A Zeroth-order Proximal Stochastic Gradient Method for Weakly Convex Stochastic Optimization

Spyridon Pougkakiotis* Dionysios S. Kalogerias†

Yale University

May 3, 2022

Abstract

In this paper we present and analyze a zeroth-order proximal stochastic gradient method suitable for the minimization of weakly convex stochastic optimization problems. We consider non-smooth 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. 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, utilizing a much simpler analysis as compared with a double Gaussian smoothing alternative recently analyzed in the literature, and under less restrictive assumptions. The proposed method is numerically compared against its (sub-)gradient-based counterparts to demonstrate its viability 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.

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 consider stochastic optimization problems of the form

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

where $F \colon \mathbb{R}^n \times \Xi \to \mathbb{R}$ is a weakly convex function in x for almost every (a.e.) $\xi \in \Xi \subset \mathbb{R}^d$ and Borel in ξ, while $g \colon \mathbb{R}^n \to \overline{\mathbb{R}} \equiv \mathbb{R} \cup \{+\infty\}$ is an extended-valued proper convex lower semi-continuous function (and hence closed), which is assumed to be proximable (that is, its proximity operator can be computed expeditiously).

* spyridon.pougkakiotis@yale.edu

† dionysis.kalogerias@yale.edu

Problem (P) is very general and appears in a variety of applications arising in signal processing (e.g. [16]), optimization (e.g. [31]), engineering (e.g. [29]), machine learning (e.g. [30]), and finance ([39]), to name a few. The reader is referred to [11, Section 2.1] and [13, Section 3.1] for a plethora of examples. Since neither $f(\cdot)$ nor $r(\cdot)$ are assumed to be smooth, standard stochastic gradient-based schemes are not applicable. In light of this, the authors in [11] analyzed various model-based stochastic sub-gradient methods 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 $\phi(\cdot)$ ([33]), which serves as a surrogate function with stationary points coinciding with those of (P).

However, there is a variety of applications in which even sub-gradient information of $f(\cdot)$ (or that of $F(\cdot, \xi)$) might not be available (e.g. [2, 7, 22]), or such a computation might be prohibitively expensive or noisy (e.g. see [1, 27, 32]). Thus, several zeroth-order schemes have been developed for the solution of stochastic optimization problems similar to (P), which require only function evaluations of $F(\cdot, \xi)$, in the absence of gradient information. Recently, the authors in [25] developed and analyzed a zeroth-order scheme based on *Gaussian smoothing* (see [34]) for the solution of stochastic compositional problems with applications to risk-averse learning, in which $r(\cdot)$ is chosen as an indicator function to a closed convex set (notice that in this case $f(\cdot)$ can model risk-aversity and thus belongs to a wider class of functions than those considered in this work). The authors in [3], based on an earlier work in [21], considered (Gaussian smoothing-based) zeroth-order schemes for non-convex stochastic optimization problems, again assuming that $r(\cdot)$ is an indicator function, and focusing on high-dimensionality issues as well as on avoiding saddle-points. A similar approach, utilizing different zeroth-order oracles, was developed and analyzed in [46]. In general, there is a plethora of zeroth-order optimization algorithms, and the interested reader is referred to [4, 10, 15, 26, 34], and the references therein.

To the best of our knowledge, the only development of zeroth-order methods for the solution of (P) can be found in the recent article [28]. The authors in [28] utilize a double Gaussian smoothing scheme, analyzed in detail in [14]. We argue herein that the use of double smoothing is essentially unnecessary, at least in conjunction with the discussion in [28]. In particular, the analysis of the proposed algorithm in [28] is substantially more complicated as compared to the analysis provided herein (cf. Section 3 and [28, Section 3]), while at the same time offering no advantage in terms of the rate bounds achieved. Additionally, in [28] 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, 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.

