

INEXACT ZERO-ORDER NONSMOOTH AND NONCONVEX STOCHASTIC COMPOSITE OPTIMIZATION AND APPLICATIONS

SPYRIDON POU GKAKIOTIS^{1,*}, DIONYSIS KALOGERIAS²

¹*Department of Mathematics, King's College London, London, England, UK*

²*Department of Electrical and Computer Engineering, Yale University, New Haven, CT, USA*

Abstract. In this paper, we present an inexact zeroth-order method suitable for the solution of nonsmooth and nonconvex stochastic composite optimization problems, in which the objective is split into a real-valued Lipschitz continuous stochastic function and an extended-valued (deterministic) proper, closed, and convex one. The algorithm operates under inexact oracles providing noisy (and biased) stochastic evaluations of the underlying finite-valued part of the objective function. We show that the proposed method converges (non-asymptotically), under very mild assumptions, close to a stationary point of an appropriate surrogate problem which is related (in a precise mathematical sense) to the original one. This, in turn, provides a new notion of approximate stationarity suitable nonsmooth and nonconvex stochastic composite optimization, generalizing conditions used in the available literature. In light of the generic oracle properties under which the algorithm operates, we showcase the applicability of the approach in a wide range of problems including large classes of two-stage nonconvex stochastic optimization and nonconvex-nonconcave minimax stochastic optimization instances, without requiring convexity of the lower level problems, or even uniqueness of the associated lower level solution maps. We showcase how the developed theory can be applied in each of these cases under general assumptions, providing algorithmic methodologies that go beyond the current state-of-the-art appearing in each respective literature, enabling the solution of problems that are out of reach for currently available methodologies.

Keywords. Nonconvex-nonconcave minimax optimization; Nonsmooth and nonconvex optimization; Nonconvex stochastic composite optimization; Two-stage stochastic programming; Zeroth-order optimization.

1. INTRODUCTION

Let $(\Omega, \mathcal{F}, \mu)$ be a complete base probability space, and consider a random vector $\xi : \Omega \rightarrow \Xi \subset \mathbb{R}^d$, and its induced Borel space $(\Xi, \mathcal{B}(\Xi), P)$, where $P : \mathcal{B}(\Xi) \rightarrow [0, 1]$ is the induced Borel measure. In this paper, we consider the following *nonsmooth, nonconvex and stochastic composite optimization* problem:

$$\min_{x \in \mathbb{R}^n} \phi(x) \triangleq \underbrace{\mathbb{E}\{F(x, \xi)\}}_{\triangleq f(x)} + r(x), \quad (\text{P})$$

*Corresponding author.

E-mail address: spyridon.pougkakiotis@kcl.ac.uk (S. Pougkakiotis), dionysis.kalogerias@yale.edu (D. Kalogerias).

Received 16 August 2025; Accepted 30 September 2025; Published online 1 January 2026.

©2026 Journal of Nonlinear and Variational Analysis

where $r: \mathbb{R}^n \rightarrow \bar{\mathbb{R}}$ is a closed, proper, convex and *proximable* function (i.e., a function the proximity operator of which can be computed expeditiously). Throughout this work, we will make use of the following blanket assumption on (P).

Assumption 1.1. The following conditions are in effect for (P):

- (A1) The function $F(x, \cdot)$ is Borel measurable for all $x \in \mathbb{R}^n$. For a.e. $\xi \in \Xi$, the function $F(\cdot, \xi): \mathbb{R}^n \rightarrow \mathbb{R}$ is $L(\xi)$ –Lipschitz continuous with $\mathbb{E}\{L^2(\xi)\} \leq G^2$, for some $G > 0$. Moreover, we have that $f(x) = \mathbb{E}\{F(x, \xi)\}$, for all $x \in \mathbb{R}^n$;
- (A2) We can draw i.i.d. samples from the law of ξ ;
- (A3) We have that r is closed, proper, convex, and proximable (as already mentioned).

1.1. Prior work and core contributions. Nonsmooth and nonconvex stochastic optimization has received a lot of attention in recent years, due to its ubiquitous presence in machine learning and artificial intelligence applications. A seminal work on tackling such problems was originally proposed in [1], in which the authors provided an interpolated normalized gradient method for the solution of (P) (assuming that $r = 0$ and that F belongs to an appropriate sub-class of Lipschitz functions), and showed that it converges non-asymptotically towards a (δ, ε) –Goldstein stationary point (for additional details on this mode of convergence, see Definition 2.1 and Section 3). This work built upon earlier developments due to Goldstein (see [2]) and led to a series of works extending these results. Indeed, an improved interpolated normalized gradient variant was later proposed in [3], showing that its non-asymptotic convergence towards a (δ, ε) –Goldstein stationary point holds for any Lipschitz continuous function F . An alternative line of work, highly related to this paper, deviated from interpolated normalized gradient schemes, considering instead (randomized) zeroth-order stochastic optimization methods (see [4, 5] and the references therein for an overview on randomized zeroth-order stochastic optimization). Indeed, as was originally identified in [6], the gradients of uniform randomized smoothed surrogates associated to (P) (again, assuming that $r = 0$) have a close (and mathematically precise) relation to the (δ, ε) –Goldstein subdifferential. In turn, they were able to show that the associated zeroth-order stochastic gradient schemes arising from such smoothing strategies also converge (non-asymptotically) in the Goldstein sense, much like interpolated normalized gradient schemes, albeit with a rate that depends on the problem dimension. Improved variants (in the sense of dimension-dependence) of the algorithm presented in [6] were later proposed in [7, 8] and then in [9]. Considerations about the inherent need for randomized smoothing were also discussed in [10]. Most works on nonsmooth and nonconvex optimization currently available in the literature focus on the unconstrained (non-composite) case (i.e., in the case where $r = 0$ in (P)). To the best of our knowledge, the case where $r \neq 0$ (and is allowed to be extended-valued) is only considered in [11] (although structured constrained formulations of (P) have been considered in other studies such as in [12]), where the authors propose a zeroth-order optimization scheme in the case where r is an indicator to a closed convex and compact set. The authors in [11] generalize (δ, ε) –Goldstein stationarity to fit their problem (cf. [11, Definition 4.2]) by appropriately extending the well-known *gradient mapping* (see [13, Section 2.2.4] for a definition of this mapping) using the Goldstein subdifferential; their proposed stationarity condition is different from the generalization proposed in this work (cf. Section 3), which combines surrogate stationarity based on both the Moreau envelope (akin to that considered in [14] for weakly convex composite optimization) and the Goldstein subdifferential (as is done in standard unconstrained nonsmooth and nonconvex optimization; e.g., see [1]). The

proposed notion of stationarity has several benefits compared to that considered in [11]. On the one hand, it readily enables the use of a generic convex regularizer r . On the other hand, it provides a natural framework for analyzing the (non-asymptotic) convergence of zeroth-order optimization schemes applied to (P), while maintaining crucial connections to the Goldstein (approximate) stationarity in the unconstrained case (i.e., when $r = 0$). More crucially, current algorithms studied in the literature, suitable for the solution of nonsmooth and nonconvex optimization problems, operate under the assumption of the availability of *exact oracles* able to evaluate the stochastic function $F(x, \xi)$, for any $x \in \mathbb{R}^n$ and a.e. $\xi \in \Xi$. As will become clear in Sections 1.2 and 5, enabling the presence of inexactness in the evaluations of $F(x, \xi)$ is of paramount importance for several applications of practical interest. Thus, this work aims at closing core gaps in the current literature of nonsmooth and nonconvex stochastic optimization, by providing a natural condition of stationarity suitable for the constrained (or composite) case, while at the same time allowing for errors in the underlying stochastic function evaluations. Furthermore, by specializing the notion of an inexact oracle in the context of zeroth-order stochastic optimization, we provide new and general conditions on the associated oracle errors. In turn, this enables us to derive improved (non-asymptotic) convergence rate bounds under very reasonable oracle error conditions, by simply utilizing the properties of zeroth-order optimization schemes.

1.2. Related applications. Problem (P) is prominent in a plethora of applications of great interest, stemming from machine learning to operational research and signal processing. Specifically, nonsmooth and nonconvex optimization involving Lipschitz continuous functions has received a lot of attention in the recent literature (e.g., see [1, 3, 6] and the references therein) due to its direct application on the training of neural networks which, when seen as compositional functions, often fail to satisfy standard assumptions like weak convexity, Lipschitz smoothness or even subdifferential regularity. Indeed, as is already mentioned in [3], most (sub)gradient-based methods rely on some form of subdifferential regularity, which fails when the function that is being optimized exhibits some “downward cusps” (e.g., see the example $f(x) = (1 - \max\{x, 0\})^2$, given in [3]). As we have already hinted earlier, one major gap in the current literature of nonsmooth and nonconvex optimization is the derivation of algorithms that are able to operate under noisy and inexact function evaluations. This is especially important in cases where the function $F(\cdot, \xi)$ appearing in (P) is itself a (possibly nonconvex) optimization problem. In this case, under fairly general conditions (e.g., see the discussion in Section 5 as well as the comprehensive exposition given in [15]), one may be able to show that $F(\cdot, \xi)$ is (possibly Lipschitz) continuous, but not necessarily differentiable or even subdifferentially regular. In this regime, the assumption that $F(\cdot, \xi)$ can be evaluated exactly, for a.e. $\xi \in \Xi$, is quite strong (since its evaluation typically occurs via the utilization of an “inner-layer” numerical optimization scheme). Two very important classes of problems that exhibit this behavior are (possibly nonconvex) two-stage stochastic programs and stochastic minimax optimization instances. Additionally, the same considerations apply in the context of hyperparameter tuning of black-box systems, the evaluation of which might be noisy and inexact (e.g., in case the objective function is evaluated via the utilization of a simulation process; the reader is referred to [16, Section 4.2] for an example of hyperparameter tuning in this context). More concretely, in the case of two-stage stochastic programming, the function $F(x, \xi)$ is defined as $F(x, \xi) = \min_{y \in \mathcal{Y}(x, \xi)} \hat{F}(x, y, \xi)$, where $\mathcal{Y}(x, \xi) \subset \mathbb{R}^m$ is the feasible set of the second-stage

variable y . Two-stage stochastic programming problems appear in a plethora of applications in operational research and engineering. While many such instances are posed in the context of convex stochastic optimization (e.g., see the detailed exposition in [17, Chapter 2]), nonconvex formulations are also highly relevant. A typical example arises in the context of beamforming optimization for wireless communication systems, in cases where the performance of the underlying network can be improved by tuning an appropriate set of parameters in a long timescale, jointly with optimizing short-time scale (i.e., recourse) variables [18, 19, 20, 21]; see also the recent line of work [22, 23, 24] within the more specialized but challenging context of intelligent reflecting surface-assisted beamforming. Another separate example of an application of two-stage stochastic optimization on certain meta-learning problems arising in the area of computer vision may be found in [25]. In the case of minimax stochastic optimization, we may separate two distinct cases. The first class of instances arises by letting $F(x, \xi) = \max_{y \in \mathcal{Y}(x, \xi)} \hat{F}(x, y, \xi)$, where $\mathcal{Y}(x, \xi) \subset \mathbb{R}^m$ is the feasible set of the *adversarial variable* y . In essence, in this formulation, the adversary is given access to instantaneous information and thus *from this point of view* the underlying stochastic minimax optimization problem is fairly similar to two-stage stochastic programming models, although structurally different. One of the most important applications of this problem formulation arises in the context of building neural networks robust to adversarial attacks (e.g., see the seminal paper [26] and numerous follow-up works). Problems of this form are practically solved via approximate stochastic hypergradient descent-type schemes (again, see [26]), although without any theoretical guarantees, despite the inherent assumption that the feasible set $\mathcal{Y}$ of the adversarial variable is independent of both x and ξ . Nonetheless, we conjecture that the stochastic hypergradient descent approach proposed in [24] in the context of nonconvex two-stage stochastic programming can possibly be adapted in this case and be shown to be non-asymptotically convergent under certain regularity and structural assumptions. The second class of stochastic minimax optimization instances, which is very well-studied in the literature, arises by assuming that the adversary only has access to ergodic information, in which case the objective function of (P) reads $F(x, \xi) = \min_{y \in \mathcal{Y}(x)} \mathbb{E}\{\hat{F}(x, y, \xi)\}$, where once again $\mathcal{Y}(x) \subset \mathbb{R}^m$ is the feasible set of the adversarial variable y . Such problems have multiple applications, especially in the context of machine learning, including generative adversarial networks (e.g., [27]), online adversarial learning (e.g., [28]), robust training of neural networks (e.g., [29]), nested optimization in reinforcement learning (e.g., see [30]), and distributionally robust optimization (e.g., see [31]), to name a few. In most of these applications, it is assumed that $\mathcal{Y}$ does not depend on x , and under assumptions like Lipschitz smoothness on $\hat{F}$ and lower level concavity, typical solution methods rely either on stochastic gradient descent-ascent variants (e.g., [32]) or the extragradient method under additional conditions (e.g., [33]), although recent works have also investigated the fully nonconvex-nonconcave setting under alternative (and even stronger) structural assumptions (e.g., see [34, 35, 36]). In this work, we showcase that a single algorithmic strategy, as proposed in this work, can be readily adapted and applied to each of these problems classes, resulting in solution methods that operate under very general assumptions, going beyond the current state-of-the-art in each respective literature, albeit at the cost of two function evaluations at each iteration (i.e., two inexact inner-problem solutions at adjacent outer-problem points). Nonetheless, we showcase that despite the added computational overhead, the proposed methodology is otherwise very efficient and can operate in regimes that are inaccessible to alternative approaches, offering strong modeling capabilities and robustness.

