CONSISTENCY ANALYSIS OF
BILEVEL DATA-DRIVEN LEARNING IN INVERSE PROBLEMS*

NEIL K. CHADA†, CLAUDIA SCHILLINGS‡, XIN T. TONG§, AND SIMON WEISSMANN¶

Abstract. One fundamental problem when solving inverse problems is how to find regularization parameters. This article considers solving this problem using data-driven bilevel optimization, i.e. we consider the adaptive learning of the regularization parameter from data by means of optimization. This approach can be interpreted as solving an empirical risk minimization problem, and we analyze its performance in the large data sample size limit for general nonlinear problems. We demonstrate how to implement our framework on linear inverse problems, where we can further show that the inverse accuracy does not depend on the ambient space dimension. To reduce the associated computational cost, online numerical schemes are derived using the stochastic gradient descent method. We prove convergence of these numerical schemes under suitable assumptions on the forward problem. Numerical experiments are presented illustrating the theoretical results and demonstrating the applicability and efficiency of the proposed approaches for various linear and nonlinear inverse problems, including Darcy flow, the eikonal equation, and an image denoising example.

Keywords. Bilevel optimization; statistical consistency; inverse problems; stochastic gradient descent; Tikhonov regularization.

AMS subject classifications. 35R30; 90C15; 62F12; 65K10.

1. Introduction

Data-driven modeling seeks to improve model accuracy and predictability by exploiting information from existing data. It has lead to a wide range of successes in deep learning, reinforcement learning, natural language processing and other areas [30,32,52]. This article is interested in its applications when solving inverse problems. Mathematically speaking, when solving inverse problems, we try to recover a $u \in \mathcal{U}$ from a perturbed data $y \in \mathcal{Y}$ where their relationship is given as

$$y = \mathcal{G}(u) + \eta. \tag{1.1}$$

In (1.1), η denotes an additive observational noise and $\mathcal{G} : \mathcal{U} \to \mathcal{Y}$ is the mapping from the parameter space $\mathcal{U}$ to the observation space $\mathcal{Y}$. Here, $\mathcal{U}$ and $\mathcal{Y}$ denote possibly infinite dimensional Banach spaces. Solutions to inverse problems have been well-studied through the use of variational and optimization methods which are well-documented in the following texts [5,28,54].

Regularization is an important aspect of the numerical treatment of inverse problems. It helps to overcome the ill-posedness problem in theory and the overfitting phenomenon in practice. It can also be interpreted as a form of a-priori knowledge in the Bayesian approach [36,53]. To implement regularization on (1.1), we estimate the

†Applied Mathematics and Computational Science Program, King Abdullah University of Science and Technology, Thuwal, 23955, KSA (neilchada123@gmail.com).

‡Institute of Mathematics, University of Mannheim, 68131 Mannheim, Germany (c.schillings@uni-mannheim.de).

§Department of Mathematics, National University of Singapore, 119077, Singapore (mattxin@nus.edu.sg).

¶Interdisciplinary Center for Scientific Computing, University of Heidelberg, 69120 Heidelberg, Germany (simon.weissmann@uni-heidelberg.de).

unknown parameter u by minimizing a regularized loss function, i.e. we consider

$$u_\lambda := \underset{u \in \mathcal{U}}{\arg\min} \ \mathcal{L}_\mathcal{Y}(\mathcal{G}(u), y) + \mathcal{S}_\lambda(u), \quad \lambda \in \mathbf{R}_+, \tag{1.2}$$

where $\mathcal{L}_\mathcal{Y} : \mathcal{Y} \times \mathcal{Y} \to \mathbf{R}_+$ is some metric in $\mathcal{Y}$ and $S_\lambda : \mathcal{U} \to \mathbf{R}_+$ is a regularization function with regularization parameter $\lambda > 0$. A common choice is Tikhonov regularization [55] which can be included in (1.2) through the penalty term $S_\lambda(u) = \frac{\lambda}{2}\|u\|_\mathcal{U}^2$. The choice of norm $\|\cdot\|_\mathcal{U}$ often models prior information on the unknown parameter. Other common forms include L_1 and total variation regularization, which are particularly useful for imaging purposes [5, 28, 41].

In (1.2), the parameter λ balances the influence of the data and the a-priori knowledge via the regularization. While expert knowledge can often provide a rough range of λ, the exact value, i.e. the λ leading to the best estimation of the unknown parameter u, is often difficult to determine. However, the parameter λ strongly influences the accuracy of the estimate and has to be properly chosen. Bilevel optimization is one way to resolve this issue [20, 24, 50, 54]. It seeks to learn the regularization parameter in a variational manner, and it can be viewed as a data–driven regularization [2]. To formulate this approach, we view the unknown parameter $U \in \mathcal{U}$ and the data $Y \in \mathcal{Y}$ in the model (1.1) as a jointly distributed random variable with distribution $\mu_{(U,Y)}$. To find the best possible regularization parameter of the model (1.1), the bilevel minimization seeks to solve

$$\begin{aligned} \lambda_* &= \underset{\lambda > 0}{\arg\min} \ F(\lambda), \quad F(\lambda) = \mathbb{E}_{\mu_{(U,Y)}}[\mathcal{L}_\mathcal{U}(u_\lambda(Y), U)], \quad \text{(upper level)} \\ u_\lambda(Y) &:= \underset{u \in \mathcal{U}}{\arg\min} \ \mathcal{L}_\mathcal{Y}(\mathcal{G}(u), Y) + \mathcal{S}_\lambda(u), \quad\quad\quad\quad \text{(lower level)} \end{aligned} \tag{1.3}$$

where $\mathcal{L}_\mathcal{U} : \mathcal{U} \times \mathcal{U} \to \mathbf{R}_+$ is some metric in the parameter space $\mathcal{U}$. The upper level problem seeks to minimize the distance between the unknown parameter U and the regularized solution corresponding to its data Y, which is computed through $u_\lambda(Y)$ in the lower level problem. To solve this (stochastic) bilevel optimization problem, we assume that we have access to training data, given through samples of $(U_i, Y_i) \sim \mu_{(U,Y)}$, and the function F in (1.3) can be approximated by its empirical Monte–Carlo approximation. The area of bilevel optimization has been applied to various methodologies for inverse problems. To motivate this, we provide various examples of the application of bilevel optimization, in the setting described by (1.3), to inverse problems and an overview of recent literature.

1.1. Motivating Examples.

1.1.1. Example 1 - PDE-constrained inverse problems.   We first consider an inverse problem (1.1) with the lower level problem formulated by a partial differential equation (PDE):

$$\begin{aligned} &\underset{u \in \mathcal{U}}{\arg\min} \ \mathcal{L}_\mathcal{Y}(\mathcal{O}(p), y) + \mathcal{S}_\lambda(u), \\ &\text{s.t. } M(u, p) = 0, \end{aligned} \tag{1.4}$$

where $u \in \mathcal{U}$ denotes the unknown parameter and $p \in \mathcal{V}$ is the state. The function $M : \mathcal{U} \times \mathcal{V} \to \mathcal{W}$ describes an underlying ODE or PDE model. The operator $\mathcal{O} : \mathcal{V} \to \mathbf{R}^K$ denotes the observation operator which maps the state p to finite dimensional observations. The Darcy's flow problem is one such example. In particular, u describes a subsurface

structure, p is the corresponding pressure field, M describes the Darcy's law, and $\mathcal{O}$ evaluates p at different locations.

In order to formulate the corresponding bilevel problem (1.3), we assume that the forward model $M(u,p)=0$ is well-posed, which means that for each $u \in \mathcal{U}$ there exists a unique $p \in \mathcal{V}$ such that $M(u,p)=0 \in \mathcal{W}$. Hence, using the solution operator $G:\mathcal{U} \to \mathcal{V}$ s.t. $M(u,G(u))=0$, we can formulate the reduced problem of (1.4) by

$$\operatorname*{arg\,min}_{u \in \mathcal{U}} \ \mathcal{L}_{\mathcal{Y}}(\mathcal{G}(u),y) + \mathcal{S}_\lambda(u), \tag{1.5}$$

where we have defined $\mathcal{G} = \mathcal{O} \circ G$. Hence, given a training data set $(u^{(j)},y^{(j)})$ we can also formulate the empirical bilevel minimization problem

$$\widehat{\lambda}_n := \operatorname*{arg\,min}_{\lambda > 0} \frac{1}{n} \sum_{j=1}^{n} \|u_\lambda(y^{(j)}) - u^{(j)}\|_2^2,$$

$$u_\lambda(y^{(j)}) := \operatorname*{arg\,min}_{u \in \mathcal{U}} \ \mathcal{L}_{\mathcal{Y}}(\mathcal{G}(u),y^{(j)}) + S_\lambda(u). \tag{1.6}$$

In terms of applications, many inverse problems arising in PDEs [3] are concerned with the recovery of an unknown which is heterogeneous. As a result it is very natural to model the unknown as Gaussian random fields. Such models include Darcy flow, the Navier–Stokes model [34] and electrical impedance tomography [6, 36]. Physical constraints such as boundary, or initial conditions are required for modeling purposes.

Holler et al. [33] consider bilevel optimization for inverse problems in the setting of (1.4). They provide theory which suggests existence of solutions and formulate their problem as an optimal control problem. This is connected with the work of Kaltenbacher [37,38] who provided a modified approach known as "all-at-once" inversion. These works have also been used in the context of deconvolution [15, 16, 51].

1.1.2. Example 2 - Image & signal processing.   Bilevel optimization is a popular solution choice for image processing problems [7, 39]. In these problems, one is interested in optimizing over an underlying image and particular areas/segments of that image. A common example of this includes image denoising which is to remove noise from an image. Another example is image deblurring where the image is commonly given as a convolution with a linear kernel $\mathcal{A}$, i.e.

$$y = \mathcal{A} * u + \eta,$$

where $*$ denotes the convolution of $\mathcal{A}$ and u, commonly expressed as

$$\mathcal{A} * u(x) = \int_{\mathbf{R}^d} \mathcal{A}(x - \tau) u(\tau) d\tau.$$

This inverse problem is also known as deconvolution. The setting of (1.3) is common for deconvolution, where their loss functions are given as

$$\widehat{\lambda}_n := \operatorname*{arg\,min}_{\lambda > 0} \frac{1}{n} \sum_{j=1}^{n} \|u_\lambda(y^{(j)}) - u^{(j)}\|_2^2,$$

$$u_\lambda(y^{(j)}) := \operatorname*{arg\,min}_{u \in \mathcal{U}} \ \mathcal{L}_{\mathcal{Y}}(\mathcal{A} * u, y^{(j)}) + \lambda \|Lu\|_2^2. \tag{1.7}$$

In (1.7), L is a regularization matrix, and the upper level problem is taken as the minimization of the empirical loss function given by a training data set $(u^{(j)},y^{(j)})$.

Commonly, λ is taken to be either a weighted function between $\mathcal{L}_{\mathcal{Y}}$ and the penalty term, or it can be viewed as the noise within a system. Common choices of L traditionally are $L = I$ or a first- or second-order operator, which can depend on the unknown or image of interest. Further detail on the choice of L and $\mathcal{A}$ are discussed in [7].

The work of De los Reyes, Schönlieb [10, 21–23] 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. Recent developments of these techniques using Bayesian methodologies for imaging can be found in [56, 57].

1.2. Our contributions. In this article, we investigate two different approaches to solve bilevel optimization and their performance on inverse problems. Firstly we formulate the offline approach of bilevel optimization as an empirical risk minimization (ERM) problem. Analyzing the performance of the ERM solution is difficult, since the loss function is random and non-convex, so numerical solutions often can only find local minima. We build a theoretical framework under these general settings, and are interested in consistency, i.e. when sample size goes to infinity. In particular, this translates to a convergence rate of the ERM solution. This framework is applied to linear inverse problems to understand the performance of bilevel optimization approach. Moreover, our results depend only on the effective dimension, but not the ambient space dimension, and hold accordingly, subject to appropriate convexity conditions. This is an important aspect in inverse problems since the underlying space can be of infinite dimension.

Secondly, we discuss how to implement stochastic gradient descent (SGD) methods on bilevel optimization. SGD is a popular optimization tool for empirical risk minimization because of its straightfoward implementation and efficiency. The low computational costs are particularly appealing in the bilevel context as finding the lower-level solution is already time consuming. Besides exact SGD, we also consider SGD with central difference approximation. This can be useful for problems with complicated forward observation models. A general consistency analysis framework is formulated for both exact SGD and approximated SGD. We demonstrate how to apply this framework to linear inverse problems.

Various numerical results are presented highlighting and verifying the theory acquired. Our results are firstly presented on various partial differential equations, both linear and nonlinear which include Darcy flow and the eikonal equation, as motivated through Example 1 in Subsection 1.1.1. We also test our theory on an image denoising example which is discussed through Example 2 in Subsection 1.1.2. In particular, we demonstrate numerically the statistical consistency result which includes the rate of convergence and we show that the learned parameter λ within each inverse problem experiment outperforms that with a fixed λ.

We emphasize that with our findings and results in this work, our focus is not on developing new methodology for bilevel learning. Instead our focus is on building a statistical understanding of bilevel learning through the notion of statistical consistency and convergence of numerical schemes.

1.3. Organization. The structure of this paper is given as follows. In Section 2 we present the bilevel optimization problem of interest in a stochastic framework, and present a statistical consistency result of the adaptive parameter. We then extend this result to the linear inverse setting with Tikhonov regularization. This will lead onto Section 3 where we discuss the stochastic gradient descent method and its application

in our bilevel optimization problem. We provide various assumptions required where we tend show in the linear setting that our parameter converges in L_2 to the minimizer. In Section 4 we test our theory on various numerical models which include both linear and nonlinear models such as Darcy flow and the eikonal equation. This will also include an image denoising example. Finally, we conclude our findings in Section 5. The appendix will contain the proofs for results in Section 2 and Section 3.

2. Regularization parameter offline recovery

In this section we discuss how to use offline bilevel optimization to recover regularization parameters. We also show that the solution is statistically consistent under suitable conditions.

2.1. Offline bilevel optimization.   Regularization parameter learning by bilevel optimization views the unknown parameter U and the data Y as a jointly distributed random variable with distribution $Z \sim \mu_{(U,Y)}$, see e.g. [2] for more details. Recall the bilevel optimization problem is given by

$$\lambda_* = \underset{\lambda \in \Lambda}{\arg\min} \ F(\lambda), \quad F(\lambda) = \mathbb{E}_{\mu_{(U,Y)}}[\mathcal{L}_{\mathcal{U}}(u_\lambda(Y), U)], \qquad \text{(upper level)}$$

$$u_\lambda(Y) := \underset{u \in \mathcal{U}}{\arg\min} \ \Psi(\lambda, u, Y), \quad \Psi(\lambda, u, y) := \mathcal{L}_{\mathcal{Y}}(\mathcal{G}(u), y) + \mathcal{S}_\lambda(u), \text{ (lower level)}$$

where $\mathcal{L}_{\mathcal{U}}$ denotes a discrepancy function in the parameter space $\mathcal{U} := \mathbf{R}^d$ and $\mathcal{L}_{\mathcal{Y}}$ denotes a discrepancy function in the observation space $\mathcal{Y} := \mathbf{R}^K$. $S_\lambda(U)$ represents the regularization with parameter $\lambda \in \Lambda$. 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 λ.

In general, we do not know the exact distribution μ in the upper level of (1.3). We consider the scenario where we have access to training data $z^{(j)} := (u^{(j)}, y^{(j)})_{j=1}^n$, which we assume to be i.i.d. samples from $\mu_{(U,Y)}$. With these data, we can approximate F in (1.3) by its empirical average:

$$\widehat{F}_n(\lambda) = \frac{1}{n} \sum_{j=1}^n \mathcal{L}_{\mathcal{U}}(u_\lambda(y^{(j)}), u^{(j)}). \tag{2.1}$$

This leads to a data-driven estimator of the regularization parameter,

$$\hat{\lambda}_n = \underset{\lambda \in \Lambda}{\arg\min} \ \widehat{F}_n(\lambda),$$
$$u_\lambda(y^{(j)}) = \underset{u \in \mathcal{U}}{\arg\min} \ \mathcal{L}_{\mathcal{Y}}(\mathcal{G}(u), y^{(j)}) + \mathcal{S}_\lambda(u). \tag{2.2}$$

This method of estimation is often known as empirical risk minimization in machine learning [9, 52]. We refer to this as "offline" since minimizing $\widehat{F}_n$ involves all n data points at each algorithmic iteration. With $\hat{\lambda}_n$ being formulated, it is of natural interest to investigate its convergence to the true parameter λ_*, when the sample size increases. Consistency analysis is of central interest in the study of statistics. In particular, if $\hat{\lambda}_n$ is the global minimum of $\widehat{F}_n$, we have the following Theorem 5.2.3 [8] from Bickel and Doksum, formulated in our notation

THEOREM 2.1. *Suppose for any $\epsilon > 0$*

$$\mathbb{P}(\sup\{\lambda \in \Lambda, |\widehat{F}_n(\lambda) - F(\lambda)|\} > \epsilon) \to 0,$$

as $n \to \infty$, $\hat{\lambda}_n$ is the global minimizer of $\widehat{F}_n$, and λ_* is the unique minimizer of F, then $\hat{\lambda}_n$ is a consistent estimator.

In more practical scenarios, the finding of $\hat{\lambda}_n$ relies on the choice of optimization algorithms. If we are using gradient-based algorithms, such as gradient descent, $\hat{\lambda}_n$ can be the global minimum of $\widehat{F}_n$ if $\widehat{F}_n$ is convex. More generally, we can only assume $\hat{\lambda}_n$ to be a stationary point of $\widehat{F}_n$, i.e. $\nabla \widehat{F}_n(\hat{\lambda}_n) = 0$. In such situations, we provide the following alternative tool replacing Theorem 2.1:

PROPOSITION 2.1.   *Suppose F is C^2, λ_* is a local minimum of F, and $\hat{\lambda}_n$ is a local minimum of $\widehat{F}_n$. Let $\mathcal{D}$ be an open convex neighborhood of λ_* in the parameter space and c_0 be a positive constant. Denote $\mathcal{A}_n$ as the event*

$$\mathcal{A}_n = \{\hat{\lambda}_n \in \mathcal{D}, \nabla^2_\lambda \widehat{F}_n(\lambda) \succeq c_0 I \text{ for all } \lambda \in \mathcal{D}\}.$$

When $\mathcal{A}_n$ takes place, the following holds:

$$|\hat{\lambda}_n - \lambda_*| \leq \frac{|\partial_\lambda \widehat{F}_n(\lambda_*) - \partial_\lambda F(\lambda_*)|}{c_0}.$$

In particular, we have

$$\mathbb{E}1_{\mathcal{A}_n}|\hat{\lambda}_n - \lambda_*| \leq \frac{\sqrt{tr(Var(\partial_\lambda f(\lambda_*, Z)))}}{c_0 \sqrt{n}}.$$

Proof.   Proof can be found in the appendix.   □

Proposition 2.1 makes two claims. From the second claim, we can see λ_n converges to λ_* at rate of $\frac{1}{\sqrt{n}}$. And with the first claim, sometimes we can have more accurate estimate on large or medium deviations. We will see how to do that in the linear inverse problem discussed below.

