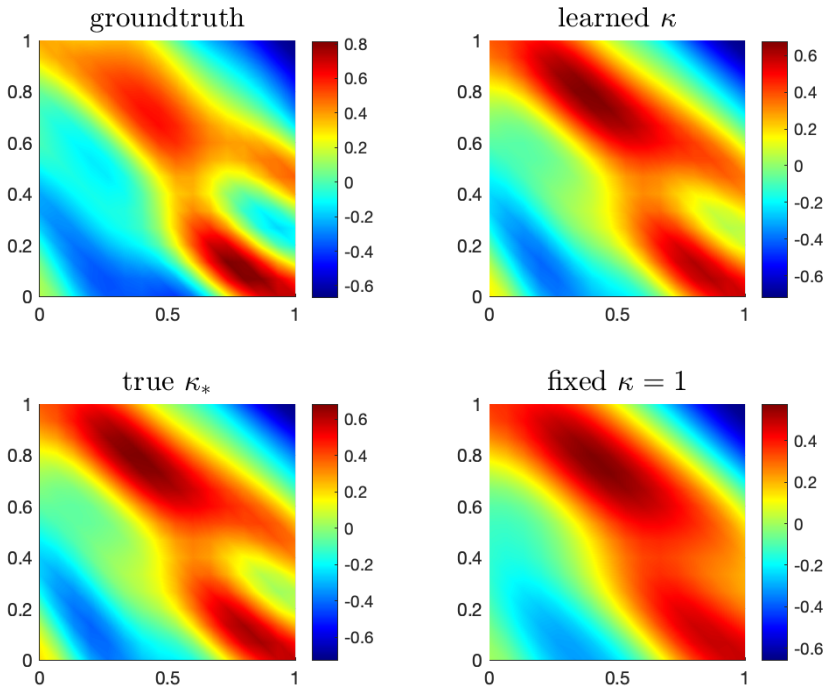

Figure 7.8: Comparison of different Tikhonov solutions for choices of the regularization parameter κ. The learned Tikhonov regularized solution corresponds to the resulting one of the SGD method Algorithm 11 for the eikonal equation.

where $(N_t)_{t\in[0,T]}$ is a Poisson process, with rate $r > 0$ and $(X_i)_{i\in\mathbb{N}}$ are i.i.d. random variables representing the jump size. Here, we have chosen $X_1 \sim \mathcal{N}(0,1)$.

We consider the inverse problem of recovering a perturbed signal of the form (7.23). The aim is to solve this problem through Tikhonov regularization with different choices of regularization parameter κ. In particular, the observed signal $\theta = (\theta_{t_1}, \dots, \theta_{t_I})^\top \in \mathbb{R}^I$ is

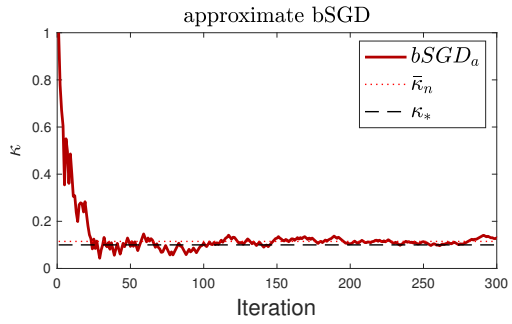

Figure 7.9: Learned regularization parameter κ_k, for the eikonal equation, resulting from the approximate bilevel SGD method Algorithm 11 with fixed derivative accuracy $h = h_0$ and the corresponding mean over the last 50 iterations $\bar{\kappa}_n$. We obtain an error $|\kappa_* - \bar{\kappa}_n|^2 = 1.9360\text{e}{-05}$.

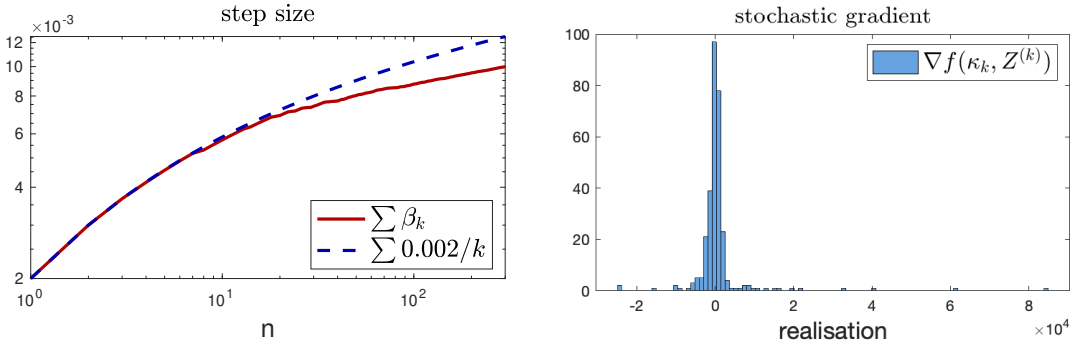

Figure 7.10: Summation of the realized adaptive step size (left) and the realized stochastic gradient $\nabla f(\kappa_k, Z^{(k)})$ (right) resulting from the online recovery, Algorithm 10 for the eikonal equation.

perturbed by white noise

$$y_{t_i} = \theta_{t_i} + \xi_{t_i}, \tag{7.24}$$

where $t_i \in \{1/I \cdot T, 2/I \cdot T, \dots, T\}$ and $\xi_{t_i} \sim \mathcal{N}(0, \sigma^2)$ are i.i.d. random variables, and the Tikhonov estimate corresponding to the lower level problem of (7.1) for given regularization parameter $\kappa > 0$ is defined by

$$\theta_\kappa(y) = (\Gamma^{-1} + \kappa R^{-1})^{-1}\Gamma^{-1}y, \tag{7.25}$$

with given regularization matrix $R \in \mathbb{R}^{I\times I}$ and $y = (y_1, \dots, y_I)^\top \in \mathbb{R}^I$. We assume to have access to training data $(\theta^{(j)}, y^{(j)})_{j=1}^n$ of (7.24) and choose the regularization parameter $\widehat{\kappa}$ according to Algorithm 10. Further, we compare the resulting estimate of the signal

$$y_{obs} = \theta^\dagger + \xi,$$

to fixed choices of $\kappa \in \{0.01,\ 0.00001\}$ and to the best possible choice $\kappa_* = \arg\min_\kappa \|\theta_\kappa(y_{obs}) - \theta^\dagger\|^2$.

For the experiment we set the rate of jumps $r = 10$ and consider the signal observed up to time $T = 1$ at $I = 1000$ observation points. For Algorithm 10, we use a training data set of size $n = 500$, we set an initial value $\kappa_0 = 0.001$ and step size $\beta_k = 0.001k^{-1}$. The Tikhonov solution (7.25) has been computed with a second-order regularization matrix $R = \Delta^{-1}$.

$\boldsymbol{\kappa}$	$1e-02$	$1e-05$	$\overline{\kappa}_n$	κ_*
error	0.0378	0.0134	0.0077	0.0073

Table 7.1: MSE over time of the reconstruction for different choices of the regularization parameter for signal denoising example.

As we can see from our results the value of $\kappa = 0.01$ oversmoothens the estimate in comparison with $\kappa = 0.00001$. This is shown in Figure 7.11. However, comparing fixed κ with the learned κ in Figure 7.12 we see an improvement, closer to the best possible κ, which is verified further through Table 7.1, where we can see the MSE over the time intervall. Both Figure 7.11 and Figure 7.12 show on the right hand side the pointwise squared error over time.

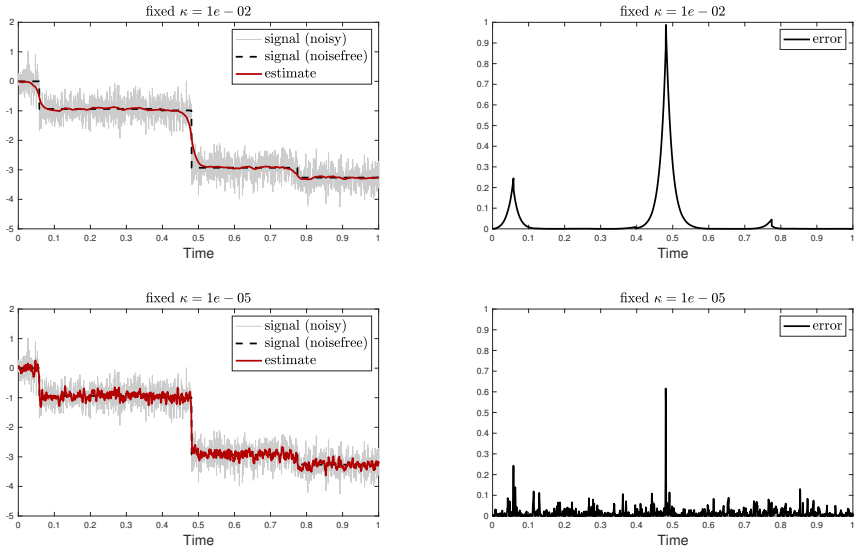

Figure 7.11: Comparison of different Tikhonov solutions for fixed choices of the regularization parameter κ for the signal denoising example.

7.2 Incorporation of neural networks approximation within inverse problems

In this part of the work, we will mainly focus on PDE-constraint inverse problems. The aim is to avoid the computation of the complex forward model by replacing it through a neural network as surrogate model. The surrogate model will be physically informed by the underlying model equation. To solve the inverse problem, we train the neural network and the unknown parameter in a one-shot fashion. By connection to the Bayesian approach to inverse problems we provide the application of the EKI method as optimizer for the training task. The formulation of the neural network based one-shot inversion will be provided in Section 7.2.1 and the application of EKI will be formulated in Sec-

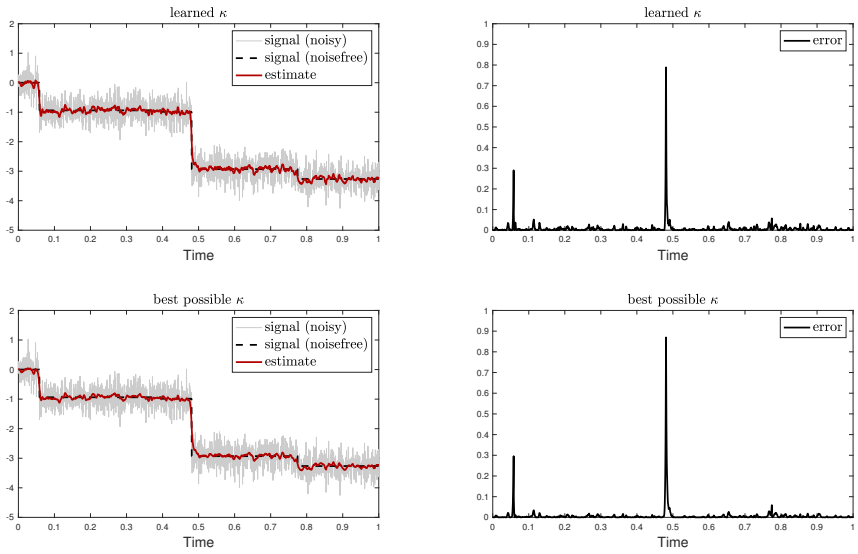

Figure 7.12: Comparison of the learned to best possible Tikhonov solutions for choices of the regularization parameter κ. The learned Tikhonov regularized solution corresponds to the resulting one of the SGD method Algorithm 10 for the signal denoising example.

tion 7.2.2. In Section 7.2.3 we briefly describe the connection to the recently invented physics-informed neural networks. The discussion will be closed in Section 7.2.4 with two numerical examples.
The corresponding constraint optimization problem is given by

$$\begin{aligned} \min_{u,p} \ & \|\mathcal{O}(p) - y\|^2_{\Gamma_{obs}} \\ \text{s.t. } & M(\theta, p) = 0\,, \end{aligned} \tag{7.26}$$

where we aim to recover the unknown parameter $\theta \in \mathcal{X}$ from the PDE solution $p \in \mathcal{V}$, which is typically defined on a domain $D \subset \mathbb{R}^d$ and will be observed at finitely many points given by

$$\mathcal{O}(p) = y \in \mathbb{R}^K.$$

This observation might also be perturbed by noise and the data misfit in (7.26) will be scaled by a suitable symmetric, positive defnite matrix $\Gamma_{obs} \in \mathbb{R}^{K\times K}$. The PDE (or ODE) model is described in $M : \mathcal{X} \times \mathcal{V} \to \mathcal{W}$. We restrict the discussion to finite dimensional spaces $\mathcal{X} = \mathbb{R}^I$, $\mathcal{V} = \mathbb{R}^v$ and $\mathcal{W} = \mathbb{R}^w$.
As we have already discussed in Section 2.1, these kind of problems are typically ill-posed and we incorporate regularization, i.e. we consider the regularized optimization problem

$$\begin{aligned} \min_{u,p} & \|\mathcal{O}(p) - y\|^2_{\Gamma_{obs}} + \kappa_1 \varphi_1(\theta) \\ \text{s.t. } & M(\theta, p) = 0\,, \end{aligned} \tag{7.27}$$

where the regularization is denoted by $\varphi_1 : \mathcal{X} \to \mathbb{R}$ and the positive scalar $\kappa_1 > 0$ is usually chosen according to prior knowledge on the unknown parameter θ. See Subsection 2.1.2 for more details on regularization of inverse problems.

We first introduce the so-called reduced problem of (7.26) and (7.27), resepectively. The forward model $M(\theta, p) = 0$ is typically a well-posed problem, in the sense that for each parameter $\theta \in \mathcal{X}$, there exists a unique state $p \in \mathcal{V}$ such that $M(\theta, p) = 0$ in $\mathcal{W}$. Introducing the solution operator $S : \mathcal{X} \to \mathcal{V}$ s.t. $M(\theta, S(\theta)) = 0$, we can reformulate the optimization problem (7.26) as an unconstrained optimization problem in form of (2.5) and (2.6)-(2.7), which is

$$\min_{\theta \in \mathcal{X}} \ \|\mathcal{O}(S(\theta)) - y\|^2_{\Gamma_{obs}} \,, \tag{7.28}$$

and

$$\min_{\theta \in \mathcal{X}} \ \|\mathcal{O}(S(\theta)) - y\|^2_{\Gamma_{obs}} + \kappa_1 \varphi_1(\theta) \,, \tag{7.29}$$

respectively. We refer to (7.28) and (7.29) as reduced formulation of the inverse problem. Note that in the reduced formulation we can now formulate again the Bayesian approach, this means we can compute the posterior distribution following (2.16) and the corresponding MAP estimator (2.25) for Gaussian prior assumption $\mathbb{Q}_0 = \mathcal{N}(m_0, C_0)$, which coincides with the choice $\varphi_1(\theta) = \|m_0 - \theta\|^2_{C_0}$.
In the following we will introduce the one-shot formulation, or sometimes also called all-at-once formulation, for inverse problems in order to incorporate neural networ approximation.