We now provide an overview of this paper. Specifically, in Section 1.3 we summarize the notation used throughout this work. Then, in Section 2 we provide necessary background material on nonsmooth optimization, variational analysis, randomized smoothing and evaluation oracles. Subsequently, in Section 3, we propose a new notion of approximate stationarity that is suitable for nonsmooth and nonconvex stochastic composite optimization. This is then used in Section 4, where we derive the proposed algorithm and show its non-asymptotic convergence under minimal assumptions. The results of Section 4 are then specialized to fit different applications in Section 5 to showcase the power and generality of the proposed methodological framework. Finally, we close this paper by collecting some conclusions in Section 6.

1.3. Notation. Throughout this work we write $\|\cdot\|$ to denote the standard Euclidean norm. Given some positive constant $\varepsilon > 0$ and some $x \in \mathbb{R}^n$, we let $\mathbb{B}_\varepsilon(x)$ denote the open ε -ball around x on $\mathbb{R}^n$, i.e., $\mathbb{B}_\varepsilon(x) \triangleq \{y \in \mathbb{R}^n \mid \|y - x\| < \varepsilon\}$. Similarly, the closed ε -ball around x on $\mathbb{R}^n$ is denoted as $\overline{\mathbb{B}}_\varepsilon(x)$. The unit sphere on $\mathbb{R}^n$ is denoted as $\mathbb{S}^{n-1} \triangleq \{x \in \mathbb{R}^n \mid \|x\| = 1\}$. We let $\overline{\mathbb{R}} \triangleq \mathbb{R} \cup \{\pm\infty\}$. A function $f: \mathbb{R}^n \rightarrow \mathbb{R}$, is said to be L -Lipschitz if for every $x, x' \in \mathbb{R}^n$ we have $|f(x) - f(x')| \leq L\|x - x'\|$. Given two real-valued functions f, g on $\mathbb{R}^n$, we denote their *integral convolution* as $(f * g)(x) \triangleq \int_{\mathbb{R}^n} f(\tau)g(x - \tau)d\tau \equiv \int_{\mathbb{R}^n} f(x - \tau)g(\tau)d\tau$ (assuming it is well-defined). Associated with integral convolution, and given a function $f: \mathbb{R}^n \rightarrow \mathbb{R}$, we define the *dilation operation* as $(\lambda \bullet f)(x) = \lambda^n f(\lambda x)$, for any $\lambda > 0$, noting that this operation dilates f , compressing it towards the origin without altering its integral over $\mathbb{R}^n$. Given a closed and proper function $f: \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$, we define its proximity operator as $\mathbf{prox}_{\lambda f}(x) \triangleq \arg \min_{w \in \mathbb{R}^n} \{f(w) + 1/(2\lambda)\|w - x\|^2\}$, where $\lambda > 0$. If $\mathbf{prox}_{\lambda f}(x)$ can be computed expeditiously (e.g., in closed-form), we say that f is *proximable*. Similarly, we define the *Moreau envelope* of f as $e_\lambda f(x) \triangleq \inf_{w \in \mathbb{R}^n} \{f(w) + 1/(2\lambda)\|w - x\|^2\}$. For some $\rho \geq 0$, we define the space of ρ -weakly convex functions as

$$\Gamma_\rho(\mathbb{R}^n) \triangleq \left\{ f: \mathbb{R}^n \rightarrow \overline{\mathbb{R}} \mid f \text{ is proper, closed, and } f + \frac{\rho}{2}\|\cdot\|^2 \text{ is convex} \right\},$$

noting that $\Gamma_0(\mathbb{R}^n)$ denotes the set of closed, proper, and convex functions. Given a proper and closed function $f: \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$, we define the *regular subdifferential* of f at $\bar{x} \in \mathbb{R}^n$, denoted as $\hat{\partial}f(\bar{x})$, as the set of all vectors $v \in \mathbb{R}^n$ that satisfy

$$f(x) \geq f(\bar{x}) + v^\top(x - \bar{x}) + o(\|x - \bar{x}\|).$$

The *limiting subdifferential* of f at $\bar{x} \in \mathbb{R}^n$, denoted as $\partial f(\bar{x})$, is defined as the set of vectors $v \in \mathbb{R}^n$ for which there exist sequences $x_k \rightarrow_f \bar{x}$ and $v_k \in \hat{\partial}f(x_k)$, with $v_k \rightarrow v$, where $x \rightarrow_f \bar{x}$ denotes f -attentive convergence. Finally, we denote the Clarke subdifferential of f at $\bar{x}$ as $\bar{\partial}f(\bar{x})$. If f is subdifferentially regular at $\bar{x}$, we have that $\partial f(\bar{x}) = \hat{\partial}f(\bar{x}) = \bar{\partial}f(\bar{x})$ (e.g., this holds for any $f \in \Gamma_\rho(\mathbb{R}^n)$).

2. PRELIMINARIES

2.1. Clarke and Goldstein subdifferentials. We begin our discussion by characterizing the Clarke subdifferential for Lipschitz functions. Its construction relies on the fact that, due to Rademacher's theorem, any Lipschitz function is almost everywhere differentiable (i.e., the subset of $\mathbb{R}^n$ in which f is non-differentiable has Lebesgue measure zero). This is done in the following proposition, which is due to Clarke [37].

Proposition 2.1 (Clarke subdifferential characterization (Lipschitz functions) [37]). *Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be an L -Lipschitz function, for some $L > 0$. Then, for any $x \in \mathbb{R}^n$ and any $g \in \bar{\partial}f(x)$, we have that $\|g\| \leq L$, and the set-valued mapping $\bar{\partial}f(\cdot)$ is upper semicontinuous. Moreover, for any $x, x' \in \mathbb{R}^n$, there exists $\lambda \in (0, 1)$ and $g \in \bar{\partial}f(\lambda x + (1 - \lambda)x')$, such that $f(x) - f(x') = g^\top(x - x')$. Finally,*

$$\bar{\partial}f(x) \equiv \text{conv} \left(\left\{ g \in \mathbb{R}^n \mid g = \lim_{x_k \rightarrow x} \nabla f(x_k) \right\} \right),$$

i.e., the Clarke subdifferential is the convex hull of all limit points of $\nabla f(x_k)$ over all sequences $\{x_k\}_{k=0}^\infty$ of differentiable points of $f(\cdot)$ which converge to x .

Consider the minimization of a general Lipschitz continuous function f . It is known that finding an ε -Clarke stationary point, in the sense that we have found an x such that $\min\{\|g\| \mid g \in \bar{\partial}f(x)\} \leq \varepsilon$, is intractable (see [1]). Instead, it has been observed that a relaxation of ε -Clarke stationarity, known as the (μ, ε) -Goldstein stationarity (see Section 3 for a definition), is computationally tractable. This relies on the so-called μ -Goldstein subdifferential, which we define next.

Definition 2.1 (μ -Goldstein subdifferential [2]). *Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be an L -Lipschitz function. Given any $x \in \mathbb{R}^n$, the μ -Goldstein subdifferential of f at x is defined by $\bar{\partial}_\mu f(x) \triangleq \text{conv} \left(\bigcup_{y \in \mathbb{B}_\mu(x)} \bar{\partial}f(y) \right)$, where $\mu > 0$ is a positive constant.*

We revisit the notion of generalized approximate stationarity in the context of nonsmooth and nonconvex optimization in Section 3.

2.2. Uniform randomized smoothing. We next introduce the notion of uniform randomized smoothing, which is obtained by the operation of integral convolution and dilation. To that end, we let $g: \mathbb{R}^n \rightarrow \mathbb{R}$ be an *integral smoothing kernel*, i.e., a bounded piecewise continuous density function (i.e., $\int_{\mathbb{R}^n} g(x) dx = 1$) that is even (i.e., $g(-x) = g(x)$), and satisfies

$$\int_{\mathbb{R}^n} \|x\| g(x) dx < +\infty.$$

Then, given some L -Lipschitz function $f: \mathbb{R}^n \rightarrow \mathbb{R}$, we define the surrogate function $f_\mu: \mathbb{R}^n \rightarrow \mathbb{R}$ as

$$\begin{aligned} f_\mu(x) &= (f * (\mu^{-1} \bullet g))(x) \equiv \int_{\mathbb{R}^n} f(x-t) (\mu^{-1} \bullet g)(t) dt \\ &= \int_{\mathbb{R}^n} f(x-\mu t) g(t) dt \equiv \mathbb{E}_g \{f(x+\mu t)\}, \end{aligned}$$

where we used a change of variables and the last equivalence follows from the symmetry of g . In this paper, we will focus on a particular mollifier function g , namely, the probability density function (p.d.f.) of a uniform random vector U over the closed unit ball on $\mathbb{R}^n$ (i.e., over $\bar{\mathbb{B}}_1(0_n)$). Specifically, let $U \sim U(\bar{\mathbb{B}}_1(0_n))$. Then, the p.d.f., say g , of U reads:

$$g_n(u) = \begin{cases} \frac{1}{c_n}, & \text{if } \|u\| \leq 1, \\ 0, & \text{otherwise} \end{cases}, \text{ with } c_n \triangleq \frac{\pi^{n/2}}{\Gamma(n/2+1)}, \Gamma(n/2+1) \triangleq \begin{cases} (n/2)!, & \text{if } n \text{ is even} \\ \sqrt{\pi} \frac{n!!}{2^{(n+1)/2}}, & \text{if } n \text{ is odd} \end{cases},$$

where $n!! = n(n-2) \cdots 2$ if n is even and $n!! = n(n-2) \cdots 1$, if n is odd. Applying the dilation operation on g with a constant $\mu^{-1} > 0$ yields the p.d.f. of a uniform random variable over the

μ -closed ball, i.e.,

$$(\mu^{-1} \bullet g)(u) = \begin{cases} \frac{1}{c_n \mu^n}, & \text{if } \|u\| \leq \mu, \\ 0, & \text{otherwise} \end{cases}.$$

Next, we provide a well-known key result which showcases the smoothing effect of integral convolution, under the assumption of Lipschitz continuity of f .

Lemma 2.1 (Uniform randomized smoothing of Lipschitz functions). *Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be an L -Lipschitz continuous function, and let $g: \mathbb{R}^n \rightarrow \mathbb{R}$ be the p.d.f. of a uniform random variable, say $U: \Omega \rightarrow \mathbb{R}^n$, over the n -dimensional unit ball $\mathbb{B}_1(0_n)$, i.e., $U \sim \mathcal{U}(\mathbb{B}_1(0_n))$. Then, the surrogate function defined as*

$$f_\mu(x) = (f * (\mu^{-1} \bullet g))(x), \quad \text{for all } x \in \mathbb{R}^n,$$

satisfies the following:

- f_μ is L -Lipschitz continuous and $|f_\mu(x) - f(x)| \leq \mu L$, for all $x \in \mathbb{R}^n$;
- f_μ is $\frac{cL\sqrt{n}}{\mu}$ -Lipschitz smooth, where $c > 0$ is a bounded constant independent of n .
Moreover, we have that

$$\begin{aligned} \nabla f_\mu(x) &= \frac{n}{\mu} \mathbb{E}_{W \sim \mathcal{U}(\mathbb{S}^{n-1})} \{f(x + \mu W)W\} \\ &\equiv \frac{n}{2\mu} \mathbb{E}_{W \sim \mathcal{U}(\mathbb{S}^{n-1})} \{(f(x + \mu W) - f(x - \mu W))W\}, \end{aligned}$$

where, as indicated above, W is a uniform random variable over the n -dimensional unit sphere $\mathbb{S}^{n-1}$;

- For all $x \in \mathbb{R}^n$, we have that $\nabla f_\mu(x) \in \bar{\partial}_\mu f(x)$, where $\bar{\partial}_\mu f(x)$ is the μ -Goldestein sub-differential of f at x .