On the other hand, Proposition 2.1 requires the existence of region $\mathcal{D}$ so that both $\hat{\lambda}_n$ and λ_* are in it, moreover $\widehat{F}_n$ needs to be strongly convex inside $\mathcal{D}$. The convexity part is necessary, since without it, there might be multiple local minima, and we will have identifiability issues. In order to apply Theorem 2.1, one needs to find $\mathcal{D}$ and bound the probability of outlier cases $\mathcal{A}_n^c$. This procedure can be nontrivial, and requires some advanced tools from probability. We demonstrate how to do so for the linear inverse problem.

2.2. Offline consistency analysis with linear observation models.   In this section we demonstrate how to apply Proposition 2.1 for linear observation models with Tikhonov regularization.

To motivate our framework, we assume $u \in \mathbf{R}^d$ and the data y is observed through a matrix $A \in \mathbf{R}^{K \times d}$

$$y = Au + \xi, \tag{2.3}$$

with Gaussian prior information $u \sim \mathcal{N}(0, \frac{1}{\lambda_*} C_0)$ 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

$$\mathcal{L}_\mathcal{Y}(\mathcal{G}(u), y) = \frac{1}{2}\|Au - y\|_\Gamma^2, \quad \mathcal{S}_\lambda(u) = \frac{\lambda}{2}\|u\|_{C_0}^2.$$

Since both of these functions are quadratic in u, the lower level optimization problem has an explicit solution

$$u_\lambda(y) = (A^\top \Gamma^{-1} A + \lambda C_0^{-1})^{-1} A^\top \Gamma^{-1} y_i.$$

If we use the root-mean-square error in the upper level to learn λ, the discrepancy function is given by

$$f(\lambda, u, y) = \|u_\lambda(y) - u\|^2.$$

and the empirical loss function is defined by

$$\widehat{F}_n(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \|u_\lambda(y_i) - u_i\|^2. \tag{2.4}$$

It is worth mentioning that $F(\lambda)$ is not convex on the real line despite that $\mathcal{G}$ is linear. The detailed calculation can be found in Remark A.1. It is for this reason, that it is necessary to introduce the local region $\mathcal{D}$, that F is convex inside, in Proposition 2.1.

In some challenging scenarios the underlying distribution of the noise and parameter is not known. However, as long as one has access to the covariances, Γ and C_0 of the underlying distribution, the bilevel optimization approach is well-defined and it can still be implemented in order to choose the regularization parameter for the inverse problem. Our theoretical results extend to the general setting, as long as u is $(0, \frac{1}{\lambda_*} C_0)$ sub-Gaussian, and ξ is $(0, \Gamma)$-sub-Gaussian.

ASSUMPTION 2.1. *A random vector z is $(0, \Sigma)$-sub-Gaussian if there exists*

$$\sigma = (\sigma_1^i, \ldots, \sigma_d^i)^\top \in \mathbf{R}^\ell,$$

such that

$$z \stackrel{d}{=} \Sigma^{1/2} \sigma,$$

where $\sigma_1, \ldots, \sigma_\ell$ are i.i.d. random variables with $\mathbb{E}\sigma_1 = 0$, $\mathbb{E}|\sigma_1|^2 = 1$ and

$$\sup_{p \geq 1} p^{-1/2} \mathbb{E}[|\sigma_1|^p]^{1/p} \leq C_\sigma,$$

for some $C_\sigma > 0$. Furthermore, we assume that the components are symmetric in the sense that $\sigma_1 \stackrel{d}{=} -\sigma_1$.

Sub-Gaussian indicates the the tail of u and ξ are not heavy, so concentration inequalities can be applied [47]. Note however that $N(0, \Gamma)$ is $(0, \Gamma)$-sub-Gaussian. Uniform distributions $\mathcal{U}[(-a, a)]$ can also be sub-Gaussian.

Another important issue in inverse problems is the notion of dimension independence. Since most applications involve models of high or even infinite dimension, it is of interest to see if the parameter recovery depends only on some *effective dimension* but not the ambient space dimension d. Here, the effective dimension is usually described by physical quantities in the inverse problem. For this paper, we assume the following:

ASSUMPTION 2.2. $\operatorname{Tr}(C_0), \|C_0\|_F$ *and* $\|A^\top \Gamma^{-1} A\|, \|A\|, \|\Gamma\|$ *are constants independent of the dimension d.*

In the subsequent development, we will refer to constants that depend only on $\operatorname{Tr}(C_0), \|C_0\|_F, \|A^\top \Gamma^{-1} A\|, \|A\|, \|\Gamma\|$ as dimension independent (DI).

Roughly speaking, in order for $\operatorname{Tr}(C_0), \|C_0\|_F$ to be DI, the spectrum of the prior covariance C_0 needs to decay to zero quickly. By assuming $\|A^\top \Gamma^{-1} A\|, \|A\|, \|\Gamma\|$ to be DI, each individual observation needs to be moderately precise. We do not have hard constraints on the number of observations, other than that they should be independent.

All these conditions are practical and can be found in [2, 36]. We will also demonstrate that they hold for our numerical example(s).

Since the formulation of u_λ involves the inversion of matrix $A^\top \Gamma^{-1} A + \lambda C_0^{-1}$, such an operation may be unstable for λ approaching ∞. When λ 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 there are lower and upper bounds such that

$$0 < \lambda_l < \frac{1}{2}\lambda_* < \frac{3}{2}\lambda_* < \lambda_u,$$

where λ_l can be chosen as a very small number and λ_u can be very large. Their values often can be obtained from physical restrictions from the inverse problem. By assuming their existence, we can restrict $\hat{\lambda}_n$ to be in the interval $\Lambda = (\lambda_l, \lambda_u)$. Now we are ready to present our main result for the offline recovery of regularization parameter. In particular, we show $\hat{\lambda}_n$ converges to λ_* with high probability.

THEOREM 2.2.   *Suppose that u is $(0, \frac{1}{\lambda_*}C_0)$-sub-Gaussian, ξ is $(0, \Gamma)$-sub-Gaussian, where $C_0 \in \mathbf{R}^{d \times d}$, $\Gamma \in \mathbf{R}^{K \times K}$ are known symmetric positive definite matrices and $\lambda_* > 0$ is unknown, and let $\lambda_n^* \in (\lambda_l, \lambda_u)$ be a local minimum of $\widehat{F}_n$. Then it holds true that $\partial_\lambda F(\lambda_*) = 0$ and there exist $C_*, c_* > 0$ such that for any $1 > \epsilon > 0$ and n,*

$$\mathbb{P}(|\hat{\lambda}_n - \lambda_*| > \epsilon, \lambda_l < \hat{\lambda}_n < \lambda_u) \leq C_* \exp(-c_* n \min(\epsilon, \epsilon^2)).$$

The values of $C_*, c_* > 0$ depend on $\lambda_l, \lambda_u, \lambda_*, C_0$ but not on n. Moreover, if Assumption 2.2 is assumed, $C_*, c_* > 0$ are also dimension independent.

Proof. (Sketch of the proof).   The basic idea of proving the consistency is to construct a open convex neighborhood $\mathcal{D}$ of λ_* in order to apply Proposition 2.1. The global convergence can be obtained by splitting

$$\mathbb{P}(|\hat{\lambda}_n - \lambda_*| > \epsilon, \lambda_l < \hat{\lambda}_n < \lambda_u) = \mathbb{P}(|\hat{\lambda}_n - \lambda_*| > \epsilon, \lambda_l < \hat{\lambda}_n < \lambda_u \,|\, \mathcal{A}_n)\mathbb{P}(\mathcal{A}_n)$$
$$+ \mathbb{P}(|\hat{\lambda}_n - \lambda_*| > \epsilon, \lambda_l < \hat{\lambda}_n < \lambda_u \,|\, \mathcal{A}_n^{\complement})\mathbb{P}(\mathcal{A}_n^{\complement}),$$

where it is left to prove the convergence of $\mathbb{P}(\mathcal{A}_n^{\complement})$ to zero, which means that $\hat{\lambda}_n$ stays in the local convexity region $\mathcal{D}$ with high probability. We give full details of the proof in the appendix.   □

Since we can obtain consistency assuming that $\hat{\lambda}_n$ is a local minimum, we do not demonstrate how to implement Theorem 2.1 for the more restrictive scenario where $\hat{\lambda}_n$ is a global minimum.

In general, knowing that $\hat{\lambda}_n$ is an accurate estimator is sufficient to guarantee that the recovery is accurate, also due to the Lipschitzness of u_λ. In particular, we can also show that the Lipschitz constant is DI:

PROPOSITION 2.2.   *Under Assumption 2.2, there is a dimension independent L so that*

$$\mathbb{E}_y \|u_\lambda(y) - u_{\lambda_*}(y)\|^2 \leq L|\lambda - \lambda_*|^2.$$

Here y is a random observation sample.

REMARK 2.1.   In practical settings it is assumed that the noise is not fully known and we can easily extend our results by changing $\Gamma \mapsto \frac{1}{\gamma}\Gamma$. We then would try to choose the ratio between regularization parameter and noise scale, this is, we can change $\lambda \mapsto \frac{\lambda}{\gamma}$ and apply again the bilevel optimization approach with "known" noise covariance Γ.

3. Regularization parameter online recovery

In this section we discuss how to implement the stochastic gradient descent (SGD) method for online solutions of the bilevel optimization. We will formulate the SGD method for general nonlinear inverse problems and state certain assumptions on the forward model and the regularization function to ensure convergence of the proposed method.

3.1. Bilevel stochastic gradient descent method.   In the offline solution of the bilevel optimization problem (2.2), one has to compute the empirical loss function $\widehat{F}_n$ or its gradient in (2.1). This involves solving the lower level problem for each training data point $(u^{(j)}, y^{(j)})$, $j = 1, \ldots, n$. When n is very large, this can be very computationally demanding. One way to alleviate this is to use the stochastic gradient descent (SGD). This has been done in the context of traditional optimization [19], where various convergence results were shown. As a result this has been applied to problems in machine learning, most notably support vector machines [17, 18], but also in a more general context without the use of SGD [29, 35]. Here we propose a SGD method to solve the bilevel optimization problem (1.3) online.

To formulate the SGD method, we first note that the gradient descent method generates iterates λ_{k+1} based on the following update rules

$$\lambda_{k+1} = \lambda_k - \beta_k \partial_\lambda F(\lambda_k),$$

where β_k is a sequence of stepsizes.

As mentioned above, the population gradient $\partial_\lambda F$ is often computationally inaccessible, and its empirical approximation $\partial_\lambda \widehat{F}_n$ is often expensive to compute. One general solution to this issue is using a stochastic approximation of $\partial_\lambda F$. Here we choose $\partial_\lambda f(\lambda_k, Z^{(k)})$, where $(Z^{(k)})_{k \in \{1, \ldots, n\}} \sim \mu_{(U,Y)}$, since it is an unbiased estimator of $\partial_\lambda F$:

$$\partial_\lambda F(\lambda_k) = \mathbb{E}_Z \partial_\lambda f(\lambda_k, Z).$$

The identity above holds by Fubini's theorem, since we assume f and its second-order derivatives are all continuous and differentiable. Comparing with $\partial_\lambda \widehat{F}_n$, $\partial_\lambda f$ 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.

We formulate the stochastic gradient descent method to solve (1.3) as Algorithm 1.

Algorithm 1 *Bilevel Stochastic Gradient Descent*

1: Input: λ_0, m, $\beta = (\beta_k)_{k=1}^n$, $\beta_k > 0$, i.i.d. sample $(Z^{(k)})_{k \in \{1, \ldots, n\}} \sim \mu_{(U,Y)}$.

    for $k = 0, \ldots, n-1$ **do**

$$\lambda_{k+1} = \chi(\lambda_k - \beta_k \partial_\lambda f(\lambda_k, Z^{(k)})), \tag{3.1}$$

    end for

2: Output: the average $\bar{\lambda}_n = \frac{1}{m} \sum_{k=n-m+1}^n \lambda_k$

In Algorithm 1, the step size β_k is a sequence decreasing to zero not too fast, so

that the Robbins–Monro conditions [46] apply:

$$\sum_{k=1}^{\infty} \beta_k = \infty, \quad \sum_{k=1}^{\infty} \beta_k^2 < \infty. \tag{3.2}$$

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 $\bar{\lambda}_n$, which has been shown to accelerate the scheme for standard SGD methods, see [45]. The projection map χ ([52] Section 14.4.1) is defined as

$$\chi(\lambda) = \arg\min_{\theta \in \Lambda} \{|\theta - \lambda|\}.$$

In other words, it maps λ to itself if $\lambda \in \Lambda$, otherwise it outputs the point in Λ that is closest to λ. Using χ ensures λ_{k+1} is still in the range of regularization parameter if Λ is closed. This operation in general shortens the distance between λ_{k+1} and λ_* when Λ is convex:

LEMMA 3.1 (Lemma 14.9 of [52]).   *If Λ is convex, then for any λ*

$$|\chi(\lambda) - \lambda_*| \leq |\lambda - \lambda_*|.$$

Proof.   Proof can be found in the appendix.   □

In particular, the stochastic gradient $\partial_\lambda f(\lambda_k, Z^{(k)})$ is given by the following lemma, which states sufficient conditions on Ψ to ensure both u_λ and f are continuously differentiable w.r.t. λ.

LEMMA 3.2.   *Suppose the lower level loss function $\Psi(\lambda, u, y)$ is C^2 for (u, λ) in a neighborhood of $(u_{\lambda_0}, \lambda_0)$ and is strictly convex in u in this neighborhood, then the function $\lambda \mapsto u_\lambda(y)$ is continuously differentiable w.r.t. λ near λ_0 and the derivative is given by*

$$\partial_\lambda u_\lambda(y) = -\left(\partial_u^2 [\Psi(\lambda, u_\lambda(y), y)]\right)^{-1} \partial_{\lambda u}^2 [\Psi(\lambda, u_\lambda(y), y)]. \tag{3.3}$$

and

$$\partial_\lambda f(\lambda, y, u) = \partial_w \mathcal{L}_\mathcal{U}(u_\lambda(y), u)^T \partial_\lambda u_\lambda(y). \tag{3.4}$$

Proof.   Proof can be found in the appendix.   □

3.2. Approximate stochastic gradient method.   In order to implement Algorithm 1, it is necessary to evaluate the gradient $\partial_\lambda f$. While Lemma 3.2 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.

One general way to find approximate gradient is applying central finite difference schemes. This involves perturbing certain coordinates in opposite direction, and use the value difference to approximate the gradient:

$$\widetilde{\partial}_\lambda f(\lambda_k, z) \approx \frac{f(\lambda_k + h_k, z) - f(\lambda_k - h_k, z)}{2h_k}, \tag{3.5}$$

where h_k is a step size. The 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

$$\mathcal{L}_\mathcal{U}(y,u) = \|y-u\|^2.$$

The exact SGD update step (3.1) can be written as

$$\lambda_{k+1} = \lambda_k - \beta_k \partial_\lambda \|u_{\lambda_k}(y^{(k)}) - u^{(k)}\|^2$$
$$= \lambda_k - \beta_k \left(\partial_\lambda u_{\lambda_k}(y^{(k)})\right)^\top (u_{\lambda_k}(y^{(k)}) - u^{(k)}).$$

In this case, it makes more sense to apply central difference scheme only on the ∂u_λ part:

$$\partial_\lambda u_\lambda(y^{(k)}) \approx \frac{u_{\lambda+h_k}(y^{(k)}) - u_{\lambda-h_k}(y^{(k)})}{2h_k} =: \widetilde{\partial}_\lambda u_\lambda(y^{(k)}). \tag{3.6}$$

Using this approximation, we formulate the approximate SGD method in the following algorithm, where we replace the exact gradient $\partial_\lambda u_\lambda(y^{(k)})$ by the numerical approximation $\widetilde{\partial}_\lambda u_\lambda(y^{(k)})$ defined in (3.6).

Here we have defined the numerical approximation of $\partial_\lambda f$ by

$$\widetilde{\partial}_\lambda f(\lambda,(y,u)) := \left(\widetilde{\partial}_\lambda u_\lambda(y)\right)^\top (u_\lambda(y) - u). \tag{3.7}$$

In most finite difference approximation schemes, the approximation error involved is often controlled by h_k. In particular, we assume the centered forward difference scheme used in either (3.5) or (3.7) yields an error of order

$$\|\mathbb{E}\widetilde{\partial}_\lambda(f(\lambda,Y,U)) - \partial_\lambda F(\lambda)\| =: \alpha_k = O(h_k^2).$$

Replacing the stochastic gradient in Algorithm 1 with its approximation, we obtain the algorithm below:

Algorithm 2 *Approximate Bilevel Stochastic Gradient Descent*

1: Input: λ_0, m, $\beta = (\beta_k)_{k=1}^n$, $\beta_k > 0$, i.i.d. sample $(Z^{(k)})_{k \in \{1,\dots,n\}} \sim \mu_{(U,Y)}$.
$\quad$ **for** $k=0,\dots,n-1$ **do**

$$\lambda_{k+1} = \chi(\lambda_k - \beta_k \widetilde{\partial}_\lambda f(\lambda_k, Z^{(k)})), \tag{3.8}$$

$\quad$ **end for**
2: Output: the average $\bar{\lambda}_n = \frac{1}{m} \sum_{k=n-m+1}^n \lambda_k$

3.3. Consistency analysis for online estimators. Next we formulate sufficient conditions that can ensure that λ_k converges in L^2 to the optimal solution λ_* of (1.3).