7.2.1 Neural networks based one-shot inversion

While the reduced formulation (7.28) assumes that the forward problem can be solved exactly in each iteration, we will follow a different approach, which simultaneously solves the forward and optimization problem. Various names for the simultaneous solution of the design and state equation exist: one-shot method, all-at-once, piggy-back iterations etc.. We refer the reader to [27] and the references therein for more details. For a theoretical analysis of one-shot methods in the context of inverse problems we refer to [127, 123].
Following the one-shot ideas, we seek to solve the problem

$$F(\theta, p) = \begin{pmatrix} M(\theta, p) \\ \mathcal{O}(p) \end{pmatrix} = \begin{pmatrix} 0 \\ y \end{pmatrix} =: \tilde{y} \,,$$

Due to the noise in the observations, we rather consider

$$y = \mathcal{O}(p) + \xi_{obs}$$

with normally distributed noise $\xi_{obs} \sim \mathcal{N}(0, \Gamma_{obs})$, $\Gamma_{obs} \in \mathbb{R}^{K \times K}$ symmetric and positive definite. Similarly, we assume that

$$0 = M(u, p) + \xi_{model} \,,$$

i.e. we assume that there could occur uncertainty in the model through an error described by $\xi_{model} \sim \mathcal{N}(0, \Gamma_{model})$, $\Gamma_{model} \in \mathbb{R}^{w \times w}$ symmetric and positive definite. This leads to the problem

$$\tilde{y} = F(\theta, p) + \begin{pmatrix} \xi_{model} \\ \xi_{obs} \end{pmatrix} .$$

Following the Bayesian approach the MAP estimate is then computed by the solution of the following minimization problem

$$\min_{\theta, p} \frac{1}{2} \|F(\theta, p) - \tilde{y}\|^2_{\Gamma} + \kappa_1 \varphi_1(\theta) + \kappa_2 \varphi_2(p),$$

where $\varphi_1 : \mathcal{X} \to \mathbb{R}$ and $\varphi_2 : \mathcal{V} \to \mathbb{R}$ are regularization functions of the parameter $\theta \in \mathcal{X}$ and the state $p \in \mathcal{V}$, $\kappa_1, \kappa_2 > 0$ and $\Gamma = \begin{pmatrix} \Gamma_{model} & 0 \\ 0 & \Gamma_{obs} \end{pmatrix} \in \mathbb{R}^{(w+K)\times(w+K)}$.

Vanishing noise and penalty methods

In order to force the forward model to be exact, i.e. the forward equation is supposed to be satisfied exactly with $M(\theta, p) = 0$, we view vanishing noise in the Bayesian setting. To do so, we consider a parametrized noise covariance model $\Gamma_{model} = \gamma\hat{\Gamma}_{model}$ for $\gamma \in \mathbb{R}_+$ and given symmetric positive definite matrix $\hat{\Gamma}_{model}$. Taking the limit $\gamma \to 0$ corresponds to the vanishing noise setting, which can be interpret as reducing the uncertainty in our model. The MAP estimate in the one-shot formulation changes to

$$\min_{\theta,p} \frac{1}{2}\|\mathcal{O}(p) - y\|^2_{\Gamma_{obs}} + \frac{\lambda}{2}\|M(u,p)\|^2_{\hat{\Gamma}_{model}} + \kappa_1\mathcal{R}_1(u) + \kappa_2\mathcal{R}_2(p) \tag{7.30}$$

with $\lambda = 1/\gamma$. We mention the close connection to penalty methods, which aim to solve constrained problems such as (7.26) by solving unconstrained optimization problems of the form (7.30) sequentially for monotonically decreasing penalty parameters λ. To make this idea rigorous, we cite the following well-known regarding the convergence of the resulting algorithm.

Proposition 7.2.1 ([20])**.** *Let the observation operator $\mathcal{O}$, the forward model M and the regularization functions φ_1, φ_2 be continuous and the feasible set $\{(\theta,p)|M(\theta,p) = 0\}$ be nonempty. For $k = 0, 1, \ldots$ let (θ_k, p_k) denote a global minimizer of*

$$\min_{\theta,p} \frac{1}{2}\|\mathcal{O}(p) - y\|^2_{\Gamma_{obs}} + \frac{\lambda_k}{2}\|M(\theta,p)\|^2_{\hat{\Gamma}_{model}} + \kappa_1\varphi_1(u) + \kappa_2\varphi_2(p)$$

with $(\lambda_k)_{k\in\mathbb{N}} \subset \mathbb{R}_+$ strictly monotonically increasing and $\lambda_k \to \infty$ for $k \to \infty$. Then every accumulation point of the sequence (θ_k, p_k) is a global minimizer of

$$\begin{aligned} &\min_{\theta,p} \frac{1}{2}\|\mathcal{O}(p) - y\|^2_{\Gamma_{obs}} + \kappa_1\varphi_1(\theta) + \kappa_2\varphi_2(p) \\ &s.t.\ M(\theta,p) = 0\,. \end{aligned}$$

This classic convergence result ensures the feasibility of the estimates, i.e. the proposed approach is able to incorporate and exactly satisfy physical constraints in the limit. We mention also the possibility to consider exact penalty terms in the objective, corresponding to different noise models in the Bayesian setting.
This setting will be the starting point of the incorporation of neural networks into the problem. Starting from this setting, we will incorporate neural networks into the problem in the following way: We replace the state p by an approximate solution by a neural network p_Υ, where $\Upsilon \in \mathbb{R}^{n_\Upsilon}$ denote the parameters of the neural network which have to be learned within this framework. We consider the following minimization problem

$$\min_{\theta,\Upsilon} \frac{1}{2}\|F(\theta, p_\Upsilon) - \tilde{y}\|^2_\Gamma + \kappa_1\varphi_1(\theta) + \kappa_2\varphi_2(p_\Upsilon, \Upsilon)\,, \tag{7.31}$$

which will be refered to neural network based one-shot formulation.

Neural networks

From a mathematical perspective, we interpret neural networks as parametrized functions which will be used to approximate general functions. In our setting, we will focus on the class of deep neural networks (DNNs) which are defined as a function $p_\Upsilon : \mathbb{R}^d \to \mathbb{R}$, where the input is defined as $x \in \mathbb{R}^d$. The number of hidden layers in the NN will be given by L and N_l will denote the corresponding amount of nodes in each layer $l \in \{1, \dots, L\}$. We define the DNN recursively: The first hidden layer is set to

$$z^1 = \sigma^*(W^1 x + b^1), \tag{7.32}$$

where $W^1 \in \mathbb{R}^{N_1 \times d}$, $b^1 \in \mathbb{R}^{N_1}$ represent the weights and the bias, the parameters of the DNN, and $z^1 \in \mathbb{R}^{N_1}$ denotes the output. From (7.32) we denote σ^* as the component-wise operation of the activation function $\sigma : \mathbb{R} \to \mathbb{R}$. In the remaining hidden layers, we define the corresponding output by

$$z^{l+1} = \sigma^*(W^{l+1} z^l + b^{l+1}), \quad l \in \{1, \dots, L-1\},$$

where, as before, $W^{l+1} \in \mathbb{R}^{N_{l+1} \times N_l}$, $b^{l+1} \in \mathbb{R}^{N_{l+1}}$ denote the input parameters and $z^{l+1} \in \mathbb{R}^{N_{l+1}}$ represents the output of the hidden layer $l+1$. Finally, the output layer is defined by

$$p_\Upsilon(x) = W^{L+1} z^L + b^{L+1}, \quad W^{L+1} \in \mathbb{R}^{N_{L+1} \times N_L},\ b^{L+1} \in \mathbb{R}^{N_{L+1}},$$

where we collect all of the parameters by $\Upsilon = \{W^l, b^l\}_{l=1,\dots,L+1}$. In Figure 7.13 we illustrate the structure of DNNs.

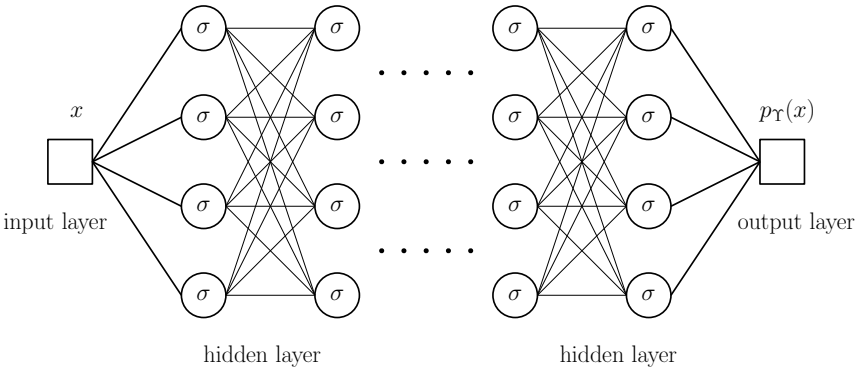

Figure 7.13: Structure of a deep neural network.

For more details on neural networks and its wide range of application, we refer the interested reader to [81, 101].
In mathematical science, neural networks have been applied in various research areas such as partial differential equations or inverse problems. For example, in [10] the authors present an overview of neural networks applied as regularization function in inverse problems. In the context of parametric PDEs, the neural network is applied to mimik the map from parameter to the PDE solution, resulting in significantly faster compution of PDE solutions [138, 196, 142]. Based on the approximation results of polynomials by feed-forward DNNs [229], the authors in [207] derive bounds on the expression rate for multivariate, real-valued functions depending holomorphically on a sequence $z = (z_j)_{j \in \mathbb{N}}$ of parameters. More specifically, the authors consider functions that admit sparse Taylor

generalized polynomial chaos (gpc) expansions, i.e. s-summable Taylor gpc coefficients. Such functions arise as response surfaces of parametric PDEs, or in a more general setting from parametric operator equations, see e.g. [205] and the references therein. Their main results is that these functions can be expressed with arbitrary accuracy $\delta > 0$ (uniform w.r. to z) by DNNs of size bounded by $C\delta^{-s/(1-s)}$ with a constant $C > 0$ independent of the dimension of the input data z. Similar results for parametric PDEs can be found in [142].
The methods in [207] motivated the work of [98] in which the authors show holomorphy of the data-to-QoI map $y \mapsto \mathbb{E}^{\mu^*}[\mathrm{QoI}]$ for additive, centered Gaussian observation noise in Bayesian inverse problems. Using the fact that holomorphy implies fast convergence of Taylor expansions, the authors derived an exponential expression rate bound in terms of the overall network size. Our approach of how to incorporate neural networs into PDE based inverse problems differs from these results, and is closely connected to so called physics-informed neural networks. See Subsection 7.2.3 for more details on this connection.

7.2.2 Application of the ensemble Kalman inversion

In the following we are going to introduce the application of EKI in order to solve the neural network based one-shot formulation (7.31). By approximating the state of the underlying PDE by a neural network, we seek to optimize the unknown parameter θ and on the other side the parameters of the neural network Υ. To do so, we define $\mathcal{H} := \mathcal{X} \times \mathbb{R}^{n_\Upsilon}$, $G : \mathcal{H} \to \mathbb{R}^{w+K}$, $G(v) = G(\theta, \Upsilon) = F(\theta, p_\Upsilon)$, $v = (\theta, \Upsilon)^\top \in \mathcal{X} \times \mathbb{R}^{n_\Upsilon}$, $q = \tilde{y} \in \mathbb{R}^{w+K}$, $\zeta = \begin{pmatrix} \xi_{model} \\ \xi_{obs} \end{pmatrix}$ and apply Algorithm 5.
This leads to the empirical summary statistics

$$\overline{(\theta, \Upsilon)}_n = \frac{1}{J}\sum_{j=1}^{J}(\theta_n^{(j)}, \Upsilon_n^{(j)}), \quad \bar{G}_n = \frac{1}{J}\sum_{j=1}^{J} G(\theta_n^{(j)}, \Upsilon_n^{(j)}),$$

$$C_n^{(\theta,\Upsilon)\tilde{y}} = \frac{1}{J}\sum_{j=1}^{J}\big((\theta_n^{(j)}, \Upsilon_n^{(j)})^\top - \overline{(\theta, \Upsilon)}_n^\top\big) \otimes \big(G\big(\theta_n^{(j)}, \Upsilon_n^{(j)}\big) - \bar{G}_n\big),$$

$$C_n^{\tilde{y}\tilde{y}} = \frac{1}{J}\sum_{j=1}^{J}\big(G(\theta_n^{(j)}, \Upsilon_n^{(j)}) - \bar{G}_n\big) \otimes \big(G(\theta_n^{(j)}, \Upsilon_n^{(j)}) - \bar{G}_n\big),$$

and following (3.4) the EKI update

$$(\theta_{n+1}^{(j)}, \Upsilon_{n+1}^{(j)})^\top = (\theta_n^{(j)}, \Upsilon_n^{(j)})^\top + C_n^{(\theta,\Upsilon)\tilde{y}}\big(C_n^{\tilde{y}\tilde{y}} + h^{-1}\Gamma\big)^{-1}\big(\tilde{y}_{n+1}^{(j)} - G(\theta_n^{(j)}, \Upsilon_n^{(j)})\big), \quad (7.33)$$

where the perturbed observation are computed as before in (3.5)

$$\tilde{y}_{n+1}^{(j)} = \tilde{y} + \xi_{n+1}^{(j)}, \quad \xi_{n+1}^{(j)} \sim \mathcal{N}(0, h^{-1}\Gamma),$$

with

$$\tilde{y} = \begin{pmatrix} 0 \\ y \end{pmatrix}, \quad \Gamma := \begin{pmatrix} \Gamma_{model} & 0 \\ 0 & \Gamma_{obs} \end{pmatrix}.$$

Figure 7.14 illustrates the basic idea of the application of the EKI to solve the neural network based one-shot formulation.

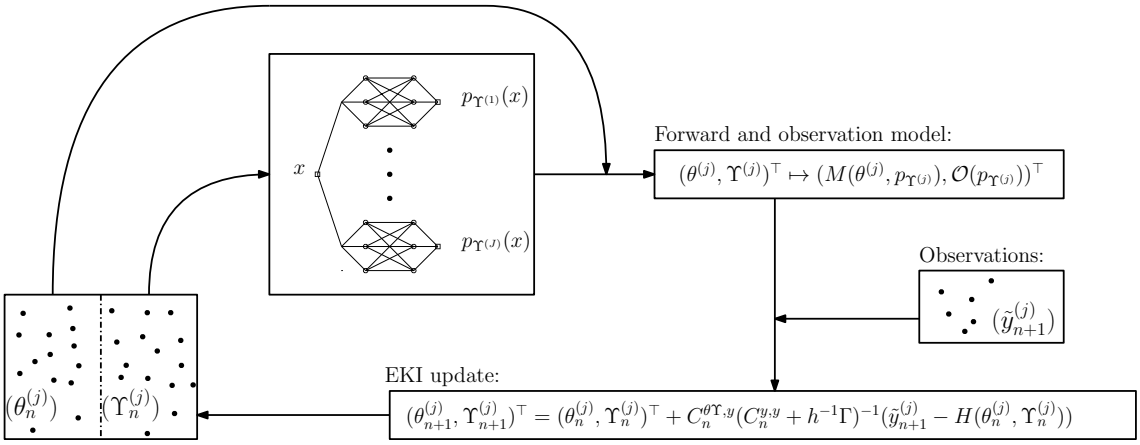

Figure 7.14: Description of the EKI applied to solve the neural network based one-shot formulation.