Moreover, at every $\bar{x} \in \mathbb{R}^n$, we have

$$\bar{\partial} f(\bar{x}) = \text{conv} \left(\limsup_{x \rightarrow \bar{x}, \mu \searrow 0} \nabla f_\mu(x) \right),$$

i.e., gradient consistency holds, noting that the outer limit is defined as

$$\limsup_{x \rightarrow \bar{x}, \mu \searrow 0} \nabla f_\mu(x) \triangleq \{v \mid \exists \{(x_k, \mu_k)\}_{k \in \mathbb{N}} \rightarrow (\bar{x}, 0), \text{ such that } \nabla f_{\mu_k}(x_k) \rightarrow v\}.$$

Proof. The first part of the lemma follows from [6, Proposition 2.2 and Theorem 3.1], where the expression for the gradient of f_μ can be shown as in [38, Lemma 2.1] (where the symmetric expression for the gradient is given in [39]; see also [40, Section 9.4]). The gradient consistency property follows from a trivial extension of [41, Theorem 9.67]. $\square$

2.3. Inexact noisy oracle for function evaluations. One crucial part of this work is that we do not assume the availability of an unbiased (exact) stochastic oracle for evaluating the function $F(x, \xi)$, give some $x \in \mathbb{R}^n$ and some $\xi \in \Xi$. Instead, we make use of a general inexact stochastic oracle, defined below.

Definition 2.2 (Inexact noisy oracle). Let Assumption 1.1 hold for problem (P). For any $x \in \mathbb{R}^n$ and any $\xi \in \Xi$, we assume the availability of an *inexact noisy oracle* which returns a measurable function $\tilde{F}(x, \xi) \triangleq F(x, \xi) + \delta(x, \xi)$, where $\delta(\cdot, \cdot)$ is some measurable random error function, such that $|\delta(x, \xi)| \leq \tilde{\delta}$, for all $x \in \mathbb{R}^n$ and a.e. $\xi \in \Xi$, where $\tilde{\delta} > 0$ is some positive constant.

Remark 2.1. As we will see later on, when discussing the applications of the proposed methodology, Definition 2.2 is consistent with what we are trying to achieve in this paper. Specifically, we assume that this “oracle” is a numerical method that is employed in order to evaluate $F(x, \xi)$ up to some error tolerance, say $\tilde{\delta} > 0$, in the sense that for all $x \in \mathbb{R}^n$ and any $\xi \in \Xi$, it returns a quantity that satisfies

$$|\tilde{F}(x, \xi) - F(x, \xi)| = |\delta(x, \xi)| \leq \tilde{\delta}.$$

Saying that this oracle returns a measurable *function* is effectively the same as assuming that the numerical algorithm that we employ is deterministic or stochastic with a fixed seed, in the sense that it always returns the same result for fixed x and ξ . If the underlying numerical algorithm (constituting the oracle) were stochastic, then we would instead have to treat $\delta(\cdot, \cdot)$ as a measurable multifunction. This is omitted for simplicity of exposition. Finally, the imposition of a uniform error bound $\tilde{\delta}$ on the oracle error is also made for simplicity. Indeed, one could instead assume that $\tilde{\delta}$ is a random variable with finite first- and second-moments. This is also omitted for brevity of exposition.

Assumption 2.1. Either of the following two conditions is in effect for (P):

(B1) The inexact noisy oracle is such that for any $x \in \mathbb{R}^n$ we have

$$\mathbb{E}_{\xi} \{\delta(x, \xi)\} = \Delta,$$

where Δ is some constant;

(B2) The function r appearing in (P) is such that $r = h + \iota_{\mathcal{X}}$, where $h \in \Gamma_0(\mathbb{R}^n)$, $\mathcal{X}$ is a convex and compact set with diameter $D > 0$, and $\iota(\cdot)$ is the indicator function defined as $\iota_{\mathcal{X}}(x) = 0$, if $x \in \mathcal{X}$, and $+\infty$ otherwise.

Remark 2.2. Let us briefly discuss the two conditions laid out in Assumption 2.1. As stated, we only require one of these two conditions to hold. Condition (B1) is very mild, and implies that the oracle error has a first-order moment independent of x . This is very natural for oracles considered in this work. To see this, assume, for example, that $F(x, \xi) = \min_{y \in \mathcal{Y}(x, \xi)} G(x, y, \xi)$, where $\mathcal{Y}(x, \xi) \subset \mathbb{R}^m$, in which case evaluating F requires the solution of an optimization problem (noting that similar problems are considered in this paper and that the discussion can symmetrically apply to maximization problems as well). Then, the oracle is a numerical algorithm that performs this optimization up to some prespecified tolerance, say $\tilde{\delta}$, in the sense that it returns $\tilde{F}(x, \xi)$, such that

$$\tilde{F}(x, \xi) - F(x, \xi) \leq \tilde{\delta},$$

for all $x \in \mathbb{R}^n$ and any $\xi \in \Xi$. Since this tolerance is independent of x , we will always observe (for any $x \in \mathbb{R}^n$) that $\tilde{F}(x, \xi) = F(x, \xi) + \delta(x, \xi)$, where, in this example, $\delta(x, \xi) \leq \tilde{\delta}$ is some positive random variable. The assumption relies on the intuition that the range of values of $\delta(\cdot, \cdot)$ should be independent of x , and requires that $\mathbb{E}_{\xi} \{\delta(x, \xi)\} = \Delta$, for any $x \in \mathbb{R}^n$. This is consistent with practice, assuming that the optimization algorithm (oracle) terminates after reaching an optimality gap of $\tilde{\delta}$, irrespectively of x . Indeed, in that case, and for any fixed x , the oracle should return an evaluation such that $\tilde{F}(x, \xi) - F(x, \xi) \leq \tilde{\delta}$ and thus averaging those evaluation differences over ξ should yield some constant Δ , and the intuition behind condition (B1) is that this constant should not depend on x (something that would naturally hold if, e.g., the minimization problem with respect to y were convex). If we assume that condition (B1) does

not hold, we instead assume, in condition **(B2)**, that x is constrained on a convex and compact set, with bounded diameter, say $D > 0$.

3. TERMINATION CRITERIA FOR NONSMOOTH AND NONCONVEX STOCHASTIC COMPOSITE OPTIMIZATION

In the context of nonsmooth and nonconvex optimization (e.g., see problem **(P)**), it is important to establish a practical and useful metric for measuring progress of an optimization algorithm. To that end, we list two important notions which will be useful in this work, and based upon which we will derive a novel notion of approximate stationarity suitable for nonsmooth and nonconvex stochastic composite optimization of the form of **(P)**.

(μ, ε) –Goldstein stationary points. Consider problem **(P)**, and assume that $r = 0$. In the context of nonsmooth and nonconvex optimization of Lipschitz continuous functions, it is well-known that finding an approximate Clarke-stationary point of f (or, equivalently, an ε –Clarke stationary point of f), i.e., a point $x \in \mathbb{R}^n$ such that

$$\min\{\|g\| \mid g \in \bar{\partial}f(x)\} \leq \varepsilon,$$

where $\varepsilon > 0$ is some pre-specified tolerance, is intractable (e.g., see [1]). Instead, following [1], one typically utilizes the notion of (μ, ε) –Goldstein stationary points. Specifically, a point $x \in \mathbb{R}^n$ is said to be a (μ, ε) –Goldstein stationary point if the following inequality is satisfied:

$$\min\{\|g\| \mid g \in \bar{\partial}_\mu f(x)\} \leq \varepsilon.$$

(μ, ε) –Moreau envelope stationary points. Next, let us consider problem **(P)**, and assume that f is ρ –weakly convex, for some $\rho > 0$ (i.e., $f \in \Gamma_\rho(\mathbb{R}^n)$). In this case, the Moreau envelope of the composite objective function, $e_\mu \phi$, is well-defined and continuously differentiable for all $\mu < \rho^{-1}$, and is known to serve as a good measure of near-stationarity (see [14]). Specifically, we say that $x \in \mathbb{R}^n$ is an (μ, ε) –Moreau envelope stationary point for **(P)** if it satisfies:

$$\|\nabla e_\mu \phi(x)\| \leq \varepsilon,$$

with the assumption that $\mu < \rho^{-1}$. In that case, one can show that if x is a (μ, ε) –Moreau envelope stationary point, then it is close to an ε –Clarke stationary point of ϕ . Specifically, if we let $\hat{x} = \text{prox}_{\mu\phi}(x)$, then the following relations hold:

$$\begin{cases} \|x - \hat{x}\| = \mu \|\nabla e_\mu \phi(x)\| \\ \phi(\hat{x}) \leq \phi(x) \\ \min\{\|g\| \mid g \in \bar{\partial}\phi(\hat{x})\} \leq \|\nabla e_\mu \phi(x)\| \end{cases}.$$

Surrogate $(\lambda, \mu, \varepsilon)$ –Moreau envelope stationary points. In this work, we will, implicitly, make use of both notions of near stationarity. Indeed, we do not assume that f in **(P)** is weakly convex, and thus we cannot make direct use of the (μ, ε) –Moreau envelope stationarity. On the other hand, we consider problems for which $r \neq 0$ and it is allowed to be extended-valued (and thus not Lipschitz continuous). As a result we cannot make direct use of the (μ, ε) –Goldstein stationarity. Instead, we will attempt to generalize both notions and combine them using an appropriate surrogate function. Specifically, by utilizing a smooth surrogate of f based on randomized uniform smoothing, we focus on solving the following ρ –weakly convex optimization

problem:

$$\min_{x \in \mathbb{R}^n} \phi_\mu(x) \triangleq f_\mu(x) + r(x), \quad f_\mu(x) \triangleq \mathbb{E}_{\xi, U \sim \mathcal{U}(\mathbb{B}_1(0))} \{F(x + \mu U, \xi)\}, \quad (\mathbf{P}_\mu)$$

where, using Lemma 2.1 and Assumption 1.1, we have that $\phi_\mu \in \Gamma_\rho(\mathbb{R}^n)$, with $\rho = cG\sqrt{n}\mu^{-1}$ (see also Lemma 4.3). Using these facts, we are now able to define the proposed notion of a *surrogate $(\lambda, \mu, \varepsilon)$ -Moreau envelope stationary point*.

Definition 3.1 (Surrogate $(\lambda, \mu, \varepsilon)$ -Moreau envelope stationarity). Consider problem $(\mathbf{P})$ and let $x \in \text{dom } r$. We say that x is a *surrogate $(\lambda, \mu, \varepsilon)$ -Moreau envelope stationary point* for $(\mathbf{P})$ if for some $\varepsilon > 0$ and some $\mu > 0$, it holds that

$$\|\nabla e_\lambda \phi_\mu(x)\| \leq \varepsilon,$$

with $\lambda < \rho^{-1}$ and $\rho = cG\sqrt{n}\mu^{-1}$, where ϕ_μ is defined as in $(\mathbf{P}_\mu)$, and ρ is the weak convexity constant of ϕ_μ .

We now proceed to explain why the surrogate $(\lambda, \mu, \varepsilon)$ -Moreau envelope stationarity is a suitable approximate stationarity condition for $(\mathbf{P})$. We start by noting that for any $x \in \text{dom } r \equiv \text{dom } \phi_\mu$,

$$\bar{\partial} \phi_\mu(x) = \hat{\partial} \phi_\mu(x) = \partial \phi_\mu(x) = \nabla f_\mu(x) + \partial r(x),$$

where the first three equalities follow from the fact that ϕ_μ is subdifferentially regular (as a weakly convex function; see [41, Example 10.32]) and the second equality follows from [41, Exercise 8.8]. Let $x^* \in \text{dom } \phi_\mu$ satisfying $\|\nabla e_\lambda \phi_\mu(x^*)\| \leq \varepsilon$ for some $\varepsilon > 0$, such that $\lambda < \rho^{-1} = \mu/(cG\sqrt{n})$. Then, we observe that if $\hat{x} = \text{prox}_{\lambda \phi_\mu}(x^*)$, we have that

$$\|x^* - \hat{x}\| \leq \lambda \varepsilon, \quad \min\{\|g\| \mid g \in \bar{\partial} \phi_\mu(\hat{x})\} \leq \varepsilon.$$

Hence, we may observe that

$$\min\{\|g\| \mid g \in \bar{\partial}_{\lambda \varepsilon} \phi_\mu(x^*)\} \leq \varepsilon.$$

Indeed, since $\|\hat{x} - x^*\| \leq \lambda \varepsilon$, and $\bar{\partial}_{\lambda \varepsilon} \phi_\mu(x^*) = \text{conv}\left(\bigcup_{y \in \mathbb{B}_{\lambda \varepsilon}(x^*)} \bar{\partial} \phi_\mu(y)\right)$, we have that

$$\hat{x} \in \mathbb{B}_{\lambda \varepsilon}(x^*), \quad \min\{\|g\| \mid g \in \bar{\partial} \phi_\mu(\hat{x})\} \leq \varepsilon.$$

Thus, it must be true that x^* is a $(\lambda \varepsilon, \varepsilon)$ -Goldstein stationary point for ϕ_μ . For example, and without loss of generality, we may assume that $\lambda \varepsilon \leq \mu$, in which case x^* is a (μ, ε) -Goldstein stationary point for ϕ_μ (noting that, in general, we expect that $\lambda \varepsilon \ll \mu$). Let us now see how this translates to the original problem in the special case where $r = 0$. In that case, we have, from [9, Lemma 4], that

$$\bar{\partial}_{\lambda \varepsilon} \phi_\mu(x^*) \subseteq \bar{\partial}_{\lambda \varepsilon + \mu} \phi(x^*),$$

or, in other words, if $r = 0$ we obtain that x^* is a $(\lambda \varepsilon + \mu, \varepsilon)$ -Goldstein stationary point for ϕ .

