SIAM J. OPTIM.
Vol. 29, No. 3, pp. 1908–1930

PROXIMALLY GUIDED STOCHASTIC SUBGRADIENT METHOD FOR NONSMOOTH, NONCONVEX PROBLEMS[*]

DAMEK DAVIS[†] AND BENJAMIN GRIMMER[†]

Abstract. In this paper, we introduce a stochastic projected subgradient method for weakly convex (i.e., uniformly prox-regular) nonsmooth, nonconvex functions—a wide class of functions which includes the additive and convex composite classes. At a high level, the method is an inexact proximal-point iteration in which the strongly convex proximal subproblems are quickly solved with a specialized stochastic projected subgradient method. The primary contribution of this paper is a simple proof that the proposed algorithm converges at the same rate as the stochastic gradient method for smooth nonconvex problems. This result appears to be the first convergence rate analysis of a stochastic (or even deterministic) subgradient method for the class of weakly convex functions. In addition, a two-phase variant is proposed that significantly reduces the variance of the solutions returned by the algorithm. Finally, preliminary numerical experiments are also provided.

Key words. nonsmooth, nonconvex, subgradient, stochastic, proximal

AMS subject classifications. 65K05, 65K10, 90C26, 90C15, 90C30

DOI. 10.1137/17M1151031

1. Introduction.

Stochastic approximation methods iteratively minimize the expectation of a family of known loss functions with respect to an unknown probability distribution. Such methods are of fundamental importance in machine learning, signal processing, statistics, and data science more broadly. For example, in machine learning, one is often interested in designing a classifier that performs well on the entire population of samples, given only a finite list of correctly labeled pairs $z_1, \ldots, z_n$ obtained from a fixed, but otherwise unknown, distribution $\mathbb{P}$. Such problems are often formulated as *population risk minimization*:

$$\text{minimize } F(x) := \begin{cases} \mathbb{E}_{z\sim\mathbb{P}}\left[f(x,z)\right] & \text{if } x \in \mathcal{X}, \\ \infty & \text{otherwise.} \end{cases} \tag{1}$$

Here, $\mathcal{X} \subseteq \mathbb{R}^d$ denotes a constraint set, while $f(x,z)$ represents the loss of the decision rule parameterized by $x \in \mathcal{X}$ on the population data z.

Much algorithmic development has been inspired by (1). Robbins and Monro's pioneering 1951 work [32] developed the first method for solving (1) when each $f(\cdot, z)$ is smooth and strongly convex and $\mathcal{X} = \mathbb{R}^d$. This and most later methods are variants of the stochastic projected (sub)gradient method, which iteratively constructs approximate solutions x_t of (1) through the recursion

$$\begin{aligned} &\text{sample } z_t \sim \mathbb{P}, \\ &\text{set } x_{t+1} = \operatorname{proj}_{\mathcal{X}}(x_t - \alpha_t \nabla_x f(x_t, z_t)), \end{aligned}$$

[*] Received by the editors October 9, 2017; accepted for publication (in revised form) April 26, 2019; published electronically July 23, 2019.
https://doi.org/10.1137/17M1151031

Funding: This material is based upon work supported by the National Science Foundation under award 1502405, and by the National Science Foundation Graduate Research Fellowship under grant DGE-1650441.

[†] Operations Research and Information Engineering, Cornell University, Ithaca, NY 14850 (dsd95@cornell.edu, bdg79@cornell.edu).

where $z_1, \ldots, z_t, \ldots$ are independently and identically distributed (i.i.d.) and α_t is an appropriate control sequence. For nonsmooth $f(\cdot, z_t)$, sample gradients are simply replaced by sample subgradients $v_t \in \partial f(x_t, z_t)$, where $\partial f(x_t, z_t)$ denotes the subdifferential in the sense of convex analysis [34].

The complexity of minimizing (1) is directly related to the regularity of $f(\cdot, z)$. For example, for convex functions $f(\cdot, z)$ the stochastic subgradient method attains expected functional accuracy ε after $O(\varepsilon^{-2})$ stochastic subgradient evaluations. For strongly convex losses, the number of stochastic subgradient evaluations drops to $O(\varepsilon^{-1})$. The interested reader may turn to the seminal work [29] for an in-depth investigation of these methods and for information-theoretic lower bounds showing such rates are unimprovable without further assumptions.

For convex functions, complexity theory does not favor smooth losses over nonsmooth losses. For nonconvex problems, the situation is less clear. In the smooth case, the seminal work of Ghadimi, Lan, and Zhang [21] develops a variant of the stochastic projected gradient method and establishes that the expected norm of the projected gradient

$$\mathbb{E}_{z_1,\ldots,z_t}\left[\|x_t - \mathrm{proj}_{\mathcal{X}}(x_t - \nabla_x \mathbb{E}_{z\sim P}\left[f(x_t, z)\right])\|^2\right], \tag{2}$$

a natural measure of stationarity, tends to zero at a controlled rate. Namely, with $O(\varepsilon^{-2})$ stochastic gradient evaluations, the algorithm produces a point with expected projected gradient norm squared less than ε.

At the time of writing the original version of this manuscript, there was no similar rate of convergence in the nonsmooth, nonconvex setting for any known subgradient-based algorithm. Part of the difficulty in establishing a complexity theory for nonsmooth, nonconvex subgradient-based methods is that the "usual criteria," namely the objective error and the norm of the gradient, can be completely meaningless. Indeed, on the one hand, one cannot expect the objective error $F(x_t) - \inf F$ to tend to zero, even in the smooth setting. On the other hand, simple examples, e.g., $F(x) = |x|$, show that $\mathrm{dist}(0, \partial F(x_t))$ can be strictly bounded below by a fixed constant for all t.

In contrast to subgradient-based methods, the "usual criteria" is meaningful for the *proximal-point method* [33], which constructs a sequence x_t of approximate minimizers through the recursion

$$x_{t+1} = \mathrm{argmin}_{x\in\mathbb{R}^d}\left\{F(x) + \frac{1}{2\gamma}\|x - x_t\|^2\right\},$$

where γ is a control parameter. Namely, it is a simple exercise to show that under minimal assumptions on F, the subdifferential distance $\mathrm{dist}(0, \partial F(x_t))$ tends to zero. Of course, each step of the proximal-point method is difficult, if not impossible, to execute without further assumptions on F.