The EKI (7.33) will be used as a derivative free optimizer of the data misfit $\|F(\theta, p_\Upsilon) - \tilde{y}\|_\Gamma^2$. The analysis presented in Chapter 3 showed that the EKI in its continuous form is able to recover the data with a finite number of particles in the limit $t \to \infty$ under suitable assumptions on the forward problem and the set of particles. In particular, the analysis assumed a linear forward problem. The limit $t \to \infty$ corresponds to the noise-free setting, as the inverse noise covariance scales with $n/N = nh$ in (3.6). As seen in Chapter 5, the application of the EKI in the inverse setting therefore often requires additional techniques such as adaptive stopping or additional regularization to overcome the ill-posedness of the minimization problem. To control the regularization of the data misfit and neural network individually, similar to (5.1) we consider the following extended system

$$\begin{aligned} F(\theta, p_\Upsilon) + \begin{pmatrix} \xi_{model} \\ \xi_{obs} \end{pmatrix} &= \tilde{y} \\ \begin{pmatrix} \theta \\ \Upsilon \end{pmatrix} + \begin{pmatrix} \xi_{param} \\ \xi_{NN} \end{pmatrix} &= 0 \end{aligned}$$

with $\xi_{model} \sim \mathcal{N}(0, 1/\lambda\, \hat{\Gamma}_{model})$, $\xi_{obs} \sim \mathcal{N}(0, \Gamma_{obs})$, $\theta \sim \mathcal{N}(\theta_0, 1/\kappa_1\, C)$ and $\Upsilon \sim \mathcal{N}(0, 1/\kappa_2\ \mathrm{Id})$. The loss function for the augmented system is therefore given by

$$\frac{1}{2}\|\mathcal{O}(p_\Upsilon) - y\|_{\Gamma_{obs}}^2 + \frac{\lambda}{2}\|M(\theta, p_\Upsilon)\|_{\hat{\Gamma}_{model}}^2 + \frac{\kappa_1}{2}\|\theta - \theta_0\|_C^2 + \frac{\kappa_2}{2}\|\Upsilon\|^2\,. \tag{7.34}$$

Assuming that the resulting forward operator

$$G(\theta, \Upsilon) = \begin{pmatrix} F(\theta, p_\Upsilon) \\ \theta \\ \Upsilon \end{pmatrix} \tag{7.35}$$

is linear, the EKI will converge to the minimum of the regularized loss function (7.34). To ensure the feasibility of the EKI estimate (w.r. to the underlying forward problem), we

propose the following algorithm using the ideas discussed in Section 7.2.1.

Algorithm 12: Penalty ensemble Kalman inversion for neural network based one-shot inversion

Input: initial ensemble $v_0^{(j)} = (\theta_0^{(j)}, \Upsilon_0^{(j)})^\top \in \mathcal{X} \times \mathbb{R}^{n_\Upsilon}, j = 1, \dots J,\ \lambda_0$.
Output: $v_N = (\theta_N, \Upsilon_N)^\top$
for $k = 0, 1, 2, \dots, N$ **do**

- Compute an approximation of the minimizer $(\theta_k, \Upsilon_k)^\top$ of

$$\min_{\theta,\Upsilon} \ \frac{1}{2}\|O(p_\Upsilon) - y\|^2_{\Gamma_{obs}} + \frac{\lambda_k}{2}\|M(\theta, p_\Upsilon)\|^2_{\hat{\Gamma}_{model}} + \frac{\kappa_1}{2}\|\theta - \theta_0\|^2_C + \frac{\kappa_2}{2}\|\Upsilon\|^2 .$$

  by solving

$$\frac{\mathrm{d}v^{(j)}}{\mathrm{d}t} = C^{vy}(v)\Gamma^{-1}(\widehat{y} - G(v^{(j)})) + C^{vy}(v^{(j)})\Gamma^{-1}\sqrt{\Gamma}\,\frac{\mathrm{d}W^{(j)}}{\mathrm{d}t}$$

  with $v^{(j)}(0) = v_0^{(j)}$ for the system (7.35), $\widehat{y} = (0, y, 0, 0)^\top$ and $\Gamma = \begin{pmatrix} 1/\lambda_k\,\Gamma_{model} & 0 \\ 0 & \Gamma_{obs} \end{pmatrix}$.

- Set $v_k = (\theta_k, \Upsilon_k)^\top = \bar{v}(T)$ for $T \to \infty$.
- Increase λ_k.
- Draw J ensemble members $v_0^{(j)}$ from $\mathcal{N}(v_k, \begin{pmatrix} C & 0 \\ 0 & I \end{pmatrix})$.

Theorem 7.2.2. *Assume that the forward operator $G : \mathcal{X} \times \mathbb{R}^{n_\Upsilon} \to \mathbb{R}^{n_G}$, $n_G := w + K + I + n_\Upsilon$,*

$$G(\theta, \Upsilon) = \begin{pmatrix} F(\theta, p_\Upsilon) \\ \theta \\ \Upsilon \end{pmatrix}$$

is linear, i.e. $F(\theta, p_\Upsilon) = L\begin{pmatrix} \theta \\ \Upsilon \end{pmatrix}$ with $L \in \mathcal{L}(\mathcal{X} \times \mathbb{R}^{n_\Upsilon}, \mathbb{R}^{w+K})$. Let $(\lambda_k)_{k\in\mathbb{N}} \subset \mathbb{R}_+$ be strictly monotonically increasing and $\lambda_k \to \infty$ for $k \to \infty$. Further, assume that the initial ensemble members are chosen so that $\mathrm{span}\{(\theta^{(j)}(0), \Upsilon^{(j)}(0))^\top, j = 1, \dots, J\} = \mathcal{X} \times \mathbb{R}^{n_\Upsilon}$. Then, Algorithm 12 generates a sequence of estimates $(\bar{\theta}_k, \bar{\Upsilon}_k)_{k\in\mathbb{N}}$, where $\bar{\theta}_k, \bar{\Upsilon}_k$ minimizes the loss function for the augmented system given by

$$\frac{1}{2}\|\mathcal{O}(p_\Upsilon) - y\|^2_{\Gamma_{obs}} + \frac{\lambda_k}{2}\|M(\theta, p_\Upsilon)\|^2_{\hat{\Gamma}_{model}} + \frac{\kappa_1}{2}\|\theta - \theta_0\|^2_C + \frac{\kappa_2}{2}\|\Upsilon\|^2$$

with given $\kappa_1, \kappa_2 > 0$. Furthermore, every accumulation point of $(\bar{\theta}_k, \bar{\Upsilon}_k)_{k\in\mathbb{N}}$ is the (unique, global) minimizer of

$$\begin{aligned} \min_{\theta,\Upsilon} \ & \frac{1}{2}\|\mathcal{O}(p_\Upsilon) - y\|^2_{\Gamma_{obs}} + \frac{\kappa_1}{2}\|\theta - \theta_0\|^2_C + \frac{\kappa_2}{2}\|\Upsilon\|^2 \\ \text{s.t.} \ & M(\theta, p_\Upsilon) = 0 \end{aligned}$$

Proof. Under the assumption of a linear forward model, the penalty function

$$\frac{1}{2}\|\mathcal{O}(p_\Upsilon) - y\|^2_{\Gamma_{obs}} + \frac{\lambda_k}{2}\|M(\theta, p_\Upsilon)\|^2_{\hat{\Gamma}_{model}} + \frac{\kappa_1}{2}\|\theta - \theta_0\|^2_C + \frac{\kappa_2}{2}\|\Upsilon\|^2$$

is strictly convex for all $k \in \mathbb{N}$, i.e. there exists a unique minimizer of the penalized problem. Choosing the initial ensemble such that $\text{span}\{(\theta^{(j)}(0), \Upsilon^{(j)}(0))^\top, j = 1, \dots, J\} = \mathcal{X} \times \mathbb{R}^{n_\Upsilon}$ ensures the convergence of the EKI estimate to the global minimizer, see [41, Theorem 3.13] and [200, Theorem 4]. The convergence of Algorithm 12 to the minimzer of the constrained problem then follows from Proposition 7.2.1. □

Remark 7.2.3. *We note that the presented convergence result is based on an assumption on the size of ensemble, which ensures the convergence to the (global) minimizer in each iteration. This is due to the well-known subspace property of the EKI, see Lemma 3.1.2. This assumption is usually not satisfied in practice , for example for large or possibly infiite-dimensional parameter / state space. However, techniques such as variance inflation, localization and adaptive ensemble choice are able to overcome the subspace property and might lead to more efficient algorithms from a computational perspective.*

Furthermore, we stress the fact that the convergence result presented above requires the linearity of the forward and observation operator (w.r. to the optimization variables), i.e. the assumption is not fulfilled when considering neural networks with a nonlinear activation function as approximation of the forward problem. However, we will demonstrate in the numerical experiments that the EKI shows promising results even in the nonlinear setting. The generalization of the theory to the nonlinear case will be subject to future work.

To accelerate the computation of the minimizer, we suggest the following variant of Algorithm 12, which increases the penalty parameter λ adaptively.

Algorithm 13: Simultaneous penalty ensemble Kalman inversion for neural network based one-shot inversion

Input: initial ensemble $v_0^{(j)} = (\theta_0^{(j)}, \Upsilon_0^{(j)})^\top \in \mathcal{X} \times \mathbb{R}^{n_\Upsilon}, j = 1, \dots J,\ \lambda_0 > 0$, $f : \mathbb{R}_{\geq 0} \to \mathbb{R}_+$.
Output: $v_T = (\bar{\theta}_T, \bar{\Upsilon}_T)^\top$
Compute an approximation of the minimizer of

$$\begin{aligned} &\min_{\theta, \Upsilon} \ \frac{1}{2}\|\mathcal{O}(p_\Upsilon) - y\|^2_{\Gamma_{obs}} + \frac{\kappa_1}{2}\|\theta - \theta_0\|^2_C + \frac{\kappa_2}{2}\|\Upsilon\|^2 \\ &\text{s.t.} \quad M(\theta, p_\Upsilon) = 0 \end{aligned}$$

by solving the following system

$$\begin{aligned} \frac{\mathrm{d}v^{(j)}}{\mathrm{d}t} &= C^{vy}(v)\Gamma^{-1}(\widehat{y} - G(v^{(j)})) + C^{vy}(v^{(j)})\Gamma^{-1}\sqrt{\Gamma}\,\frac{\mathrm{d}W^{(j)}}{\mathrm{d}t} \\ \frac{\mathrm{d}\lambda}{\mathrm{d}t} &= f(\lambda) \end{aligned}$$

with $v^{(j)}(0) = v_0^{(j)}$ for the system (7.35), $\lambda(0) = \lambda_0$, $\widehat{y} = (0, y, 0, 0)^\top$ and $\Gamma = \begin{pmatrix} 1/\lambda\,\Gamma_{model} & 0 \\ 0 & \Gamma_{obs} \end{pmatrix}$.

We note that the function f has to be chosen in a way, such that the scaling parameter λ is monotonically increasing. In our numerical examples, we have chosen as an example $\frac{\mathrm{d}\lambda}{\mathrm{d}t} = 1/\lambda$, i.e. $f(\lambda) = 1/\lambda$.

7.2.3 Connection to physics-informed neural networks

In combination with a one-shot approach for the training of the neural network parameters, our method is closely related to the physics-informed neural networks (PINNs). In [181, 182] the authors consider PDEs of the form

$$f(t,x) := p_t + N(p,\lambda) = 0, \quad t \in [0,T],\, x \in D\,,$$

where N is a nonlinear differential operator parametrized by λ. The authors replace p by a neural network p_Υ and use automatic differentiation to construct the function $f_\Upsilon(t,x)$. The neural network parameters are then obtained by minimizing $\mathrm{MSE} = \mathrm{MSE}_p + \mathrm{MSE}_f$, where

$$\mathrm{MSE}_p := \frac{1}{N_p}\sum_{i=1}^{N_p} |p(t_p^i, x_p^i) - p^i|^2\,, \quad \mathrm{MSE}_f := \frac{1}{N_f}\sum_{i=1}^{N_f} |f(t_f^i, x_f^i)|^2\,,$$

where N_p is the number of training data points (t_p^i, x_p^i, p^i) for the PDE solution $p(t,x)$ and N_f the number of collocation points (t_f^i, x_f^i) for $f(t,x)$ respectively. For the minimization a L-BFGS method is used. The parameters λ of the differential operator turn into parameters of the neural network f_θ and can be learned by minimizing the MSE.

In [228] the authors consider so called Bayesian neural networks (BNNs), where the neural network parameters are updated according to Bayes' theorem. Hereby the initial distribution on the network parameters serves as prior distribution. The likelihood requires the PDE solution, which is obtained by concatenating the Bayesian neural network with a physics-informed neural network, which they call Bayesian physics-informed neural networks (B-PINNs). For the estimation of the posterior distributions they use the Hamiltonian Monte Carlo method and variational inference. In contrast to the PINNs, the Bayesian framework allows them to quantify the aleatoric uncertainty associated with noisy data. In addition to that, their numerical experiments indicate that B-PINNs beat PINNs in case of large noise levels on the observations. In contrast to that, our proposed method is based on the MAP estimate and remains exact in the small noise limit. We have propose a derivative free optimization method, the EKI, which shows promising results (also compared to quasi-Newton methods) without requiring derivatives w.r. to the weights and design parameters. More details on this can be found in the following subsection.

7.2.4 Numerical results

We present two numerical experiments in order to illustrate the introduced one-shot formulation. In the first example we consider the linear one-dimensional example introduced in Example 2.1.15. We compare a black-box method, quasi-Newton method for the one-shot formulation and EKI for the one-shot formulation. Further, we also consider the quasi-Newton and EKI for the neural network based one-shot inversion. We numerically explore the convergence behavior of the EKI for the neural networks based one-shot inversion (Algorithm 13) also for a nonlinear forward model.

The second experiment is concerned with a linear two-dimensional problem which investigates the potential of the EKI for neural network based inversion in the higher-dimensional setting.

One-dimensional example

We consider the problem presented in Example 2.1.15, which is devoted to recover the unknown data $\theta^\dagger$ from noisy observations

$$y = \mathcal{O}(p^\dagger) + \eta^\dagger,$$

where $p^\dagger = \mathfrak{L}^{-1}(\theta^\dagger)$ is the solution of the one-dimensional elliptic equation

$$\begin{cases} -\frac{\mathrm{d}^2 p}{\mathrm{d}x^2}(x) + p(x) = \theta^\dagger(x), & x \in D := (0, \pi), \\ p(x) = 0, & x \in \partial D, \end{cases}$$

with operator $\mathcal{O}$ observing the dynamical system at $K = 2^3 - 1$ equispaced observation points $x_i = \frac{i}{2^4} \cdot \pi$, $i = 1, \dots, K$ and mesh size $h = 2^{-6}$.
We again assume that the unknown parameter θ is Gaussian with $\theta \sim \mathcal{N}(0, C_0)$, where $C_0 = \beta(-\frac{\mathrm{d}^2}{\mathrm{d}x^2})^{-\nu}$ for $\beta = 5,\ \nu = 1.5$. Further, we assume a observational noise covariance $\Gamma_{obs} = 0.1 \cdot I_{n_y}$, a model error covariance $\hat{\Gamma}_{model} = 100 \cdot I_{n_u}$ and we set a regularization parameter $\kappa_1 = 0.002$, while we turn off the regularization on p, i.e. we set $\kappa_2 = 0$. The feed-forward DNN has $L = 3$ hidden layers with $N_1 = N_2 = 10$ size of hidden layers and $N_0 = N_L = 1$ size of the input and output layer. We set the sigmoid function $\sigma^*(x) = \frac{1}{1+e^{-x}}$ as activation function.
The EKI method is based on the deterministic formulation (3.13), which will be solved with the `MATLAB` function `ode45` up to time $T = 10^{10}$. We initialize the ensemble of particles $(\theta^{(j)})$, $(\theta^{(j)}, p^{(j)})$ and $(\theta^{(j)}, \Upsilon^{(j)})$ respectively, as i.i.d. samples with $\theta_0^{(j)}$ drawn from the prior distribution $\mathcal{N}(0, C_0)$, $p_0^{(j)}$ drawn from $\mathcal{N}(0, 5\,\mathrm{Id}_v)$ and the DNN weights drawn from $\mathcal{N}(0, \mathrm{Id}_{n_\Upsilon})$. These samples are all independent from each other.
We compare the results to a classical gradient-based method, which will be a quasi-Newton method with BFGS updates, as implemented by `MATLAB`.
To summarize the considered method we introduce the following abbreviations:

1. reduced formulation: explicit solution (redTik).
2. one-shot formulation: we compare the performance of the EKI with Algorithm 12 (osEKI_1), the EKI with Algorithm 13 (osEKI_2) and the quasi-Newton method with Algorithm 12 (osQN_1).
3. neural network based one-shot formulation: we compare the performance of the EKI with Algorithm 13 (nnosEKI_2) and the quasi-Newton method with Algorithm 12 (nnosQN_1).