Remark 3.1. Let us also briefly discuss the relation of the Goldstein subdifferential of the surrogate function ϕ_μ to the original function ϕ . Since ϕ_μ is ρ -weakly convex, with $\rho = cG\sqrt{n}\mu^{-1}$, then (using Lemma 2.1) any $v \in \bar{\partial} \phi_\mu(x)$ satisfies, for all $x, y \in \text{dom } \phi_\mu \equiv \text{dom } \phi$:

$$\begin{aligned} \phi(y) + \mu G &\geq \phi_\mu(y) \geq \phi_\mu(x) + v^\top(y - x) - \frac{\rho}{2}\|y - x\|^2 \\ &\geq \phi(x) + v^\top(y - x) - \frac{\rho}{2}\|y - x\|^2, \end{aligned}$$

thus implying that $v \in \bar{\partial}_{\mu G}^{\text{eps}} \phi(x)$, where $\bar{\partial}_{\mu G}^{\text{eps}} \phi(x)$ denotes the *epsilon-(regular) subdifferential* of ϕ at x , with G being the Lipschitz continuity constant of ϕ . Hence, for any $x \in \text{dom } \phi$, we can easily show that the $\lambda \varepsilon$ –Goldstein subdifferential of the surrogate ϕ_μ (which is essentially employed in our proposed approximate stationarity condition) satisfies

$$\bar{\partial}_{\lambda \varepsilon} \phi_\mu(x) = \text{conv} \left(\bigcup_{y \in \mathbb{B}_{\lambda \varepsilon}(x)} \bar{\partial} \phi_\mu(y) \right) \subseteq \text{conv} \left(\bigcup_{y \in \mathbb{B}_{\lambda \varepsilon}(x)} \bar{\partial}_{\mu G}^{\text{eps}} \phi(y) \right).$$

In other words, the $\lambda \varepsilon$ –Goldstein subdifferential of the surrogate ϕ_μ , at some $x \in \text{dom } \phi$, is a subset of an enlargement of the $\lambda \varepsilon$ –Goldstein subdifferential of ϕ , obtained by substituting, in the definition of the Goldstein subdifferential, the regular subdifferential by the μG –epsilon-regular subdifferential (as defined above). Although the latter construct might be an artificially large set in general, our discussion above showcases that the particular subset obtained by our proposed approximate stationarity condition is a sensible and appropriate choice in the context of nonsmooth and nonconvex stochastic composite optimization problems considered herein.

4. AN INEXACT ZERO-ORDER METHOD FOR (P)

We are now ready to derive and analyze our proposed algorithm for the solution of (P), under Assumption 1.1, where F is only accessible via an inexact noisy oracle (cf. Definition 2.2). The algorithm will be analyzed under two general assumptions, by further imposing Assumption 2.1 (i.e., either imposing a natural first-moment stationarity property of the oracle, or assuming that the optimization of (P) is performed over a convex and compact set $\mathcal{X}$). We begin by stating the proposed method in Algorithm Z-iProxSG.

Algorithm Z-iProxSG Zeroth-order inexact Proximal Stochastic Gradient method

Input: $x_0 \in \text{dom}(r)$, a sequence $\{\alpha_t\}_{t \geq 0} \subset \mathbb{R}_+$, $\mu > 0$, and $T > 0$.

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

Sample (i.i.d.) $\xi_t \in \Xi$, $W_t \sim \mathcal{U}(\mathbb{S}^{n-1})$.

Compute and store two oracle evaluations: $\tilde{F}(x_t + \mu W_t, \xi_t)$, $\tilde{F}(x_t - \mu W_t, \xi_t)$.

Set

$$x_{t+1} = \text{prox}_{\alpha_t r}(x_t - \alpha_t G(x_t, \xi_t, W_t; \mu)),$$

where $G_t \equiv G(x_t, \xi_t, W_t; \mu) \triangleq \frac{n}{2\mu} (\tilde{F}(x_t + \mu W_t, \xi_t) - \tilde{F}(x_t - \mu W_t, \xi_t)) W_t$.

end for

Sample $t^* \in \{0, \dots, T\}$ according to $\mathbb{P}(t^* = t) = \frac{\alpha_t}{\sum_{i=0}^T \alpha_i}$.

return x_{t^*} .

4.1. Technical results. We begin by stating certain important technical results that will be instrumental in analyzing the non-asymptotic convergence of Algorithm Z-iProxSG. We start by bounding the quantity $\mathbb{E}\{\|G_t\|^2 \mid \xi_0, W_0, \dots, \xi_{t-1}, W_{t-1}\}$, where $G_t \equiv G(x_t, \xi_t, W_t; \mu)$ corresponds to the (biased and noisy) stochastic gradient estimator appearing in Algorithm Z-iProxSG, at iteration $t \geq 0$. For simplicity of notation, we define the expectation operator with respect to the filtration up to time t as $\mathbb{E}_{[t]} \{\cdot\} \triangleq \mathbb{E} \{\cdot \mid \xi_0, W_0, \dots, \xi_{t-1}, W_{t-1}\}$.

Lemma 4.1. Consider problem (P) and let Assumption 1.1 hold. Let also $\{x_t\}_{t=0}^T$ be the sequence of iterates generated by Algorithm Z-iProxSG. Then, we have that

$$\mathbb{E}_{[t]} \{ \|G_t\|^2 \} \equiv \mathbb{E}_{[t]} \{ \|G(x_t, \xi_t, W_t; \mu)\|^2 \} \leq 32\sqrt{2\pi}nG^2 + \frac{n^2}{\mu^2}2\tilde{\delta}^2,$$

where $G > 0$ is the constant appearing in Assumption 1.1 and $\tilde{\delta} > 0$ is the oracle error bound given in Definition 2.2.

Proof. The proof follows by extending the proof of [6, Lemma E.1], upon noting that our stochastic oracle $\tilde{F}$ is noisy and biased. We start by noting that

$$\begin{aligned} \mathbb{E}_{[t]} \{ \|G_t\|^2 \} &= \mathbb{E}_{[t]} \left\{ \left\| \frac{n}{2\mu} (\tilde{F}(x_t + \mu W_t, \xi_t) - \tilde{F}(x_t - \mu W_t, \xi_t)) W_t \right\|^2 \right\} \\ &= \frac{n^2}{4\mu^2} \mathbb{E}_{[t]} \left\{ \|W_t\|^2 (F(x_t + \mu W_t, \xi_t) + \delta(x_t + \mu W_t, \xi_t) - F(x_t - \mu W_t, \xi_t) - \delta(x_t - \mu W_t, \xi_t))^2 \right\} \\ &\leq \frac{n^2}{2\mu^2} \left(\mathbb{E}_{[t]} \left\{ \|W_t\|^2 (F(x_t + \mu W_t, \xi_t) - F(x_t - \mu W_t, \xi_t))^2 \right\} \right. \\ &\quad \left. + \mathbb{E}_{[t]} \left\{ \|W_t\|^2 (\delta(x_t + \mu W_t, \xi_t) - \delta(x_t - \mu W_t, \xi_t))^2 \right\} \right), \end{aligned}$$

where the inequality follows from the identity $(a + b)^2 \leq 2a^2 + 2b^2$. In order to bound the first term in the right hand side of the previous inequality, we follow exactly the analysis in [6, Lemma E.1], to obtain that

$$\mathbb{E}_{[t]} \left\{ \|W_t\|^2 (F(x_t + \mu W_t, \xi_t) - F(x_t - \mu W_t, \xi_t))^2 \right\} \leq \frac{64\sqrt{2\pi}\mu^2 G^2}{n}. \quad (1)$$

For the second term, we use the definition of the oracle (cf. Definition 2.2) to obtain

$$\mathbb{E}_{[t]} \left\{ \|W_t\|^2 (\delta(x_t + \mu W_t, \xi_t) - \delta(x_t - \mu W_t, \xi_t))^2 \right\} \leq 4\tilde{\delta}^2, \quad (2)$$

where we used the identity $(a - b)^2 \leq 2a^2 + 2b^2$, the fact that $\|W_t\|^2 = 1$ (since $W_t \sim \mathcal{U}(\mathbb{S}^{n-1})$), as well as the uniform bound on the oracle noise, $|\delta(x, \xi)| \leq \tilde{\delta}$, for any $x \in \mathbb{R}^n$ and a.e. $\xi \in \Xi$. By combining (1) and (2), we obtain the result. $\square$

Lemma 4.2. Consider problem (P) and let Assumption 1.1 hold. Let also $\{x_t\}_{t=0}^T$ be the sequence of iterates generated by Algorithm Z-iProxSG. Then, for any $r \in \mathbb{R}^n$, we have that:

- If condition (B1) of Assumption 2.1 holds,

$$\mathbb{E}_{[t]} \left\{ r^\top G_t \right\} = \mathbb{E}_{[t]} \left\{ r^\top \nabla f_\mu(x_t) \right\};$$

- If, instead, condition (B2) of Assumption 2.1 holds, and $r = x_1 - x_2$, for $x_1, x_2 \in \mathcal{X}$,

$$\mathbb{E}_{[t]} \left\{ r^\top G_t \right\} \geq \mathbb{E}_{[t]} \left\{ r^\top \nabla f_\mu(x_t) \right\} - \frac{n}{\mu} \tilde{\delta} D,$$

where D is the diameter of $\mathcal{X}$.

Proof. We begin by proving the first part of the result, which relies on condition **(B1)** of Assumption 2.1, which implies that for any $x \in \mathbb{R}^n$, $\mathbb{E}_\xi \{\delta(x, \xi)\} = \Delta$, for some constant Δ . Then, for any $r \in \mathbb{R}^n$, we have

$$\begin{aligned} \mathbb{E}_{[t]} \left\{ r^\top G_t \right\} &= \frac{n}{2\mu} \mathbb{E}_{[t]} \left\{ r^\top (F(x + \mu W_t, \xi_t) - F(x - \mu W_t, \xi_t)) W_t \right\} \\ &\quad + \frac{n}{2\mu} \mathbb{E}_{[t]} \left\{ r^\top (\delta(x + \mu W_t, \xi_t) - \delta(x - \mu W_t, \xi_t)) W_t \right\}. \end{aligned}$$

From [39, Lemma 8] and the symmetry of W_t , we obtain that the first term above satisfies

$$\frac{n}{2\mu} \mathbb{E}_{[t]} \left\{ r^\top (F(x + \mu W_t, \xi_t) - F(x - \mu W_t, \xi_t)) W_t \right\} = \mathbb{E}_{[t]} \left\{ r^\top \nabla f_\mu(x_t) \right\},$$

noting that this holds irrespectively of whether condition **(B1)** or **(B2)** of Assumption 2.1 holds. For the second term, using condition **(B1)**, we have

$$\begin{aligned} &\frac{n}{2\mu} \mathbb{E}_{[t]} \left\{ r^\top (\delta(x + \mu W_t, \xi_t) - \delta(x - \mu W_t, \xi_t)) W_t \right\} \\ &= \frac{n}{2\mu} \mathbb{E}_{[t]} \left\{ \mathbb{E}_{[t]} \left\{ r^\top (\delta(x + \mu W_t, \xi_t) - \delta(x - \mu W_t, \xi_t)) W_t \mid W_t \right\} \right\} = 0. \end{aligned}$$

To prove the second part of the lemma, using condition **(B2)** instead, we observe that

$$\frac{n}{2\mu} \mathbb{E}_{[t]} \left\{ r^\top (\delta(x + \mu W_t, \xi_t) - \delta(x - \mu W_t, \xi_t)) W_t \right\} \leq \frac{n}{2\mu} \|r\| 2\tilde{\delta} \leq \frac{n}{\mu} \tilde{\delta} D,$$

where we used the fact that $|\delta(x, \xi)| \leq \tilde{\delta}$ for all $x \in \mathbb{R}^n$ and a.e. $\xi \in \Xi$, and $r = x_1 - x_2$, with $x_1, x_2 \in \mathcal{X}$, noting that $\mathcal{X}$ has diameter $D > 0$ (from condition **(B2)**). $\square$

Lemma 4.3. Fix some $\mu > 0$ and let Assumption 1.1 hold. Then, $\phi_\mu \triangleq f_\mu + r \in \Gamma_\rho(\mathbb{R}^n)$, where $\rho = \frac{cG\sqrt{n}}{\mu}$, with c a bounded constant independent of n , and G the constant given in Assumption 1.1. Moreover, if we let $\hat{x}_t \triangleq \text{prox}_{\bar{\rho}^{-1}\phi_\mu}(x_t)$, where x_t is the iterate generated by Algorithm Z-iProxSG at time $t \geq 0$ and $\bar{\rho} \in (\rho, 2\rho]$, then,

$$\hat{x}_t = \text{prox}_{\alpha_t r}(\alpha_t \bar{\rho} x_t - \alpha_t \nabla f_\mu(\hat{x}_t) + \zeta_t \hat{x}_t),$$

where α_t is the step-size of Algorithm Z-iProxSG at time t and $\zeta_t \triangleq 1 - \alpha_t \bar{\rho}$.

