A zeroth-order proximal stochastic gradient method for weakly convex stochastic optimization

Pougkakiotis, Spyridon; Kalogerias, Dionysis

Published in:
SIAM Journal on Scientific Computing

DOI:
10.1137/22M1494270

Publication date:
2023

This is an Author Accepted manuscript of an article published by SIAM Journal on Scientific Computing. The version of record Pougkakiotis, S., & Kalogerias, D. (2023). A zeroth-order proximal stochastic gradient method for weakly convex stochastic optimization. SIAM Journal on Scientific Computing, 45(5) is available online at: https://doi.org/10.1137/22M1494270

A ZEROTH-ORDER PROXIMAL STOCHASTIC GRADIENT METHOD FOR WEAKLY CONVEX STOCHASTIC OPTIMIZATION

SPYRIDON POUGKAKIOTIS* AND DIONYSIOS S. KALOGERIAS†

Abstract.
In this paper we analyze a zeroth-order proximal stochastic gradient method suitable for the minimization of weakly convex stochastic optimization problems. We consider nonsmooth and nonlinear stochastic composite problems, for which (sub-)gradient information might be unavailable. The proposed algorithm utilizes the well-known Gaussian smoothing technique, which yields unbiased zeroth-order gradient estimators of a related partially smooth surrogate problem (in which one of the two nonsmooth terms in the original problem's objective is replaced by a smooth approximation). This allows us to employ a standard proximal stochastic gradient scheme for the approximate solution of the surrogate problem, which is determined by a single smoothing parameter, and without the utilization of first-order information. We provide state-of-the-art convergence rates for the proposed zeroth-order method using minimal assumptions. The proposed scheme is numerically compared against alternative zeroth-order methods as well as a stochastic sub-gradient scheme on a standard phase retrieval problem. Further, we showcase the usefulness and effectiveness of our method for the unique setting of automated hyper-parameter tuning. In particular, we focus on automatically tuning the parameters of optimization algorithms by minimizing a novel heuristic model. The proposed approach is tested on a proximal alternating direction method of multipliers for the solution of $\mathcal{L}_1/\mathcal{L}_2$-regularized PDE-constrained optimal control problems, with evident empirical success.

Key words. Zeroth-order optimization, weakly convex stochastic optimization, stochastic gradient descent, hyper-parameter tuning, composite optimization

MSC codes. 90C15, 90C56, 90C30

1. Introduction.

We are interested in the solution of stochastic weakly convex optimization problems that are not necessarily smooth. Let $(\Omega, \mathscr{F}, P)$ be any complete base probability space, and consider a random vector $\xi : \Omega \to \mathbb{R}^d$. We are interested in stochastic optimization problems of the form

$$\text{(P)} \qquad \min_{x\in\mathbb{R}^n} \phi(x) \coloneqq f(x) + r(x), \qquad f(x) \coloneqq \mathbb{E}_\xi \left[F\left(x, \xi\right)\right],$$

where $F\colon \mathbb{R}^n \times \Xi \to \mathbb{R}$ is Borel in ξ, f is weakly convex, while $r\colon \mathbb{R}^n \to \overline{\mathbb{R}} \equiv \mathbb{R}\cup\{+\infty\}$ is a proper convex lower semi-continuous function (and hence closed), which is assumed to be proximable (that is, its proximity operator can be computed analytically).

Problem (P) is very general and appears in a variety of applications arising in signal processing (e.g. [18]), optimization (e.g. [33]), engineering (e.g. [31]), machine learning (e.g. [32]), and finance ([43]), to name a few. The reader is referred to [13, Section 2.1] and [15, Section 3.1] for a plethora of examples. Since neither f nor r are assumed to be smooth, standard stochastic gradient-based schemes are not applicable. In light of this, the authors in [13] analyzed various model-based stochastic sub-gradient methods (using a standard generalization of the convex subdifferential) for the efficient solution of (P) and were able to show that convergence is achieved in the sense of near-stationarity of the Moreau envelope of ϕ ([36]), which serves as a surrogate function with stationary points coinciding with those of (P). Given an approximate solution to (P), the Moreau envelope offers a way to approximately measure its distance from stationarity in the absence of differentiability. Indeed, a nearly stationary point for the Moreau envelope is close to a nearly stationary point for the problem under consideration (see [13, Section 2.2] or Section 3.1).

*Department of Electrical Engineering, Yale University (spyridon.pougkakiotis@yale.edu).
†Department of Electrical Engineering, Yale University (dionysis.kalogerias@yale.edu.

However, there is a variety of applications in which even sub-gradient information of f (or that of $F(\cdot,\xi)$) might not be available due to the lack of sufficient knowledge about the function (e.g. [2, 8, 24]), or such a computation might be prohibitively expensive or noisy (e.g. see [1, 29, 35]). Thus, several zeroth-order schemes have been developed for the solution of stochastic optimization problems similar to (P), requiring only function evaluations of $F(\cdot,\xi)$. Such methods utilize zeroth-order gradient estimates of an appropriate (closely related) surrogate function $F_\mu(\cdot,\xi)$ which depends on a smoothing parameter $\mu > 0$.

Zeroth-order methods have a long history within the field of optimization (e.g. see the seminal paper on the well-known simultaneous perturbation stochastic approximation (SPSA) [49], the well-known Matyas' method [3, 34, 46], or the more recent discussion in [12, Chapter 1]). However, the relatively recent works on the *Gaussian and uniform smoothing* techniques for convex [16, 38] and differentiable non-convex programming [23] have sparked a lot of interest in the literature. Following these developments, the authors in [27] developed and analyzed a zeroth-order scheme based on the Gaussian smoothing (see [38]) for the solution of stochastic compositional problems with applications to risk-averse learning, in which r is chosen as an indicator function to a compact convex set. The authors in [4], based on the earlier work in [23], considered (Gaussian smoothing-based) zeroth-order schemes for non-convex Lipschitz smooth stochastic optimization problems, again assuming that r is an indicator function, and focusing on high-dimensionality issues as well as on avoiding saddle-points. We note that the class of non-convex Lipschitz smooth functions is encompassed within the class of weakly convex ones and hence the class of functions appearing in (P) is strictly wider (see Proposition 2.3). In general, there is a plethora of zeroth-order optimization algorithms, and the interested reader is referred to [5, 12, 17, 28, 38, 49, 54], and the references therein.

To the best of our knowledge, the only developments on zeroth-order methods for the solution of (P) can be found in the recent articles given in [30, 37]. The authors in [30] utilize a double Gaussian smoothing scheme, which was originally proposed for convex functions in [16]. We argue herein that the use of double smoothing is essentially unnecessary, at least in conjunction with the discussion in [30]. In particular, the analysis of the proposed algorithm in [30] is substantially more complicated as compared to the analysis provided herein (cf. Section 3 and [30, Section 3]), while at the same time offering no advantage in terms of the rate bounds achieved (both here as well as in [30] an $\mathcal{O}(\sqrt{n}\epsilon^{-4})$ rate is shown; cf. Theorem 3.4 and [30, Theorem 1]). Additionally, in [30] it is assumed that the iterates produced by the proposed algorithm remain bounded, an assumption that is not required in our analysis. Further, as we show in Section 4.1, the double smoothing approach, except from the fact that it requires the tuning of two smoothing parameters, does not exhibit better convergence behaviour in practice as compared to the proposed method herein. On the other hand, the authors in [37] present an adaptive zeroth-order method for problems of the form of (P) using a uniform smoothing scheme. However, the analysis in the aforementioned paper yields a worse dependence on the problem dimensions n than that obtain herein, while at the same time requires certain additional restrictive assumptions (in particular, an $\mathcal{O}(n^2\epsilon^{-4})$ convergence rate is shown, cf. Theorem 3.4 and [37, Corollary 19], and the authors assume that the iterates lie in a compact set and that the function $F(\cdot,\xi)$ is Lipschitz continuous with a constant that does not depend on ξ; neither of these is assumed in our analysis).

Instead, in this paper we develop and analyze a zeroth-order proximal stochastic gradient method for the solution of (P), utilizing standard (single) Gaussian smoothing (see [38]). Following the developments in [13], we analyze the algorithm and show that it obtains an ϵ-stationary solution to the Moreau envelope of an appropriate *surrogate problem* in at most $\mathcal{O}(\sqrt{n}\epsilon^{-4})$ iterations; a state-of-the-art bound of the same order as the bound achieved by sub-gradient schemes (see [13]), up to a constant term depending on the square root of the dimension of x (i.e. $\sqrt{n}$). This rate matches the one shown in [30] for the double Gaussian smoothing scheme, however, the proposed analysis is significantly easier, and does not assume boundedness of the iterates, which is required for the analysis in [30]. Additionally, given any near-stationary solution to the surrogate problem for which the convergence analysis is performed, we show that it is a near-stationary solution for the Moreau envelope of the original problem. Such a connection is easy to establish when r is an indicator function (e.g. see [27]), however not so obvious for general closed convex functions r that are studied here. Indeed, this was not considered in [30]. A rate directly related to the Moreau envelope of the original problem is given in the analysis in [37] (where a uniform smoothing scheme is studied), however, the analysis in the aforementioned work utilizes additional restrictive assumptions to achieve this (as previously mentioned, boundedness of the problem's domain and Lipschitz continuity of $F(\cdot,\xi)$ with a uniform Lipschitz constant for all ξ), while an $\mathcal{O}(n^2\epsilon^{-4})$ rate is shown (i.e. a significantly worse dependence on the problem dimensions n).

In order to empirically stress the viability and usefulness of the proposed approach, we consider two problems. Initially, we test our method on several phase-retrieval instances taken from [13], and compare its numerical behaviour against a sub-gradient model-based scheme developed in [13], as well zeroth-order stochastic gradient schemes based on the double Gaussian smoothing, the uniform smoothing, and the SPSA. The observed numerical behaviour confirms the theory, in that the proposed zeroth-order method converges consistently at a rate that is slower only by a constant factor than that exhibited by the sub-gradient scheme, while it is competitive against all other zeroth-order schemes. Subsequently, we showcase that the practical performance of the proposed algorithm is seemingly identical to that achieved by the double smoothing zeroth-order scheme analyzed in [30], even if the two smoothing parameters of the latter are tuned.

Next, we consider a very important application of zeroth-order (or in general derivative-free) optimization; that is hyper-parameter tuning. This is a very old problem (traditionally appearing in the industry, e.g. see [8], and often solved by hand via exhausting or heuristic random search schemes) that has seen a surge in importance in light of the recent developments in artificial intelligence and machine learning. There is a wide literature on this subject, which can only briefly be mentioned here. The most common approaches are based on Bayesian optimization techniques (e.g. see [6, 7, 22]), although derivative-free schemes have also been considered (e.g. see [2]). In certain special cases, application specific automated tuning strategies have also been investigated (e.g. see [10, 21, 42]). Given the importance of hyper-parameter tuning, there have been developed several heuristic software packages for this purpose, such as the Nevergrad toolkit (see [25]). In this paper, we consider the problem of tuning the parameters of optimization algorithms. To that end, we derive a novel heuristic model, the minimization of which yields the hyper-parameters that minimize the residual reduction of an optimization algorithm that depends on them, after a fixed given number of iterations, for an arbitrary class of optimization problems (assumed to follow an unknown distribution from which we can sample). Focusing on a proximal alternating direction method of multipliers (pADMM), we tune its penalty parameter for two problem classes; the optimal control of the Poisson equation as well as the optimal control of the convection-diffusion equation. In both cases we numerically verify the efficient performance of the pADMM with the "learned" hyper-parameter when considering out-of-sample instances. The MATLAB implementation is provided.

Notation. We denote by $\langle\cdot,\cdot\rangle$ the inner product in $\mathbb{R}^n$, and given a vector $x \in \mathbb{R}^n$, $\|x\|_2$ denotes the induced Euclidean norm. Given a complete probability space $(\Omega, \mathscr{F}, P)$, where $\mathscr{F}$ is a sigma algebra and P is a probability measure, we denote by $\mathcal{L}_p(\Omega, \mathscr{F}, P; \mathbb{R})$, for some $p \in [1, +\infty)$, the space of all $\mathscr{F}$-measurable functions $\varphi\colon \Omega \to \mathbb{R}$ such that $\left(\int_\Omega |\varphi(\omega)|^p \, dP(\omega)\right)^{1/p} < +\infty$. Given a random vector $Z\colon \Omega \to \mathbb{R}^d$, and a random function $\varphi\colon \mathbb{R}^d \to \overline{\mathbb{R}}$, we denote the expected value as $E_Z[\varphi(Z)] = \int_\Omega \varphi\left(Z(\omega)\right) dP(\omega)$, where the subscript is employed to stress that the expectation is taken with respect to the random variable Z. Finally, given a function $\varphi\colon \mathbb{R}^n \to \mathbb{R}^m$, we say that φ is Lipschitz continuous on a set $X \subset \mathbb{R}^n$ if there is a constant $c \geq 0$ such that $\|\varphi(x_1) - \varphi(x_2)\|_2 \leq c\|x_1 - x_2\|_2$, for all $x_1,\ x_2 \in X$. If φ is Lipschitz continuous on a neighbourhood of every point of X (potentially with different Lipschitz constants), then it is said that φ is locally Lipschitz continuous on X.