Figure 7.15 shows the increasing sequence of λ used for Algorithm 12 and the quasi-Newton method and Algorithm 13 (over time).

One-shot inversion We compare the one-shot inversion based on the FEM approximation of the forward model in our 1d example. The following results illustrate the convergence of the EKI and numerically investigate the performance of Algorithm 13.
In Figure 7.16 we see the difference of the estimates given by EKI with Algorithm 12 (osEKI_1), the EKI with Algorithm 13 (osEKI_2) and the quasi-Newton method with Algorithm 12 (osQN_1) compared to the Tikhonov solution and the truth (on the left-hand side) and in the observation space (on the right-hand side). All of the three methods result

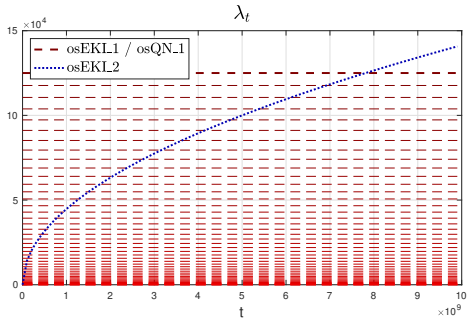

Figure 7.15: Scaling parameter λ depending on time for Algorithm 1, $\lambda_k = k^3$ for $k = 1, 2, \ldots, 50$, and Algorithm 2, $\mathrm{d}\lambda/\mathrm{d}t = 1/\lambda$.

in an excellent approximation of the Tikhonov solution. As the forward model is linear, we expect the quasi-Newton method as well as the EKI with Algorithm 12 to converge to the regularized solution. Similarly the EKI with Algorithm 13 shows good performance while reducing the computational effort significantly compared to Algorithm 12.

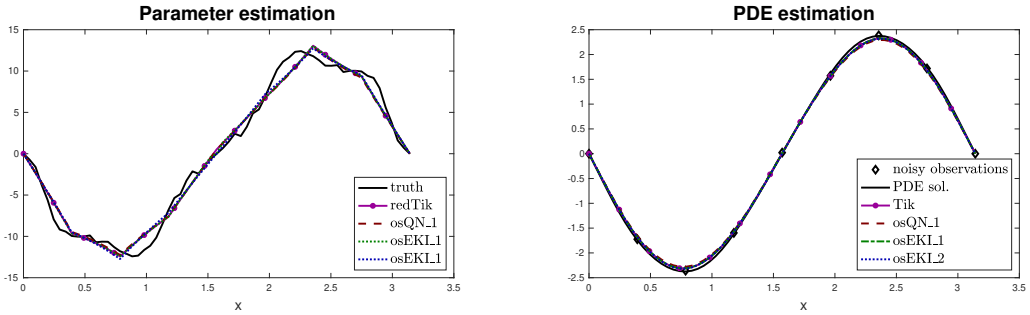

Figure 7.16: Comparison of parameter estimation given by EKI with Algorithm 12 (osEKI_1), the EKI with Algorithm 13 (osEKI_2) and the quasi-Newton method with Algorithm 12 (osQN_1) compared to the Tikhonov solution and the truth (on the left hand side) and in the observation space (on the right hand side).

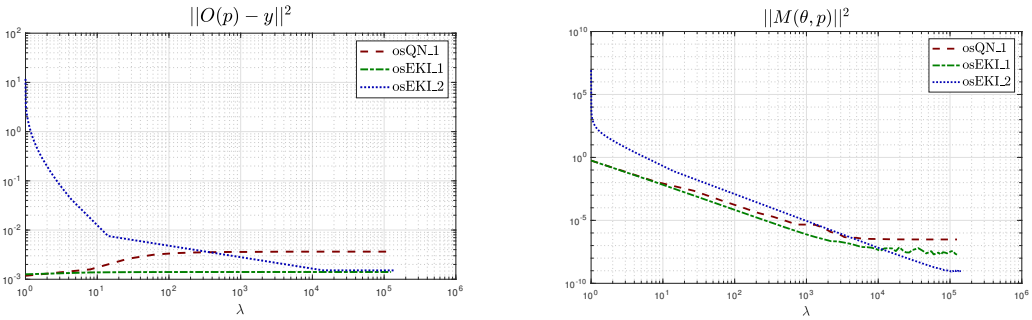

Figure 7.17: Comparison of the data misfit given by EKI with Algorithm 12 (osEKI_1), the EKI with Algorithm 13 (osEKI_2) and the quasi-Newton method with Algorithm 12 (osQN_1) (on the left hand side) and residual of the forward problem (on the right hand side), both w.r. to κ.

By comparing the data misfit and the residuals of the forward problem, which are shown

in Figure 7.17, we illustrate the very good performance of the EKI (for both algorithms) with feasibility of the estimate (w.r.t. the forward problem) in the range of 10^{-10}.

One-shot method with neural network approximation In our next experiment we replace the forward problem by a neural network approximation in the one-shot setting. Our focus will be on Algorithm 13, as it has shown promising results in the previous experiment.

While the EKI for the neural network based one-shot inversion leads to very good approximation results of the regularized solution, cp. Figure 7.18, the quasi-Newton method performs slightly worse, which might be attributed to the nonlinearity introduced by the neural network approximation.

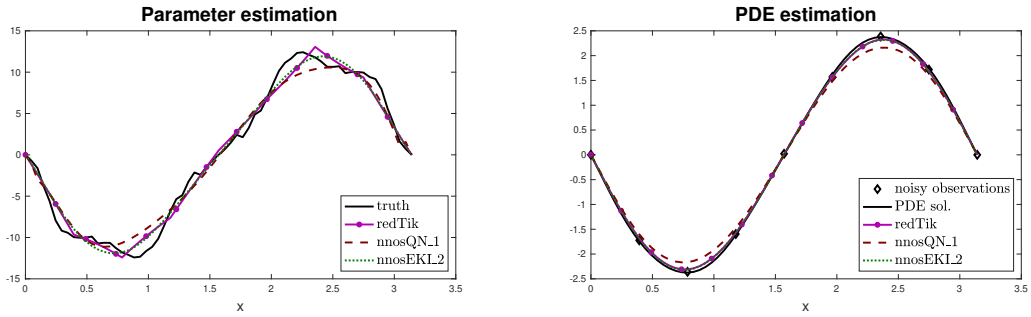

Figure 7.18: Comparison of parameter estimation given by the EKI with Algorithmn 13 (nnosEKI_2) and the quasi-Newton method with Algorithm 12 (nnosQN_1) for the neural network based one-shot inversion compared to the Tikhonov solution and the truth (on the left hand side) and in the observation space (on the right hand side).

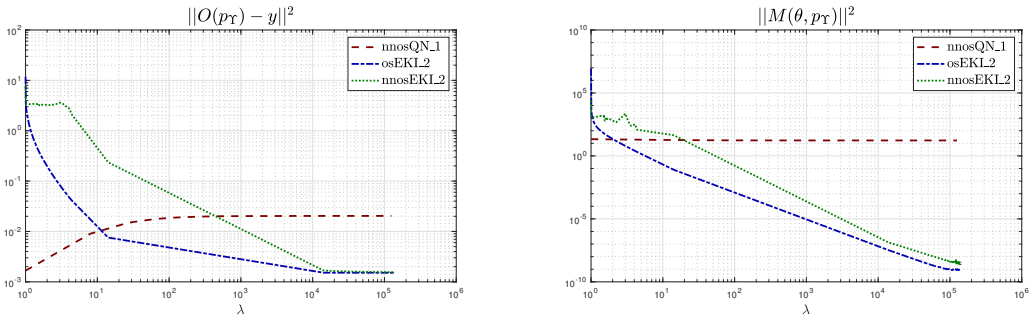

Figure 7.19: Comparison of the data misfit given by the EKI with Algorithm 13 (nnosEKI_2) and the quasi-Newton method with Algorithm 12 (nnosQN_1) for the neural network based one-shot inversion compared to EKI with Algorithm 13 (osEKI_2) from the previous experiment (on the left hand side) and residual of the forward problem (on the right hand side), both w.r. to λ.

We compare the data misfit and residual of the forward problem in Figure 7.19, which show an excellent convergence behaviour of the EKI for the neural network based one-shot optimization, while the quasi-Newton method does not converge to a feasible estimate.

Nonlinear forward model We consider in the following a nonlinear forward model of the form

$$\begin{cases} -\nabla \cdot (\exp(\theta^\dagger) \cdot \nabla p) = 10, & x \in D := (0, \pi), \\ p = 0, & x \in \partial D, \end{cases}$$

Note that the mapping from the unknown parameter function to the state is nonlinear. We use the same discretization as in the linear problem. The unknown parameter $\theta^\dagger$ is assumed to be Gaussian with zero mean and $C_0 = \beta(-\frac{\mathrm{d}^2}{\mathrm{d}x^2})^{-\nu}$ where we choose $\beta = 1$, $\nu = 2$. Further, we set $\Gamma_{obs} = 0.0001 \cdot \mathrm{Id}_K$, $\hat{\Gamma}_{model} = 10 \cdot \mathrm{Id}_I$, $\kappa_1 = 2$ and $\kappa_2 = 0$. Furthermore, the structure of the feed-forward DNN remains the same as in the linear case.

We compare the one-shot method with neural network approximation resulting from the EKI with Algorithm 13 with the Tikhonov solution of the reduced formulation, which has been approximated by a quasi-Newton method. We determine the scaling parameter λ in Algorithm 13 by the ODE $\mathrm{d}\lambda/\mathrm{d}t = 1$, i.e. the scaling parameter grows linearly. Similarly to the linear case, we find that the one-shot method with neural network approximation leads to a good approximation of the Tikhonov solution for the reduced model, cp. Figure 7.20.

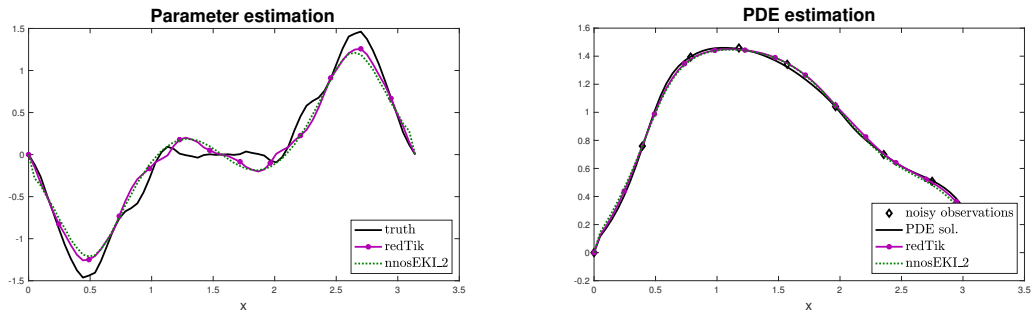

Figure 7.20: Comparison of parameter estimation given by the EKI with Algorithm 13 (osEKI_2) and the Tikhonov solution (on the left hand side) and corresponding PDE solution (on the right hand side).

In Figure 7.21, we observe that the penalty parameter λ drives the estimate towards feasibilty, i.e. towards the solution of the constrained optimization problem.

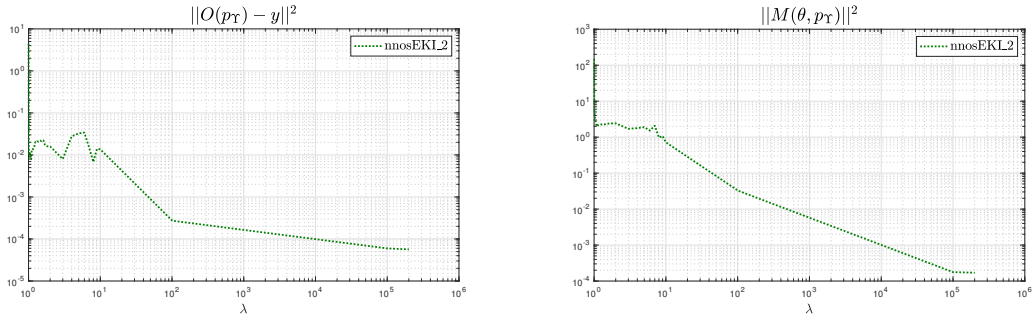

Figure 7.21: Data misfit given by the EKI with Algorithm 13 (osEKI_2) for the neural network based one-shot inversion compared (on the left hand side) and residual of the forward problem (on the right hand side), both w.r. to λ.

Two-dimensional example

In the second example we consider the two-dimensional Poisson equation

$$\begin{cases} -\Delta p = \theta^\dagger, & x \in D := (0,1)^2, \\ p = 0, & x \in \partial D, \end{cases} \tag{7.36}$$

for which we consider again the problem of recovering the unknown source term $\theta^\dagger$ from noisy observations

$$y = \mathcal{O}(p^\dagger) + \eta^\dagger,$$

where $p^\dagger$ denotes the solution of (7.36). Our observation operator $\mathcal{O}$ observes $K = 50$ randomly picked observation points x_i, $i = 1, \dots, K$, which can be seen in Figure 7.22. The forward model (7.36) has been approximated numerically with continuous, piecewise linear finite element basis functions on a mesh with 95 grid points in the interior of D and 40 grid points on ∂Dusing the `MATLAB Partial Differential Equation Toolbox`. The approximated solution operator is again denoted by $S \in \mathbb{R}^{I \times I}$, with $I = 95$. The unknown parameter θ is again modelled as Gaussian random field with (discretized) covariance operator

$$C_0 = \beta \cdot (\tau^2 \cdot \mathrm{Id} - \Delta)^{-\alpha}$$

for $\beta = 100$, $\alpha = 2$ and $\tau = 1$. We assume the observational noise covariance to be $\Gamma_{obs} = 0.01 \cdot \mathrm{Id}_K$ and the model covariance $\hat{\Gamma}_{model} = 0.1 \cdot \mathrm{Id}_I$. The regularization parameter has been set to $\kappa_1 = 0.002$ and again $\kappa_2 = 0$. We build up the feed-forward DNN with $L = 3$ layers, $N_1 = N_2 = 10$ hidden neurons, $N_0 = 2$ input neurons and $N_L = 1$ output neuron, and sigmoid activation function. The EKI method is again based on the deterministic formulation (3.13) and initialized with $J = 300$ particles drawn i.i.d. from the prior.
We display the truth and the corresponding PDE solution in Figure 7.22.