Proof. We start by noting that, from Lemma 2.1, f_μ is ρ -Lipschitz smooth (where we used the fact that f is G -Lipschitz continuous from Assumption 1.1), and thus ρ -weakly convex (see, e.g., [42, Proposition 4.12]). Then, from [42, Proposition 4.1], we have that $\phi_\mu = f_\mu + r$ must also be ρ -weakly convex, since $r \in \Gamma_0(\mathbb{R}^n)$ from Assumption 1.1. Finally, by letting $\hat{x}_t = \text{prox}_{\bar{\rho}^{-1}\phi_\mu}(x_t)$, we have, by definition, that

$$\begin{aligned} \alpha_t \bar{\rho}(x_t - \hat{x}_t) \in \alpha_t \partial r(\hat{x}_t) + \alpha_t \nabla f_\mu(\hat{x}_t) &\Leftrightarrow \alpha_t \bar{\rho} x_t - \alpha_t \nabla f_\mu(\hat{x}_t) + \zeta_t \hat{x}_t \in \hat{x}_t + \alpha_t \partial r(\hat{x}_t) \\ &\Leftrightarrow \hat{x}_t = \text{prox}_{\alpha_t r}(\alpha_t \bar{\rho} x_t - \alpha_t \nabla f_\mu(\hat{x}_t) + \zeta_t \hat{x}_t), \end{aligned}$$

where $\zeta_t = 1 - \alpha_t \bar{\rho}$. $\square$

4.2. Convergence analysis. We are now ready to derive a non-asymptotic convergence analysis of Algorithm **Z-iProxSG**. We will analyze the algorithm based on the surrogate problem $(\mathbf{P}_\mu)$, i.e. $\min_{x \in \mathbb{R}^n} \phi_\mu(x)$, for some fixed $\mu > 0$. Upon noting, from Lemma 4.3, that ϕ_μ is weakly convex, the analysis will follow by extending the analysis given in [16], by allowing inexact oracle evaluations. Then, we will briefly discuss conditions on the oracle error that allow us to retrieve convergence rates appearing in the literature in the context of exact stochastic oracles (matching the rates currently available only in the unconstrained case, i.e., in the case where $r = 0$).

Lemma 4.4. *Consider problem $(\mathbf{P})$ and let Assumption 1.1 hold. Let also $\{x_t\}_{t=0}^T$ be the sequence of iterates of Algorithm **Z-iProxSG**. Set $\bar{\rho} \in (\rho, 2\rho]$, where $\rho = \frac{cG\sqrt{n}}{\mu}$ and choose $\alpha_t \in (0, \bar{\rho}^{-1}]$, for any $t \geq 0$. Then:*

- *If condition **(B1)** of Assumption 2.1 holds, the following inequality is satisfied:*

$$\mathbb{E}_{[t]} \{ \|x_{t+1} - \hat{x}_t\|^2 \} \leq (1 - (2\alpha_t(\bar{\rho} - \rho))) \|x_t - \hat{x}_t\|^2 + 8\alpha_t^2 \left(16\sqrt{2\pi n}G^2 + \frac{n^2}{\mu^2} \tilde{\delta}^2 \right).$$

- *If, instead, condition **(B2)** of Assumption 2.1 holds, then the following inequality is satisfied:*

$$\begin{aligned} \mathbb{E}_{[t]} \{ \|x_{t+1} - \hat{x}_t\|^2 \} &\leq (1 - (2\alpha_t(\bar{\rho} - \rho))) \|x_t - \hat{x}_t\|^2 \\ &\quad + 2\zeta_t \alpha_t \frac{n}{\mu} \tilde{\delta} D + 4\alpha_t^2 \left(16\sqrt{2\pi n}G^2 + \frac{n^2}{\mu^2} \tilde{\delta}^2 \right). \end{aligned}$$

Proof. By definition, we have that $\hat{x}_t = \mathbf{prox}_{\bar{\rho}^{-1}\phi_\mu}(x_t)$. Thus,

$$\begin{aligned} \mathbb{E}_{[t]} \{ \|x_{t+1} - \hat{x}_t\|^2 \} &= \mathbb{E}_{[t]} \left\{ \left\| \mathbf{prox}_{\alpha_t r}(x_t - \alpha_t G_t) - \mathbf{prox}_{\alpha_t r}(\alpha_t \bar{\rho} - \alpha_t \nabla f_\mu(x_t) + \zeta_t \hat{x}_t) \right\|^2 \right\} \\ &\leq \mathbb{E}_{[t]} \left\{ \left\| (x_t - \alpha_t G_t) - (\alpha_t \bar{\rho} - \alpha_t \nabla f_\mu(x_t) + \zeta_t \hat{x}_t) \right\|^2 \right\} \\ &= \zeta_t^2 \|x_t - \hat{x}_t\|^2 - 2\zeta_t \alpha_t \mathbb{E}_{[t]} \left\{ (x_t - \hat{x}_t)^\top (G_t - \nabla f_\mu(\hat{x}_t)) \right\} + \alpha_t^2 \mathbb{E}_{[t]} \{ \|G_t - \nabla f_\mu(\hat{x}_t)\|^2 \}, \end{aligned}$$

where the first equality follows from Algorithm **Z-iProxSG** and from Lemma 4.3, and the inequality follows from the non-expansiveness of the proximity operator of r . Next, we separate two cases.

Case 1: Under condition **(B1)** of Assumption 2.1, by utilizing Lemmata 4.1 and 4.2, we have

$$\begin{aligned} \mathbb{E}_{[t]} \{ \|x_{t+1} - \hat{x}_t\|^2 \} &\leq \zeta_t^2 \|x_t - \hat{x}_t\|^2 - 2\zeta_t \alpha_t (x_t - \hat{x}_t)^\top (\nabla f_\mu(x_t) - \nabla f_\mu(\hat{x}_t)) + 8\alpha_t^2 \left(16\sqrt{2\pi n}G^2 + \frac{n^2}{\mu^2} \tilde{\delta}^2 \right) \\ &\leq \zeta_t^2 \|x_t - \hat{x}_t\|^2 + 2\zeta_t \alpha_t \rho \|x_t - \hat{x}_t\|^2 + 8\alpha_t^2 \left(16\sqrt{2\pi n}G^2 + \frac{n^2}{\mu^2} \tilde{\delta}^2 \right) \\ &= (1 - (2\alpha_t(\bar{\rho} - \rho) + \alpha_t^2 \bar{\rho}(2\rho - \bar{\rho}))) \|x_t - \hat{x}_t\|^2 + 8\alpha_t^2 \left(16\sqrt{2\pi n}G^2 + \frac{n^2}{\mu^2} \tilde{\delta}^2 \right), \end{aligned}$$

where, in the last inequality, we used the fact that f_μ is ρ -weakly convex, with $\rho = cG\sqrt{n}\mu^{-1}$ (cf. Lemma 4.3), which implies (e.g., from [42, Proposition 4.10]) that

$$(x_1 - x_2)^\top (\nabla f_\mu(x_1) - \nabla f_\mu(x_2)) \geq -\rho \|x_1 - x_2\|^2, \quad \text{for all } x_1, x_2 \in \mathbb{R}^n.$$

The first inequality then follows by noting that $\bar{\rho} \leq 2\rho$.

Case 2: Similarly, under condition **(B2)** of Assumption 2.1, by utilizing once again Lemmata 4.1 and 4.2, we have

$$\begin{aligned}
& \mathbb{E}_{[t]} \left\{ \|x_{t+1} - \hat{x}_t\|^2 \right\} \\
& \leq \zeta_t^2 \|x_t - \hat{x}_t\|^2 - 2\zeta_t \alpha_t (x_t - \hat{x}_t)^\top (\nabla f_\mu(x_t) - \nabla f_\mu(\hat{x}_t)) + 2\zeta_t \alpha_t \frac{n}{\mu} \tilde{\delta} D \\
& \quad + 8\alpha_t^2 \left(16\sqrt{2\pi}nG^2 + \frac{n^2}{\mu^2} \tilde{\delta}^2 \right) \\
& \leq \zeta_t^2 \|x_t - \hat{x}_t\|^2 + 2\zeta_t \alpha_t \rho \|x_t - \hat{x}_t\|^2 + 2\zeta_t \alpha_t \frac{n}{\mu} \tilde{\delta} D + 8\alpha_t^2 \left(16\sqrt{2\pi}nG^2 + \frac{n^2}{\mu^2} \tilde{\delta}^2 \right) \\
& = (1 - (2\alpha_t(\bar{\rho} - \rho) + \alpha_t^2 \bar{\rho}(2\rho - \bar{\rho}))) \|x_t - \hat{x}_t\|^2 + 2\zeta_t \alpha_t \frac{n}{\mu} \tilde{\delta} D + 8\alpha_t^2 \left(16\sqrt{2\pi}nG^2 + \frac{n^2}{\mu^2} \tilde{\delta}^2 \right),
\end{aligned}$$

where D is the diameter of $\mathcal{X}$ (cf. condition **(B2)** of Assumption 2.1) and $\tilde{\delta}$ is the bound on the oracle noise (cf. Definition 2.2). $\square$

We are now ready to derive the non-asymptotic (ergodic) convergence rate of Algorithm **Z-iProxSG** in terms of the magnitude of the gradient of the Moreau envelope of ϕ_μ . We will provide two different rates, based on either condition **(B1)** or **(B2)** of Assumption 2.1.

Theorem 4.1. Fix $\mu > 0$ and consider problem **(P)** by letting Assumption 1.1 hold. Let also $\{x_t\}_{t=0}^T$ be the sequence of iterates of Algorithm **Z-iProxSG**, with x_{t^*} being the point returned by the method. If condition **(B1)** of Assumption 2.1 holds, then by letting $\Phi \geq e_{(2\rho)-1} \phi_\mu(x_0) - \min_{x \in \mathbb{R}^n} \phi_\mu(x)$ (with $\Phi > 0$) and choosing

$$\alpha_t = \sqrt{\frac{\Phi}{8cGn^{3/2}\mu^{-1} \left(16\sqrt{2\pi}G^2 + \frac{n}{\mu^2} \tilde{\delta}^2 \right) (T+1)}},$$

we obtain that

$$\mathbb{E} \left\{ \left\| \nabla e_{(2\rho)-1} \phi_\mu(x_{t^*}) \right\|^2 \right\} \leq 4 \sqrt{\frac{8cG\Phi n^{3/2}\mu^{-1} \left(16\sqrt{2\pi}G^2 + \frac{n}{\mu^2} \tilde{\delta}^2 \right)}{T+1}},$$

where c is a constant independent of n (cf. Lemma 2.1), G is the Lipschitz continuity constant of f (cf. Assumption 1.1), and $\tilde{\delta}$ is the bound on the oracle noise (cf. Definition 2.2).

Alternatively, if condition **(B2)** of Assumption 2.1 holds instead, then, for the same choice of step-size, we obtain

$$\mathbb{E} \left\{ \left\| \nabla e_{(2\rho)-1} \phi_\mu(x_{t^*}) \right\|^2 \right\} \leq 4 \sqrt{\frac{8cG\Phi n^{3/2}\mu^{-1} \left(16\sqrt{2\pi}G^2 + \frac{n}{\mu^2} \tilde{\delta}^2 \right)}{T+1}} + 4\tilde{\delta}D\frac{n}{\mu},$$

where D is the diameter of $\mathcal{X}$ (cf. Assumption 2.1).