Structure of the article. The rest of this paper is organized as follows. In Section 2 we introduce some notation as well as preliminary notions of significant importance for the developments in this paper. In Section 3 we derive and analyze the proposed zeroth-order proximal stochastic gradient method for the solution of (P). In Section 4 we present some numerical results, and in Section 5 we derive our conclusions.

2. Preliminaries.

In this section, we introduce some preliminary notions that will be used throughout this paper. In particular, we first discuss certain core properties of stochastic weakly convex functions of the form of f. Subsequently, we introduce the Gaussian smoothing (e.g. see [27, 38]), which provides a smooth surrogate for f in (P). In turn, this can be used to obtain zeroth-order optimization schemes; such methods are only allowed to access a zeroth-order oracle (i.e. only sample-function evaluations are available). In turn, the Gaussian smoothing guides us in the choice of minimal assumptions on the stochastic part of the objective function in (P). Finally, we introduce the proximity operator, as well as certain core properties of it. These notions will then be used to derive a zeroth-order proximal stochastic gradient method in Section 3.

2.1. Stochastic weakly convex functions.

Let us briefly discuss some core properties of the well-studied class of weakly convex functions. For a detailed study on the properties of these functions (and of related sets), the reader is referred to [52], and the references therein. Below we define the class of weakly convex functions for completeness.

DEFINITION 2.1. *Let $f\colon \mathbb{R}^n \mapsto \mathbb{R}$. It is said to be ρ-weakly convex, for some $\rho > 0$, if for any $x_1,\ x_2 \in \mathbb{R}^n$, and any $\lambda \in [0,1]$, it holds that*

$$f\left(\lambda x_1 + (1-\lambda)x_2\right) \leq \lambda f(x_1) + (1-\lambda)f(x_2) + \frac{\lambda(1-\lambda)\rho}{2}\left\|x_1 - x_2\right\|_2^2.$$

In what follows, we make use of a standard generalization of the well-known convex subdifferential (which consists of all global affine under-estimators of a convex function at a given point). Specifically, we consider the subdifferential that consists of all global concave quadratic under-estimators (see [13, Section 2.2]). In particular, given a locally Lipschitz continuous function $f\colon \mathbb{R}^n \mapsto \overline{\mathbb{R}}$, and some $x \in \operatorname{dom}(f)$, we define the generalized subdifferential $\partial f(x)$ as the set of all vectors $v \in \mathbb{R}^n$ satisfying

$$f(y) \geq f(x) + \langle v, y - x \rangle + o\left(\|y - x\|_2\right), \qquad \text{as } y \to x,$$

and set $\partial f(x) = \emptyset$ for any $x \notin \mathrm{dom}(f)$. A more general definition, based on the Clarke generalized directional derivative (see [11]), can be found in [52, Section 1]. We note that the mapping $x \mapsto \partial f(x)$ of a weakly convex function f inherits many properties of the subgradient mapping of a convex function (see [52, Section 4]), and reduces to the standard convex subdifferential if f is a convex function. In the following proposition we state some important properties holding for weakly convex functions.

PROPOSITION 2.2. *Any ρ-weakly convex function $f \colon \mathbb{R}^n \mapsto \mathbb{R}$ is locally Lipschitz continuous and regular in the sense of Clarke, and thus directionally differentiable. Furthermore, it is bounded below, and there exists $z \in \mathbb{R}^n$ such that*

$$f(x_2) \geq f(x_1) + \langle z, x_2 - x_1 \rangle - \frac{\rho}{2} \|x_2 - x_1\|_2^2.$$

Moreover, the latter holds for any $z \in \partial f(x_1)$. Finally, the map $x \mapsto f(x) + \frac{\rho}{2}\|x\|_2^2$ is convex and

$$\langle z_1 - z_2, x_1 - x_2 \rangle \geq -\rho \|x_1 - x_2\|_2^2,$$

for all $x_1, x_2 \in \mathbb{R}^n$, $z_1 \in \partial f(x_1)$, and $z_2 \in \partial f(x_2)$.

Proof. The proof can be found in [52, Propositions 4.4, 4.5, and 4.8]. ☐

PROPOSITION 2.3. *Any continuously differentiable function $f \colon \mathbb{R}^n \to \mathbb{R}$, with globally ρ-Lipschitz gradient, where $\rho > 0$, is ρ-weakly convex.*

Proof. The proof follows trivially from Proposition 2.2, see [52, Proposition 4.12].☐

2.2. Gaussian smoothing. Let us introduce the notion Gaussian smoothing. To that end, we follow the notation adopted in [27]. Let $f \colon \mathbb{R}^n \to \mathbb{R}$ be a Borel function, and $U \sim \mathcal{N}\left(0_n, I_n\right)$ a normal random vector, where I_n is the identity matrix of size n. Given a non-negative smoothing parameter $\mu \geq 0$, the Gaussian smoothing of f is defined as

$$f_\mu(\cdot) \coloneqq \mathbb{E}_U\left[f\left((\cdot) + \mu U\right)\right],$$

assuming that the expectation is well-defined and finite for all $x \in \mathbb{R}^n$. The precise conditions on $F(x, \xi)$ (in (P)) for this to hold will be given later in this section. Let $\mathcal{N} \colon \mathbb{R}^n \to \mathbb{R}$, with a slight abuse of notation, be the standard Gaussian density in $\mathbb{R}^n$, that is the mapping $x \mapsto \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}x^\top x}$. Then, we can observe that:

$$f_\mu(x) = \int f(x + \mu u) \mathcal{N}\left(u\right) du = \mu^{-n} \int f(v) \mathcal{N}\left(\frac{v - x}{\mu}\right) dv,$$

where the second equality holds via introducing an integration variable $v = x + \mu u$. The second characterization yields the following expressions for the gradient of f_μ (assuming it exists):

$$\begin{aligned}
\nabla f_\mu(x) &= \mu^{-(n+2)} \int f(v) \mathcal{N}\left(\frac{v - x}{\mu}\right)(v - x) dv \\
&= \mu^{-1} \int f(x + \mu u) \mathcal{N}\left(u\right) u du \\
&= \mathbb{E}_U\left[\frac{f\left(x + \mu U\right) - f(x)}{\mu} U\right] \\
&= \mathbb{E}_U\left[\frac{f\left(x + \mu U\right) - f\left(x - \mu U\right)}{2\mu} U\right],
\end{aligned}$$

where $U \sim \mathcal{N}(0_n, I_n)$. The second equality follows from a change of variables, the third from the properties of the standard Gaussian, while the last one can be trivially shown by direct computation (e.g. see [38]).

In what follows, we impose certain assumptions on the function F given (implicitly) in (P), in order to guarantee that its Gaussian smoothing is well-defined and satisfies several properties of interest.

ASSUMPTION 2.4. *Let $F\colon \mathbb{R}^n \times \Xi \to \mathbb{R}$ satisfy the following properties:*

(C1) *$F(x,\cdot) \in \mathcal{L}_2(\Omega, \mathscr{F}, P; \mathbb{R})$, and is Borel for any $x \in \mathbb{R}^n$.*

(C2) *The function $f(x) = \mathbb{E}_\xi[F(x,\xi)]$ is ρ-weakly convex for some $\rho \geq 0$.*

(C3) *There exists a positive random variable $C(\xi)$ such that $\sqrt{\mathbb{E}_\xi[C(\xi)^2]} < \infty$, and for all $x_1,\ x_2 \in \mathbb{R}^n$, and a.e. $\xi \in \Xi$, the following holds:*

$$|F(x_1,\xi) - F(x_2,\xi)| \leq C(\xi)\|x_1 - x_2\|_2.$$

Remark 2.5. In view of (**C1**) in Assumption 2.4, we can infer that f is well-defined and finite for any x. In fact, this can be shown with a weaker condition in place of (**C1**), that is, if we were to assume that $F(x,\cdot) \in \mathcal{L}_1(\Omega, \mathscr{F}, P; \mathbb{R})$ for any $x \in \mathbb{R}^n$. The stronger assumption will be utilized in Lemma 2.6. Furthermore, from [45, Theorem 7.44], under (**C1**) and (**C3**), it follows that there exists a constant $L_{f,0} > 0$, such that f is $L_{f,0}$-Lipschitz continuous on $\mathbb{R}^n$. Again, this holds even if we weaken assumption (**C3**), and only require that $\mathbb{E}_\xi[C(\xi)] < \infty$, however, the stronger form of this assumption is utilized in Lemma 2.6.

Under Assumption 2.4, we will provide certain properties of the surrogate function f_μ, as presented in [38].

LEMMA 2.6. *Let Assumption 2.4 hold. Then, f_μ is ρ-weakly convex, and there exists a constant $L_{f_\mu,0} \leq L_{f,0}$ such that f_μ is $L_{f_\mu,0}$-Lipschitz continuous on $\mathbb{R}^n$. Additionally, for any $\mu \geq 0$, we obtain*

$$|f_\mu(x) - f(x)| \leq \mu L_{f,0} n^{\frac{1}{2}}, \qquad \text{for any } x \in \mathbb{R}^n, \tag{2.1}$$

while for any $\mu > 0$, f_μ is Lipschitz continuously differentiable with

$$\nabla f_\mu(x) = \mathbb{E}_U\left[\frac{f(x+\mu U) - f(x)}{\mu}U\right] = \mathbb{E}_{U,\xi}\left[\frac{F(x+\mu U,\xi) - F(x,\xi)}{\mu}U\right], \tag{2.2}$$

where $U,\ \xi$ are statistically independent. Additionally, we have that

$$\mathbb{E}_{U,\xi}\left[\left\|\frac{F(x+\mu U,\xi) - F(x,\xi)}{\mu}U\right\|_2^2\right] \leq (n^2+2n)L_{f,0}^2. \tag{2.3}$$

Proof. Weak convexity of the surrogate can be obtained by [27, Lemma 5.2]. For a proof of (2.1), as well as the first equality of (2.2), the reader is referred to [38, Appendix, Proof of Theorem 1]. The second equality in (2.2), in light of (**C3**) of Assumption 2.4, follows by Fubini's theorem (we should note that with a slight abuse of notation, the second expectation in (2.2) is taken with respect to the product measure of the two corresponding random vectors U and ξ). Following the developments in

[27, Lemma 5.4], we show (2.3). In particular, we have

$$\begin{aligned}
\mathbb{E}_{U,\xi}\left[\left\|\frac{F\left(x+\mu U,\xi\right)-F(x,\xi)}{\mu}U\right\|_2^2\right] &= \frac{1}{\mu^2}\mathbb{E}_{U,\xi}\left[\left|F\left(x+\mu U,\xi\right)-F(x,\xi)\right|^2\|U\|_2^2\right] \\
&= \frac{1}{\mu^2}\mathbb{E}_{U}\left[\mathbb{E}_{\xi}\left[\left|F\left(x+\mu U,\xi\right)-F(x,\xi)\right|^2\|U\|_2^2\middle|U\right]\right] \\
&= \frac{1}{\mu^2}\mathbb{E}_{U}\left[\mathbb{E}_{\xi}\left[\left|F\left(x+\mu U,\xi\right)-F(x,\xi)\right|^2\middle|U\right]\|U\|_2^2\right] \\
&\leq L_{f,0}^2\mathbb{E}_U\left[\|U\|_2^4\right] = (n^2+2n)L_{f,0}^2,
\end{aligned}$$

where in the second equality we used the tower property, while in the last line we employed (**C3**), and evaluated the 4-th moment of the χ-distribution. ◻

2.3. Proximal point and the Moreau envelope.

At this point, we briefly discuss certain well-known notions for completeness. More specifically, given a closed function $p\colon \mathbb{R}^n \to \overline{\mathbb{R}}$, and a positive penalty $\lambda > 0$, we define the proximal point

$$\mathbf{prox}_{\lambda p}(u) \coloneqq \arg\min_x \left\{p(x) + \frac{1}{2\lambda}\|u-x\|_2^2\right\},$$

as well as the corresponding Moreau envelope

$$p^{\lambda}(u) \coloneqq \min_x \left\{p(x) + \frac{1}{2\lambda}\|x-u\|_2^2\right\} = p\left(\mathbf{prox}_{\lambda p}(u)\right) + \frac{1}{2\lambda}\left\|\mathbf{prox}_{\lambda p}(u) - u\right\|_2^2.$$

We can show (e.g. see [13, 36]) that if p is ρ-weakly convex, for some $\rho > 0$, then p_λ is continuously differentiable for any $\lambda \in \left(0, \rho^{-1}\right)$, with

$$\nabla p^{\lambda}(u) = \lambda^{-1}\left(u - \mathbf{prox}_{\lambda p}(u)\right).$$

The Moreau envelope has been used as a smooth penalty function for line-search in Newton-like methods (e.g. see [39]). More recently, it was noted in [13, Section 2.2] that the norm of its gradient (that is $\|\nabla p^{\lambda}(u)\|_2$) can serve as a near-stationarity measure for nonsmooth optimization. The latter approach is adopted in this paper, and thus, we will later on derive a convergence analysis of the proposed zeroth-order proximal stochastic gradient method based on the magnitude of the gradient of an appropriate Moreau envelope.