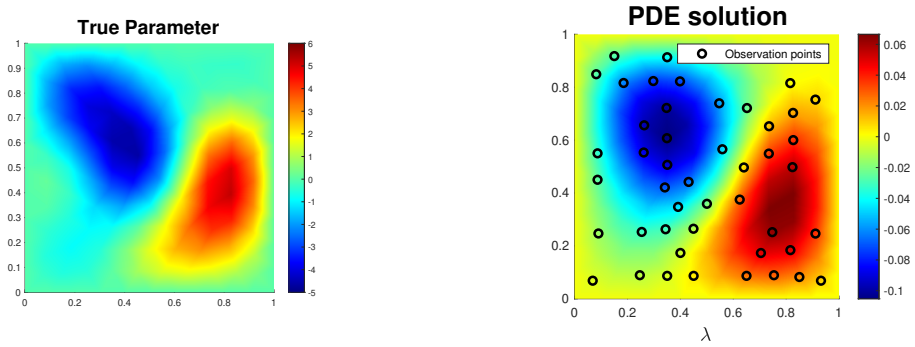

Figure 7.22: Ground truth (left hand side) and the corresponding PDE solution (right hand side).

We compare the neural network based one-shot formulation solved by the EKI with Algorithm 13 to the explicit Tikhonov solution of the reduced formulation of the corresponding inverse problem. We determine the scaling parameter λ in Algorithm 13 by the ODE $\mathrm{d}\lambda/\mathrm{d}t = 1/\lambda^2$. We demonstrate in Figure 7.23 that the EKI leads to a comparable solution.
Figure 7.24 shows that the proposed approach leads to a feasible solution w.r.t. the forward problem

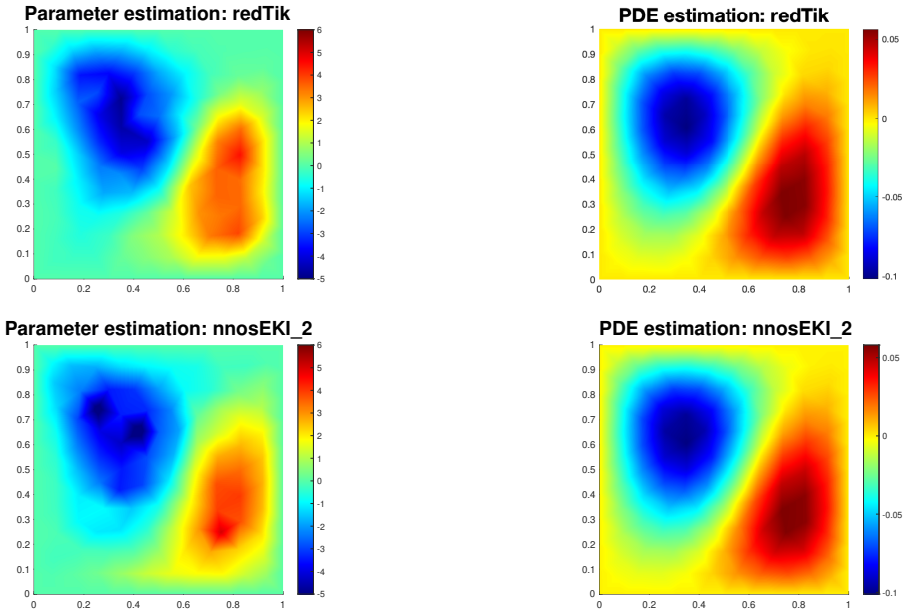

Figure 7.23: Comparison of parameter estimation given by the EKI with Algorithm 13 (osEKI_2) (below) and the Tikhonov solution (above) (on the left hand side) and corresponding PDE solution (on the right hand side).

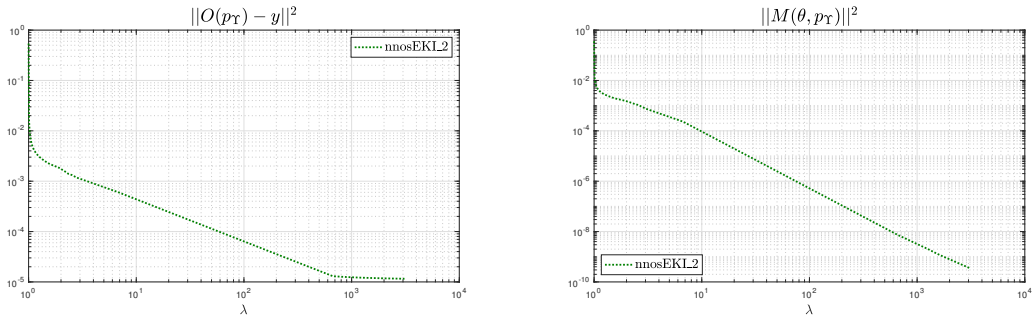

Figure 7.24: Data misfit given by the EKI with Algorithm 13 (osEKI_2) for the neural network based one-shot inversion compared (on the left hand side) and residual of the forward problem (on the right hand side), both w.r.t. λ.

8 Conclusion and outlook

We close this thesis by summarizing the discussed methods and giving a brief overview on interesting future work.

Chapter 3

We have started our discussion in Chapter 3 with the introduction of the ensemble Kalman filter applied to inverse problem and its perspective as derivativefree optimization method. We have shown well-posedness of the method by providing existence of unique strong solutions of the underyling system of stochastic differential equations. Under the assumption of linear forward maps, we have quantified the ensemble collapse and the accuracy of the method.
There are mainly two open points which should be considered for future work. The first point is to verifying the continuous-time limit (3.9). This can be done by providing convergence results of the discrete method (3.8) to the system of SDEs in a weak or strong sense. The second point is the extension of the provided convergence theory from linear to nonlinear forward maps.

Chapter 4

We have focused on the incorporation of box-constraints into ensemble Kalman inversion. The ideas were based on the projected gradient descent methods and its continuous-time limit. We have formulated a transformed method leading to a descent direction of the method.
A natural extension of this method could be the incorporation of nonlinear constraints, as well as an extension to nonlinear forward maps. Furthermore, the proposed methods can be applied to a hierarchical version of the ensemble Kalman inversion [40].

Chapter 5

In Chapter 5 we have extended the theoretical results for the ensemble Kalman inversion to the case of noisy observations in the inverse problem. We have formulated the Tikhonov regularized ensemble Kalman inversion and provided well-posedness and convergence results for linear forward maps and fixed regularization parameter. Furthermore, we have presented a row of ideas of how to adapt the regularization parameter including data-driven regularization and the MAP formulation of the Bayesian inverse problem.
Again a natural question to ask is could the theory be extended to a nonlinear setting. Since the theoretical results were based on fixed regularization parameters, similar convergence results might be proven for the adaptive regularization choice. Another direction would be to introduce other common forms of regularization, such as L_1 regularization.

Chapter 6

In this Chapter, we have discussed various particle based sampling methods. These methods included deterministic and stochastic interacting particle approximations to the Fokker–Planck formulation of overdamped Langevin dynamics. By preconditioning with the empirical covariance, we were able to formulate derivative free variants. Furthermore, we have proposed a localised preconditioning approach in order to make the resulting posterior approximation more accurate.
The presented methods of both the deterministic and stochastic interacting particle formulations can be combined with stochastic gradient descent methods [28] as well as stochastic gradient Langevin dynamics methods [227].

Chapter 7

We have provided two example of machine learning approaches for inverse problems

In the first part, we have applied bilevel optimization in order to choose the regularization parameter by learning from the data. This approach has been applied to minimization based inverse problems. We have provided bot offline and online consistency results for the empirical risk minimization problem. While we have analysed the empirical loss in the offline setting, we have formulated the stochastic gradient descent method for solving the problem online.
One possible extension might be the consideration of a Bayesian approach of the underlying bilevel optimization problem. In particular, this could be related to hierarchical learning [174]. Another potential direction is to understand statistical consistency from other choices of regularization, such as L_1 or total variation. Moreover, one could apply alternative optimizers, as for example the ensemble Kalman inversion as derivative free optimization method, to solve the bilevel optimization problem. Finally, a very interesting direction to go is the small noise limit, i.e. proving convergence of the regularized solution to the underlying minimizing norm solution in some sense.

Furthermore, we have considered the neural network based one-shot formulation for inverse problems. The neural network has been applied as surrogate model of the underlying complex forward model. We have applied the ensemble Kalman inversion in order to train the neural network and the unknown parameter in a one-shot fashion.
Several questions for future work arise. For example, a promising direction to go is to provide theoretical analysis of the neural network based one-shot inversion applying recent expressivity results for neural networks in PDE based problems. Moreover, a comparison to state-of-the-art optimization methods in the machine learning community should be discussed.

Bibliography

[1] S. Agapiou, S. Larsson, and A. M. Stuart. Posterior contraction rates for the Bayesian approach to linear ill-posed inverse problems. *Stochastic Processes and their Applications*, 123(10):3828 – 3860, 2013. ISSN 0304-4149. doi: https://doi.org/10.1016/j.spa.2013.05.001. URL http://www.sciencedirect.com/science/article/pii/S0304414913001427.

[2] S. Agapiou, O. Papaspiliopoulos, D. Sanz-Alonso, and A. M. Stuart. Importance sampling: Intrinsic dimension and computational cost. *Statistical Science*, 32:405–431, 2015.

[3] D. J. Albers, P.-A. Blancquart, M. E. Levine, E. E. Seylabi, and A. Stuart. Ensemble Kalman methods with constraints. *Inverse Problems*, 35(9):095007, aug 2019. doi: 10.1088/1361-6420/ab1c09. URL https://doi.org/10.1088%2F1361-6420%2Fab1c09.

[4] N. Amor, G. Rasool, and N. C. Bouaynaya. Constrained state estimation - a review. *arXiv: Signal Processing*, 2018.

[5] J. L. Anderson. An adaptive covariance inflation error correction algorithm for ensemble filters. *Tellus A*, 59(2):210–224, 2007. doi: 10.1111/j.1600-0870.2006.00216.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-0870.2006.00216.x.

[6] J. L. Anderson. Spatially and temporally varying adaptive covariance inflation for ensemble filters. *Tellus A*, 61(1):72–83, 2009. doi: 10.1111/j.1600-0870.2008.00361.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-0870.2008.00361.x.

[7] S. W. Anzengruber and R. Ramlau. Morozov's discrepancy principle for Tikhonov-type functionals with nonlinear operators. *Inverse Problems*, 26(2):025001, dec 2009. doi: 10.1088/0266-5611/26/2/025001. URL https://doi.org/10.1088%2F0266-5611%2F26%2F2%2F025001.

[8] D. Armbruster, M. Herty, and G. Visconti. A stabilization of a continuous limit of the ensemble Kalman filter. *ArXiv*, abs/2006.15390, 2020.

[9] N. d. F. Arnaud Doucet and N. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer New York, 2001. ISBN 9780387951461. doi: 10.1007/978-1-4757-3437-9. URL https://www.springer.com/de/book/9780387951461.

[10] S. Arridge, P. Maass, O. Öktem, and C.-B. Schönlieb. Solving inverse problems using data-driven models. *Acta Numerica*, 28:1–174, 2019. doi: 10.1017/S0962492919000059.

[11] A. Bakushinskii. On the principle of iterative regularization. *USSR Computational Mathematics and Mathematical Physics*, 19(4):256 – 260, 1979. ISSN 0041-5553. doi: https://doi.org/10.1016/0041-5553(79)90173-3. URL http://www.sciencedirect.com/science/article/pii/0041555379901733.

[12] A. Bakushinskii. Remarks on choosing a regularization parameter using the quasi-optimality and ratio criterion. *USSR Computational Mathematics and Mathematical Physics*, 24(4):181 – 182, 1984. ISSN 0041-5553. doi: https://doi.org/10.1016/0041-5553(84)90253-2. URL http://www.sciencedirect.com/science/article/pii/0041555384902532.

[13] H. T. Banks and K. Kunisch. *Estimation Techniques for Distributed Parameter Systems*. Birkhauser Boston, Inc., USA, 1989. ISBN 0817634339.

[14] M. Benning and M. Burger. Modern regularization methods for inverse problems. *Acta Numerica*, 27:1–111, 2018. doi: 10.1017/S0962492918000016.

[15] K. Bergemann and S. Reich. A localization technique for ensemble Kalman filters. *Q. J. R. Meteorological Soc.*, 136:701–707, 2010.

[16] K. Bergemann and S. Reich. A mollified ensemble Kalman filter. *Quarterly Journal of the Royal Meteorological Society*, 136(651):1636–1643, 2010.

[17] K. Bergemann and S. Reich. An ensemble Kalman–Bucy filter for continuous data assimilation. *Meteorolog. Zeitschrift*, 21:213–219, 2012.

[18] M. Bertero and P. Boccacci. *Introduction to Inverse Problems in Imaging*. CRC Press, 1998. ISBN 9781439822067. URL https://books.google.de/books?id=C02wLTkCtR0C.

[19] D. Bertsekas. Projected Newton methods for optimization problems with simple constraints. *SIAM Journal on Control and Optimization*, 20(2):221–246, 1982. doi: 10.1137/0320018. URL https://doi.org/10.1137/0320018.

[20] D. P. Bertsekas. *Nonlinear programming*. Athena Scientific, 2nd edition, Sept. 2008. ISBN 1886529000. URL http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/1886529000.

[21] A. Beskos, F. Pinski, J. Sanz-Serna, and A. Stuart. Hybrid Monte Carlo on Hilbert spaces. *Stochastic Processes and their Applications*, 121(10):2201 – 2230, 2011. ISSN 0304-4149. doi: https://doi.org/10.1016/j.spa.2011.06.003. URL http://www.sciencedirect.com/science/article/pii/S0304414911001396.

[22] A. Beskos, A. Jasra, E. A. Muzaffer, and A. M. Stuart. Sequential Monte Carlo methods for Bayesian elliptic inverse problems. *Statistics and Computing*, 25(4):727–737, 2015. doi: 10.1007/s11222-015-9556-7. URL https://doi.org/10.1007/s11222-015-9556-7.

[23] P. Bickel and K. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics*. Number Bd. 1 in Mathematical Statistics: Basic Ideas and Selected Topics. Prentice Hall, 2001. ISBN 9780138503635. URL https://books.google.de/books?id=8poZAQAAIAAJ.

[24] A. N. Bishop, P. D. Moral, K. Kamatani, and B. Rémillard. On one-dimensional Riccati diffusions. *The Annals of Applied Probability*, 29(2):1127–1187, apr 2019. doi: 10.1214/18-aap1431. URL https://doi.org/10.1214%2F18-aap1431.

[25] D. Blömker, C. Schillings, and P. Wacker. A strongly convergent numerical scheme from ensemble Kalman inversion. *SIAM Journal on Numerical Analysis*, 56(4): 2537–2562, 2018. doi: 10.1137/17M1132367. URL https://doi.org/10.1137/17M1132367.

[26] T. Bonesky. Morozov's discrepancy principle and Tikhonov-type functionals. *Inverse Problems*, 25(1):015015, dec 2008. doi: 10.1088/0266-5611/25/1/015015. URL https://doi.org/10.1088%2F0266-5611%2F25%2F1%2F015015.

[27] A. Borzi and V. Schulz. *Computational optimization of systems governed by partial differential equations*. Society for Industrial and Applied Mathematics, USA, 2012. ISBN 1611972043.

