Computational Optimization and Applications
https://doi.org/10.1007/s10589-025-00745-1

Derivative-free stochastic bilevel optimization for inverse problems

Mathias Staudigl[1] · **Simon Weissmann**[1] · **Tristan van Leeuwen**[2,3]

Received: 27 November 2024 / Accepted: 17 October 2025

Abstract
Inverse problems are key issues in several scientific areas, including signal processing and medical imaging. Data-driven approaches for inverse problems aim for learning model and regularization parameters from observed data samples, and investigate their generalization properties when confronted with unseen data. This approach dictates a statistical approach to inverse problems, calling for stochastic optimization methods. In order to learn model and regularisation parameters simultaneously, we develop in this paper a stochastic bilevel optimization approach in which the lower level problem represents a variational reconstruction method formulated as a convex non-smooth optimization problem, depending on the observed sample. The upper level problem represents the learning task of the regularisation parameters. Combining the lower level and the upper level problem leads to a stochastic non-smooth and non-convex optimization problem, for which standard gradient-based methods are not straightforward to implement. Instead, we develop a unified and flexible methodology, building on a derivative-free approach, which allows us to solve the bilevel optimization problem only with samples of the objective function values. We perform a complete complexity analysis of this scheme. Numerical results on signal denoising and experimental design demonstrate the computational efficiency and the generalization properties of our method.

Keywords Inverse problems · Data-driven design · Derivative-free optimization · Gaussian smoothing

1 Introduction

Bilevel optimization is a very important optimization methodology for solving inverse problems [5, 21]. The strength of bilevel optimization is that it allows to endogenously learn hyper-parameters, which otherwise would have to be tuned manually.

Extended author information available on the last page of the article

Published online: 31 October 2025

A very prominent instantiation of this is the task of learning regularization parameters [30, 32, 35]. A mathematical formulation of this problem is to first define a variational reconstruction method involving a data fidelity function $x \mapsto \mathcal{L}(\mathrm{K}(x), \xi)$, where $\xi \in \Xi$ is the observed data, and $\mathrm{K} : \mathcal{X} \to \mathcal{D}$ is the forward operator, mapping model parameters x to observations in $\mathcal{D}$. We then define the reconstruction operator $x^*(y, \cdot) : \Xi \to \mathcal{X}$ as a solution of the optimization problem

$$x^*(y,\xi) \in \operatorname*{argmin}_{x\in\mathcal{X}} \{\mathcal{L}(\mathrm{K}(x),\xi) + \mathcal{S}_y(x)\} \qquad \text{for all } (y,\xi) \in \mathcal{Y}\times\Xi. \tag{1.1}$$

The function $\mathcal{S}_y : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is a parameter-dependent regularizer, that avoids overfitting and imposes a-priori known structure into the model parameter. Choosing this parameter $y \in \mathcal{Y}$ a-priori is a severe bottleneck in the effective solution of the underlying inverse problem and poses significant practical challenges. Traditionally, this problem of *hyperparameter tuning* has been heuristically solved and generally requires a large number of solutions of this variational problem for a pre-defined grid of parameter values y. Bilevel optimization replaces this heuristic search procedure by a disciplined optimization approach which selects model parameters on par with regularization parameters, given the data sample representing the inverse problem. However, the bilevel methodology is not only useful for solving the hyperparameter learning problem. It also has a significant impact for other inverse problems in which the forward operator itself exhibits a dependence on model parameters. This is generically the case in *optimal experimental design*. In this framework we address the question of where and when to take measurements, which variables to include, and what experimental conditions should be employed. Mathematically, this leads to a forward model K_y which depends on a vector of design parameters $y \in \mathcal{Y}$, which have to be chosen before the variational model is solved. Hence, problem (1.1) needs to be modified to

$$x^*(y,\xi) \in \operatorname*{argmin}_{x\in\mathcal{X}} \{\mathcal{L}(\mathrm{K}_y(x),\xi) + \mathcal{S}_y(x)\} \qquad \text{for all } (y,\xi) \in \mathcal{Y}\times\Xi. \tag{1.2}$$

To obtain a generic set-up for learning selected components of (1.2) from data, we adopt a supervised learning approach [2]: We are given random variables $\xi = (\xi_1, \xi_2) \in \Xi_1 \times \Xi_2 = \Xi$, defined on a fixed probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where the first component contains model parameters, and the second component are the observations. This random element lives in some measurable space Ξ with joint distribution P_ξ. Our aim is to learn the model parameters $x^*(y, \xi_2)$ (as a function of regularization parameters and data), and regularization parameters $y^* \in \mathcal{Y}$ simultaneously so that they are optimal given the expected risk defined in terms of the loss function and the data. Following [2], this leads to the stochastic bilevel formulation

$$\begin{aligned} &y^* \in \operatorname*{argmin}_{y\in\mathcal{Y}} \{\mathbb{E}_\xi[F(x^*(y,\xi_2),\xi_1)] + r_1(y)\} \\ &\text{s.t.: } x^*(y,\xi_2) \in \operatorname*{argmin}_{x\in\mathcal{X}} \{g(x,y,\xi_2) + r_2(x)\}. \end{aligned} \tag{1.3}$$

The *upper level objective* $\mathbb{E}_\xi[F(x^*(y,\xi_2),\xi_1)] + r_1(y)$ contains an expectation-valued part involving a tracking-type function $F : \mathcal{X} \times \Xi_1 \to \mathbb{R}$, usually assumed to be sufficiently smooth, and a regularizer/penalty function $r_1(y)$, i.e. chosen to promote a-priori known structure in the parameter vector. The *lower level objective* $g(x,y,\xi_2) + r_2(x)$ is a variational model for obtaining model parameters, as a function of the realized data $\xi_2 \in \Xi_2$ and the tunable hyperparameter $y \in \mathcal{Y}$.

Example 1.1 (Bilevel Learning) The bilevel learning approach for inverse problem is a statistical learning methodology to select the regularization parameter based on a variational formulation. The unknown parameter and the corresponding observation are modeled as jointly distributed random variables $(\xi_1,\xi_2) : \Omega \to \Xi_1 \times \Xi_2$, defined as

$$\xi_2(\omega) \triangleq \mathrm{K}_{y_1}\xi_1(\omega) + Z(\omega), \quad \omega \in \Omega,$$

where $Z \in L^\infty(\Omega;\Xi_2)$ denotes measurement noise. In this model $\Xi_1 \triangleq \mathcal{X}$ is the data space and $\Xi_2 \triangleq \mathcal{D}$ is the set of observations. The forward operator may depend explicitly on a hyperparameter $y_1 \in \mathcal{Y}_1$. In order to build a reconstruction of the unknown parameter $\xi_1(\omega)$ for a fixed $\omega \in \Omega$, we consider the variational problem

$$\min_{x\in\mathcal{X}} \; g(x,(y_1,y_2),\xi_2(\omega)) + r_2(x), \quad g(x,(y_1,y_2),\xi_2(\omega)) \triangleq \mathcal{L}(\mathrm{K}_{y_1}(x),\xi_2(\omega)) + \mathcal{S}_{y_2}(x)\,,$$

where $\mathcal{L} : \Xi_2 \times \Xi_2 \to \mathbb{R}$ is a data fidelity function, $\mathcal{S}_{y_2} : \mathcal{X} \to \mathbb{R}$ is a regularization function with regularization parameter $y_2 \in \mathcal{Y}$ and $r_2(\cdot)$ is another regularization function reflecting a-priori knowledge about the data. The reconstruction highly depends on the choice of the hyperparameter vector $(y_1,y_2) \in \mathcal{Y} \triangleq \mathcal{Y}_1 \times \mathcal{Y}_2$, and the overall goal in bilevel learning is to choose these hyperparameters based on the stochastic bilevel optimization problem (1.3). This approach has been investigated in many previous studies (see e.g. [19, 32, 43]). We will provide more details about this application in Sect. 8. A typical upper-level objective function in this context is $F(x,\xi_1) = \frac{1}{2}\,||x-\xi_1||^2_{\mathcal{X}}$, and the bilevel problem (1.3) becomes

$$\begin{aligned}\min_{y=(y_1,y_2)} \; & \mathbb{E}_{(\xi_1,\xi_2)}\left[\frac{1}{2}\,||x^*(y,\xi_2)-\xi_1||^2_{\mathcal{X}}\right] + r_1(y)\\ \text{s.t.:}\; & x^*(y,\xi_2) \in \operatorname*{argmin}_{x\in\mathcal{X}}\{\mathcal{L}(\mathrm{K}_{y_1}(x),\xi_2(\omega)) + \mathcal{S}_{y_2}(x) + r_2(x)\}.\end{aligned}$$

◇

1.1 Challenges and related literature

Directly solving the stochastic bilevel optimization problem (1.3) is challenging for at least two reasons: First, in order to solve the upper level problem, we need to know a solution of the lower level problem. However, this is just our variational inverse

problem, and thus is typically a large-scale optimization problem itself (although very often convex). Even if this can entail computational challenges, it can in principle be overcome via state-of-the-art convex programming techniques; The second challenge that arises is how to optimize the upper level objective function, which is only available as an implicit function of the lower level solution mapping $x^*(y, \xi)$. This problem becomes even more pronounced when the lower level solution is not unique. While non-uniqueness could be dealt with penalty methods (see e.g. [36, 40]), the presence of stochastic perturbations in the problem data, renders also this approach challenging. Instead, in this paper we investigate in detail solution methods for settings in which the lower level mapping can be solved up to some accuracy at reasonable computational costs, and then use this mapping to construct a simple optimization method that avoids delicate issues such as computing gradients, or even higher-order information of the upper level objective. Specifically, we make the following standing hypothesis throughout this paper[1]:

Standing Hypothesis *For all $(y, \xi_2) \in \mathcal{Y} \times \Xi_2$ the lower level problem*

$$\min_{x \in \mathcal{X}} \{g(x, y, \xi_2) + r_2(x)\} \tag{LL}$$

admits a unique solution $x^*(y, \xi_2)$, which is a measurable function of the data ξ_2.

Working under this hypothesis, the main remaining question is how to effectively solve the upper level problem

$$\min_{y \in \mathcal{Y}} \Psi(y) \triangleq \mathbb{E}_\xi[F(x^*(y, \xi_2), \xi_1)] + r_1(y). \tag{1.4}$$

The challenge within this formulation lies in the fact that the first function $y \mapsto \mathbb{E}_\xi[F(x^*(y, \xi_2), \xi_1)]$ is expectation-valued (hence hard to evaluate) and in general non-smooth and non-convex. The lack of regularity properties makes a direct gradient-based approach less qualified, without even talking about the difficulties in computing a gradient (aka the *hypergradient* [20, 29]) of this composite function. The key complications arising in this formulation are (i) the dependence of the lower level solution $x^*(y, \xi_2)$ on the random variable ξ_2, (ii) the potential non-smoothness of the lower level variational problem, (iii) the non-smoothness of the upper level problem. All three complications make any attempt to adapt standard methods for solving bilevel optimization problems complicated. One main technical contribution of this paper is to construct a practically efficient strategy for solving the stochastic bilevel problem (1.3) building on a *zeroth-order stochastic oracle model* for estimating the hypergradient, allowing for bias in the random estimator, and inexactness of the solution of the lower level problem. Although this setting received a significant amount of attention recently, mainly driven from applications in machine learning such as meta-learning [45], hyper-parameter optimization [22, 47] and reinforcement learning [33], the composite setting embodied in (1.4) is complicating the hypergra-

[1] A more precise formulation of this hypothesis will be given in Sect. 3.

dient estimation task a lot. The survey [39] gives a comprehensive state-of-the-art overview.

1.1.1 Stochastic bilevel optimization

The bilevel instance (1.3) differs from the typical machine learning setting in our requirement that the lower level problem needs to be solved for any realization of the random variable ξ_2. In machine learning, the typically encountered formulation has no non-smooth terms and no explicit constraints:

$$\min_{y\in\mathbb{R}^d} \psi(y)\triangleq f(x^*(y),y) \quad \text{s.t.: } x^*(y)\in \operatorname*{argmin}_{x\in\mathbb{R}^n} g(x,y)\,,$$

where $f(x,y)\triangleq\mathbb{E}[F(x,y,\xi_1)]$ and $g(x,y)\triangleq\mathbb{E}[G(x,y,\xi_2)]$. Under strong regularity conditions the hyperobjective ψ is smooth enough so that its gradient can be characterized by the implicit function theorem

$$\nabla\psi(y) = \nabla_y f(x^*(y),y) - \nabla^2_{xy} g(x^*(y),y)\left[\nabla^2_{xx} g(x^*(y),y)\right]^{-1}\nabla_x f(x^*(y),y).$$

In the composite non-smooth setting arising in inverse problems, and which is of interest in this paper, there is no hope that a similar formula for the hypergradient can be defined. For numerical approximation methods departing from this approach, see [26, 33].

Recently, [12] propose a stochastic zeroth-order method for a class of stochastic mathematical programs under equilibrium constraints, in which the lower-level problem is described by the solution set of a stochastic variational inequality, and the upper-level problem is a stochastic unconstrained optimization problem. We extend this setting to the non-smooth proximal framework in both the upper and the lower-level problem. This is a non-trivial extension, since it requires a fundamentally different analysis of the iteration complexity of the method in terms of the prox-gradient mapping (cf. (4.12)). Moreover, we provide complexity estimates on the criticality measure represented by the prox-gradient mapping via an integrated smoothing and zeroth-order optimization scheme, without any a-priori convexity assumptions on the hyperobjective.

1.1.2 Zeroth-order stochastic optimization

The numerical solution of stochastic optimization problems requires the availability of a stochastic oracle. In low informational settings such as simulation-based, or black-box optimization, an attractive stochastic oracle is one that relies only on noisy function queries. Such zeroth-order methods have been studied in the literature under the name of derivative-free optimization [10, 48], Bayesian optimization [23], and optimization with bandit feedback [8, 18]. Moreover, gradient-free methods received a lot of attention within mathematical imaging [19, 20], and scientific computing [34, 44], as well as in machine learning and computational statistics [1, 17, 24]. We discuss the connection to the most important references in the following.

[6] performs a detailed comparison of different derivative-free methods based on noisy function evaluations, assuming that the noise component is additive and with zero mean and bounded range. They established conditions on the gradient estimation errors that guarantee convergence to a neighborhood of the solution. We perform a complexity analysis of a derivative free method in which the function values are noisy evaluations of the hyperobjective of the bilevel problem (1.3), without a uniformly bounded noise assumption. Instead, we only assume standard variance bounds in L^p, for some $p \geq 2$.

[3] provide an in-depth analysis of zero-order estimators for solving general stochastic optimization problems, using a Frank-Wolfe method, a stochastic proximal gradient method, or a higher-order method building on the cubic regularization globalization technique. Their general complexity statements are not immediately transferable to our problem, since we solve a stochastic bilevel problem, with potentially inexact feedback between the upper and the lower level problem. This noisy and inexact feedback mechanism leads to an additional bias in the gradient estimator, which needs to be carefully balanced in order to prove convergence guarantees of the method.

1.2 Main contributions and outline

Our main results can be summarized as follows:

1. Under weak regularity assumption on the hyperobjective $h(y) = \mathbb{E}[F(x^*(y, \xi_2), \xi_1)]$ (essentially only Lipschitz continuity), we derive an iteration complexity statement in terms of the proximal gradient mapping for the Gaussian smoothed objective h_η. In particular, we give complexity statements assuming that the lower level problem can be solved exactly, or inexactly, with a controlled precision in an L^p sense.
2. We particularize this result in the convex case to obtain a complexity statement in terms of the original objective function optimality gap.
3. To relate the complexity statement derived for the smoothed hyperobjective, we define a notion for a relaxed stationary point, using a fuzzy version of the Goldstein subgradient, originally introduced in [28] for Lipschitz continuous mathematical programs. This allows us to transfer the complexity statements derived in pervious sections for the smoothed prox-gradient mapping to a criticality measure involving the Goldstein subgradient.

The remainder of the manuscript is structured as follows. We introduce our notation and some known results, used in the analysis, in Sect. 2. Section 3 presents the formulation of the stochastic bilevel optimization problem with the corresponding assumptions. In Sect. 4, we introduce our proposed zeroth-order optimization method. Section 5 begins the convergence analysis in a non-convex setting with a fixed smoothing parameter, covering both exact and inexact lower level solutions. We then proceed to Sect. 6, where we analyze the convex case and quantify the smoothing error. Section 7 addresses the explicit complexity and relaxed stationarity for non-convex problems. In Sect. 8, we apply our algorithm to linear inverse prob-

lems, with a particular focus on imaging. Finally, we conclude the main body of the manuscript with a summary in Sect. 9. For clarity, most of the proofs are deferred to Appendices A–C.

2 Notation and preliminaries

For a finite dimensional real vector space $\mathcal{E}$, we denote by $\mathcal{E}^*$ its dual space. The value of a linear function $s \in \mathcal{E}^*$ at point $x \in \mathcal{E}$ is denoted by $s(x) \triangleq \langle s, x \rangle$. We endow the spaces $\mathcal{E}$ and $\mathcal{E}^*$ with Euclidean norms $||x|| = \langle Bx, x \rangle^{1/2}$ and $||s||_* = \langle s, B^{-1}s \rangle^{1/2}$, where $B = B^*$ represents the Riesz isomorphism, i.e. a positive definite linear operator from $\mathcal{E}$ to $\mathcal{E}^*$. For a subset $C \subset \mathcal{E}$ we define the distance of $x \in \mathcal{E}$ to C by $\operatorname{dist}(x, C) \triangleq \inf_{z \in C} ||x - z||$. The closed ball with center x and radius $r > 0$ is denoted as $\mathbb{B}(x, r)$. The convex hull of a set X is denoted as $\operatorname{Conv}(X)$. If Ω is a topological space, we denote by $\mathcal{B}(\Omega)$ the Borel σ-algebra. In this paper, we consider functions with different levels of smoothness. We say a function $h : \mathcal{E} \to \mathbb{R}$ belongs to class $\mathrm{C}^{0,0}(\mathcal{E})$ if there exists a constant $\operatorname{lip}_0(h) > 0$ such that

$$|h(x_1) - h(x_2)| \leq \operatorname{lip}_0(h) \, ||x_1 - x_2|| \,, \qquad \forall x_1, x_2 \in \mathcal{E} \,,$$

h belongs to class $\mathrm{C}^{1,1}(\mathcal{E})$ if there exists a constant $\operatorname{lip}_1(h) > 0$

$$||\nabla h(x_1) - \nabla h(x_2)||_* \leq \operatorname{lip}_1(h) \, ||x_1 - x_2|| \,, \qquad \forall x_1, x_2 \in \mathcal{E}.$$

For $h \in \mathrm{C}^{1,1}(\mathcal{E})$, we have the Lipschitz descent Lemma [41, Lemma 1.2.3]

$$h(x_2) \leq h(x_1) + \langle \nabla h(x_1), x_2 - x_1 \rangle + \frac{\operatorname{lip}_1(h)}{2} \, ||x_2 - x_1||^2 \,, \quad \forall x_1, x_2 \in \mathcal{E}. \tag{2.1}$$

For extended real-valued convex functions $h : \mathcal{E} \to [-\infty, \infty]$, we define its (effective) domain $\operatorname{dom}(h) = \{y \in \mathcal{Y} | h(y) < \infty\}$. The convex subdifferential is the set-valued mapping $\partial h(y) \triangleq \{v \in \mathcal{E}^* | h(\tilde{y}) \geq h(y) + \langle v, \tilde{y} - y \rangle \quad \forall \tilde{y} \in \mathcal{E}\}$. Elements of the set $\partial h(y)$ are called subgradients, and the inequality defining the set is called the subgradient inequality. A convex function is called *proper* if it never attains the value $-\infty$.

Definition 2.1 Let $\delta \geq 0$. For a convex function $h : \mathcal{E} \to (-\infty, +\infty]$, the δ-subdifferential $\partial_\delta h(y)$ the set of vectors $v \in \mathcal{E}^*$ satisfying

$$h(\tilde{y}) \geq h(y) - \delta + \langle v, \tilde{y} - y \rangle \qquad \forall \tilde{y} \in \mathcal{E}.$$

Note that the above definition reduces naturally to the convex subdifferential by setting $\delta = 0$.

Definition 2.2 The proximal operator of a closed convex and proper function $g : \mathcal{E} \to (-\infty, \infty]$ is defined by

$$\text{prox}_g(x) \triangleq \underset{u \in \mathcal{E}}{\text{argmin}} \{g(u) + \frac{1}{2} ||u - x||^2\}. \tag{2.2}$$

The prox-operator is always 1-Lipschitz (non-expansive) [4]. We also make use of the Pythagorean identity on the Euclidean space $\mathcal{E}$ with inner product $\langle B\cdot, \cdot\rangle$:

$$2\langle y - u, B(x - y)\rangle = ||x - u||^2 - ||x - y||^2 - ||y - u||^2 . \tag{2.3}$$

For $p \in [1, \infty]$, let $L^p(\Omega, \mathcal{F}, \mathbb{P}; \mathcal{E})$ be the set of all random variables for which the integral $\mathbb{E}_{\mathbb{P}}[|f|^p] \triangleq \int_\Omega |f(\omega)|^p \, \mathrm{d}\mathbb{P}(\omega)$ exists and is finite. This is a Banach space with norm $|f|_p \triangleq (\mathbb{E}_{\mathbb{P}}[|f|^p])^{1/p}$.

3 Problem formulation

We denote by $(\mathcal{X}, ||\cdot||_{\mathcal{X}})$ and $(\mathcal{Y}, ||\cdot||_{\mathcal{Y}})$ finite dimensional Euclidean vector spaces, with dual spaces $(\mathcal{X}^*, ||\cdot||_{\mathcal{X}^*}), (\mathcal{Y}^*, ||\cdot||_{\mathcal{Y}^*})$. Let $(\Omega_0, \mathcal{A}, \mathbb{P}_0)$ be a complete probability space, carrying random elements $\xi_1 \in L^0(\Omega_0, \mathcal{A}_0, \mathbb{P}_0; \Xi_1)$ and $\xi_2 \in L^0(\Omega_0, \mathcal{A}_0, \mathbb{P}_0; \Xi_2)$ taking values in a measurable space $(\Xi_i, \mathcal{B}(\Xi_i)), i = 1, 2$. We define $\xi(\omega) \triangleq (\xi_1(\omega), \xi_2(\omega))$, and denote the distribution of this random element as $\mathrm{P}_\xi \triangleq \mathbb{P}_0 \circ \xi^{-1}$. Accordingly, the marginal distributions are defined as $\mathrm{P}_{\xi_1}(A) \triangleq \mathrm{P}_\xi(A \times \Xi_2)$ and $\mathrm{P}_{\xi_2}(B) \triangleq \mathrm{P}_\xi(\Xi_1 \times B)$ for $A \in \mathcal{B}(\Xi_1)$ and $B \in \mathcal{B}(\Xi_2)$, respectively.

Remark 3.1 Throughout this paper we abuse notation in that we do not notationally distinguish a random variable ξ from its realization (hitherto also denoted by ξ). We belief this common abuse of notation is simplifying the notation and its meaning should be clear from the context.

3.1 The hyperobjective program

In problem (1.3), the variable $y \in \mathcal{Y}$ (i.e. the learning parameters) is chosen before the event ω is realized, whereas x is a decision variable (i.e. the model parameters) that is implemented just-in-time, given $y \in \mathcal{Y}$ and the realization $\xi_2(\omega) \in \Xi$. A solution of the lower-level optimization problems constitutes therefore of a feedback mapping $x^*(\cdot, \xi_2) \in L^\infty(\mathcal{Y}; \mathcal{X})$, satisfying a measurability property with respect to the noise variable:

$$\omega \mapsto x^*(y, \xi_2(\omega)) \in L^0(\Omega, \mathcal{A}_0, \mathbb{P}_0; L^\infty(\mathcal{Y}; \mathcal{X})).$$

In particular, by the Doob-Dynkin Lemma, the mapping $\omega \mapsto x^*(y, \xi_2(\omega))$ is $\sigma(\xi_2)$ -measurable, for all $y \in \mathcal{Y}$. The following standing assumption shall apply throughout the paper.