3. A zeroth-order proximal stochastic gradient method.

In this section we derive a zeroth-order proximal stochastic gradient method suitable for the solution of problems of the form of (P). Let us employ the following assumption:

ASSUMPTION 3.1. *Let $F(x,\xi)$ be defined as in (P) satisfying Assumption 2.4. Additionally, we assume that r is a proper (i.e.* $\text{dom}(r) \neq \emptyset$*) closed convex function (and thus lower semi-continuous), and proximable (that is, its proximity operator can be evaluated analytically). Finally, we can generate two statistically independent random sequences $\{U_i\}_{i=0}^{\infty}$, $\{\xi_i\}_{i=0}^{\infty}$, such that each $U_i \sim \mathcal{N}\left(0_n, I_n\right)$ and ξ_i is i.i.d., respectively.*

In light of Assumption 3.1, and by utilizing Lemma 2.6, we can quantify the quality of the approximation of $\phi(x)$ by $\phi_\mu(x) \coloneqq f_\mu(x) + r(x)$, for any $x \in \mathbb{R}^n$. Additionally, we know that f_μ is smooth, even if f is not. Thus, we can derive an optimization algorithm for the minimization of ϕ_μ (which can utilize stochastic gradient approximations for the smooth function f_μ), and then retrieve an approximate solution to the original problem, where the approximation accuracy can be directly controlled by the smoothing parameter μ. Thus, we analyze a zeroth-order stochastic optimization method for the solution of the following surrogate problem

$$(\mathrm{P}_\mu) \qquad \min_x \; \phi_\mu(x) \coloneqq f_\mu(x) + r(x),$$

where $f_\mu(x) = \mathbb{E}_U\left[f\left(x+\mu U\right)\right]$, $\mu > 0$, and f, r are as in (P). The method is summarized in Algorithm Z-ProxSG.

Algorithm Z-ProxSG Zeroth-Order Proximal Stochastic Gradient

Input: $x_0 \in \mathrm{dom}(r)$, a sequence $\{\alpha_t\}_{t\geq 0} \subset \mathbb{R}_+$, $\mu > 0$, and $T > 0$.
for $(t = 0, 1, 2, \ldots, T)$ **do**
Sample ξ_t, $U_t \sim \mathcal{N}\left(0_n, I_n\right)$, and set

$$x_{t+1} = \mathbf{prox}_{\alpha_t r}\left(x_t - \alpha_t G\left(x_t, U_t, \xi_t\right)\right),$$

where $G\left(x_t, U_t, \xi_t\right) \coloneqq \mu^{-1}\left(F\left(x_t + \mu U_t, \xi_t\right) - F(x_t, \xi_t)\right) U_t$.
end for
Sample $t^* \in \{0, \ldots, T\}$ according to $\mathbb{P}(t^* = t) = \frac{\alpha_t}{\sum_{i=0}^T \alpha_i}$.
return x_{t^*}.

3.1. Convergence analysis. In what follows, we derive the convergence analysis for Algorithm Z-ProxSG. We obtain the rate of the proposed algorithm for finding a nearly-stationary solution to the surrogate problem (P_μ) (see Theorem 3.4), and then by utilizing Lemma 2.6, we argue that a nearly-stationary solution of the surrogate problem is nearly-stationary for the Moreau envelope of problem (P) (see Theorem 3.6). The analysis follows closely the developments in [13, Section 3.2].

Let us first introduce some notation. Set $\bar{\rho} \in (\rho, 2\rho]$, where ρ is the weak-convexity constant of $f(\cdot)$. We define $\hat{x}_t \coloneqq \mathbf{prox}_{\bar{\rho}^{-1}\phi_\mu}(x_t)$, and $\delta_t \coloneqq 1 - \alpha_t\bar{\rho}$. The auxiliary point $\hat{x}_t$ is the "optimal" proximal step at iteration t. In Lemma 3.3, we show how far is the new iterate of Algorithm Z-ProxSG (in expectation) from this "optimal" proximal step. In turn, this bound is then utilized in Theorem 3.4 to show convergence in terms of reduction of the gradient norm of the surrogate Moreau envelope. The following lemma introduces a useful property of this auxiliary point.

LEMMA 3.2. *For any $t \geq 0$, and any iterate x_t of Algorithm Z-ProxSG, we obtain*

$$\hat{x}_t = \mathbf{prox}_{\alpha_t r}\left(\alpha_t \bar{\rho} x_t - \alpha_t \nabla f_\mu(x_t) + \delta_t \hat{x}_t\right).$$

Proof. See Appendix A.1. ∎

Following [13], we derive a descent property for the iterates.

LEMMA 3.3. *Let Assumption 3.1 hold, set $\bar{\rho} \in (\rho, 2\rho]$, and choose $\alpha_t \in (0, 1/\bar{\rho}]$, for any $t \geq 0$. Then, the following inequality holds:*

$$\mathbb{E}^t_{U,\xi}\left[\|x_{t+1} - \hat{x}_t\|_2^2\right] \leq \|x_t - \hat{x}_t\|_2^2 + 4(n^2+2n)\alpha_t^2 L_{f,0}^2 - 2\alpha_t(\bar{\rho} - \rho)\|x_t - \hat{x}_t\|_2^2,$$

where $\mathbb{E}^t_{U,\xi}\left[\cdot\right] \equiv \mathbb{E}_{U,\xi}\left[\cdot | U_{t-1}, \xi_{t-1}, \ldots, U_0, \xi_0\right]$.

Proof. We have

$$
\begin{aligned}
&\mathbb{E}^t_{U,\xi}\left[\|x_{t+1}-\hat{x}_t\|_2^2\right]\\
&\quad= \mathbb{E}^t_{U,\xi}\left[\left\|\mathbf{prox}_{\alpha_t r}\left(x_t-\alpha_t G\left(x_t,U_t,\xi_t\right)\right)-\mathbf{prox}_{\alpha_t r}\left(\alpha_t\bar{\rho}x_t-\alpha_t\nabla f_\mu(\hat{x}_t)+\delta_t\hat{x}_t\right)\right\|_2^2\right]\\
&\quad\leq \mathbb{E}^t_{U,\xi}\left[\|(x_t-\alpha_t G\left(x_t,U_t,\xi_t\right))-(\alpha_t\bar{\rho}x_t-\alpha_t\nabla f_\mu(\hat{x}_t)+\delta_t\hat{x}_t)\|_2^2\right]\\
&\quad= \delta_t^2\|x_t-\hat{x}_t\|_2^2-2\delta_t\alpha_t\mathbb{E}^t_{U,\xi}\left[\langle x_t-\hat{x}_t, G\left(x_t,U_t,\xi_t\right)-\nabla f_\mu(\hat{x}_t)\rangle\right]\\
&\qquad+\alpha_t^2\mathbb{E}^t_{U,\xi}\left[\|G\left(x_t,U_t,\xi_t\right)-\nabla f_\mu(\hat{x}_t)\|_2^2\right]\\
&\quad\leq \delta_t^2\|x_t-\hat{x}_t\|_2^2-2\delta_t\alpha_t\left\langle x_t-\hat{x}_t,\nabla f_\mu(x_t)-\nabla f_\mu(\hat{x}_t)\right\rangle+4(n^2+2n)\alpha_t^2L_{f,0}^2\\
&\quad\leq \delta_t^2\|x_t-\hat{x}_t\|_2^2+2\delta_t\alpha_t\rho\|x_t-\hat{x}_t\|_2^2+4(n^2+2n)\alpha_t^2L_{f,0}^2\\
&\quad= \left(1-\left(2\alpha_t(\bar{\rho}-\rho)+\alpha_t^2\bar{\rho}(2\rho-\bar{\rho})\right)\right)\|x_t-\hat{x}_t\|_2^2+4(n^2+2n)\alpha_t^2L_{f,0}^2,
\end{aligned}
$$

where the first equality follows from Lemma 3.2, the first inequality follows from non-expansiveness of the proximal operator (e.g. see [44, Theorem 12.12]), the second inequality follows from the triangle inequality and (2.3), while the third inequality follows from weak convexity of f_μ (see Proposition 2.2). Since $\bar{\rho}\leq 2\rho$, the result follows. ◻

We can now establish the convergence rate of Algorithm Z-ProxSG, in terms of the magnitude of the gradient of the Moreau envelope of the surrogate problem's objective function.

THEOREM 3.4. *Let Assumption 3.1 hold. Let also $\{x_t\}_{t=0}^T$ be the sequence of iterates produced by Algorithm Z-ProxSG, with x_{t^*} being the point that the algorithm returns. For any $t\geq 0$, $\mu>0$, and for any $\bar{\rho}\in(\rho,2\rho]$, it holds that*

$$
\begin{aligned}
\mathbb{E}_{U,\xi}\left[\phi_\mu^{1/\bar{\rho}}(x_{t+1})\right]\leq \mathbb{E}_{U,\xi}\left[\phi_\mu^{1/\bar{\rho}}(x_t)\right]-\frac{\alpha_t(\bar{\rho}-\rho)}{\bar{\rho}}\mathbb{E}_{U,\xi}\left[\left\|\nabla\phi_\mu^{1/\bar{\rho}}(x_t)\right\|_2^2\right]\\
+2(n^2+2n)\bar{\rho}\alpha_t^2L_{f,0}^2,
\end{aligned}
\tag{3.1}
$$

and x_{t^*} satisfies

$$
\mathbb{E}_{U,\xi}\left[\left\|\nabla\phi_\mu^{1/\bar{\rho}}(x_{t^*})\right\|_2^2\right]\leq\frac{\bar{\rho}}{\bar{\rho}-\rho}\frac{\left(\phi_\mu^{1/\bar{\rho}}(x_0)-\min_x\ \phi_\mu(x)\right)+2(n^2+2n)\bar{\rho}L_{f,0}^2\sum_{t=0}^T\alpha_t^2}{\sum_{t=0}^T\alpha_t}.
\tag{3.2}
$$

In particular, letting $\bar{\rho}=2\rho$, $\Delta\geq\phi_\mu^{1/\bar{\rho}}(x_0)-\min_x\ \phi_\mu(x)$, and setting

$$
\alpha_t=\frac{1}{2}\min\left\{\frac{1}{\rho},\sqrt{\frac{\Delta}{(n^2+2n)\rho L_{f,0}^2(T+1)}}\right\},
\tag{3.3}
$$

in Algorithm Z-ProxSG, yields:

$$
\mathbb{E}_{U,\xi}\left[\left\|\nabla\phi_\mu^{1/(2\rho)}(x_{t^*})\right\|_2^2\right]\leq 8\max\left\{\frac{\Delta\rho}{T+1},L_{f,0}\sqrt{\frac{\Delta\rho n(n+2)}{T+1}}\right\}.
\tag{3.4}
$$

Proof. Using the definition of the Moreau envelope, we have

$$
\begin{aligned}
\mathbb{E}^t_{U,\xi}\left[\phi_\mu^{1/\bar{\rho}}(x_{t+1})\right]&\leq\mathbb{E}^t_{U,\xi}\left[\phi_\mu(\hat{x}_t)+\frac{\bar{\rho}}{2}\|\hat{x}_t-x_{t+1}\|_2^2\right]\\
&\leq\phi_\mu(\hat{x}_t)+\frac{\bar{\rho}}{2}\left[\|x_t-\hat{x}_t\|_2^2+4(n^2+2n)\alpha_t^2L_{f,0}^2-2\alpha_t(\bar{\rho}-\rho)\|x_t-\hat{x}_t\|_2^2\right]\\
&=\phi_\mu^{1/\bar{\rho}}(x_t)+\bar{\rho}\left[2(n^2+2n)\alpha_t^2L_{f,0}^2-\alpha_t(\bar{\rho}-\rho)\|x_t-\hat{x}_t\|_2^2\right],
\end{aligned}
$$

where the second inequality follows from Lemma 3.3, and the equality follows from the definition of $\hat{x}_t$. Then, (3.1) is derived by taking the expectation with respect to the filtration (all the data observed so far, i.e. $U_{t-1}, \xi_{t-1}, \ldots, U_0, \xi_0$). Inequality (3.2) can be obtained as in [13, Section 3], by rearranging and utilizing the closed form of the gradient of the associated Moreau envelope.