[28] L. Bottou, F. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60:223–311, 2018.

[29] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004. ISBN 0521833787.

[30] K. Bredies, K. Kunisch, and T. Pock. Total generalized variation. *SIAM Journal on Imaging Sciences*, 3(3):492–526, 2010. doi: 10.1137/090769521. URL https://doi.org/10.1137/090769521.

[31] M. Burger and S. Osher. Convergence rates of convex variational regularization. *Inverse Problems*, 20(5):1411–1421, jul 2004. doi: 10.1088/0266-5611/20/5/005. URL https://doi.org/10.1088%2F0266-5611%2F20%2F5%2F005.

[32] M. Burger, K. Papafitsoros, E. Papoutsellis, and C.-B. Schönlieb. Infimal convolution regularisation functionals of BV and Lp spaces. the case p=∞. In *System Modelling and Optimization*, 2015.

[33] M. Burger, K. Papafitsoros, E. Papoutsellis, and C.-B. Schönlieb. Infimal convolution regularisation functionals of BV and Lp spaces. part i: The finite p case. *Journal of Mathematical Imaging and Vision*, 55(3):343–369, 2016. doi: 10.1007/s10851-015-0624-6. URL https://doi.org/10.1007/s10851-015-0624-6.

[34] M. D. Butala, R. A. Frazin, Y. Chen, and F. Kamalabadi. Tomographic imaging of dynamic objects with the ensemble Kalman filter. *IEEE Transactions on Image Processing*, 18(7):1573–1587, July 2009. ISSN 1057-7149. doi: 10.1109/TIP.2009.2017996.

[35] F. Cakoni and D. Colton. A survey in mathematics for industry: Open problems in the qualitative approach to inverse electromagnetic scattering theory.

European Journal of Applied Mathematics, 16(3):411–425, 2005. doi: 10.1017/S0956792505005978.

[36] L. Calatroni, C. Chung, J. C. de los Reyes, C.-B. Schönlieb, and T. Valkonen. Bilevel approaches for learning of variational imaging models. *ArXiv*, abs/1505.02120, 2015.

[37] A. Carrassi, M. Bocquet, L. Bertino, and G. Evensen. Data assimilation in the geosciences: An overview of methods, issues, and perspectives. *WIREs Climate Change*, 9(5):e535, 2018. doi: 10.1002/wcc.535. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/wcc.535`.

[38] J. Carrillo, K. Craig, and F. Patacchini. A blob method for diffusion. *Calculus of Variations and Partial Differential Equations*, 58:1–53, 2019.

[39] N. K. Chada and X. T. Tong. Convergence acceleration of ensemble Kalman inversion in nonlinear settings. *ArXiv*, abs/1911.02424, 2019.

[40] N. K. Chada, M. A. Iglesias, L. Roininen, and A. M. Stuart. Parameterizations for ensemble Kalman inversion. *Inverse Problems*, 34(5):055009, 2018. URL `http://stacks.iop.org/0266-5611/34/i=5/a=055009`.

[41] N. K. Chada, A. M. Stuart, and X. T. Tong. Tikhonov regularization within ensemble Kalman inversion. *SIAM Journal on Numerical Analysis*, 58(2):1263–1294, 2020. doi: 10.1137/19M1242331. URL `https://doi.org/10.1137/19M1242331`.

[42] K. Chadan, D. Colton, L. Päivärinta, and W. Rundell. *An Introduction to Inverse Scattering and Inverse Spectral Problems*. Society for Industrial and Applied Mathematics, 1997. doi: 10.1137/1.9780898719710. URL `https://epubs.siam.org/doi/abs/10.1137/1.9780898719710`.

[43] A. Chambolle and T. Pock. An introduction to continuous optimization for imaging. *Acta Numerica*, 25:161–319, 2016. doi: 10.1017/S096249291600009X.

[44] A. Chambolle, V. Caselles, D. Cremers, M. Novaga, and T. Pock. *An Introduction to Total Variation for Image Analysis*, pages 263 – 340. De Gruyter, Berlin, Boston, 16 Jul. 2010. ISBN 9783110226157. doi: https://doi.org/10.1515/9783110226157.263. URL `https://www.degruyter.com/view/book/9783110226157/10.1515/9783110226157.263.xml`.

[45] Y. Chen and D. S. Oliver. Ensemble randomized maximum likelihood method as an iterative ensemble smoother. *Mathematical Geosciences*, 44(1):1–26, 2012. doi: 10.1007/s11004-011-9376-z. URL `https://doi.org/10.1007/s11004-011-9376-z`.

[46] A. Chernov, H. Hoel, K. Law, F. Nobile, and R. Tempone. Multilevel ensemble Kalman filtering for spatially extended models. *ArXiv e-prints*, Aug. 2016.

[47] J. Chung and M. I. Español. Learning regularization parameters for general-form Tikhonov. *Inverse Problems*, 33(7):074004, jun 2017. doi: 10.1088/1361-6420/33/7/074004. URL `https://doi.org/10.1088%2F1361-6420%2F33%2F7%2F074004`.

[48] K. A. Cliffe, M. B. Giles, R. Scheichl, and A. L. Teckentrup. Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients. *Comput. Vis. Sci.*, 14(1):3–15, Jan. 2011. ISSN 1432-9360.

[49] A. Cohen, R. DeVore, and C. Schwab. Convergence rates of best n-term Galerkin approximations for a class of elliptic sPDEs. *Foundations of Computational Mathematics*, 10(6):615–646, 2010. doi: 10.1007/s10208-010-9072-2. URL https://doi.org/10.1007/s10208-010-9072-2.

[50] A. Cohen, R. A. DeVore, and C. Schwab. Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDEs. *Analysis and Applications*, 09:11–47, 2011.

[51] D. Colton, R. Ewing, and W. Rundell. *Inverse Problems in Partial Differential Equations*. Proceedings in Applied Mathematics. Society for Industrial and Applied Mathematics, 1990. ISBN 9780898712520. URL https://books.google.de/books?id=eAi3GJ2x4i8C.

[52] S. L. Cotter, M. Dashti, J. C. Robinson, and A. M. Stuart. Bayesian inverse problems for functions and applications to fluid mechanics. *Inverse Problems*, 25(11):115008, oct 2009. doi: 10.1088/0266-5611/25/11/115008. URL https://doi.org/10.1088%2F0266-5611%2F25%2F11%2F115008.

[53] S. L. Cotter, M. Dashti, and A. M. Stuart. Approximation of bayesian inverse problems for PDEs. *SIAM Journal on Numerical Analysis*, 48(1):322–345, 2010. doi: 10.1137/090770734. URL https://doi.org/10.1137/090770734.

[54] S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White. MCMC methods for functions: Modifying old algorithms to make them faster. *Statist. Sci.*, 28(3):424–446, 08 2013. doi: 10.1214/13-STS421. URL https://doi.org/10.1214/13-STS421.

[55] N. Couellan and W. Wang. Bi-level stochastic gradient for large scale support vector machine. *Neurocomputing*, 153:300 – 308, 2015. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2014.11.025. URL http://www.sciencedirect.com/science/article/pii/S0925231214015586.

[56] N. Couellan and W. Wang. Uncertainty-safe large scale support vector machines. *Computational Statistics & Data Analysis*, 109:215 – 230, 2017. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2016.12.008. URL http://www.sciencedirect.com/science/article/pii/S0167947316302985.

[57] N. Couellan and W. Wang. On the convergence of a stochastic approximation method for structured bi-level optimization. Nov. 2018. URL https://hal.archives-ouvertes.fr/hal-01932372. working paper or preprint.

[58] M. Dashti. Besov priors for Bayesian inverse problems. *Inverse Problems & Imaging*, 6(2):183–200, 2012. ISSN 1930-8337. doi: 10.3934/ipi.2012.6.183. URL http://aimsciences.org//article/id/900b5bce-4a6c-4da1-8b93-0b87da3f7df9.

[59] M. Dashti and A. M. Stuart. Uncertainty quantification and weak approximation of an elliptic inverse problem. *SIAM Journal on Numerical Analysis*, 49(6):2524–2542, 2011. doi: 10.1137/100814664. URL https://doi.org/10.1137/100814664.

[60] M. Dashti and A. M. Stuart. *The Bayesian Approach to Inverse Problems*, pages 311–428. Springer International Publishing, Cham, 2017. ISBN 978-3-319-12385-1. doi: 10.1007/978-3-319-12385-1_7. URL https://doi.org/10.1007/978-3-319-12385-1_7.

[61] M. Dashti, K. J. H. Law, A. M. Stuart, and J. Voss. MAP estimators and their consistency in Bayesian nonparametric inverse problems. *Inverse Problems*, 29(9): 095017, sep 2013. doi: 10.1088/0266-5611/29/9/095017. URL `https://doi.org/10.1088%2F0266-5611%2F29%2F9%2F095017`.

[62] F. Daum and J. Huang. Particle filter for nonlinear filters. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5920–5923, 2011.

[63] J. C. De los Reyes, C. B. Schönlieb, and T. Valkonen. Bilevel parameter learning for higher-order total variation regularisation models. *Journal of Mathematical Imaging and Vision*, 57(1):1–25, 2017. doi: 10.1007/s10851-016-0662-8. URL `https://doi.org/10.1007/s10851-016-0662-8`.

[64] J. de Wiljes, S. Reich, and W. Stannat. Long-time stability and accuracy of the ensemble Kalman–bucy filter for fully observed processes and small measurement noise. *SIAM Journal on Applied Dynamical Systems*, 17(2):1152–1181, 2018. doi: 10.1137/17M1119056.

[65] K. Deckelnick, C. M. Elliott, and V. Styles. Numerical analysis of an inverse problem for the eikonal equation. *Numerische Mathematik*, 119(2):245, 2011. doi: 10.1007/s00211-011-0386-z. URL `https://doi.org/10.1007/s00211-011-0386-z`.

[66] P. Degond and F.-J. Mustieles. A deterministic approximation of diffusion equations using particles. *SIAM J. Sci. Comput.*, 11:293–310, 1990.

[67] P. Del Moral and J. Tugaut. On the stability and the uniform propagation of chaos properties of ensemble Kalman Bucy filters. *The Annals of Applied Probability*, 28 (2):790–850, 04 2018. doi: 10.1214/17-AAP1317.

[68] M. D'Elia, J. C. de los Reyes, and A. Trujillo. Bilevel parameter optimization for nonlocal image denoising models. *arXiv: Optimization and Control*, 2019.

[69] Z. Ding and Q. Li. Ensemble Kalman inversion: mean-field limit and convergence analysis. *arXiv: Numerical Analysis*, 2019.

[70] Z. Ding and Q. Li. Ensemble Kalman sampling: mean-field limit and convergence analysis. *ArXiv*, abs/1910.12923, 2019.

[71] Z. Ding, Q. Li, and J. Lu. Ensemble Kalman inversion for nonlinear problems: weights, consistency, and variance bounds. *ArXiv*, abs/2003.02316, 2020.

[72] D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224, 1995. ISSN 01621459. URL `http://www.jstor.org/stable/2291512`.

[73] S. Duane, A. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216 – 222, 1987. ISSN 0370-2693. doi: https://doi.org/10.1016/0370-2693(87)91197-X. URL `http://www.sciencedirect.com/science/article/pii/037026938791197X`.

[74] A. Duncan, N. Nüsken, and L. Szpruch. On the geometry of Stein variational gradient descent. *ArXiv*, abs/1912.00894, 2019.

[75] A. A. Emerick and A. C. Reynolds. Investigation of the sampling performance of ensemble-based methods with a simple reservoir model. *Computational Geosciences*, 17:325–350, 2013.

[76] H. Engl, M. Hanke, and G. Neubauer. *Regularization of Inverse Problems*. Mathematics and Its Applications. Springer Netherlands, 1996. ISBN 9780792341574. URL https://books.google.de/books?id=DF7R_fVLuM8C.

[77] H. W. Engl. On the choice of the regularization parameter for iterated Tikhonov regularization of iii-posed problems. *Journal of Approximation Theory*, 49(1):55 – 63, 1987. ISSN 0021-9045. doi: https://doi.org/10.1016/0021-9045(87)90113-4. URL http://www.sciencedirect.com/science/article/pii/0021904587901134.

[78] H. W. Engl, K. Kunisch, and A. Neubauer. Convergence rates for Tikhonov regularisation of non-linear ill-posed problems. *Inverse Problems*, 5(4):523–540, aug 1989. doi: 10.1088/0266-5611/5/4/007. URL https://doi.org/10.1088%2F0266-5611%2F5%2F4%2F007.

[79] O. G. Ernst, B. Sprungk, and H.-J. Starkloff. Analysis of the ensemble and polynomial chaos Kalman filters in Bayesian inverse problems. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1):823–851, 2015. doi: 10.1137/140981319.

[80] G. Evensen. The Ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean Dynamics*, 53(4):343–367, Nov 2003.

[81] J. Fan, C. Ma, and Y. Zhong. A selective overview of deep learning. *ArXiv*, abs/1904.05526, 2019.

[82] L. Franceschi, P. Frasconi, S. Salzo, R. Grazzi, and M. Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1568–1577, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/franceschi18a.html.

[83] A. Garbuno-Inigo, F. Hoffmann, W. Li, and A. M. Stuart. Interacting Langevin diffusions: Gradient structure and ensemble Kalman sampler. *SIAM Journal on Applied Dynamical Systems*, 19(1):412–441, 2020. doi: 10.1137/19M1251655. URL https://doi.org/10.1137/19M1251655.

[84] L. Gawarecki. *Stochastic Differential Equations in Infinite Dimensions with Applications to Stochastic Partial Differential Equations*. Probability and Its Applications. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 9783642161940.

[85] A. E. Gelfand and A. F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990. ISSN 01621459. URL http://www.jstor.org/stable/2289776.

[86] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, 1984.

[87] M. B. Giles. Multilevel Monte Carlo methods. *Acta Numerica*, 24:259–328, 2015. doi: 10.1017/S096249291500001X.

[88] J. Goodman and J. Weare. Ensemble samplers with affine invariance. *Comm. Appl. Math. and Comput. Science*, 5:65–80, 2010.

[89] M. Grasmair. Linear convergence rates for Tikhonov regularization with positively homogeneous functionals. *Inverse Problems*, 27(7):075014, jun 2011. doi: 10.1088/0266-5611/27/7/075014. URL https://doi.org/10.1088%2F0266-5611%2F27%2F7%2F075014.

[90] C. Groetsch and J. King. Extrapolation and the method of regularization for generalized inverses. *Journal of Approximation Theory*, 25(3):233 – 247, 1979. ISSN 0021-9045. doi: https://doi.org/10.1016/0021-9045(79)90014-5. URL http://www.sciencedirect.com/science/article/pii/0021904579900145.

[91] J. Hadamard. Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton University Bulletin*, pages 49–52, 1902.

[92] M. Hairer, A. M. Stuart, and J. Voss. Analysis of SPDEs arising in path sampling part ii: The nonlinear case. *Ann. Appl. Probab.*, 17(5-6):1657–1706, 10 2007. doi: 10.1214/07-AAP441. URL https://doi.org/10.1214/07-AAP441.

[93] M. Hanke, A. Neubauer, and O. Scherzer. A convergence analysis of the Landweber iteration fornonlinear ill-posed problems. *Numerische Mathematik*, 72(1):21–37, 1995. doi: 10.1007/s002110050158. URL https://doi.org/10.1007/s002110050158.