Assumption 1 $r_1 : \mathcal{Y} \to (-\infty, \infty]$ is a closed convex and proper function.

Assumption 2 $F : \mathcal{X} \times \Xi_1 \to \mathbb{R}$ is a Carathéodory function:

(a) $\omega \mapsto F(x, \xi_1(\omega))$ is $\sigma(\xi_1)$-measurable for every $x \in \mathcal{X}$;
(b) $x \mapsto F(x, \xi_1)$ is continuous for almost every $\xi_1 \in \Xi_1$.

Assumption 3 For all $x \in \mathcal{X}$, the value $\mathbb{E}_{\mathbb{P}_0}[F(x, \xi_1)]$ is finite. There exists a positive valued random variable $\mathrm{lip}_0(F(\cdot, \xi_1)) : \Omega \to (0, \infty)$ such that $|\mathrm{lip}_0(F(\cdot, \xi_1))|_1 < \infty$, and for all $x_1, x_2 \in \mathcal{X}$ it holds that

$$|F(x_1, \xi_1) - F(x_2, \xi_1)| \leq \mathrm{lip}_0(F(\cdot, \xi_1)) \, ||x_1 - x_2||_{\mathcal{X}} \, . \tag{3.1}$$

Assumption 3 implies that $x \mapsto f(x) \triangleq \mathbb{E}_{\mathbb{P}_0}[F(x, \xi_1)]$ is Lipschitz continuous [46, Thm.7.44], with Lipschitz constant $\mathrm{lip}_0(f) \triangleq |\mathrm{lip}_0(F(\cdot, \xi_1))|_1$. In particular, the function $x \mapsto f(x)$ is measurable.

Assumption 4 $r_2 : \mathcal{X} \to (-\infty, \infty]$ is proper, closed and convex. For all $y \in \mathrm{dom}\,(r_1)$, the function $x \mapsto g(x, y, \xi_2)$ is continuously differentiable and convex.

Assumption 5 For all $(y, \xi_2) \in \mathrm{int}\,\mathrm{dom}\,(r_1) \times \Xi_2$ the parameterized variational inequality

$$\text{Find } x \in \mathcal{X} \text{ such that } 0 \in \nabla_x g(x, y, \xi_2) + \partial r_2(x) \tag{3.2}$$

has a unique solution $x^*(y, \xi_2)$, enjoying the following properties:

(S.1) $\omega \mapsto x^*(y, \xi_2(\omega))$ is measurable, uniformly in $y \in \mathrm{int}\,\mathrm{dom}\,(r_1)$;
(S.2) $y \mapsto x^*(y, \xi_2)$ is Lipschitz continuous on $\mathrm{int}\,\mathrm{dom}\,(r_1)$, for almost all $\xi_2 \in \Xi_2$.

Our set of assumptions correspond to typical hypothesis that have been used in oracle-based approaches to bilevel problems. Assumption 4 is a structural assumption on the data, which reflects the typical structure of variational formulations of inverse problems. Assumption 5 are technical assumptions which are needed to carry out our derivative-free approach. Measurability (S.1) of the reconstruction is arguably a minimal assumption. Lipschitz continuity is a more restrictive assumption, which is essentially an a-priori hypothesis on the solution regularity of the lower level problem. An important special case where Lipschitz continuity of the reconstruction operator is obtained when $\nabla_x g(x, y, \xi)$ is uniformly Lipschitz and r_2 is an indicator function of a closed convex set. In this case, Theorem 2B.1 and Corollary 2B.3 of [15] establishes the Lipschitz continuity of the reconstruction operator. Another set

of transparent conditions is described in Theorem 1 of [19], which considers smooth lower level problems. We need this assumption to obtain the Lipschitz property of the implicit function $y \mapsto H(y, \xi)$.

Combining Assumptions 3 and 5, we can define the stochastic hyperobjective

$$H : \mathcal{Y} \times \Xi \to \mathbb{R}, \quad (y, \xi) \mapsto H(y, \xi) \triangleq F(x^*(y, \xi_2), \xi_1). \tag{3.3}$$

Note that $H(\cdot, \xi) \in \mathrm{C}^{0,0}(\mathcal{Y})$. In order to bound the variance of our gradient estimator, we need an a-priori assumption on the integrability of the random Lipschitz modulus.

Assumption 6 We assume that $|\mathrm{lip}_0(H(\cdot, \xi))|_2 < \infty$.

Thanks to the inherited measurability, we can leverage Fubini's theorem to obtain $h(y) \triangleq \mathbb{E}_{\mathbb{P}}[H(y, \xi)] = \int_{\Xi_2} f(x^*(y, w_2))\, \mathrm{dP}_{\xi_2}(w_2)$. The fact that $f \in \mathrm{C}^{0,0}(\mathcal{Y})$ combined with (S.2) allows us to conclude $h \in \mathrm{C}^{0,0}(\mathcal{Y})$.

Absorbing the lower level solution into the upper level, we arrive at the reduced formulation of the upper level optimization problem

$$\Psi^{\,\mathrm{Opt}} \triangleq \inf_{y \in \mathcal{Y}} \{\Psi(y) \triangleq h(y) + r_1(y)\}, \tag{3.4}$$

which is commonly known in bilevel optimization as the *hyperobjective optimization problem*.

3.2 Approximate stationarity conditions

The hyperobjective program (3.4) is a non-convex and non-smooth optimization problem, involving a Lipschitz continuous function $y \mapsto h(y)$, and a convex composite term $y \mapsto r_1(y)$. As is typical in non-convex optimization, our aim is to localize a specific class of approximate stationary points, as we are about to define in this section. For a locally Lipschitz function $h : \mathcal{Y} \to \mathbb{R}$, the generalized directional derivative in the sense of Clarke [9] of h at $y \in \mathcal{Y}$ in direction $u \in \mathcal{Y}$ is defined as

$$h^\circ(y; u) \triangleq \limsup_{y' \to y, t \to 0^+} \frac{h(y' + tu) - h(y')}{t}.$$

The Clarke subdifferential of h at y is the set

$$\partial_{\mathrm{C}} h(y) \triangleq \{s \in \mathcal{Y}^* | h^\circ(y, u) \geq \langle s, u \rangle \quad \forall u \in \mathcal{Y}\}.$$

The primary goal of non-smooth non-convex optimization is the search for stationary points. A point $y \in \mathcal{Y}$ is called (Clarke)-*stationary* for $\Psi = h + r$ if the inclusion

$$0 \in \partial_C h(y) + \partial r_1(y)$$

is satisfied.

Definition 3.1 Given $\varepsilon > 0$, a point $y^* \in \mathcal{Y}$ is called an ε-stationary point of (3.4) if

$$\operatorname{dist}\left(0, \partial_C \Psi(y^*)\right) \leq \varepsilon. \tag{3.5}$$

Recently, a series of papers challenged the question whether optimization algorithms are able to identify ε-stationary points in finite time. [50] provided a definite negative answer to this question, by demonstrating that no first-order method is able to identify ε-stationary points in finite time. Therefore, we will content ourselves with a more modest stationarity notion.

Definition 3.2 ([28]) For any $\delta > 0$, the Goldstein δ-subdifferential of h at $y \in \mathcal{Y}$ is the set

$$\partial_{\mathrm{G}}^{\delta} h(y) \triangleq \operatorname{Conv}\left(\bigcup_{\tilde{y} \in \mathbb{B}(y,\delta)} \partial_{\mathrm{C}} h(\tilde{y})\right). \tag{3.6}$$

We employ the Goldstein subdifferential for relating the stationarity measures of a smoothed auxiliary model, with stationarity with respect to the original problem. As such, our proposal of an approximate stationary point combines the definitions of [13, 14] for stochastic subgradient methods, and [38] for zeroth-order methods.

Definition 3.3 For any $(\varepsilon, \delta) > 0$, we call a random variable $y^* \in L^0(\Omega, \mathcal{F}, \mathbb{P}; \mathcal{Y})$ an (ε, δ)-stationary point of (1.4) if

$$\mathbb{E}\left[\operatorname{dist}\left(y^*, \{y \mid \operatorname{dist}\left(0, \partial_G^{\delta} h(y) + \partial r_1(y)\right)^2 \leq \varepsilon\}\right)^2\right] \leq \varepsilon. \tag{3.7}$$

4 Derivative free randomized proximal gradient method

4.1 Gaussian smoothing of the implicit function

To simplify the notation, we write $||u||_{\mathcal{Y}} \triangleq ||u|| \triangleq \sqrt{\langle Bu, u\rangle}$, given the Riesz mapping $B = B^* \succ 0$ from $\mathcal{Y}$ to $\mathcal{Y}^*$. We denote the dimension of the Euclidean space $\mathcal{Y}$ by n. The n-dimensional Lebesgue measure on $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ is denoted by $\mathrm{Leb}_{\mathcal{Y}}$, and we typically write $\mathrm{d}y$, instead of $\mathrm{dLeb}_{\mathcal{Y}}(y)$. We define the Gaussian Lebesgue density on $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}), \mathrm{Leb}_{\mathcal{Y}})$ as $\pi_\eta(z|y) \triangleq \frac{\sqrt{\det(B)}}{(2\pi)^{n/2}\eta^n} \exp\left(-\frac{1}{2\eta^2}||z-y||^2\right)$.

Given a function $h : \mathcal{Y} \to \mathbb{R}$ and a positive parameter $\eta > 0$, for any $\eta > 0$ we define the Gaussian smoothing of h as the convolution

$$h_\eta(y) \triangleq (h \circledast \pi_\eta)(y) = \int_{\mathcal{Y}} h(z)\pi_\eta(z|y)\mathrm{d}z. \tag{4.1}$$

Let us introduce an independent probability space $(\Omega_1, \mathcal{A}_1, \mathbb{P}_1)$. We say $U : (\Omega_1, \mathcal{A}_1) \to (\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ is a standard Gaussian random variable on $\mathcal{Y}$, denoted as $U \sim \mathrm{N}(0, \mathrm{Id}_{\mathcal{Y}})$, if $\mathbb{P}_1 \circ U^{-1}$ admits the density $\pi_1(\cdot|0) \equiv \pi$ on $\mathcal{Y}$ with respect to $\mathrm{Leb}_{\mathcal{Y}}$. Via the change of variables $z = y + \eta u$, we can rewrite the above integral as

$$h_\eta(y) = \int_{\mathcal{Y}} h(y + \eta u)\pi(u)\mathrm{d}u = \mathbb{E}_{\mathbb{P}_1}[h(y + \eta U)].$$

For $\eta > 0$, the function $y \mapsto h_\eta(y)$ is differentiable and $\eta > 0$ plays the role of a smoothing parameter. Using the expression above, we immediately deduce the formula for the gradient (see Appendix A, eq. (A.3)) as

$$\nabla h_\eta(y) = \mathbb{E}_{\mathbb{P}_1}\left[\frac{h(y + \eta U)}{\eta} BU\right] = \mathbb{E}_{\mathbb{P}_1}\left[\frac{h(y + \eta U) - h(y)}{\eta} BU\right]. \tag{4.2}$$

Specifically, we leverage upon the work [42], and use the following estimates.[2]

Lemma 4.1 *Let $h \in \mathrm{C}^{0,0}(\mathcal{Y})$. Then $h_\eta \in \mathrm{C}^{0,0}(\mathcal{Y})$ and $\mathrm{lip}_0(h_\eta) \leq \mathrm{lip}_0(h)$ for all $\eta > 0$.*

Lemma 4.2 *([42], Theorem 1) Let $h \in \mathrm{C}^{0,0}(\mathcal{Y})$ and $\eta > 0$. Then for all $y \in \mathcal{Y}$ it holds*

$$|h_\eta(y) - h(y)| \leq \eta \mathrm{lip}_0(h)\sqrt{n}\,.$$

Lemma 4.3 *Let $h \in \mathrm{C}^{0,0}(\mathcal{Y})$ and $\eta > 0$. Then $h_\eta \in \mathrm{C}^{1,1}(\mathcal{Y})$ with $\mathrm{lip}_1(h_\eta) = \frac{\sqrt{n}}{\eta}\mathrm{lip}_0(h)$. Moreover, for all $y \in \mathcal{Y}$, there holds*

$$||\nabla h_\eta(y)||_*^2 \leq \mathrm{lip}_0(h)^2(4 + n)^2. \tag{4.3}$$

In the convex case, we report a classical relation between the gradients of the Gaussian smoothed function and the δ-subdifferential.

Lemma 4.4 *([42], Theorem 2) If $h \in \mathrm{C}^{0,0}(\mathcal{Y})$ and convex, then, for all $y \in \mathcal{Y}$ and $\eta > 0$, we have*

$$\nabla h_\eta(y) \in \partial_\delta h(y), \quad for \delta = \eta \mathrm{lip}_0(h)\sqrt{n} \tag{4.4}$$

where $\partial_\delta h$ is the δ-subdifferential (cf. Definition 2.1).

[2] For being self-contained, we provide proofs of these facts in Appendix A.

The next proposition establishes a quantitative connection between the gradients of the smoothed function h_η and the Goldstein δ-subgradient. This is the key tool to relate complexity estimates of the smoothed objective with the original, unsmoothed, objective.

Proposition 4.5 *([38], Theorem 3.6) Let $h \in \mathrm{C}^{0,0}(\mathcal{Y})$ and $\mathcal{D} \subset \mathcal{Y}$ a convex compact set. Then, for all $\delta > 0$ and for all $\varepsilon > 0$, it holds that*

$$\nabla h_\eta(y) \in \partial_{\mathrm{G}}^{\delta} h(y) + \varepsilon \mathbb{B}_{\mathcal{Y}} \qquad \forall \eta \in (0, \bar{\eta}], \forall y \in \mathcal{D}.$$

where $\mathbb{B}_{\mathcal{Y}}$ denotes the unit ball in $\mathcal{Y}$, $\bar{\eta} \triangleq \min\{1, \delta/\Gamma\}$, $\Gamma \triangleq \left[-n\mathcal{W}_{-1}\left(\frac{-\nu^{2/n}}{2\pi e}\right)\right]^{1/2}$ and $\nu \triangleq \min\{\frac{\varepsilon}{4\,\mathrm{lip}_0(h)}, (2\pi)^{n/2} - \frac{1}{2}\}$. $\mathcal{W}_{-1}$ is the negative branch of the Lambert W-function, i.e. of the inverse of $x \mapsto xe^x$, $x \in \mathbb{R}$.

Remark 4.1 Since $\nu \leq (2\pi)^{\frac{n}{2}} - \frac{1}{2}$, we have $\frac{\nu^{2/n}}{2\pi e} < \frac{1}{e}$, and hence $\mathcal{W}_{-1}\left(\frac{\nu^{2/n}}{2\pi e}\right) < -1$. Thus, $\Gamma \in (\sqrt{n}, \infty)$. $\diamond$

4.2 Zeroth-order gradient estimator of the implicit function

The first step in our construction is the design of a zeroth-order gradient estimator. This requires a solution of the lower-level problem. We discuss two different settings. First, we consider the case in which the solution of the lower level problem is available exactly. This is a very common assumption in stochastic bilevel optimization; see e.g. [11, 12, 39], as well as [19] for inverse problems. We then relax this assumption by allowing for controllable errors in the lower level solution. This scenario is more realistic, but also more challenging since the inexact model introduces an additional bias in the stochastic gradient estimator. We account for this additional difficulty by presenting two different complexity estimates, one for the exact and one for the inexact case, respectively.

4.3 Exact lower level solution

Consider the implicit function $h : \mathcal{Y} \to \mathbb{R}$ given by $h(y) = \mathbb{E}_{\mathbb{P}}[H(y, \cdot)]$, where $H(y, \xi) = F(x^*(y, \xi_2), \xi_1)$ is the hyperobjective, defined in (3.3). We have $h \in \mathrm{C}^{0,0}(\mathcal{Y})$, so that its Gaussian smoothing with parameter $\eta > 0$ satisfies $h_\eta \in \mathrm{C}^{1,1}(\mathcal{Y})$. Let $u \in \mathcal{Y}$ represent a direction and $\delta > 0$ a parameter. We define the finite-difference estimator

$$\hat{\nabla}_{(u,\eta)} H(y, \xi) \triangleq \frac{H(y + \eta u, \xi) - H(y, \xi)}{\eta} Bu = \frac{F(x^*(y + \eta u, \xi_2), \xi_1) - F(x^*(y, \xi_2), \xi_1)}{\eta} Bu.$$

If $u^{(m)} = \{u^1, \dots, u^m\}$ is an m-tuple of directions in $\mathcal{Y}$ and $\xi^{(m)} = \{\xi^1, \dots, \xi^m\}$ are m-i.i.d copies of the random variable ξ, then we define the random gradient estimator, based on finite differences of the subsampled hyperobjective:

$$V_\eta(y, u^{(m)}, \xi^{(m)}) \triangleq \frac{1}{m} \sum_{i=1}^{m} \hat{\nabla}_{(u^i,\eta)} H(y, \xi^i). \tag{4.5}$$

To realize this estimator on a sufficiently large common probability space, we build the typical product space enlargement $(\Omega, \mathcal{A}, \mathbb{P}) = (\Omega_0 \times \Omega_1, \mathcal{A}_0 \otimes \mathcal{A}_1, \mathbb{P}_0 \times \mathbb{P}_1)$. On this extended setup, we abuse notation and identify the random element ξ and U as measurable functions on $(\Omega, \mathcal{A})$ by means of the following notational convention:

$$\xi(\omega) = \xi(\omega_0) \text{ and } U(\omega) = U(\omega_1) \qquad \forall \omega \in \Omega.$$

Let $U^{(m)} \triangleq (U^1, \dots, U^m)$ be an iid random sample of Gaussian $\mathcal{Y}$-valued random vectors and $\xi^{(m)} \triangleq (\xi^1, \dots, \xi^m)$ an iid sample of ξ, assumed to be independent of each other. Define the random estimator

$$\hat{V}_{\eta,m}(y, \omega) \triangleq V_\eta(y, U^{(m)}(\omega), \xi^{(m)}(\omega)) \qquad \forall \omega \in \Omega. \tag{4.6}$$

Given a positive smoothing parameter $\eta > 0$, we are iteratively solving the stochastic composite optimization problem

$$\begin{aligned} &\min_{y \in \mathcal{Y}} \Psi_\eta(y) \quad \text{with} \quad \Psi_\eta(y) \triangleq h_\eta(y) + r_1(y) \quad \text{and} \\ &h_\eta(y) = \mathbb{E}_{\mathbb{P}}[F(x^*(y + \eta U, \xi_2), \xi_1)]. \end{aligned} \tag{4.7}$$

In the following, we assume that $\Psi_\eta^{\mathrm{Opt}} \triangleq \inf_{y \in \mathcal{Y}} \Psi_\eta(y) > -\infty$. The smooth part of this composite minimization problem is the Gaussian smoothing of the hyperobjective h, and r_1 is a closed convex and proper regularizing term.

4.4 Inexact lower level solution

We now define a relaxation of the stochastic oracle, allowing for computational errors in the lower level solution.

Definition 4.6 (Inexact lower level solution) Given $p \geq 2$ and $\beta \geq 0$, we call a mapping $x^\beta \in L^\infty(\mathcal{Y} \times \Xi; \mathcal{X})$ a β-optimal solution of the lower level problem (LL) if

$$\mathbb{E}\left[\left|\left| x^\beta(y, \xi) - x^*(y, \xi) \right|\right|_{\mathcal{X}}^p \right]^{1/p} \leq \beta. \tag{4.8}$$

Remark 4.2 We note that an inexact solution can readily be obtained by embedding our main iteration in a double-loop algorithmic strategy in which the inner loop is some fast solver that returns an approximate solution of the lower level problem, for

fixed parameters (y, ξ). The exact formulation of such an inner loop solver should be adapted to the nature of the lower level optimization problem. We treat the lower-level problem essentially as an oracle, which can be queried at any position (y, ξ_2) and returning us some feedback $x^\beta(y, \xi)$. Hence, we do not need to specify a specific numerical scheme employed to realize this oracle. However, our approach can easily be embedded in a double loop architecture in which an inner loop constructs an approximate lower level solution $x^\beta(y, \xi_2)$, and the outer level is our scheme (Algorithm 1). Such double-loop structures are a very popular solution strategy in stochastic bilevel optimization (cf. [12, 36] and references therein).

Remark 4.3 Inexactness of lower level solutions in bilevel optimization has been investigated in [19, 20] in deterministic regimes. Our notion takes into consideration the potential noisy nature of the data.

Given the inexact lower level solution mapping, we accordingly define the inexact hyperobjective as

$$H^\beta(y, \xi) \triangleq F(x^\beta(y, \xi_2), \xi_1) \quad \text{for all } y \in \mathcal{Y},\ (\xi_1, \xi_2) \in \Xi_1 \times \Xi_2.$$

The resulting biased random gradient estimator is given by

$$\begin{aligned} \hat{\nabla}_{(u,\eta)} H^\beta(y, \xi) \triangleq & \frac{H^\beta(y + \eta u, \xi) - H^\beta(y, \xi)}{\eta} \\ Bu = & \frac{F(x^\beta(y + \eta u, \xi_2), \xi_1) - F(x^\beta(y, \xi_2), \xi_1)}{\eta} Bu, \end{aligned}$$

and replace the multi-point random gradient estimator by

$$V_\eta^\beta(y, u^{(m)}, \xi^{(m)}) \triangleq \frac{1}{m} \sum_{i=1}^{m} \hat{\nabla}_{(u^i,\eta)} H^\beta(y, \xi^i). \tag{4.9}$$

As in the exact case, in order to reduce notational clutter, we will adopt the simplified notation $\hat{V}_{\eta,m}^\beta(y, \omega) \triangleq V_\eta^\beta(y, U^{(m)}(\omega), \xi^{(m)}(\omega))$ for the multi-point random gradient estimator based on the zeroth-order oracle.

4.5 The algorithmic scheme

Since $h_\eta \in \mathrm{C}^{1,1}(\mathcal{Y})$, we are in the classical proximal-gradient framework, which is defined in terms of the fixed point iteration

$$\bar{y}^+ = T_{\eta,t}(y) \triangleq \operatorname{prox}_{tr_1}(y - tB^{-1}\nabla h_\eta(y)),$$

where $t \in [0, \infty)$ is a step size parameter, $\eta > 0$ and $y \in \mathcal{Y}$.

Since we have no direct access to the gradient $\nabla h_\eta(y)$, we define a stochastic approximation using the operator $P_t : \mathcal{Y} \times \mathcal{Y}^* \to \mathcal{Y}$ defined by

$$P_t(y, v) \triangleq \operatorname{prox}_{tr_1}(y - tB^{-1}v) \qquad \forall (y, v) \in \mathcal{Y} \times \mathcal{Y}^*. \tag{4.10}$$

Clearly, $P_t(y, \nabla h_\eta(y)) = T_{\eta,t}(y)$ for $y \in \mathcal{Y}$.

Our numerical scheme for solving (1.3) is a derivative-free stochastic implementation of the proximal-gradient method. The random gradient estimator either employs the finite-difference estimator (4.5) (Method 'ExactLL'), or (4.9). The Pseudo-code of the resulting scheme `DFProxGrad` is reported in Algorithm 1.