That said, it should be noted that in [14] tighter dependence on the dimension n is shown for the norm of the gradient estimates obtained from the double smoothing approach, assuming that the smoothing parameters are appropriately tuned, as compared with the dependence obtained using the standard single smoothing [34]. Whether the use of the double smoothing approach can result in convergence rates with better dependence on the dimension, as compared to that obtained in this work, and especially without imposing additional assumptions, is still an open problem. This was not attempted here, mainly due to the lack of numerical evidence supporting superiority of the double Gaussian smoothing scheme.

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 [34]). Following the developments in [11], 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 [11]), up to a constant term depending on the square root of the dimension of x (i.e. $\sqrt{n}$). As discussed above, a similar result was also shown in [28], however, under additional assumptions and using a substantially more complicated analysis. 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 $r(\cdot)$ is an indicator function (e.g. see [25]), however not so obvious for general closed convex functions $r(\cdot)$ that are studied here. Indeed, this was not considered in [28].

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 problem instances taken from [11], and compare its numerical behaviour against the sub-gradient model-based schemes developed in [11]. Subsequently, we showcase that the practical performance of the proposed algorithm is almost identical to that achieved by the double smoothing zeroth-order scheme analyzed in [28]. Overall, the observed 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 sub-gradient schemes.

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 [7], 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 [5, 6, 20]), 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 [9, 19, 38]). Given the importance of hyper-parameter tuning, there have been developed several heuristic software packages for this purpose, such as the Nevergrad toolkit (see [23]). 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.

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 introduce certain core properties of stochastic weakly convex functions of the form of f. Subsequently, we discuss *Gaussian smoothing* (e.g. see [25, 34]), which provides a smooth surrogate for $f(\cdot)$ 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). 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 [44], and the references therein.

Proposition 1. *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 for any two $x_1,\ x_2 \in \mathbb{R}^n$ and any $\lambda \in [0,1]$ we have*

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

Additionally, there exists $z \in \mathbb{R}^n$ such that

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

Moreover, the latter holds for any $z \in \partial f(x_1)$. Finally

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

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 [44, Propositions 4.4, 4.5, and 4.8]. □

Proposition 2. *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 1, and can be found in [44, Proposition 4.12]. □

2.2 Gaussian smoothing

Let us introduce the notion Gaussian smoothing. To that end, we follow the notation adopted in [25]. 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 $\mu \ge 0$, 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 $\mathbb{R}^n$. 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):

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

where $U \sim \mathcal{N}\left(0_n, I_n\right)$. 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 [34]).

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

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

(**C2**) *$F\left(x,\xi\right)$ is Borel in $\xi \in \Xi$ and ρ-weakly convex in $x \in \mathbb{R}^n$ for some $\rho \geq 0$, i.e. $F\left(\cdot,\xi\right)$ is Borel-measurable, and the map $x \mapsto F\left(x,\xi\right) + \frac{\rho}{2}\|x\|_2^2$ is convex for a.e. $\xi \in \Xi$.*

(**C3**) *There exists a positive random variable $C(\xi)$ such that $\sqrt{\mathbb{E}_\xi\left[C(\xi)^2\right]} < \infty$, 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 1. *In view of* (**C1**) *and* (**C2**) *in Assumption 1, we can infer that $f(\cdot)$ 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 $F(x,\cdot) \in \mathcal{L}_1\left(\Omega, \mathscr{F}, P; \mathbb{R}\right)$ for any $x \in \mathbb{R}^n$. The stronger assumption will be utilized later on. Furthermore, from* [41, Theorem 7.44]*, under* (**C1**) *and* (**C3**)*, we can show that there must exist a constant $L_{f,0} > 0$, such that $f(\cdot)$ is $L_{f,0}$-Lipschitz continuous on $\mathbb{R}^n$. Again, this holds even if we weaken assumption* (**C3**)*, and only require that $\mathbb{E}_\xi\left[C(\xi)\right] < \infty$, however, the stronger form of this assumption will be utilized later on. Finally, from* (**C2**)*, it is trivial to show that $f(\cdot)$ is ρ-weakly convex.*

Under Assumption 1, we will provide certain properties of the surrogate function $f_\mu(\cdot)$, as presented in [34].