The search for an appropriate class of functions F for which each proximal subproblem may be (approximately) executed naturally leads us to the deceptively simple, yet surprisingly broad, class of ρ-*weakly convex* functions. Formally, a function $F : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is ρ-*weakly convex* if

$$\text{the assignment } x \mapsto F(x) + \frac{\rho}{2}\|x\|^2 \text{ is convex.}$$

For example, any C^2 function on a compact convex set becomes convex after adding the quadratic $\frac{|\lambda|}{2}\|\cdot\|^2$, where λ is the minimal eigenvalue of its Hessian across all points in the set. In the nonsmooth setting, this class includes all *convex composite losses*

$$h(c(x)),$$

where h is convex and L-Lipschitz and c is C^1 with β-Lipschitz Jacobian; such functions are known to be βL-weakly convex [12, Lemma 4.2]. The additive composite class is another widely used, much studied class of weakly convex functions, formed from all sums

$$g(x) + r(x),$$

where r is closed and convex and g is C^1 with β-Lipschitz gradient; such functions are known to be β-weakly convex. For further examples of weakly convex functions, see [9, section 2.1], which includes formulations of robust phase retrieval, covariance matrix estimation, blind deconvolution, sparse dictionary learning, robust principal component analysis, and conditional value at risk. We provide several further examples in section 2.1. It is important to note that none of these applications are covered by the seminal work of Ghadimi, Lan, and Zhang [21], which assumes an additive composite objective form.

Contributions. In this paper, we develop the first known complexity guarantees for a subgradient-based method for a general class of nonsmooth, nonconvex losses in stochastic optimization. The guarantees in this paper apply to ρ-weakly convex losses F. Our algorithm, called the proximally guided stochastic subgradient method (PGSG; see Algorithm 2), follows an inner–outer-loop strategy that may be compactly and informally summarized as

$$x_{t+1} = \varepsilon\text{-argmin}_{x \in \mathbb{R}^d} \left\{ F(x) + \rho \|x - x_t\|^2 \right\} \qquad \text{(in expectation).} \tag{3}$$

The outer loop of PGSG is governed by the approximate proximal-point method applied to the population risk F. Due to ρ-weak convexity of F, the inner-loop subproblem is a *strongly convex* stochastic optimization problem. Thus, by classical complexity theory, approximate solutions to the inner-loop subproblems may be quickly found. We then turn our attention to establishing complexity guarantees.

As stated before, simple examples show that one cannot expect the iterates produced by a subgradient-based algorithm themselves to be ε-stationary because $\operatorname{dist}(0, \partial F(x_t))$ may be bounded below for all t. Instead, we introduce the following convergence measure: a (random) point is an *ε-solution* if

$$\mathbb{E}\left[\operatorname{dist}(\bar{x}, \{x \mid \operatorname{dist}(0, \partial F(x))^2 \leq \varepsilon\})^2\right] \leq \varepsilon, \tag{4}$$

where ∂F denotes the subdifferential of F in the sense of variational analysis [35]; see section 3. We then show that when both inner and outer loops are coupled together appropriately, an outer-loop iterate chosen uniformly at random is an *ε-solution* after $O(\varepsilon^{-2})$ stochastic subgradient evaluations. The nearly stationary point nearby $\bar{x}$ is itself a solution to a *strongly convex* stochastic optimization problem. Thus, it is in principle obtainable to any desired degree of accuracy; see Remark 1.

Having established expectation guarantees, we turn our attention to probabilistic guarantees. Namely, following [20] (which considers the smooth case), we say a (random) point $\bar{x}$ is an *(ε, Λ)-solution* if

$$\mathbb{P}\left(\operatorname{dist}(\bar{x}, \{x \mid \operatorname{dist}(0, \partial F(x))^2 \leq \varepsilon\})^2 \leq \varepsilon\right) \geq 1 - \Lambda.$$

Markov's inequality shows that PGSG finds an (ε, Λ)-solution $\bar{x}$ after

$$O\left(\frac{1}{\Lambda^2\varepsilon^2}\right)$$

stochastic subgradient evaluations. To improve this complexity, we introduce a 2-phase algorithm, called 2PGSG, which produces an (ε, Λ)-solution after

$$O\left(\frac{\log(1/\Lambda)}{\varepsilon^2} + \frac{\log(1/\Lambda)}{\Lambda\varepsilon}\right)$$

stochastic subgradient evaluations, substantially reducing the variance of our solution estimate. The technique for achieving this improvement is somewhat different to what was proposed for the smooth case in [20]. The challenge in establishing the result is that we no longer have unbiased estimates of subgradients at nearly stationary points. Indeed, the iterates produced by subgradient methods are only nearby nearly stationary points and are not nearly stationary themselves.

Finally, we turn our attention to a more practical variant of PGSG, which does not assume that the weak convexity constant ρ is known. In this setting, a simple idea—letting the outer-loop step size tend to infinity—results in a point $\bar{x}$ which satisfies (4) after $O(\varepsilon^{2/(1-\beta)})$ stochastic subgradient evaluations, where $\beta \in (0, 1)$ is a user defined meta-parameter. We mention that the seminal work of Ghadimi, Lan, and Zhang [21] also assumes knowledge of the weak convexity constant ρ; in their setting ρ is simply the Lipschitz constant of the gradient.

We validate our results with some preliminary numerical experiments on the population objective of a robust real phase retrieval problem. We also discuss several more examples of weakly convex functions in section 2.1.

1.1. Related work.

Stochastic gradient methods. The convergence rates presented in [20] match known rates for the stochastic gradient method in nonconvex optimization. There, the standard stochastic gradient method may be used without modification. Interestingly, recent work has developed methods which ensure that $\mathbb{E}\|\nabla F\|^2 \le \varepsilon$ after at most $O(\varepsilon^{-3/2})$ oracle calls [17]. This shows a surprising gap between smooth and nonsmooth, nonconvex optimization not present in the convex case.

Stochastic proximal-gradient methods. For additive composite problems

$$\text{minimize}\left\{\mathbb{E}_z\left[f(x, z)\right] + r(x)\right\},$$

one often employs stochastic proximal-gradient methods, which require, at every iteration, a (potentially costly) evaluation of the mapping

$$\mathbf{prox}_r(y) = \operatorname{argmin}\left\{r(x) + \frac{1}{2}\|x - y\|^2\right\}.$$

These methods achieve expected projected gradient norm ε, as in (2), after $O(\varepsilon^{-2})$ stochastic gradient evaluations [21]. These methods have also been extended to regularizers that are arbitrary closed prox-bounded functions r [40], a setting which we do not cover.