Proof. From Lemma 4.4, we have that

$$\begin{aligned} \mathbb{E}_{[t]} \left\{ e_{\bar{\rho}-1} \phi_{\mu}(x_{t+1}) \right\} &\leq \mathbb{E}_{[t]} \left\{ \phi_{\mu}(\hat{x}_t) + \frac{\bar{\rho}}{2} \|\hat{x}_t - x_{t+1}\|^2 \right\} \\ &\leq \phi_{\mu}(\hat{x}_t) + \frac{\bar{\rho}}{2} \left(\|x_t - \hat{x}_t\|^2 - 2\alpha_t(\bar{\rho} - \rho) \|x_t - \hat{x}_t\|^2 + 8\alpha_t^2 \left(16\sqrt{2\pi n}G^2 + \frac{n^2}{\mu^2} \tilde{\delta}^2 \right) \right) \\ &= e_{\bar{\rho}-1} \phi_{\mu}(x_t) + \bar{\rho} \left(-\alpha_t(\bar{\rho} - \rho) \|x_t - \hat{x}_t\|^2 + 4\alpha_t^2 \left(16\sqrt{2\pi n}G^2 + \frac{n^2}{\mu^2} \tilde{\delta}^2 \right) \right), \end{aligned}$$

where the first inequality follows from the definition of the Moreau envelope of ϕ_{μ} , and the equality follows from the definition of $\hat{x}_t$ (cf. Lemma 4.3). By the definition of $\hat{x}_t$, and since ϕ_{μ} is ρ -weakly convex and $\bar{\rho} > \rho$, we may use [43, Theorem 3.4] to obtain that

$$\nabla e_{\bar{\rho}-1} \phi_{\mu}(x_t) = \frac{1}{\bar{\rho}} (x_t - \hat{x}_t).$$

Using the previous fact, we next take expectation with respect to the filtration $\xi_0, W_0, \dots, \xi_{t-1}, W_{t-1}$ and use the law of total expectation to obtain

$$\begin{aligned} \mathbb{E} \left\{ e_{\bar{\rho}-1} \phi_{\mu}(x_{t+1}) \right\} &\leq \mathbb{E} \left\{ e_{\bar{\rho}-1} \phi_{\mu}(x_t) \right\} - \frac{\alpha_t(\bar{\rho} - \rho)}{\bar{\rho}} \left\| \nabla e_{\bar{\rho}-1} \phi_{\mu}(x_t) \right\|^2 \\ &\quad + 4\alpha_t^2 \bar{\rho} \left(16\sqrt{2\pi n}G^2 + \frac{n^2}{\mu^2} \tilde{\delta}^2 \right). \end{aligned}$$

By unrolling the above recursion, we have

$$\begin{aligned} \mathbb{E} \left\{ e_{\bar{\rho}-1} \phi_{\mu}(x_{t+1}) \right\} &\leq e_{\bar{\rho}-1} \phi_{\mu}(x_0) - \frac{\bar{\rho} - \rho}{\bar{\rho}} \sum_{t=0}^T \alpha_t \left\| \nabla e_{\bar{\rho}-1} \phi_{\mu}(x_t) \right\|^2 \\ &\quad + 4\bar{\rho} \left(16\sqrt{2\pi n}G^2 + \frac{n^2}{\mu^2} \tilde{\delta}^2 \right) \sum_{t=0}^T \alpha_t^2. \end{aligned}$$

We can lower bound the left-hand side of the above inequality by $\phi_{\mu}^* \triangleq \min_{x \in \mathbb{R}^n} \phi_{\mu}(x)$, and re-arrange to get

$$\frac{1}{\sum_{t=0}^T \alpha_t} \sum_{t=0}^T \alpha_t \left\| \nabla e_{\bar{\rho}-1} \phi_{\mu}(x_t) \right\|^2 \leq \frac{\bar{\rho}}{\bar{\rho} - \rho} \frac{e_{\bar{\rho}-1} \phi_{\mu}(x_0) - \phi_{\mu}^* + 4\bar{\rho} \left(16\sqrt{2\pi n}G^2 + \frac{n^2}{\mu^2} \tilde{\delta}^2 \right) \sum_{t=0}^T \alpha_t^2}{\sum_{t=0}^T \alpha_t}.$$

We observe that the left-hand side of the last inequality is nothing else than $\mathbb{E} \left\{ \left\| \nabla e_{\bar{\rho}-1} \phi_{\mu}(x_{t^*}) \right\|^2 \right\}$, where x_{t^*} is the iterate that Algorithm **Z-iProxSG** returns. To complete the proof, we set $\bar{\rho} = 2\rho$, we let $\Phi \geq e_{(2\rho)-1} \phi_{\mu}(x_0) - \phi_{\mu}^*$ such that $\Phi > 0$, and set $\alpha_t = \gamma/\sqrt{T+1}$, for some $\gamma > 0$ to obtain

$$\mathbb{E} \left\{ \left\| \nabla e_{\bar{\rho}-1} \phi_{\mu}(x_{t^*}) \right\|^2 \right\} \leq 2 \frac{\Phi + 8\rho \left(16\sqrt{2\pi n}G^2 + \frac{n^2}{\mu^2} \tilde{\delta}^2 \right)}{\gamma\sqrt{T+1}}.$$

Then, we minimize over γ , which yields that

$$\gamma = \sqrt{\frac{\Phi}{8\rho \left(16\sqrt{2\pi n}G^2 + \frac{n^2}{\mu^2} \tilde{\delta}^2 \right)}},$$

and completes the first part of the proof, upon noting that $\rho = cG\sqrt{n}\mu^{-1}$. For the second part of the proof, we assume (without loss of generality) that $\rho^{-1} > \alpha_t$ (where α_t is given in the statement of the theorem). Then, $0 < \zeta_t \leq 1$ (where ζ_t is defined in Lemma 4.3) and the “descent” recursion carries an additional term of the form $4\frac{n}{\mu}\tilde{\delta}D\alpha_t$ (cf. Lemma 4.4). The result then follows immediately by performing the same analysis as before. $\square$

Remark 4.1. Let us now briefly discuss the result of Theorem 4.1. To that end, we need to separate two cases, i.e., depending on whether condition (B1) or (B2) of Assumption 2.1 holds. In the former case, which is really general and highly relevant to the applications considered herein, it suffices to enforce that $\tilde{\delta} = \mathcal{O}(\mu/\sqrt{n})$ to retrieve the same convergence rate as that obtained in [6, Theorem 3.2], in the context of *unconstrained* nonsmooth and nonconvex stochastic optimization. On the other hand, under condition (B2) of Assumption 2.1, we instead need to enforce that $\tilde{\delta} = \mathcal{O}(\varepsilon^2\mu/n)$ in order to retrieve the convergence rate of [6, Theorem 3.2]. Another important point, already briefly mentioned in Section 2.3 (cf. Remark 2.1), is that the analysis could be extended to accommodate for stochastic upper bounds on the oracle error (i.e., allowing $\tilde{\delta}$ to be a function of the underlying randomness, and force it to have finite first- and second-moment rather than enforcing it to be uniformly bounded). This could be done using a similar methodology as that presented in [24], and would reveal an averaged-over-the-iterates error propagation. This is omitted here for brevity of exposition.

5. SELECTED APPLICATIONS

In this section, we showcase the applicability of the proposed algorithm in two wide classes of problems, namely, two-stage stochastic programming, and stochastic minimax optimization, while also briefly mentioning certain additional applications that may be of interest to the wider academic community.

5.1. Two-stage stochastic programming. On our usual probability space $(\Omega, \mathcal{F}, P)$, we consider a random vector $\xi: \Omega \rightarrow \Xi \subset \mathbb{R}^d$, and its induced Borel space $(\Xi, \mathcal{B}(\Xi), P)$. In this section, we consider nonconvex two-stage stochastic programming problems of the form

$$\min_{x \in \mathcal{X}} \mathbb{E}_\xi \left\{ \min_{y \in \mathcal{Y}(\xi)} \widehat{F}(x, y, \xi) \right\}, \quad (2SP)$$

where $\widehat{F}: \mathbb{R}^n \times \mathbb{R}^m \times \Xi \rightarrow \mathbb{R}$ is Borel in $\xi \in \Xi$, and continuous on $\mathbb{R}^n \times \mathbb{R}^m$ for a.e. $\xi \in \Xi$. We let $\mathcal{X} \subset \mathbb{R}^n$ be a convex and compact set, and the multifunction $\mathcal{Y}: \Xi \rightrightarrows \mathbb{R}^m$ be compact-valued and measurable with respect to $\mathcal{B}(\Xi)$ (and thus, the indicator function $\iota_{\mathcal{Y}(\xi)}(\cdot)$ is random lower semicontinuous; cf. [17, Definition 9.47]).

Remark 5.1. Let us note that problem (2SP) is not stated in its full generality. Specifically, one could consider a constraint set $\mathcal{Y}(\xi)$, for the second-stage problem, which also depends x . Then, in order to show that Assumption 1.1 holds for (2SP), we would have to use the perturbation analysis machinery from, e.g., [15]. This is omitted here in the interest of clarity and brevity. Nonetheless, even in its current simplified form, problem (2SP) is already very general (with a plethora of important applications, as stated in the introduction) and its solution under minimal conditions remains a challenge.

Regularity conditions and assumptions. In keeping with the notation of **(P)**, we let $F(x, \xi) \triangleq \min_{y \in \mathcal{Y}(\xi)} \hat{F}(x, y, \xi)$. In order to ensure that problem **(2SP)** is well-defined, we will implicitly make the minimal assumption that $\hat{F}(x, y^*(x, \xi(\cdot)), \xi(\cdot)) \in \mathcal{L}_1(\Omega, \mathcal{F}, P; \mathbb{R})$ for any measurable selection $y^*(x, \xi(\cdot)) \in \arg \min_{y \in \mathcal{Y}(\xi(\cdot))} \hat{F}(x, y, \xi(\cdot))$. Throughout this section, we will employ the following blanket assumption on **(2SP)**.

Assumption 5.1. The following conditions are in effect for **(2SP)**:

- (C1) For a.e. $\xi \in \Xi$, the function $\hat{F}(\cdot, y, \xi): \mathbb{R}^n \rightarrow \mathbb{R}$ is differentiable for every $y \in \mathcal{Y}$ and $\nabla_x \hat{F}(\cdot, \cdot, \xi)$ is continuous on $\mathbb{R}^n \times \mathcal{Y}$, for a.e. $\xi \in \Xi$. Moreover, for all $(x, y) \in \mathbb{R}^n \times \mathbb{R}^m$, the function $\hat{F}(x, y, \cdot)$ is Borel measurable;
- (C2) The set $\mathcal{X} \subset \mathbb{R}^n$ is convex and compact and the multifunction $\mathcal{Y}: \Xi \rightrightarrows \mathbb{R}^m$ is compact-valued and Borel measurable;
- (C3) We can draw i.i.d. samples from the law of ξ ;
- (C4) For a.e. $\xi \in \Xi$, the function $F(x, \xi)$ satisfies condition **(A1)** of Assumption 1.1.

Let us now briefly consider the conditions imposed in Assumption 5.1. Specifically, the only condition that requires verification is **(C4)**. We proceed to argue that this condition is indeed minimal. Specifically, for a.e. $\xi \in \Xi$, and by using conditions **(C1)**–**(C2)** of Assumption 5.1, we may employ Danskin's theorem (e.g., see [17, Theorem 9.26]), which implies that $F(\cdot, \xi)$ is $L(\xi)$ –Lipschitz continuous on $\mathcal{X}$ (since $\mathcal{X}$ is assumed to be compact). In other words, condition **(C4)** merely enforces that $\mathbb{E}\{L^2(\xi)\} \leq G^2$, for some $G > 0$.

Applying Algorithm Z-iProxSG to (2SP). Next, we discuss the compatibility of **(2SP)** (alongside Assumption 5.1) with the developments in Section 4, and in particular with the oracle definition (cf. Definition 2.2 and its associated Assumption 2.1). Let us begin by noting that condition **(B2)** of Assumption 2.1 is already satisfied since we have assumed that $\mathcal{X}$ is a convex and compact set. We next discuss the plausibility of condition **(B1)** instead, while discussing the compatibility of the inexact noisy oracle given in Definition 2.2 with **(2SP)**. We start by noting that Assumption 5.1 does not enforce convexity of the second-stage problem, i.e., of $\min_{y \in \mathcal{Y}(\xi)} \hat{F}(x, y, \xi)$, given some $(x, \xi) \in \mathbb{R}^n \times \Xi$. Nonetheless, the oracle definition implicitly assumes that we can consistently employ some algorithm (e.g., a numerical method) that is able to find, for any $x \in \mathcal{X}$ and a.e. $\xi \in \Xi$, a solution $\tilde{y}(x, \xi) \in \mathcal{Y}(\xi)$ such that

$$F(x, \tilde{y}(x, \xi), \xi) - \min_{y \in \mathcal{Y}(\xi)} F(x, y, \xi) = \delta(x, \xi) \leq \tilde{\delta},$$

where, in this case, the absolute value is obsolete (by the definition of problem **(2SP)**). In what follows, we argue that the proposed algorithm provides a general-purpose solution method for two-stage stochastic programming problems that go far beyond what has already been considered in the literature. To that end, we discuss Assumption 5.1 by separating cases, first considering lower-level convexity, and then discussing the general nonconvex lower-level case.