Lemma 2.1. *Let Assumption 1 hold. Then, $f_\mu(\cdot)$ is ρ-weakly convex, and there exists a constant $L_{f_\mu,0} \leq L_{f,0}$ such that $f_\mu(\cdot)$ 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 $\mu > 0$, $f_\mu(\cdot)$ is Lipschitz continuously differentiable, with gradient given as

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

where U, ξ are statistically independent. The Lipschitz constant of the gradient is

$$L_{f_\mu,1} = \frac{n^{\frac{1}{2}}}{\mu} L_{f,0}.$$

Additionally, we have that

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

Proof. Weak convexity of the surrogate can be obtained by [25, Lemma 5.2]. For a proof of (2.1), as well as the first equality of (2.2), the reader is referred to [34, Appendix]. The second equality in (2.2), in light of (**C3**) of Assumption 1, 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 [25, 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 derive 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 [11, 33]) 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 [35]). More recently, it was noted in [11, Section 2.2] that the norm of its gradient (that is $\|\nabla p^\lambda(u)\|_2$) can serve as a near-stationarity measure for non-smooth 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 2. *Let $F(x,\xi)$ as defined in (P) and Assumption 1 hold for $F(x,\xi)$. Additionally, we assume that $r(\cdot)$ is a closed convex function (and thus lower semi-continuous), proper (i.e. $\text{dom}(r) \neq \emptyset$), and proximable (that is, its proximity operator can be evaluated expedetiously). Finally, we can generate two statistically independent random sequences $\{U_i\}_{i=0}^{\infty}$, $\{\xi_i\}_{i=0}^{\infty}$, such that each $U_i \sim \mathcal{N}(0_n, I_n)$ and ξ_i is i.i.d., respectively.*

In light of Assumption 2, and by utilizing Lemma 2.1, 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 $f_\mu(\cdot)$ is smooth, even if $f(\cdot)$ is not. Thus, we can derive an optimization algorithm for the minimization of $\phi_\mu(\cdot)$ (which can utilize stochastic gradient approximations for the smooth function $f_\mu(\cdot)$), 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

$$\min_{x} \ \phi_\mu(x) = f_\mu(x) + r(x), \tag{P$_\mu$}$$

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

Algorithm Z-ProxSG Zeroth-Order Proximal Stochastic Gradient

Input: $x_0 \in \text{dom}(r)$, 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, $U_t \sim \mathcal{N}(0_n, I_n)$, and set

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

  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. Once an approximate solution to the surrogate problem is found, we utilize Lemma 2.1, in order to assess the quality of this solution for the original problem (P). The analysis follows closely the developments in [11, Section 3.2].

Let us first introduce some notation. Set $\bar{\rho} > \rho$, where ρ is the weak-convexity constant of $F(\cdot, \xi)$. We define $\hat{x}_t \coloneqq \mathbf{prox}_{\bar{\rho}^{-1}\phi_\mu}(x_t)$, and $\delta_t \coloneqq 1 - \alpha_t \bar{\rho}$.

Lemma 3.1. *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 [11], we derive a descent property for the iterates.

Lemma 3.2. *Let Assumption 2 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 $\mathbb{E}^t_{U,\xi}[\cdot] \equiv \mathbb{E}_{U,\xi}[\cdot | U_{t-1}, \xi_{t-1}, \ldots, U_0, \xi_0]$.

Proof. We have

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

where the first equality follows from Lemma 3.1, the first inequality follows from non-expansiveness of the proximal operator (e.g. see [40, 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 1). Since $\bar{\rho} \leq 2\rho$, the result follows. □

We can now establish convergence of Algorithm Z-ProxSG, in terms of magnitude of the gradient of the Moreau envelope of the surrogate problem's objective function $\phi_\mu^{1/\bar{\rho}}(\cdot)$.