Evaluating the proximal mapping of r could be substantially more expensive than computing a subgradient. For example, if $r = \|\cdot\|_2$ is the spectral norm on $\mathbb{R}^{n\times n}$, then its proximal mapping requires a full singular value decomposition. In contrast, a subgradient may be computed from a single maximal eigenvector.

Another advantage of stochastic subgradient methods over stochastic proximal-gradient methods is that multiple nonsmooth functions may be present in the objective function F. The same is not true for stochastic proximal-gradient methods: even if two functions r_1 and r_2 have simple proximal operators, the proximal operator of the sum $r = r_1 + r_2$ can be quite complex. Similarly, the proximal operator of an expectation $\mathbb{E}_z[r(x,z)]$ could be intractable.

Stochastic methods for convex composite. Recently [14], a method was proposed for finding stationary points of the convex composite problem in which $f(x,z) = h(c(x,z),z)$. The first method adapts the prox-linear algorithm [4, 5, 6, 11, 13, 19, 25] to the stochastic setting: given x_t, sample z_t and form x_{t+1} as the solution to the convex problem

$$x_{t+1} = \operatorname{argmin}_{x\in X}\left\{h(c(x_t,z_t) + \nabla c(x_t,z_t)(x - x_t), z_t) + \frac{1}{2\gamma_t}\|x - x_t\|^2\right\}, \tag{5}$$

where $\gamma_t = \theta(1/\sqrt{t})$. The second proposed method is a straightforward application of the stochastic projected subgradient method [28]. Both methods are shown to almost surely converge to stationary points, but no rates of convergence are given.

We remark that the convergence proof presented in [14] is complex, being based on the highly nontrivial theory of nonconvex differential inclusions. We believe there is a benefit to having a simple proof of convergence, albeit for a slightly different subgradient method, and this is what we provide in this paper.

Further work on minimizing convex composite problems appears in [26, 38, 39]. This series of papers analyzes nested expectations: $F(x) = \mathbb{E}_v[h(\mathbb{E}_w[c(x,w) \mid v], v)]$. Although the stochastic structure considered in these papers is more general than what we consider in problem (1), the assumptions made on F are much stronger than our assumptions on F. In particular, the authors prove rates under the assumption that (a) F is convex, (b) F is strongly convex, or (c) F is nonconvex, but *differentiable* with Lipschitz continuous gradient. For case (c), the authors propose an algorithm that finds an ε-stationary point of F after $O(\varepsilon^{-2.25})$ gradient evaluations [39] (in particular, they consider unconstrained problems).

Inexact proximal-point methods in nonconvex optimization. The idea of using the inexact proximal-point method to guide a nonconvex optimization algorithm to stationary points is not new. For example, Hare and Sagastizabal [22, 23] propose a method for computing inexact proximal points which then enables the analysis of a nonconvex bundle method. The more recent work [31] exploits linearly convergent algorithms for solving the proximal subproblems. In contrast for the subproblems considered in this work, there are no linearly convergent stochastic subgradient algorithms capable of minimizing the proximal-point step.

Subgradient methods for weakly convex problems. This paper is not the first to consider subgradient methods under weak convexity. For example, the early work [30] proves subsequential convergence of the (nonprojected) subgradient method for weakly convex *deterministic* problems. However, no rates were given in that work.

Almost sure convergence of stochastic subgradient methods for nonconvex problems. Convergence to stationary points of stochastic subgradient methods in nonsmooth, nonconvex optimization has previously been attained in several different scenarios, some of which are more general than the scenario considered in problem (1) [15, 16, 37]. No rates of convergence were given in these works. In contrast, the novelty of the proposed approach lies in the attained rate of convergence, which matches the best-known rates of convergence for smooth nonconvex stochastic optimization [20].

Rates of convergence in stochastic weakly convex optimization. Since the first draft of this paper appeared on arXiv in July 2017,[1] several works appearing in 2018 have established convergence of the standard stochastic projected subgradient method under weak convexity [9]. The obtained rates (in expectation) are essentially the same as those obtained in this paper, namely they are of the form presented in (4). The authors of [9] do not provide any probabilistic guarantees.

1.2. Outline.

Section 2 presents notation and several basic results used in this paper, as well as further examples of weakly convex functions. Section 3 presents our convergence analysis under the assumption that ρ is known. Section 3.2 presents our probabilistic guarantees. Section 3.3 presents our convergence analysis when ρ is unknown. Section 4 presents preliminary numerical results obtained on a robust phase retrieval problem.

2. Notation and basic results.

Most of the notation and concepts we use in this paper can be found in [7, 35]. Our main probabilistic assumption is that we work in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathbb{R}^d$ is equipped with the Borel σ-algebra, which we use to define measurable mappings.

For a given function $F : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, we let

$$\text{dom } F = \{x \in \mathbb{R}^d \mid F(x) < \infty\}, \qquad \text{epi } F = \{(x,t) \in \mathbb{R}^d \times \mathbb{R} \mid f(x) \leq t\}.$$

We say a function is *closed* if epi F is a closed set. We say a function is *proper* if dom $F \neq \emptyset$.

Let $F \colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a proper closed function. At any point $x \in \text{dom } F$, we let

$$\partial F(x) = \{v \in \mathbb{R}^d \mid \left(\forall y \in \mathbb{R}^d\right)\ F(y) \geq F(x) + \langle v, y - x\rangle + o(\|y - x\|)\}$$

denote the *Fréchet subdifferential* of F at x. On the other hand, if $x \notin \text{dom } F$ we let $\partial F(x) = \emptyset$. It is an easy exercise to show that at any local minimizer x of F, we have the inclusion $0 \in \partial F(x)$.

For the class of weakly convex functions, all elements of the subdifferential generate quadratic underestimators of the function F, as the following proposition shows. The equivalences are based on [8, Theorem 3.1].

Proposition 2.1 (subgradients of weakly convex functions). *Suppose that $F : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is a closed function. Then the following are equivalent.*

1. *F is ρ-weakly convex. That is, $F + \frac{\rho}{2}\|\cdot\|^2$ is convex.*
2. *For any $x, y \in \mathbb{R}^d$ with $v \in \partial F(x)$, we have*

$$F(y) \geq F(x) + \langle v, y - x\rangle - \frac{\rho}{2}\|y - x\|^2. \tag{6}$$

3. *For all $x, y \in \mathbb{R}^d$ and $\alpha \in [0,1]$, we have*

$$F(\alpha x + (1-\alpha)y) \leq \alpha F(x) + (1-\alpha) f(y) + \frac{\rho\alpha(1-\alpha)}{2}\|x - y\|^2.$$

2.1. Examples of weakly convex functions.

As stated in the introduction, the class of weakly convex functions is broad. In the nonsmooth setting, this class includes all *convex composite losses*

$$h(c(x)),$$

[1] Available at https://arxiv.org/abs/1707.03505v4.

where h is convex and L-Lipschitz and c is C^1 with β-Lipschitz Jacobian; such functions are known to be βL-weakly convex [12, Lemma 4.2]. Several popular weakly convex formulations are presented in [9, section 2.1]. We now discuss several further examples.