Finally, by setting α_t as in (3.3), separating cases, and plugging the respective expressions in (3.2), yields (3.4) and completes the proof. ∎

The previous theorem provides an $\mathcal{O}\left(\sqrt{n}\epsilon^{-4}\right)$ convergence rate of Algorithm Z-ProxSG for finding an ϵ-stationary point of the Moreau envelope corresponding to (P_μ), i.e. $\phi_\mu^{1/(2\rho)}$. Let us notice that in the case where f is a convex function we can specialize Theorem 3.4 and obtain an $\mathcal{O}(\sqrt{n}\epsilon^{-2})$ convergence rate (noticing that any convex function is also ρ-weakly convex for any $\rho > 0$). This can be done by following the developments in [13, Section 4.1]. However, this is omitted for brevity of exposition.

In what follows, we would like to assess the quality of such a solution for the original problem (P). To that end, we will utilize Lemma 2.6. Before we proceed, let us provide certain well–known properties of the Moreau envelope, which indicate that it serves as a measure of closeness to optimality. We can observe (see [13, Section 2.2]) that for any $x \in \mathbb{R}^n$, and $\hat{x} \coloneqq \mathbf{prox}_{\lambda\phi_\mu}(x)$, the following hold:

$$\|\hat{x} - x\|_2 = \lambda \left\|\nabla \phi_\mu^\lambda(x)\right\|_2, \quad \phi_\mu\left(\hat{x}\right) \leq \phi_\mu(x), \quad \text{dist}\left(0; \partial\phi_\mu(\hat{x})\right) \leq \left\|\nabla \phi_\mu^\lambda(x)\right\|_2,$$

where, given any closed set $\mathcal{A} \subset \mathbb{R}^n$, $\text{dist}\left(z; \mathcal{A}\right) \coloneqq \inf_{z' \in \mathcal{A}} \|z - z'\|_2$. In other words, a near-stationary point of $\phi_\mu^{1/(2\rho)}$ is close to a near-stationary point of ϕ_μ. We expect that if $\mathbb{E}_{U,\xi}\left[\left\|\nabla \phi_\mu^{1/\bar{\rho}}(x_{t^*})\right\|_2\right] \leq \epsilon$, for some small $\epsilon > 0$, then there will exist a small $\delta(\epsilon) > 0$ such that $\mathbb{E}_{U,\xi}\left[\text{dist}\left(0, \partial\phi_\mu(x_{t^*})\right)\right] \leq \delta(\epsilon)$. Indeed, this is a standard assumption used in the literature (e.g. see [13, 30, 28]). The direct relation between δ and ϵ is not known in general, but in some cases this can be measured. For example, if $\partial\phi_\mu$ is a sub-Lipschitz continuous mapping (see [44, Definition 9.27]) or if r is an indicator function to a compact convex set (see [27]), then we obtain that $\delta = \mathcal{O}(\epsilon)$.

In what follows, assuming that $\mathbb{E}_{U,\xi}\left[\text{dist}\left(0, \partial\phi_\mu(x_{t^*})\right)\right] \leq \delta$, for some small $\delta > 0$, we show that $\mathbb{E}_{U,\xi}\left[\left\|\nabla \phi^{1/\bar{\rho}}(x_{t^*})\right\|_2^2\right] \leq \mathcal{O}\left(\delta^2 + \sqrt{n}\mu\right)$. To that end, in the following lemma we relate the Moreau envelope of the original problem's objective function ϕ^λ to the surrogate ϕ_μ in (P_μ).

Lemma 3.5. *Let Assumption 3.1 hold. Given any $x \in \mathbb{R}^n$, any $\bar{\rho} \in (\rho, 2\rho]$, and any $\mu > 0$, we have that*

$$\langle x - \tilde{x}, v_\mu \rangle \geq \frac{\bar{\rho} - \rho}{\bar{\rho}^2} \left\|\nabla \phi^{1/\bar{\rho}}(x)\right\|_2^2 - 2\mu L_{f,0} n^{\frac{1}{2}},$$

where $\tilde{x} \coloneqq \mathbf{prox}_{\bar{\rho}^{-1}\phi}(x)$, $\phi^{1/\bar{\rho}}$ is the Moreau envelope of ϕ in (P), and $v_\mu \in \partial\phi_\mu(x)$.

Proof. See Appendix A.2. ∎

Theorem 3.6. *Let Assumption 3.1 hold. Let x_δ be any δ-stationary point of problem (P_μ), that is, there exists $v_\mu \in \partial\phi_\mu(x_\delta)$, such that $\|v_\mu\|_2 \leq \delta$ (equivalently, $\text{dist}\left(0, \partial\phi_\mu(x_\delta)\right) \leq \delta$). Given any $\bar{\rho} \in (\rho, 2\rho]$, and any $\mu > 0$, we have that $|\phi\left(x_\delta\right) - \phi_\mu\left(x_\delta\right)| \leq \mu L_{f,0} n^{\frac{1}{2}}$. Moreover,*

$$\left\|\nabla \phi^{1/\bar{\rho}}(x_\delta)\right\|_2^2 \leq \frac{\bar{\rho}^2}{\bar{\rho} - \rho}\left(\frac{\delta^2}{\bar{\rho} - \rho} + 4\mu L_{f,0} n^{\frac{1}{2}}\right).$$

In particular, assuming that $\mathbb{E}_{U,\xi}\left[\operatorname{dist}\left(0, \partial\phi_\mu(x_{t^*})\right)\right] \leq \delta$*, where* x_{t^*} *is returned by Algorithm Z-ProxSG, we obtain that*

$$\mathbb{E}_{U,\xi}\left[\left\|\nabla\phi^{1/\bar{\rho}}(x_{t^*})\right\|_2^2\right] \leq \frac{\bar{\rho}^2}{\bar{\rho}-\rho}\left(\frac{\delta^2}{\bar{\rho}-\rho} + 4\mu L_{f,0} n^{\frac{1}{2}}\right).$$

Proof. The first part of the lemma follows immediately from the definition of ϕ_μ and Lemma 2.6.

From Lemma 3.5, we have that

$$\langle x_\delta - \tilde{x}_\delta, v_\mu\rangle \geq \frac{\bar{\rho}-\rho}{\bar{\rho}^2}\left\|\nabla\phi^{1/\bar{\rho}}(x_\delta)\right\|_2^2 - 2\mu L_{f,0} n^{\frac{1}{2}}, \tag{3.5}$$

where $\tilde{x}_\delta \coloneqq \mathbf{prox}_{\bar{\rho}^{-1}\phi}(x_\delta)$. From the triangle inequality, we obtain

$$\left\|\nabla\phi^{1/\bar{\rho}}(x_\delta)\right\|_2^2 - \frac{\delta\bar{\rho}}{\bar{\rho}-\rho}\left\|\nabla\phi^{1/\bar{\rho}}(x_\delta)\right\|_2 - \frac{2\bar{\rho}^2\mu L_{f,0} n^{\frac{1}{2}}}{\bar{\rho}-\rho} \leq 0,$$

where we used the definition of $\tilde{x}_\delta$, the expression of the gradient of $\phi^{1/\bar{\rho}}(x_\delta)$, and the assumption that $\|v_\mu\|_2 \leq \delta$. For ease of presentation, we introduce some notation. Let $u \coloneqq \left\|\nabla\phi^{1/\bar{\rho}}(x_\delta)\right\|_2$, $\beta \coloneqq -\frac{\delta\bar{\rho}}{\bar{\rho}-\rho}$, and $\gamma \coloneqq -\frac{2\bar{\rho}^2\mu L_{f,0} n^{\frac{1}{2}}}{\bar{\rho}-\rho}$. We proceed by finding an upper bound for u, so that the previous inequality is satisfied. This is trivial, since we can equate this inequality to zero, and find the most-positive solution of the quadratic equation in u. Indeed, it is easy to see that

$$u \leq \frac{1}{2}\left(-\beta + \sqrt{\beta^2 - 4\gamma}\right).$$

Thus we easily obtain $u^2 \leq \left(\beta^2 - 2\gamma\right)$. The first bound then follows immediately by plugging the values of β and γ.

Finally, by assuming that $\mathbb{E}_{U,\xi}\left[\operatorname{dist}\left(0, \partial\phi_\mu(x_{t^*})\right)\right] \leq \delta$, substituting x_{t^*} in (3.5), taking total expectations and repeating the previous analysis, yields the second bound and completes the proof. ∎

Remark 3.7. Let us notice that the convergence rate in Theorem 3.4 is given in terms of the expected squared gradient norm of the surrogate Moreau envelope evaluated at the output of Algorithm Z-ProxSG, that is $\mathbb{E}_{U,\xi}\left[\left\|\nabla\phi_\mu^{1/\bar{\rho}}(x_{t^*})\right\|_2^2\right]$. This is in line with the results presented in [30], however, the authors of the aforementioned paper did not investigate the error introduced by considering the surrogate problem. In this paper, we attempted to do this in Theorem 3.6. Ideally, we would like to provide a rate on $\mathbb{E}_{U,\xi}\left[\left\|\nabla\phi^{1/\bar{\rho}}(x_{t^*})\right\|_2^2\right]$. In the special cases where r is an indicator function to a compact convex set or $\partial\phi$ is a sub-Lipschitz mapping, this can be done easily (e.g. see [27, Section 6.4.2]). In the general case, and without additional restrictive assumption (as in [37]), we are able to show that any near-stationary point for the surrogate problem is near-stationary for the Moreau envelope of the original function, with the approximation improving for smaller values of μ. Thus, assuming that x_{t^*} is near-stationary in expectation for the surrogate problem (P_μ), we were able to show that it will be near-stationary in expectation for the Moreau envelope corresponding to (P).

4. Numerical results. In this section we provide numerical evidence for the effectiveness of the proposed approach. Firstly, we run the method on certain phase retrieval instances taken from [13] and compare the proposed zeroth-order approach, outlined in Algorithm Z-ProxSG, against the double smoothing zeroth-order proximal stochastic gradient method analyzed in [30], a uniform smoothing zeroth-order method (e.g. see [37]), the simultaneous perturbation stochastic approximation method (originally proposed in [49]), as well as the stochastic sub-gradient method proposed and analyzed in [13], noting that the latter method is significantly more difficult to employ (and implement) in the general case, since it assumes knowledge of sub-gradient information. In order to obtain a meaningful comparison, all zeroth-order schemes are using a constant step-size and constant smoothing parameter. For completeness, the four algorithms used in our comparison are outlined in Algorithm DSZ-ProxSG, UniZ-ProxSG, SPSA, and ProxSSG, respectively. Next, we verify that the proposed approach performs almost identically to the method outlined in [30], while being easier to tune and analyze (and additionally requiring n less flops per iteration).

Subsequently, we employ the proposed algorithm for the important task of tuning the parameters of optimization algorithms in order to obtain good and consistent behaviour for a wide range of optimization problems. We note that this problem can only be tackled by zeroth-order schemes, since there is no availability of first-order information. In particular, we employ a proximal alternating direction method of multipliers (pADMM) for the solution of PDE-constrained optimization instances. It is well-known that the behaviour of ADMM is heavily affected by the choice of its penalty parameter, and thus, we employ Algorithm Z-ProxSG in order to find a nearly optimal value (in a sense to be described) for this parameter that allows the method to behave well for similar (out-of-sample) PDE-constrained optimization instances. To our knowledge, the heuristic model proposed for achieving this task is novel and highly effective.

The code is written in MATLAB and can be found on GitHub [1]. The experiments were run on MATLAB 2019a, on a PC with a 2.2GHz Intel core i7 processor (hexa-core), 16GM RAM, using the Windows 10 operating system.