Theorem 3.3. *Let Assumption 2 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 $\bar{\rho} > \rho$, it holds that*

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

and x_{t^*} satisfies

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

In particular, letting $\bar{\rho} = 2\rho$, $\Delta \leq \phi_\mu^{1/\bar{\rho}}(x_0) - \min\limits_x \phi_\mu(x)$, 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. Following the developments in [11], we have

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

where the first inequality follows from the definition of the Moreau envelope, the second from Lemma 3.2, 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. Inequality (3.2) can be obtained as in [11, 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 expression in (3.2), yields (3.4) and completes the proof. □

The previous theorem shows convergence of Algorithm Z-ProxSG to an approximate solution of (P_μ). 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 2.1. 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 easily show that for any $x \in \mathbb{R}^n$, and $\hat x \coloneqq \mathbf{prox}_{\lambda\phi_\mu}(x)$, the following hold:

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

where, given any closed set $\mathcal{A} \subset \mathbb{R}^n$, $\mathrm{dist}\,(z;\mathcal{A}) \coloneqq \inf_{z'\in\mathcal{A}}\|z - z'\|_2$.

In the following lemma, we relate the Moreau envelope of the original problem's objective function $\phi^\lambda(\cdot)$ to the smooth surrogate $\phi_\mu(\cdot)$ in (P_μ).

Lemma 3.4. *Let Assumption 2 hold. Given any iterate x_t of Algorithm Z-ProxSG, any $\bar\rho \in (\rho, 2\rho]$, and any $\mu > 0$, we have that*

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

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

Proof. See Apendix A.2. □

Theorem 3.5. *Let Assumption 2 hold. Let x_ϵ be an ϵ-stationary point of problem (P_μ), that is, there exists $v_\mu \in \partial\phi_\mu(x_\epsilon)$, such that $\|v_\mu\|_2 \leq \epsilon$ (i.e. $\mathrm{dist}\,(0, \partial\phi_\mu(x_\epsilon)) \leq \epsilon$). Given any $\bar\rho \in (\rho, 2\rho]$, and any $\mu > 0$, we have that $|\phi(x_\epsilon) - \phi_\mu(x_\epsilon)| \leq \mu L_{f,0} n^{\frac{1}{2}}$. Moreover,*

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

Proof. The first part of the lemma follows immediately from the definition of $\phi_\mu(\cdot)$ and Lemma 2.1.

From Lemma 3.4, we have that

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

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

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

where we used the definition of $\tilde{x}_\epsilon$, the expression of the gradient of $\phi^{1/\bar{\rho}}(x_\epsilon)$, and the assumption that $\|v_\mu\|_2 \leq \epsilon$. For ease of presentation, we introduce some notation. Let $u \coloneqq \left\|\nabla \phi^{1/\bar{\rho}}(x_\epsilon)\right\|_2$, $\beta \coloneqq -\frac{\epsilon\bar{\rho}}{\bar{\rho}-\rho}$, 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 result then follows immediately by plugging the values of β and γ. □

Remark 2. *Let us notice that the convergence rate in Theorem 3.3 is given in terms of the squared gradient norm of the Moreau envelope of the surrogate $\phi_\mu(\cdot)$. This is in line with the results presented in [28], 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 3.5. Ideally, we would like to bound $\|\nabla \phi^{1/\bar{\rho}}(x_{t^*})\|_2^2$. In the special case where $r(\cdot)$ is an indicator function to a closed convex set, this can be done as in [25, Section 6.4.2]. In the general case, we were able to bound $\|\nabla \phi^{1/\bar{\rho}}(x_\epsilon)\|_2^2$, where x_ϵ is any ϵ-stationary point of problem (P_μ).*

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 [11] and compare the proposed zeroth-order approach, outlined in Algorithm Z-ProxSG, against the double smoothing zeroth-order proximal stochastic gradient method analyzed in [28], as well as the stochastic sub-gradient and the stochastic proximal point method proposed and analyzed in [11], noting that the latter two methods are significantly more difficult to employ (and implement) in the general case, since they assume availability of (sub-) gradient information. For completeness, the three algorithms are outlined in Algorithm DSZ-ProxSG, S-PPM, ProxSSG, respectively. We verify that the proposed approach performs almost identically to the method outlined in [28], 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: $x_0 \in \text{dom}(r)$, 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