Example 1 (censored block model). The censored block model [1] is a variant of the standard stochastic block model [2], which seeks to detect two communities in a partially observed graph. Mathematically, we encode such communities by forming the "community matrix" $M = \bar{\theta}\bar{\theta}^T \in \{-1,1\}^d$, where $\bar{x} \in \{-1,1\}^d$ is a membership vector in which $\bar{x}_i = 1$ if node i is in the first community and $\bar{x}_i = -1$ otherwise. In the censored block model, we observe a randomly corrupted version $\hat{M}$ of the matrix M:

$$\hat{M} = \begin{cases} 0 & \text{with probability } 1-p, \\ M_{ij} & \text{with probability } p(1-\epsilon), \\ -M_{ij} & \text{with probability } p\epsilon. \end{cases}$$

Then our task is to recover M given only $\hat{M}$. We may formulate this problem in the following convex composite form:

$$F(x) = \sum_{ij \mid \hat{M}_{ij} \neq 0} |x_i x_j - \hat{M}_{ij}|.$$

Notice that absolute value function encourages the matrix xx^T to agree with $\hat{M}$ in most of its nonzero entries—the bulk of which are equal to M_{ij}—due to the sparsity promoting behavior of the nonsmooth absolute value function.

Example 2 (robust phase retrieval). Phase retrieval is a common task in computational science; applications include imaging, X-ray crystallography, and speech processing. Given a set of tuples $\{(a_i, b_i)\}_{i=1}^m \subset \mathbb{R}^d \times \mathbb{R}$, the (real) phase retrieval problem seeks a vector $x \in \mathbb{R}^d$ satisfying $(a_i^T x)^2 = b_i$ for each index $i = 1, \ldots, m$. This problem is NP-hard [18]. Strictly speaking, phase retrieval is a feasibility problem. However, when the set of measurements $\{b_i\}$ is corrupted by gross outliers, one considers the following "robust" phase retrieval objective:

$$F(x) = \frac{1}{n}\sum_{i=1}^n |\langle a_i, x\rangle^2 - b_i|.$$

Notice that this nonsmooth objective is given in convex composite form, and therefore it is weakly convex.

Example 3 (nonsmooth trimmed estimation). Let $f_1, \ldots, f_n$ be Lipschitz continuous, convex loss functions on $\mathbb{R}^d$. The goal of trimmed estimation [3, 27, 36] is to fit a model while simultaneously detecting and removing "outlier" objectives f_i. Mathematically, we fix a number $h \in \{1, \ldots, n\}$ indicating the number of "inliers," and formulate the problem as follows:

$$\text{minimize}_{x\in\mathbb{R}^d, w\in\mathbb{R}^n} \ \sum_{i=1}^n w_i f_i(x)$$
$$\text{subject to } w_i \in [0,1] \text{ and } \sum_{i=1}^n w_i = h.$$

One can see that for fixed x, the only objective values that contribute to the sum are those that are among the h-minimal elements of the set $\{f_1(x), \dots, f_n(x)\}$. In the appendix, we provide a short proof that this objective is weakly convex. Notice that it is in general nonconvex, despite each f_i being convex.

3. Proximally guided stochastic subgradient method.

In this section, we formalize the proposed algorithm. First we slightly generalize the problem considered in the introduction, namely we assume that

$$\text{minimize}_{x\in\mathbb{R}^d}\ F(x) = \begin{cases} f(x) & \text{if } x \in \mathcal{X}, \\ \infty & \text{otherwise,} \end{cases} \tag{7}$$

where f is a closed ρ-weakly convex function. Weak convexity of f implies that each of the proximal subproblems $\min_{x\in\mathbb{R}^d}\{F(x) + (1/2\gamma)\|x - x_t\|^2\}$ is

$$\mu := \gamma^{-1} - \rho$$

strongly convex. Next we introduce a stochastic subgradient oracle and a basic assumption on F.

Assumption A. Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and equip $\mathbb{R}^d$ with the Borel σ-algebra. Then we assume that

(A1) it is possible to generate i.i.d. realizations $z_1, z_2, \dots$ from $\mathbb{P}$;

(A2) there is an open set $U \subseteq \mathbb{R}^d$ containing $\mathcal{X}$ and a measurable mapping $G : U \times \Omega \to \mathbb{R}^d$ such that $\mathbb{E}_z[G(x, z)] \in \partial f(x)$;

(A3) there is a constant $L \geq 0$ such that for all $x \in U$, we have $\mathbb{E}_z[\|G(x, z)\|^2] \leq L^2$.

Assumption A is standard in the literature on stochastic subgradient methods. In particular, assumptions (A1) and (A2) are identical to assumptions (A1) and (A2) in [28], while assumption (A3) is identical to [28, equation (2.5)]. A useful consequence of (A3) is that f itself is Lipschitz.

LEMMA 3.1 (Lipschitz continuity of f [9, section 3.2]). *Suppose that assumption* (A3) *holds. Then f is L-Lipschitz continuous on U.*

The main workhorse of PGSG is a stochastic subgradient method for solving regularized subproblems $\min_{x\in\mathbb{R}^d}\{F(x) + (1/2\gamma)\|x - x_t\|^2\}$ induced by the proximal-point method. We now state this method.

Algorithm 1 Projected stochastic subgradient method for proximal-point subproblems PSSM$(y_0, G, \gamma, \{\alpha_t\}, J)$.

```
Input: $y_0 \in \mathcal{X}$, quadratic multiplier $\gamma > 0$, maximum iterations $J \in \mathbb{N}$, nonnegative
    step-size sequence $\{\alpha_t\}$.
for $j = 0, \dots, J-2$ do
    sample $z_j$ and set $v_j = G(y_j, z_j) + \frac{1}{\gamma}(y_j - y_0)$,
    $y_{j+1} \leftarrow \text{proj}_{\mathcal{X}}(y_{t,j} - \alpha_j v_j)$
end for
Output: $\tilde{y} = \frac{2}{J(J+1)} \sum_{j=0}^{J-1} (j+1) y_j$.
```

Before introducing the proximally guided stochastic subgradient (PGSG) method, we introduce two necessary algorithm parameters:

$$j_t \geq \frac{11}{\gamma^2\mu^2}, \tag{8}$$

$$\alpha_j = \frac{2}{\mu\left(j+2+\dfrac{36}{\gamma^4\mu^4(j+1)}\right)}. \tag{9}$$

The algorithm now follows.