Require: $y_0 \in \operatorname{dom}(r_1)$ and terminal time $N \in \mathbb{N}$. Let $(\alpha_k)_{k\geq 0}, (\beta_k)_{k\geq 0}, (\eta_k)_{k\geq 0} \subset (0, \infty)$, and $(m_k)_{k\geq 0}$ be a sequence in $\mathbb{N}$.
for $k = 0, \dots, N-1$ **do**
  if Method = 'ExactLL' **then**
    Compute $\hat{V}_{k+1} \triangleq V_\eta(y_k, U^{(m_{k+1})}, \xi^{(m_{k+1})})$
  else
    Compute $\hat{V}_{k+1} \triangleq V_\eta^{\beta_k}(y_k, U^{(m_{k+1})}, \xi^{(m_{k+1})})$
  end if
  Update $y_{k+1} = P_{\alpha_k}(y_k, \hat{V}_{k+1})$
end for

Algorithm 1 Derivative-free approximate prox-grad algorithm (`DFProxGrad`)

Remark 4.4 Algorithm `DFProxGrad` under Method 'ExactLL' requires numerical parameters $(\alpha_k)_{k=0}^{N-1}$ and a batch size sequence $(m_k)_{k=1}^{N}$. The inexact regime requires additionally a user-defined sequence $(\beta_k)_{k=0}^{N-1}$ as an additional input, which defines the error tolerance of the lower level solution mapping involved in the construction of the estimator $\hat{V}_{k+1}$. Although our algorithm is defined over a fixed time window, the exact instantiation of the parameters (e.g. step size, inexactness of lower level solutions and sampling rate) is independent of the terminal time N. Our complexity results Theorem 5.2, Theorem 5.6 and Corollary 5.4, Corollary 5.7 contain explicit expressions for these sequences culminating in good complexity bounds. An exception is Corollary 6.2, which requires an step-size schedule $(\alpha_k)_{k=0}^{N-1}$ explicitly depending on N. Our theoretical result ensures that the algorithm can be run up to time N and then continued without requiring a restart, since the step sizes do not depend on N. This stands in contrast to many theoretical guarantees for stochastic algorithms, which often rely on fixing a step size depending on N rather than considering the diminishing step sizes that we employ here.

4.6 Gap functions

In order to derive performance guarantees when running `DFProxGrad`, we need to introduce certain merit functions. The first, and most obvious merit function to use would be the observed difference in the objective function values $\Psi(y_N) - \Psi^{\mathrm{Opt}}$. Since we have no access to Ψ, but rather its smoothed counterpart Ψ_η, a conceptually implementable merit function based on the objective function gap is

$$\Delta_\eta(y_N) \triangleq \Psi_\eta(y_N) - \Psi_\eta^{\mathrm{Opt}}. \tag{4.11}$$

However, since the smoothed implicitly defined objective function $y \mapsto \Psi_\eta(y)$ is in general non-convex, measuring the distance to the global optimum value is practically not very relevant. Instead, we are focussing on functions which measure the distance to stationarity of points produced by the algorithm.

The *prox-gradient mapping* is the operator $\mathcal{G}_{\eta,t} : \mathcal{Y} \to \mathcal{Y}$ defined by

$$\mathcal{G}_{\eta,t}(y) \triangleq \frac{1}{t}(y - T_{\eta,t}(y)). \tag{4.12}$$

Indeed, thanks to the smoothing, one can show that a small norm of the prox-gradient mapping implies that approximate stationarity applies [16].[3]

The stochastic analogue to the prox-gradient mapping is the random operator $\tilde{\mathcal{G}}_{\eta,t} : \mathcal{Y} \times \Omega \to \mathcal{Y}$,

$$\tilde{\mathcal{G}}_{\eta,t}(y,\omega) \triangleq \frac{1}{t}\left(y - P_t(y, \hat{V}_{\eta,m}(y,\omega))\right). \tag{4.13}$$

Note that if $r_1 = 0$, then $\tilde{\mathcal{G}}_{\eta,t}(y,\omega) = \hat{V}_{\eta,m}(y,\omega)$ for all $(y,\omega) \in \mathcal{Y} \times \Omega$. For the complexity analysis of the inexact regime, we have to adapt the definition of the gradient mapping accordingly to

$$\tilde{\mathcal{G}}^{\beta}_{\eta,t}(y,\omega) \triangleq \frac{1}{t}\left(y - P_t(y, \hat{V}^{\beta}_{\eta,m}(y,\omega))\right). \tag{4.14}$$

4.7 Properties of the gradient estimator with exact lower level solutions

In this section we work out some a-priori error estimates on the random gradient estimator (4.5). Whenever convenient, we suppress the dependence on ω, and simply write $\hat{V}_{\eta,m}(y) \equiv V_\eta(y, U^{(m)}, \xi^{(m)})$. The first Lemma shows that our random estimator is unbiased in terms of the gradient operator of the smoothed function h_η.

Lemma 4.7 *For all $y \in \mathcal{Y}$, we have $\mathbb{E}_{\mathbb{P}}[\hat{V}_{\eta,m}(y)] = \nabla h_\eta(y)$ and*

$$\mathbb{E}_{\mathbb{P}}\left[\left|\left|\hat{V}_{\eta,m}(y)\right|\right|_*^2\right] - \|\nabla h_\eta(y)\|_*^2 \leq \frac{\mathrm{s}^2}{m},$$

where we defined $\mathrm{s} \triangleq (4+n)\,|\mathrm{lip}_0(H(\cdot,\xi))|_2$.

Proof See Appendix B. □

[3] See Appendix C for a self-contained proof.

Remark 4.5 We point out that our random estimator $\hat{V}_{\eta,m}(y)$ is an unbiased estimator of the gradient of the smoothed function h_η. It is not unbiased with respect to first-order information of the original function h. Furthermore, the variance of the estimator scales inversely with the batch size m, and scales quadratically with the dimension n. We absorb this dependency in the constant s, which will be used throughout our derived estimates.

We define the error process

$$\Delta W_{\eta,m}(y,\omega) \triangleq \hat{V}_{\eta,m}(y,\omega) - \nabla h_\eta(y) \qquad \forall (y,\omega) \in \mathcal{Y} \times \Omega. \tag{4.15}$$

An immediate corollary of Lemma 4.7 is that the error process defines essentially a martingale difference sequence:

$$\mathbb{E}_{\mathbb{P}}[\Delta W_{\eta,m}(y)] = 0, \text{ and} \tag{4.16}$$

$$\mathbb{E}_{\mathbb{P}}\left[||\Delta W_{\eta,m}(y)||_*^2\right] = \mathbb{E}_{\mathbb{P}}\left[\left|\left|\hat{V}_{\eta,m}(y)\right|\right|_*^2\right] - ||\nabla h_\eta(y)||_*^2 \leq \frac{\mathrm{s}^2}{m}. \tag{4.17}$$

Moreover, the error process can be used to estimate the prox-gradient mapping as follows:

Lemma 4.8 *We have*

$$||\mathcal{G}_{\eta,t}(y)||^2 \leq 2\left|\left|\tilde{\mathcal{G}}_{\eta,t}(y)\right|\right|^2 + 2\,||\Delta W_{\eta,m}(y)||_*^2 \qquad a.s.. \tag{4.18}$$

Proof Using the non-expansiveness of the prox-operator, we obtain

$$\begin{aligned}
||\mathcal{G}_{\eta,t}(y)||^2 &= \left|\left|\frac{1}{t}[y - P_t(y, \hat{V}_{\eta,m}(y))] + \frac{1}{t}[P_t(y, \hat{V}_{\eta,m}(y)) - T_{\eta,t}(y)]\right|\right|^2 \\
&\leq 2\left|\left|\tilde{\mathcal{G}}_{\eta,t}(y)\right|\right|^2 + \frac{2}{t^2}\left|\left|P_t(y, \hat{V}_{\eta,m}(y)) - T_{\eta,t}(y)\right|\right|^2 \\
&\leq 2\left|\left|\tilde{\mathcal{G}}_{\eta,t}(y)\right|\right|^2 + 2\left|\left|B^{-1}(\hat{V}_{\eta,m}(y) - \nabla h_\eta(y))\right|\right|^2 \\
&= 2\left|\left|\tilde{\mathcal{G}}_{\eta,t}(y)\right|\right|^2 + 2\,||\Delta W_{\eta,m}(y))||_*^2.
\end{aligned}$$

□

4.8 Properties of the gradient estimator with inexact lower level solutions

The inexactness of the solution of the lower-level problem will have its trace on the variance of the random estimator. The bias can be described by means of the following error decomposition.

Lemma 4.9 *For all $y \in \mathcal{Y}$ and $\beta > 0$, it holds*

$$\begin{aligned}\mathbb{E}_{\mathbb{P}}\left[\hat{V}^{\beta}_{\eta,m}(y)\right] =& \nabla h_{\eta}(y) \\ &+ \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{\mathbb{P}}\left[\frac{F(x^{\beta}(y+\eta U^i,\xi^i),\xi_1^i) - F(x^*(y+\eta U^i,\xi^i),\xi_1^i)}{\eta}BU^i\right],\end{aligned} \tag{4.19}$$

and

$$\begin{aligned}&\frac{1}{m}\sum_{i=1}^{m}\left|\left|\mathbb{E}_{\mathbb{P}}\left[\frac{F(x^{\beta}(y+\eta U^i,\xi^i),\xi_1^i) - F(x^*(y+\eta U^i,\xi_2^i),\xi_1^i)}{\eta}BU^i\right]\right|\right|_* \\ &\quad\leq \frac{\sqrt{n}\,|\mathrm{lip}_0(F(\cdot,\xi_1)|_1}{\eta}\mathbb{E}_{\mathbb{P}}\left[\,||x^{\beta}(y+\eta U,\xi_2) - x^*(y+\eta U,\xi_2)||_{\mathcal{X}}^p\right]^{\frac{1}{p}}.\end{aligned} \tag{4.20}$$

Proof See Appendix B. □

Let $(y_k)_k$ be the stochastic process whose sample paths are generated via Algorithm 1. The natural filtration associated with this process is $\mathcal{F}_k \triangleq \sigma(y_1,\ldots,y_k)$. Along the sample paths of this process, we can perform the following error decomposition of the random gradient estimators:

$$\hat{V}^{\beta}_{k+1} = \hat{V}_{k+1} - a_{k+1} + b_{k+1}, \tag{4.21}$$

with

$$\begin{aligned}a_{k+1} &\triangleq \frac{1}{m_{k+1}}\sum_{i=1}^{m_{k+1}}\frac{F(x^{\beta_k}(y_k,\xi^i_{2,k+1}),\xi^i_{1,k+1}) - F(x^*(y_k,\xi^i_{2,k+1}),\xi^i_{1,k+1})}{\eta}BU^i_{k+1},\\ b_{k+1} &\triangleq \frac{1}{m_{k+1}}\sum_{i=1}^{m_{k+1}}\frac{F(x^{\beta_k}(y_k+\eta U^i_{k+1},\xi^i_{2,k+1}),\xi^i_{1,k+1}) - F(x^*(y_k+\eta U^i_{k+1},\xi^i_{2,k+1}),\xi^i_{1,k+1})}{\eta}BU^i_{k+1}.\end{aligned}$$

Note that $\mathbb{E}(a_{k+1}|\mathcal{F}_k) = 0$, and we can derive a bound in $L^2(\mathbb{P})$ as the following Lemma shows.

Lemma 4.10 *Let be $p > 2$ the exponent from Definition 4.6. There exists a constant $C_F > 0$, such that*

$$\mathbb{E}\left[||a_{k+1}||_*^2\,|\mathcal{F}_k\right] \leq C_F\frac{\beta_k^2}{\eta^2}, \text{and}\mathbb{E}\left[||b_{k+1}||_*^2\,|\mathcal{F}_k\right] \leq C_F\frac{\beta_k^2}{\eta^2}\,. \tag{4.22}$$

Proof See Appendix B.3. □

5 Complexity analysis for the non-convex case

5.1 Exact lower level solution

We begin our convergence analysis in the non-convex setting, focusing on cases where the lower-level problem can be solved exactly. Our first Lemma provides an estimate on the per-iteration function progress in terms of the smoothed hyperobjective Ψ_η.

Lemma 5.1 *Consider the sequence $(y_k)_{k=0}^N$ generated by Algorithm (1) with gradient estimator (4.5). Then, for all $\eta > 0$, we have*

$$\begin{aligned}\Psi_\eta(y_{k+1}) - \Psi_\eta(y_k) \leq & -\alpha_k \left|\left|\tilde{\mathcal{G}}_{\eta,\alpha_k}(y_k)\right|\right|^2 \left(1 - \frac{\alpha_k \mathrm{lip}_1(h_\eta)}{2}\right) \\ & + \alpha_k \langle \Delta W_{k+1}, \mathcal{G}_{\eta,\alpha_k}(y_k)\rangle + \alpha_k \left|\left|\Delta W_{k+1}\right|\right|_*^2\end{aligned} \tag{5.1}$$

for all $k = 0, \ldots, N-1$.

Proof See Appendix B.4. □

Set

$$E_{k+1} \triangleq ||\Delta W_{k+1}||_*^2 + \langle \Delta W_{k+1}, \mathcal{G}_{\alpha_k}(y_k)\rangle \quad \text{and} \quad \Psi_\eta^{\mathrm{Opt}} \triangleq \min_{y\in\mathcal{Y}} \Psi_\eta(y).$$

Summing (5.1) from $k = 0, \ldots, N-1$, we obtain

$$\begin{aligned}&\sum_{k=0}^{N-1} \alpha_k \left(1 - \frac{\mathrm{lip}_1(h_\eta)\alpha_k}{2}\right) \left|\left|\tilde{\mathcal{G}}_{\eta,\alpha_k}(y_k)\right|\right|^2 \leq \Psi_\eta(y_1) - \Psi_\eta(y_{N+1}) \\ &\quad + \sum_{k=0}^{N-1} \alpha_k E_{k+1} \leq \Psi_\eta(y_1) - \Psi_\eta^{\mathrm{Opt}} + \sum_{k=0}^{N-1} \alpha_k E_{k+1}.\end{aligned}$$

Let $\mathcal{F}_k \triangleq \sigma(y_0, \ldots, y_k)$ denote the natural filtration up to time k of the process, so that

$$\mathbb{E}_k[E_{k+1}] \triangleq \mathbb{E}[E_{k+1}|\mathcal{F}_k] = \mathbb{E}[||\Delta W_{k+1}||_*^2 \,|\mathcal{F}_k] \leq \frac{\mathrm{s}^2}{m_{k+1}}, \quad \text{a.s.}.$$

Therefore, using the law of iterated expectations, we obtain

$$\mathbb{E}\left[\sum_{k=0}^{N-1} \alpha_k \left(1 - \frac{\mathrm{lip}_1(h_\eta)\alpha_k}{2}\right) \left|\left|\tilde{\mathcal{G}}_{\eta,\alpha_k}(y_k)\right|\right|^2\right] \leq \Psi_\eta(y_1) - \Psi_\eta^{\mathrm{Opt}} + \sum_{k=0}^{N-1} \frac{\alpha_k \mathrm{s}^2}{m_{k+1}}. \tag{5.2}$$

This yields our first main result in this paper:

Theorem 5.2 *Fix $N \in \mathbb{N}$ arbitrary and consider step sizes $(\alpha_k)_{k\geq 0}$ chosen in such a way that $\alpha_k \in (0, 2/\mathrm{lip}_1(h_\eta)]$, with $\alpha_k < 2/\mathrm{lip}_1(h_\eta)$ for at least one $k \in \{0, \ldots, N-1\}$. Let $(y_k)_{k=0}^N$ be generated by Algorithm 1 with gradient estimator (4.5). On $(\Omega, \mathcal{F}, \mathbb{P})$ define an independent random variable $\kappa : \Omega \to \{0, \ldots, N-1\}$ with probability mass function*

$$p(k) \triangleq \mathbb{P}(\kappa = k) \triangleq \frac{\alpha_k - \alpha_k^2 \mathrm{lip}_1(h_\eta)/2}{\sum_{t=0}^{N-1}(\alpha_t - \alpha_t^2 \mathrm{lip}_1(h_\eta)/2)}, \quad k \in \{0, \ldots, N-1\}. \tag{5.3}$$

Then

$$\mathbb{E}\left[\left|\left|\tilde{\mathcal{G}}_{\eta,\alpha_\kappa}(y_\kappa)\right|\right|^2\right] \leq \frac{\Psi_\eta(y_1) - \Psi_\eta^{\mathrm{Opt}} + \sum_{k=0}^{N-1} \frac{\alpha_k \mathrm{s}^2}{m_{k+1}}}{\sum_{t=0}^{N-1}(\alpha_t - \alpha_t^2 \mathrm{lip}_1(h_\eta)/2)}. \tag{5.4}$$

Proof Using eq. (5.2), together with the observation that

$$\mathbb{E}\left[\left|\left|\tilde{\mathcal{G}}_{\eta,\alpha_\kappa}(y_\kappa)\right|\right|^2\right] = \sum_{k=0}^{N-1} \frac{\alpha_k - \alpha_k^2 \mathrm{lip}_1(h_\eta)/2}{\sum_{t=0}^{N-1}(\alpha_t - \alpha_t^2 \mathrm{lip}_1(h_\eta)/2)} \mathbb{E}\left[\left|\left|\tilde{\mathcal{G}}_{\eta,\alpha_k}(y_k)\right|\right|^2\right],$$

the thesis follows. □

A few remarks are in order.

Remark 5.1 Due to the ratio $\frac{\alpha_k}{m_{k+1}}$, there is a trade-off between too aggressive step-sizes and the size of the mini-batches. In fact, consider an arbitrary step-size rule with $\alpha_k \leq \frac{1}{\mathrm{lip}_1(h_\eta)}$. This bound implies $\frac{\mathrm{lip}_1(h_\eta)}{2}\alpha_k^2 \leq \frac{1}{2}\alpha_k$. Therefore, the numerator in our complexity bound (5.4) can be simplified to

$$\mathbb{E}\left[\left|\left|\tilde{\mathcal{G}}_{\eta,\alpha_\kappa}(y_\kappa)\right|\right|^2\right] \leq \frac{\Psi_\eta(y_1) - \Psi_\eta^{\mathrm{Opt}} + \sum_{k=0}^{N-1} \frac{\alpha_k \mathrm{s}^2}{m_{k+1}}}{\sum_{t=0}^{N-1}(\alpha_t/2)}.$$

Concretely, choosing $\alpha_k = \frac{2\theta}{\mathrm{lip}_1(h_\eta)\sqrt{k+1}}$ with $\theta \in (0, 1/2)$, and mini-batches $m_{k+1} = \mathrm{a}\sqrt{k+1}$, with $\mathrm{a} > 0$, we can recover the typical $\mathcal{O}(\log(N)/\sqrt{N})$ complexity estimate for proximal gradient methods. Indeed, such a step size choice yields the iteration complexity upper bound

$$\mathbb{E}\left[\left|\left|\tilde{\mathcal{G}}_{\eta,\alpha_\kappa}(y_\kappa)\right|\right|^2\right] \leq \frac{\frac{\mathrm{lip}_1(h_\eta)}{\beta}(\Psi_\eta(y_1) - \Psi_\eta^{\mathrm{Opt}}) + \frac{2\mathrm{s}^2}{\mathrm{a}}(1 + \log(N))}{\sqrt{N}}.$$

On the contrary, if a constant step size and constant mini-batch estimation strategy is adopted, then we see that convergence with respect to our merit function can only

happen up to a plateau, a well-known fact when using stochastic approximation [7, 25]. Specifically, taking constant mini-batches $m_{k+1} = m$ and constant step-sizes $\alpha_k = \frac{2\beta}{\mathrm{lip}_1(h_\eta)}$ for all $k \in \{0, \dots, N-1\}$ and some $\beta \in (0, 1/2)$, then our complexity bound is readily seen to become

$$\mathbb{E}\left[\left|\left|\tilde{\mathcal{G}}_{\eta,\alpha_\kappa}(y_\kappa)\right|\right|^2\right] \leq \frac{\mathrm{lip}_1(h_\eta)(\Psi_\eta(y_1) - \Psi_\eta^{\mathrm{Opt}})}{\beta(2-\beta)N} + \frac{2\mathrm{s}^2}{(2-\beta)m}.$$

Remark 5.2 Theorem 5.2 gives a finite-time complexity estimate of the gap function evaluated at a randomly chosen iterate. Hence, the complexity bound should be interpreted as a criticality measure of an averaged output based on the sequence generated by the stochastic process. Such performance estimates in terms of averaged quantities are typical in stochastic programming [24, 37].

Our next result is a complexity estimate in terms of the prox-gradient mapping involving the deterministic gradient ∇h_η, instead of the stochastic approximation.

Corollary 5.3 *Under the same assumptions as in Theorem 5.2 we assume additionally that the step sizes α_k are chosen such that $\alpha_k \in (0, 2/\mathrm{lip}_1(h_\eta)]$, with $\alpha_k < 2/\mathrm{lip}_1(h_\eta)$ for at least one $k \in \{0, \dots, N-1\}$. Let $(y_k)_{k=0}^N$ be generated by Algorithm 1 with gradient estimator (4.5) and let $\kappa : \Omega \to \{0, \dots, N-1\}$ be the discrete random variable with distribution (5.3). Then,*

$$\mathbb{E}[||\mathcal{G}_{\eta,\alpha_\kappa}(y_\kappa)||^2] \leq \frac{4(\Psi_\eta(y_1) - \Psi_\eta^{\mathrm{Opt}})}{\sum_{t=0}^{N-1}(2\alpha_t - \alpha_t^2 \mathrm{lip}_1(h_\eta))} + \frac{\sum_{k=0}^{N-1} \frac{2\mathrm{s}^2}{m_{k+1}}(4\alpha_k - \alpha_k^2 \mathrm{lip}_1(h_\eta))}{\sum_{t=0}^{N-1}(2\alpha_t - \alpha_t^2 \mathrm{lip}_1(h_\eta))}. \quad (5.5)$$

Proof From Lemma 4.8 we readily obtain

$$\begin{aligned}\frac{\alpha_k}{2}\left(1 - \frac{\alpha_k \mathrm{lip}_1(h_\eta)}{2}\right)||\mathcal{G}_{\eta,\alpha_k}(y_k)||^2 \leq & \alpha_k\left(1 - \frac{\alpha_k \mathrm{lip}_1(h_\eta)}{2}\right)\left|\left|\tilde{\mathcal{G}}_{\eta,\alpha_k}(y_k)\right|\right|^2 \\ & + \alpha_k\left(1 - \frac{\alpha_k \mathrm{lip}_1(h_\eta)}{2}\right)||\Delta W_{k+1}||_*^2.\end{aligned}$$

Consequently, using (5.4):

$$\begin{aligned}\frac{1}{2}\mathbb{E}[||\mathcal{G}_{\eta,\alpha_\kappa}(y_\kappa)||^2] &= \frac{1}{2}\sum_{k=0}^{N-1}\frac{\alpha_k-\alpha_k^2\mathrm{lip}_1(h_\eta)/2}{\sum_{t=0}^{N-1}(\alpha_t-\alpha_t^2\mathrm{lip}_1(h_\eta)/2)}\mathbb{E}[||\mathcal{G}_{\eta,\alpha_k}(y_k)||^2]\\ &\le \sum_{k=0}^{N-1}\frac{\alpha_k-\alpha_k^2\mathrm{lip}_1(h_\eta)/2}{\sum_{t=0}^{N-1}(\alpha_t-\alpha_t^2\mathrm{lip}_1(h_\eta)/2)}\mathbb{E}[\left|\left|\tilde{\mathcal{G}}_{\eta,\alpha_k}(y_k)\right|\right|^2]\\ &\quad+\sum_{k=0}^{N-1}\frac{\alpha_k(1-\frac{\alpha_k\mathrm{lip}_1(h_\eta)}{2})\mathbb{E}[||\Delta W_{k+1}||_*^2]}{\sum_{t=0}^{N-1}(\alpha_t-\alpha_t^2\mathrm{lip}_1(h_\eta)/2)}\\ &\le \frac{\Psi_\eta(y_1)-\Psi_\eta^{\mathrm{Opt}}+\sum_{k=0}^{N-1}\frac{\alpha_k\mathrm{s}^2}{m_{k+1}}}{\sum_{t=0}^{N-1}(\alpha_t-\alpha_t^2\mathrm{lip}_1(h_\eta)/2)}+\sum_{k=0}^{N-1}\frac{\alpha_k(1-\frac{\alpha_k\mathrm{lip}_1(h_\eta)}{2})\frac{\mathrm{s}^2}{m_{k+1}}}{\sum_{t=0}^{N-1}(\alpha_t-\alpha_t^2\mathrm{lip}_1(h_\eta)/2)}\\ &= \frac{\Psi_\eta(y_1)-\Psi_\eta^{\mathrm{Opt}}}{\sum_{t=0}^{N-1}(\alpha_t-\alpha_t^2\mathrm{lip}_1(h_\eta)/2)}+\frac{\sum_{k=0}^{N-1}\alpha_k\frac{\mathrm{s}^2}{m_{k+1}}(2-\frac{\alpha_k\mathrm{lip}_1(h_\eta)}{2})}{\sum_{t=0}^{N-1}(\alpha_t-\alpha_t^2\mathrm{lip}_1(h_\eta)/2)}.\end{aligned}$$

□

Corollary 5.4 *For a time window $N \ge 2$, we choose the step size $\alpha_k = \frac{2\theta}{\mathrm{lip}_1(h_\eta)\sqrt{k+1}}, \theta \in (0, 1/2), k \ge 0$, and the sampling rate $m_{k+1} = \mathrm{a}\sqrt{k+1}, \mathrm{a} > 0$. Let $(y_k)_{k=0}^N$ be generated by Algorithm 1 with gradient estimator (4.5). Then, we have*

$$\mathbb{E}[||\mathcal{G}_{\eta,\alpha_\kappa}(y_\kappa)||^2] \le \frac{2\mathrm{lip}_1(h_\eta)(\Psi_\eta(y_1)-\Psi_\eta^{\mathrm{Opt}})}{\theta\sqrt{N}}+\frac{\frac{8\mathrm{s}^2}{\mathrm{a}}(1+\log(N))}{\sqrt{N}}.$$

The total number of calls to the stochastic oracle and lower level solutions to find a point $y \in \mathcal{Y}$ such that $\mathbb{E}[||\mathcal{G}_\eta(y)||^2] \le \varepsilon$ is bounded by $\mathcal{O}(\varepsilon^{-3})$.