[94] P. C. Hansen. The truncated svd as a method for regularization. *BIT Numerical Mathematics*, 27(4):534–553, 1987. doi: 10.1007/BF01937276. URL https://doi.org/10.1007/BF01937276.

[95] P. C. Hansen. Analysis of discrete ill-posed problems by means of the L-curve. *SIAM Review*, 34(4):561–580, 1992. doi: 10.1137/1034115. URL https://doi.org/10.1137/1034115.

[96] J. Harlim. *Data–driven computational methods*. Cambridge University Press, Cambridge, 2018.

[97] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970. ISSN 00063444. URL http://www.jstor.org/stable/2334940.

[98] L. Herrmann, C. Schwab, and J. Zech. Deep ReLU neural network expression rates for data-to-QoI maps in Bayesian PDE inversion. Technical Report 2020-02, Seminar for Applied Mathematics, ETH Zürich, Switzerland, 2020. URL https://www.sam.math.ethz.ch/sam_reports/reports_final/reports2020/2020-02.pdf.

[99] M. Herty and G. Visconti. Kinetic methods for inverse problems. *Kinetic & Related Models*, 12(1937-5093 2019 5 1109):1109, 2019. ISSN 1937-5093. doi: 10.3934/krm.2019042. URL http://aimsciences.org//article/id/33f5318c-8118-4246-8d4b-30fd7af4bb1e.

[100] M. Herty and G. Visconti. Continuous limits for constrained ensemble Kalman filter. *Inverse Problems*, 36(7):075006, jul 2020. doi: 10.1088/1361-6420/ab8bc5. URL https://doi.org/10.1088%2F1361-6420%2Fab8bc5.

[101] C. F. Higham and D. J. Higham. Deep learning: An introduction for applied mathematicians. *SIAM Review*, 61(4):860–891, 2019. doi: 10.1137/18M1165748. URL https://doi.org/10.1137/18M1165748.

[102] H. Hoel, K. Law, and R. Tempone. Multilevel ensemble Kalman filtering. *SIAM Journal on Numerical Analysis*, 54(3):1813–1839, 2016. doi: 10.1137/15M100955X. URL https://doi.org/10.1137/15M100955X.

[103] B. Hofmann, B. Kaltenbacher, C. Pöschl, and O. Scherzer. A convergence rates result for Tikhonov regularization in Banach spaces with non-smooth operators. *Inverse Problems*, 23(3):987–1010, apr 2007. doi: 10.1088/0266-5611/23/3/009. URL https://doi.org/10.1088%2F0266-5611%2F23%2F3%2F009.

[104] G. Holler, K. Kunisch, and R. C. Barnard. A bilevel approach for parameter learning in inverse problems. *Inverse Problems*, 34(11):115012, sep 2018. doi: 10.1088/1361-6420/aade77. URL https://doi.org/10.1088%2F1361-6420%2Faade77.

[105] J. Hu, K. Fennel, J. P. Mattern, and J. Wilkin. Data assimilation with a local ensemble Kalman filter applied to a three-dimensional biological model of the middle atlantic bight. *Journal of Marine Systems*, 94:145 – 156, 2012. ISSN 0924-7963. doi: https://doi.org/10.1016/j.jmarsys.2011.11.016.

[106] P. Hungerländer. Regularization of inverse problems via box constrained minimization. *Inverse Problems & Imaging*, 14(1930-8337 2020 3 437):437, 2020. ISSN 1930-8337. doi: 10.3934/ipi.2020021. URL http://aimsciences.org//article/id/c8b66b0b-99fb-4047-bd83-d2dd46fd0707.

[107] K. V. Huynh and B. Kaltenbacher. Some application examples of minimization based formulations of inverse problems and their regularization. *Inverse Problems & Imaging*, 0(1930-8337 2020 0 31), 2020. ISSN 1930-8337. doi: 10.3934/ipi.2020074. URL http://aimsciences.org//article/id/f6a19d4f-7cc3-4d49-a7d3-6e6a318f3035.

[108] M. Iglesias, M. Park, and M. V. Tretyakov. Bayesian inversion in resin transfer molding. *Inverse Problems*, 34(10):105002, jul 2018. doi: 10.1088/1361-6420/aad1cc. URL https://doi.org/10.1088%2F1361-6420%2Faad1cc.

[109] M. A. Iglesias. Iterative regularization for ensemble data assimilation in reservoir models. *Computational Geosciences*, 19(1):177–212, Feb 2015. ISSN 1573-1499. doi: 10.1007/s10596-014-9456-5.

[110] M. A. Iglesias. A regularizing iterative ensemble Kalman method for PDE-constrained inverse problems. *Inverse Problems*, 32(2):025002, jan 2016. doi: 10.1088/0266-5611/32/2/025002. URL https://doi.org/10.1088%2F0266-5611%2F32%2F2%2F025002.

[111] M. A. Iglesias and Y. Yang. Adaptive regularisation for ensemble Kalman inversion with applications to non-destructive testing and imaging. *ArXiv*, abs/2006.14980, 2020.

[112] M. A. Iglesias, K. J. H. Law, and A. M. Stuart. Ensemble Kalman methods for inverse problems. *Inverse Problems*, 29(4):045001, mar 2013. doi: 10.1088/0266-5611/29/4/045001. URL https://doi.org/10.1088%2F0266-5611%2F29%2F4%2F045001.

[113] A. Inigo, N. Nüsken, and S. Reich. Affine invariant interacting Langevin dynamics for Bayesian inference. Technical report, University of Potsdam, 2019.

[114] V. Isakov. On inverse problems in secondary oil recovery. *European Journal of Applied Mathematics*, 19:459–478, 2008.

[115] S. Jenni and P. Favaro. Deep bilevel learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[116] R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29:1–17, 1998.

[117] C. Kahane, J. Kahane, and B. Bollobas. *Some Random Series of Functions*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 1993. ISBN 9780521456029. URL https://books.google.de/books?id=mquwmHIdT3UC.

[118] J. Kaipio and E. Somersalo. *Statistical and computational inverse problems*. Applied mathematical sciences; Volume 160. New York, NY, 2010. ISBN 9781441919649.

[119] R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35–45, 03 1960. ISSN 0021-9223. doi: 10.1115/1.3662552. URL https://doi.org/10.1115/1.3662552.

[120] R. E. Kálmán and R. S. Bucy. New results in linear filtering and prediction theory. *Journal of Basic Engineering*, 83:95–108, 1961.

[121] E. Kalnay. *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, 2002. doi: 10.1017/CBO9780511802270.

[122] B. Kaltenbacher. Some Newton-type methods for the regularization of nonlinear ill-posed problems. *Inverse Problems*, 13(3):729–753, jun 1997. doi: 10.1088/0266-5611/13/3/012. URL https://doi.org/10.1088%2F0266-5611%2F13%2F3%2F012.

[123] B. Kaltenbacher. Regularization based on all-at-once formulations for inverse problems. *SIAM J. Numer. Anal.*, 54:2594–2618, 2016.

[124] B. Kaltenbacher. Minimization based formulations of inverse problems and their regularization. *SIAM Journal on Optimization*, 28(1):620–645, 2018. doi: 10.1137/17M1124036. URL https://doi.org/10.1137/17M1124036.

[125] B. Kaltenbacher, A. Neubauer, and O. Scherzer. *Iterative Regularization Methods for Nonlinear Ill-Posed Problems*. De Gruyter, Berlin, Boston, 2008. ISBN 978-3-11-020827-6. doi: https://doi.org/10.1515/9783110208276. URL https://www.degruyter.com/view/title/32605.

[126] B. Kaltenbacher, A. Kirchner, and B. Vexler. Adaptive discretizations for the choice of a Tikhonov regularization parameter in nonlinear inverse problems. *Inverse Problems*, 27(12):125008, nov 2011. doi: 10.1088/0266-5611/27/12/125008. URL https://doi.org/10.1088%2F0266-5611%2F27%2F12%2F125008.

[127] B. Kaltenbacher, A. Kirchner, and B. Vexler. Goal oriented adaptivity in the IRGNM for parameter identification in PDEs: II. all-at-once formulations. *Inverse Problems*, 30(4):045002, feb 2014. doi: 10.1088/0266-5611/30/4/045002. URL https://doi.org/10.1088%2F0266-5611%2F30%2F4%2F045002.

[128] R. Kandepu, L. Imsland, and B. A. Foss. Constrained state estimation using the unscented Kalman filter. *16th Mediterranean Conference on Control and Automation*, 2008. doi: 10.1109/MED.2008.4602001. URL https://ieeexplore.ieee.org/document/4602001.

[129] N. Kantas, A. Beskos, and A. Jasra. Sequential Monte Carlo methods for high-dimensional inverse problems: A case study for the Navier–Stokes equations. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):464–489, 2014. doi: 10.1137/130930364. URL https://doi.org/10.1137/130930364.

[130] I. Karatzas and S. Shreve. *Brownian Motion and Stochastic Calculus*. Graduate Texts in Mathematics. Springer New York, 1991. ISBN 9780387976556. URL https://books.google.de/books?id=ATNy_Zg3PSsC.

[131] I. Kasanick'y and J. Mandel. On well-posedness of Bayesian data assimilation and inverse problems in Hilbert space. *arXiv: Probability*, 2017.

[132] A. W. Kędzierawski. The inverse scattering problem for time-harmonic acoustic waves in an inhomogeneous medium with complex refraction index. *Journal of Computational and Applied Mathematics*, 47(1):83 – 100, 1993. ISSN 0377-0427. doi: https://doi.org/10.1016/0377-0427(93)90092-P. URL http://www.sciencedirect.com/science/article/pii/037704279390092P.

[133] D. Kelly, K. Law, and A. M. Stuart. Well-posedness and accuracy of the ensemble Kalman filter in discrete and continuous time. *Nonlinearity*, 27(10):2579, 2014. URL http://stacks.iop.org/0951-7715/27/i=10/a=2579.

[134] D. Kelly, A. J. Majda, and X. T. Tong. Nonlinear stability of the ensemble Kalman filter with adaptive covariance inflation. *ArXiv e-prints*, July 2015.

[135] D. Kelly, A. J. Majda, and X. T. Tong. Concrete ensemble Kalman filters with rigorous catastrophic filter divergence. *Proceedings of the National Academy of Sciences*, 2015. ISSN 0027-8424. doi: 10.1073/pnas.1511063112.

[136] D. Kelly, A. J. Majda, and X. T. Tong. Nonlinear stability and ergodicity of ensemble based Kalman filters. *Nonlinearity*, 29(2):657, 2016. URL http://stacks.iop.org/0951-7715/29/i=2/a=657.

[137] R. Z. Khasminskii. *Stochastic stability of differential equations. Transl. by D. Louvish. Ed. by S. Swierczkowski.* Monographs and Textbooks on Mechanics of Solids and Fluids. Mechanics: Analysis, 7. Alphen aan den Rijn, The Netherlands; Rockville, Maryland, USA. Sijthoff & Noordhoff, 1980.

[138] Y. KHOO, J. LU, and L. YING. Solving parametric PDE problems with artificial neural networks. *European Journal of Applied Mathematics*, page 1–15, 2020. doi: 10.1017/S0956792520000182.

[139] B. T. Knapik, A. W. van der Vaart, and J. H. van Zanten. Bayesian inverse problems with Gaussian priors. *The Annals of Statistics*, 39(5):2626–2657, 2011. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/41713591.

[140] N. B. Kovachki and A. M. Stuart. Ensemble Kalman inversion: a derivative-free technique for machine learning tasks. *Inverse Problems*, 35(9):095005, aug 2019. doi: 10.1088/1361-6420/ab1c3a. URL https://doi.org/10.1088%2F1361-6420%2Fab1c3a.

[141] C. Kravaris and J. H. Seinfeld. Identification of parameters in distributed parameter systems by regularization. In *The 22nd IEEE Conference on Decision and Control*, pages 50–55, 1983.

[142] G. Kutyniok, P. Petersen, M. Raslan, and R. Schneider. A theoretical analysis of deep neural networks and parametric PDEs. *arXiv preprint arXiv:1904.00377*, 2019.

[143] E. Kwiatkowski and J. Mandel. Convergence of the square root ensemble Kalman filter in the large ensemble limit. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1):1–17, 2015. doi: 10.1137/140965363.

[144] T. Lange and W. Stannat. On the continuous time limit of ensemble square root filters. *arXiv: Probability*, 2019.

[145] T. Lange and W. Stannat. On the continuous time limit of the ensemble Kalman filter. *Math. Comp.*, 90(327):233–265, 2021.

[146] M. Lassas. Discretization-invariant Bayesian inversion and Besov space priors. *Inverse Problems & Imaging*, 3(1):87–122, 2009. ISSN 1930-8337. doi: 10.3934/ipi.2009.3.87. URL http://aimsciences.org//article/id/2c169f6d-50f3-4e01-9fc0-24c836905147.

[147] K. Law, A. M. Stuart, and K. Zygalakis. *Data Assimilation: A Mathematical Introduction*. Texts in Applied Mathematics. Springer International Publishing, 2016. ISBN 9783319366876. URL https://books.google.de/books?id=hVJwtAEACAAJ.

[148] K. Law, H. Tembine, and R. Tempone. Deterministic mean-field ensemble Kalman filtering. *SIAM Journal on Scientific Computing*, 38(3):A1251–A1279, 2016. doi: 10.1137/140984415.

[149] F. Le Gland, V. Monbet, and V.-D. Tran. Large sample asymptotics for the ensemble Kalman filter. Research Report RR-7014, INRIA, 2009. URL https://hal.inria.fr/inria-00409060.

[150] B. Leimkuhler, C. Matthews, and J. Weare. Ensemble preconditioning for Markov chain Monte Carlo simulations. *Stat. Comput.*, 28:277–290, 2018.

[151] F. Lindgren, H. Rue, and J. Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011. doi: 10.1111/j.1467-9868.2011.00777.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2011.00777.x.

[152] J. Liu, J. Liu, and S. Jun. *Monte Carlo Strategies in Scientific Computing*. Springer Series in Statistics. Springer New York, 2001. ISBN 9780387952307. URL https://books.google.de/books?id=Dk_ou-gqnHQC.

[153] Q. Liu and D. Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 2378–2386, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

[154] G. J. Lord, C. E. Powell, and T. Shardlow. *An Introduction to Computational Stochastic PDEs*. Cambridge Texts in Applied Mathematics. Cambridge University Press, 2014. doi: 10.1017/CBO9781139017329.

[155] J. C. D. los Reyes. Image denoising: Learning the noise model via nonsmooth PDE-constrained optimization, 2013. ISSN 1930-8337. URL http://aimsciences.org//article/id/2bd95fb1-cc0f-48e7-8a05-46e5b1a62840.

[156] X.-p. Luo and J. Yang. Regularization and iterative methods for monotone inverse variational inequalities. *Optimization Letters*, 8(4):1261–1272, 2014. doi: 10.1007/s11590-013-0653-2. URL https://doi.org/10.1007/s11590-013-0653-2.

[157] A. J. Majda and J. Harlim. *Filtering Complex Turbulent Systems*. Cambridge University Press, 2012. doi: 10.1017/CBO9781139061308.

[158] A. J. Majda and X. T. Tong. Performance of ensemble Kalman filters in large dimensions. *Communications on Pure and Applied Mathematics*, 71(5):892–937, 2018. doi: 10.1002/cpa.21722.