PROPOSITION 3.1. *Suppose that there is a convex region $\mathcal{D} \subset \Lambda$ and a constant $c > 0$ such that*

$$(\lambda - \lambda_*)\partial_\lambda F(\lambda) > c|\lambda - \lambda_*|^2. \tag{3.9}$$

for all $\lambda \in \mathcal{D}$ and there are constants a, $b > 0$ such that for all $\lambda \in \mathcal{D}$ it holds true that

$$\mathbb{E}[|\widetilde{\partial}_\lambda f(\lambda, Z)|^2] < a + b|\lambda - \lambda_*|^2. \tag{3.10}$$

Also the bias in the approximated SGD is bounded by

$$|\mathbb{E}\widetilde{\partial}_\lambda f(\lambda_k, Z_k) - \partial_\lambda F(\lambda_k)|^2 \leq \alpha_k. \tag{3.11}$$

Let $\mathcal{A}_n$ be the event that $\lambda_k \in \mathcal{D}$. 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

$$\mathbb{E}1_{\mathcal{A}_n}|\lambda_n - \lambda_*|^2 \leq \left(\mathbb{E}Q_0 + 2a\sum_{j=1}^\infty \beta_j^2 \right) C_n + \frac{\alpha_0}{c^2}.$$

Here

$$C_n = \min_{k \leq n} \max \left\{ \prod_{j=k+1}^n (1 - c\beta_j), a\beta_k/c \right\} \tag{3.12}$$

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

$$\mathbb{E}1_{\mathcal{A}_n}|\lambda_n - \lambda_*|^2 \leq \left(\mathbb{E}Q_0 + 2(a + D/c)\sum_{j=1}^\infty \beta_j^2 \right) C_n.$$

   Proof.   Proof can be found in the appendix.   □

REMARK 3.1.   We note that the above result also leads to similar convergence of the average estimator $\bar{\lambda}_n$ since by Jensen's inequality

$$|\bar{\lambda}_n - \lambda_*|^2 \leq \frac{1}{m} \sum_{k=n-m+1}^n |\lambda_k - \lambda_*|^2.$$

 Further, for standard SGD methods the averaging step has been shown to lead to the highest possible convergence rate under suitable assumptions. Interested readers can refer to [45] for more details.

REMARK 3.2.   Most of our discussion focused on the case where unlimited independent data points $Z^{(k)} = (u^{(k)}, y^{(k)})$ are accessible. Sometimes in practice this is not possible, and we have access to $\{Z^{(k)}, k = 1, \ldots, M\}$. SGD in this case can still be implemented to minimize the empirical loss function $F(\lambda) = \frac{1}{M}\sum_{k=1}^M f(\lambda, Z^{(k)})$. To do so, we just need to let $\mu_{U,Y} = \frac{1}{M}\sum_{k=1}^K \delta_{u^{(k)}, y^{(k)}}$. So when we apply SGD and draw a random sample from $\mu_{U,Y}$, we essentially draw a random index i from $\{1, \ldots, M\}$ and apply SGD to the data point $Z^{(i)}$. Proposition 3.1 still holds and shows that SGD will converge to the minimizer of the empirical loss function.

   3.4. Consistency analysis with linear inverse problem.   We consider again the linear inverse problem from Section 2.2

$$y = Au + \xi,$$

but we do not state specific assumptions on the distribution without $\mathbb{E}[|u|^4] < \infty$ and $\mathbb{E}[|A^\top \Gamma^{-1} \xi|^4] < \infty$.

We formulate the online convergence for the corresponding bilevel optimization with least-square data misfit and Tikhonov regularization, i.e.

$$\mathcal{L}_{\mathcal{Y}}(Au, y) = \frac{1}{2} \|Au - y\|_\Gamma^2, \quad S_\lambda(u) = \frac{\lambda}{2} \|u\|_{C_0}^2.$$

THEOREM 3.1.   *Let* $\beta = (\beta_k)_{k \in \mathbb{N}}$ *be a sequence of step sizes with* $\beta_k > 0$, $\sum\limits_{k=1}^{\infty} \beta_k = \infty$, *and* $\sum\limits_{k=1}^{\infty} \beta_k^2 < \infty$. *Furthermore, let* $\mathbb{E}[|u|^4] < \infty$ *and* $\mathbb{E}[|A^\top \Gamma^{-1} \xi|^4] < \infty$. *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 1, converge to* λ_* *in the sense*

$$\mathbb{E}|\lambda_n - \lambda_*|^2 \le BC_n,$$

(2) *the iterates generated from the approximate SGD, Algorithm 2 with formula (3.7) and* $h_k = h$, *converge to* λ_* *up to an error of order* $\mathcal{O}(h^4)$, *i.e.*

$$\mathbb{E}|\lambda_n - \lambda_*|^2 \le B(C_n + h^4).$$

If we use decaying finite difference stepsize $h_k \le h \beta_k^{1/4}$, *then the error can be further bounded by*

$$\mathbb{E}|\lambda_n - \lambda_*|^2 \le BC_n.$$

Moreover, if Assumption 2.2 is assumed, the constants B, C_n *are dimension independent.*

Proof.   Proof can be found in the appendix.   □

REMARK 3.3.   While in the offline setting the proof of the consistency result for the linear sub-Gaussian setting was heavily relying on the sub-Gaussian assumption on u and ξ, in the online setting we are able to extend the result to a wider range of distributions of u and ξ. For our proof of Theorem 3.1 in Appendix B.3 we only need to assume that $\mathbb{E}[|u|^4] < \infty$ and $\mathbb{E}[|A^\top \Gamma^{-1} \xi|^4] < \infty$. Hence, it can also be applied to general linear inverse problems without sub-Gaussian assumption on the unknown parameter or sub-Gaussian assumption on the noise.

4. Numerical results

In this section our focus will be directed on testing the results of the previous sections. We will present various inverse problems to our theory, which will be based on partial differential equations, both linear and nonlinear, which includes a linear 2D Laplace equation, a 2D Darcy flow from geophysical sciences and a 2D eikonal equation which arises in wave propagation. As a final numerical experiment, related to the examples discussed in Section 1, 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 1 with exact derivative (3.4).
- bSGD_a: Application of the bilevel SGD, Algorithm 2 with derivative approximation (3.7) for fixed $h_k = h_0$ in (3.6).

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 3.1 has already implied this, since the difference between the two schemes is of order h_0^{-4}, which is often smaller than the error from the numerical forward map solver or the use of $\bar{\lambda}_n$. For this reason, we do not demonstrate this scheme in our numerics.

4.1. Linear example: 2D Laplace equation.   We consider the following forward model

$$\begin{cases} -\Delta p(x) & = u(x), \quad x \in X, \\ p(x) & = 0, \qquad x \in \partial X, \end{cases} \tag{4.1}$$

with Lipschitz domain $X = [0,1]^2$ and consider the corresponding inverse problem of recovering the unknown $u^\dagger$ from observation of (4.1), described through

$$y = \mathcal{O}(p) + \eta, \tag{4.2}$$

where $\eta \sim \mathcal{N}(0,\Gamma)$ is the measurement noise and p is the solution of (4.1). We solve the PDE in weak form, where $\mathcal{A}^{-1} : \mathcal{U} \to \mathcal{V}$, with $\mathcal{U} = L^\infty(X)$ and $\mathcal{V} = H_0^1(X) \cap H^2(X)$, denotes the solution operator for (4.1) and $\mathcal{O} : \mathcal{V} \to \mathbf{R}^K$ denotes the observation map taking measurements at K randomly chosen points in X, i.e. $\mathcal{O}(p) = (p(x_1), \ldots, p(x_K))^\top$, for $p \in \mathcal{V}$, $x_1, \ldots, x_K \in X$. For our numerical setting $K = 250$ points have been observed, which is illustrated in Figure 4.1. We can express this problem as a linear inverse problem in the reduced form (2.3) by

$$y^\dagger = A u^\dagger + \eta \in \mathbf{R}^K, \tag{4.3}$$

where $A = \mathcal{O} \circ \mathcal{A}^{-1}$ is the forward operator which takes measurements of (4.1). The forward model (4.1) is solved numerically on a uniform mesh with 32×32 grid points in X by a finite element method with continuous, piecewise linear finite element basis functions.

We assume that our unknown parameter $u^\dagger$ follows a Gaussian distribution $\mathcal{N}(0, \frac{1}{\lambda_*} C_0)$ with covariance

$$C_0 = \beta \cdot (\tau^2 I - \Delta)^{-\alpha}, \tag{4.4}$$

with Laplacian operator Δ equipped with Dirichlet boundary conditions, known β, $\tau > 0$, $\alpha > 1$ and unknown $\lambda^* > 0$. To sample from the Gaussian distribution, we consider the truncated Karhunen-Loève (KL) expansion [40], which is a series representation for $u \sim \mathcal{N}(0, C_0)$, i.e.

$$u(x) = \sum_{i=1}^\infty \xi_i \sqrt{\frac{1}{\lambda_*}} \sigma_i \varphi_i(x), \tag{4.5}$$

where $(\sigma_i, \varphi_i)_{i \in \mathbb{N}}$ are the eigenvalues and eigenfunctions of the covariance operator C_0 and $\xi = (\xi_i)_{i \in \mathbb{N}}$ is an i.i.d. sequence with $\xi_1 \sim \mathcal{N}(0,1)$ i.i.d.. Here, we have sampled
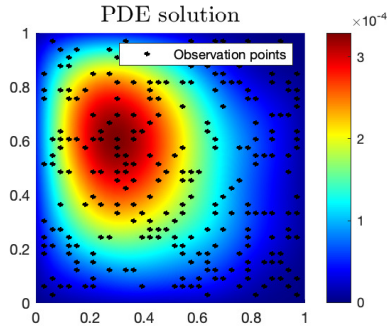
Fig. 4.1: *Reference PDE solution for the Laplace equation of the underlying unknown parameter $u^\dagger$, and the corresponding randomized observation points $x_1,\ldots,x_K \in X$.*
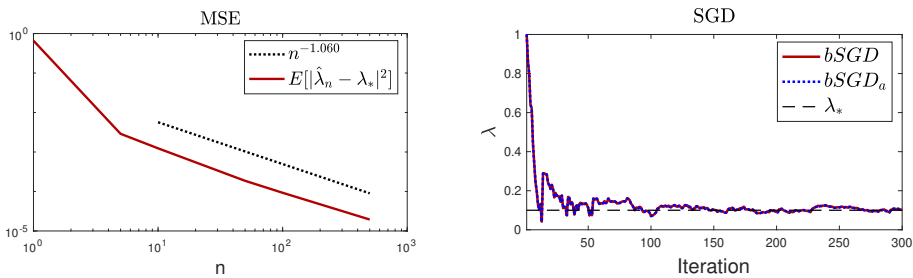

Fig. 4.2: *MSE (left) resulting from the offline recovery depending on training data size. Learned regularization parameter λ_k (right) resulting from the online recovery, Algorithm 1 for the Laplace equation.*

from the KL expansion for the discretized C_0 on the uniform mesh. Furthermore, we assume to have access to training data $(u^{(j)}, y^{(j)} = Au^{(j)} + \eta^{(j)})_{j=1,\ldots,n}$, $n \in \mathbb{N}$, which we will use to learn the unknown scaling parameter λ_* before solving the inverse problem. For the numerical experiment we set $\beta = 100$, $\tau = 0.1$, $\alpha = 2$ and $\lambda_* = 0.1$. After learning the regularization parameter, we will compare the estimated parameter through the different results of the Tikhonov minimum

$$u_{\lambda_i}(y^\dagger) = (A^\top \Gamma^{-1} A + \lambda_i \cdot C_0)^{-1} A^\top y^\dagger,$$

for $\lambda_1 = \widehat{\lambda}$ learned from the training data, $\lambda_2 = \lambda_*$ and fixed $\lambda_3 = 1$. We have used the MATLAB function `fmincon` to recover the regularization parameter offline by solving the empirical optimization problem

$$\widehat{\lambda}_n \in \underset{\lambda > 0}{\arg\min} \frac{1}{n} \sum_{j=1}^{n} |u_\lambda(y^{(j)}) - u^{(j)}|^2. \tag{4.6}$$

We use $M = 1000$ samples of training data to construct Monte–Carlo estimates of $\mathbb{E}[|\widehat{\lambda}_n - \lambda^\dagger|^2]$. While the computation of the empirical loss function can be computationally demanding, we also apply the proposed online recovery in the form of the SGD method to learn the regularization parameter λ by running Algorithm 1 with chosen step size $\beta_k = 200/k$, range of regularization parameter $\Lambda = [0.0001, 10]$ and initial value $\lambda_0 = 1$. The resulting iterate λ_k can be seen in Figure 4.2 on the right side.
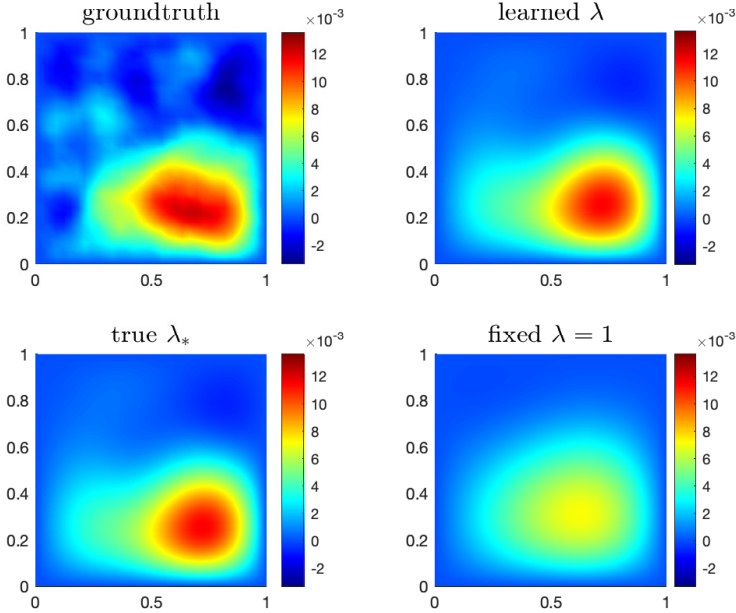
Fig. 4.3: *Comparison of different Tikhonov solutions for choices of regularization parameter λ_i. The learned Tikhonov regularized solution corresponds to the resulting one of the SGD method Algorithm 1 for the Laplace equation.*

From the numerical experiments for the linear example we observe that the numerics match our derived theory. In the offline recovery setting, this is first evident in Figure 4.2 on the left side. We compare the MSE with the theoretical rate, which seems to decay at the same rate. The online recovery is highlighted by the right plot in Figure 4.2 which demonstrates the convergence towards λ_* as the iterations progress. Further, we show the result of the approximate bSGD method Algorithm 2 for fixed chosen $h_k = 0.01$ in (3.6). As the derivative approximation (3.6) is closely exact, we see very similar good performance of the approximate bSGD method.

Finally, Figure 4.3 shows the recovery of the underlying unknown through different choices of λ. It verifies that the adaptive learning of λ outperforms that of fixed regularization parameter $\lambda = 1$.

4.1.1. Dimension independent experiments.   Next, we are going to analyze the independence of dimension in the bilevel optimization approach. Our setup is similar as discussed before, but considers the domain $X = [0,1]$.

We solve the forward model numerically on a uniform mesh for different choices of mesh sizes $h \in \{2^{-5}, 2^{-6}, 2^{-7}, 2^{-8}\}$ by a finite element method with continuous, piecewise-linear ansatz functions, where the same $K = 5$ observation points have been observed on each mesh. We assume that the underlying parameter $u^\dagger$ follows a Gaussian distribution $\mathcal{N}(0, \frac{1}{\lambda_*} C_0)$ with $C_0 = (-\Delta)^{-1}$ and apply again the truncated KL expansion up to a fixed truncation index, but considering discretized versions φ_i^h of $\varphi_i(x)$ on each level h.

In Figure 4.4, we compare the MSE for the resulting estimates $\widehat{\lambda}_n$ for different choices of sample size n depending on the dimension d. Here, we use again $M = 1000$

samples of training data to construct Monte-Carlo estimates of the MSE $\mathbb{E}[|\widehat{\lambda}_n - \lambda_*|^2]$.
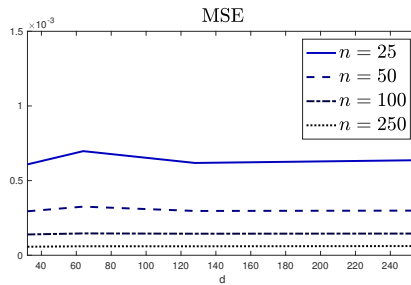

Fig. 4.4: *MSE resulting from the offline recovery depending on dimension of the parameter space d and evaluated for different choices of training data size n.*
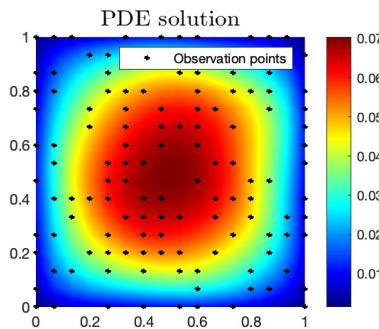

Fig. 4.5: *Reference PDE solution for Darcy flow of the underlying unknown parameter $u^\dagger$ and the corresponding randomized observation points $x_1,\ldots,x_K \in X$.*

4.2. Nonlinear example: 2D Darcy flow.   We now consider the following elliptic PDE which arises in the study of subsurface flow known as Darcy flow. The forward model is concerned using the log-permeability $\log u \in L^\infty(X) =: \mathcal{U}$ to solve for the pressure $p \in H_0^1(X) \cap H^2(X) =: \mathcal{V}$ from

$$\begin{cases} -\nabla \cdot (\exp(u)\nabla p) &= f, \quad x \in X \\ p &= 0, \quad x \in \partial X \end{cases} \tag{4.7}$$

with domain $X = [0,1]^2$ and known scalar field $f \in \mathbf{R}$. We again consider the corresponding inverse problem of recovering the unknown $u^\dagger$ from observation of (4.7), described through

$$y = \mathcal{O}(p) + \eta, \tag{4.8}$$

where $\mathcal{O}: \mathcal{V} \to \mathbf{R}^K$ denotes the linear observation map, which takes again measurements at K randomly chosen points in X, i.e. $\mathcal{O}(p) = (p(x_1),\ldots,p(x_K))^\top$, for $p \in \mathcal{V}$, $x_1,\ldots,x_K \in X$. For our numerical setting we choose $K = 125$ observational points, which can again be seen in Figure 4.5. The measurement noise is denoted by $\eta \in \mathcal{N}(0,\Gamma)$, for $\Gamma \in \mathbf{R}^{K \times K}$ symmetric and positive definite.