Algorithm 2 Proximally guided stochastic subgradient method PGSG($y_0, G, \gamma, \{j_t\}, T$).

```
Input: $x_0 \in \mathcal{X}$, weak convexity constant $\rho > 0$, $\gamma \in (0, 1/\rho)$, maximum iterations
    $T \in \mathbb{N}$, maximum inner loop iteration $\{j_t\}$ satisfying (8).
Define the step-size sequence $\{\alpha_j\}$ as in (9).
for $t = 0, \dots, T-2$ do
    $x_{t+1} = \mathrm{PSSM}(x_t, G, \gamma, \{\alpha_j\}, j_t)$
end for
Output: $x_R$, where $R$ is sampled uniformly from $\{0, \dots, T-1\}$.
```

As stated in the introduction, PGSG employs an inner–outer-loop strategy, which is shown in Algorithm 2. The outer loop executes $T-1$ approximate proximal-point steps, resulting in the iterates $\{x_t\}$. The inner loop, shown in Algorithm 1, approximately solves the proximal-point subproblem, which is now strongly convex, using a stochastic subgradient method for strongly convex optimization [24]. Beyond its use in governing the outer-loop dynamics of PGSG, the proximal-point subproblems also lead to a natural measure of stationarity.

Indeed, for all $t \in \mathbb{N}$, define the proximal point

$$\hat{x}_t := \mathrm{argmin}_{x \in \mathbb{R}^d} \left\{ F(x) + \frac{1}{2\gamma}\|x - x_t\|^2 \right\}. \tag{10}$$

Note that $\hat{x}_t$ exists and is unique by the μ-strong convexity of the proximal subproblem. We stress that this point, although in principle obtainable via convex optimization, is never computed. Instead it is only used to formulate convergence guarantees. To that end, the following lemma shows that the gap $\gamma^{-1}\|x_t - \hat{x}_t\|$ is a natural measure of stationarity.

LEMMA 3.2 (convergence criteria). *Let $F : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a proper closed function. Let $x \in \mathbb{R}^d$. If*

$$\hat{x} \in \mathrm{argmin}_{y \in \mathbb{R}^d} \left\{ F(y) + \frac{1}{2\gamma}\|y - x\|^2 \right\},$$

then we have the bound

$$\mathrm{dist}(x, \{y \in \mathbb{R}^d \mid \mathrm{dist}(0, \partial F(y))^2 \leq \gamma^{-2}\|x - \hat{x}\|^2\}) \leq \|x - \hat{x}\|^2. \tag{11}$$

Proof. As $\hat{x}$ is a minimizer, we have

$$0 \in \partial \left[F(\cdot) + \frac{1}{2\gamma}\|\cdot - x\|^2 \right](y) = \partial F(y) + \frac{1}{\gamma}(y - x),$$

where the equality follows by the sum rule for a smooth additive term $(2\gamma)^{-1}\|\cdot - x\|^2$ [35]. Thus, we have the inclusion $\hat{x} \in \{y \in \mathbb{R}^d \mid \mathrm{dist}(0, \partial F(y))^2 \leq \gamma^{-2}\|x - \hat{x}\|^2\}$, which leads to the desired conclusion. ∎

Based on this lemma, the iterate x_t is ε-close to an ε-stationary point in expectation whenever

$$\mathbb{E}\|x_t - \hat{x}_t\|^2 \leq \min\{\varepsilon, \gamma^2\varepsilon\}.$$

Establishing this fact is the main technical goal of the following theorem.

Theorem 3.3 (convergence of PGSG). *Let $x_0 \in \mathcal{X}$, consider any $T \in \mathbb{N}$, and let $x_R = \mathrm{PGSG}(x_0, G, \gamma, \{j_t\}, T)$. Define the quantity*

$$\mathcal{B}_{T,\{j_t\}} := \frac{4}{T\mu}\left(F(x_0) - \inf F + \sum_{t=0}^{T-1} \frac{72L^2}{\mu(j_t+1)}\right).$$

Then $\mathbb{E}\|x_R - \hat{x}_R\|^2 \leq \mathcal{B}_{T,\{j_t\}}$. Consequently, we have the following bound:

$$\mathbb{E}\left[\mathrm{dist}(x_R, \{x \mid \mathrm{dist}(0, \partial F(x))^2 \leq \gamma^{-2}\mathcal{B}_{T,\{j_t\}}\})^2\right] \leq \mathcal{B}_{T,\{j_t\}}.$$

In particular, given $\Delta \geq F(x_0) - \inf F$, and setting

$$j_t := \left\lceil \max\left(\frac{576L^2}{\mu^2 \min\{\varepsilon, \varepsilon\gamma^2\}}, \frac{11}{\gamma^2\mu^2}\right)\right\rceil \qquad \text{and} \qquad T := \left\lceil \frac{4\Delta}{\mu \min\{\varepsilon, \varepsilon\gamma^2\}}\right\rceil,$$

we have

$$\mathbb{E}\left[\mathrm{dist}(x_R, \{x \mid \mathrm{dist}(0, \partial F(x))^2 \leq \varepsilon\})^2\right] \leq \varepsilon.$$

The total number of stochastic oracle evaluations required to compute this point is bounded by $j_t \cdot T = O(\Delta L^2 \varepsilon^{-2})$.