[159] X. Mao. *Stochastic Differential Equations and Applications*. Horwood series in mathematics & applications. Horwood Pub., 2008. ISBN 9781904275343. URL https://books.google.de/books?id=vlaI_vmm5QMC.

[160] Y. Marzouk, T. Moselhy, M. Parno, and A. Spantini. *Sampling via Measure Transport: An Introduction*. Springer International Publishing, New York, 2016. ISBN 978-3-319-11259-6. doi: 10.1007/978-3-319-11259-6_23-1. URL https://doi.org/10.1007/978-3-319-11259-6_23-1.

[161] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.*, 21 (6):1087–1092, June 1953. doi: 10.1063/1.1699114.

[162] P. Moral. *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. Probability and Its Applications. Springer New York, 2004. ISBN 9780387202686. URL https://books.google.de/books?id=8LypfuG8ZLYC.

[163] V. A. Morozov. On the solution of functional equations by the method of regularization. *Dokl. Akad. Nauk SSSR*, 167(3):510–512, 1966.

[164] M. Z. Nashed and G. Wahba. Regularization and approximation of linear operator equations in reproducing kernel spaces. *Bulletin of the American Mathematical Society*, 80:1213–1218, 1974.

[165] M. Z. Nashed and G. Wahba. Generalized inverses in reproducing kernel spaces: An approach to regularization of linear operator equations. *SIAM Journal on Mathematical Analysis*, 5(6):974–987, 1974. doi: 10.1137/0505095. URL https://doi.org/10.1137/0505095.

[166] M. Z. Nashed and G. Wahba. Convergence rates of approximate least squares solutions of linear integral and operator equations of the first kind. *Mathematics of Computation*, 28(125):69–80, 1974. ISSN 00255718, 10886842. URL http://www.jstor.org/stable/2005817.

[167] F. Natterer and F. Wübbeling. *Mathematical Methods in Image Reconstruction*. Society for Industrial and Applied Mathematics, 2001. doi: 10.1137/1.9780898718324. URL https://epubs.siam.org/doi/abs/10.1137/1.9780898718324.

[168] A. Neubauer. An a posteriori parameter choice for Tikhonov regularization in Hilbert scales leading to optimal convergence rates. *SIAM Journal on Numerical Analysis*, 25(6):1313–1326, 1988. doi: 10.1137/0725074. URL https://doi.org/10.1137/0725074.

[169] E. Nummelin. *General Irreducible Markov Chains and Non-Negative Operators*. Cambridge Tracts in Mathematics. Cambridge University Press, 1984. doi: 10.1017/CBO9780511526237.

[170] N. Nüsken and S. Reich. Note on interacting Langevin diffusion: Gradient structure and ensemble Kalman sampler. Technical Report arXiv:1908.10890v1, University of Potsdam, 2019.

[171] B. Oksendal. *Stochastic Differential Equations: An Introduction with Applications*. Universitext. Springer Berlin Heidelberg, 2013. ISBN 9783662028476. URL https://books.google.de/books?id=_6_rCAAAQBAJ.

[172] D. S. Oliver, A. C. Reynolds, and N. Liu. *Inverse theory for petroleum reservoir characterization and history matching*. Cambridge University Press, 2008.

[173] D. S. Oliver, A. C. Reynolds, and N. Liu. *Inverse Theory for Petroleum Reservoir Characterization and History Matching*. Cambridge University Press, 2008. doi: 10.1017/CBO9780511535642.

[174] O. Papaspiliopoulos, G. O. Roberts, and M. Sköld. A general framework for the parametrization of hierarchical models. *Statist. Sci.*, 22(1):59–73, 02 2007. doi: 10.1214/088342307000000014. URL https://doi.org/10.1214/088342307000000014.

[175] S. Pathiraja and S. Reich. Discrete gradients for computational Bayesian inference. *Journal of Computational Dynamics*, 6:236–251, 2019.

[176] V. I. Paulsen and M. Raghupathi. *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2016. doi: 10.1017/CBO9781316219232.

[177] G. Pavliotis. *Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations*. Texts in Applied Mathematics. Springer New

York, 2014. ISBN 9781493913220. URL https://books.google.de/books?id=mpHFoAEACAAJ.

[178] L. E. Payne. *Improperly Posed Problems in Partial Differential Equations*. Society for Industrial and Applied Mathematics, 1975. doi: 10.1137/1.9781611970463. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611970463.

[179] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *Siam Journal on Control and Optimization*, 30:838–855, 1992.

[180] G. D. Prato. *An Introduction to Infinite-Dimensional Analysis*. Springer Berlin Heidelberg, 2006. ISBN 978-3-540-29021-6. doi: 10.1007/3-540-29021-4. URL https://link.springer.com/book/10.1007/3-540-29021-4.

[181] M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics informed deep learning (Part I): Data-driven solutions of nonlinear partial differential equations. *arXiv preprint arXiv:1711.10561*, 2017.

[182] M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics informed deep learning (Part II): Data-driven discovery of nonlinear partial differential equations. *arXiv preprint arXiv:1711.10566*, 2017.

[183] S. Reich. A dynamical systems framework for intermittent data assimilation. *BIT Numerical Mathematics*, 51(1):235–249, Mar 2011. ISSN 1572-9125. doi: 10.1007/s10543-010-0302-4. URL http://dx.doi.org/10.1007/s10543-010-0302-4.

[184] S. Reich. A dynamical systems framework for intermittent data assimilation. *BIT Numer Math*, 51:235–249, 2011.

[185] S. Reich. Data assimilation: The Schrödinger perspective. *Acta Numerica*, 28: 635–711, 2019. doi: 10.1017/S0962492919000011.

[186] S. Reich and C. Cotter. *Probabilistic Forecasting and Bayesian Data Assimilation*. Cambridge University Press, 2015. doi: 10.1017/CBO9781107706804.

[187] E. Resmerita. Regularization of ill-posed problems in Banach spaces: convergence rates. *Inverse Problems*, 21(4):1303–1314, jun 2005. doi: 10.1088/0266-5611/21/4/007. URL https://doi.org/10.1088%2F0266-5611%2F21%2F4%2F007.

[188] Ring, Wolfgang. Structural properties of solutions to total variation regularization problems. *ESAIM: M2AN*, 34(4):799–810, 2000. doi: 10.1051/m2an:2000104. URL https://doi.org/10.1051/m2an:2000104.

[189] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[190] G. O. Roberts and J. S. Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268, 1998. doi: 10.1111/1467-9868.00123. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00123.

[191] G. O. Roberts, A. Gelman, and W. R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.*, 7(1):110–120, 02 1997. doi: 10.1214/aoap/1034625254. URL https://doi.org/10.1214/aoap/1034625254.

[192] L. Roininen, J. M. J. Huttunen, and S. Lasanen. Whittle-Matérn priors for Bayesian statistical inversion with applications in electrical impedance tomography. *Inverse Problems & Imaging*, 8(1930-8337 2014 2 561):561, 2014. ISSN 1930-8337. doi: 10.3934/ipi.2014.8.561. URL http://aimsciences.org//article/id/466c718a-c026-4ea2-bdd5-699793c40890.

[193] M. Rudelson and R. Vershynin. Hanson-Wright inequality and sub-Gaussian concentration. *Electron. Commun. Probab.*, 18:9 pp., 2013. doi: 10.1214/ECP.v18-2865. URL https://doi.org/10.1214/ECP.v18-2865.

[194] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259 – 268, 1992. ISSN 0167-2789. doi: https://doi.org/10.1016/0167-2789(92)90242-F. URL http://www.sciencedirect.com/science/article/pii/016727899290242F.

[195] G. Russo. Deterministic diffusion of particles. *Comm. Pure Appl. Math.*, 43:697–733, 1990.

[196] N. D. Santo, S. Deparis, and L. Pegolotti. Data driven approximation of parametrized PDEs by reduced basis and neural networks. *ArXiv*, abs/1904.01514, 2020.

[197] V. Sazonov. A remark on characteristic functionals. *Theory of Probability & Its Applications*, 3(2):188–192, 1958. doi: 10.1137/1103018. URL https://doi.org/10.1137/1103018.

[198] O. Scherzer. Convergence rates of iterated Tikhonov regularized solutions of nonlinear iii —posed problems. *Numerische Mathematik*, 66(1):259–279, 1993. doi: 10.1007/BF01385697. URL https://doi.org/10.1007/BF01385697.

[199] R. L. Schilling. *Measures, Integrals and Martingales*. Cambridge University Press, 2005. doi: 10.1017/CBO9780511810886.

[200] C. Schillings and A. M. Stuart. Analysis of the ensemble Kalman filter for inverse problems. *SIAM Journal on Numerical Analysis*, 55(3):1264–1290, 2017. doi: 10.1137/16M105959X. URL https://doi.org/10.1137/16M105959X.

[201] C. Schillings and A. M. Stuart. Convergence analysis of ensemble Kalman inversion: the linear, noisy case. *Applicable Analysis*, 97(1):107–123, 2018. doi: 10.1080/00036811.2017.1386784.

[202] M. Schmidt, D. Kim, and S. Sra. *Projected Newton-type methods in machine learning*, pages 305–330. MIT Press, Cambridge, MA, USA, Dec. 2011.

[203] T. Schneider, S. Lan, A. Stuart, and J. Teixeira. Earth system modeling 2.0: A blueprint for models that learn from observations and targeted high-resolution simulations. *Geophysical Research Letters*, 44(24):12,396–12,417, 2017. doi: 10.1002/

2017GL076101. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017GL076101.

[204] T. Schuster, B. Kaltenbacher, B. Hofmann, and K. Kazimierski. *Regularization Methods in Banach Spaces*. Radon series on computational and applied mathematics. De Gruyter, 2012. ISBN 9783110255249. URL https://books.google.de/books?id=oBnoygAACAAJ.

[205] C. Schwab. QMC Galerkin discretization of parametric operator equations. In J. Dick, F. Y. Kuo, G. W. Peters, and I. H. Sloan, editors, *Monte Carlo and Quasi-Monte Carlo Methods 2012*, pages 613–629, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-41095-6.

[206] C. Schwab and A. M. Stuart. Sparse deterministic approximation of Bayesian inverse problems. *Inverse Problems*, 28(4):045003, mar 2012. doi: 10.1088/0266-5611/28/4/045003. URL https://doi.org/10.1088%2F0266-5611%2F28%2F4%2F045003.

[207] C. Schwab and J. Zech. Deep learning in high dimension: Neural network expression rates for generalized polynomial chaos expansions in UQ. *Analysis and Applications*, 17(01):19–55, 2019. doi: 10.1142/S0219530518500203. URL https://doi.org/10.1142/S0219530518500203.

[208] D. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. A Wiley-interscience publication. Wiley, 1992. ISBN 9780471547709. URL https://books.google.de/books?id=7crCUS_F2ocC.

[209] J. Sethian. *Level Set Methods and Fast Marching Methods: Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 1999. ISBN 9780521645577. URL https://books.google.de/books?id=ErpOoynE4dIC.

[210] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014. doi: 10.1017/CBO9781107298019.

[211] V. Shikhman and O. Stein. Constrained optimization: Projected gradient flows. *Journal of Optimization Theory and Applications*, 140:117–130, 09 2009. doi: 10.1007/s10957-008-9445-8.

[212] D. Simon. Kalman filtering with state constraints: a survey of linear and nonlinear algorithms. *IET Control Theory Applications*, 4(8):1303–1318, 2010.

[213] L. D. Simon, M. Iglesias, B. Jones, and C. Wood. Quantifying uncertainty in thermophysical properties of walls by means of Bayesian inversion. *Energy and Buildings*, 177:220 – 245, 2018. ISSN 0378-7788. doi: https://doi.org/10.1016/j.enbuild.2018.06.045. URL http://www.sciencedirect.com/science/article/pii/S0378778817334035.

[214] J. T. Slagel, J. Chung, M. Chung, D. Kozak, and L. Tenorio. Sampled Tikhonov regularization for large linear inverse problems. *Inverse Problems*, 35(11):114008, oct 2019. doi: 10.1088/1361-6420/ab2787. URL https://doi.org/10.1088%2F1361-6420%2Fab2787.

[215] A. M. Stuart. Inverse problems: A Bayesian perspective. *Acta Numerica*, 19:451–559, 2010. doi: 10.1017/S0962492910000061.

[216] T. J. Sullivan. *Introduction to Uncertainty Quantification*, volume 63. Springer International Publishing, 2015. ISBN 978-3-319-23394-9. doi: 10.1007/978-3-319-23395-6.

[217] A. Taghvaei and P. Mehta. Accelerated flow for probability distributions. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6076–6085, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL `http://proceedings.mlr.press/v97/taghvaei19a.html`.

[218] A. Taghvaei and P. G. Mehta. Gain function approximation in the feedback particle filter. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 5446–5452, 2016.

[219] A. Tarantola and B. Valette. Inverse problems - quest for information. *Journal of Geophysics*, pages 159–170, 1982.

[220] L. Tierney. Markov chains for exploring posterior distributions. *Ann. Statist.*, 22 (4):1701–1728, 12 1994. doi: 10.1214/aos/1176325750. URL `https://doi.org/10.1214/aos/1176325750`.

[221] L. Tierney. A note on Metropolis-Hastings kernels for general state spaces. *Ann. Appl. Probab.*, 8(1):1–9, 02 1998. doi: 10.1214/aoap/1027961031. URL `https://doi.org/10.1214/aoap/1027961031`.

[222] X. Tong, A. Majda, and D. Kelly. Nonlinear stability of the ensemble Kalman filter with adaptive covariance inflation. *Communications in Mathematical Sciences*, 14 (5):1283–1313, 2016. ISSN 1539-6746. doi: 10.4310/CMS.2016.v14.n5.a5.

[223] X. T. Tong. Performance analysis of local ensemble Kalman filter. *Journal of Nonlinear Science*, 28(4):1397–1442, Aug 2018. ISSN 1432-1467. doi: 10.1007/s00332-018-9453-2. URL `https://doi.org/10.1007/s00332-018-9453-2`.

[224] S. J. Vollmer. Posterior consistency for Bayesian inverse problems through stability and regression results. *Inverse Problems*, 29(12):125011, nov 2013. doi: 10.1088/0266-5611/29/12/125011. URL `https://doi.org/10.1088%2F0266-5611%2F29%2F12%2F125011`.

[225] D. Wang, Y. Chen, and X. Cai. State and parameter estimation of hydrologic models using the constrained ensemble Kalman filter. *Monthly Resources Research*, 45(11), 2009. doi: 10.1029/2008WR007401.

[226] L. Wasserman. *All of Nonparametric Statistics*. Springer Texts in Statistics. Springer New York, 2006. ISBN 9780387306230. URL `https://books.google.de/books?id=MRFlzQfRg7UC`.

[227] M. Welling and Y. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning, ICML'11*, pages 681–688. Omnipress, 2011.

[228] L. Yang, X. Meng, and G. E. Karniadakis. B-PINNs: Bayesian physics-informed neural networks for forward and inverse PDE problems with noisy data. *arXiv preprint arXiv:2003.06097v1*, 2020.

[229] D. Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114, 2017. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2017.07.002. URL `http://www.sciencedirect.com/science/article/pii/S0893608017301545`.

[230] Y. Zhang, N. Liu, and D. Oliver. Ensemble filter methods with perturbed observations applied to nonlinear problems. *Comput Geosciences*, 14(2), 2010.