We formulate the inverse problem through

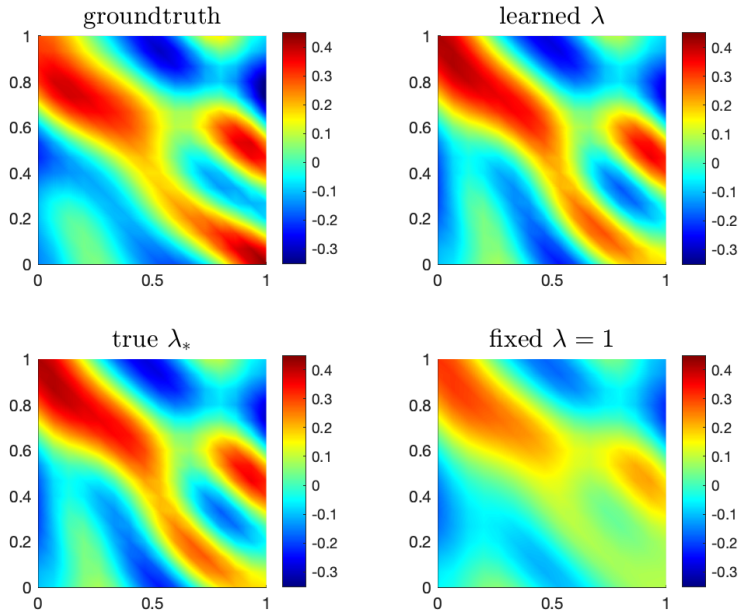$$y^\dagger = \mathcal{G}(u^\dagger) + \eta, \tag{4.9}$$

Fig. 4.6: *Comparison of different Tikhonov solutions for choices of the regularization parameter λ. The learned Tikhonov regularized solution corresponds to the resulting one of the SGD method Algorithm 2 for Darcy flow.*
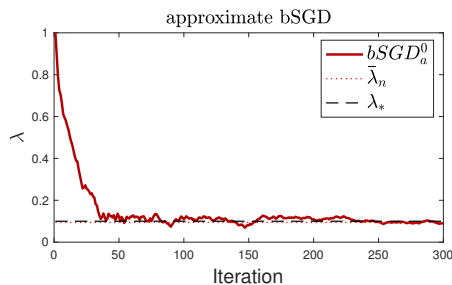

Fig. 4.7: *Learned regularization parameter λ_k, for Darcy flow, resulting from the approximate bilevel SGD method Algorithm 2 with fixed derivative accuracy $h = h_0$ and the corresponding mean over the last 50 iterations $\bar{\lambda}_n$. We obtain an error $|\lambda_* - \bar{\lambda}_n|^2 = 3.3640\mathrm{e}{-05}$.*

with $\mathcal{G} = \mathcal{O} \circ G$, where $G : \mathcal{U} \to \mathcal{V}$ denotes the solution operator of (4.7), solving the PDE (4.7) in weak form. The forward problem (4.7) has been solved by a second-order centered finite difference method on a uniform mesh with 256 grid points.

We assume that $u^\dagger$ follows the Gaussian distribution $\mathcal{N}(0, \frac{1}{\lambda_*} C_0)$ with a covariance operator (4.4) prescribed with Neumann boundary condition. Similar as before, $\beta, \tau > 0$ and $\alpha > 1$ are known, while $\lambda_* > 0$ is unknown. This time, in order to infer the unknown parameter, we use the KL expansion and do estimation of the coefficients ξ. See also [11, 31] for more details. Therefore we truncate (4.5) up to d and consider the nonlinear

map $\mathcal{G} : \mathbf{R}^d \to \mathbf{R}^K$, with $\mathcal{G}(\xi) = G(u^\xi(\cdot))$ and

$$u^\xi(\cdot) = \sum_{i=1}^d \xi_i \sqrt{\frac{1}{\lambda_*} \sigma_i} \varphi_i(\cdot).$$

This implies our unknown parameter is given by $\xi \in \mathbf{R}^d$ and we set a Gaussian prior on ξ with $\mathcal{N}(0, \frac{1}{\lambda_*} I)$, where $\lambda_* > 0$ is unknown.

We again assume to have access to training data $(\xi^{(j)}, y^{(j)})_{j=1,\dots,n}$, $n \in \mathbb{N}$, where $\xi^{(j)} \sim \Xi \sim \mathcal{N}(0, \frac{1}{\lambda_*} I)$ and we aim to solve the original bilevel optimization problem

$$\widehat{\lambda} \in \underset{\lambda > 0}{\arg\min} \ \mathbb{E}[\|u_\lambda(Y) - \Xi\|^2], \quad u_\lambda(Y) = \underset{\xi \in \mathbf{R}^d}{\arg\min} \ \frac{1}{2} \|\mathcal{G}(\xi) - Y\|_\Gamma^2 + \frac{\lambda}{2} \|\xi\|_I^2. \tag{4.10}$$

The corresponding empirical optimization problem is given by

$$\widehat{\lambda}^n \in \underset{\lambda > 0}{\arg\min} \ \frac{1}{n} \sum_{j=1}^n \|u_\lambda(y^{(j)}) - \xi^{(j)}\|^2, \quad u_\lambda(y^{(j)}) = \underset{\xi \in \mathbf{R}^d}{\arg\min} \ \frac{1}{2} \|\mathcal{G}(\xi) - y^{(j)}\|_\Gamma^2 + \frac{\lambda}{2} \|\xi\|_I^2,$$

$$\tag{4.11}$$

for a given size of the training data n. In comparison to the linear setting, we are not able to compute the Tikhonov minimum analytically for each observation $y^{(j)}$, as we require more computational power to solve (4.11). We will solve (4.11) online by application of Algorithm 2, where we will approximate the derivative of the forward model by centered difference method (3.6). We keep the accuracy of the numerical approximation fixed to $h_k = 0.01$.

For our numerical results we choose $d = 25$ coefficients in the KL expansion and the noise covariance $\Gamma = \gamma^2 I$ with $\gamma = 0.001$. For the prior model set $\beta = 10$, $\alpha = 2$, $\tau = 3$ and the true scaling parameter $\lambda_* = 0.1$.

For the SGD method we have chosen a step size $\beta_k = 0.001 k^{-1}$. The learned parameter moves fast into the direction of the true λ_*, and oscillates around this value, where the variance reduces with the iterations, as seen in Figure 4.7.

Finally, Figure 4.6 highlights again the importance and improvements of choosing the right regularization parameters.

4.3. Nonlinear example: Eikonal equation. We also seek to test our theory on the eikonal equation, which is concerned with wave propagation. Given a slowness or inverse velocity function $s(x) \in C^0(\bar{X}) =: \mathcal{U}$, characterizing the medium, and a source location $x_0 \in X$, the forward eikonal equation is to solve for travel time $T(x) \in C^0(\bar{X}) =: \mathcal{V}$ satisfying

$$\begin{cases} |\nabla T(x)| &= s(x), \quad x \in X \setminus \{x_0\}, \\ T(x_0) &= 0, \\ \nabla T(x) \cdot \nu(x) &\geq 0, \qquad x \in \partial X. \end{cases} \tag{4.12}$$

The forward solution $T(x)$ represents the shortest travel time from x_0 to a point in the domain X. The Soner boundary condition imposes that the wave propagates along the unit outward normal $\nu(x)$ on the boundary of the domain. The model equation (4.12) is of the form (1.4) with an additional constraint arising from the Soner boundary condition.

The inverse problem for (4.12) is to determine the speed function $s = \exp(u)$ from measurements of the shortest travel time $T(x)$. The data is assumed to take the form

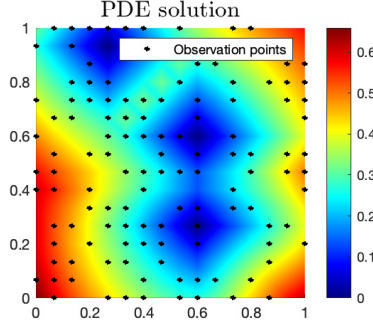$$y = \mathcal{O}(T) + \eta, \tag{4.13}$$

Fig. 4.8: *Reference PDE solution for the eikonal equation of the underlying unknown parameter $u^\dagger$, and the corresponding randomized observation points $x_1, \ldots, x_K \in X$.*
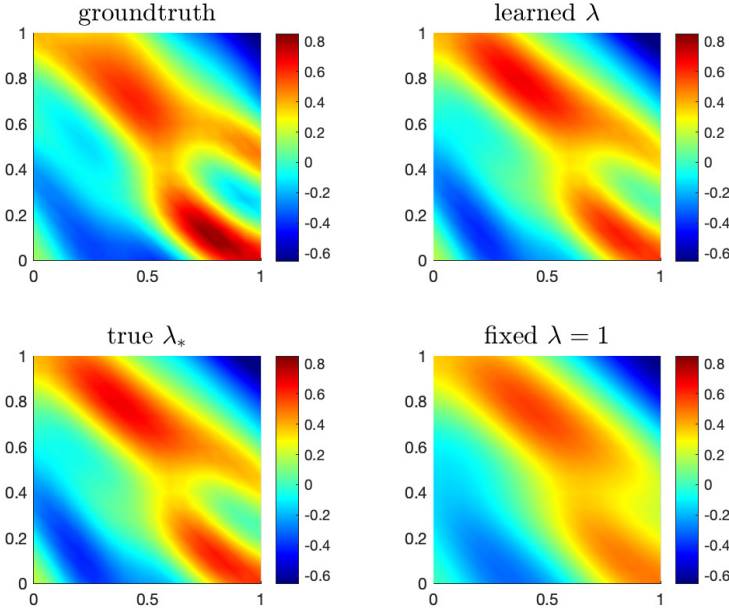

Fig. 4.9: *Comparison of different Tikhonov solutions for choices of the regularization parameter λ. The learned Tikhonov regularized solution corresponds to the resulting one of the SGD method Algorithm 2 for the eikonal equation.*

where $\mathcal{O}: \mathcal{V} \to \mathbf{R}^K$ denotes the linear observation map, which takes again measurements at $K = 125$ randomly chosen grid points in X, i.e. $\mathcal{O}(p) = (T(x_1), \ldots, T(x_K))^\top$, for $T \in \mathcal{Z}$, $x_1, \ldots, x_K \in X$. The observed points can be seen in Figure 4.8. The measurement noise is again denoted by $\eta \in \mathcal{N}(0, \Gamma)$, for $\Gamma \in \mathbf{R}^{K \times K}$ symmetric and positive definite. Again we formulate the inverse problem through

$$y^\dagger = \mathcal{G}(u^\dagger) + \eta, \qquad (4.14)$$

with $\mathcal{G} = \mathcal{O} \circ G$, where $G: \mathcal{U} \to \mathcal{V}$ denotes the solution operator of (4.7). As before we will assume our unknown $u^\dagger$ is distributed according to a mean-zero Gaussian with

covariance structure (4.4). For this numerical example we set $\beta=1$, $\tau=0.1$, $\alpha=2$ and $\lambda_*=0.1$. We truncate the KL expansion such that the unknown parameters $\xi\in\mathbf{R}^d$ with $d=25$. For the eikonal equation we take a similar approach to Section 4.2, that is we use the SGD described through Algorithm 2. For the SGD method we have chosen an adaptive step size

$$\beta_k=\min\left(0.002,\frac{\lambda_0}{|\partial_\lambda f(\lambda_k,Z^{(k)})|}\right)k^{-1}.$$

Here, the chosen step size β_k provides a bound on the maximal moved step in each SGD step, i.e.

$$|\beta_k\cdot\partial_\lambda f(\lambda_k,Z^{(k)})|\le\lambda_0/k.\qquad(4.15)$$

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 (3.2) is still satisfied. Therefore, we will also show the resulting $\sum_{k=1}^n\beta_k$ and the realisation of the stochastic gradient $\partial f_\lambda(\lambda_k,Z^{(k)})$ in Figure 4.11.
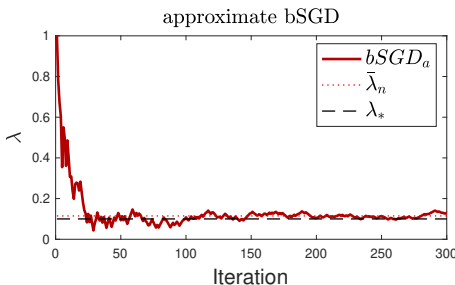

Fig. 4.10: *Learned regularization parameter λ_k, for the eikonal equation, resulting from the approximate bilevel SGD method Algorithm 2 with fixed derivative accuracy $h=h_0$ and the corresponding mean over the last 50 iterations $\bar{\lambda}_n$. We obtain an error $|\lambda_*-\bar{\lambda}_n|^2=1.9360\mathrm{e}-05$.*
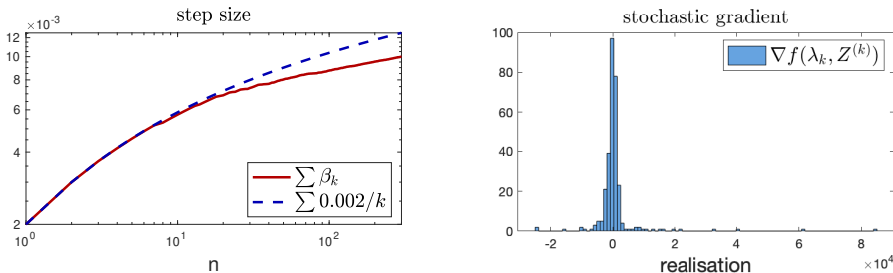

Fig. 4.11: *Summation of the realized adaptive step size (left) and the realized stochastic gradient $\nabla f(\lambda_k,Z^{(k)})$ (right) resulting from the online recovery, Algorithm 1 for the eikonal equation.*

Our setting for the parameter choices of our prior and for the bilevel-optimization problem remain the same. To discretize (4.12) on a uniform mesh with 256 grid points we use a fast marching method, described by the work of Sethian [27, 49].

As we observe the numerical experiments, Figure 4.9 highlights that using the learned λ_n provides recoveries almost identical to that of using the true λ_*. For both
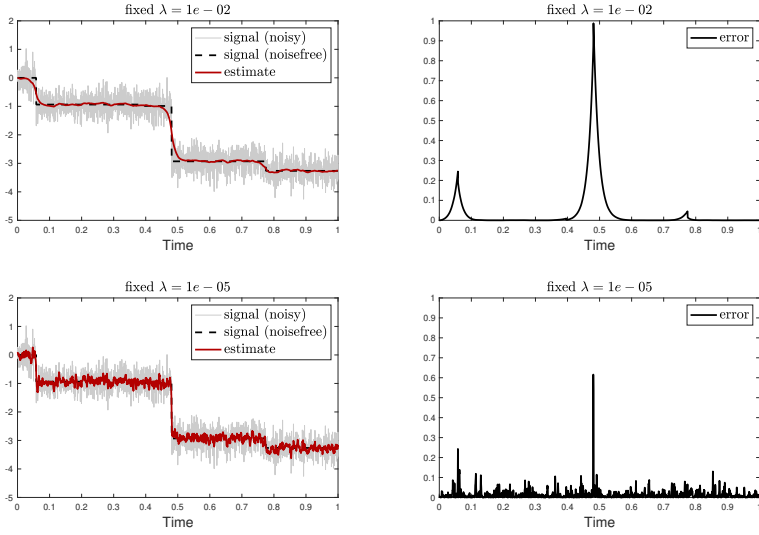
Fig. 4.12: *Comparison of different Tikhonov solutions for fixed choices of the regularization parameter λ for the signal denoising example.*
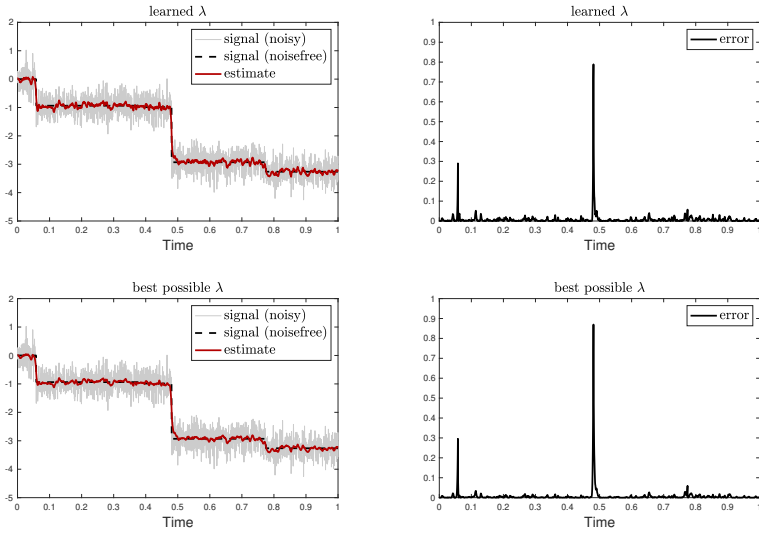

Fig. 4.13: *Comparison of the learned to best possible Tikhonov solutions for choices of the regularization parameter λ. The learned Tikhonov regularized solution corresponds to the resulting one of the SGD method Algorithm 1 for the signal denoising example.*

cases we see an improvement over the case $\lambda = 1$ which is what we expected and have seen throughout our experiments. This is verified through Figure 4.10 where we see oscillations of the learned λ_k around the true λ_*, until approximately 100 iterations where it starts to become stable. Finally from Figure 4.11 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 $\nabla_\lambda f(\lambda_k, Z^{(k)})$. Figure 4.11 also shows the histrogram of the stochastic gradient and its rare realized large values.

4.4. Signal denoising example.   We now consider implementing our methods on image denoising, which is discussed in Section 1 and Subsection 1.1.2. We are interested in denoising a 1D compound Poisson process of the form

$$u_t = \sum_{i=1}^{N_t} X_i, \tag{4.16}$$

where $(N_t)_{t \in [0,T]}$ is a Poisson process, with rate $r > 0$ and $(X_i)_{i=1}^{N_t}$ are i.i.d. random variables representing the jump size. Here, we have chosen $X_1 \sim \mathcal{N}(0,1)$. We consider the task of recovering a perturbed signal of the form (4.16) through Tikhonov regularization with different choices of regularization parameter λ. In particular, the observed signal $u = (u_{t_1}, \ldots, u_{t_d})^\top \in \mathbf{R}^d$ is perturbed by white noise

$$y_{t_i} = u_{t_i} + \eta_{t_i}, \tag{4.17}$$

where $t_i \in \{1/d \cdot T, 2/d \cdot T, \ldots, T\}$ and $\eta_{t_i} \sim \mathcal{N}(0, \sigma^2)$ are i.i.d. random variables, and the Tikhonov estimate corresponding to the lower level problem of (1.7) for given regularization parameter $\lambda > 0$ is defined by

$$u_\lambda(y) = (\Gamma^{-1} + \lambda L^{-1})^{-1} \Gamma^{-1} y, \tag{4.18}$$

with given regularization matrix $L \in \mathbf{R}^{d \times d}$ and $y = (y_1, \ldots, y_d)^\top \in \mathbf{R}^d$. We assume to have access to training data $(u^{(j)}, y^{(j)})_{j=1}^n$ of (4.17) and choose the regularization parameter $\widehat{\lambda}$ according to Algorithm 1. Further, we compare the resulting estimate of the signal

$$y_{obs} = u^\dagger + \eta,$$

to fixed choices of $\lambda \in \{0.01, 0.00001\}$ and to the best possible choice $\lambda_* = \arg\min_\lambda \|u_\lambda(y_{obs}) - u^\dagger\|^2$.