Proof We start with recalling a simple integral bound. Note that

$$\sum_{t=1}^{N}\frac{1}{\sqrt{t}} \ge \int_0^N \frac{1}{\sqrt{x+1}}\,\mathrm{d}x = 2\sqrt{N+1}-2 \ge \sqrt{N}$$

for $N \ge 2$. Using this bound, the specific choices for the step sizes and the mini-batch size, lead to the following inequalities:

$$\begin{aligned}\mathbb{E}[||\mathcal{G}_{\eta,\alpha_\kappa}(y_\kappa)||^2] &\leq \frac{4(\Psi_\eta(y_1)-\Psi_\eta^{\mathrm{Opt}})}{\sum_{t=0}^{N-1}\alpha_t}+\frac{\sum_{k=0}^{N-1}\frac{2\mathrm{s}^2}{m_{k+1}}(4\alpha_k-\alpha_k^2\mathrm{lip}_1(h_\eta))}{\sum_{t=0}^{N-1}\alpha_t}\\ &\leq \frac{4(\Psi_\eta(y_1)-\Psi_\eta^{\mathrm{Opt}})}{\sum_{t=0}^{N-1}\alpha_t}+\frac{\sum_{k=0}^{N-1}\frac{8\mathrm{s}^2}{m_{k+1}}\alpha_k}{\sum_{t=0}^{N-1}\alpha_t}\\ &\leq \frac{2\mathrm{lip}_1(h_\eta)(\Psi_\eta(y_1)-\Psi_\eta^{\mathrm{Opt}})}{\beta\sqrt{N}}+\frac{\frac{8\mathrm{s}^2}{\mathrm{a}}(1+\log(N))}{\sqrt{N}}.\end{aligned}$$

Hence, by choosing $N\in O(\varepsilon^{-2})$ we ensure that $\mathbb{E}[||\mathcal{G}_{\eta,\alpha_\kappa}(y_\kappa)||^2]\in O(\varepsilon^{-1})$. Hence, the iteration complexity of the method is bounded by $\mathcal{O}(\varepsilon^{-2})$. Now, to bound the oracle complexity, note that in each iteration of Algorithm 1 we need m_{k+1} Gaussian vectors U and the same number of random vectors $\xi=(\xi_1,\xi_2)$ to construct the random vector $\sum_{i=1}^{m_{k+1}}\frac{H(y^k+\eta U^i_{k+1},\xi^i_{k+1})}{\eta}BU^i_{k+1}$. We therefore have m_{k+1} calls of the stochastic function $H(\cdot,\xi)$ in every single iteration. The total number of calls is thus $\sum_{k=0}^{N-1}m_{k+1}=\mathrm{a}\sum_{k=1}^{N}\sqrt{k}\leq\frac{2\mathrm{a}}{3}N^{3/2}$ As $N\in\mathcal{O}(\varepsilon^{-2})$, the oracle complexity is upper bounded by $\mathcal{O}(\varepsilon^{-3})$. Similarly, in every iteration we need m_{k+1} solutions of the lower level problem. Hence, by the above computation, the total number of lower level solves is bounded by $\mathcal{O}(\varepsilon^{-3})$. □

5.2 Inexact lower level solution

Using this merit function and the definition of the error increment

$$\Delta W^\beta_{k+1}\triangleq\hat{V}^{\beta_k}_{k+1}-\nabla h_\eta(y_k)=\hat{V}_{k+1}-a_{k+1}+b_{k+1}-\nabla h_\eta(y_k)=\Delta W_{k+1}-a_{k+1}+b_{k+1},$$

we can repeat the one-step analysis of the exact case to obtain the bound

$$\begin{aligned}\Psi_\eta(y_{k+1})-\Psi_\eta(y_k)\leq&-\alpha_k\left(1-\frac{\alpha_k\mathrm{lip}_1(h_\eta)}{2}\right)||\tilde{\mathcal{G}}^{\beta_k}_{\eta,\alpha_k}(y_k)||^2\\&+\alpha_k\langle\Delta W^\beta_{k+1},\mathcal{G}_{\eta,\alpha_k}(y_k)\rangle+\alpha_k\left|\left|\Delta W^\beta_{k+1}\right|\right|^2_*.\end{aligned}\tag{5.6}$$

Lemma 4.8 generalizes in the inexact case in the following way:

Lemma 5.5 *We have*

$$||\mathcal{G}_{\eta,t}(y)||^2\leq 2\left|\left|\tilde{\mathcal{G}}^\beta_{\eta,t}(y)\right|\right|^2+2\left|\left|\Delta W^\beta_{\eta,m}(y)\right|\right|^2_*\qquad a.s..\tag{5.7}$$

Proof The assertion follows line by line as in Lemma 4.8 by replacing $V_{\eta,m}(y)$ with $V^\beta_{\eta,m}(y)$. □

Using this lemma directly in (5.6), we see that for $\alpha_k\in(0,2/\mathrm{lip}_1(h_\eta)]$

$$\Psi_\eta(y_{k+1}) - \Psi_\eta(y_k) \leq -\frac{\alpha_k}{2}\left(1 - \frac{\alpha_k \text{lip}_1(h_\eta)}{2}\right)||\mathcal{G}_{\eta,\alpha_k}(y_k)||^2 + \alpha_k\langle \Delta W^\beta_{k+1}, \mathcal{G}_{\eta,\alpha_k}(y_k)\rangle + \alpha_k\left|\left|\Delta W^\beta_{k+1}\right|\right|^2_* + \alpha_k\left(1 - \frac{\alpha_k \text{lip}_1(h_\eta)}{2}\right)\left|\left|\Delta W^\beta_{k+1}\right|\right|^2_*.$$

Applying Young's inequality of the inner product, we conclude that for arbitrary $\delta > 0$

$$\Psi_\eta(y_{k+1}) - \Psi_\eta(y_k) \leq -\frac{\alpha_k}{2}\left(1 - \frac{1}{\delta} - \frac{\alpha_k \text{lip}_1(h_\eta)}{2}\right)||\mathcal{G}_{\eta,\alpha_k}(y_k)||^2 + \alpha_k\left(2 + \frac{\delta}{2} - \frac{\alpha_k \text{lip}_1(h_\eta)}{2}\right)\left|\left|\Delta W^\beta_{k+1}\right|\right|^2_*.$$

Rearranging this expression and summing both sides from $k = 0$ to $N-1$, we remain with

$$\sum_{k=0}^{N-1}\frac{\alpha_k}{2}\left(\frac{\delta-1}{\delta} - \frac{\alpha_k \text{lip}_1(h_\eta)}{2}\right)||\mathcal{G}_{\eta,\alpha_k}(y_k)||^2 \leq \Psi_\eta(y_1) - \Psi_\eta^{\text{Opt}} + \sum_{k=0}^{N-1}\alpha_k\left(\frac{4+\delta}{2} - \frac{\alpha_k \text{lip}_1(h_\eta)}{2}\right)\left|\left|\Delta W^\beta_{k+1}\right|\right|^2_*. \tag{5.8}$$

Since $\left|\left|\Delta W^\beta_{k+1}\right|\right|^2_* \leq 3\,||\Delta W_{k+1}||^2_* + 3\,||a_{k+1}||^2_* + 3\,||b_{k+1}||^2_*$, we can take iteratively conditional expectations to obtain the main complexity bound for the inexact regime.

Theorem 5.6 *Suppose that the step sizes α_k are chosen such that $\alpha_k \in (0, \frac{2(\delta-1)}{\delta \text{lip}_1(h_\eta)}]$, with $\alpha_k < \frac{2(\delta-1)}{\delta \text{lip}_1(h_\eta)}$ for at least one $k \in \{1,\ldots,N\}$. Let $(y_k)_{k=0}^N$ be generated by Algorithm 1 with inexact gradient estimator (4.9), $\delta > 1$ and $r, s \geq 1$ such that $\frac{2s(r-1)}{r} = p \geq 2$, where p is the exponent in Definition 4.6. On $(\Omega, \mathcal{F}, \mathbb{P})$ define an independent random variable $\kappa : \Omega \to \{0,\ldots,N-1\}$ with probability mass function*

$$p(k) = \mathbb{P}(\kappa = k) \triangleq \frac{\alpha_k\frac{\delta-1}{\delta} - \alpha_k^2 \text{lip}_1(h_\eta)/2}{\sum_{t=0}^{N-1}(\alpha_t\frac{\delta-1}{\delta} - \alpha_t^2 \text{lip}_1(h_\eta)/2)} \quad \forall k \in \{0,\ldots,N-1\}.$$

Let $D_k \triangleq \frac{3(4+\delta)}{2}\left(\frac{s^2}{m_{k+1}} + C_F\frac{2\beta_k^2}{\eta^2}\right)$ for $k = 0,\ldots,N-1$. Then,

$$\frac{1}{2}\mathbb{E}[||\mathcal{G}_{\eta,\alpha_\kappa}(y_\kappa)||^2] \leq \frac{\Psi_\eta(y_1) - \Psi_\eta^{\mathrm{Opt}}}{\sum_{t=0}^{N-1}(\alpha_t \frac{\delta-1}{\delta} - \alpha_t^2 \mathrm{lip}_1(h_\eta)/2)} + \frac{\sum_{k=0}^{N-1}\alpha_k D_k}{\sum_{t=0}^{N-1}(\alpha_t \frac{\delta-1}{\delta} - \alpha_t^2 \mathrm{lip}_1(h_\eta)/2)}. \tag{5.9}$$

Proof Starting from (5.8) the proof follows the same lines as the proof of Corollary 5.3. □

Similarly as in the exact case, we again find a trade-off between aggressive step-sizes and the size of the mini-batches. However, in the inexact computational model, we additionally observe a trade-off between the step-size schedule and the accuracy tolerance β_k in the lower level problem. In order to ensure convergence, the estimate developed in Theorem 5.6 reveals the condition $\sum_{k=0}^{\infty}\alpha_k\beta_k^2 < \infty$. This observation allows us to design explicit parameter sequences with interpretable complexity bounds. In addition, similarly as in the exact case, we require an upper bound on the step-size α_k to ensure the divergence of the denominator. These conditions lead to the following refined error estimate.

Corollary 5.7 *Under the same conditions as in Theorem 5.6, let be $\delta > 1$ and consider a step-size $\alpha_k \leq \frac{\min\{\delta-1,1\}}{\delta \mathrm{lip}_1(h_\eta)}$. Let $(y_k)_{k=0}^N$ be generated by Algorithm 1 with gradient estimator (4.9). Then*

$$\frac{1}{2}\mathbb{E}[||\mathcal{G}_{\eta,\alpha_\kappa}(y_\kappa)||^2] \leq \frac{\Psi_\eta(y_1) - \Psi_\eta^{\mathrm{Opt}}}{\frac{\delta-1}{2\delta}\sum_{t=0}^{N-1}\alpha_t} + \frac{\sum_{k=0}^{N-1}\alpha_k D(\frac{1}{m_{k+1}} + \beta_k^2)}{\frac{\delta-1}{2\delta}\sum_{t=0}^{N-1}\alpha_t},$$

where $D \triangleq \frac{3(4+\delta)}{2}\max\{\mathrm{s}^2, \frac{2}{\eta^2}C_F\}$.

Proof We observe that $D_k \leq D(\frac{1}{m_{k+1}} + \beta_k^2)$ and with the bound on α_k we have

$$\alpha_t\frac{\delta-1}{\delta} - \alpha_t^2\mathrm{lip}_1(h_\eta)/2 \geq \alpha_t(\delta-1)(\frac{1}{\delta} - \frac{1}{2\delta}) = \alpha_t\frac{\delta-1}{2\delta}.$$

Combining these estimates with the bound (5.9) verifies the assertion. □

The constants appearing in the upper complexity bound can be well balanced via a judicious choice of δ. For instance, setting $\delta = 2$, the step-size policy $\alpha_k = \frac{2\theta}{\mathrm{lip}_1(h_\eta)\sqrt{k}}$ with $\theta \in (0, 1/2)$, and choosing the sampling rate $m_k = \mathrm{a}\sqrt{k}$ and the accuracy tolerance $\beta_k = \mathrm{b}k^{-\frac{1}{4}}$ with constants $\mathrm{a}, \mathrm{b} > 0$, we obtain the overall complexity estimate

$$\frac{1}{2}\mathbb{E}[||\mathcal{G}_{\eta,\alpha_\kappa}(y_\kappa)||^2] \leq \frac{\frac{\mathrm{lip}_1(h_\eta)}{\theta}(\Psi_\eta(y_1) - \Psi_\eta^{\mathrm{Opt}}) + 2D(\frac{1}{\mathrm{a}} + \mathrm{b}^2)(1+\log(N))}{\sqrt{N}},$$

which resembles those of the exact oracle case.

6 The convex case with inexact lower level solution

We now turn to the case in which the implicit function h is convex. In this special setting, the smoothed function h_η is also convex and Lipschitz continuous. By the subgradient inequality, we have for all $y \in \mathcal{Y}$ and $g \in \partial h(y)$

$$h_\eta(y) = \mathbb{E}[h(y + \eta U)] \geq \mathbb{E}[h(y) + \langle g, \eta U \rangle] = h(y). \tag{6.1}$$

Moreover, in the convex case, it holds true that $\nabla h_\eta(y)$ always belongs to some δ -subdifferential of the function h (cf. Lemma 4.4). In this section, we make an additional boundedness assumption on the bilevel problem.

Assumption 7 The domain $\mathrm{dom}\,(r_1)$ is bounded.

Theorem 6.1 *Assume that the implicit function $y \mapsto h(y)$ is convex and Assumption 7 holds. Assume that the step-size policy $(\alpha_k)_{k\geq 0}$ satisfies*

$$0 < \alpha_{N-1} \leq \alpha_{N-2} \leq \ldots \leq \alpha_1 \leq \frac{1}{\mathrm{lip}_1(h_\eta)} and \alpha_k + \alpha_{k-1} \leq \frac{1}{\mathrm{lip}_1(h_\eta)}, \quad forall\ k = 1, \ldots, N-1. \tag{6.2}$$

Let $(y_k)_{k=0}^N$ be generated by Algorithm 1 with gradient estimator (4.9) and let $\kappa : \Omega \to \{0, \ldots, N-1\}$ be an independent random variable, with probability mass function

$$p(k) = \mathbb{P}(\kappa = k) \triangleq \frac{\mathrm{a}_k}{A_N}, \quad A_N \triangleq \sum_{t=0}^{N-1} \mathrm{a}_t,\ \mathrm{a}_k \triangleq \alpha_k - \alpha_k^2 \mathrm{lip}_1(h_\eta). \tag{6.3}$$

Then, we have

$$\mathbb{E}[\Psi(y_\kappa) - \Psi^{\mathrm{Opt}}] \leq \frac{\sum_{k=0}^{N-1} \frac{\alpha_k^2}{m_{k+1}} D_k + \frac{M\sqrt{C_F}}{\eta} \sum_{k=0}^{N-1} \mathrm{a}_k \beta_k + M/2 + \alpha_1 \Delta\Psi_1}{A_N} + \eta\sqrt{n}\mathrm{lip}_0(h), \tag{6.4}$$

where $D_k \triangleq 3\left(\mathrm{s}^2/2 + \frac{\beta_k^2 m_{k+1}}{\eta^2} C_F\right)$ and $M \triangleq \sup_{y_1, y_2 \in \mathrm{dom}\,(r_1)} ||y_1 - y_2||^2$.

Proof Let y^* denote a solution of the original problem (3.4). Let $(\alpha_k)_k$ be a sequence of step-sizes, satisfying $0 \leq \alpha_k < \frac{1}{\mathrm{lip}_1(h_\eta)}$. For $\eta > 0$ we then have

$$\Psi_\eta(y_{k+1}) - \Psi_\eta(y^*) = h_\eta(y_{k+1}) - h_\eta(y_k) + h_\eta(y_k) - h_\eta(y^*) + r_1(y_{k+1}) - r_1(y_k).$$

Using the descent property (2.1) and the convexity of the smoothed implicit function h_η, we deduce that

$$h_\eta(y_{k+1}) - h_\eta(y_k) \leq \langle \nabla h_\eta(y_k), y_{k+1} - y_k \rangle + \frac{\mathrm{lip}_1(h_\eta)}{2} ||y_{k+1} - y_k||^2, \text{ and}$$
$$h_\eta(y_k) - h_\eta(y^*) \leq \langle \nabla h_\eta(y_k), y_k - y^* \rangle.$$

Recall that $\Delta W^\beta_{k+1} = \hat{V}^\beta_{k+1} - \nabla h_\eta(y_k)$. Then, we continue from the above with

$$\begin{aligned}\Psi_\eta(y_{k+1}) - \Psi_\eta(y^*) \leq & \langle \Delta W^\beta_{k+1}, y_k - y_{k+1} \rangle + \frac{\mathrm{lip}_1(h_\eta)}{2} ||y_{k+1} - y_k||^2 \\ & + \langle \Delta W^\beta_{k+1}, y^* - y_k \rangle + r_1(y_{k+1}) - r_1(y^*) + \langle \hat{V}^\beta_{k+1}, y_{k+1} - y^* \rangle.\end{aligned}$$

By definition of the point y_{k+1}, we have

$$r_1(y^*) \geq r_1(y_{k+1}) + \frac{1}{\alpha_k} \langle B(y_k - y_{k+1}), y^* - y_{k+1} \rangle - \langle \hat{V}^\beta_{k+1}, y^* - y_{k+1} \rangle.$$

Combining these two estimates, we can continue with

$$\begin{aligned}\Psi_\eta(y_{k+1}) - \Psi_\eta(y^*) &\leq \langle \Delta W^\beta_{k+1}, y_k - y_{k+1} \rangle + \frac{\mathrm{lip}_1(h_\eta)}{2} ||y_{k+1} - y_k||^2 + \langle \Delta W^\beta_{k+1}, y_k - y^* \rangle \\ &\quad + \frac{1}{\alpha_k} \langle B(y_k - y_{k+1}), y_{k+1} - y^* \rangle \\ &= \langle \Delta W^\beta_{k+1}, y_k - y_{k+1} \rangle + \frac{\mathrm{lip}_1(h_\eta)}{2} ||y_{k+1} - y_k||^2 + \langle \Delta W^\beta_{k+1}, y^* - y_k \rangle \\ &\quad + \frac{1}{\alpha_k} \left[\frac{1}{2} ||y_k - y^*||^2 - \frac{1}{2} ||y_{k+1} - y_k||^2 - \frac{1}{2} ||y_{k+1} - y^*||^2 \right].\end{aligned}$$

Note that $ax - \frac{bx^2}{2} \leq \frac{a^2}{2b}$ for all $x \geq 0$, implying that

$$\begin{aligned}&\langle \Delta W^\beta_{k+1}, y_k - y_{k+1} \rangle + \frac{\mathrm{lip}_1(h_\eta)\alpha_k - 1}{2\alpha_k} ||y_{k+1} - y_k||^2 \\ &\qquad \leq \left|\left| \Delta W^\beta_{k+1} \right|\right|_* \cdot ||y_{k+1} - y_k|| + \frac{\mathrm{lip}_1(h_\eta)\alpha_k - 1}{2\alpha_k} ||y_{k+1} - y_k||^2 \\ &\qquad \leq \frac{\alpha_k}{2(1 - \alpha_k \mathrm{lip}_1(h_\eta))} \left|\left| \Delta W^\beta_{k+1} \right|\right|^2_*.\end{aligned}$$

Thus, multiplying both sides in the penultimate display by $(\alpha_k - \alpha_k^2 \mathrm{lip}_1(h_\eta))$, we can continue the bound by

$$\begin{aligned}&(\alpha_k - \alpha_k^2 \mathrm{lip}_1(h_\eta))[\Psi_\eta(y_{k+1}) - \Psi_\eta(y^*)] \\ &\quad \leq \frac{\alpha_k^2}{2} \left|\left| \Delta W^\beta_{k+1} \right|\right|^2_* + (\alpha_k - \alpha_k^2 \mathrm{lip}_1(h_\eta)) \langle \Delta W^\beta_{k+1}, y^* - y_k \rangle \\ &\quad + (1 - \alpha_k \mathrm{lip}_1(h_\eta)) \left[\frac{1}{2} ||y_k - y^*||^2 - \frac{1}{2} ||y_{k+1} - y^*||^2 \right].\end{aligned}$$

Using (6.1), we note that $\Psi_\eta(y_{k+1}) \geq \Psi(y_{k+1})$. Additionally, Lemma 4.2 yields $\Psi_\eta(y^*) \geq \Psi^{\mathrm{Opt}} - \eta \mathrm{lip}_0(h)\sqrt{n}$. This allows us to bound the objective function gap by

$$
\begin{aligned}
&(\alpha_k - \alpha_k^2 \mathrm{lip}_1(h_\eta))[\Psi(y_{k+1}) - \Psi^{\mathrm{Opt}}] \\
&\quad \leq (\alpha_k - \alpha_k^2 \mathrm{lip}_1(h_\eta))[\Psi_\eta(y_{k+1}) - \Psi_\eta(y^*)] + \eta \mathrm{lip}_0(h)\sqrt{n}(\alpha_k - \alpha_k^2 \mathrm{lip}_1(h_\eta)) \\
&\leq \frac{\alpha_k^2}{2} \left|\left| \Delta W_{k+1}^\beta \right|\right|_*^2 + (\alpha_k - \alpha_k^2 \mathrm{lip}_1(h_\eta)) \langle \Delta W_{k+1}^\beta, y^* - y_k \rangle \\
&\quad + (1 - \alpha_k \mathrm{lip}_1(h_\eta)) \left[\frac{1}{2} ||y_k - y^*||^2 - \frac{1}{2} ||y_{k+1} - y^*||^2 \right] \\
&\quad + \eta \mathrm{lip}_0(h)\sqrt{n}(\alpha_k - \alpha_k^2 \mathrm{lip}_1(h_\eta)).
\end{aligned}
$$