- (1) The first case that is naturally covered in our setup is the case where $\mathcal{Y}(\xi)$ is a convex set, and for any $x \in \mathcal{X}$ and a.e. $\xi \in \Xi$, $\hat{F}(x, \cdot, \xi)$ is a convex function. Let us note that this does not imply that **(2SP)** is a convex problem, since we have not enforced convexity on $\hat{F}(\cdot, y, \xi)$. Additionally, and unlike most approaches in the literature, the methodology works without imposing any regularity conditions on the second-stage problem (such as,

e.g., Slater's or Robinson's constraint qualifications), neither uniqueness of the second-stage optimal solution for a fixed pair (x, ξ) . In the latter case, i.e., under the assumption of uniqueness of the second-stage problem's optimal solution (which, for example, follows under the assumption of strong convexity of $\hat{F}(x, \cdot, \xi)$), one could invoke Danskin's theorem (again, see [17, Theorem 9.26]) to show that $F(\cdot, \xi)$ is differentiable (in which case, one could attempt to solve (2SP) by utilizing stochastic hypergradient descent; e.g., see the developments in [22], which focus on two-stage programming problems arising in wireless communication systems). In the general framework of stochastic bilevel optimization, which subsumes two-stage stochastic programming, hypergradient descent schemes that rely on lower-level strong convexity have been well-studied. We refer the reader to [44, 45], and the references therein, for additional details. In this general setting, we can assume that the lower level problem can be solved to any accuracy, thus making our oracle as accurate as needed. Thus, following our discussion in Section 4.2, we may readily enforce that $\tilde{\delta} = \mathcal{O}(\mu/\sqrt{n})$ (by making use of an appropriate convex numerical optimization solver), thus retrieving the convergence rate achieved in [6, Theorem 3.2]. Obviously, if condition (B1) of Assumption 2.1 is not satisfied, it instead suffices to enforce that $\tilde{\delta} = \mathcal{O}(\varepsilon^2/\mu)$ (since condition (B2) of Assumption 2.1 is readily satisfied) to obtain the same rate. Finally, let us observe that condition (B1) of Assumption 2.1 is very natural in this case. Indeed, as already discussed in Remark 2.2, we can call a numerical optimization solver for the lower level problem (i.e., for $\min_{y \in \mathcal{Y}(\xi)} \hat{F}(x, y, \xi)$) and enforce that it returns a solution of prescribed accuracy, for any $x \in \mathcal{X}$. Thus, the discussion in Remark 2.2 readily applies in this context.

- (2) In general, the conditions given in Assumption 5.1 do not exclude the case where $\hat{F}(x, \cdot, \xi)$ is nonconvex. In this case, the oracle given in Definition 2.2 is still consistent and general enough. Indeed, we do not specify the magnitude of $\tilde{\delta}$ in this definition (which refers to the upper bound on the oracle error). Thus, the proposed algorithm works as intended also in this case. The difference to the convex lower-level case is that we can no longer control the magnitude of $\tilde{\delta}$ to an arbitrary degree (unless further structure is imposed to the lower-level problem). Thus, we cannot expect to retrieve the same convergence rates as those derived in [6, Theorem 3.2], and would instead have to settle for an approximately stationary point, with the approximation accuracy directly dependent on the oracle error bound $\tilde{\delta}$. Concerning condition (B1) of Assumption 2.1, the situation is less clear compared with the convex lower-level case. Specifically, the discussion given in Remark 2.2 is no longer necessarily applicable. Instead, what this condition implies is that the lower-level problem, while nonconvex, is “equally hard”, irrespectively of $x \in \mathcal{X}$. In other words, this condition implicitly assumes that the lower-level problem can be solved to a similar accuracy, irrespectively of the outer-level parameter vector x . While this is not a strong assumption, it is not readily verifiable; for that reason, Algorithm Z-iProxSG was analyzed also under condition (B2) of Assumption 2.1, which automatically holds in this case.

Comparison with alternative solution methods. Let us now compare the proposed methodology with alternative optimization schemes that have been devised in the literature to solve problems of the form of (2SP). There are currently two classes of methods suitable for the

solution of (2SP) in the available literature, namely, stochastic successive convex approximation (SSCA) optimization and stochastic hypergradient descent schemes. Specifically, there is a long line of works focusing on SSCA-type methods for the solution of nonconvex two-stage stochastic programs studied herein, which were heavily utilized in the context of optimization over wireless communication networks and resource allocation (e.g., see [46, 47, 48] and the references therein). Such methods, which are typically classified as “two-timescale schemes”, rely on successive convex surrogates and approximation of the problem statistics during the optimization process, which incurs high computational costs as well as unrealistic assumptions for their theoretical grounding (which does not include non-asymptotic guarantees). Many of the drawbacks of SSCA schemes were later addressed in a line of work focused on stochastic hypergradient descent schemes (see [22, 23, 24]), which avoid the use of any problem statistics (enabling the online execution of the associated algorithms) while also providing much stronger theoretical guarantees compared with SSCA-type methods. Specifically, the work in [23] provided a detailed theoretical analysis of these stochastic hypergradient schemes under fairly general assumptions, under the condition that the second-stage problem is solved exactly. This was later relaxed in [24], which allowed for inexact evaluations of the objective function of (2SP). Nonetheless, both approaches require weak convexity and differentiability of $F(\cdot, \xi)$, which in turn can only be established under some strong assumptions on the second-stage problem (i.e., the minimization with respect to y). Most notably, the strong second-order sufficient conditions at each optimal solution of the second-stage problem stand out, since they imply a local solution uniqueness property for the second-stage problem, which is known to fail in many circumstances in nonconvex optimization (e.g., see [15]). Additionally, while [24] allows for oracle errors, there is still a requirement that the distance between the retrieved approximate solution to the second-stage problem (returned by the oracle) is “close” to some optimal solution. In this work, we instead only require that the objective values of these two points are close, which is much more consistent in the context of nonconvex optimization. Indeed, in order to guarantee that this “closeness” of the oracle point to some optimal solution, required in [24], is satisfied, the authors had to restrict the class of functions $\hat{F}(\cdot, \cdot, \xi)$ to those that are real-analytic. While this class is fairly rich, it is significantly more limited compared with the functions included in this work, under Assumption 5.1. As expected, assuming that the conditions required by stochastic hypergradient schemes are satisfied for (2SP), one may obtain better rates compared with those derived herein, and the corresponding algorithms require a single oracle evaluation at each outer stochastic hypergradient iteration. Nonetheless, under an additional oracle evaluation, we showcase that the proposed approach given in Algorithm Z-iProxSG can operate under significantly more general assumptions, thus substantially advancing the known capabilities of numerical optimization for nonconvex two-stage stochastic programming.

5.2. Stochastic minimax optimization. Next, we consider general stochastic minimax optimization problems. To that end, we will separate two cases, which typically require distinct solution methods and are naturally applied in different contexts. Specifically, we first consider the case of minimax stochastic optimization problems in which the “adversary” has complete access to instantaneous information, while in the second case, we will assume that the “adversary” only has access to ergodic information. A typical application of great importance for the former formulation is that of building deep learning models that are resistant to adversarial attacks (e.g., see [26]), while the latter formulation typically appears in applications involving

generative adversarial networks, distributionally robust optimization or robust training of neural networks, among others (e.g., see [27, 29, 31]).

5.2.1. Adversary with instantaneous information. We first consider stochastic minimax optimization problems of the following form:

$$\min_{x \in \mathcal{X}} \mathbb{E}_{\xi} \left\{ \max_{y \in \mathcal{Y}(\xi)} \hat{F}(x, y, \xi) \right\}, \quad (\text{MM-I})$$

where, as in the two-stage programming case, we assume that the feasible set of the adversarial variable y is independent of x but may depend on the random vector ξ (noting that typical applications assume that $\mathcal{Y}$ is also independent of ξ ; e.g., see [26]). Once again, in keeping with the notation of (P), we let $F(x, \xi) \triangleq \max_{y \in \mathcal{Y}(\xi)} \hat{F}(x, y, \xi)$. Furthermore, to ensure that problem (MM-I) is well-defined, we will again implicitly make the minimal assumption that $\hat{F}(x, y^*(x, \xi(\cdot)), \xi(\cdot)) \in \mathcal{L}_1(\Omega, \mathcal{F}, P; \mathbb{R})$ for any measurable selection $y^*(x, \xi(\cdot)) \in \arg \max_{y \in \mathcal{Y}(\xi(\cdot))} \hat{F}(x, y, \xi(\cdot))$. Let us note the similarity between problem (2SP) and (MM-I). Indeed, the only difference between the two formulations is the maximization with respect to y in the latter case, compared with the minimization present in the former case. In light of this, our entire discussion given in Section 5.1 applies readily also in this case, and thus Algorithm Z-iProxSG can be immediately utilized to solve (MM-I), while being theoretically supported under Assumption 5.1. At this point, it is important to note that the presence of a maximization (instead of a minimization) term in the objective function of (MM-I) does not change anything structurally important in the context of Assumption 5.1. Indeed, as we did in Section 5.1, we can once again apply Danskin's theorem to show that conditions (C1)–(C2) imply $L(\xi)$ –Lipschitz continuity of $F(\cdot, \xi)$, for a.e. $\xi \in \Xi$, which in turn yields that condition (C4) is merely an assumption on the boundedness of second-moment of the associated Lipschitz constant random function $L(\cdot)$. This example, when paired with the example presented in Section 5.1, immediately showcases the power of the scheme presented in Algorithm Z-iProxSG. Indeed, the same algorithmic strategy can be readily employed to solve two distinct optimization problems that are traditionally challenging, under the same minimal conditions collected in Assumption 5.1. What is especially interesting in this case is the juxtaposition of the proposed methodology (applied in the context of instantaneous stochastic minimax optimization) and the methodology presented in [26], which lacks any serious theoretical support (let alone under the minimal set of assumptions laid out herein). Nonetheless, as we have also stated in the introduction, an adaptation of the stochastic hypergradient descent scheme proposed in [24] could potentially be theoretically supported in this case under certain regularity and structural conditions (albeit stronger compared with those laid out in Assumption 5.1). This is left as a future research direction, open to further consideration.

5.2.2. Adversary with ergodic information. We consider stochastic (nonconvex-nonconcave) minimax optimization problems of the form:

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \mathbb{E}_{\xi} \{ \hat{F}(x, y, \xi) \} \triangleq \hat{f}(x, y), \quad (\text{MM-E})$$

where the feasible set for the adversarial variable, i.e. $\mathcal{Y}$, is now independent of both x and ξ , and the adversary is assumed to only have access to ergodic information. Similar to the previous two applications, we let $F(x, \xi) \triangleq \hat{F}(x, y^*(x), \xi)$, where $y^*(x) \in \arg \max_{y \in \mathcal{Y}} \{ \mathbb{E}_{\xi} \{ \hat{F}(x, y, \xi) \} \}$ is some selection. Let us observe that, for any two selections $y_1^*(x)$, $y_2^*(x)$, the objective function

value of problem (P) is the same, i.e. $\mathbb{E}_\xi \{F(x, y_1^*(x), \xi)\} = \mathbb{E}_\xi \{F(x, y_2^*(x), \xi)\}$. This detail will be important in order to show that problem (MM-E) can be cast in the form of (P) and satisfy the conditions of Assumption 1.1 under mild assumptions. Indeed, let us now briefly discuss Assumption 1.1 in the context of the following problem:

$$\min_{x \in \mathcal{X}} f(x) = \mathbb{E}_\xi \{\hat{F}(x, y^*(x), \xi)\},$$

where $y^*(x) \in \arg \max_{y \in \mathcal{Y}} \{\mathbb{E}_\xi \{\hat{F}(x, y, \xi)\}\}$ is an arbitrary selection. We note that under the assumption of compactness of $\mathcal{Y}$, the differentiability of $\hat{f}(\cdot, y)$ for any $y \in \mathcal{Y}$, and the continuity of $\nabla_x \hat{f}(x, y)$ on $\mathbb{R}^n \times \mathcal{Y}$, we would obtain that $f(\cdot)$ is Lipschitz continuous on $\mathcal{X}$. The requirement of Assumption 1.1 is slightly stronger, in that it enforces Lipschitz continuity of $F(\cdot, y^*(\cdot), \xi)$ rather than of f (alongside the second-moment condition of the associated Lipschitz continuity constant). For example, the former could be guaranteed under the following assumptions on $\hat{F}$, without the requirement that $\mathcal{Y}$ is compact:

- $\hat{F}(x, y, \xi)$ is twice-differentiable with respect to y for all $x \in \mathcal{X}$ and a.e. $\xi \in \Xi$ and the Hessian (w.r.t. y) is continuous jointly in (x, y) ;
- $\hat{F}(x, y, \xi)$ is Lipschitz continuous with respect to x , uniformly in y , for a.e. $\xi \in \Xi$.