For the experiment we set the rate of jumps $r = 10$ and consider the signal observed up to time $T = 1$ at $d = 1000$ observation points. For Algorithm 1, we use a training data set of size $n = 500$, we set an initial value $\lambda_0 = 0.001$ and step size $\beta_k = 0.001 k^{-1}$. The Tikhonov solution (4.18) has been computed with a second-order regularization matrix $L = \Delta^{-1}$. As we can see from our results the value of $\lambda = 0.001$ oversmoothens the estimate in comparison with $\lambda = 0.00001$. This is shown in Figure 4.12. However comparing fixed λ with the learned λ in Figure 4.13 we see an improvement, closer to the best possible λ, which is verified further through Table 4.1, where we can see the MSE over the time interval. Both Figure 4.12 and Figure 4.13 show on the right-hand side the pointwise squared error over time.

$\boldsymbol{\lambda}$	$1e-02$	$1e-05$	$\overline{\lambda}_n$	λ_*
error	0.0378	0.0134	0.0077	0.0073

Table 4.1: *MSE over time of the reconstruction for different choices of the regularization parameter for signal denoising example.*

5. Conclusion

In this work we have provided new insights into the theory of bilevel learning for inverse problems. In particular our focus was on deriving statistical consistency results with respect to the data limit of the regularization parameter λ. This was considered for both the offline and online representations of the bilevel problem. For the online version we used and motivated stochastic gradient descent as the choice of optimizer, as it is well known to reduce the computational time required compared to other methodologies. To test our theory we ran numerical experiments on various PDEs which not only verified the theory, but clarified that adapting the regularization parameter λ outperforms that of a fixed value. Our results in this article provide numerous directions for future, both practically and analytically.

- One direction is to consider a fully Bayesian approach, or understanding, to bilevel learning. In the context of statistical inverse problems, this could be related to treating λ as a hyperparameter of the underlying unknown. This is referred to as hierarchical learning [44] which aims to improve the overall accuracy of the reconstruction [1, 25].

- Another potential direction is to understand statistical consistency from other choices of regularization. Answering this for other penalty terms, such as L_1, total variation and adversarial [42] (based on neural networks), is of importance and interest in numerous applications [2]. A potential first step in this direction would be to consider the well-known elastic-net regularization [28], which combines both L_1 and Tikhonov regularization. Of course to consider this one would need to modify the assumptions on convexity.

- One could propose using alternative optimizers, which provide a lower computational cost. A natural choice would be derivative-free optimization [26]. One potential optimizer could be ensemble Kalman inversion [13, 14], a recent derivative-free methodology, which is of particular interest to the authors. In particular as EKI has been used in hierarchical settings [11, 12], the reduction in cost could be combined with the hierarchical motivation discussed above.

- As a final avenue, it would be of interest to make extensions to both the infinite-dimensional setting, and bilevel problems involving multiple hyperparameters. For the former, initial work has been done on this, in the context of imaging [10, 43]. For the latter, to the best of the authors' knowledge, this has been primarily considered in machine learning applications [4, 29], which should be extendable in our bilevel data-driven setting.

Appendix A. Proofs of offline consistency analysis.

A.1. General framework. We start with the proof for the general framework:

Proof. (Proof of Proposition 2.1.) To simplify the mathematical notation, we use z to denote the data couple (u, y), and use f to denote the data loss function

$$f(\lambda, z) = \mathcal{L}_{\mathcal{U}}(u_\lambda(y), u).$$

When $\hat{\lambda}_n \in \mathcal{D}$, we apply the fundamental theorem of calculus on $\partial_\lambda \widehat{F}_n$, and find

$$\partial_\lambda \widehat{F}_n(\lambda_*) = \partial_\lambda \widehat{F}_n(\hat{\lambda}) + \int_0^1 \partial_\lambda^2 \widehat{F}_n(s\lambda_* + (1-s)\hat{\lambda}_n)(\lambda_* - \hat{\lambda}_n)ds = A_F(\hat{\lambda}_n - \lambda_*),$$

where

$$A_F := \int_0^1 \partial_\lambda^2 \widehat{F}_n((1-s)\hat{\lambda}_n + s\lambda_*)ds \succeq c_o I.$$

Note that

$$\mathbf{0} = \partial_\lambda F(\lambda_*) = \partial_\lambda F(\lambda_*) - \partial_\lambda \widehat{F}_n(\lambda_*) + \partial_\lambda \widehat{F}_n(\lambda_*)$$
$$= A_F(\hat{\lambda}_n - \lambda_*) + \partial_\lambda F(\lambda_*) - \partial_\lambda \widehat{F}_n(\lambda_*).$$

We can reorganize this as

$$-\left(\partial_\lambda \widehat{F}_n(\lambda_*) - \partial_\lambda F(\lambda_*)\right) = A_F(\hat{\lambda}_n - \lambda_*),$$

As a consequence, we now have a formula for the point estimation error $\lambda_* - \hat{\lambda}_n$.

$$\|\lambda_* - \hat{\lambda}\| = \left\|A_F^{-1}\left(\partial_\lambda \widehat{F}_n(\lambda_*) - \partial_\lambda F(\lambda_*)\right)\right\| \leq c_0^{-1}\left\|\partial_\lambda \widehat{F}_n(\lambda_*) - \partial_\lambda F(\lambda_*)\right\|. \qquad (A.1)$$

Note that by using $\partial_\lambda \mathbb{E}f(\lambda, Z) = \mathbb{E}\partial_\lambda f(\lambda, Z)$, see [48, Theorem 12.5],

$$\partial_\lambda \widehat{F}_n(\lambda_*) - \partial_\lambda F(\lambda_*) = \frac{1}{n}\sum_{i=1}^n \partial_\lambda f(\lambda_*, z_i) - \mathbb{E}\partial_\lambda f(\lambda_*, Z).$$

So

$$\mathbb{E}\|\partial_\lambda \widehat{F}_n(\lambda_*) - \partial_\lambda F(\lambda_*)\|^2 = \frac{1}{n}\text{tr}(\text{Var}(\partial_\lambda f(\lambda_*, Z))).$$

And our second claim follows by Cauchy-Schwarz

$$\mathbb{E}1_{\mathcal{A}_n}\|\partial_\lambda \widehat{F}_n(\lambda_*) - \partial_\lambda F(\lambda_*)\| \leq \sqrt{\mathbb{E}\|\partial_\lambda \widehat{F}_n(\lambda_*) - \partial_\lambda F(\lambda_*)\|^2}.$$

$\square$

A.2. Formulas for the linear inverse problem. Next, we apply Proposition 2.1 to the linear inverse problem under Assumption 2.1.

The solution of the Tikhonov regularized optimization problem (without assuming any distribution on y_i and ξ_i respectively) in the linear setting can be written as

$$u_\lambda(y_i) = (A^\top\Gamma^{-1}A + \lambda C_0^{-1})^{-1}A^\top\Gamma^{-1}y_i,$$

and we consider the difference

$$\tilde{u}_i - u_\lambda(y_i) \overset{d}{=} -(A^\top\Gamma^{-1}A + \lambda C_0^{-1})^{-1}A^\top\Gamma^{-1}(Au_i - \Gamma^{1/2}\xi_i) + u_i,$$

where we have defined $\xi_i = \Gamma^{-1/2}(Au_i - y_i) \sim \mathcal{N}(0, I)$. Denote $D := C_0^{1/2}A^\top\Gamma^{-1/2}$, $\Omega_0 = C_0^{-1}$ and $v_i = \Omega_0^{1/2}u_i$, and note that

$$(A^\top\Gamma^{-1}A + \lambda C_0^{-1})^{-1}A^\top\Gamma^{-1}(Au_i - \Gamma^{1/2}\xi_i) + u_i$$

$$\begin{aligned}
&= (A^\top \Gamma^{-1} A + \lambda C_0^{-1})^{-1} (\lambda C_0^{-1} u_i + A^\top \Gamma^{-1/2} \xi_i) \\
&= C_0^{1/2} (C_0^{1/2} A^\top \Gamma^{-1} A C_0^{1/2} + \lambda I)^{-1} C_0^{1/2} (\lambda C_0^{-1} u_i + A^\top \Gamma^{-1/2} \xi_i) \\
&= C_0^{1/2} (DD^\top + \lambda I)^{-1} (\lambda v_i + D\xi_i).
\end{aligned}$$

Therefore we define

$$Q_\lambda = (DD^\top + \lambda I)^{-1},$$

and the data loss can be written as

$$\begin{aligned}
f(\lambda, z) &= \mathrm{Tr}(Q_\lambda C_0 Q_\lambda (\lambda v + D\xi)(\lambda v + D\xi)^\top) \\
&= \mathrm{Tr}(Q_\lambda C_0 Q_\lambda (\lambda^2 vv^\top + 2\lambda D\xi v^\top + D\xi \xi^\top D^\top)).
\end{aligned}$$

We further define the following quantities

$$P_1 = Q_\lambda C_0 Q_\lambda, \quad P_2 = \frac{\partial P_1}{\partial \lambda} = -(Q_\lambda^2 C_0 Q_\lambda + Q_\lambda C_0 Q_\lambda^2),$$

$$P_3 = \frac{\partial P_2}{\partial \lambda} = 2(Q_\lambda^3 C_0 Q_\lambda + Q_\lambda^2 C_0 Q_\lambda^2 + Q_\lambda C_0 Q_\lambda^3),$$

$$P_4 = \frac{\partial P_3}{\partial \lambda} = -6(Q_\lambda^4 C_0 Q_\lambda + Q_\lambda^3 C_0 Q_\lambda^2 + Q_\lambda^2 C_0 Q_\lambda^3 + Q_\lambda C_0 Q_\lambda^4).$$

Note that $\|Q_\lambda\| \leq \lambda^{-1}$, so we have

$$|\mathrm{Tr}(Q_\lambda^k C_0 Q_\lambda^j)| = |\mathrm{Tr}(C_0 Q_\lambda^{j+k})| \leq \frac{1}{\lambda^{j+k}} \mathrm{Tr}(C_0),$$

$$\|Q_\lambda^k C_0 Q_\lambda^j\| \leq \|Q_\lambda\|^{j+k} \|C_0\| \leq \frac{1}{\lambda^{j+k}} \|C_0\|,$$

$$\|Q_\lambda^k C_0 Q_\lambda^j\|_F \leq \|Q_\lambda^k\| \|C_0 Q_\lambda^j\|_F \leq \|Q_\lambda\|^{j+k} \|C_0\|_F \leq \frac{1}{\lambda^{j+k}} \|C_0\|_F.$$

In conclusion, for function T being $T(A) = |\mathrm{Tr}(A)|$ or $T(A) = \|A\|$ or $T(A) = \|A\|_F$, we all have

$$T(P_k) \leq (\tfrac{2}{\lambda})^{k+1} T(C_0).$$

In particular, under Assumption 2.2, $T(P_k)$ will be bounded by constants independent of the dimension.

Using these notations, we have

$$\partial_\lambda f(\lambda, z) = \mathrm{Tr}\left(P_2(\lambda^2 vv^\top + 2\lambda D\xi v^\top + D\xi \xi^\top D^\top) + P_1(2\lambda vv^\top + 2D\xi v^\top)\right).$$

$$\partial_\lambda^2 f(\lambda, z) = \mathrm{Tr}\left(P_3(\lambda^2 vv^\top + 2\lambda D\xi v^\top + D\xi \xi^\top D^\top) + 4P_2(\lambda vv^\top + D\xi v^\top) + 2P_1 vv^\top\right).$$

$$\partial_\lambda^3 f(\lambda, z) = \mathrm{Tr}\left(P_4(\lambda^2 vv^\top + 2\lambda D\xi v^\top + D\xi \xi^\top D^\top) + 6P_3(\lambda vv^\top + D\xi v^\top) + 6P_2 vv^\top\right).$$

A.3. Pointwise consistency analysis.   To apply Proposition 2.1, it is necessary to show that the gradient of $\widehat{F}_n(\lambda)$ is a good approximation of $\partial F(\lambda)$ at $\lambda = \lambda_*$ with high probability. This is actually true for general λ.

To show this, we start by showing that the sample covariances are consistent.

LEMMA A.1.   *Let $z_i = (z_1^i, \ldots, z_d^i)^\top \in \mathbf{R}^d$ with z_l^i i.i.d. random variables in $i = 1, \ldots, n$ as well as in $l = 1, \ldots, d$ with $\mathbb{E}z_1^1 = 0$, $\mathbb{E}|z_1^1|^2 = 1$ and*

$$\sup_{p \geq 1} p^{-1/2} \mathbb{E}[|z_1^1|^p]^{1/p} \leq C_v,$$

and $\xi = (\xi_1^i, \ldots, \xi_K^i)^\top \in \mathbf{R}^K$ with ξ_l^i i.i.d. random variables in $i = 1, \ldots, n$ as well as in $l = 1, \ldots, K$ with $\mathbb{E}\xi_1^1 = 0$, $\mathbb{E}|\xi_1^1|^2 = 1$ and

$$\sup_{p \geq 1} p^{-1/2} \mathbb{E}[|\xi_1^1|^p]^{1/p} \leq C_\xi.$$

For

$$C_n = \frac{1}{n} \sum_{i=1}^n z_i z_i^\top, \quad B_n = \frac{1}{n} \sum_{i=1}^n z_i \xi_i^\top,$$

the following holds

$$\mathbb{P}(|\mathrm{Tr}(\Sigma C_n) - \mathrm{Tr}(\Sigma)| > t) \leq 2 \exp\left(-cn \min\left(\frac{t^2}{\|\Sigma\|_F^2}, \frac{t}{\|\Sigma\|}\right)\right),$$

$$\mathbb{P}(|\mathrm{Tr}(\Sigma B_n)| > t) \leq 2 \exp\left(-cn \min\left(\frac{t^2}{\|\Sigma\|_F^2}, \frac{t}{\|\Sigma\|}\right)\right).$$

Moreover, there is a universal constant C such that

$$\mathbb{E}|\mathrm{Tr}(\Sigma C_n)|^2 \leq 2\mathrm{Tr}(\Sigma)^2 + C\left(\frac{\|\Sigma\|_F^3}{n^{3/2}} + \frac{\|\Sigma\|^3}{n^3}\right).$$

$$\mathbb{E}|\mathrm{Tr}(\Sigma B_n)|^2 \leq C\left(\frac{\|\Sigma\|_F^3}{n^{3/2}} + \frac{\|\Sigma\|^3}{n^3}\right).$$

Proof.   Note

$$\mathrm{Tr}(\Sigma z_i z_i^\top) = z_i^\top \Sigma z_i.$$

We define the block-diagonal matrix $D_\Sigma \in \mathbf{R}^{nd \times nd}$ which consists of n blocks of Σ, and $Z = [z_1; z_2; \cdots; z_n] \in \mathbf{R}^{nd}$. Note that

$$\mathrm{Tr}(\Sigma C_n) = Z^\top (\tfrac{1}{n} D_\Sigma) Z.$$

By the Hanson–Wright inequality [47, Theorem 1.1], we obtain for some constants c_0 and K,

$$\mathbb{P}(|\mathrm{Tr}(\Sigma C_n) - \mathrm{Tr}(\Sigma)| > t) \leq 2 \exp\left(-c_0 \min\left(\frac{t^2}{K^4 \|\tfrac{1}{n} D_\Sigma\|_F^2}, \frac{t}{K^2 \|\tfrac{1}{n} D_\Sigma\|}\right)\right).$$

Note that

$$\|\tfrac{1}{n} D_\Sigma\|_F^2 = \frac{1}{n^2} \|D_\Sigma\|_F^2 = \frac{1}{n} \|\Sigma\|_F^2,$$

$$\|\tfrac{1}{n}D_\Sigma\| = \frac{1}{n}\|D_\Sigma\| = \frac{1}{n}\|\Sigma\|.$$

So the first assertion is proved. For the second claim we first note that

$$\mathrm{Tr}(\Sigma\xi_i z_i^\top) = z_i^\top \Sigma\xi_i = \left[z_i^\top, \xi_i^\top\right] Q \begin{bmatrix} z_i \\ \xi_i \end{bmatrix}, \quad Q = \begin{bmatrix} 0 & \Sigma \\ 0 & 0 \end{bmatrix} \in \mathbf{R}^{(d+d_y)\times(d+d_y)}.$$

Consider then a block-diagonal matrix $D_Q \in \mathbf{R}^{n(d+y)\times n(d+d_y)}$ which consists of n blocks of Q, and $Z = [z_1; \xi_1; z_2; \xi_2; \cdots; z_n; \xi_n] \in \mathbf{R}^{n(d+d_y)}$. Then we can verify that

$$\mathrm{Tr}(\Sigma B_n) = Z^\top (\tfrac{1}{n}D_Q)Z.$$

By the Hanson–Wright inequality [47, Theorem 1.1], we have

$$\mathbb{P}(|\mathrm{Tr}(\Sigma B_n)| > t) \leq 2\exp\left(-c\min\left(\frac{t^2}{K^4\|\tfrac{1}{n}D_Q\|_F^2}, \frac{t}{K^2\|\tfrac{1}{n}D_Q\|}\right)\right).$$

Again we note that

$$\|\tfrac{1}{n}D_Q\|_F^2 = \frac{1}{n^2}\|D_Q\|_F^2 = \frac{1}{n}\|\Sigma\|_F^2,$$

$$\|\tfrac{1}{n}D_Q\| = \frac{1}{n}\|D_Q\| = \frac{1}{n}\|Q\|,$$

and finally end up with

$$\mathbb{P}(|\mathrm{Tr}(\Sigma B_n)| > t) \leq 2\exp\left(-cn\min\left(\frac{t^2}{K^4\|\Sigma\|_F^2}, \frac{t}{K^2\|\Sigma\|}\right)\right).$$

For the bounds of second moments, let $T = \mathrm{Tr}(\Sigma C_n) - \mathrm{Tr}(\Sigma)$, then

$$\mathbb{E}|\mathrm{Tr}(\Sigma C_n)|^2 \leq 2|\mathrm{Tr}(\Sigma)|^2 + 2\mathbb{E}T^2.$$

Then note that by the probability bound,

$$\begin{aligned}
\mathbb{E}T^2 &\leq \int_0^\infty t^2 \mathbb{P}(|T| > t)dt \\
&\leq 2\int t^2 \exp\left(-cn\min\left(\frac{t^2}{\|\Sigma\|_F^2}, \frac{t}{\|\Sigma\|}\right)\right)dt \\
&\leq 2\int_0^\infty t^2 \exp\left(-cn\frac{t^2}{\|\Sigma\|_F^2}\right)dt + 2\int_0^\infty t^2 \exp\left(-cn\frac{t}{\|\Sigma\|}\right)dt.
\end{aligned}$$

We let $s = \frac{t\sqrt{cn}}{\|\Sigma\|_F}$ and find

$$\int_0^\infty t^2 \exp\left(-cn\frac{t^2}{\|\Sigma\|_F^2}\right)dt = \frac{\|\Sigma\|_F^3}{(cn)^{\frac{3}{2}}}\int_0^\infty s^2 \exp\left(-s^2\right)ds.$$

We then let $s = cn\frac{t}{\|\Sigma\|}$ and find

$$\int_0^\infty t^2 \exp\left(-cn\frac{t}{\|\Sigma\|}\right)dt = \frac{\|\Sigma\|^3}{(cn)^3}\int_0^\infty s^2 \exp(-s)ds.$$

So there is a universal constant C such that

$$\mathbb{E}T^2 \leq \frac{C}{2}\left(\frac{\|\Sigma\|_F^3}{n^{3/2}} + \frac{\|\Sigma\|^3}{n^3}\right).$$

The bound for $\mathbb{E}|\mathrm{Tr}(\Sigma B_n)|$ follows identically.   □

By the previous result we obtain the following convergence results.