Algorithm S-PPM Stochastic Proximal Point

Input: $x_0 \in \text{dom}(r)$, 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

$$x_{t+1} = \mathbf{prox}_{\alpha_t(F(\cdot,\xi)+r(\cdot))}\left(x_t\right).$$

end for

Algorithm ProxSSG Proximal Stochastic Sub-Gradient

Input: $x_0 \in \text{dom}(r)$, 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 $G\left(x_t, \xi_t\right) \in \partial F(x, \xi)$.
end for

4.1 Phase retrieval

Let us first focus on the solution of phase retrieval problems. Following [11], 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, 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 [11], this is a weakly convex optimization problem. We attempt to solve it using Algorithms Z-ProxSG, DSZ-ProxSG, ProxSSG, and S-PPM. For this specific instance, we can explicitly compute the sub-gradient of each summand of the objective,

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

as well as evaluate the proximal operator appearing in Algorithm S-PPM. The explicit expressions for the phase retrieval problem can be found in [11, Section 5.1].

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 and DSZ-ProxSG, 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 [11, Section 3], we set $\alpha_t = \frac{1}{2\sqrt{T}}$. On the other hand, Algorithm S-PPM is known to be very robust with respect to the step-size. After some preliminary testing, we confirmed that the choice $\alpha_t = 10^{-1}$ worked well, and it was adopted for all the experiments. Finally, Algorithms Z-ProxSG, DSZ-ProxSG are quite robust with respect to the choice of the smoothing parameter μ ($\mu_1,\ \mu_2$, respectively). For Algorithm Z-ProxSG this was set to $\mu = 5 \cdot 10^{-10}$. For Algorithm DSZ-ProxSG, by loosely following the theory in [14, 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 the two zeroth-order schemes.

We set up 9 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 three methods for T iterations. In Figure 1, we present the average convergence profiles with 95% confidence intervals for each of the three methods.

We can draw several useful observations from Figure 1. Firstly, we can observe that when m is significantly larger than d, the stochastic proximal point method is able to find a global solution very consistently. Overall, this method exhibits the best behaviour (as expected), however, it is generally much more expensive to employ. On the other hand, while the convergence of the two zeroth-order schemes is slower, as compared to the convergence of the first-order schemes (as we expected from the theory), the obtained solution is comparable to that yielded from the proximal stochastic sub-gradient scheme. Finally, as we already mentioned, the two zeroth-order schemes have a very similar behaviour, which was expected as we used similar values for the smoothing parameters. Thus the scheme proposed in this paper is more attractive due to its simplicity. In order to verify this behaviour for a wide-range of smoothing parameter values, 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 in each case, are reported in Figure 2.

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 [8] for an introductory review of ADMMs), although we conjecture that the same technique can be employed for tuning a much wider range of optimization methods.