To bound the terms on the right-hand side, we first use error decomposition (4.21) to bound the first addendum by $\left|\left| \Delta W_{k+1}^\beta \right|\right|_*^2 \leq 3 ||\Delta W_{k+1}||_*^2 + 3 ||a_{k+1}||_*^2 + 3 ||b_{k+1}||_*^2$, as well as the second addendum $\langle \Delta W_{k+1}^\beta, y^* - y_k \rangle = \langle \Delta W_{k+1}, y^* - y_k \rangle - \langle a_{k+1}, y^* - y_k \rangle + \langle b_{k+1}, y^* - y_k \rangle$. Hence, taking conditional expectations on both sides, we continue with

$$
\begin{aligned}
\mathbb{E}_k \left[(\alpha_k - \alpha_k^2 \mathrm{lip}_1(h_\eta))(\Psi(y_{k+1}) - \Psi(y^*)) \right] \leq & \frac{3\alpha_k^2}{2} \frac{s^2}{m_{k+1}} \\
& + 3\alpha_k^2 \frac{n\beta_k^2}{\eta^2} C_F + \mathbb{E}_k \left[(\alpha_k - \alpha_k^2 \mathrm{lip}_1(h_\eta)) \langle b_{k+1}, y_k - y^* \rangle \right] \\
& + (1 - \alpha_k \mathrm{lip}_1(h_\eta)) \mathbb{E}_k \left[\frac{1}{2} ||y_k - y^*||^2 - \frac{1}{2} ||y_{k+1} - y^*||^2 \right] \\
& + \eta \mathrm{lip}_0(h)\sqrt{n}(\alpha_k - \alpha_k^2 \mathrm{lip}_1(h_\eta)) \\
\leq & \frac{3\alpha_k^2}{2} \frac{s^2}{m_{k+1}} + 3\alpha_k^2 \frac{n\beta_k^2}{\eta^2} C_F + M(\alpha_k - \alpha_k^2 \mathrm{lip}_1(h_\eta)) \mathbb{E}_k \left[||b_{k+1}||_* \right] \\
& + (1 - \alpha_k \mathrm{lip}_1(h_\eta)) \mathbb{E}_k \left[\frac{1}{2} ||y_k - y^*||^2 - \frac{1}{2} ||y_{k+1} - y^*||^2 \right] \\
& + \eta \mathrm{lip}_0(h)\sqrt{n}(\alpha_k - \alpha_k^2 \mathrm{lip}_1(h_\eta)),
\end{aligned}
$$

where the second inequality uses Cauchy-Schwarz and the bound $M \geq ||y_k - y^*||^2$, which holds thanks to Assumption 7. Since the step size sequence $(\alpha_k)_k$ is non-decreasing and satisfies condition (6.2), we can continue to obtain

$$
\begin{aligned}
&\sum_{k=0}^{N-1} (1 - \alpha_k \mathrm{lip}_1(h_\eta)) \left(\frac{1}{2} ||y_k - y^*||^2 - \frac{1}{2} ||y_{k+1} - y^*||^2 \right) \\
&\quad = (1 - \alpha_1 \mathrm{lip}_1(h_\eta)) \frac{1}{2} ||y_1 - y^*||^2 + \sum_{k=2}^{N} \mathrm{lip}_1(h_\eta)(\alpha_k - \alpha_{k+1}) \frac{1}{2} ||y_{k+1} - y^*||^2 \\
&\qquad - (1 - \mathrm{lip}_1(h_\eta)\alpha_N) \frac{1}{2} ||y_{N+1} - y^*||^2 \\
&\quad \leq (1 - \alpha_N \mathrm{lip}_1(h_\eta)) \frac{M}{2}.
\end{aligned}
$$

Next, calling $\Delta\Psi_k \triangleq \Psi(y_k) - \Psi^{\mathrm{Opt}}$ and $\mathrm{a}_k \triangleq \alpha_k - \alpha_k^2 \mathrm{lip}_1(h_\eta)$, we deduce that

$$\sum_{k=0}^{N-1} \mathrm{a}_k \Delta\Psi_k = \sum_{k=0}^{N-1} \mathrm{a}_k \Delta\Psi_{k+1} + \sum_{k=0}^{N-1} \mathrm{a}_k (\Delta\Psi_k - \Delta\Psi_{k+1})$$

and

$$\begin{aligned}
\sum_{k=0}^{N-1} \mathrm{a}_k (\Delta\Psi_k - \Delta\Psi_{k+1}) &= \sum_{k=0}^{N-1} \mathrm{a}_k \Delta\Psi_k - \sum_{k=0}^{N-1} \mathrm{a}_k \Delta\Psi_{k+1} \\
&= \mathrm{a}_1 \Delta\Psi_1 + \sum_{k=1}^{N-1} \mathrm{a}_k \Delta\Psi_k - \sum_{k=0}^{N-1} \mathrm{a}_k \Delta\Psi_{k+1} \\
&\leq \mathrm{a}_1 \Delta\Psi_1 + \sum_{k=1}^{N-1} \mathrm{a}_{k-1} \Delta\Psi_k - \sum_{k=0}^{N-1} \mathrm{a}_k \Delta\Psi_{k+1} \leq \mathrm{a}_1 \Delta\Psi_1.
\end{aligned}$$

The third inequality uses the relation $\mathrm{a}_k \leq \mathrm{a}_{k-1}$.[4] Taking full expectations and summing from $k = 0, \ldots, N-1$, we continue the above bound

$$\begin{aligned}
\mathbb{E}\left[\sum_{k=0}^{N-1} (\alpha_k - \alpha_k^2 \mathrm{lip}_1(h_\eta)) \Delta\Psi_k\right] &\leq \mathbb{E}\left[\sum_{k=0}^{N-1} (\alpha_k - \alpha_k^2 \mathrm{lip}_1(h_\eta)) \Delta\Psi_{k+1}\right] + (\alpha_1 - \mathrm{lip}_1(h_\eta)\alpha_1^2)\Delta\Psi_1 \\
&\leq \sum_{k=0}^{N-1} \left(\frac{3\alpha_k^2}{2} \frac{\mathrm{s}^2}{m_{k+1}} + \frac{3\alpha_k^2 \beta_k^2}{\eta^2} C_F\right) + (1 - \alpha_N \mathrm{lip}_1(h_\eta)) \frac{M}{2} \\
&\quad + \eta \mathrm{lip}_0(h) \sqrt{n} \sum_{k=0}^{N-1} (\alpha_k - \alpha_k^2 \mathrm{lip}_1(h_\eta)) \\
&\quad + M \sum_{k=0}^{N-1} (\alpha_k - \alpha_k^2 \mathrm{lip}_1(h_\eta)) \mathbb{E}\left(||b_{k+1}||_*\right) \\
&\quad + (\alpha_1 - \alpha_1^2 \mathrm{lip}_1(h_\eta)) \Delta\Psi_1 \\
&\leq \sum_{k=0}^{N-1} \left(\frac{3\alpha_k^2}{2} \frac{\mathrm{s}^2}{m_{k+1}} + \frac{3\alpha_k^2 \beta_k^2}{\eta^2} C_F\right) \\
&\quad + M \sum_{k=0}^{N-1} (\alpha_k - \alpha_k^2 \mathrm{lip}_1(h_\eta)) \frac{\beta_k}{\eta} \sqrt{C_F} + (1 - \alpha_N \mathrm{lip}_1(h_\eta)) \frac{M}{2} \\
&\quad + \eta \mathrm{lip}_0(h) \sqrt{n} \sum_{k=0}^{N-1} (\alpha_k - \alpha_k^2 \mathrm{lip}_1(h_\eta)) + (\alpha_1 - \alpha_1^2 \mathrm{lip}_1(h_\eta)) \Delta\Psi_1.
\end{aligned}$$

[4] This can be deduced as follows: Since $\alpha_k \leq \alpha_{k-1}$ one easily sees that

$$\mathrm{a}_k - \mathrm{a}_{k-1} = (\alpha_k - \alpha_{k-1}) - \mathrm{lip}_1(h_\eta)(\alpha_k^2 - \alpha_{k-1}^2) = (\alpha_k - \alpha_{k-1})\left(1 - \mathrm{lip}_1(h_\eta)(\alpha_k + \alpha_{k-1})\right) \leq 0.$$

Therefore, defining $D_k \triangleq 3\left(\mathrm{s}^2/2 + \frac{\beta_k^2 m_{k+1}}{\eta^2} C_F\right)$, and constructing an independent random variable $\kappa : \Omega \to \{0, \ldots, N-1\}$ with density function

$$p(k) = \mathbb{P}(\kappa = k) = \frac{\mathrm{a}_k}{\sum_{t=0}^{N-1} \mathrm{a}_t} \equiv \frac{\mathrm{a}_k}{A_N}, \quad A_N \triangleq \sum_{t=0}^{N-1} \mathrm{a}_t,$$

we obtain in a similar fashion as in the proof of Theorem 5.2

$$\mathbb{E}[\Delta\Psi_\kappa] \leq \frac{\sum_{k=0}^{N-1} \frac{\alpha_k^2}{m_{k+1}} D_k + \frac{M\sqrt{C_F}}{\eta} \sum_{k=0}^{N-1} \mathrm{a}_k \beta_k + M/2 + \alpha_1 \Delta\Psi_1}{A_N} + \eta\sqrt{n}\mathrm{lip}_0(h).$$

□

Similar to the analysis in the non-convex case, we can simplify the complexity bound of Theorem 6.1 via a judicious selection of parameters.

Corollary 6.2 *Under the same conditions as in Theorem 6.1, let be $\delta > 1$ and $\Delta\Psi_k \triangleq \Psi(y_k) - \Psi^{\mathrm{Opt}}$. Then*

$$\mathbb{E}[\Delta\Psi_\kappa] \leq \frac{\sum_{k=0}^{N-1} \alpha_k^2 (\frac{1}{m_{k+1}} + \beta_k^2)\bar{D} + \frac{M\sqrt{C_F}}{\eta} \sum_{k=0}^{N-1} \alpha_k \beta_k + M/2 + \alpha_1 \Delta\Psi_1}{\sum_{k=0}^{N-1} \alpha_k/2} + \eta\sqrt{n}\mathrm{lip}_0(h). \tag{6.5}$$

In particular, for fixed time horizon N, choosing step size $\alpha_k = \frac{\alpha_0}{\sqrt{N}}$, the constant mini-batch $m_{k+1} = m \geq 1$, and $\beta_k = \frac{a}{\sqrt{k+a}}$, $a \geq 1$, as well as $\eta = \frac{1}{\sqrt{N}}$. we obtain

$$\mathbb{E}[\Delta\Psi_\kappa] \leq \frac{\frac{\bar{D}\alpha_0}{m} + \frac{\bar{D}\alpha_0}{N} a^2 \log(N+a) + \frac{M\sqrt{C_F}}{\eta} \frac{2\alpha_0 a\sqrt{N+a}}{\sqrt{N}} + \frac{M}{2} + \alpha_1 \Delta\Psi_1}{\frac{\sqrt{N}}{2}} + \frac{\sqrt{n}\mathrm{lip}_0(h)}{\sqrt{N}}. \tag{6.6}$$

Proof First we note that

$$\sum_{k=0}^{N-1} \frac{\alpha_k^2}{m_{k+1}} D_k \leq \bar{D} \sum_{k=0}^{N-1} \alpha_m^2 (\frac{1}{m_{k+1}} + \beta_k^2),$$

where $\bar{D} \triangleq 3\max\{\mathrm{s}^2/2, \frac{C_F}{\eta^2}\}$; Second $\sum_{k=0}^{N-1} \mathrm{a}_k \beta_k \leq \sum_{k=0}^{N-1} \alpha_k \beta_k$. Moreover, by choosing the step size $\alpha_k \leq \frac{1}{2\mathrm{lip}_1(h_\eta)}$, we see that $\mathrm{a}_k = \alpha_k - \alpha_k^2 \mathrm{lip}_1(h_\eta) \geq \frac{\alpha_k}{2}$. Combining all these estimates, we arrive at (6.5). For fixed time horizon N, choose

step size $\alpha_k = \frac{\alpha_0}{\sqrt{N}}$, the constant mini-batch $m_{k+1} = m \geq 1$, and $\beta_k = \frac{a}{\sqrt{k+a}}$, $a \geq 1$, as well as $\eta = \frac{1}{\sqrt{N}}$. Substituting these numbers into expression (6.5), we immediately obtain (6.6). □

It is worth noting that the fixed step sizes $\alpha_k = \frac{\alpha_0}{\sqrt{N}}$ can be replaced by diminishing step sizes of the order $\frac{1}{\sqrt{k}}$.

7 Explicit complexity and relaxed stationarity

The previous results provided a finite-time complexity estimate in terms of the gradient mapping of the proximal gradient algorithm, involving the Gaussian smoothed objective. It is intuitive that a small proximal gradient in the smoothed regime should imply an approximate stationary point in the *original* optimization problem, when the smoothing parameter is sufficiently small. In this section we make this intuition precise and relate our complexity estimate from Theorem 5.2 to a complexity estimate with respect to a relaxed stationary point.

Fix $\eta > 0$ and define $\alpha_1 = \frac{2\theta}{\mathrm{lip}_1(h_\eta)}$. Define the auxiliary process $(\hat{y}_k)_{k\geq 1}$ by

$$\hat{y}_k \triangleq P_{\alpha_1}(y_k, \hat{V}_{k+1}) = \operatorname*{argmin}_u \{r_1(u) + \frac{1}{2\alpha_1}\left|\left|u - (y_k - \alpha_1 B^{-1}\hat{V}_{k+1})\right|\right|^2\}.$$

This point is uniquely characterized by the optimality condition

$$y_k - \alpha_1 B^{-1}\hat{V}_{k+1} \in \hat{y}_k + \alpha_1 B^{-1}\partial r_1(\hat{y}_k),$$

or equivalently

$$\hat{y}_k + \alpha_1 B^{-1}\mathcal{D}(y_k) \in \hat{y}_k + \alpha_1 B^{-1}\partial r_1(\hat{y}_k) \Leftrightarrow \mathcal{D}(y_k) \in \partial r_1(\hat{y}_k),$$

where

$$\mathcal{D}(y_k) \triangleq B\left(\frac{y_k - \hat{y}_k}{\alpha_1}\right) - \nabla h_\eta(\hat{y}_k) + (\nabla h_\eta(\hat{y}_k) - \nabla h_\eta(y_k)) + (\nabla h_\eta(y_k) - \hat{V}_{k+1}).$$

This yields

$$B\left(\frac{y_k - \hat{y}_k}{\alpha_1}\right) + (\nabla h_\eta(\hat{y}_k) - \nabla h_\eta(y_k)) - \Delta W_{k+1} \in \partial r_1(\hat{y}_k) + \nabla h_\eta(\hat{y}_k).$$

From now on we continue our developments with Assumption 7 in place. Choose $\varepsilon_1 > 0, \varepsilon_2 > 0$, and $\eta < \bar{\eta}$ (depending on $\varepsilon_1, \varepsilon_2$), as defined in Proposition 4.5, so that

$$B\left(\frac{y_k-\hat{y}_k}{\alpha_1}\right)+(\nabla h_\eta(\hat{y}_k)-\nabla h_\eta(y_k))-\Delta W_{k+1}\in\partial r_1(\hat{y}_k)+\partial_G^{\varepsilon_2}h_\eta(\hat{y}_k)+\frac{\varepsilon_1}{3}\mathbb{B}_{\mathcal{Y}}.$$

Therefore, using Lemma 4.3, we arrive at

$$\begin{aligned}\operatorname{dist}\left(0,\partial_G^{\varepsilon_2}h(\hat{y}_k)+\partial r_1(\hat{y}_k)\right)^2 &\leq \frac{6}{\alpha_1^2}\left\|y_k-\hat{y}_k\right\|^2+3\left\|\nabla h_\eta(\hat{y}_k)-\nabla h_\eta(y_k)\right\|_*^2\\ &\quad+3\left\|\Delta W_{k+1}\right\|_*^2+\frac{2\varepsilon_1^2}{3}\\ &\leq\left(\frac{6}{\alpha_1^2}+3\frac{n\operatorname{lip}_0(h)^2}{\eta^2}\right)\left\|y_k-\hat{y}_k\right\|^2+3\left\|\Delta W_{k+1}\right\|_*^2+\frac{2\varepsilon_1^2}{3}.\end{aligned}$$

Next, we relate the auxiliary process $(\hat{y}_k)_k$ to the stochastic sequence $(y_k)_k$ generated by Algorithm 1. To that end, observe that

$$\left\|y_k-\hat{y}_k\right\|\leq\left\|y_k-T_{\eta,\alpha_1}(y_k)\right\|+\left\|T_{\eta,\alpha_1}(y_k)-\hat{y}_k\right\|=\alpha_1\left\|\mathcal{G}_{\eta,\alpha_1}(y_k)\right\|+\alpha_1\left\|\Delta W_{k+1}\right\|_*.$$

Combining this estimate and Lemma 4.3 with the penultimate display, we arrive at

$$\begin{aligned}\operatorname{dist}\left(0,\partial_G^{\varepsilon_2}h(\hat{y}_k)+\partial r_1(\hat{y}_k)\right)^2 &\leq\left(12+\frac{6n\operatorname{lip}_0(h)^2}{\eta^2}\alpha_1^2\right)\left\|\mathcal{G}_{\eta,\alpha_1}(y_k)\right\|^2\\ &\quad+\left(15+\frac{6n\operatorname{lip}_0(h)^2}{\eta^2}\alpha_1^2\right)\left\|\Delta W_{k+1}\right\|_*^2+\frac{2\varepsilon_1^2}{3}\\ &\leq\left(12+24\beta^2\right)\left\|\mathcal{G}_{\eta,\alpha_1}(y_k)\right\|^2+\left(15+24\beta^2\right)\left\|\Delta W_{k+1}\right\|_*^2+\frac{2\varepsilon_1^2}{3}\\ &\leq 18\left\|\mathcal{G}_{\eta,\alpha_1}(y_k)\right\|^2+21\left\|\Delta W_{k+1}\right\|_*^2+\frac{2\varepsilon_1^2}{3}.\end{aligned}$$

Adopting a non-increasing step size regime in Algorithm 1, we can leverage the monotonicity result of the prox-gradient mapping with respect to the step size, described in Appendix C, so that for all $k\in\{0,1,\ldots,N-1\}$

$$\operatorname{dist}\left(0,\partial_G^{\varepsilon_2}h(\hat{y}_k)+\partial r_1(\hat{y}_k)\right)^2\leq 18\left\|\mathcal{G}_{\eta,\alpha_k}(y_k)\right\|^2+21\left\|\Delta W_{k+1}\right\|_*^2+\frac{2\varepsilon_1^2}{3}. \quad (7.1)$$

From these preparatory calculations, we can state the next relation between the complexity analysis in terms of the prox-gradient mapping (Corollary 5.3), and our definition of an (ε,δ)-stationary point (Definition 3.3).

Theorem 7.1 *Given positive parameters $(\varepsilon_1,\varepsilon_2)$ and let Assumption 7 together with all assumptions formulated in Corollary 5.4 be true. Pick $\eta\in(0,\bar{\eta}]$ so that the gradient estimate of Proposition 4.5 for the given pair $(\varepsilon_1,\varepsilon_2)$ applies. Let $(y_k)_{k=0}^N$ be the stochastic process generated by Algorithm 1 with gradient estimator (4.5) and Let $(\hat{y}_k)_{k=0}^N$ be the auxiliary process constructed recursively with*

$$\hat{y}_0 = y_0 and \hat{y}_k = P_{\alpha_1}(y_k, \hat{V}_{k+1}) \quad \forall k = 1, \dots, N-1. \tag{7.2}$$

If $\kappa : \Omega \to \{0, \dots, N-1\}$ is the random variable with law defined in Theorem 5.2, then for $N \geq 2$ chosen sufficiently large so that

$$\frac{36\mathrm{lip}_1(h_\eta)[\Psi_\eta(y_1) - \Psi_\eta^{\mathrm{Opt}}]}{\beta\sqrt{N}} + \frac{\frac{228\mathrm{s}^2}{\mathrm{a}}(1+\log(N))}{\sqrt{N}} \leq \frac{\varepsilon_1^2}{3}. \tag{7.3}$$

Then,

$$\mathbb{E}\left[\,\mathrm{dist}\,(0, \partial_G^{\varepsilon_2} h(\hat{y}_\kappa) + \partial r_1(\hat{y}_\kappa))^2\right] \leq \varepsilon_1^2,$$

i.e. the algorithm delivers an $(\varepsilon_1, \varepsilon_2)$-stationary point in the sense of Definition 3.3.

Proof Continuing from (7.1) and using (5.5), we readily deduce

$$\begin{aligned}
&\mathbb{E}\left[\,\mathrm{dist}\,(0, \partial_G^{\varepsilon_2} h(\hat{y}_\kappa) + \partial r_1(\hat{y}_\kappa))^2\right] \\
&\quad \leq \frac{2\varepsilon_1^2}{3} + 21\sum_{k=0}^{N-1} \frac{\frac{\mathrm{s}^2}{m_{k+1}}(2\alpha_k - \alpha_k^2 \mathrm{lip}_1(h_\eta))}{\sum_{t=0}^{N-1}(2\alpha_t - \alpha_t^2 \mathrm{lip}_1(h_\eta))} \\
&\quad + 18\left[\frac{4(\Psi_\eta(y_1) - \Psi_\eta^{\mathrm{Opt}})}{\sum_{t=0}^{N-1}(2\alpha_t - \alpha_t^2 \mathrm{lip}_1(h_\eta))} + \frac{\sum_{k=0}^{N-1} 2\frac{\mathrm{s}^2}{m_{k+1}}(4\alpha_k - \alpha_k^2 \mathrm{lip}_1(h_\eta))}{\sum_{t=0}^{N-1}(2\alpha_t - \alpha_t^2 \mathrm{lip}_1(h_\eta))}\right] \\
&\quad \leq \frac{2\varepsilon_1^2}{3} + \frac{72(\Psi_\eta(y_1) - \Psi_\eta^{\mathrm{Opt}})}{\sum_{t=1}^{N-1}(2\alpha_t - \alpha_t^2 \mathrm{lip}_1(h_\eta))} + \sum_{k=0}^{N-1} \frac{57\frac{\mathrm{s}^2}{m_{k+1}}(4\alpha_k - \alpha_k^2 \mathrm{lip}_1(h_\eta))}{\sum_{t=0}^{N-1}(2\alpha_t - \alpha_t^2 \mathrm{lip}_1(h_\eta))}.
\end{aligned}$$

Choose $\alpha_k = \frac{2\beta}{\mathrm{lip}_1(h_\eta)\sqrt{k}}$, so that $2\alpha_k - \alpha_k^2 \mathrm{lip}_1(h_\eta) \geq \alpha_k$ for all $k \geq 0$. Additionally, choosing $m_k = \mathrm{a}\sqrt{k}$, and following the computations performed in Corollary 5.3, we continue the previous display as

$$\mathbb{E}\left[\,\mathrm{dist}\,(0, \partial_G^{\varepsilon_2} h(\hat{y}_\kappa) + \partial r_1(\hat{y}_\kappa))^2\right] \leq \frac{36\mathrm{lip}_1(h_\eta)[\Psi_\eta(y_1) - \Psi_\eta^{\mathrm{Opt}}]}{\beta\sqrt{N}} + \frac{2\varepsilon_1^2}{3} + \frac{228\frac{\mathrm{s}^2}{\mathrm{a}}(1+\log(N))}{\sqrt{N}}.$$

Choosing N so large that (7.3) holds, the thesis follows. □

Remark 7.1 Since $\mathrm{lip}_1(h_\eta) = O(1/\eta)$ and Proposition 4.5 shows that $\eta = O(\varepsilon_2)$, the implied iteration complexity by Theorem 7.1 is on the order of maginitude of $N^{-1/2} \sim \frac{\varepsilon_1^2 \eta}{3}$, so that $N \sim \frac{9}{\varepsilon_1^4 \varepsilon_2^2}$. Choosing $\varepsilon \equiv \varepsilon_1 = \varepsilon_2$ therefore yields the leading order of ε^{-6} for the iteration complexity.