LEMMA A.2.   *The empirical loss function $\widehat{F}_n$ is C^3 in λ, and for any $\lambda \in (\lambda_l, \lambda_r)$, there exist constants C, $c > 0$ such that for all $\varepsilon > 0$*

$$\mathbb{P}(|\partial_\lambda \widehat{F}_n(\lambda) - (1 - \lambda/\lambda_*)\mathrm{Tr}(P_2 DD^\top)| > \varepsilon) \leq C\exp(-nc\min\{\varepsilon, \varepsilon^2\}),$$

and

$$\mathbb{P}(|\partial_\lambda^2 \widehat{F}_n(\lambda) - \mathrm{Tr}\left((\frac{\lambda^2}{\lambda_*} - \lambda)P_3 + (\frac{4\lambda}{\lambda_*} - 3\lambda)P_2 + \frac{2}{\lambda_*}P_1\right)| > \varepsilon) \leq C\exp(-cn\min\{\varepsilon, \varepsilon^2\}).$$

Under Assumption 2.2, both C and c are independent of ambient dimension d.

Proof.   Since $\widehat{F}_n(\lambda) = \frac{1}{n}\sum_{i=1}^n f(\lambda, \zeta_i)$, if we let

$$C_v = \frac{1}{n}\sum_{i=1}^n v_i v_i^\top, \quad B = \frac{1}{n}\sum_{i=1}^n \xi_i v_i^\top, \quad C_\xi = \frac{1}{n}\sum_{i=1}^n \xi_i \xi_i^\top,$$

then

$$\partial_\lambda \widehat{F}_n(\lambda) = \mathrm{Tr}\left(P_2(\lambda^2 C_v + 2\lambda DB + DC_\xi D^\top) + P_1(2\lambda C_v + 2DB)\right) \tag{A.2}$$

$$= \mathrm{Tr}\left((P_2 \lambda^2 + 2\lambda P_1)C_v + (2P_1 + 2\lambda P_2)DB + D^\top P_2 DC_\xi\right) \tag{A.3}$$

We note that

$$\mathbb{E}\partial_\lambda \widehat{F}_n(\lambda) = \mathrm{Tr}\left(P_2(\lambda^2/\lambda_* I + DD^\top) + 2\lambda P_1/\lambda_*\right)$$

$$= \mathrm{Tr}\left(-(Q_\lambda^2 C_0 Q_\lambda + Q_\lambda C_0 Q_\lambda^2)(\frac{\lambda^2}{\lambda_*} I + DD^\top) + 2\frac{\lambda}{\lambda_*} Q_\lambda C_0 Q_\lambda\right)$$

$$= \mathrm{Tr}\left((1 - \frac{\lambda}{\lambda_*})P_2 DD^\top\right).$$

Moreover, in (A.2), $\partial_\lambda \widehat{F}_n$ can be written as sum of $\mathrm{Tr}(\Sigma_1 C_v), \mathrm{Tr}(\Sigma_2 B)$ and $\mathrm{Tr}(\Sigma_3 C_\xi)$ for certain matrices Σ such as

$$\Sigma_1 = (P_2 \lambda^2 + 2\lambda P_1), \quad \Sigma_2 = (2P_1 + 2\lambda P_2)D, \quad \Sigma_3 = D^\top P_2 D.$$

Note that for any random variables A_k

$$\mathbb{P}\left(|\sum_{k=1}^m (A_k - \mathbb{E}A_k)| > \varepsilon\right) \leq \sum_{k=1}^m \mathbb{P}(|A_k - \mathbb{E}A_k| > \varepsilon/m).$$

Therefore we can apply Lemma A.1 at each trace term, and bound its probability of deviating from its mean. Therefore, we can find constants C_1, c such that

$$\mathbb{P}(|\partial_\lambda \widehat{F}_n(\lambda) - (1 - \lambda/\lambda_*)\mathrm{Tr}(P_2 DD^\top)| > \varepsilon) \leq C_1 \exp\left(-cn\min(\varepsilon^2, \varepsilon)\right).$$

If Assumption 2.2 is assumed, note that

$$\|Q_\lambda\| \le \frac{1}{\lambda}\|C_0\| \le \frac{1}{\lambda_l}\|C_0\|,$$

for all $\lambda \ge \lambda_l$ respectively. So using norm inequalities $\|AB\| \le \|A\|\|B\|$ and $\|AB\|_F \le \|A\|\|B\|_F$, we can verify that all $P_i, i = 1,\ldots,4$ have dimension independent operator norm and Frobenius norms. Therefore all $\Sigma_i, i = 1,\ldots,3$ have dimension independent operator norm and Frobenius norms. So C_1 and c are dimension independent.

For the second claim,

$$\partial_\lambda^2 f(\lambda, z) = \text{Tr}\bigg(P_3(\lambda^2 vv^\top + 2\lambda D\xi v^\top + D\xi\xi^\top D^\top) + 4P_2(\lambda vv^\top + D\xi v^\top) + 2P_1 vv^\top \bigg).$$

we find

$$\partial_\lambda^2 \widehat{F}_n(\lambda) = \text{Tr}\bigg((\lambda^2 P_3 + 4\lambda P_2 + 2P_1)C_v + (2\lambda P_3 + 4P_2)DB + D^\top P_3 DC_\xi \bigg). \qquad \text{(A.4)}$$

Therefore,

$$\mathbb{E}\partial_\lambda^2 \widehat{F}_n(\lambda) = \text{Tr}\big((\lambda^2 P_3 + 4\lambda P_2 + 2P_1)/\lambda_* + DD^\top P_3 \big).$$

The deviation probability can also be obtained by analyzing matrices

$$\Sigma_1' = (\lambda^2 P_3 + 4\lambda P_2 + 2P_1), \quad \Sigma_2' = (2\lambda P_3 + 4P_2)D, \quad \Sigma_3' = D^\top P_3 D.$$

Note that

$$\text{Tr}(Q_\lambda^{-1} P_3) = \text{Tr}(2Q_\lambda^2 C_0 Q_\lambda + Q_\lambda^2 C_0 Q_\lambda + Q_\lambda C_0 Q_\lambda^2 + 2Q_\lambda C_0 Q_\lambda^2) = -3P_2.$$

$$\text{Tr}(\lambda P_2 + 2P_1) = \text{Tr}((Q_\lambda^{-1} - \lambda I)Q_\lambda^2 C_0 Q_\lambda + Q_\lambda C_0 Q_\lambda^2(Q_\lambda^{-1} - \lambda I)) = -\text{Tr}(DD^\top P_2).$$

So the average can also be written as

$$\begin{aligned}
\mathbb{E}\partial_\lambda^2 \widehat{F}_n(\lambda) &= \text{Tr}\big((\lambda(\lambda I + DD^\top)P_3 + 4\lambda P_2 + 2P_1)/\lambda_* + (1 - \lambda/\lambda_*)DD^\top P_3 \big) \\
&= \text{Tr}\big((-3\lambda P_2 + 4\lambda P_2 + 2P_1)/\lambda_* + (1 - \lambda/\lambda_*)DD^\top P_3 \big) \\
&= \text{Tr}\big(-DD^\top P_2/\lambda_* + (1 - \lambda/\lambda_*)DD^\top P_3 \big) \\
&= \text{Tr}\big(DD^\top((1 - \lambda/\lambda_*)P_3 - P_2/\lambda_*) \big).
\end{aligned}$$

Likewise, we can obtain

$$\partial_\lambda^3 \widehat{F}_n(\lambda) = \text{Tr}\bigg((\lambda^2 P_4 + 6\lambda P_3 + 6P_2)C_v + (2\lambda P_4 + 6P_2)DB + D^\top P_4 DC_\xi \bigg),$$

and

$$\partial_\lambda^3 \widehat{F}_n(\lambda) = \text{Tr}\bigg((\lambda^2 P_4 + 6\lambda P_3 + 6P_2)/\lambda_* + D^\top P_4 D \bigg). \qquad \text{(A.5)}$$

$$\square$$

REMARK A.1.   It is worthwhile to note that

$$\partial_\lambda^2 F(\lambda) = \mathbb{E}\partial_\lambda^2 \widehat{F}_n(\lambda) = \text{Tr}\big(DD^\top((1 - \lambda/\lambda_*)P_3 - P_2/\lambda_*) \big),$$

is not always positive, and it can be negative if λ 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.

A.4. Consistency analysis within an interval.   To apply Proposition 2.1, it is also necessary to show that the $\widehat{F}_n(\lambda)$ is strongly convex in a local region/interval. This can be done using a chaining argument in probability theory.

First, we show that the empirical loss function has bounded derivatives with high probability.

LEMMA A.3.   *There exists an $L>0$ and c that only depend on $\mathrm{Tr}(C_0), \|C_0\|_F$ and $\|D\|$ such that the following holds true*

$$\mathbb{P}\left(\max_{\lambda_l \leq \lambda \leq \lambda_u} |\partial_\lambda^k \widehat{F}_n(\lambda)| > L, k=1,2,3\right) \leq 6\exp(-nc).$$

Under Assumption 2.2, L and c are dimension-independent constants.

Proof.   Recall that $\|Q_\lambda\| \leq \frac{1}{\lambda_l}$. From (A.2), (A.4) and (A.5), and Lemma A.4, we have

$$\mathbb{P}(\max_{\lambda_l \leq \lambda \leq \lambda_u} |\partial_\lambda^k \widehat{F}_n(\lambda) - \mathbb{E}\widehat{F}_n(\lambda)| > t) \leq 2\exp\left(-cn\min\left(\frac{t^2}{\|\Sigma_k\|_F^2}, \frac{t}{\|\Sigma_k\|}\right)\right),$$

for each $k=1,2,3$. So using union bound we find that

$$\mathbb{P}(\max_{\lambda_l \leq \lambda \leq \lambda_u} |\partial_\lambda^k \widehat{F}_n(\lambda) - \mathbb{E}\widehat{F}_n(\lambda)| > t, k=1,2,3) \leq 6\exp\left(-cn\min\left(\frac{t^2}{\|\Sigma_k\|_F^2}, \frac{t}{\|\Sigma_k\|}\right)\right).$$

Here each Σ_k consists of matrices of form $P_j S$ or $S P_j$ where $j=1,2,3,4$ and $S=I, D$ or $DD^\top$. Then because

$$\|P_j S\| \leq \|P_j\|\|S\| \leq \frac{\|C_0\|\|S\|}{\lambda_l^{j+1}}, \quad \|P_j S\|_F \leq \frac{\|C_0\|_F\|S\|}{\lambda_l^{j+1}}.$$

So we see that c can depend on $\|C_0\| \leq \mathrm{Tr}(C_0), \|C_0\|_F$ and $\|D\|$. Meanwhile, $\mathbb{E}\partial_\lambda^k \widehat{F}_n$ is a linear sum of some $\mathrm{Tr}(P_j)$. While

$$|\mathrm{Tr}(P_j)| \leq (\tfrac{2}{\lambda_l})^{j+1}\mathrm{Tr}(C_0).$$

So L can also be taken as a constant that depends only on $\mathrm{Tr}(C_0), \|C_0\|_F$ and $\|D\|$. This concludes our proof.   □

Next, we show that if a function is bounded at each fixed point with high probability, it is likely to be bounded on a fixed interval if it is Lipschitz.

LEMMA A.4.   *Suppose $f_n(\lambda)$ is a C^1 function of λ and the following is true for some interval $I=[\lambda_l, \lambda_u]$*

$$\mathbb{P}(f_n(\lambda) > a) \leq C\exp(-nc_a) \quad \forall \lambda_l \leq \lambda \leq \lambda_u.$$

Then

$$\mathbb{P}\left(\max_{\lambda \in I} f_n(\lambda) > 2a, \max_{\lambda \in I}|\partial f_n(\lambda)| \leq M\right) \leq a^{-1}|\lambda_u - \lambda_l| MC\exp(-nc_a).$$

Alternatively, suppose $f_n(\lambda)$ is a C^1 function of λ and the following is true for some interval I

$$\mathbb{P}(f_n(\lambda) < a) \leq \exp(-nc_a) \quad \forall \lambda_l \leq \lambda \leq \lambda_u.$$

Then

$$\mathbb{P}\left(\min_{\lambda\in I} f_n(\lambda) < a/2, \max_{\lambda\in I} |\partial f_n(\lambda)| \leq M\right) \leq 2a^{-1}|\lambda_u - \lambda_l| MC \exp(-nc_a).$$

Proof.   Pick $\lambda_i = \lambda_l + \frac{2a}{|M|} i$ for $i = 0, \ldots, \lfloor \frac{|\lambda_u - \lambda_l| M}{2a} \rfloor$. Then $\lambda_l \leq \lambda_i \leq \lambda_u$, and for any $\lambda_l \leq \lambda \leq \lambda_u$, $|\lambda - \lambda_i| \leq \frac{a}{M}$ for some λ_i. Note that if $|\partial f_n(\lambda)| \leq M$, and $f_n(\lambda_i) \leq a$, for all i, then for any $\lambda_l \leq \lambda \leq \lambda_u$,

$$f_n(\lambda) \leq f_n(\lambda_i) + (\lambda_i - \lambda)\partial_\lambda f_n(\lambda) \leq a + \frac{a}{M}M = 2a.$$

Consequentially, by union bound

$$\mathbb{P}\left(\min_{\lambda\in I} f_n(\lambda) > 2a, \max_{\lambda\in I} |\partial f_n(\lambda)| \leq M\right) \leq \mathbb{P}\left(f_n(\lambda_i) > a \text{ for some } i\right)$$
$$\leq a^{-1}|\lambda_u - \lambda_l| MC \exp(-nc_a).$$

The same argument can be applied to show the second claim, except that we choose $\lambda_i = c + \frac{a}{|M|}$. □

The next lemma indicates that the loss function is strongly convex within $\mathcal{D}$ with high probability.

LEMMA A.5.   *Assume that the largest eigenvalue of $DD^\top$ is λ_D and let $\lambda \in \mathcal{D} := [\frac{5}{6}\lambda_*, \frac{7}{6}\lambda_*]$. Then for some constants $c, C > 0$,*

$$\mathbb{P}(\min_{\lambda\in\mathcal{D}} \partial_\lambda^2 \widehat{F}_n(\lambda) < H_*/4) \leq \frac{C}{\min\{H_*, 1\}} \exp\left(-cn\min(H_*^2, H_*, 1)\right),$$

with

$$H_* = H_*(\lambda_D, \lambda_*) = \frac{\lambda_D^2}{(\lambda_D + 2\lambda_*)^2 \lambda_* \|A^\top \Gamma^{-1} A\|} > 0.$$

Proof.   Denote

$$\mathcal{A} = DD^\top((1 - \lambda/\lambda_*)P_3 - P_2/\lambda_*),$$

and v_i being the eigenvector of $DD^\top$ corresponds to eigenvalue λ_i. Note that v is also the eigenvector Q_λ with eigenvalue $(\lambda_i + \lambda)^{-1}$, then

$$v_i^\top \mathcal{A} v_i = \left(\frac{6(1 - \lambda/\lambda_*)}{(\lambda_i + \lambda)^3} + \frac{2}{(\lambda_i + \lambda)^2 \lambda_*}\right) \lambda_i v_i^\top C_0 v_i.$$

When $\lambda \in \mathcal{D}$, if $7/6\lambda_* \geq \lambda > \lambda_*$

$$\frac{6(1 - \lambda/\lambda_*)}{(\lambda_i + \lambda)^3} + \frac{2}{(\lambda_i + \lambda)^2 \lambda_*} \geq -\frac{1}{(\lambda_i + \lambda)^2 \lambda} + \frac{2}{(\lambda_i + \lambda)^2 \lambda_*} \geq \frac{1}{(\lambda_i + \lambda)^2 \lambda_*}.$$

If $\lambda \leq \lambda_*$, the same relation also holds. Then note that if v_i are all the eigenvectors of $DD^\top$ with eigenvalues λ_i, while λ_i are decreasing,

$$\text{Tr}\mathcal{A} = \sum_{i=1}^d v_i^\top \mathcal{A} v_i \geq \sum_{i=1}^d \frac{\lambda_i v_i^\top C_0 v_i}{(\lambda_i + 2\lambda_*)^2 \lambda_*} = \frac{\lambda_1 v_1^\top C_0 v_1}{(\lambda_1 + 2\lambda_*)^2 \lambda_*}.$$

Finally, note that

$$\lambda_D = \lambda_1 = v_1^\top D D^\top v_1 = v_1^\top C_0^{1/2} A^\top \Gamma^{-1} A C_0^{1/2} v_1 \leq \|A^\top \Gamma^{-1} A\| \|C_0^{1/2} v_1\|^2.$$

So

$$\text{Tr}\mathcal{A} \geq \frac{\lambda_D^2}{(\lambda_D + 2\lambda_*)^2 \lambda_* \|A^\top \Gamma^{-1} A\|} = H_*.$$

for $\lambda \in \mathcal{D}$ and we set $\varepsilon = H_*/2 > 0$ to apply Corollary A.2. We obtain some C_1, c

$$C_1 \exp\left(-cn\min(H_*^2, H_*)\right)$$
$$\geq \mathbb{P}\left(|\partial_\lambda^2 \widehat{F}_n(\lambda) - \frac{2}{\lambda_*}\text{Tr}\left((3\lambda_* I - 2\lambda I + D D^\top)D D^\top Q_\lambda^4\right)| > H_*/2\right)$$
$$\geq \mathbb{P}\left(\partial_\lambda^2 \widehat{F}_n(\lambda) < \frac{2}{\lambda_*}\text{Tr}\left((3\lambda_* I - 2\lambda I + D D^\top)D D^\top Q_\lambda^4\right) - H_*/2\right)$$
$$\geq \mathbb{P}(\partial_\lambda^2 \widehat{F}_n(\lambda) < H_*/2).$$

By Lemma A.3 there exists an $L > 0$ and c_1 such that

$$\mathbb{P}\left(\max_{\lambda \in \mathcal{D}} |\partial_\lambda^3 \widehat{F}_n(\lambda|) > L\right) \leq 6\exp(-nc_1),$$

and by Lemma A.4 it holds true that

$$\frac{C_2}{\min(H_*, 1)} \exp\left(-cn\min(H_*^2, H_*, 1)\right)$$
$$\geq \mathbb{P}\left(\min_{\lambda \in \mathcal{D}} \partial_\lambda^2 \widehat{F}_n(\lambda) < H_*/4, \max_{\lambda \in \mathcal{D}} |\partial_\lambda^3 \widehat{F}_n(\lambda)| \leq M\right),$$

for some $C_2 > 0$. We define the sets $A_n := \{\min_{\lambda \in \mathcal{D}} \partial_\lambda^2 \widehat{F}_n(\lambda) < H_*/4\}$ and $B_n := \{\max_{\lambda \in \mathcal{D}} |\partial_\lambda^3 \widehat{F}_n(\lambda)| \leq L\}$, and we obtain

$$\mathbb{P}(\min_{\lambda \in \mathcal{D}} \partial_\lambda^2 \widehat{F}_n(\lambda) < H_*/4) = \mathbb{P}(A_n \mid B_n)\mathbb{P}(B_n) + \mathbb{P}(A_n \mid B_n^{\complement})\mathbb{P}(B_n^{\complement})$$
$$\leq \mathbb{P}(A_n \cap B_n) + \mathbb{P}(B_n^{\complement})$$
$$\leq \frac{C}{\min(H_*, 1)} \exp\left(-cn\min(H_*^2, H_*, 1)\right).$$

$\square$

The last lemma indicates that the empirical loss function is unlikely to have local minima outside $[\frac{2}{3}\lambda_*, \frac{4}{3}\lambda_*]$.