Algorithm DSZ-ProxSG Double Smoothing Z-ProxSG

Input: $x_0 \in \mathrm{dom}(r)$, a sequence $\{\alpha_t\}_{t\geq 0} \subset \mathbb{R}_+$, $\mu_1 \geq 2\mu_2 > 0$, and $T > 0$.
for $(t = 0, 1, 2, \ldots, T)$ **do**
Sample ξ_t, $U_{t,1}$, $U_{t,2} \sim \mathcal{N}(0_n, I_n)$, and set

$$x_{t+1} = \mathbf{prox}_{\alpha_t r}\left(x_t - \alpha_t G\left(x_t, U_{t,1}, U_{t,2}, \xi_t\right)\right),$$

where

$$G\left(x_t, U_{t,1}, U_{t,2}, \xi_t\right) = \mu_2^{-1}\left(F\left(x_t + \mu_1 U_{t,1} + \mu_2 U_{t,2}, \xi_t\right) - F(x_t + \mu_1 U_{t,1}, \xi_t)\right) U_{t,2}.$$

end for

4.1. Phase retrieval. Let us first focus on the solution of phase retrieval problems. Following [13], we generate standard Gaussian measurements $a_i \sim \mathcal{N}(0, I_d)$ for $i = 1, \ldots, m$, a target signal $\bar{x}$ as well as a starting point x_0 on the unit sphere. Then,

[1] https://github.com/spougkakiotis/Z-ProxSG

Algorithm UniZ-ProxSG Uniform Z-ProxSG

Input: $x_0 \in \mathrm{dom}(r) \subset \mathbb{R}^d$, a sequence $\{\alpha_t\}_{t\geq 0} \subset \mathbb{R}_+$, $\mu > 0$, and $T > 0$.
for $(t = 0, 1, 2, \ldots, T)$ **do**
Sample ξ_t, and U_t uniformly from the d-dimensional ball, and set

$$x_{t+1} = \mathbf{prox}_{\alpha_t r}\left(x_t - \alpha_t G\left(x_t, U_t, \xi_t\right)\right),$$

where

$$G\left(x_t, U_t, \xi_t\right) = \frac{d}{\mu}\left(F\left(x_t, \xi_t\right) - F(x_t + \mu U_t, \xi_t)\right) U_t.$$

end for

Algorithm SPSA Simultaneous Perturbation Stochastic Approximation

Input: $x_0 \in \mathrm{dom}(r)$, a sequence $\{\alpha_t\}_{t\geq 0} \subset \mathbb{R}_+$, $\mu_1 \geq 2\mu_2 > 0$, and $T > 0$.
for $(t = 0, 1, 2, \ldots, T)$ **do**
Sample ξ_t, and U_t from a d-dimensional Bernoulli distribution, and set

$$x_{t+1} = \mathbf{prox}_{\alpha_t r}\left(x_t - \alpha_t G\left(x_t, U_t, \xi_t\right)\right),$$

with

$$G\left(x_t, U_t, \xi_t\right) = \frac{F\left(x_t + \mu U_t, \xi_t\right) - F(x_t - \mu U_t, \xi_t)}{2\mu U_t},$$

where the division is component-wise.
end for

Algorithm ProxSSG Proximal Stochastic Sub-Gradient

Input: $x_0 \in \mathrm{dom}(r)$, a sequence $\{\alpha_t\}_{t\geq 0} \subset \mathbb{R}_+$, and $T > 0$.
for $(t = 0, 1, 2, \ldots, T)$ **do**
Sample ξ_t, and set

$$x_{t+1} = \mathbf{prox}_{\alpha_t r}\left(x_t - \alpha_t G\left(x_t, \xi_t\right)\right),$$

where $G\left(x_t, \xi_t\right) \in \partial F(x_t, \xi_t)$.
end for

by setting $b_i = \langle a_i, \bar{x}\rangle^2$, for $i = 1, \ldots, m$, we want to solve

$$\min_{x\in\mathbb{R}^d} f(x) = \frac{1}{m}\sum_{i=1}^{m}\left|\langle a_i, x\rangle^2 - b_i\right|.$$

As discussed in [13], this is a weakly convex optimization problem. We attempt to solve it using Algorithms Z-ProxSG, DSZ-ProxSG, UniZ-ProxSG, SPSA, and ProxSSG. For this specific instance, we can explicitly compute the sub-gradient appearing in Algorithm ProxSSG. Specifically, as shown in [13, Section 5.1], the subdifferential of the function $f_i(x) \coloneqq |\langle a_i, x\rangle^2 - b_i|$ reads

$$\partial f_i(x) = 2\langle a_i, x\rangle \cdot \begin{cases} \mathrm{sign}\left(\langle a_i, x\rangle^2 - b_i\right), & \text{if } \langle a_i, x\rangle \neq 0, \\ [-1, 1], & \text{otherwise} \end{cases}.$$

At each iteration of Algorithm ProxSSG we choose the sub-gradient that yields the highest objective value reduction.

Before proceeding with the experiments, let us discuss some implementation details. Each of the tested algorithms is heavily affected by the choice of the step-size α_t. We choose this parameter to be constant. For Algorithms Z-ProxSG, DSZ-ProxSG, UniZ-ProxSG, and SPSA, by loosely following the theory in Section 3, we set it to $\alpha_t = \frac{1}{2d\sqrt{T}}$ for all $t \geq 0$. Similarly, for Algorithm ProxSSG, following [13, Section 3], we set $\alpha_t = \frac{1}{2\sqrt{T}}$. Finally, Algorithms Z-ProxSG, UniZ-ProxSG, and SPSA are quite robust with respect to the choice of the smoothing parameter μ (or $\mu_1,\ \mu_2$, for Algorithm DSZ-ProxSG). For Algorithms Z-ProxSG, UniZ-ProxSG, and SPSA this was set to $\mu = 5 \cdot 10^{-10}$. From Theorem 3.6 we observe that the smaller the value of μ, the better the quality of the obtained solution (in terms of closeness to a stationary point of the Moreau envelope of the objective function). Indeed, there is no "optimal" value for μ and hence we set it to an as small as possible value, considering numerical accuracy issues that can arise due to finite machine precision. For Algorithm DSZ-ProxSG, by loosely following the theory in [16, Section 2.2], we set $\mu_1 = 5 \cdot 10^{-7},\ \mu_2 = 5 \cdot 10^{-10}$. Notice that we enforce $\mu = \mu_2$ in order to observe a comparable numerical behaviour between all zeroth-order schemes.

We set up 6 optimization problems, with varying sizes (d, m). In every case, the maximum number of iterations is set as $T = 2 \cdot 10^3 \cdot m$. The random seed of MATLAB was set to *"shuffle"*, which is initiated based on the current time. For each pair of sizes we produce 15 instances and run each of the five methods for T iterations. In Figure 1, we present the average convergence profiles with 95% confidence intervals for each method.

We can draw several useful observations from Figure 1. Firstly, while the convergence of the zeroth-order schemes is slower, as compared to the convergence of the sub-gradient scheme (as we expected from the theory), the obtained solutions are comparable for all algorithms. On the other hand, all zeroth-order schemes have a very similar behaviour, which was expected as we used similar values for the smoothing parameters. Let us notice that the theory in Section 3.1 can easily be altered to apply for Algorithm UniZ-ProxSG, since the Gaussian and the uniform smoothing techniques are very similar (see, for example, the analysis in [16]). Algorithm SPSA seems to behave equally well, compared to the other zeroth-order schemes, however, no convergence analysis is available in the literature for problems of the form of (P). Standard convergence analyses for SPSA are available for (stochastic) convex programming instances, allowing adaptive choices for the step-size α_t as well as the smoothing parameter μ. However, the adaptive choices proposed in [48] for convex programming did not deliver convergence for the phase retrieval instances solved here, thus we tuned this algorithm in the same way we tuned all the other zeroth-order schemes. In order to verify that Algorithms Z-ProxSG and DSZ-ProxSG behave seemingly identically even if we tune the ratio μ_1/μ_2, we set $(d, m) = (40, 60)$ and run the two zeroth-order methods using various values of (μ_1, μ_2), always ensuring that $\mu = \mu_2$. The results, which are averaged over 15 randomly generated instances, are reported in Figure 2.

We note that the authors in [16] show that for convex programming instances a proper tuning of the ratio μ_1/μ_2 can lead to a better convergence rate for the double-smoothing as compared to the single smoothing, in terms of its dependence on the dimension of the problem (noting that this has not been shown for weakly convex problems of the form of (P) in [30]). As we observe in Figure 2, varying this ratio

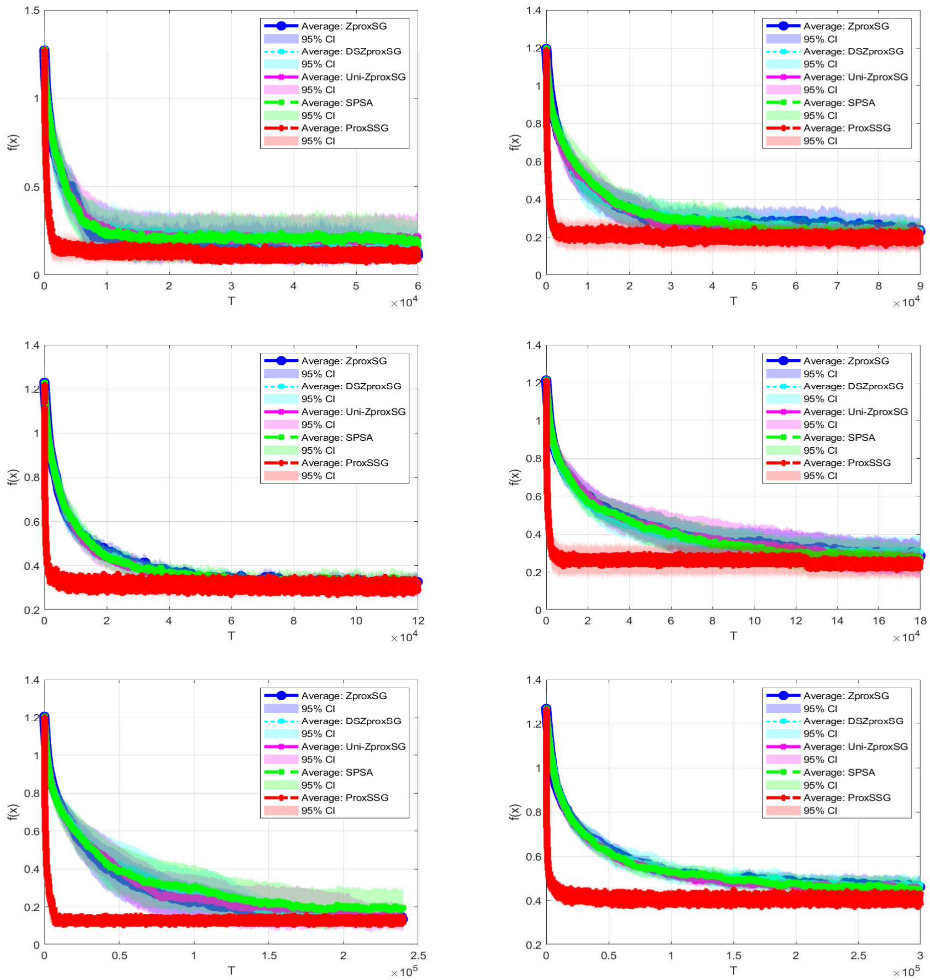

FIG. 1. *Convergence profiles for Z-ProxSG, DSZ-ProxSG, Uni-ZproxSG, SPSA and ProxSSG: average objective function value (lines) and 95% confidence intervals (shaded regions) vs number of iterations. The upper row corresponds, from left to right, to* $(d, m) = (10, 30)$, $(20, 45)$. *The middle row corresponds, from left to right, to* $(d, m) = (40, 60)$, $(35, 90)$. *The lower row corresponds, from left to right, to* $(d, m) = (30, 120)$, $(80, 150)$.

does not seem to have any actual effect in practice, since we observe that for a wide range of values for μ_1/μ_2 the double-Gaussian smoothing method behaves seemingly identically.

Notice that we could obtain better results by extensively tuning α_t and T for each instance, however, we provided general values that seem to exhibit a very consistent behaviour for all of the presented schemes.

4.2. Hyper-parameter tuning for optimization methods. Next, we consider the problem of tuning hyper-parameters of optimization algorithms, so as to improve their robustness and efficiency over a chosen set of optimization instances. The discussion in this section will be restricted to the case of an alternating direction method of multipliers (see [9] for an introductory review of ADMMs), although we

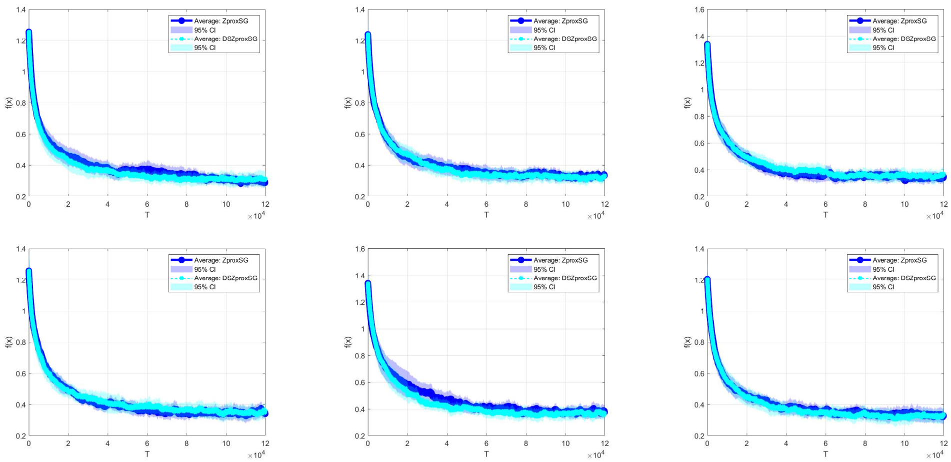

Fig. 2. *Convergence profiles for Z-ProxSG, DSZ-ProxSG: average objective function value (lines) and 95% confidence intervals (shaded regions) vs number of iterations, for* $(d, m) = (40, 60)$*. The upper row corresponds, from left to right, to* $(\mu_1, \mu_2) = (10^{-x}, 10^{-y})$*,* $x = 4, 5, 6$*,* $y = 7$*. The lower row corresponds, from left to right, to* $(\mu_1, \mu_2) = (10^{-x}, 10^{-y})$*,* $x = 6, 7, 8$*,* $y = 9$*. In each case we set* $\mu = \mu_2$*.*

conjecture that the same technique can be employed for tuning a much wider range of optimization methods.