Under these assumptions, we may utilize Robinson's implicit function theorem (e.g., see [49, Theorem 2B.1]) to show that, for all $x \in \mathcal{X}$, there exists a locally Lipschitz continuous selection $y^*(x)$, which in turn implies that $F(x, y^*(x), \xi)$ is locally Lipschitz continuous. Lipschitz continuity is then retrieved by assuming that $\mathcal{X}$ is compact. Note that the fact that the single-valued locally Lipschitz localization $y^*(x)$ cannot necessarily be found in practice is not a problem for the proposed methodology. Indeed, since $f(x)$ has the same value for all selections, and since our algorithm operates under the assumption that $F(x, \xi)$ can only be evaluated inexactly, we may define F using any measurable selection $y^*(\cdot)$; in turn, this can ensure that F satisfies the conditions of Assumption 1.1 under mild assumptions. Once again, we see that the proposed algorithmic framework is readily applicable to problems of the form of (MM-E), and its nonasymptotic convergence guarantees hold under less restrictive assumptions compared with alternative approaches provided in the literature (e.g., see the developments in [34, 35, 36] and the references therein). We note, however, that the case in which the lower-level (maximization) problem is concave is typically best handled using stochastic gradient descent-ascent schemes (e.g., see the developments in [32] and the references therein), assuming that the sample gradients of $\hat{F}$ can be readily computed (which is not a requirement for the method proposed herein). Overall, we observe that the proposed approach is highly versatile and enables the approximate solution of intractable optimization instances under very general assumption that are out of reach for currently available gradient-based methodologies.

5.3. Additional applications. Let us observe that while Sections 5.1–5.2 focus on cases where $F(\cdot, \xi)$ represents the value function of some optimization problem, the proposed algorithm is applicable in a plethora of other settings in which the presence of inexact oracles for the evaluation of the objective function of (P) remains crucial. Indeed, a natural example includes cases in which the evaluation of $F(\cdot, \xi)$ requires the utilization of some numerical simulation of a real-world process (e.g., involving the solution of discretized partial differential equations, among many other examples). In this case, the evaluation oracles remain inexact and problem (P) enables one to solve general stochastic parametric problems under the minimal assumption

of Lipschitz continuity. Applications of this form appear in several real-world domains, and are typically classified as hyperparameter tuning problems (e.g., see [16, Section 4.2] for an example problem in the context of hyperparameter tuning of algorithmic parameters). For simplicity of exposition, we refer the reader to [50, Chapter 4] for a detailed discussion on several application instances of this form.

6. CONCLUSIONS

In this work, we derive a zeroth-order method suitable for the solution of general nonsmooth and nonconvex stochastic composite optimization problems in which the real-valued part of the objective is Lipschitz continuous while the extended-valued one is closed, proper, and convex. The algorithm is shown to converge, non-asymptotically, close to a stationary point under minimal assumptions, where near-stationarity is controlled using a novel optimality measure proposed herein (generalizing notions that are currently available in the literature). Importantly, the algorithm is able to operate under general stochastic oracles, providing inexact and biased evaluations of the stochastic objective function.

In light of the generality of the proposed algorithm, we showcase its ability of handling (in a theoretically supported manner) large classes of two-stage stochastic programming as well as nonconvex-nonconcave stochastic minimax optimization problems, in regimes that are out-of-reach of alternative optimization methods that are currently available in the literature. Specifically, we demonstrate the versatility of the proposed methodology by juxtaposing the assumptions required to establish its non-asymptotic ergodic convergence in several challenging applications against the assumptions required by alternative state-of-the-art approaches appearing in the literature.

Acknowledgments

Dionysis Kalogerias was supported by the National Natural Science Foundation under Grant No. 2242215.

REFERENCES

- [1] J. Zhang, H. Lin, S. Jegelka, S. Sra, and A. Jadbabaie. Complexity of finding stationary points of nonconvex nonsmooth functions. In: *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 11173–11182. PMLR, 2020.
- [2] A. A. Goldstein. Optimization of Lipschitz continuous functions. *Math. Program.* 13 (1977), 14–22.
- [3] D. Davis, D. Drusvyatskiy, Y.T. Lee, S. Padmanabhan, and G. Ye. A gradient sampling method with complexity guarantees for Lipschitz functions in high and low dimensions. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, vol. 35, pp. 6692–6703. Curran Associates, Inc., 2022.
- [4] Y.E. Nesterov and V.G. Spokoiny. Random gradient-free minimization of convex functions. *Found. Comput. Math.* 17 (2017), 527–566.
- [5] J.C. Duchi, P.L. Bartlett, and M.J. Wainwright. Randomized smoothing for stochastic optimization. *SIAM J. Optim.* 22 (2012), 674–701.
- [6] T. Lin, Z. Zheng, and M. Jordan. Gradient-free methods for deterministic and stochastic nonsmooth nonconvex optimization. In : *Advances in Neural Information Processing Systems*, vol. 35, pp. 26160–26175, 2022.
- [7] L. Chen, J. Xu, and L. Luo. Faster gradient-free algorithms for nonsmooth nonconvex stochastic optimization. In: *Proceedings of the 40th International Conference on Machine Learning, ICML’23*. 2023.

- [8] A. Cutkosky, H. Mehta, and F. Orabona. Optimal, stochastic, non-smooth, non-convex optimization through online-to-non-convex conversion. In *Proceedings of the 40th International Conference on Machine Learning, ICML'23*. JMLR.org, 2023.
- [9] G. Kornowski and O. Shamir. An algorithm with optimal dimension-dependence for zero-order nonsmooth nonconvex stochastic optimization. *J. Mach. Learn. Res.* 25 (2024), 1-14.
- [10] M. Jordan, G. Kornowski, T. Lin, O. Shamir, and M. Zampetakis. Deterministic nonsmooth nonconvex optimization. In: Gergely Neu and Lorenzo Rosasco, (ed.) *Proceedings of Thirty Sixth Conference on Learning Theory*, vol. 195 of *Proceedings of Machine Learning Research*, pp. 4570–4597, PMLR, 2023.
- [11] Z. Liu, C. Chen, L. Luo, and B.K.H. Low. Zeroth-order methods for constrained nonconvex nonsmooth stochastic optimization. In: *Proceedings of the 41st International Conference on Machine Learning, ICML'24*. 2024.
- [12] B. Grimmer and Z. Jia. Goldstein stationarity in lipschitz constrained optimization. *Optim. Lett.* 19 (2025), 425-435.
- [13] Y. Nesterov. *Lectures on Convex Optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer, 2018.
- [14] D. Davis and D. Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM J. Optim.* 29 (2019), 207-239.
- [15] J. F. Bonnans and A. Shapiro. *Perturbation Analysis of Optimization Problems*. Springer Series in Operations Research and Financial Engineering. Springer New York, NY, 2000.
- [16] S. Pougkakiotis and D. Kalogierias. A zeroth-order proximal stochastic gradient method for weakly convex stochastic optimization. *SIAM J. Sci. Comput.* 45 (2023), A2679–A2702.
- [17] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*, Third Edition. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2021.
- [18] A. Liu, R. Yang, T.Q.S. Quek, and M.J. Zhao. Two-stage stochastic optimization via primal-dual decomposition and deep unrolling. *IEEE Trans. Signal Process.* 69 (2021), 3000-3015.
- [19] X. Zhai, G. Han, Y. Cai, and L. Hanzo. Beamforming design based on two-stage stochastic optimization for RIS-assisted over-the-air computation systems. *IEEE Int. Things J.* 9 (2022), 5474-5488.
- [20] Y. Zhao, Y. Teng, A. Liu, and V. Lau. Two-timescale joint UL/DL dictionary learning and channel estimation in massive MIMO systems. In: *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, pp. 5408–5413, 2022.
- [21] Y. Zhao, Y. Teng, A. Liu, X. Wang, and V.K.N. Lau. Joint UL/DL dictionary learning and channel estimation via two-timescale optimization in massive MIMO systems. *IEEE Trans. Wireless Commun.* 23 (2024), 2369-2382.
- [22] H. Hashmi, S. Pougkakiotis, and D. Kalogierias. Model-free learning of optimal beamformers for passive IRS-assisted sumrate maximization. In: *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.
- [23] H. Hashmi, S. Pougkakiotis, and D. Kalogierias. Model-free learning of two-stage beamformers for passive IRS-aided network design. *IEEE Trans. Signal Proces.* 72 (2024), 652-669.
- [24] S. Pougkakiotis, H. Hashmi, and D. Kalogierias. Data-driven learning of two-stage beamformers in passive IRS-assisted systems with inexact oracles. *arXiv:2410.24154*, 2024.
- [25] X. Qin, X. Song, and S. Jiang. Bi-level meta-learning for few-shot domain generalization. In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15900–15910, 2023.
- [26] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In: *International Conference on Learning Representations (ICLR)*, 2018.
- [27] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2672–2680, 2014.
- [28] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, Cambridge, 2006.
- [29] A. Sinha, H. Namkoong, and J. C. Duchi. Certifying some distributional robustness with principled adversarial training. In: *International Conference on Learning Representations*, 2018.

- [30] A. Tamar, Y. Glassner, and S. Mannor. Optimizing the CVaR via sampling. In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAI'15*, pp. 2993–2999, 2015.
- [31] H. Rahimian and S. Mehrotra. Distributionally robust optimization: A review. *ArXiv*, abs/1908.05659, 2019.
- [32] T. Lin, C. Jin, and M. I. Jordan. Two-timescale gradient descent ascent algorithms for nonconvex minimax optimization. *J. Mach. Learn. Res.* 26 (2015), 1-45.
- [33] K. Emmanouilidis, R. Vidal, and N. Loizou. Stochastic extragradient with random reshuffling: Improved convergence for variational inequalities. In: *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3682–3690. PMLR, 2024.
- [34] C. Daskalakis, N. Golowich, S. Skoulakis, and E. Zampetakis. Stay-on-the-ridge: Guaranteed convergence to local minimax equilibrium in nonconvex-nonconcave games. In: *Proceedings of the 36th Conference on Learning Theory (COLT)*, volume 195 of *Proceedings of Machine Learning Research*, pp. 5146–5198, 2023.
- [35] J. Diakonikolas, C. Daskalakis, and M. I. Jordan. Efficient methods for structured nonconvex-nonconcave min-max optimization. In: *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2746–2754, 2021.
- [36] B. Grimmer, H. Lu, P. Worah, and V. Mirrokni. The landscape of the proximal point method for nonconvex–nonconcave minimax optimization. *Math. Program.* 191 (2022), 1-35.
- [37] F. H. Clarke. *Optimization and Nonsmooth Analysis*. Society for Industrial and Applied Mathematics, Philadelphia, 1990.
- [38] A. D. Flaxman, A. T. Kalai, and H. B. McMahan. Online convex optimization in the bandit setting: Gradient descent without a gradient. In: *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '05*, pp. 385–394. Society for Industrial and Applied Mathematics, Philadelphia, 2005.
- [39] O. Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *J. Mach. Learn. Res.* 18 (2017), 1-11.
- [40] A. Nemirovsky and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. John Wiley and Sons, New York, 1983.
- [41] R. T. Rockafellar and R. J-B Wets. *Variational Analysis*, volume 317. Springer, Heidelberg, 2004.
- [42] J.-P. Vial. Strong and weak convexity of sets and functions. *Math. Oper. Res.* 8 (1983), 231-259.
- [43] H. Attouch and D. Aze. Approximation and regularization of arbitrary functions in Hilbert spaces by the Lasry-Lions method. *Annales de l'Institut Henri Poincaré C, Analyse non linéaire*, 10 (1993), 289-312.
- [44] X. Chen, T. Xiao, and K. Balasubramanian. Optimal algorithms for stochastic bilevel optimization under relaxed smoothness conditions. *J. Mach. Learn. Res.* 25 (2024), 1-51.
- [45] G. Riccardo, P. Massimiliano, and S. Saverio. Convergence properties of stochastic hypergradients. In: *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3826–3834. PMLR, 2021.
- [46] M. Hong, Q. Li, and Y.-F. Liu. Decomposition by successive convex approximation: A unifying approach for linear transceiver design in heterogeneous networks. *IEEE Trans. Wireless Commun.* 15 (2016), 1377-1392.
- [47] G. Scutari, F. Facchinei, P. Song, D. P. Palomar, and J.-S. Pang. Decomposition by partial linearization: Parallel optimization of multi-agent systems. *IEEE Trans. Signal Proces.* 62 (2014), 641-656.
- [48] Y. Yang, G. Scutari, D. P. Palomar, and M. Pesavento. A parallel decomposition method for nonconvex stochastic multi-agent optimization problems. *IEEE Trans. Signal Proces.* 64 (2016), 2949-2964.
- [49] A. L. Dontchev and R. T. Rockafellar. *Implicit Functions and Solution Mappings*. Springer Monographs in Mathematics. Springer, Dordrecht, 2009.
- [50] W. B. Powell. *Reinforcement Learning and Stochastic Optimization: A Unified Framework for Sequential Decisions*. John Wiley & Sons, Hoboken, NJ, 2022.