LEMMA A.6.   *Assume again that the largest eigenvalue of $D D^\top$ is λ_D. Let*

$$L_* = \frac{2\lambda_D^2}{3(\lambda_D + \lambda_u)^3 \|A^\top \Gamma^{-1} A\|}.$$

There are constants c, C such that

$$\mathbb{P}\left(\min_{\lambda_u \geq \lambda > \frac{4}{3}\lambda_*} \partial_\lambda \widehat{F}_n(\lambda) < L_*/4\right) \leq \frac{C}{\min\{L_*, 1\}} \exp\left(-cn\min\left(L_*^2, L_*, 1\right)\right),$$

and

$$\mathbb{P}\left(\min_{\frac{2}{3}\lambda_* \geq \lambda > \lambda_l} \partial_\lambda \widehat{F}_n(\lambda) > -L_*/4\right) \leq \frac{C}{\min\{L_*, 1\}} \exp\left(-cn \min\left(L_*^2, L_*, 1\right)\right).$$

Proof. We first note that with v being the leading eigenvector of $DD^\top$,

$$-\mathrm{Tr}(P_2 DD^\top) \geq v^\top (Q_\lambda^2 C_0 Q_\lambda + Q_\lambda C_0 Q_\lambda^2) v \geq 2\lambda_D \frac{v^\top C_0 v}{(\lambda_D + \lambda_u)^3}$$

$$\geq \frac{2\lambda_D^2}{(\lambda_D + \lambda_u)^3 \|A^\top \Gamma^{-1} A\|} =: 3L_*.$$

For $\lambda > \frac{4}{3}\lambda_*$ we have

$$(1 - \lambda/\lambda_*)\mathrm{Tr}(P_2 DD^\top) \geq L_*.$$

We set $\varepsilon = L_*/2$ to apply Lemma A.2 and obtain

$$C\exp\left(-nc\min(L_*^2, L_*)\right) \geq \mathbb{P}(|\partial_\lambda \widehat{F}_n(\lambda) - (1 - \lambda/\lambda_*)\mathrm{Tr}(Q_\lambda^3)| > L_*/2)$$

$$\geq \mathbb{P}(\partial_\lambda \widehat{F}_n(\lambda) < (1 - \lambda/\lambda_*)\mathrm{Tr}(Q_\lambda^3) - L_*/2)$$

$$\geq \mathbb{P}(\partial_\lambda \widehat{F}_n(\lambda) < L_*/2).$$

Similarly as in Lemma A.5, we use Lemma A.3 and Lemma A.4 to obtain the first assertion by using

$$(1 - \lambda/\lambda_*)\mathrm{Tr}(P_2 DD^\top) \leq -L_* < 0,$$

for $\lambda < \frac{2}{3}\lambda_*$. □

A.5. Summarizing argument.   Finally, we are ready to prove Theorem 2.2.

Proof. (Proof of Theorem 2.2.) Denote $\mathcal{D} = [\frac{2}{3}\lambda_*, \frac{4}{3}\lambda_*]$,

$$H_* = \frac{\lambda_D^2}{(\lambda_D + 2\lambda_*)^2 \lambda_* \|A^\top \Gamma^{-1} A\|} > 0,$$

and the events

$$B = \{\lambda_l < \hat{\lambda}_n < \lambda_u\}, \quad \mathcal{A}_n = \{\hat{\lambda}_n \in \mathcal{D}, \partial_\lambda^2 \widehat{F}_n(\lambda) \geq \tfrac{1}{4}H_* \text{ for all } \lambda \in \mathcal{D}\}.$$

First we decompose

$$\mathbb{P}(|\hat{\lambda}_n - \lambda_*| > \varepsilon, B) = \mathbb{P}(|\hat{\lambda}_n - \lambda_*| > \varepsilon, B \,|\, \mathcal{A}_n) \cdot \mathbb{P}(\mathcal{A}_n)$$

$$+ \mathbb{P}(|\hat{\lambda}_n - \lambda_*| > \varepsilon, B \,|\, \mathcal{A}_n^{\complement}) \cdot \mathbb{P}(\mathcal{A}_n^{\complement})$$

$$\leq \mathbb{P}(|\hat{\lambda}_n - \lambda_*| > \varepsilon, B \,|\, \mathcal{A}_n) + \mathbb{P}(B \cap \mathcal{A}_n^{\complement}).$$

In the last step we have used $\mathbb{P}(\hat{\lambda}_n \leq \lambda_u) = 1$. By Proposition 2.1

$$\mathbb{P}(|\hat{\lambda}_n - \lambda_*| > \varepsilon, B \,|\, \mathcal{A}_n) \leq \mathbb{P}\left(|\partial_\lambda \widehat{F}_n(\lambda_*) - \partial_\lambda F(\lambda_*)| > \frac{1}{4}H_*\epsilon, B\right)$$

$$= \mathbb{P}\left(|\partial_\lambda \widehat{F}_n(\lambda_*)| > \frac{1}{4}H_*\epsilon, B\right),$$

which can be bounded by Lemmas A.2 and A.4

$$\mathbb{P}(|\hat{\lambda}_n - \lambda_*| > \varepsilon \,|\, \mathcal{A}_n) \le C_1 \exp(-nc_1 \min\{\epsilon, \epsilon^2\}),$$

for some C_1, c_1.

We bound the probability $\mathbb{P}(\mathcal{A}_n^{\complement})$ by

$$\mathbb{P}(B \cap \mathcal{A}_n^{\complement}) \le \mathbb{P}(B, \hat{\lambda}_n \notin \mathcal{D}) + \mathbb{P}(\{\partial_\lambda^2 \widehat{F}_n(\lambda) \ge H_*/4 \text{ for all } \lambda \in \mathcal{D}\}^{\complement}),$$

and study both terms separately. Note first, by Lemma A.6, for some constants $C_2, c_2 > 0$ the following holds

$$\begin{aligned}
\mathbb{P}(B, \hat{\lambda}_n \notin \mathcal{D}) &\le \mathbb{P}(\partial_\lambda \widehat{F}_n(\lambda) = 0 \text{ for some } \lambda \in (\lambda_l, \lambda_u) \setminus \mathcal{D}) \\
&\le C_2 \exp(-c_2 n).
\end{aligned}$$

Second, by Lemma A.5, for some constants $C_3, c_3 > 0$ we obtain

$$\mathbb{P}(\{\partial_\lambda^2 \widehat{F}_n(\lambda) \ge H_*/4 \text{ for all } \lambda \in \mathcal{D}\}^{\complement}) \le C_3 \exp(-c_3 n).$$

So there exist some constants $C_*, c_* > 0$ such that

$$\mathbb{P}(|\hat{\lambda}_n - \lambda_*| > \epsilon) \le C_* \exp(-c_* n \min(\epsilon, \epsilon^2)).$$

$\square$

Proof. (**Proof of Proposition 2.2.**) As for the last claim, note that

$$u_\lambda(y) = (A^\top \Gamma^{-1} A + \lambda C_0^{-1})^{-1} A^\top \Gamma^{-1} y = C_0^{1/2} Q_\lambda y,$$

moreover by the chain rule, there is a w between λ and λ_*, so that

$$\|Q_\lambda - Q_{\lambda_*}\| = |\lambda - \lambda_*| \|Q_w^2\| \le \frac{|\lambda - \lambda_*|}{\lambda_l^2},$$

and

$$\begin{aligned}
\mathbb{E}_y \|u_\lambda(y) - u_{\lambda_*}(y)\|^2 &= \mathbb{E}_y \operatorname{Tr}(C_0(Q_\lambda - Q_{\lambda_*}) y y^T (Q_\lambda - Q_{\lambda_*})) \\
&= \operatorname{Tr}(C_0(Q_\lambda - Q_{\lambda_*})(A C_0 A^\top / \lambda_* + \Gamma)(Q_\lambda - Q_{\lambda_*})) \\
&\le \frac{|\lambda - \lambda_*|^2}{\lambda_l^4} \|A C_0 A^\top / \lambda_* + \Gamma\| \operatorname{Tr}(C_0).
\end{aligned}$$

This concludes our proof. $\square$

Appendix B. Online consistency analysis.

B.1. Stochastic gradient decent.
We start by verifying Lemma 3.2.

Proof. (**Proof of Lemma 3.2.**) We apply the implicit function theorem to prove this statement. For fixed $y \in \mathbf{R}^K$, we define the function

$$\varphi(\lambda, u) := \nabla_u \Psi(\lambda, u, y).$$

Since $(\lambda, u) \mapsto \Psi(\lambda, u, y)$ is strictly convex, we have that for all (λ, u) near $(\lambda_0, u_{\lambda_0})$ the Jacobian of φ w.r.t. u is invertible, i.e.

$$D_u \varphi(\lambda, u) = \nabla_u^2 \Psi(\lambda, u, y) > 0.$$

Set $\bar{\lambda} \in \mathbf{R}^d$ arbitrary, then for $\bar{u} = u_{\bar{\lambda}}(y)$ it holds true that

$$\varphi(\bar{\lambda}, \bar{u}) = 0,$$

and by the implicit function theorem there exists an open neighborhood $\mathcal{D} \subset \mathbf{R}^d$ of λ_0 with $\bar{\lambda} \in \mathcal{D}$ such that there exists a unique continuously differentiable function $\bar{U} : \mathcal{D} \to \mathbf{R}^d$ with $\bar{U}(\bar{\lambda}) = \bar{u}$ and

$$\varphi(\lambda, \bar{U}(\lambda)) = 0,$$

for all $\lambda \in \Lambda$, i.e. $\bar{U}$ maps all $\lambda \in \Lambda$ to the corresponding regularized solution $\bar{U}(\lambda) = u_\lambda(y)$. Further, the partial derivatives of $\bar{U}$ are given by

$$\frac{\partial \bar{U}}{\partial \lambda_i}(\lambda) = -\left[D_u \varphi(\lambda, \bar{U}(\lambda))\right]^{-1} \left[\frac{\partial \varphi}{\partial \lambda_i}(\lambda, \bar{U}(\lambda))\right].$$

Since the choice of $\bar{\lambda} \in \mathbf{R}^d$ is arbitrary, it follows that $\lambda \mapsto u_\lambda(y)$ is continuously differentiable with derivative given by

$$\partial_\lambda u_\lambda(y) = -\left(\partial_u^2 \left[\Psi(\lambda, u_\lambda(y), y)\right]\right)^{-1} \partial_{\lambda u}^2 \left[\Psi(\lambda, u_\lambda(y), y)\right].$$

The computation of $\partial_\lambda f$ can be obtained by the chain rule.   □

B.2. General consistency analysis framework.
Proof. **(Proof of Proposition 3.1.)**
Note that

$$\Delta_{k+1} = \chi(\lambda_k - \beta_k \widetilde{\partial}_\lambda f(\lambda_k, Z_k)) - \lambda_*,$$

and apply Lemma 3.1

$$|\Delta_{k+1}|^2 = |\chi\left(\lambda_k - \beta_k \widetilde{\partial}_\lambda f(\lambda_k, Z_k)\right) - \lambda_*|^2 \leq |\lambda_k - \beta_k \widetilde{\partial}_\lambda f(\lambda_k, Z_k) - \lambda_*|^2$$
$$= |\Delta_k - \beta_k \widetilde{\partial}_\lambda f(\lambda_k, Z_k)|^2$$
$$= |\Delta_k - \beta_k \partial_\lambda F(\lambda_k, Z_k) - \beta_k \delta_k - \beta_k \xi_k|^2,$$

where

$$\delta_k = \mathbb{E}_k \widetilde{\partial} f(\lambda_k, Z) - \partial F(\lambda_k), \quad \xi_k = \widetilde{\partial}_\lambda f(\lambda_k, Z_k) - \mathbb{E}_k \widetilde{\partial}_\lambda f(\lambda_k, Z),$$

is the bias and noise in the stochastic gradient, 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 $\tau = \inf\{k \geq 0 \mid \lambda_k \in \mathcal{D}\}$. Next, we note that

$$\mathbb{E}_k |\partial f(\lambda_k, Z_k)|^2 = |\partial_\lambda F(\lambda_k) + \delta_k|^2 + \mathbb{E}_k |\xi_k|^2.$$

So if $\tau \geq k$,

$$\mathbb{E}_k |\Delta_{k+1}|^2 \leq |\Delta_k|^2 - 2\beta_k \Delta_k^T (\partial_\lambda F(\lambda_k) + \delta_k) + \beta_k^2 |\partial_\lambda F(\lambda_k) + \delta_k|^2 + \mathbb{E}_k |\xi_k|^2$$
$$\leq |\Delta_k|^2 - 2\beta_k \Delta_k^T \partial_\lambda F(\lambda_k) + 2\beta_k |\Delta_k||\delta_k| + \beta_k^2(a + b|\Delta_k|^2)$$
$$\leq |\Delta_k|^2 - 2c\beta_k |\Delta_k|^2 + \frac{1}{2} c\beta_k |\Delta_k|^2 + \frac{2}{c}\beta_k |\delta_k|^2 + \beta_k^2 a + b\beta_k^2 |\Delta_k|^2$$
$$\leq (1 - 1.5c\beta_k + b\beta_k^2)|\Delta_k|^2 + (a\beta_k + 2\alpha_k/c)\beta_k.$$

Since $\beta_k < c/2b$, we have

$$\mathbb{E}_k 1_{\tau \geq k+1} |\Delta_{k+1}|^2 \leq \mathbb{E}_k 1_{\tau \geq k} |\Delta_{k+1}|^2 \leq 1_{\tau \geq k} (1 - c\beta_k)|\Delta_k|^2 + (a\beta_k + 2\alpha_k/c)\beta_k.$$

Let $Q_k = 1_{\tau_k \geq k} |\Delta_k|^2$, then we 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 Gronwall's inequality

$$\mathbb{E}Q_n \leq a \sum_{k=1}^{n} \left(\prod_{j=k+1}^{n} (1 - c\beta_j)\beta_k^2 \right) + \frac{2}{c} \sum_{k=1}^{n} \left(\prod_{j=k+1}^{n} (1 - c\beta_j)\beta_k \alpha_k \right) + \exp\left(-c \sum_{j=1}^{n} \beta_j \right) \mathbb{E}Q_0. \tag{B.1}$$

Next we look at the second term of (B.1). 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, (B.1) 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 (B.1) 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,

$$\sum_{k=1}^{k_0} \prod_{j=k+1}^{n} (1 - c\beta_j)\beta_k^2 \leq \sum_{k=1}^{k_0} \prod_{j=k_0+1}^{n} (1 - c\beta_j)\beta_k^2$$

$$\leq \prod_{j=k_0+1}^{n} (1 - c\beta_j) \sum_{k=1}^{\infty} \beta_k^2 \leq C_n.$$

Also

$$\sum_{k=k_0+1}^{n}\prod_{j=k+1}^{n}(1-c\beta_j)\beta_k^2 \leq \frac{1}{c}\beta_{k_0}\sum_{k=1}^{k_0}\prod_{j=k+1}^{n}(1-c\beta_j)c\beta_k$$

$$\leq \frac{1}{c}\beta_{k_0}\sum_{k=1}^{k_0}\left(\prod_{j=k+1}^{n}(1-c\beta_j)-\prod_{j=k}^{n}(1-c\beta_j)\right)$$

$$\leq \frac{1}{c}\beta_{k_0}=C_n.$$

The sum of the previous two inequalities leads to

$$\sum_{k=1}^{n}\left(\prod_{j=k+1}^{n}(1-c\beta_j)\beta_k^2\right)\leq 2C_n.$$

Finally

$$\exp(-c\sum_{j=1}^{n}\beta_j)\mathbb{E}Q_0 \leq \exp(-c\sum_{j=k_0+1}^{n}\beta_j)\mathbb{E}Q_0 \leq C_n.$$

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 increase to ∞, and β_{k_n} will decay to zero. Meanwhile,

$$C_n \leq \min\left\{\prod_{j=k+1}^{n}(1-c\beta_j),\beta_{k_n}\right\} \leq \min\left\{\sqrt{\prod_{j=1}^{n}(1-c\beta_j)},\beta_{k_n}\right\},$$

which will decay to zero when $n \to \infty$. □

B.3. Application to linear inverse problems.