Remark 1 (obtaining a nearly stationary point). As stated, the theorem indicates that x_R is nearby a nearly stationary point. The proof of Lemma 3.2 shows that one can in principle obtain the nearly stationary point $\hat{x}_R$ by solving the *strongly convex* stochastic optimization problem

$$\hat{x}_R = \mathrm{argmin}_{x \in \mathbb{R}^d}\left\{F(x) + \tfrac{1}{2\gamma}\|x - x_R\|^2\right\},$$

which is solvable to any desired degree of accuracy (in expectation). Furthermore, Lemma 3.2 shows that one can estimate the degree of stationarity of $\hat{x}_R$ via the bound $\mathrm{dist}(0, \partial F(\hat{x}_R))^2 \leq \gamma^{-2}\|x_R - \hat{x}_R\|^2$. In particular, given an estimate $\tilde{x}_R \approx \hat{x}_R$, we have the bound

$$\mathrm{dist}(0, \partial F(\hat{x}_R))^2 \leq 2\gamma^{-2}\|x_R - \tilde{x}_R\|^2 + 2\gamma^{-2}\|\tilde{x}_R - \hat{x}_R\|^2,$$

which indicates that $2\gamma^{-2}\|x_R - \tilde{x}_R\|^2$ may serve as a bound on the true stationarity of $\hat{x}_R$ (up to tolerance $2\gamma^{-2}\|\tilde{x}_R - \hat{x}_R\|^2$).

3.1. Proof of Theorem 3.3.

Throughout the proof we will need the following bound on the proximal-point step.

Lemma 3.4 (bounded step lengths). *Let $\gamma > 0$, $x \in \mathcal{X}$, and suppose that*

$$\hat{x} \in \mathrm{argmin}_{y \in \mathbb{R}^d}\left\{F(y) + \tfrac{1}{2\gamma}\|y - x\|^2\right\}.$$

Then $\gamma^{-1}\|x - \hat{x}\| \leq 2L$.

Proof. Note that

$$\frac{1}{2\gamma}\|x-\hat{x}\|^2 \leq F(x)-F(\hat{x}) \leq L\|x-\hat{x}\|,$$

where Lipschitz continuity follows from Lemma 3.1. Divide both sides of the inequality by $\frac{1}{2}\|x-\hat{x}\|$ to get the result. ∎

We now analyze one inner loop of Algorithm 2. This inner loop may be interpreted as a variant of the stochastic projected subgradient method applied to the strongly convex optimization problem

$$\text{minimize}_{x\in\mathbb{R}^d}\ F_y(x) := F(x) + \tfrac{1}{2\gamma}\|x-y\|^2.$$

We note that the following proof is similar in outline to that of [24], but the results of that work are not sufficient for our purposes.

PROPOSITION 3.5 (analysis of PSSM). *Let $y \in \mathcal{X}$ and let $\hat{y}$ be the unique minimizer of $F_y(x)$ over all $x \in \mathbb{R}^d$. Set $\tilde{y} = \mathrm{PSSM}(y, G, \gamma, \{\alpha_j\}, J)$. Then if $\gamma \in (0, 1/\rho)$ and $\{\alpha_j\}$ is chosen as in* (9)*, we have*

$$\begin{aligned}
\mathbb{E}\left[F_y(\tilde{y}) - F_y(\hat{y})\right] &\leq \frac{72L^2}{\mu(J+1)} + \frac{30\|y-\hat{y}\|^2}{\gamma^4\mu^3 J(J+1)},\\
\mathbb{E}\left[\|\tilde{y}-\hat{y}\|^2\right] &\leq \frac{144L^2}{\mu^2(J+1)} + \frac{60\|y-\hat{y}\|^2}{\gamma^4\mu^4 J(J+1)},\\
\mathbb{E}\left[\|y-\tilde{y}\|^2\right] &\leq \frac{288L^2}{\mu^2(J+1)} + \left(2 + \frac{120}{\gamma^4\mu^4 J(J+1)}\right)\|y-\hat{y}\|^2.
\end{aligned}$$

On the other hand, if $0 < \alpha_j \leq 2\gamma$ for all j, but $\{\alpha_j\}$ and γ are otherwise unconstrained, we have

$$\mathbb{E}\left[\|y-\tilde{y}\|\right] \leq L\sum_{i=0}^{J-1}\alpha_i.$$

Proof. Since $\hat{y} \in \mathcal{X}$ and $\mathrm{proj}_{\mathcal{X}}$ is nonexpansive, we have

$$\begin{aligned}
\|y_{j+1}-\hat{y}\|^2 &\leq \|y_j - \alpha_j v_j - \hat{y}\|^2\\
&= \|y_j-\hat{y}\|^2 - 2\alpha_j\langle y_j-\hat{y}, v_j\rangle + \alpha_j^2\|v_j\|^2. \qquad (12)
\end{aligned}$$

To proceed further, we must bound $\|v_j\|^2$. To that end, recall that F_y is μ-strongly convex. Therefore, for any $x \in \mathcal{X}$,

$$\begin{aligned}
\mathbb{E}_z\left\|G(x,z)+\frac{1}{\gamma}(x-y)\right\|^2 &= \mathbb{E}_z\left\|G(x,z)-\frac{1}{\gamma}(\hat{y}-y)+\frac{1}{\gamma}(x-\hat{y})\right\|^2\\
&\leq \mathbb{E}_z 3\|G(x,z)\|^2 + 3\left\|\frac{1}{\gamma}(\hat{y}-y)\right\|^2 + 3\left\|\frac{1}{\gamma}(x-\hat{y})\right\|^2\\
&\leq 15L^2 + 3\left\|\frac{1}{\gamma}(x-\hat{y})\right\|^2\\
&\leq 15L^2 + \frac{6}{\gamma^2\mu}(F_y(x)-F_y(\hat{y})),
\end{aligned}$$

where the first inequality follows from Jensen's inequality, the second inequality uses (A3) twice and Lemma 3.4, and the third inequality follows from the strong convexity.

Returning to (12), we let $\bar{v}_j = \mathbb{E}_j v_j \in \partial F_y(y_j)$, where $\mathbb{E}_j[\cdot]$ denotes the expectation conditioned on $y_1, \ldots, y_j$. Now, we take the conditional expectation of both sides of the equation, which yields

$$\begin{aligned}
\mathbb{E}_j\|y_{j+1} - \hat{y}\|^2 &\le \mathbb{E}_j\|y_j - \hat{y}\|^2 - 2\alpha_j\langle y_j - \hat{y}, \bar{v}_j\rangle + \alpha_j^2\mathbb{E}_j\|v_j\|^2 \\
&\le \mathbb{E}_j\|y_j - \hat{y}\|^2 + \alpha_j^2\left(15L^2 + \frac{6}{\gamma^2\mu}\mathbb{E}_j F_y(y_j) - F_y(\hat{y})\right) \\
&\quad - 2\alpha_j\left(\mathbb{E}_j F_y(y_j) - F_y(\hat{y}) + \frac{\mu}{2}\mathbb{E}_j\|y_j - \hat{y}\|^2\right) \\
&= (1 - \alpha_j\mu)\,\mathbb{E}_j\|y_j - \hat{y}\|^2 + 15\alpha_j^2 L^2 \\
&\quad - \left(2\alpha_j - \frac{6\alpha_j^2}{\gamma^2\mu}\right)(\mathbb{E}_j F_y(y_j) - F_y(\hat{y})) \\
&\le (1 - \alpha_j\mu)\,\mathbb{E}_{t,j}\|y_j - \hat{y}\|^2 + 15\alpha_j^2 L^2 - \alpha_j\left(\mathbb{E}_j F_y(y_j) - F_y(\hat{y})\right),
\end{aligned}$$