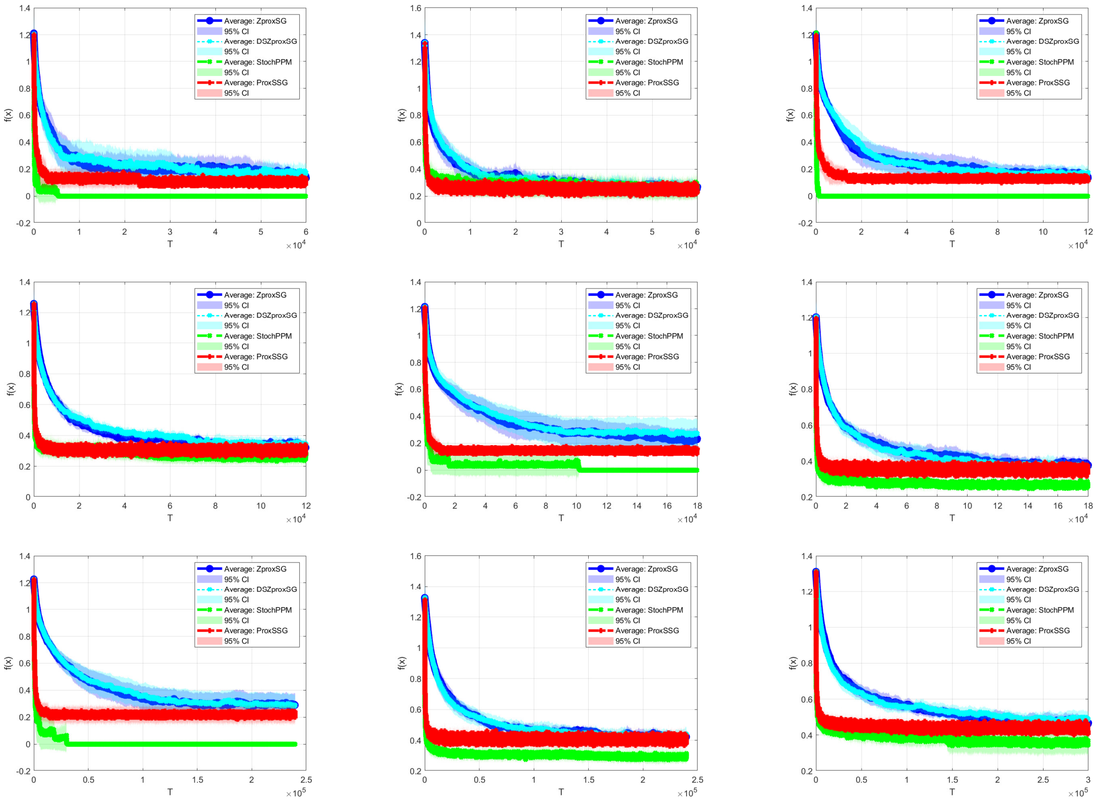

Figure 1: Convergence profiles for Z-ProxSG, DSZ-ProxSG, S-PPM, and ProxSSG: average objective function value (lines) and 95% confidence intervals (shaded regions) vs number of iterations. The upper row corresponds, from left to right, to $(d, m) = (10, 30)$, $(20, 30)$, $(20, 60)$. The middle row corresponds,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

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:

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

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

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

D denotes a linear differential operator, $\boldsymbol{x}$ is a 2-dimensional spatial variable, and $\beta_1, \ \beta_2 \geq 0$ 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

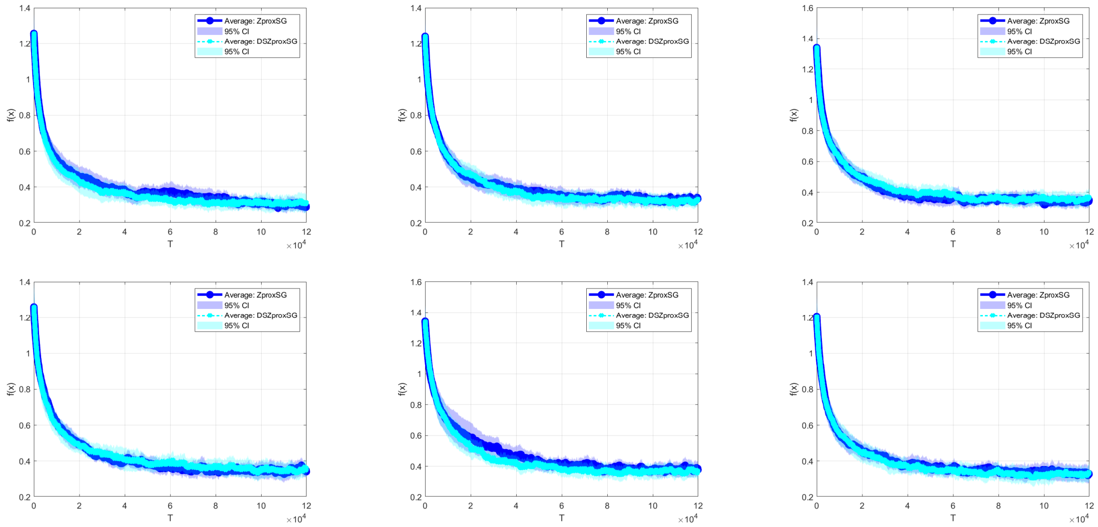

Figure 2: Convergence profiles for Z-ProxSG, DSZ-ProxSG: average objective function value (lines) and 95% confidence intervals (shaded regions) vs number of iterations, for $(d, m) = (40, 60)$. The upper row corresponds, from left to right, to $(\mu_1, \mu_2) = (10^{-x}, 10^{-y})$, $x = 4, 5, 6$, $y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

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 [36], 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 [36, 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 [36]).