8 Numerical experiments

In our numerical experiments, we consider the bilevel learning approach to inverse problems and specify two case studies:

- Experiment 1: One-dimensional signal denoising.
- Experiment 2: Optimal experimental design.

Both experiments fall naturally within our general learning framework, as described in the next subsections.

8.1 One-dimensional signal denoising

In the first experiment, we consider a simple one-dimensional image denoising problem, inspired by [19]. The goal is to reconstruct a random one dimensional piecewise constant signal, observed at discretely sampled points and perturbed with Gaussian observational noise. Specifically, we represent the signal as a random vector $\xi_1 = (\xi_1(t_1), \ldots, \xi_1(t_n))$ with $n = 256$ sample points, corrupted by Gaussian white noise

$$\xi_2(t) = \xi_1(t) + \sigma Z(t) \quad t \in \{t_1, \ldots, t_n\}.$$

$(Z(t))_{t\in\{t_1,\ldots,t_n\}}$ are independent and identically distributed $\mathrm{N}(0,1)$ random variables, and $\sigma = \sqrt{0.001}$. For the data generating process, we choose the sampling times uniformly by $t_i = \frac{i}{n}$, $i = 1, \ldots, n$,

$$\xi_1(t_i, \omega) = 1_{[C(\omega),R(\omega)]}(t_i),$$

where C, R are two independent uniformly distributed random variables with $C \sim \mathrm{U}([\frac{1}{8}, \frac{1}{4}])$ and $R \sim \mathrm{U}([\frac{3}{8}, \frac{7}{8}])$. As a variational model for reconstructing the signal from the noisy observations, we consider the loss function

$$\min_{x\in\mathbb{R}^n} g(x, (\lambda, \tau, \nu), \xi_2) \triangleq \frac{1}{2} ||x - \xi_2||^2 + \frac{\lambda}{2} ||\mathrm{L}x||^2 + \tau \mathrm{TV}_\nu(x), \tag{8.1}$$

where $(\lambda, \tau, \nu) \in \mathbb{R}^3_+$ is the hyperparameter vector of our problem. $\mathrm{L} \in \mathbb{R}^{n\times n}$ is a symmetric positive definite regularization matrix, chosen as $\mathrm{L}^2 = 0.01^2\Delta$, with Δ being the discrete Laplace operator for the 1D-signal, and $\lambda > 0$ is the Tikhonov regularization parameter. In addition to the Tikhonov regularization, we consider a smoothed Total Variation regularization

$$\mathrm{TV}_\nu(x) \triangleq \sum_i \sqrt{|x_{i+1} - x_i|^2 + \nu^2} \qquad \nu > 0, \text{ a learned smoothing parameter.}$$

Bilevel learning problems with this particular structure have also been studied in [19], and our numerical implementation follows the setup described in there. In par-

ticular, it is reported in [19], that the lower level problem (8.1) is μ-strongly convex and ℓ-smooth with

$$\mu \geq \lambda \cdot \mathrm{e}_{\min}(\mathrm{L}^2) \quad \text{and} \quad \ell \leq 1 + \frac{\tau\partial}{\nu} + \lambda\|\mathrm{L}^2\|,$$

where $\mathrm{e}_{\min}(\mathrm{L}^2) > 0$ denotes the smallest eigenvalue of L and $\partial > 0$ is a constant arising due to the spatial discretization of the Total Variation. Hence, in order to approximately retrieve the reconstruction operator $x^*(y, \xi_2)$, we can implement a gradient descent scheme achieving a full control over the inexactness in the lower level solution, as defined in Definition 4.6. More precisely, when implementing gradient descent with constant step size $\frac{1}{\ell}$, we achieve accuracy β using the stopping criterion $\frac{\|\nabla_x g(x,y,\xi_2)\|^2}{\mu^2} \leq \beta$ after at most $\lceil \frac{\log(\beta)}{\log((q-1)/(q+1))} \rceil$ iterations, where $q = \frac{\ell}{\mu}$ is the condition number of the problem.

Implementation and validation For numerically solving the upper level problem, we've introduced the parametrization

$$y \in \mathbb{R}^3 \mapsto (\lambda, \tau, \nu) = (10^{y_1}, 10^{y_2}, 10^{y_3}) = (\lambda(y_1), \tau(y_2), \nu(y_3)), \tag{8.2}$$

with the additional restriction $y = (y_1, y_2, y_3) \in [-7, 7]^3$. This leads to the upper-level problem

$$\min_{y\in[-7,7]^3} \mathbb{E}[\underbrace{\|x^*(y,\xi_2) - \xi_1\|^2}_{=F(x^*(y,\xi_2),\xi_1)}] + \underbrace{10^{-6}\left(\frac{1 + \frac{\partial\cdot\tau(y_2)}{\nu(y_3)} + \lambda(y_1)\|\mathrm{L}^2\|}{\lambda(y_1)\mathrm{e}_{\min}(\mathrm{L}^2)}\right)^2}_{=r_2(y)}. \tag{8.3}$$

In every iteration k of our main scheme, we first solve the lower level problem (8.1) up to accuracy $\beta_k = \frac{\beta_0}{\sqrt{k+1}}$, $\beta_0 > 0$ using gradient descent. We then increase the batch size of the random gradient estimator (4.9) by $m_k = \sqrt{k} \cdot m_0$, $m_0 \in \mathbb{N}$. We adopt the step-size policy $\alpha_k = \frac{\alpha_0}{\sqrt{k+1}}$, $\alpha_0 > 0$. The smoothing parameter is fixed at level $\eta = 0.01$. In order to numerically confirm the convergence of the generated trajectory, we plot the summation over the random operator $\tilde{\mathcal{G}}^{\beta_k}_{\eta,k}$ scaled by the step-size policy, i.e. we demonstrate that $\Delta_k := \sum_{s=0}^{k} \alpha_s \tilde{\mathcal{G}}^{\beta_s}_{\eta,s}$ remains bounded. To verify the generalization properties of the method, we generate a validation data set independent of the data set applied in the application of Algorithm 1, defined as i.i.d. sample $(\xi_1^{i,\mathrm{val}}, \xi_2^{i,\mathrm{val}})_{i=1}^{m_{\mathrm{val}}}$, $m_{\mathrm{val}} \in \mathbb{N}$, of (ξ_1, ξ_2). We then plot the normalized empirical errors in the upper level

$$\mathrm{Err}_i(y) \triangleq \frac{\|x^*(y, \xi_2^{i,\mathrm{val}}) - \xi_1^{i,\mathrm{val}}\|}{\|\xi_1^{i,\mathrm{val}}\|} \quad \forall i \in \{1, \ldots, m_{\mathrm{val}}\}. \tag{8.4}$$

The lower level solution $x^*(y, \xi_2^{i,\mathrm{val}})$ is obtained via the gradient descent method with tolerance $\beta = 10^{-7}$. In this visualization, we compare the generalization error for different choices of y with the resulting learned parameters by Algorithm 1. We set $\alpha_0 = 1$, $\beta_0 = 0.01$ and $m_0 = 1$ as parameters in Algorithm `DFProxGrad`, and terminate after $N = 700$ iterations.

In Fig. 1a–c we plot the resulting regularization parameters generated by Algorithm 1, where we observe that all three parameters converge. This result can also be observed from Fig. 1d, where we demonstrate that the summation over the random operators $\tilde{\mathcal{G}}_{\eta,k}^{\beta_k}$ remains bounded. The resulting reconstruction of the signal using the learned regularization parameters for solving the lower level problem (8.1) is plotted in Fig. 2b. As comparison, in Fig. 2c–f, we plot the reconstructions of the signal using different regularization parameters chosen by hand. In all four cases, we have chosen a smoothing parameter $\nu = 10^{-3}$. The comparison of the reconstruction already conveys some evidence that our learned regularization parameter using Algorithm 1 outperforms the fixed regularization parameters. This empirical evidence is further demonstrated in Fig. 3 where we compare the generalization error (8.4) over validation data set independent of the training data set.

8.2 Image reconstruction based on the radon transform

In X-ray tomography, the forward operator is a discretization of the Radon transform [31], where data are collected at various angles $\theta \in [0, \pi)$. The unknown x represents a 2D image, and the measurements $d(\theta) = \mathrm{A}_\theta x$ for one angle represents the line integrals of that image along straight lines at angle θ. We discretize this line integral, so that A_θ is numerically represented as a $1 \times n_x$ row vector, acting on model parameters $x \in \mathbb{R}^{n_x}$. Collecting a large number of angles $\theta \in [0, \pi)$ leads to a well-posed inverse problem and generally yields a good reconstruction. For practical applications it is of interest to reduce the number of angles, dictating the use of additional regularization to fill in the missing information. To model such a situation, we let $n \in \mathbb{N}$ denote the maximal number of angles from which we can observe the 2D image. Let $\theta = (\theta_1, \ldots, \theta_d) \in [0, \pi)^d$ the vector of angles, and the *full forward model*

$$\mathrm{K} = \begin{pmatrix} \mathrm{A}_{\theta_1} \\ \vdots \\ \mathrm{A}_{\theta_d} \end{pmatrix} \in \mathbb{R}^{d \times n_x}, \quad x \in \mathbb{R}^{n_x} \mapsto \mathrm{K}x \in \mathbb{R}^d$$

mapping the image $x \in \mathbb{R}^{n_x}$ to data in $\mathbb{R}^d$. Let $\xi_1 \in L^\infty(\Omega, \mathbb{R}^{n_x})$ represent the random vector of model parameters and let $Z \in L^\infty(\Omega; \mathbb{R}^d)$ be an independent observational noise. Define

$$D(\omega) \triangleq \mathrm{K}\xi_1(\omega) + Z(\omega).$$

The random vector $D \in \mathbb{R}^d$ corresponds to d noise-contaminated perspectives of the image, representing the measurements available in a design using all d angles simul-

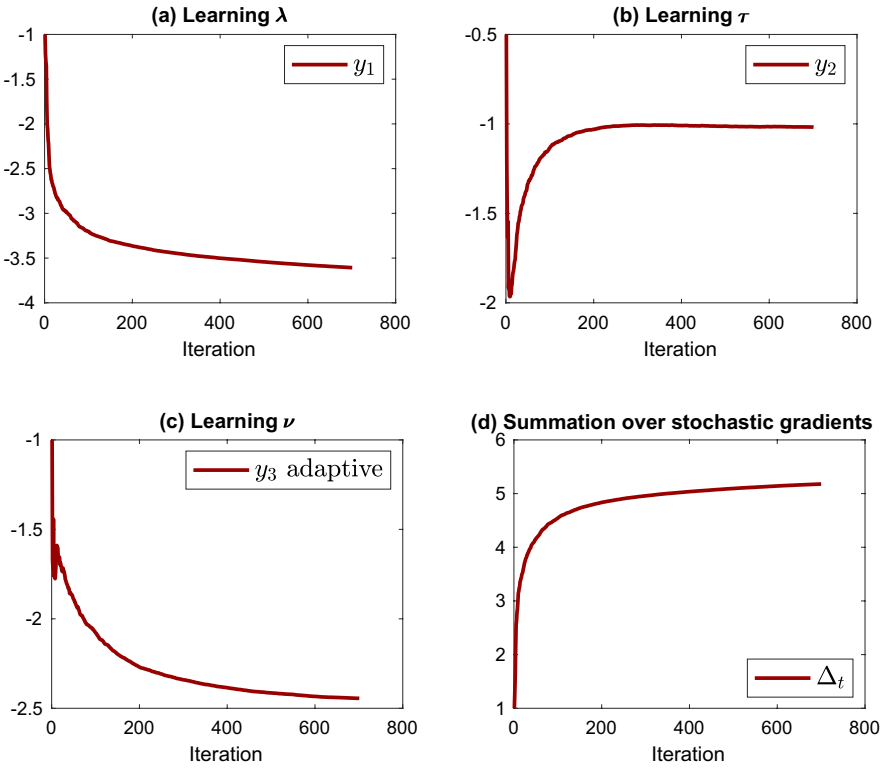


Fig. 1 **a**–**c** Learned parameters (y_1, y_2, y_3) resulting from Algorithm 1 and **d** summation over the random operators $\tilde{\mathcal{G}}_{\eta,t}^{\beta_t}$

taneously. To model an experimental design problem in which we can only take measurements from $1 \leq q \leq d$ chosen angles, we have to define a selection procedure from the coordinates of the vector D. Hence, given a fixed integer $1 \leq q \leq d$ (typically much smaller than d), we define $\mathcal{S}_{q,d} \triangleq \{J \subset \{1, 2, \ldots, d\} \mid |J| = q\}$. Each element $J \in \mathcal{S}_{q,d}$ gives rise to a q-tuple of observations

$$U_{J(\omega)} D(\omega) = (\mathrm{A}_{\theta_j} \xi_1(\omega) + Z_j(\omega))_{j \in J} \in \mathbb{R}^q, \tag{8.5}$$

where $U_J \in \mathbb{R}^{q \times d}$ is a matrix selecting the components contained in J out of the random vector $D = \mathrm{K}\xi_1 + Z$. Hence, using our statistical formulation of the experimental design problem, the random pair (ξ_1, ξ_2) consists of model parameters $\xi_1 \in \mathbb{R}^{n_x} \equiv \Xi_1$ and observations $\xi_2 = (J, U_J D) \in \Xi_2 \equiv \mathcal{S}_{q,d} \times \mathbb{R}^q$.

Our aim is to learn the best angles to be used for taking measurements via a specific variational formulation of the experimental design problem, given by

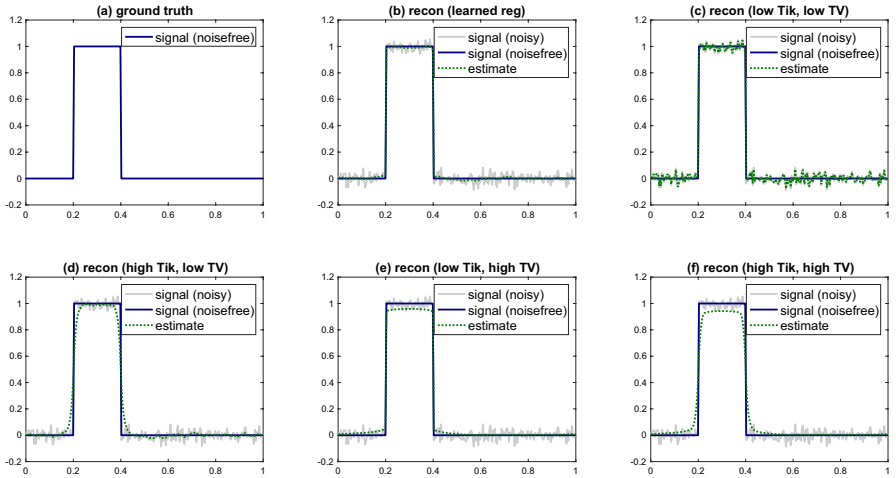
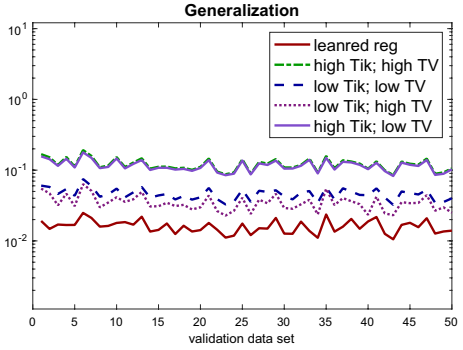


Fig. 2 **a** Ground truth signal and **b** reconstruction of the signal using the learned regularization parameters (λ, τ, ν) after $N = 300$ iterations. As comparison we show the reconstruction **c** using low Tikhonov regularization with $\lambda = 10^{-3}$ and low TV regularization with $\tau = 10^{-3}$, **d** using high Tikhonov regularization with $\lambda = 10^{-1}$ and low TV regularization with $\tau = 10^{-3}$, **e** using low Tikhonov regularization with $\lambda = 10^{-3}$ and high TV regularization with $\tau = 1$, and **f** using high Tikhonov regularization with $\lambda = 10^{-1}$ and high TV regularization with $\tau = 1$. In (**c**–**f**) we have fixed the smoothing parameter $\nu = 10^{-3}$

Fig. 3 Pointwise generalisation error in the upper level (8.4) over the validation data set $(\xi_1^{i,\mathrm{val}}, \xi_2^{i,\mathrm{val}})_{i=1}^{m_{\mathrm{val}}}$, $m_{\mathrm{val}} = 50$. We plot the errors for the different choices of regularisation parameters from Fig. 2



$$\min_{x\in\mathbb{R}^{n_x}_+} g(x,(w,\lambda,\tau,\nu),\xi_2) = \min_{x\in\mathbb{R}^{n_x}} \{g(x,(w,\lambda,\tau,\nu),\xi_2) + \delta_{\mathbb{R}^{n_x}_+}\}, \text{where}$$

$$g(x,(w,\lambda,\tau,\nu),\xi_2) \triangleq ||U_J \mathrm{K}x - U_J D||_2^2 + \frac{\lambda}{2}||x||^2 + \tau \mathrm{TV}_\nu(x) = ||\mathrm{K}x - D||^2_{U_J^\top U_J} + \frac{\lambda}{2}||x||^2 + \tau \mathrm{TV}_\nu(x),$$

$$\mathrm{TV}_\nu(x) \triangleq \sum_{i,j} \sqrt{|x_{i+1,j} - x_{i,j}|^2 + |x_{i,j+1} - x_{i,j}|^2 + \nu^2},$$

in which $x \in \mathbb{R}^{n_x}_+$ represents a two-dimensional discretized image of size $\sqrt{n_x} \times \sqrt{n_x}$ px. Note that we incorporate state constraints to the lower level solution forcing the images x to remain non-negative, which is implemented using the projected gradient method. The loss function depends on hyperparameters $y = (w, \lambda, \tau, \nu)$, which are trained via the minimization of the upper level loss function

$$\Psi(y) = \mathbb{E}[F(x^*(y,\xi_2),\xi_1)] = \mathbb{E}[\|x^*(y,\xi_2) - \xi_1\|^2]\,.$$

The interpretation of the hyperparameters λ, τ, ν is clear. The hyperparameter $w \in \mathbb{R}^d$ is used to parametrize the probability distribution over angles. To explain this transformation, define the set of probability mass functions $\mathcal{P} \triangleq \{p = (p_1, \dots, p_d) \in \mathbb{R}^d \mid p_i \geq 0, \sum_{i=1}^d p_i = 1\}$, in which $p_i \in [0,1]$ describes the probability of selecting the line integral A_{θ_i} in the experimental design. The weight vector $w = (w_1, \dots, w_d) \in \mathbb{R}^d$ determines the distribution over the random sample $J \in \mathcal{S}_{q,d}$ via the soft-max parametrization

$$p_i = \frac{\exp(w_i)}{\sum_{j=1}^d \exp(w_j)}, \quad i = 1, \dots, d.$$

Having constructed this probability distribution, we sample $J \in \mathcal{S}_{q,d}$ without replacement according to $(p_1, \dots, p_d)$.

Implementation and validation In our concrete experiment, we assume that we are allowed to pick $q = 6$ angles out of a pool of $d = 64$ possible angles $\theta_i = \frac{(i-1)\pi}{d}$. The goal is to reconstruct images of size 64×64 px given the noisy measurements generated by (8.5), where (Z_i) are independent and identical distributed according to $\mathrm{N}(0, 0.01^2)$. Our set of images consist of randomly generated triangles of varying size, rotation in the space and varying gray levels ranging from 0.5 to 1. The angles and direction of the triangles are kept fixed. In Fig. 4, we show i.i.d. realization of 16 different images.

In order to enhance the reconstruction accuracy we have implemented the OED problem of choosing the best possible policy over the set of all possible angles [49]. The regularization parameters (λ, τ, ν) are again parametrized as in (8.2). Including the vector of weights $(w_1, \dots, w_d)$, we abuse notation and identify the tuple of hyperparameters by the list

$$y = (y_1, \dots, y_d, 10^{y_{d+1}}, 10^{y_{d+2}}, 10^{y_{d+3}}),$$

so that our upper level problem (8.3) is a minimisation problem over a space $\mathcal{Y}$ of dimension $n = d + 3 = 67$.

In our numerical implementation, we have chosen the uniform distribution $(1/d, \dots, 1/d)$ (i.e. $w = 0 \in \mathbb{R}^d$, corresponding to a naive selection mechanism of angles) as initial condition. The same distribution is used as comparison in our validation over the validation data set. Algorithm 1 with inexact lower level solution is applied with $\alpha_0 = 0.2$, $\beta_0 = 0.1$ and $m_0 = 1$, and terminated after $N = 2000$ iterations. The generalization performance is illustrated in Fig. 6, where we have applied various configurations of regularization parameters together with the uniform policy. Among fixed choices of regularization parameters, we've also implemented the bilevel learning approach for selecting the regularization parameters (λ, τ, ν) with a fixed uniform policy. For validation, we use an i.i.d. sample $(\xi_1^{i,\mathrm{val}})_{i=1}^{m_{\mathrm{val}}}$, $m_{\mathrm{val}} \in \mathbb{N}$, of ξ_1 and plot the normalized empirical errors in the upper level

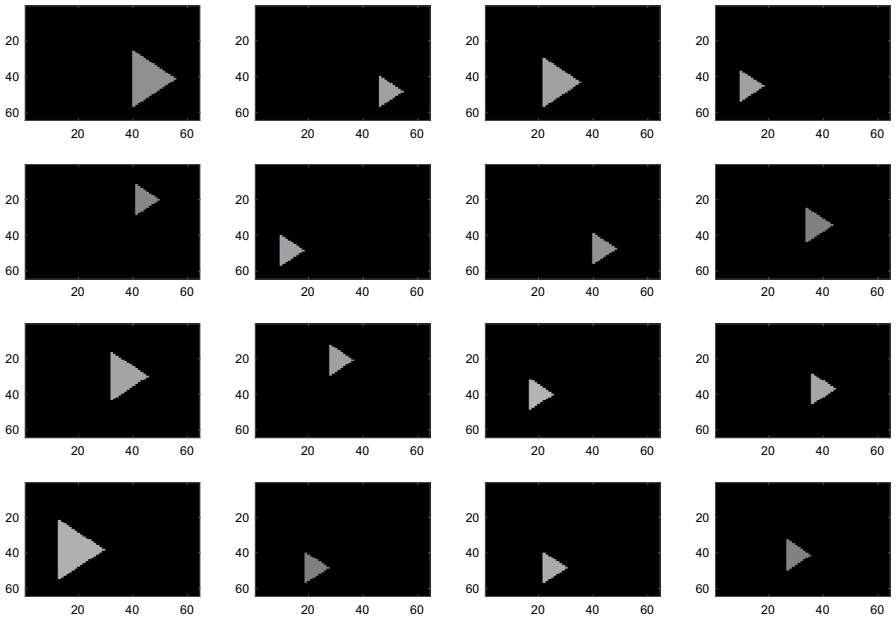


Fig. 4 Realisations of the random triangles in Example 8.2

$$\mathrm{Err}_i(y) := \frac{\|x^*(y, \xi_2^{i,\mathrm{val}}) - \xi_1^{i,\mathrm{val}}\|}{\|\xi_1^{i,\mathrm{val}}\|} \quad \forall i \in \{1, \ldots, m_{\mathrm{val}}\}. \tag{8.6}$$

where $\xi_2^{i,\mathrm{val}}$ is generated according to (8.5). The lower level solution $x^*(y, \xi_2^{i,\mathrm{val}})$ is again obtained by gradient descent with targeted accuracy $\beta = 10^{-7}$.