where the second inequality uses our bound on $\mathbb{E}_z\|G(x,z) + \gamma^{-1}(x-y)\|^2$ and the strong convexity of F_y, and the third inequality is a consequence of the bound

$$\frac{6\alpha_j}{\gamma^2\mu} = \frac{2\mu(j+2)(6/\gamma^2\mu)}{(\mu(j+2))^2 + \frac{36}{\gamma^4\mu^2}\frac{j+2}{j+1}} \le \frac{2\mu(j+2)(6/\gamma^2\mu)}{(\mu(j+2))^2 + (6/\gamma^2\mu)^2} \le 1.$$

Multiplying by $(j+1)/\alpha_j$, we find that

$$\begin{aligned}
(j+1)\alpha_j^{-1}\mathbb{E}_j\|y_{j+1} - \hat{y}\|^2 \le (j+1)\left(\alpha_j^{-1} - \mu\right)\mathbb{E}_j\|y_j - \hat{y}\|^2 + 15(j+1)\alpha_j L^2 \\
- (j+1)\left(\mathbb{E}_j F_y(y_j) - F_y(\hat{y})\right).
\end{aligned}$$

By our choice of α_j, we have $(j+1)\alpha_j^{-1} = (j+2)(\alpha_{j+1}^{-1} - \mu)$. Therefore, summing the previous inequality, we have

$$0 \le \left(\alpha_0^{-1} - \mu\right)\|y - \hat{y}\|^2 + 15L^2\sum_{j=0}^{J-1}(j+1)\alpha_j - \sum_{j=0}^{J-1}(j+1)\left(\mathbb{E}_j F_y(y_j) - F_y(\hat{y})\right).$$

Therefore, noting that $\sum_{j=0}^{j_t-1}(j+1)\alpha_j \le 2j_t/\mu$ and $\alpha_0^{-1} - \mu = 18/(\gamma^4\mu^3)$, and using the convexity of F_y, we deduce that

$$\mathbb{E}\left(F_y(\tilde{y}) - F_y(\hat{y})\right) \le \frac{36\|y - \hat{y}\|^2}{\gamma^4\mu^3 J(J+1)} + \frac{60L^2}{\mu(J+1)}.$$

The first distance bound then follows as a direct consequence of the strong convexity of F_y, while the second follows from the convexity of $\|\cdot\|^2$.

Finally, we now work in the case in which γ may be strictly greater than $1/\rho$. We claim that for all $j = 0, \ldots, J-1$, we have $\mathbb{E}[\|y_j - y_0\|] \le L\sum_{i=0}^{j}\alpha_i$. Indeed, this is clearly true for $j = 0$. Inductively, we also have

$$\begin{aligned}
\mathbb{E}_j\|y_{j+1} - x_t\| &\le \mathbb{E}_j\|y_j - \alpha_j\left(G(y_j, z_j) + (y_j - x_t)/\gamma\right) - x_t\| \\
&\le |1 - \alpha_j/\gamma| \cdot \mathbb{E}_j\|y_j - x_t\| + \alpha_j\mathbb{E}_{\Xi_0}\|G(y_j, z_j)\| \\
&\le \mathbb{E}_j\|y_j - x_t\| + \alpha_j L,
\end{aligned}$$

where the first inequality follows by nonexpansiveness of $\mathrm{proj}_{\mathcal{X}}$ and the third follows from the inequality $0 < \alpha_j \le 2\gamma$. Applying the law of expectation completes the inductive step. Therefore, we have

$$\mathbb{E}\left[\|y-\tilde{y}\|\right] \le \mathbb{E}\left[\frac{2}{J(J+1)}\sum_{j=0}^{J-1}(j+1)\|y_j-y\|\right] \le \frac{2}{J(J+1)}\sum_{j=0}^{J-1}(j+1)\left(L\sum_{i=0}^{j}\alpha_i\right)$$
$$\le L\sum_{i=0}^{J-1}\alpha_i,$$

as desired. □

We now give the proof of Theorem 3.3.

Proof of Theorem 3.3. By the strong convexity of the proximal-point subproblem, we have

$$F(\hat{x}_t) \le F(x_t) - \left(\frac{1}{2\gamma}+\frac{\mu}{2}\right)\|\hat{x}_t - x_t\|^2.$$

Then by Proposition 3.5, we have the following bound:

$$\mathbb{E}_t\left[F\left(x_{t+1}\right)\right] \le F(\hat{x}_t) + \frac{1}{2\gamma}\|\hat{x}_t - x_t\|^2 + \frac{72L^2}{\mu(j_t+1)} + \frac{30\|x_t-\hat{x}_t\|^2}{\gamma^4\mu^3 j_t(j_t+1)}$$
$$\le F(x_t) + \frac{72L^2}{\mu(j_t+1)} - \left(\frac{\mu}{2} - \frac{30}{\gamma^4\mu^3 j_t(j_t+1)}\right)\|x_t-\hat{x}_t\|^2,$$

where $\mathbb{E}_t[\cdot]$ denotes the expectation conditioned on $x_1,\ldots,x_t$. Rearranging, using the lower bound on j_t (which makes the multiple of $\|\hat{x}_t - x_t\|^2$ larger than $\mu/4$ as $30/121 < 1/4$), applying the law of total expectation, and summing, we find that

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|x_t-\hat{x}_t\|^2\right] \le \frac{4}{T\mu}\left(F(x_0) - \inf F + \sum_{t=0}^{T-1}\frac{72L^2}{\mu(j_t+1)}\right),$$

as desired. To complete the proof, apply Lemma 3.2. □

3.2. Probabilistic guarantees. In the previous section, we developed expected complexity results which describe the average behavior of the PGSG over multiple runs. We are also interested in giving guarantees for a single run of an algorithm. Thus, in this section we recall the notion of an (ε,Λ)-solution given in the introduction: a random variable $\bar{x}$ is called an *(ε,Λ)-solution* if

$$\mathbb{P}\left(\operatorname{dist}(\bar{x},\{x \mid \operatorname{dist}(0,\partial F(x))^2 \le \varepsilon\})^2 \le \varepsilon\right) \ge 1-\Lambda.$$