4.2.1. Proximal ADMM for PDE-constrained optimization. In this section, we are interested in the solution of optimization problems with partial differential equation (PDE) constraints via a proximal alternating direction method of multipliers (pADMM). We note that various other applications would be suitable for the presented method, however, we restrict the problem pool for ease of presentation.

We consider optimal control problems of the following form:

$$
\begin{aligned}
\min_{\mathrm{y},\mathrm{u}} \quad & \mathrm{J}\left(\mathrm{y}(\boldsymbol{x}), \mathrm{u}(\boldsymbol{x})\right), \\
\text{s.t.} \quad & \mathrm{Dy}(\boldsymbol{x}) - \mathrm{u}(\boldsymbol{x}) = \mathrm{g}(\boldsymbol{x}), \\
& \mathrm{u_a}(\boldsymbol{x}) \le \mathrm{u}(\boldsymbol{x}) \le \mathrm{u_b}(\boldsymbol{x}),
\end{aligned} \tag{4.1}
$$

where $(\mathrm{y}, \mathrm{u}) \in \mathcal{H}_1(\mathrm{K}) \times \mathcal{L}_2(\mathrm{K})$, $\mathrm{J}\left(\mathrm{y}(\boldsymbol{x}), \mathrm{u}(\boldsymbol{x})\right)$ is a convex functional defined as

$$
\mathrm{J}\left(\mathrm{y}(\boldsymbol{x}), \mathrm{u}(\boldsymbol{x})\right) \coloneqq \frac{1}{2}\|\mathrm{y} - \bar{\mathrm{y}}\|^2_{\mathcal{L}_2(\mathrm{K})} + \frac{\beta_1}{2}\|\mathrm{u}\|^2_{\mathcal{L}_1(\mathrm{K})} + \frac{\beta_2}{2}\|\mathrm{u}\|^2_{\mathcal{L}_2(\mathrm{K})}, \tag{4.2}
$$

D denotes a linear differential operator, $\boldsymbol{x}$ is a 2-dimensional spatial variable, and $\beta_1,\ \beta_2 \ge 0$ denote the regularization parameters of the control variable.

The problem is considered on a given compact spatial domain $K \subset \mathbb{R}^2$ with boundary ∂K, and is equipped with Dirichlet boundary conditions. The algebraic inequality constraints are assumed to hold a.e. on K. We further note that $\mathrm{u_a}$ and $\mathrm{u_b}$ are chosen as constants, although a more general formulation would be possible. In what follows, we consider two classes of state equations (i.e. the equality constraints in (4.1)): the Poisson's equation, as well as the convection–diffusion equation. For the Poisson optimal control, by following [40], we set the desired state as $\bar{\mathrm{y}} = \sin(\pi x_1)\sin(\pi x_2)$. For the convection-diffusion, which reads as $-\epsilon\Delta\mathrm{y} + \mathrm{w}\cdot\nabla\mathrm{y} = \mathrm{u}$, where w is the wind vector given by $\mathrm{w} = [2\mathrm{x}_2(1-\mathrm{x}_1)^2, -2\mathrm{x}_1(1-\mathrm{x}_2^2)]^\top$, we set the desired state as $\bar{\mathrm{y}} = \exp(-64((\mathrm{x}_1-0.5)^2+(\mathrm{x}_2-0.5)^2))$ with zero boundary conditions (e.g. see [40, Section 5.2]). The diffusion coefficient ϵ is set as $\epsilon = 0.05$. In both cases, we set $K = (0,1)^2$, $\mathrm{u_a} = -2$, and $\mathrm{u_b} = 1.5$ (see [40]).

We solve problem (4.1) via a *discretize-then-optimize* strategy. We employ the Q1 finite element discretization implemented in IFISS[2] (see [19, 20]). This yields a sequence of ℓ_1-regularized convex quadratic programming problems of the following form:

$$\min_{x\in\mathbb{R}^n} c^\top x + \frac{1}{2}x^\top Q x + \|Dx\|_1 + \delta_{\mathcal{K}}(x), \qquad \text{s.t. } Ax = b, \tag{4.3}$$

where $A \in \mathbb{R}^{m\times n}$ models the linear constraints, $D \in \mathbb{R}^{n\times n}$ is a diagonal matrix, and $\mathcal{K}$ models the restrictions on the discretized control variables. We note that the discretization of the smooth part of the objective of problem (4.1) follows a standarad Galekrin approach (e.g. see [51]), while the $\mathcal{L}_1$ term is discretized by the *nodal quadrature rule* as in [47, 53] (which achieves a first-order convergence–see [53]).

We can reformulate problem (4.3) by introducing an auxiliary variable $w \in \mathbb{R}^n$, as follows

$$\min_{x\in\mathbb{R}^n, w\in\mathbb{R}^n} c^\top x + \frac{1}{2}x^\top Q x + \|Dw\|_1 + \delta_{\mathcal{K}}(w), \qquad \text{s.t. } Ax = b, \quad w - x = 0. \tag{4.4}$$

Given a penalty $\sigma > 0$, we associate the following augmented Lagrangian to (4.4)

$$\begin{aligned} L_\sigma(x, w, y_1, y_2) &\coloneqq c^\top x + \frac{1}{2}x^\top Q x + g(w) + \delta_{\mathcal{K}}(w) - y_1^\top(Ax - b) - y_2^\top(w - x) \\ &\qquad + \frac{\sigma}{2}\|Ax - b\|^2 + \frac{\sigma}{2}\|w - x\|^2. \end{aligned}$$

Let an arbitrary positive definite matrix R_x be given, and assume the notation $\|x\|_{R_x}^2 = x^\top R_x x$. Also, given a convex set $\mathcal{K}$, let $\Pi_{\mathcal{K}}(\cdot)$ denote the Euclidian projection onto $\mathcal{K}$. We now provide (in Algorithm pADMM) a proximal ADMM for the approximate solution of (4.4).

Algorithm pADMM Proximal Alternating Direction Method of Multipliers

Input: $\sigma > 0$, $R_x \succ 0$, $\gamma \in \left(0, \frac{1+\sqrt{5}}{2}\right)$, $(x_0, w_0, y_{1,0}, y_{2,0}) \in \mathbb{R}^{3n+m}$.

for $(t = 0, 1, 2, \ldots)$ **do**

$w_{t+1} = \arg\min_w \{L_\sigma(x_t, w, y_{1,t}, y_{2,t})\} \equiv \Pi_{\mathcal{K}}\left(\mathbf{prox}_{\sigma^{-1}g}\left(x_t + \sigma^{-1}y_{2,t}\right)\right).$

$x_{t+1} = \arg\min_x \left\{L_\sigma(x, w_{t+1}, y_{1,t}, y_{2,t}) + \frac{1}{2}\|x - x_t\|_{R_x}^2\right\}.$

$y_{1,t+1} = y_{1,t} - \gamma\sigma(Ax_{t+1} - b).$

$y_{2,t+1} = y_{2,t} - \gamma\sigma(w_{t+1} - x_{t+1}).$

end for

We notice that under feasibility and convexity assumptions on (4.4), Algorithm pADMM is able to achieve global convergence potentially at a linear rate, assuming strong convexity (see [14]), even in cases where R_x is not positive definite [26]. Here we assume that R_x is positive definite, and we employ it as a means of reducing the

[2] https://personalpages.manchester.ac.uk/staff/david.silvester/ifiss/default.htm

memory requirements of Algorithm pADMM. More specifically, given some constant $\hat{\sigma} > 0$, such that $\hat{\sigma} I_n - \text{Off}(Q) \succ 0$, we define

$$R_x = \hat{\sigma} I_n - \text{Off}(Q),$$

where $\text{Off}(B)$ denotes the matrix with zero diagonal and off-diagonal elements equal to the off-diagonal elements of B. We note that this method was employed in [41] as a means of obtaining a starting point for a semi-smooth Newton-proximal method of multipliers, suitable for the solution of (4.3).

In the experiments to follow, Algorithm pADMM uses the zero vector as a starting point, while the step-size is set to the value $\gamma = 1.618$. The penalty parameter σ is given to the algorithm by the user, and this is later utilized to tune the method over an appropriate set of problem instances. We expect that different values for σ should be chosen when considering Poisson and convection-diffusion problems. Thus, in the following subsection we tune Algorithm pADMM for each of the two problem-classes separately.

4.2.2. Automated tuning: problem formulation and numerical results. Given a positive number k, we consider a general stochastic optimization problem of the following form

$$\text{(4.5)} \qquad \min_{\sigma \in \mathbb{R}} f(\sigma; k) \coloneqq \mathbb{E}\left[F(\sigma, \xi; k)\right] + \delta_{[\sigma_{\min}, \sigma_{\max}]}\left(\sigma\right), \qquad \xi \sim P,$$

where $f(\sigma; k) =$"expected residual reduction of Algorithm pADMM after k iterations, given the penalty parameter σ, for discretized problems of the form of (4.3) originating from a distribution P". We assume that $\xi \in \Xi \subset \mathbb{R}^d$, where a sample ξ is a specific problem instance of the form of (4.3). In particular, we consider two different tuning problems, and thus two different distributions $P_1,\ P_2$. Sampling either of the two distributions $P_1,\ P_2$ yields a problem of the form of (4.3) with arbitrary (but sensible) values for the regularization parameters $\beta_1,\ \beta_2 > 0$, as well as a randomly chosen (grid-based) problem size. For P_1, the linear constraints model the Poisson equation, while for P_2 the convection-diffusion equation. The values for the remaining problem parameters (i.e. control bounds, desired states, wind vector, and diffusion coefficient) are given in the previous subsection.

Remark 4.1. Notice that the choice of $f(\cdot; k)$ in (4.5) has multiple motivations. Firstly, by choosing a small value for k (e.g. 10 or 15), we can ensure that each run of Algorithm pADMM will not take excessive time (since one run of the algorithm corresponds to a sample-function evaluation within Algorithm Z-ProxSG). Additionally, the scale of $f(\cdot; k)$ is expected to be comparable for very different classes of problems. Indeed, assuming that Algorithm pADMM does not diverge (which could only happen if an infeasible instance was tackled), we expect that in most cases $0 \leq f(\cdot; k) \leq C$, where $C = \mathcal{O}(1)$ is a small positive value, irrespectively of the problem under consideration, since we measure the residual reduction. However, it should be noted that this is a heuristic. Indeed, finding the parameter value that yields the fastest residual reduction in the first k iterations does not necessarily yield an optimal convergence behaviour in the long-run. Nonetheless, we can always increase the value of k at the expense of a more expensive meta-tuning. In both cases considered here, this was not required.

Finally, we note that the constraints in (4.5) arise from prior information that we might have about the class of problems that we consider. It is well-known that very small or very large values for the penalty parameter of the ADMM tend to perform poorly (e.g. see the discussions in [9, Section 3.4.1.] or [50]). Thus, some limited preliminary experimentation can determine suitable values for $\sigma_{\min}$ and $\sigma_{\max}$ for each problem class that is considered. In the experiments to follow we set $\sigma_{\min} = 10^{-2}$ and $\sigma_{\max} = 10^2$.

In order to find an approximate solution to (4.5), we need to define a representative discrete training set from the space of optimization problems produced by P_1 (or P_2, respectively). To that end, we will use a discrete training set $\hat{\Xi} = \{\xi_1, \ldots, \xi_m\} \subset \Xi$, which yields the following problem

$$\min_{\sigma \in \mathbb{R}} f(\sigma; k) \coloneqq \frac{1}{m} \sum_{j=1}^{m} F(\sigma, \xi_j; k) + \delta_{[\sigma_{\min}, \sigma_{\max}]}(\sigma) . \tag{4.6}$$

Once an approximate solution to (4.6) is found, we can test its quality on out-of-sample PDE-constrained optimization instances. For both problem classes (i.e. Poisson and convection-diffusion optimal control), we construct 80 optimization instances. In particular, we define the sets

$$\mathcal{B}_1 \coloneqq \{0, 10^{-2}, 10^{-4}, 10^{-6}\},\ \mathcal{B}_2 \coloneqq \{0, 10^{-2}, 10^{-4}, 10^{-6}\},$$
$$\mathcal{M} \coloneqq \{(2^3+1)^2, (2^4+1)^2, (2^5+1)^2, (2^6+1)^2, (2^7+1)^2\},$$

where $\mathcal{B}_1$ ($\mathcal{B}_2$, respectively) contains potential values for β_1 (β_2, respectively), while $\mathcal{M}$ contains potential problem sizes. At each iteration t of Algorithm Z-ProxSG, we sample uniformly $\beta_{t,1} \in \mathcal{B}_1$, $\beta_{t,2} \in \mathcal{B}_2$, and $n_t \in \mathcal{M}$, and use the triple $\xi = (\beta_{t,1}, \beta_{t,2}, n_t)$ to generate an optimization instance. Then, $F(\cdot, \xi; k)$ can be evaluated by running Algorithm pADMM on this instance for k iterations and subsequently computing the residual reduction. In the following runs of Algorithm Z-ProxSG, we set $\mu = 5 \cdot 10^{-10}$, and $T = 200 \cdot m$, where $m = |\mathcal{B}_1| \cdot |\mathcal{B}_2| \cdot |\mathcal{M}| = 80$.