We solve problem (4.1) via a *discretize-then-optimize* strategy. We employ the Q1 finite element discretization implemented in IFISS[2] (see [17, 18]). This yields a sequence of ℓ_1-regularized convex quadratic programming problems of the following form:

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

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 [43]), while the $\mathcal{L}_1$ term is discretized by the *nodal quadrature rule* as in [42, 45] (an approximation that achieves a first-order convergence–see [45]).

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

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

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

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

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

Let an arbitrary positive definite matrix R_x be given, and assume the notation $\|x\|_{R_x}^2 = x^\top R_x x$. 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 certain standard assumptions on (4.4), Algorithm pADMM is able to achieve linear convergence (see [12]), even in cases where R_x is not positive definite [24]. Here we assume that R_x is positive definite, and we employ it as a means of reducing the memory requirements of Algorithm pADMM. More specifically, given some constant $\hat{\sigma} > 0$, such that $\hat{\sigma} I_n - \mathrm{Off}(Q) \succ 0$, we define

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

where $\mathrm{Off}(B)$ 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 [37] 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

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

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 3. *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, 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. 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.

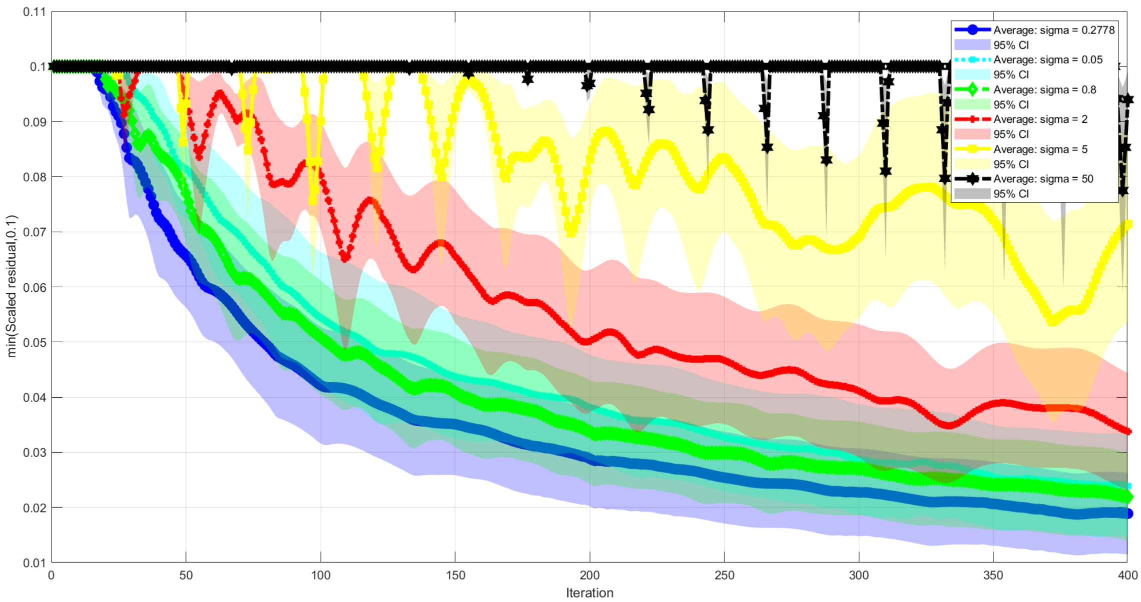

Figure 3: Convergence profiles for pADMM with varying penalty parameter σ: average residual reduction (lines) and 95% confidence intervals (shaded regions) vs number of pADMM iterations. The algorithm is run over 40 randomly selected Poisson optimal control problems.