Overall, we observe a significant improvement by applying our learned policy. The resulting reconstructions for the different choices of regularization parameters and policies are shown in Fig. 5. These reconstructions further demonstrate the significant improvement through the proposed OED approach based on the stochastic bilevel optimization approach.

9 Conclusion

In this paper we've studied a zeroth-order gradient method for a particular class of stochastic bilevel programs which arise naturally in data-driven learning of inverse problems. Our complexity estimates adapt to smoothing and inexact solutions of the lower level problem. Our theoretical and numerical results display the favourable properties of our scheme. In future work, we plan to continue this line of research along the following directions:

- Higher-order numerical methods: The merit function employed in this paper is a stationary point. In non-convex optimization, an important question is whether our method is able to avoid saddle-points. For this, we plan to develop stochas-

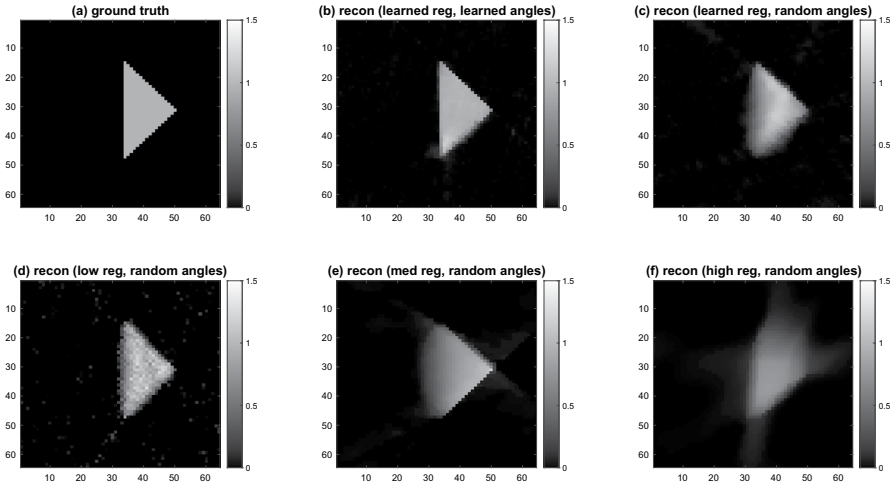


Fig. 5 **a** Ground truth image and **b** reconstruction of the image using the learned regularization parameters (λ, τ, ν) and the learned distribution p after $N = 2000$ iterations of `DFProxGrad`. As comparison we show the reconstruction of the image **c** using the learned regularization parameters (λ, τ, ν) after $N = 2000$ iterations and a fixed uniform policy, **d** using low regularization with $\lambda = 10^{-9}$, $\tau = 10^{-9}, \nu = 10^{-2}$ and uniform policy, **e** using medium regularization with $\lambda = 10^{-3}, \tau = 10^{-3}$, $\nu = 10^{-2}$ and uniform policy, **f** using high regularization with $\lambda = 10^{-2}, \tau = 10^{-2}, \nu = 10^{-2}$ and uniform policy

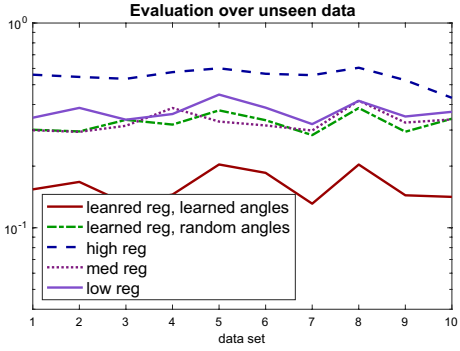


Fig. 6 Pointwise generalization error in the upper level $\mathrm{Err}_i(p, \lambda, \tau, \nu)$ (Eq. (8.6)) over the validation data set $(\xi_i^{\mathrm{val}})_{i=1}^{m_{\mathrm{val}}}$, $m_{\mathrm{val}} = 10$. We plot the errors for the different choices of regularization parameters from Fig. 5

tic Newton methods, employing derivative-free gradient estimation strategies, as done in this paper.

- Weakening the Lipschitz continuity assumptions of the hyperobjective. Interesting recent results in this direction are reported in [38].
- Construction of the random estimator: In this paper we adopt a Monte-Carlo approach to estimate the directional deriviative using iid Gaussian directions. It would be interesting to include more structure in this sampling approach. Quasi- or Multi-level Monte Carlo approaches would be interesting new stochastic simulation approaches to reduce the computational costs [27].

Properties of the Gaussian smoothing

Let $\mathcal{E}$ be a finite-dimensional real vector space, and define $M_p \triangleq \mathbb{E}[||U||^p]$.

Lemma A.1 *([42], Lemma 1) We have $M_0 = 1, M_2 = n$ and for $p \in [0, 2]$,*

$$M_p \leq n^{p/2}. \tag{A.1}$$

If $p \geq 2$, then

$$n^{p/2} \leq M_p \leq (p+n)^{p/2}. \tag{A.2}$$

Proof of Lemma 4.1 For all $y_1, y_2 \in \mathcal{Y}$, we have

$$\begin{aligned} |h_\eta(y_1) - h_\eta(y_2)| &= |\mathbb{E}_{\mathbb{P}_1}[h(y_1 + \eta U)] - \mathbb{E}_{\mathbb{P}_1}[h(y_2 + \eta U)]| \\ &= |\mathbb{E}_{\mathbb{P}_1}[h(y_1 + \eta U) - h(y_2 + \eta U)]| \\ &\leq \mathbb{E}_{\mathbb{P}_1}\left[|h(y_1 + \eta U) - h(y_2 + \eta U)|\right] \\ &\leq \mathrm{lip}_0(h)\,||y_1 - y_2||\,. \end{aligned}$$

□

Proof of Lemma 4.2 For any $y \in \mathcal{Y}$ we have

$$|h_\eta(y) - h(y)| \leq \mathbb{E}_{\mathbb{P}_1}\left[|h(y + \eta U) - h(y)|\right] \leq \eta \mathrm{lip}_0(h) \mathbb{E}_{\mathbb{P}_1}\left[||U||\right] = \eta \mathrm{lip}_0(h)\sqrt{n}.$$

□

Proof of Lemma 4.3 Let Υ denote the normalizing constant of the Gaussian density $\pi_\eta(\bullet|y)$. Using the formula (4.1), for any $y \in \mathcal{Y}$, we can directly differentiate under the integral to obtain

$$\begin{aligned} \nabla h_\eta(y) &= \frac{1}{\Upsilon \eta^n} \int_{\mathcal{Y}} h(z) \exp\left(-\frac{1}{2\eta^2}\,||z-y||^2\right) \frac{B(z-y)}{\eta^2}\,\mathrm{d}z \\ &= \frac{1}{\Upsilon} \int_{\mathcal{Y}} \frac{1}{\eta} h(y + \eta u) \exp\left(-\frac{1}{2}\,||u||^2\right) Bu\,\mathrm{d}u \\ &= \mathbb{E}_{\mathbb{P}_1}\left[\frac{h(y+\eta U) - h(y)}{\eta} BU\right] \\ &= \mathbb{E}_{\mathbb{P}_1}\left[\frac{h(y+\eta U)}{\eta} BU\right]. \end{aligned} \tag{A.3}$$

Now let $y_1, y_2 \in \mathcal{Y}$ so that

$$
\begin{aligned}
||\nabla h_\eta(y_1) - \nabla h_\eta(y_2)||_* &\leq \mathbb{E}_{\mathbb{P}_1}\left[\left|\frac{h(y_1+\eta U)-h(y_2+\eta U)}{\eta}\right| ||BU||_*\right] \\
&\leq \mathrm{lip}_0(h)\frac{||y_1-y_2||}{\eta}\mathbb{E}_{\mathbb{P}_1}\left[||U||\right] \\
&\leq \mathrm{lip}_0(h)\frac{||y_1-y_2||}{\eta}\sqrt{n}
\end{aligned}
$$

where the last inequality uses [42, Lemma1]. To obtain the bound on the gradient norm, we continue from the first relation, showing that

$$
\begin{aligned}
||\nabla h_\eta(y)||_*^2 &\leq \mathbb{E}_{\mathbb{P}_1}\left[\left|\frac{h(y+\eta U)-h(y)}{\eta}\right|^2 ||BU||_*^2\right] \\
&\leq \mathrm{lip}_0(h)^2\mathbb{E}_{\mathbb{P}_1}\left[||U||^2\cdot||BU||_*^2\right] \\
&= \mathrm{lip}_0(h)^2\mathbb{E}_{\mathbb{P}_1}\left[||U||^4\right] \leq \mathrm{lip}_0(h)^2(4+n)^2.
\end{aligned}
$$

The last equality uses again [42, Lemma1]. □

Technical proofs

Proof of Lemma 4.7

Given $y \in \mathcal{Y}$, we use the law of iterated expectations to compute

$$
\begin{aligned}
\mathbb{E}_{\mathbb{P}}\left[\hat{V}_{\eta,m}(y)\right] &= \mathbb{E}_{\mathbb{P}}\left[\frac{1}{m}\sum_{i=1}^m \hat{\nabla}_{(U^i,\eta)}H(y,\xi^i)\right] \\
&= \mathbb{E}_{\mathbb{P}}\left[\frac{1}{m}\sum_{i=1}^m \mathbb{E}_{\mathbb{P}}\left(\hat{\nabla}_{(U^i,\eta)}H(y,\xi^i)|\sigma(U^i)\right)\right] \\
&= \frac{1}{m}\sum_{i=1}^m \mathbb{E}_{\mathbb{P}}\left[\frac{h(y+\eta U^i)-h(y)}{\eta}BU^i\right] \\
&= \nabla h_\eta(y)
\end{aligned}
$$

where the last equality uses eq. (4.2). For the second bound, observe that

$$\mathbb{E}_{\mathbb{P}}\left[\left|\left|\hat{V}_{\eta,m}(y)\right|\right|_*^2\right] = \mathbb{E}_{\mathbb{P}}\left[\left|\left|\frac{1}{m}\sum_{i=1}^{m}\left(\hat{\nabla}_{(U^i,\eta)}H(y,\xi^i) - \nabla h_\eta(y)\right) + \nabla h_\eta(y)\right|\right|_*^2\right]$$
$$= \frac{1}{m^2}\mathbb{E}_{\mathbb{P}}\left[\left|\left|\sum_{i=1}^{m}\left(\hat{\nabla}_{(U^i,\eta)}H(y,\xi^i) - \nabla h_\eta(y)\right)\right|\right|_*^2\right] + ||\nabla h_\eta(y)||_*^2.$$

Define the centered random variable $X_i \triangleq \hat{\nabla}_{(U^i,\eta)}H(y,\xi^i) - \nabla h_\eta(y)$ for $1 \le i \le m$, to obtain an i.i.d collection of zero-mean random variables in $\mathcal{Y}^*$. Therefore, we can continue from the last line of the previous display by noting that

$$\mathbb{E}_{\mathbb{P}}\left[\left|\left|\sum_{i=1}^{m}X_i\right|\right|_*^2\right] = \mathbb{E}_{\mathbb{P}}\left[\langle B^{-1}\sum_{i=1}^{m}X_i, \sum_{i=1}^{m}X_i\rangle\right]$$
$$= \sum_{i,j}\mathbb{E}_{\mathbb{P}}\left[\langle B^{-1}X_i, X_j\rangle\right] = \sum_{i=1}^{m}\mathbb{E}_{\mathbb{P}}\left[\langle B^{-1}X_i, X_i\rangle\right]$$
$$= \sum_{i=1}^{m}\mathbb{E}_{\mathbb{P}}[||X_i||_*^2].$$

Since $\mathbb{E}_{\mathbb{P}}[||X_i||_*^2] = \mathbb{E}_{\mathbb{P}}\left[\left|\left|\hat{\nabla}_{(U,\eta)}H(y,\xi)\right|\right|_*^2\right] - ||\nabla h_\eta(y)||_*^2$, it follows

$$\mathbb{E}_{\mathbb{P}}\left[\left|\left|\hat{V}_{\eta,m}(y)\right|\right|_*^2\right] \le \frac{1}{m}\mathbb{E}_{\mathbb{P}}\left[\left|\left|\hat{\nabla}_{(U,\eta)}H(y,\xi)\right|\right|_*^2\right] + (1-\frac{1}{m})\|\nabla h_\eta(y)\|_*^2.$$

$$\mathbb{E}_{\mathbb{P}}\left[\left|\left|\hat{\nabla}_{(U,\eta)}H(y,\xi)\right|\right|_*^2\right] = \mathbb{E}_{\mathbb{P}}\left[\left|\left|\frac{H(y+\eta U,\xi)-H(y,\xi)}{\eta}BU\right|\right|_*^2\right]$$
$$= \mathbb{E}_{\mathbb{P}}\left[\left|\frac{H(y+\eta U,\xi)-H(y,\xi)}{\eta}\right|^2 \cdot ||BU||_*^2\right]$$
$$\le \mathbb{E}_{\mathbb{P}}\left[\mathrm{lip}_0(H(\cdot,\xi))^2\,||U||^4\right] \overset{\text{Lemma } A.1}{\le} |\mathrm{lip}_0(H(\cdot,\xi))|_2^2\,(4+n)^2.$$

Proof of Lemma 4.9

For arbitrary $y \in \mathcal{Y}$ we compute

$$
\begin{aligned}
\mathbb{E}_{\mathbb{P}}\left[\hat{V}^{\beta}_{\eta,m}(y)\right] &= \mathbb{E}_{\mathbb{P}}\left[\frac{1}{m}\sum_{i=1}^{m}\hat{\nabla}_{(U^i,\eta)}H^{\beta}(y,\xi^i)\right] \\
&= \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{\mathbb{P}}\left[\hat{\nabla}_{(U^i,\eta)}H^{\beta}(y,\xi^i)\right] \\
&= \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{\mathbb{P}}\left[\frac{F(x^{\beta}(y+\eta U^i,\xi_2^i),\xi_1^i)-F(x^{\beta}(y,\xi_2^i),\xi_1^i)}{\eta}BU^i\right] \\
&= \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{\mathbb{P}}\left[\frac{F(x^{*}(y+\eta U^i),\xi_2^i),\xi_1^i)-F(x^{*}(y,\xi_2^i),\xi_1^i)}{\eta}BU^i\right] \\
&\quad + \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{\mathbb{P}}\left[\frac{F(x^{\beta}(y+\eta U^i),\xi_2^i),\xi_1^i)-F(x^{*}(y+\eta U^i,\xi_2^i),\xi_1^i)}{\eta}BU^i\right] \\
&\quad - \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{\mathbb{P}}\left[\frac{F(x^{\beta}(y,\xi_2^i),\xi_1^i)-F(x^{*}(y,\xi_2^i),\xi_1^i)}{\eta}BU^i\right].
\end{aligned}
$$

From Lemma 4.7, we deduce that

$$
\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{\mathbb{P}}\left[\frac{F(x^{*}(y+\eta U^i,\xi_2^i),\xi_1^i)-F(x^{*}(y,\xi_2^i),\xi_1^i)}{\eta}BU^i\right] = \nabla h_{\eta}(y),
$$

and by mutual independence of U^i from $\xi^i = (\xi_1^i, \xi_2^i)$

$$
\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{\mathbb{P}}\left[\frac{F(x^{\beta}(y,\xi_2^i),\xi_1^i)-F(x^{*}(y,\xi_2^i),\xi_1^i)}{\eta}BU^i\right] = 0.
$$

For the second assertion, we apply Lipschitz continuity of F (Assumption 3), the iid assumption on the random pair (U^i, ξ^i), and Hölder's inequality to obtain

$$
\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m}\|\mathbb{E}_{\mathbb{P}}\Big[\frac{F(x^{\beta}(y+\eta U^i,\xi_2^i),\xi_1^i)-F(x^*(y+\eta U^i),\xi_1^i)}{\eta}BU^i\Big]\|_*\\
&\leq \mathbb{E}_{\mathbb{P}}\Big[\|\frac{F(x^{\beta}(y+\eta U,\xi_2),\xi_1)-F(x^*(y+\eta U,\xi_2),\xi_1)}{\eta}BU\|_*\Big]\\
&=\frac{1}{\eta}\mathbb{E}_{\mathbb{P}}\Big[\left|F(x^{\beta}(y+\eta U,\xi_2),\xi_1)-F(x^*(y+\eta U,\xi_2),\xi_1)\right|\|BU\|_*\Big]\\
&\leq \frac{1}{\eta}\mathbb{E}_{\mathbb{P}}\Big[\mathrm{lip}_0(F(\cdot,\xi_1))\left|\left|x^{\beta}(y+\eta U,\xi_2)-x^*(y+\eta U,\xi_2)\right|\right|_{\mathcal{X}}\cdot\|BU\|_*\Big]\\
&\leq \frac{1}{\eta}\mathbb{E}_{\mathbb{P}}[\mathrm{lip}_0(F(\cdot,\xi_1))]\cdot\mathbb{E}_{\mathbb{P}}\Big[\left|\left|x^{\beta}(y+\eta U,\xi_2)-x^*(y+\eta U,\xi_2)\right|\right|_{\mathcal{X}}\cdot\|BU\|_*\Big]\\
&\leq \frac{|\mathrm{lip}_0(F(\cdot,\xi_1)|_1}{\eta}\mathbb{E}_{\mathbb{P}}\Big[\left|\left|x^{\beta}(y+\eta U,\xi_2)-x^*(y+\eta U,\xi_2)\right|\right|_{\mathcal{X}}^{p}\Big]^{\frac{1}{p}}\cdot\mathbb{E}_{\mathbb{P}}\Big[||U||^{\frac{p-1}{p}}\Big]^{\frac{p}{p-1}}\\
&\leq \frac{\sqrt{n}\,|\mathrm{lip}_0(F(\cdot,\xi_1)|_1}{\eta}\mathbb{E}_{\mathbb{P}}\Big[\left|\left|x^{\beta}(y+\eta U,\xi_2)-x^*(y+\eta U,\xi_2)\right|\right|_{\mathcal{X}}^{p}\Big]^{\frac{1}{p}}.
\end{aligned}
$$

Proof of Lemma 4.10

We have

$$
\begin{aligned}
||a_{k+1}||_* &= \frac{1}{m_{k+1}}\left|\left|\sum_{i=1}^{m_{k+1}}\left[\frac{F(x^{\beta_k}(y_k,\xi_{2,k+1}^i),\xi_{1,k+1}^i)-F(x^*(y_k,\xi_{2,k+1}^i),\xi_{1,k+1}^i)}{\eta}\right]BU_{k+1}^i\right|\right|_*\\
&\leq \frac{1}{m_{k+1}}\sum_{i=1}^{m_{k+1}}\left|\frac{F(x^{\beta_k}(y_k,\xi_{2,k+1}^i),\xi_{1,k+1}^i)-F(x^*(y_k,\xi_{2,k+1}^i),\xi_{1,k+1}^i)}{\eta}\right|\left|\left|U_{k+1}^i\right|\right|\\
&\leq \frac{1}{m_{k+1}}\sum_{i=1}^{m_{k+1}}\frac{\mathrm{lip}_0(F(\cdot,\xi_{1,k+1}^i))}{\eta}\left|\left|x^{\beta_k}(y_k,\xi_{2,k+1}^i)-x^*(y_k,\xi_{2,k+1}^i)\right|\right|_{\mathcal{X}}\cdot\left|\left|U_{k+1}^i\right|\right|/,.
\end{aligned}
$$

Hence, by Jensen's inequality and the tower property and the independence of the triple $(\xi_{1,k+1}^i,\xi_{2,k+1}^i,U_{k+1}^i)$, we obtain

$$
\begin{aligned}
\mathbb{E}[||a_{k+1}||_*^2|\mathcal{F}_k] &\leq \frac{1}{\eta^2 m_{k+1}^2}\mathbb{E}\left[\left(\sum_{i=1}^{m_{k+1}}\mathrm{lip}_0(F(\cdot,\xi_{1,k+1}^i))\left|\left|x^{\beta_k}(y_k,\xi_{2,k+1}^i)-x^*(y_k,\xi_{2,k+1}^i)\right|\right|_{\mathcal{X}}\cdot\left|\left|U_{k+1}^i\right|\right|\right)^2|\mathcal{F}_k\right]\\
&\leq \frac{1}{\eta^2 m_{k+1}}\sum_{i=1}^{m_{k+1}}\mathbb{E}\left[\mathrm{lip}_0(F(\cdot,\xi_{1,k+1}^i))^2\left|\left|x^{\beta_k}(y_k,\xi_{2,k+1}^i)-x^*(y_k,\xi_{2,k+1}^i)\right|\right|_{\mathcal{X}}^2\cdot\left|\left|U_{k+1}^i\right|\right|^2|\mathcal{F}_k\right]\\
&= \frac{n}{\eta^2 m_{k+1}}\sum_{i=1}^{m_{k+1}}\mathbb{E}\left[\mathrm{lip}_0(F(\cdot,\xi_{1,k+1}^i))^2|\mathcal{F}_k\right]\cdot\mathbb{E}\left[\left|\left|x^{\beta_k}(y_k,\xi_{2,k+1}^i)-x^*(y_k,\xi_{2,k+1}^i)\right|\right|_{\mathcal{X}}^2|\mathcal{F}_k\right]\\
&\leq \frac{n\,|\mathrm{lip}_0(F(\cdot,\xi_1)|_2^2}{\eta^2 m_{k+1}}\sum_{i=1}^{m_{k+1}}\mathbb{E}\left[\left|\left|x^{\beta_k}(y_k,\xi_{2,k+1}^i)-x^*(y_k,\xi_{2,k+1}^i)\right|\right|_{\mathcal{X}}^p|\mathcal{F}_k\right]^{2/p}\\
&\leq \frac{n\,|\mathrm{lip}_0(F(\cdot,\xi_1)|_2^2}{\eta^2}\beta_k^2 \triangleq C_F\frac{\beta_k^2}{\eta^2},
\end{aligned}
$$

where $p\geq 2$, is the exponent from Definition 4.6. We can bound the $L^2(\mathbb{P})$-norm for the bias term b_{k+1} in a similar way. First, observe that

$$||b_{k+1}||_* \leq \frac{1}{m_{k+1}} \sum_{i=1}^{m_{k+1}} \left| \frac{F(x^{\beta_k}(y_k + \eta U^i_{k+1}, \xi^i_{2,k+1}), \xi^i_{1,k+1}) - F(x^*(y_k + \eta U^i_{k+1}, \xi^i_{2,k+1}), \xi^i_{1,k+1})}{\eta} \right| ||U^i_{k+1}||$$
$$\leq \frac{1}{\eta m_{k+1}} \sum_{i=1}^{m_{k+1}} \mathrm{lip}_0(F(\cdot, \xi^i_{1,k+1})) \left|\left|x^{\beta_k}(y_k + \eta U^i_{k+1}, \xi^i_{2,k+1}) - x^*(y_k + \eta U^i_{k+1}, \xi^i_{2,k+1})\right|\right|_{\mathcal{X}} \cdot ||U^i_{k+1}||.$$