Proof. (**Proof of Theorem 3.1.**) We will set $\mathcal{D}=\Lambda=[\lambda_l,\lambda_u]$. Then because χ always brings λ_k back into $\mathcal{D}$, the event $\mathcal{A}$ always happens.

Recall that

$$\partial_\lambda f(\lambda,z)=\mathrm{Tr}\left(P_2(\lambda^2 vv^\top + 2\lambda D\xi v^\top + D\xi\xi^\top D^\top)+P_1(2\lambda vv^\top + 2D\xi v^\top)\right).$$

$$\partial_\lambda^2 f(\lambda,z)=\mathrm{Tr}\left(P_3(\lambda^2 vv^\top + 2\lambda D\xi v^\top + D\xi\xi^\top D^\top)+4P_2(\lambda vv^\top + D\xi v^\top)+2P_1 vv^\top\right).$$

$$\partial_\lambda^3 f(\lambda,z)=\mathrm{Tr}\left(P_4(\lambda^2 vv^\top + 2\lambda D\xi v^\top + D\xi\xi^\top D^\top)+6P_3(\lambda vv^\top + D\xi v^\top)+6P_2 vv^\top\right).$$

$$(\text{B.2})$$

We have seen in the proof of Lemma A.6, that

$$-\partial_\lambda F(\lambda)=(\lambda/\lambda_* -1)\mathrm{Tr}(P_2 DD^\top).$$

Multiplication with $(\lambda - \lambda_*)$ gives

$$-(\lambda - \lambda_*)\partial_\lambda F(\lambda) = (\lambda - \lambda_*)^2 \mathrm{Tr}(P_2 D^T D / \lambda_*).$$

Note that if v is the eigenvector of $DD^\top$ with maximum eigenvalue λ_D

$$\mathrm{Tr}(P_2 DD^\top / \lambda_*) \geq v^\top P_2 DD^\top v / \lambda_* = \frac{\lambda_D v^\top C_0 v}{\lambda_*(\lambda_D + \lambda)^3} \geq \frac{\lambda_D^2}{\lambda_*(\lambda_D + \lambda_u)^3 \|A^\top \Gamma^{-1} A\|}.$$

So if we take c as

$$c = \frac{2\lambda_D^2}{\lambda_*(\lambda_D + \lambda_u)^3 \|A^\top \Gamma^{-1} A\|},$$

(3.9) is verified.

Next, we note that by Taylor's theorem, there are some w_k, w_k' between $\lambda_k - h_k$ and $\lambda_k + h_k$ such that

$$|\widetilde{\partial}_\lambda f(\lambda_k, Z) - \partial_\lambda f(\lambda_k, Z)| = \frac{1}{6} h_k^2 |\partial_{\lambda^3}^3 f(w_k, Z) + \partial_{\lambda^3}^3 f(w_k', Z)|.$$

Therefore

$$\mathbb{E}|\widetilde{\partial}_\lambda f(\lambda_k, Z) - \partial_\lambda f(\lambda_k, Z)|^2 = \frac{1}{18} h_k^4 (\mathbb{E}|\partial_{\lambda^3}^3 f(w_k, Z)|^2 + \mathbb{E}|\partial_{\lambda^3}^3 f(w_k', Z)|^2).$$

We will show that there is a dimension-free constant B_λ that may depend on λ such that

$$\mathbb{E}|\partial_{\lambda^3}^3 f(w_k, Z)|^2 \leq B_\lambda, \quad \text{and} \quad \mathbb{E}|\partial_{\lambda^3}^3 f(w_k', Z)|^2 \leq B_\lambda. \tag{B.3}$$

This comes from the fact that each component of $\partial_{\lambda^3}^3 f(w_k, Z)$ can be written as $\mathrm{Tr}(\Sigma C_v)$ or $\mathrm{Tr}(\Psi B)$ or $\mathrm{Tr}(\Sigma C_\xi)$, with some Σ and Ψ. Here we define

$$C_v = vv^\top, \quad B = \xi v^\top, \quad C_\xi = \xi \xi^\top.$$

Then Lemma A.4 with $n = 1$ shows that for some universal constant C

$$\mathbb{E}(\Sigma C_v) \leq 2|\mathrm{Tr}(\Sigma)|^2 + C(\|\Sigma\|_F^3 + \|\Sigma\|^3).$$

$$\mathbb{E}(\Sigma C_\xi) \leq 2|\mathrm{Tr}(\Sigma)|^2 + C(\|\Sigma\|_F^3 + \|\Sigma\|^3).$$

$$\mathbb{E}(\Psi B) \leq C(\|\Psi\|_F^3 + \|\Psi\|^3).$$

Meanwhile, the Σ matrices in $\partial_{\lambda^3}^3 f(w_k, Z)$ are of the form P_j or $D^\top P_j D$, which we know have bounded operator, trace and Frobenius norms, from the proof of Proposition 3.1, and the Ψ matrix is of the form $P_j D$, so $\|P_j D\|_F \leq \|P_j\|_F \|D\|, \|P_j D\| \leq \|P_j\| \|D\|$. So we can conclude that there is a dimension-free constant B, such that (B.3) holds.

Therefore (3.11) is verified by Jensen's inequality

$$|\mathbb{E}\widetilde{\partial}_\lambda f(\lambda_k, Z) - \partial_\lambda F(\lambda_k)|^2 = |\mathbb{E}\widetilde{\partial}_\lambda f(\lambda_k, Z) - \mathbb{E}\partial_\lambda f(\lambda_k, Z)|^2$$

$$\leq \mathbb{E}|\widetilde{\partial}_\lambda f(\lambda_k, Z) - \partial_\lambda f(\lambda_k, Z)|^2 \leq \frac{1}{9} h_k^4 B_\lambda.$$

Finally, because $\lambda \in [\lambda_l, \lambda_u]$, so B_λ can be bounded as well.

To prove that (3.10) is satisfied, note that by Young's inequality

$$\mathbb{E}|\widetilde{\partial}_\lambda f(\lambda_k, Z)|^2 \leq 2\mathbb{E}|\widetilde{\partial}_\lambda f(\lambda_k, Z) - \partial_\lambda f(\lambda_k, Z)|^2 + 2\mathbb{E}|\partial_\lambda f(\lambda_k, Z)|^2.$$

Since we have already bounded $\mathbb{E}|\widetilde{\partial}_\lambda f(\lambda_k, Z) - \partial_\lambda f(\lambda_k, Z)|^2$, it suffices to bound $\mathbb{E}|\partial_\lambda f(\lambda_k, Z)|^2$ by a dimension-independent constant A_λ. But each component of $\partial_\lambda f(w_k, Z)$ can be written as $\text{Tr}(\Sigma C_v)$ or $\text{Tr}(\Psi B)$ or $\text{Tr}(\Sigma C_\xi)$, with some Σ and Ψ. Then Lemma A.4 with $n = 1$ shows that for some universal constant C

$$\mathbb{E}(\Sigma C_v) \leq 2|\text{Tr}(\Sigma)|^2 + C(\|\Sigma\|_F^3 + \|\Sigma\|^3).$$

$$\mathbb{E}(\Sigma C_\xi) \leq 2|\text{Tr}(\Sigma)|^2 + C(\|\Sigma\|_F^3 + \|\Sigma\|^3).$$

$$\mathbb{E}(\Psi B) \leq C(\|\Psi\|_F^3 + \|\Psi\|^3).$$

Meanwhile, the Σ matrices in $\partial_{\lambda^3}^3 f(w_k, Z)$ are of the form P_j or $D^\top P_j D$, which we know have bounded operator, trace and Frobenius norms and the Ψ matrix is of the form $P_j D$, so $\|P_j D\|_F \leq \|P_j\|_F \|D\|, \|P_j D\| \leq \|P_j\| \|D\|$. So we can conclude that there is a dimension-free constant A_λ, such that $\mathbb{E}|\partial_\lambda f(\lambda_k, Z)|^2 \leq A_\lambda$. Finally, noting that $\lambda \in [\lambda_l, \lambda_u]$, we have our conclusion.   □

REFERENCES

[1] S. Agapiou, J.M. Bardsley, O. Papaspiliopoulos, and A.M. Stuart, *Analysis of the Gibbs sampler for hierarchical inverse problems*, SIAM/ASA J. Uncertain. Quantif., 2(1):511–544, 2014. 4.2

[2] S. Arridge, P. Maass, O. Öktem, and C.B. Schönlieb, *Solving inverse problems using data-driven models*, Acta Numer., 28:1–174, 2019. 1, 2.1, 2.2, 4.2

[3] Y.Y. Belov, *Inverse Problems for Partial Differential Equations*, Inverse and Ill-Posed Problems Series, De Gruyter, 32, 2002. 1.1.1

[4] K.P. Bennett, J. Hu, G. Kunapuli, and J.S. Pang, *Model selection via bilevel optimization*, IEEE International Joint Conference on Neural Networks (IJCNN), 2006. 4.2

[5] M. Benning and M. Burger, *Modern regularization methods for inverse problems*, Acta Numer., 27:1–111, 2018. 1, 1

[6] L. Borcea, *Electrical impedance tomography*, Inverse Probl., 18(6):R99, 2002. 1.1.1

[7] M. Burger, A.C.G. Mennucci, S. Osher, and M. Rumpf, *Level Set and PDE Based Reconstruction Methods in Imaging*, Springer, 2008. 1.1.2, 1.1.2

[8] P.J. Bickel and K.A. Doksum, *Mathematical Statistics: Basic Ideas and Selected Topics*, Cambridge University Press, 1971. 2.1

[9] L. Bottou, *On-line learning and stochastic approximations*, in D. Saad (eds.), On-line Learning in Neural Networks, Cambridge University Press, USA, 9–42, 1999. 2.1

[10] L. Calatroni, C. Chung, J.C. De los Reyes, C.-B. Schönlieb, and T. Valkonen, *Bilevel approaches for learning of variational imaging models*, in M. Bergounioux, G. Peyré, C. Schnörr, J.-B. Caillau and T. Haberkorn (eds.), Variational Methods, De Gruyter, 2016. 1.1.2, 4.2

[11] N.K. Chada, M.A. Iglesias, L. Roininen, and A.M. Stuart, *Parameterizations for ensemble Kalman inversion*, Inverse Probl., 34(5):055009, 2018. 4.2, 4.2

[12] N.K. Chada, *Analysis of hierarchical ensemble Kalman inversion*, arXiv preprint, arXiv:1801.00847, 2018. 4.2

[13] N.K. Chada, Y. Ying, and D. Sanz-Alonso, *Derivative and ensemble iterative Kalman methods: A unified perspective with some new variants*, Found. Data Sci., 3(3):331-369, 2021. 4.2

[14] N.K. Chada, A.M. Stuart, and X.T. Tong, *Tikhonov regularization within ensemble Kalman inversion*, SIAM J. Numer. Anal., 58(2):1263–1294, 2020. 4.2

[15] J. Chung, M. Chung, and D.P. O'Leary, *Designing optimal spectral filters for inverse problems*, SIAM J. Sci. Comput., 33(6):3132–3152, 2011. 1.1.1

[16] J. Chung, M.I. Espanol, and T. Nguyen, *Optimal regularization parameters for general-form Tikhonov regularization*, arXiv preprint, arXiv:1407.1911, 2014. 1.1.1

[17] N. Couellan and W. Wang, *Bi-level stochastic gradient for large scale support vector machine*, Neurocomputing, 153(4):300–308, 2015. 3.1

[18] N. Couellan and W. Wang, *Uncertainty-safe large scale support vector machines*, Comput. Stat. Data Anal., 109:215–230, 2017. 3.1

[19] N. Couellan and W. Wang, *On the convergence of stochastic bi-level gradient methods*, preprint, hal-01932372, 2018. 3.1

[20] B. Colson, P. Marcotte, and G. Savard, *An overview of bilevel optimization*, Ann. Oper. Res., 153:235–256, 2007. 1

[21] M. D'Elia, J.C. De los Reyes, A.M. Trujillo, *Bilevel parameter optimization for learning nonlocal image denoising models*, arXiv preprint, arxiv:1912.02347, 2020. 1.1.2

[22] J.C. De los Reyes and C.-B. Schönlieb, *Image denoising: Learning the noise model via nonsmooth PDE-constrained optimization*, Inverse Probl., 7:1183–1214, 2013. 1.1.2

[23] J.C. De los Reyes, C.-B. Schönlieb, and T. Valkonen, *Bilevel parameter learning for higher-order total variation regularisation models*, J. Math. Imaging Vis., 57:1–25, 2017. 1.1.2

[24] S. Dempe, *Foundations of Bilevel Programming*, Springer, 2002. 1

[25] M. Dunlop, M.A. Iglesias, and A.M. Stuart, *Hierarchical Bayesian level set inversion*, Stat. Comput., 27:1555–1584, 2017. 4.2

[26] M.J. Ehrhardt and L. Roberts, *Inexact derivative-free optimization for bilevel learning*, J. Math. Imaging Vis., 63:580–600, 2021. 4.2

[27] C.M. Elliott, K. Deckelnick, and V. Styles, *Numerical analysis of an inverse problem for the Eikonal equation*, Numer. Math., 119:245, 2011. 4.3

[28] H.W. Engl, K. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Math. Appl., Kluwer Academic Publishers Group, Dordrecht, 375, 1996. 1, 1, 4.2

[29] L. Franceschi, P. Frasconi, S. Salzo, R. Grazzi, and M. Pontil, *Bilevel programming for hyper-parameter optimization and meta-learning*, International Conference on Machine Learning, 1563–1572, 2018. 3.1, 4.2

[30] J.H. Friedman, R. Tibshirani, and T. Hastie, *Elements of Statistical Learning*, Springer, 2001. 1

[31] A. Garbuno–Inigo, F. Hoffmann, W. Li, and A.M. Stuart, *Interacting Langevin diffusions: Gradient structure and ensemble Kalman sampler*, SIAM J. Appl. Dyn. Syst., 19(1):412–441, 2020. 4.2

[32] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, The MIT Press, 2016. 1

[33] G. Holler, K. Kunisch, and R.C. Barnard, *A bilevel approach for parameter learning in inverse problems*, Inverse Probl., 34:115012, 2018. 1.1.1

[34] 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 J. Uncertain. Quantif., 2:464–489, 2014. 1.1.1

[35] S. Jenni and P. Favaro, *Deep bilevel learning*, Eur. Conf. Comput. Vis., 632–648, 2018. 3.1

[36] J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, Springer Verlag, NewYork, 2004. 1, 1.1.1, 2.2

[37] B. Kaltenbacher, A. Kirchner, and B. Vexler, *Goal-oriented adaptivity in the IRGNM for parameter identification in PDEs II: all-at-once formulations*, Inverse Probl., 30:045002, 2014. 1.1.1

[38] B. Kaltenbacher, *Regularization based on all-at-once formulations for inverse problems*, SIAM J. Numer. Anal., 54:2594–2618, 2016. 1.1.1

[39] K. Kunisch and T. Pock, *A bilevel optimization approach for parameter learning in variational models*, SIAM J. Imaging Sci., 6:938–983, 2013. 1.1.2

[40] G. Lord, C.E. Powell, and T. Shardlow, *An Introduction to Computational Stochastic PDEs*, Cambridge Texts in Applied Mathematics, 2014. 4.1

[41] S. Lu and S.V. Pereverzev, *Regularization Theory for Ill-posed Problems. Selected Topics*, Inverse and Ill-Posed Problems Series, Walter de Gruyter GmbH, 58, 2013. 1

[42] S. Lunz, O. Öktem, and C.-B. Schönlieb, *Adversarial regularizers in inverse problems*, Proc. Int. Conf. Neur. Inf. Process. Syst., 8516–8525, 2018. 4.2

[43] A. Maugeri and L. Scrimali, *A new approach to solve convex infinite-dimensional bilevel problems: application to the pollution emission price problem*, J. Optim. Theory Appl., 169:370–387, 2016. 4.2

[44] O. Papaspiliopoulos, G.O. Roberts, and M. Sköld, *A general framework for the parametrization of hierarchical models*, Stat. Sci., 22(1):59–73, 2007. 4.2

[45] B.T. Polyak and A.B. Juditsky, *Acceleration of stochastic approximation by averaging*, SIAM J. Control Optim., 30(4):838–855, 1992. 3.1, 3.1

[46] H. Robbins and S. Monro, *A stochastic approximation method*, Ann. Math. Stat., 22:400–407, 1951. 3.1

[47] M. Rudelson and R. Vershynin, *Hanson-Wright inequality and sub-Gaussian concentration*, Elec-

tron. Commun. Probab, 18:1–9, 2013. 2.2, A.3

[48] R. Schilling, *Measures, Integrals and Martingales*, Cambridge University Press, Second Edition, 2017. A.1

[49] J.A. Sethian, *Level Set Methods and Fast Marching Methods*, Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, 1999. 4.3

[50] A. Sinha, P. Malo, and K. Deb, *A Review on bilevel optimization: from classical to evolutionary approaches and applications*, IEEE Trans. Evol. Comput., 22(2):276–295, 2018. 1

[51] J.T. Slagel, J. Chung, M. Chung, D. Kozak, and L. Tenorio, *Sampled Tikhonov regularization for large linear inverse problems*, Inverse Probl., 33:114008, 2019. 1.1.1

[52] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press, 2014. 1, 2.1, 3.1, 3.1

[53] A.M. Stuart, *Inverse problems: A Bayesian perspective*, Acta Numer., 19:451–559, 2010. 1

[54] A. Tarantola, *Inverse Problem Theory and Methods for Model Parameter Estimation*, Elsevier, 1987. 1, 1

[55] A.N. Tikhonov, A.S. Leonov, and A.G. Yagola, *Nonlinear ill-posed inverse problems*, in A.G. Ramm (eds.), Inverse Problems, Tomography, and Image Processing, Springer, Boston, MA, 1998. 1

[56] A.F. Vidal, V. De Bortoli, M. Pereyra, and A. Durmus, *Maximum likelihood estimation of regularisation parameters in high-dimensional inverse problems: an empirical Bayesian approach. Part I: Methodology and Experiments*, arXiv preprint, arxiv:1911.11709, 2020. 1.1.2

[57] V. De Bortoli, A. Durmus, A.F. Vidal, and M. Pereyra, *Maximum likelihood estimation of regularisation parameters in high-dimensional inverse problems: an empirical Bayesian approach Part II: Theoretical Analysis*, arXiv preprint, arXiv:2008.05793, 2020. 1.1.2