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^* 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.

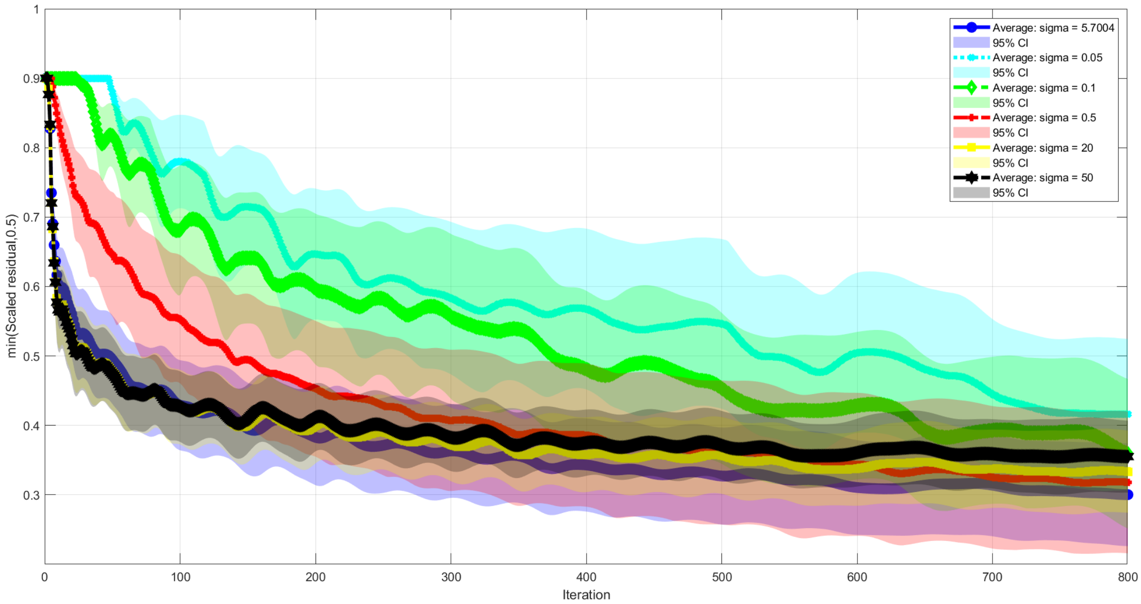

Figure 4: Convergence profiles for pADMM with varying penalty parameter σ: average residual reduction (lines) and 95% confidence intervals (shaded regions) vs number of pADMM iterations. The algorithm is run over 40 randomly selected convection-diffusion optimal control problems.

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. setting it to the number of iterations that we are willing to let ADMM run), 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 where 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.

A Appendix

A.1 Proof of Lemma 3.1

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

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

□

A.2 Proof of Lemma 3.4

Proof. Following [25, 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 1, 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 $x_1=x_t$ and $x_2=\tilde{x}_t\coloneqq\mathbf{prox}_{\bar{\rho}^{-1}\phi}(x_t)$, and by noting that $\bar{\rho}>\rho$, we obtain

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

However, we know that the map $x\mapsto\left(\phi(\cdot)+\frac{\bar{\rho}}{2}\|\cdot-x_t\|_2^2\right)$ is strongly convex with parameter $\bar{\rho}-\rho$, and is minimized at $\tilde{x}_t$, and thus $\phi(x_t)+\frac{\bar{\rho}}{2}\|x_t-x_t\|_2^2-\left(\phi(\tilde{x}_t)+\frac{\bar{\rho}}{2}\|\tilde{x}_t-x_t\|_2^2\right)\ge\frac{\bar{\rho}-\rho}{2}\|x_t-\tilde{x}_t\|_2^2$. Hence, we obtain

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

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] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] R. J. and O. Teytaud, *Nevergrad - A gradient-free optimization platform*. https://GitHub.com/FacebookResearch/Nevergrad, 2018.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.