Poisson optimal control. Let us first consider Poisson optimal control problems. We apply Algorithm Z-ProxSG to find an approximate solution of (4.6), with $k = 15$. We choose σ^* as the last iteration of Algorithm Z-ProxSG, which in this case turned out to be $\sigma^* = 0.2778$. Then, in order to evaluate the quality of this penalty, we run Algorithm pADMM on 40 randomly-chosen out-of-sample Poisson optimal control problems for different penalty values $\sigma \in [\sigma_{\min}, \sigma_{\max}]$, including σ^*. In particular, in order to create out-of-sample instances, we define the sets

$$\hat{\mathcal{B}}_1 \coloneqq \{10^{-3}, 5 \cdot 10^{-3}, 10^{-5}, 5 \cdot 10^{-5}\},\ \hat{\mathcal{B}}_2 \coloneqq \{10^{-3}, 5 \cdot 10^{-3}, 10^{-5}, 5 \cdot 10^{-5}\},$$
$$\hat{\mathcal{M}} \coloneqq \{(2^3+1)^2, (2^4+1)^2, (2^5+1)^2, (2^6+1)^2, (2^7+1)^2, (2^8+1)^2\},$$

These correspond to 96 optimization instances, that were not used during the zeroth-order meta-tuning. The averaged convergence profiles (measuring the scaled residual versus the ADMM iteration) are summarized in Figure 3.

In Figure 3 we observe that out of the 6 different values for σ, Algorithm pADMM exhibits the most consistent behaviour when using the value that Algorithm Z-ProxSG suggested as "optimal". The next two best-performing values were $\sigma = 0.8$, $\sigma = 0.05$, and one can observe these are the ones closest to $\sigma^* = 0.2778$. Let us notice that the $y-$axis in Figure 3 only shows values less than 0.1. This was enforced for readability purposes.

Optimal control of the convection-diffusion equation. We now consider the optimal control of the convection-diffusion equation. As before, we apply Algorithm Z-ProxSG to find an approximate solution of (4.6), with $k = 15$. We choose σ^*

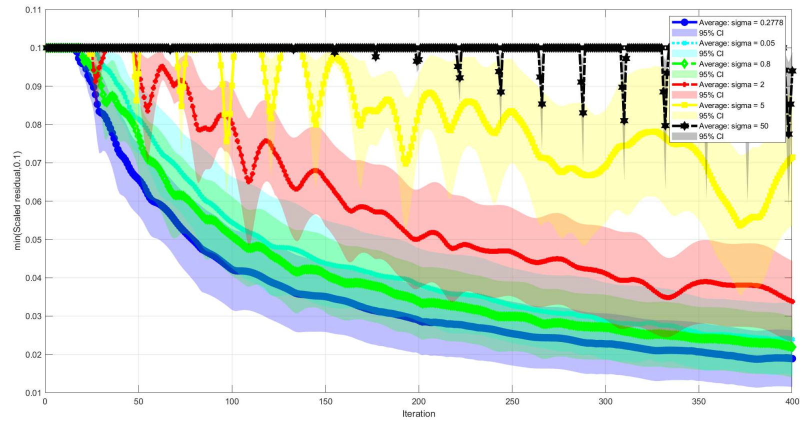

FIG. 3. *Convergence profiles for pADMM with varying penalty parameter σ: average residual reduction (lines) and 95% confidence intervals (shaded regions) vs number of pADMM iterations. The algorithm is run over 40 randomly selected (out-of-sample) Poisson optimal control problems.*

as the last iteration of Algorithm Z-ProxSG, which in this case turned out to be $\sigma^* = 5.7004$. We evaluate the quality of this penalty by running Algorithm pADMM on 40 randomly-chosen out-of-sample convection-diffusion optimal control problems for different penalty values $\sigma \in [\sigma_{\min}, \sigma_{\max}]$, including σ^*. As before these instances are created by sampling the previously defined sets $\hat{\mathcal{B}}_1$, $\hat{\mathcal{B}}_2$ and $\hat{\mathcal{M}}$. The averaged convergence profiles (measuring the scaled residual versus the ADMM iteration) are summarized in Figure 4.

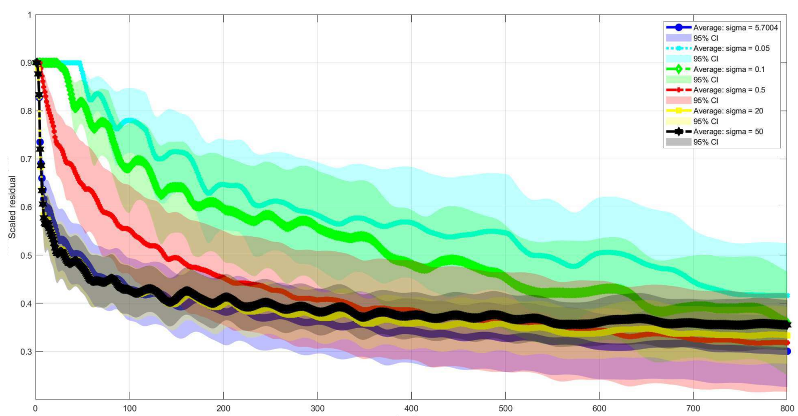

FIG. 4. *Convergence profiles for pADMM with varying penalty parameter σ: average residual reduction (lines) and 95% confidence intervals (shaded regions) vs number of pADMM iterations. The algorithm is run over 40 randomly selected (out-of-sample) convection-diffusion optimal control problems.*

Based on the results shown in Figure 4 we can observe that Algorithm Z-ProxSG is indeed able to find a value for σ that approximately minimizes the residual reduction of the ADMM during the first k iterations. However, as already noted, that this is not necessarily the optimal choice when running Algorithm pADMM for a much larger number of iterations. We expect that in many cases (e.g. as in the optimal control of the Poisson equation) the first few iterations of the ADMM are sufficient to predict the behaviour of the algorithm in later iterations. On the other hand, from the convection-diffusion instances we observe that a very steep residual reduction during the first ADMM iterations (e.g. observed when $\sigma = 50$ or $\sigma = 20$) does not necessarily result in the minimum achievable residual reduction after a large number of ADMM iterations. Of course this could be taken into account by increasing the value of k (e.g. the users might set it equal to the number of iterations that they are willing to let ADMM run for the specific application at hand), but it should be noted that this would result in more expensive sample-function evaluations of problem (4.5). Other heuristics could also improve the generalization performance of the model in (4.5) (such as employing different starting point strategies for the ADMM runs during the "training"). However, the focus of this paper prevents us from investigating this matter any further. Most importantly, in both problem classes, we were able to observe that Algorithm Z-ProxSG succeeds in finding an approximate solutions to (4.5), yielding efficient versions of Algorithm pADMM.

5. Conclusions.

In this paper we have derived and analyzed a zeroth-order proximal stochastic gradient method suitable for the solution of weakly convex stochastic optimization problems. We demonstrated that, under standard assumptions, the algorithm is guaranteed to converge to a near-stationary solution of the problem at a rate comparable to that achieved by similar sub-gradient schemes. The theoretical results were consistently verified numerically on certain phase-retrieval instances, supporting the viability of the proposed approach. Finally, we developed a novel heuristic model for the calculation of "optimal" hyper-parameters of optimization algorithms for an arbitrary given class of problems. Using the latter, we were able to showcase that the proposed zeroth-order algorithm can be efficiently employed for hyper-parameter tuning problems, yielding very promising results.

Appendix A. Appendix.

A.1. Proof of Lemma 3.2.

Proof. From the definition of $\hat{x}_t$ we have

$$\begin{aligned}\alpha_t\bar{\rho}\left(x_t - \hat{x}_t\right) \in \alpha_t\partial r\left(\hat{x}_t\right) + \alpha_t\nabla f_\mu(\hat{x}_t) &\Leftrightarrow \alpha_t\bar{\rho}x_t - \alpha_t\nabla f_\mu(\hat{x}_t) + \delta_t\hat{x}_t \in \hat{x}_t + \alpha_t\partial r\left(\hat{x}_t\right)\\ &\Leftrightarrow \hat{x}_t = \mathbf{prox}_{\alpha_t r}\left(\alpha_t\bar{\rho}x_t - \alpha_t\nabla f_\mu(x_t) + \delta_t\hat{x}_t\right).\end{aligned}$$

This completes the proof. ∎

A.2. Proof of Lemma 3.5.

Proof. Following [27, Lemma 5.2], we begin by noticing that for any $x_1,\ x_2 \in \mathbb{R}^n$ the following holds

$$\begin{aligned}\phi(x_1) - \phi(x_2) &= \phi_\mu(x_1) + \phi(x_1) - \phi_\mu(x_1) - \phi_\mu(x_2) - \phi(x_2) + \phi_\mu(x_2)\\ &\le \phi_\mu(x_1) - \phi_\mu(x_2) + 2\sup_{x\in\mathbb{R}^n}|\phi_\mu(x) - \phi(x)|\\ &\le \phi_\mu(x_1) - \phi_\mu(x_2) + 2\mu L_{f,0}n^{\frac{1}{2}},\end{aligned}$$

where the second inequality follows from (2.1). On the other hand, given $v_\mu \in \partial\phi_\mu(x_t)$, from ρ-weak convexity of $\phi_\mu(\cdot)$, and by utilizing Proposition 2.2, we obtain

$$\begin{aligned}\langle x_1 - x_2, v_\mu\rangle &\ge \phi_\mu(x_1) - \phi_\mu(x_2) - \frac{\rho}{2}\|x_1 - x_2\|_2^2\\ &\ge \phi(x_1) - \phi(x_2) - \frac{\rho}{2}\|x_1 - x_2\|_2^2 - 2\mu L_{f,0}n^{\frac{1}{2}},\end{aligned}$$

for any $x_1,\ x_2 \in \mathbb{R}^n$. By letting $x_1 = x$ and $x_2 = \tilde{x} \coloneqq \mathbf{prox}_{\bar{\rho}^{-1}\phi}(x)$, and by noting that $\bar{\rho} > \rho$, we obtain

$$\begin{aligned}\langle x - \tilde{x}, v_\mu\rangle &\geq \phi(x) - \phi(\tilde{x}) - \frac{\rho}{2}\|x - \tilde{x}\|_2^2 - 2\mu L_{f,0} n^{\frac{1}{2}} \\ &\equiv \phi(x) + \frac{\bar{\rho}}{2}\|x - x\|_2^2 - \Big(\phi(\tilde{x}) + \frac{\bar{\rho}}{2}\|\tilde{x} - x\|_2^2\Big) \\ &\quad + \frac{\bar{\rho} - \rho}{2}\|\tilde{x} - x\|_2^2 - 2\mu L_{f,0} n^{\frac{1}{2}}\end{aligned}$$

However, we know that the map $y \mapsto \big(\phi(y) + \frac{\bar{\rho}}{2}\|y - x\|_2^2\big)$ is strongly convex with parameter $\bar{\rho} - \rho$, and is minimized at $\tilde{x}$, and thus

$$\phi(x) + \frac{\bar{\rho}}{2}\|x - x\|_2^2 - \Big(\phi(\tilde{x}) + \frac{\bar{\rho}}{2}\|\tilde{x} - x\|_2^2\Big) \geq \frac{\bar{\rho} - \rho}{2}\|x - \tilde{x}\|_2^2.$$

Hence, we obtain

$$\begin{aligned}\langle x - \tilde{x}, v_\mu\rangle &\geq (\bar{\rho} - \rho)\|\tilde{x} - x\|_2^2 - 2\mu L_{f,0} n^{\frac{1}{2}} \\ &\equiv \frac{\bar{\rho} - \rho}{\bar{\rho}^2}\|\nabla \phi^{1/\bar{\rho}}(x)\|_2^2 - 2\mu L_{f,0} n^{\frac{1}{2}},\end{aligned}$$

where the last equivalence follows from the characterization of the gradient of the Moreau envelope, as well as the definition of $\tilde{x}_t$, and completes the proof. ∎

REFERENCES

[1] P. Alberto, F. Nogueira, H. Rocha, and L. N. Vicente, *Pattern search methods for user-provided points: Application to molecular geometry problems*, SIAM Journal on Optimization, 14 (2004), pp. 1216–1236, https://doi.org/10.1137/S1052623400377955.