Using Jensen's inequality and Hölder's inequality as in the previous estimate, we see for $s \geq 1$,

$$\mathbb{E}[||b_{k+1}||_*^2 |\mathcal{F}_k]$$
$$\leq \frac{1}{\eta^2 m_{k+1}} \sum_{i=1}^{m_{k+1}} \mathbb{E}\left[\mathrm{lip}_0(F(\cdot, \xi^i_{1,k+1}))^2 \left|\left|x^{\beta_k}(y_k + \eta U^i_{k+1}, \xi^i_{2,k+1}) - x^*(y_k + \eta U^i_{k+1}, \xi^i_{2,k+1})\right|\right|^2_{\mathcal{X}} \cdot ||U^i_{k+1}||^2 |\mathcal{F}_k\right]$$
$$= \frac{|\mathrm{lip}_0(F(\cdot, \xi^i_1))|_2^2}{\eta^2 m_{k+1}} \sum_{i=1}^{m_{k+1}} \mathbb{E}\left[\left|\left|x^{\beta_k}(y_k + \eta U^i_{k+1}, \xi^i_{2,k+1}) - x^*(y_k + \eta U^i_{k+1}, \xi^i_{2,k+1})\right|\right|^2_{\mathcal{X}} \cdot ||U^i_{k+1}||^2 |\mathcal{F}_k\right]$$
$$\leq \frac{|\mathrm{lip}_0(F(\cdot, \xi^i_1))|_2^2}{\eta^2 m_{k+1}} \sum_{i=1}^{m_{k+1}} \mathbb{E}\left[\left|\left|x^{\beta_k}(y_k + \eta U^i_{k+1}, \xi^i_{2,k+1}) - x^*(y_k + \eta U^i_{k+1}, \xi^i_{2,k+1})\right|\right|^{2s}_{\mathcal{X}} |\mathcal{F}_k\right]^{1/s} \cdot \mathbb{E}[||U^i_{k+1}||^{2r}]^{1/r}$$

for $\frac{1}{s} + \frac{1}{r} = 1$. Choosing $2\,s = p$, we obtain

$$\mathbb{E}[||b_{k+1}||_*^2 |\mathcal{F}_k]$$
$$\leq \frac{|\mathrm{lip}_0(F(\cdot, \xi^i_1))|_2^2}{\eta^2 m_{k+1}} \sum_{i=1}^{m_{k+1}} \mathbb{E}\left[\left|\left|x^{\beta_k}(y_k + \eta U^i_{k+1}, \xi^i_{2,k+1}) - x^*(y_k + \eta U^i_{k+1}, \xi^i_{2,k+1})\right|\right|^p_{\mathcal{X}} |\mathcal{F}_k\right]^{2/p} \cdot \mathbb{E}[\left|\left|U^i_{k+1}\right|\right|^{\frac{2p}{p-2}}]^{\frac{p-2}{p}}$$
$$\leq \frac{n\,|\mathrm{lip}_0(F(\cdot, \xi^i_1))|_2^2}{\eta^2} \beta_k^2 = C_F \frac{\beta_k^2}{\eta^2}.$$

Proof of Lemma 5.1

The optimality condition for the iterate y_{k+1} gives

$$B\left(\frac{y_k - y_{k+1}}{\alpha_k}\right) \in \hat{V}_{k+1} + \partial r_1(y_{k+1}).$$

This means that there exists $\rho_{k+1} \in \partial r_1(y_{k+1})$ satisfying

$$\hat{V}_{k+1} + \rho_{k+1} = B\left(\frac{y_k - y_{k+1}}{\alpha_k}\right).$$

Since $r_1(\cdot)$ is convex, the convex subgradient inequality gives for all $u \in \mathcal{Y}$,

$$\begin{aligned} r_1(u) &\geq r_1(y_{k+1}) - \langle \hat{V}_{k+1} - B\left(\frac{y_k - y_{k+1}}{\alpha_k}\right), u - y_{k+1} \rangle \\ &= r_1(y_{k+1}) - \langle \hat{V}_{k+1}, u - y_{k+1} \rangle + \frac{1}{\alpha_k} \langle B(y_{k+1} - y_k), y_{k+1} - u \rangle. \end{aligned} \tag{B.1}$$

Set $u = y_k$ to obtain

$$
\begin{aligned}
r_1(y_k) &\geq r_1(y_{k+1}) - \langle \hat{V}_{k+1}, y_k - y_{k+1} \rangle + \frac{1}{\alpha_k} \left|\left| y_{k+1} - y_k \right|\right|^2 \\
&= r_1(y_{k+1}) - \alpha_k \langle \hat{V}_{k+1}, \tilde{\mathcal{G}}_{\eta,\alpha_k}(y_k) \rangle + \alpha_k \left|\left| \tilde{\mathcal{G}}_{\eta,\alpha_k}(y_k) \right|\right|^2 .
\end{aligned}
$$

The descent property (2.1) for $h_\eta \in \mathrm{C}^{1,1}(\mathcal{Y})$ gives

$$
\begin{aligned}
h_\eta(y_{k+1}) &\leq h_\eta(y_k) + \langle \nabla h_\eta(y_k), y_{k+1} - y_k \rangle + \frac{\mathrm{lip}_1(h_\eta)}{2} \left|\left| y_{k+1} - y_k \right|\right|^2 \\
&= h_\eta(y_k) - \alpha_k \langle \nabla h_\eta(y_k), \tilde{\mathcal{G}}_{\eta,\alpha_k}(y_k) \rangle + \frac{\alpha_k^2 \mathrm{lip}_1(h_\eta)}{2} \left|\left| \tilde{\mathcal{G}}_{\eta,\alpha_k}(y_k) \right|\right|^2 \\
&= h_\eta(y_k) - \alpha_k \langle \hat{V}_{k+1}, \tilde{\mathcal{G}}_{\eta,\alpha_k}(y_k) \rangle + \frac{\alpha_k^2 \mathrm{lip}_1(h_\eta)}{2} \left|\left| \tilde{\mathcal{G}}_{\eta,\alpha_k}(y_k) \right|\right|^2 \\
&\quad + \alpha_k \langle \hat{V}_{k+1} - \nabla h_\eta(y_k), \tilde{\mathcal{G}}_{\eta,\alpha_k}(y_k) \rangle \\
&\leq h_\eta(y_k) - \alpha_k \left|\left| \tilde{\mathcal{G}}_{\eta,\alpha_k}(y_k) \right|\right|^2 - (r_1(y_{k+1}) - r_1(y_k)) + \alpha_k \langle \Delta W_{k+1}, \tilde{\mathcal{G}}_{\eta,\alpha_k}(y_k) \rangle \\
&\quad + \frac{\alpha_k^2 \mathrm{lip}_1(h_\eta)}{2} \left|\left| \tilde{\mathcal{G}}_{\eta,\alpha_k}(y_k) \right|\right|^2 ,
\end{aligned}
$$

where we have used (B.1) in the last inequality. Rearranging the last inequality yields

$$
\begin{aligned}
\Psi_\eta(y_{k+1}) - \Psi_\eta(y_k) &\leq -\alpha_k \left|\left| \tilde{\mathcal{G}}_{\eta,\alpha_k}(y_k) \right|\right|^2 \left(1 - \frac{\alpha_k \mathrm{lip}_1(h_\eta)}{2}\right) + \alpha_k \langle \Delta W_{k+1}, \mathcal{G}_{\eta,\alpha_k}(y_k) \rangle \\
&\quad + \alpha_k \langle \Delta W_{k+1}, \tilde{\mathcal{G}}_{\eta,\alpha_k}(y_k) - \mathcal{G}_{\eta,\alpha_k}(y_k) \rangle \\
&\leq -\alpha_k \left|\left| \tilde{\mathcal{G}}_{\eta,\alpha_k}(y_k) \right|\right|^2 \left(1 - \frac{\alpha_k \mathrm{lip}_1(h_\eta)}{2}\right) + \alpha_k \langle \Delta W_{k+1}, \mathcal{G}_{\eta,\alpha_k}(y_k) \rangle \\
&\quad + \alpha_k \left|\left| \Delta W_{k+1} \right|\right|_* \cdot \left|\left| \tilde{\mathcal{G}}_{\eta,\alpha_k}(y_k) - \mathcal{G}_{\eta,\alpha_k}(y_k) \right|\right| ,
\end{aligned}
$$

where the Cauchy-Schwarz inequality in the last inequality is employed. Using the non-expansiveness of the prox-operator, we obtain

$$
\left|\left| \tilde{\mathcal{G}}_{\eta,\alpha_k}(y_k) - \mathcal{G}_{\eta,\alpha_k}(y_k) \right|\right| \leq \left|\left| B^{-1}(\nabla h_\eta(y_k) - V_{k+1}) \right|\right| = \left|\left| \nabla h_\eta(y_k) - V_{k+1} \right|\right|_* = \left|\left| \Delta W_{k+1} \right|\right|_*^2 .
$$

Hence, we can continue the previous display as

$$
\Psi_\eta(y_{k+1}) - \Psi_\eta(y_k) \leq -\alpha_k \left|\left| \tilde{\mathcal{G}}_{\eta,\alpha_k}(y_k) \right|\right|^2 \left(1 - \frac{\alpha_k \mathrm{lip}_1(h_\eta)}{2}\right) + \alpha_k \langle \Delta W_{k+1}, \mathcal{G}_{\eta,\alpha_k}(y_k) \rangle + \alpha_k \left|\left| \Delta W_{k+1} \right|\right|_*^2 .
$$

Properties of the prox-gradient mapping

Monotonicity of the prox-gradient mapping

Consider the function $\varphi_y : \alpha \mapsto \frac{1}{\alpha} \left|\left| y - T_{\eta,\alpha}(y) \right|\right|$. For $y \in \mathrm{zer}\,(\partial r_1 + \nabla h_\eta) \triangleq \{ y \in \mathcal{Y} \mid \partial r_1(y) + \nabla h_\eta(y) = 0 \}$, we have $\varphi_y(\alpha) = 0$ for

all $\alpha > 0$. We next prove a classical monotonicity result with respect to the parameter α of this mapping.

Proposition C.1 *If $y \notin \operatorname{zer}(\partial r_1 + \nabla h_\eta)$, then it holds*

$$\alpha_1 > \alpha_2 > 0 \Rightarrow \varphi_y(\alpha_1) < \varphi_y(\alpha_2). \tag{C.1}$$

Proof To simplify notation, let us define $\bar{y}(\alpha) := T_{\eta,\alpha}(y)$. This point satisfies the monotone inclusion (Fermat's optimality principle)

$$\frac{1}{\alpha} B(y - \bar{y}(\alpha)) - \nabla h_\eta(y) \in \partial r_1(\bar{y}(\alpha)).$$

Hence, for $\alpha_1 > \alpha_2 > 0$, the maximal monotonicity of the subdifferential ∂r_1 yields

$$\langle \frac{1}{\alpha_1} B(y - \bar{y}(\alpha_1)) - \frac{1}{\alpha_2} B(y - \bar{y}(\alpha_2)), \bar{y}(\alpha_1) - \bar{y}(\alpha_2) \rangle \geq 0.$$

Rearranging,

$$\begin{aligned}
0 \leq & \frac{1}{\alpha_1} \langle B(y - \bar{y}(\alpha_1)), \bar{y}(\alpha_1) - y \rangle + \frac{1}{\alpha_1} \langle B(y - \bar{y}(\alpha_1)), y - \bar{y}(\alpha_2) \rangle \\
& - \frac{1}{\alpha_2} \langle B(y - \bar{y}(\alpha_2)), y - \bar{y}(\alpha_1) \rangle \\
& - \frac{1}{\alpha_2} \langle B(y - \bar{y}(\alpha_2)), y - \bar{y}(\alpha_2) \rangle \\
= & - \frac{1}{\alpha_1} ||\bar{y}(\alpha_1) - y||^2 - \frac{1}{\alpha_2} ||\bar{y}(\alpha_2) - y||^2 \\
& + \left(\frac{1}{\alpha_1} + \frac{1}{\alpha_2} \right) \langle B(y - \bar{y}(\alpha_1)), y - \bar{y}(\alpha_2) \rangle .
\end{aligned}$$

Consequently,

$$\begin{aligned}
\alpha_1 \varphi_y(\alpha_1)^2 + \alpha_2 \varphi_y(\alpha_2)^2 &\leq (\alpha_1 + \alpha_2) \langle B \left(\frac{y - \bar{y}(\alpha_1)}{\alpha_1} \right), \frac{y - \bar{y}(\alpha_2)}{\alpha_2} \rangle \\
&\leq \frac{\alpha_1 + \alpha_2}{2} \left(\varphi_y(\alpha_1)^2 + \varphi_y(\alpha_2)^2 \right).
\end{aligned}$$

This, in turn leads to,

$$(\alpha_1 - \alpha_2) \left(\varphi_y(\alpha_1)^2 - \varphi_y(\alpha_2)^2 \right) \leq 0 .$$

□

Approximate stationarity

In this appendix, we derive the important relation between the norm of the prox-gradient mapping and stationary points. Consider the smoothed implicit function $\Psi_\eta = h_\eta(y) + r_1(y)$. The prox-gradient mapping is defined as $\mathcal{G}_{\eta,t}(y) = \frac{1}{t}(y - P_t(y, \nabla h_\eta(y))$, where h_η is the Gaussian smoothing of the function h. Since ∇h_η is a Lipschitz continuous operator, the optimality condition defining the point $\bar{y}_t^+ = P_t(y, \nabla h_\eta(y))$ is

$$\begin{aligned} &0 \in t\partial r_1(\bar{y}_t^+) + t[-B\mathcal{G}_{\eta,t}(y) + \nabla h_\eta(y)] \\ \Leftrightarrow &\mathcal{G}_{\eta,t}(y) + B^{-1}(\nabla h_\eta(\bar{y}_t^+) - \nabla h_\eta(y)) \in B^{-1}\partial\Psi_\eta(\bar{y}_t^+)\,. \end{aligned}$$

Hence,

$$\begin{aligned} \operatorname{dist}(0, \partial\Psi_\eta(\bar{y}_t^+)) &\le ||\mathcal{G}_{\eta,t}(y)|| + \frac{\sqrt{n}}{\eta}\operatorname{lip}_0(h)\left|\left|\bar{y}_t^+ - y\right|\right| \\ &\le \left(1 + \frac{t\sqrt{n}}{\eta}\operatorname{lip}_0(h)\right)||\mathcal{G}_{\eta,t}(y)||\,. \end{aligned}$$

In particular, choosing $t \triangleq \frac{\eta}{\sqrt{n}\operatorname{lip}_0(h)}$, the above relation implies $\operatorname{dist}(0, \partial\Psi_\eta(\bar{y}_t^+)) \le 2\,||\mathcal{G}_{\eta,t}(y)||$.

Acknowledgements The first author thanks the FMJH Program Gaspard Monge for optimization and operations research and their interactions with data science for financial support and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 556222748.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

1. Agarwal, A., Dekel, O., Xiao, L.: Optimal algorithms for online convex optimization with multi-point bandit feedback. In 23rd Conference on Learning Theory, pages 28–40, (2010)
2. Arridge, S., Maass, P., Öktem, O., Schönlieb, C.-B.: Solving inverse problems using data-driven models. Acta Numer. **28**, 1–174 (2019)
3. Balasubramanian, K., Ghadimi, S.: Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points. Found. Comput. Math. **22**(1), 35–76 (2022)

4. Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces. Springer - CMS Books in Mathematics, (2016)
5. Benning, M., Burger, M.: Modern regularization methods for inverse problems. Acta Numer. **27**, 1–111 (2018)
6. Berahas, A.S., Cao, L., Choromanski, K., Scheinberg, K.: A theoretical and empirical comparison of gradient approximations in derivative-free optimization. Found. Comput. Math. **22**(2), 507–560 (2022)
7. Bottou, L., Curtis, F., Nocedal, J.: Optimization methods for large-scale machine learning. SIAM Rev. **60**(2), 223–311 (2018)
8. Cesa-Bianchi, N., Lugosi, G.: Prediction, learning, and games. Cambridge University Press, Cambridge (2006)
9. Clarke, F.H.: Optimization and nonsmooth analysis. SIAM, (1990)
10. Conn, A.R., Scheinberg, K., Vicente, L.N.: Introduction to derivative-free optimization. SIAM, (2009)
11. Cui, S., Shanbhag, U.V., Staudigl, M.: A regularized variance-reduced modified extragradient method for stochastic hierarchical games. arXiv preprint arXiv:2302.06497, (2023)
12. Cui, S., Shanbhag, U.V., Yousefian, F.: Complexity guarantees for an implicit smoothing-enabled method for stochastic mpecs. Mathematical Programming, (2022)
13. Davis, D., Drusvyatskiy, D.: Stochastic model-based minimization of weakly convex functions. SIAM J. Optim. **29**(1), 207–239 (2019)
14. Davis, D., Grimmer, B.: Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. SIAM J. Optim. **29**(3), 1908–1930 (2019)
15. Dontchev, A.L., Rockafellar, R.T.: Implicit functions and solution mappings: A view from variational analysis, volume 11. Springer, (2009)
16. Drusvyatskiy, D., Paquette, C.: Efficiency of minimizing compositions of convex functions and smooth maps. Math. Program. **178**, 503–558 (2019)
17. Duchi, J.C., Jordan, M.I., Wainwright, M.J., Wibisono, A.: Optimal rates for zero-order convex optimization: The power of two function evaluations. IEEE Trans. Inf. Theory **61**(5), 2788–2806 (2015)
18. Duvocelle, B., Mertikopoulos, P., Staudigl, M., Vermeulen, D.: Multiagent online learning in time-varying games. Mathematics of Operations Research, 2023/01/31 (2022)
19. Ehrhardt, M.J., Roberts, L.: Inexact derivative-free optimization for bilevel learning. J. Math. Imaging Vis. **63**(5), 580–600 (2021)
20. Ehrhardt, M.J., Roberts, L.: Analyzing inexact hypergradients for bilevel learning. IMA J. Appl. Math. **89**(1), 254–278 (2024)
21. Engl, H.W., Hanke, M., Neubauer, G.: Regularization of inverse problems. Mathematics and Its Applications. Springer, Netherlands (1996)
22. Franceschi, L., Frasconi, P., Salzo, S., Grazzi, R., Pontil, M.: Bilevel programming for hyperparameter optimization and meta-learning. In International conference on machine learning, pages 1568–1577. PMLR, (2018)
23. Garnett, R.: Bayesian optimization. Cambridge University Press, Cambridge (2023)
24. Ghadimi, Saeed, Lan, Guanghui: Stochastic first- and zeroth-order methods for nonconvex stochastic programming. SIAM J. Optim. **23**(4), 2341–2368 (2013)
25. Ghadimi, S., Lan, G., Zhang, H.: Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. Math. Program. **155**(1–2), 267–305 (2016)
26. Ghadimi, S., Wang, M.: Approximation methods for bilevel programming. arXiv preprint arXiv:1802.02246, (2018)
27. Giles, M.B.: Multilevel Monte Carlo methods. Acta Numer. **24**, 259–328 (2015)
28. Goldstein, A.A.: Optimization of Lipschitz continuous functions. Math. Program. **13**, 14–22 (1977)
29. Grazzi, R., Franceschi, L., Pontil, M., Salzo, S.: On the iteration complexity of hypergradient computation. International Conference on Machine Learning, pages 3748–3758, (2020)
30. Haber, E., Tenorio, L.: Learning regularization functionals–a supervised training approach. Inverse Prob. **19**(3), 611 (2003)
31. Hansen, P.C., Jørgensen, J., Lionheart, W.R.B.: Computed tomography: algorithms, insight, and just enough theory. Society for Industrial and Applied Mathematics, Philadelphia, PA, (2021)
32. Holler, G., Kunisch, K., Barnard, R.C.: A bilevel approach for parameter learning in inverse problems. Inverse Prob. **34**(11), 115012 (2018)

33. Hong, M., Wai, H.-T., Wang, Z., Yang, Z.: A two-timescale stochastic algorithm framework for bilevel optimization: complexity analysis and application to actor-critic. SIAM J. Optim. **33**(1), 147–180 (2023)
34. Kozak, D., Molinari, C., Rosasco, L., Tenorio, L., Villa, S.: Zeroth-order optimization with orthogonal random directions. Mathematical Programming, (2022)
35. Kunisch, K., Pock, T.: A bilevel optimization approach for parameter learning in variational models. SIAM. J. Imaging. Sci. **6**(2), 938–983 (2013)
36. Kwon, J., Kwon, D., Wright, S., Nowak, R.D.: A fully first-order method for stochastic bilevel optimization. PMLR, (2023)
37. Lan, G.: First-order and Stochastic Optimization Methods for Machine Learning. Springer Series in the Data Sciences, Springer Nature (2020)
38. Lei, M., Pong, T.K., Sun, S., Yue, M.-C.: Subdifferentially polynomially bounded functions and gaussian smoothing-based zeroth-order optimization. arXiv preprint arXiv:2405.04150, (2024)
39. Liu, R., Gao, J., Zhang, J., Meng, D., Lin, Z.: Investigating bi-level optimization for learning and vision from a unified perspective: a survey and beyond. IEEE Trans. Pattern Anal. Mach. Intell. **44**(12), 10045–10067 (2022)
40. Lu, Z., Mei, S.: First-order penalty methods for bilevel optimization. SIAM J. Optim. **34**(2), 1937–1969 (2024)
41. Nesterov, Y.: Lectures on Convex Optimization, volume 137 of Springer optimization and its applications. Springer International Publishing, (2018)
42. Nesterov, Y., Spokoiny, V.: Random gradient-free minimization of convex functions. Found. Comput. Math. **17**(2), 527–566 (2017)
43. Ochs, P., Ranftl, R., Brox, T., Pock, T.: Techniques for gradient-based bilevel optimization with nonsmooth lower level problems. J. Math. Imaging Vision **56**, 175–194 (2016)
44. Pougkakiotis, S., Kalogerias, D.: A zeroth-order proximal stochastic gradient method for weakly convex stochastic optimization. SIAM J. Sci. Comput. **45**(5), A2679–A2702 (2023)
45. Rajeswaran, A., Finn, C., Kakade, S.M.: and Sergey Levine. Meta-learning with implicit gradients. Adv. Neural Inf. Process. Syst. 32, (2019)
46. Shapiro, A., Dentcheva, D., Ruszczyński, A.: Lectures on stochastic programming. Society for Industrial and Applied Mathematics, 2017/12/29 (2009)
47. Sinha, A., Khandait, T., Mohanty, R.: A gradient-based bilevel optimization approach for tuning regularization hyperparameters. Optim. Lett. **18**(6), 1383–1404 (2024)
48. Spall, J.C.: Introduction to stochastic search and optimization: estimation, simulation, and control. John Wiley & Sons, (2005)
49. Wang, T., Lucka, F., van Leeuwen, T.: Sequential experimental design for x-ray CT using deep reinforcement learning. IEEE Trans. Comput. Imaging **10**, 953–968 (2024)
50. Zhang, J., Lin, H., Jegelka, S., Sra, S., Jadbabaie, A.: Complexity of finding stationary points of nonconvex nonsmooth functions. International conference on machine learning, pages 11173–11182, (2020)

Authors and Affiliations

Mathias Staudigl[1] · Simon Weissmann[1] · Tristan van Leeuwen[2,3]

✉ Mathias Staudigl
mathias.staudigl@uni-mannheim.de

Simon Weissmann
simon.weissmann@uni-mannheim.de

Tristan van Leeuwen
T.van.Leeuwen@cwi.nl

[1] Department of Mathematics, Mannheim University, B6 26, 68159 Mannheim, Germany

[2] Centrum Wiskunde & Informatica, Science Park Amsterdam 123, 1098 XG Amsterdam, The Netherlands

[3] Utrecht University, Heidelberglaan 8, 3584 CS Utrecht, The Netherlands