[2] C. Audet and D. Orban, *Finding optimal algorithmic parameters using derivative-free optimization*, SIAM Journal on Optimization, 17 (2006), pp. 642–664, https://doi.org/10.1137/040620886.

[3] N. Baba, *Convergence of a random optimization method for constrained optimization problems*, Journal of Optimization Theory and Applications, 33 (1981), pp. 451–461, https://doi.org/10.1007/BF00935752.

[4] K. Balasubramanian and S. Gadhimi, *Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points*, Foundations of Computational Mathematics, 22 (2022), pp. 35–76, https://doi.org/10.1007/s10208-021-09499-8.

[5] K. Balasubramanian and S. Ghadimi, *Zeroth-order (non)-convex stochastic optimization via conditional gradient and gradient updates*, in Advances in Neural Information Processing Systems, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds., vol. 31, Curran Associates, Inc., 2018, https://proceedings.neurips.cc/paper/2018/file/36d7534290610d9b7e9abed244dd2f28-Paper.pdf.

[6] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, *Algorithms for hyper-parameter optimization*, in Advances in Neural Information Processing Systems, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, eds., vol. 24, Curran Associates, Inc., 2011, https://proceedings.neurips.cc/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf.

[7] J. Bergstra and Y. Bengio, *Random search for hyper-parameter optimization*, Journal of Machine Learning Research, 13 (2012), pp. 281–305, http://jmlr.org/papers/v13/bergstra12a.html.

[8] A. J. Booker, J. E. Dennis, P. D. Frank, D. B. Serafini, and V. Torczon, *Optimization using surrogate objectives on a helicopter test example*, Birkhäuser Boston, Boston, MA, 1998, pp. 49–58, https://doi.org/10.1007/978-1-4612-1780-0_3.

[9] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Foundations and Trends in Machine Learning, 3 (2010), pp. 1–122, https://doi.org/10.1561/2200000016.

[10] D. Calvetti, P. C. Hansen, and L. Reichel, *L-curve curvature bounds via Lanczos bidiagonalization*, Electronic Transactions on Numerical Analysis, 14 (2002), pp. 20–35.
[11] F. H. Clarke, *Generalized gradients and applications*, Transactions of the American Mathematical Society, 205 (1975), pp. 247–262.
[12] A. R. Conn, K. Scheinberg, and L. N. Vicent, *Introduction to Derivative-Free Optimization*, MOS-SIAM Series on Optimization, SIAM & Mathematical Optimization Society, Philadelphia, 2009, https://doi.org/10.1137/1.9780898718768.
[13] D. Davis and D. Drusvyatskiy, *Stocahstic model-based minimization of weakly convex functions*, SIAM Journal on Optimization, 29 (2019), pp. 207–239, https://doi.org/10.1137/18M1178244.
[14] W. Deng and W. Yin, *On the global and linear convergence of the generalized alternating direction method of multipliers*, Journal of Scientific Computing, 66 (2016), pp. 889–916, https://doi.org/10.1007/s10915-015-0048-x.
[15] D. Drusvyatskiy and C. Paquette, *Efficiency of minimizing compositions of convex functions and smooth maps*, Mathematical Programming, 178 (2019), pp. 503–558, https://doi.org/10.1007/s10107-018-1311-3.
[16] J. C. Duchi, M. I. Jordan, W. M. J., and A. Wibisono, *Optimal rates for zero-order convex optimization: The power of two function evaluations*, IEEE Transactions on Information Theory, 61 (2015), pp. 2788–2806, https://doi.org/10.1109/TIT.2015.2409256.
[17] D. Dvinskikh, V. Tominin, Y. Tominin, and A. Gasnikov, *Gradient-free optimization for non-smooth minimax problems with maximum value of adversarial noise*, 2022, https://arxiv.org/abs/arXiv:2202.06114.
[18] Y. C. Eldar and S. Mendelson, *Phase retrieval: Stability and recovery guarantees*, Applied and Computational Harmonic Analysis, 36 (2014), pp. 473–494, https://doi.org/10.1016/j.acha.2013.08.003.
[19] H. C. Elman, A. Ramage, and D. J. Silvester, *Algorithm 866: IFISS, a Matlab toolbox for modelling incompressible flow*, ACM Transactions on Mathematical Software, 33 (2007), p. 14, https://doi.org/10.1145/1236463.1236469.
[20] H. C. Elman, A. Ramage, and D. J. Silvester, *IFISS: A computational laboratory for investigating incompressible flow problems*, SIAM Review, 52 (2014), pp. 261–273, https://doi.org/10.1137/120891393.
[21] C. Fenu, L. Reichel, G. Rodriguez, and H. Sadok, *GCV for Tikhonov regularization by partial SVD*, BIT, 57 (2017), pp. 1019–1039, https://doi.org/10.1007/s10543-017-0662-0.
[22] M. Feurer and F. Hutter, *Hyperparameter optimization*, Springer International Publishing, 2019, pp. 3–33, https://doi.org/10.1007/978-3-030-05318-5_1.
[23] S. Ghadimi and G. Lan, *Stochastic first- and zeroth-order methods for nonconvex stochastic programming*, SIAM Journal on Optimization, 23 (2013), pp. 2341–2368, https://doi.org/10.1137/120880811.
[24] N. J. Higham, *Optimization by direct search in matrix computations*, SIAM Journal on Matrix Analysis and Applications, 14 (1993), pp. 317–333, https://doi.org/10.1137/0614023.
[25] R. J. and O. Teytaud, *Nevergrad - A gradient-free optimization platform.* https://GitHub.com/FacebookResearch/Nevergrad, 2018.
[26] F. Jiang, Z. Wu, and X. Cai, *Generalized ADMM with optimal indefinite proximal term for linearly constrained convex optimization*, Journal of Industrial & Management Optimization, 16 (2020), pp. 835–856, https://doi.org/10.3934/jimo.2018181.
[27] D. S. Kalogerias and W. B. Powell, *Zeroth-order stochastic compositional algorithms for risk-aware learning*, SIAM Journal on Optimization, 32 (2022), https://doi.org/10.1137/20M1315403.
[28] D. Kozak, C. Molinari, L. Rosasco, L. Tenorio, and S. Villa, *Zeroth order optimization with orthogonal random directions*, 2021, https://arxiv.org/abs/arXiv:2107.03941v2.
[29] H. Kumar, D. S. Kalogerias, G. J. Pappas, and A. Ribeiro, *Actor-only deterministic policy gradient via zeroth-order gradient oracles in action space*, in 2021 IEEE International Symposium on Information Theory (ISIT), 2021, pp. 1676–1681, https://doi.org/10.1109/ISIT45174.2021.9518023.
[30] V. Kungurtsev and F. Rinaldi, *A zeroth order method for stochastic weakly convex optimization*, Computational Optimization and Applications, 80 (2021), pp. 731–753, https://doi.org/10.1007/s10589-021-00313-3.
[31] S. Ling and T. Strohmer, *Self-calibration and biconvex compressive sensing*, Inverse Problems, 31 (2015), p. 115002, https://doi.org/10.1088/0266-5611/31/11/115002.
[32] J. Mairal, J. Ponce, G. Sapiro, A. Zisserman, and F. Bach, *Supervised dictionary learning*, in Advances in Neural Information Processing Systems, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, eds., vol. 21, Curran Associates, Inc., 2008, https:

//proceedings.neurips.cc/paper/2008/file/c0f168ce8900fa56e57789e2a2f2c9d0-Paper.pdf.

[33] C. Malivert, *Méthode de descente sur un fermé non convexe*, in Analyse non convexe (Pau, 1977), no. 60 in Mémoires de la Société Mathématique de France, Société mathématique de France, 1979, pp. 113–124, https://doi.org/10.24033/msmf.264.

[34] J. Matyas, *Random optimization*, Automation and Remote Control, 26 (1965), pp. 246–253.

[35] J. C. Meza and M. L. Martinez, *Direct search methods for the molecular conformation problem*, Journal of Computational Chemistry, 15 (1994), pp. 627–632, https://doi.org/10.1002/jcc.540150606.

[36] J.-J. Moreau, *Proximité et dualité dans un espace hilbertien*, Bulletin de la Société Mathématique de France, 93 (1965), pp. 273–299, https://doi.org/10.24033/bsmf.1625.

[37] P. Nazari, D. A. Tarzanagh, and G. Michailidis, *Adaptive first- and zeroth-order methods for weakly convex stochastic optimization problems*, 2020, https://arxiv.org/abs/arXiv:2005.09261v2.

[38] Y. Nesterov and V. Spokoiny, *Random gradient-free minimization of convex functions*, Foundations of Computational Mathematics, 17 (2017), pp. 527–566, https://doi.org/10.1007/s10208-015-9296-2.

[39] P. Patrinos and A. Bemporad, *Proximal Newton methods for convex composite optimization*, in 52nd IEEE Conference on Decision and Control, 2013, pp. 2358–2363, https://doi.org/10.1109/CDC.2013.6760233.

[40] J. W. Pearson, M. Porcelli, and M. Stoll, *Interior-point methods and preconditioning for PDE-constrained optimization problems involving sparsity terms*, Numerical Linear Algebra with Applications, 27 (2019), p. e2276, https://doi.org/10.1002/nla.2276.

[41] S. Pougkakiotis and J. Gondzio, *A semismooth Newton-proximal method of multipliers for ℓ_1-regularized convex quadratic programming*, 2022, https://arxiv.org/abs/arXiv:2201.10211.

[42] M. Pragliola, L. Calatroni, A. Lanza, and F. Sgallari, *Residual whiteness principle for automatic parameter selection in ℓ_1-ℓ_2 image super-resolution problems*, in Scale Space and Variational Methods in Computer Vision, A. Elmoataz, J. Fadili, Y. Quéau, J. Rabin, and L. Simon, eds., Springer International Publishing, 2021, pp. 476–488, https://doi.org/10.1007/978-3-030-75549-2_38.

[43] R. T. Rockafellar and S. Uryasev, *Optimization of conditional value-at-risk*, Journal of Risk, 2 (2000), pp. 21–41, https://doi.org/10.21314/JOR.2000.038.

[44] R. T. Rockafellar and R. J. B. Wets, *Variational Analysis*, vol. 317 of Grundlehren der mathematischen Wissenschaften, Springer-Verlag Berlin Heidelberg, 1998, https://doi.org/10.1007/978-3-642-02431-3.

[45] A. Shapiro, D. Dentcheva, and A. Ruszczy'nski, *Lectures on Stochastic Programming: Modeling and Theory*, MOS-SIAM Series on Optimization, SIAM & Mathematical Optimization Society, Philadelphia, 2014, https://doi.org/10.1137/1.9781611973433.

[46] F. J. Solis and R. J.-B. Wets, *Minimization by random search techniques*, Mathematics of Operations Research, 6 (1981), pp. 19–30, https://doi.org/10.1287/moor.6.1.19.

[47] X. Song, B. Chen, and B. Yu, *An efficient duality-based approach for PDE-constrained sparse optimization*, Computational Optimization and Applications, 69 (2018), pp. 461–500, https://doi.org/10.1007/s10589-017-9951-4.

[48] J. Spall, *Implementation of the simultaneous perturbation algorithm for stochastic optimization*, IEEE Transactions on Aerospace and Electronic Systems, 34 (1998), pp. 817–823, https://doi.org/10.1109/7.705889.

[49] J. C. Spall, *Multivariate stochastic approximation using simultaneous perturbation gradient approximation*, IEEE Transactions on Automatic Control, 37 (1992), pp. 332–341, https://doi.org/10.1109/9.119632.

[50] A. Teixeira, E. Ghadimi, I. Shames, H. Sandberg, and M. Johansson, *Optimal scaling of the admm algorithm for distributed quadratic programming*, in 52nd IEEE Conference on Decision and Control, 2013, pp. 6868–6873, https://doi.org/10.1109/CDC.2013.6760977.

[51] F. Tröltzsch, *Optimal Control of Partial Differential Equations: Theory, Methods and Applications*, vol. 112 of Graduate Studies in Mathematics, American Mathematical Society, 2010, https://doi.org/10.1090/gsm/112.

[52] J.-P. Vial, *Strong and weak convexity of sets and functions*, Mathematics of Operations Research, 8 (1983), pp. 231–259, https://doi.org/10.1287/moor.8.2.231.

[53] G. Wachsmuth and D. Wachsmuth, *Convergence and regularization results for optimal control problems with sparsity functional*, ESAIM: Control, Optimisation and Calculus of Variations, 17 (2011), pp. 858–886, https://doi.org/10.1051/cocv/2010027.

[54] Y. Wang, S. Du, S. Balakrishnan, and A. Singh, *Stochastic zeroth-order optimization in high dimensions*, in Proceedings of the Twenty-First International Conference on Artificial

Intelligence and Statistics, A. Storkey and F. Perez-Cruz, eds., vol. 84 of Proceedings of Machine Learning Research, PMLR, 2018, pp. 1356–1365, https://proceedings.mlr.press/v84/wang18e.html.