Theorem 3.3 together with Markov's inequality implies that x_R, generated with

$$j_t := \left\lceil \max\left(\frac{576L^2}{\mu^2\min\{\varepsilon\Lambda,\varepsilon\Lambda\gamma^2\}}, \frac{12}{\gamma^2\mu^2}\right)\right\rceil \qquad \text{and} \qquad T := \left\lceil \frac{4\Delta}{\mu\min\{\varepsilon\Lambda,\varepsilon\Lambda\gamma^2\}}\right\rceil,$$

where $\Delta \ge F(x_0) - \inf F$, is an (ε,Λ)-solution after

$$j_t\cdot T = O(\Delta L^2(\varepsilon\Lambda)^{-2}) \tag{13}$$

stochastic oracle evaluations. In this section, we develop a two-stage algorithm that significantly improves the dependence on Λ in this bound.

The method we propose proceeds in two phases. In the first phase, multiple independent copies of PGSG are called, resulting in candidates $x_{R^1}, \ldots, x_{R^S}$. For each of the candidates, we then compute an approximate proximal point $\tilde{x}_{R^s} \approx \hat{x}_{R^s}$. In the second phase, we select one of the candidates $x_{R^{\bar{s}}}$ based on the size of $\gamma^{-1}\|x_{R^s} - \tilde{x}_{R^s}\|$, a proxy for the true proximal step length. We will see that such a point is an (ε, Λ)-solution, and the total number of stochastic oracle evaluations has a much better dependence on Λ.

Before we introduce the algorithm, let us define three parameters

$$j_t := \left\lceil \max\left\{ \frac{576L^2}{\mu^2 \min\{\varepsilon/24, \varepsilon\gamma^2/24\}}, \frac{11}{\gamma^2\mu^2} \right\} \right\rceil, \qquad T := \left\lceil \frac{4\Delta}{\mu \min\{\varepsilon/24, \varepsilon\gamma^2/24\}} \right\rceil, \tag{14}$$

and

$$J := \left\lceil \max\left\{ \frac{48L^2\sqrt{2}}{\mu \min\{\varepsilon, \varepsilon\gamma^2\}} \cdot \frac{S}{\Lambda}, \frac{11}{\gamma^2\mu^2} \cdot \sqrt{\frac{S}{\Lambda}} \right\} \right\rceil,$$

where $\Delta \geq F(x_0) - \inf F$. The algorithm now follows.

Algorithm 3 Two-phase proximally guided stochastic subgradient method 2PGSG(x_0, G, γ, J, S).

```
Input: $x_0 \in \mathcal{X}$, weak convexity constant $\rho > 0$, $\gamma \in (0, 1/\rho)$, stochastic subgradient
    iteration $J \in \mathbb{N}$, number of copies $S \in \mathbb{N}$.
Define the maximum iterations $T$ as in (14).
Define the maximum inner-loop iteration $\{j_t\}$ as in (14).
Define the step-size sequence $\{\alpha_j\}$ as in (9).
Optimization phase.
for $s = 1, \ldots, S$ do
    set $x_{R^s} = \mathrm{PGSG}(x_0, G, \gamma, \{j_t\}, T)$,
    set $\tilde{x}_{R^s} = \mathrm{PSSM}(x_{R^s}, G, \gamma, \{\alpha_j\}, J)$
end for
Post-optimization phase.
Choose $x^* = x_{R^{\bar{s}}}$ from the candidate list $\{x_r^s\}_{s=1}^S$ such that
        $\bar{s} = \mathrm{argmin}_{s=1,\ldots,S} \ \|x_{R^s} - \tilde{x}_{R^s}\|$.
Output: $x^*$.
```

The analysis of this algorithm requires a bound on the expectation of $\|x_{R^s} - \tilde{x}_{R^s}\|^2$ and $\|\tilde{x}_{R^s} - \hat{x}_{R^s}\|^2$, which we now provide.

Lemma 3.6. *Let x_{R^s} be generated as in Algorithm* 3. *Then*

$$\mathbb{E}\left[\|x_{R^s} - \tilde{x}_{R^s}\|^2\right] \leq \frac{1}{4}\min\{\varepsilon, \gamma^2\varepsilon\},$$
$$\mathbb{E}\left[\|\tilde{x}_{R^s} - \hat{x}_{R^s}\|^2\right] \leq \frac{\Lambda}{4S}\min\{\varepsilon, \gamma^2\varepsilon\}.$$

Proof. By Proposition 3.5 and Theorem 3.3, the bound holds:

$$\begin{aligned}\mathbb{E}\left[\|x_{R^s}-\tilde{x}_{R^s}\|^2\right] &\le \frac{288L^2}{\mu^2(J+1)}+\left(2+\frac{120}{\gamma^4\mu^4J(J+1)}\right)\mathbb{E}\left[\|x_{R^s}-\hat{x}_{R^s}\|^2\right]\\ &\le \frac{288L^2}{\mu^2(J+1)}+\left(2+\frac{120}{\gamma^4\mu^4J(J+1)}\right)\mathcal{B}_{T,\{j_t\}}\\ &\le \frac{\Lambda}{8S}\min\{\varepsilon,\gamma^2\varepsilon\}/8+\min\{\varepsilon,\gamma^2\varepsilon\}/8\le\min\{\varepsilon,\gamma^2\varepsilon\}/4,\end{aligned}$$

which proves the first bound.

On the other hand, Proposition 3.5 and Theorem 3.3 imply that

$$\begin{aligned}\mathbb{E}\left[\|\tilde{x}_{R^s}-\hat{x}_{R^s}\|^2\right] &\le \frac{144L^2}{\mu^2(J+1)}+\frac{60}{\gamma^4\mu^4J(J+1)}\mathbb{E}\left[\|x_{R^s}-\hat{x}_{R^s}\|^2\right]\\ &\le \frac{144L^2}{\mu^2(J+1)}+\frac{60}{\gamma^4\mu^4J(J+1)}\mathcal{B}_{T,\{j_t\}}\\ &\le \frac{\Lambda}{8S}\min\{\varepsilon,\gamma^2\varepsilon\}+\frac{\Lambda}{8S}\min\{\varepsilon,\gamma^2\varepsilon\}=\frac{\Lambda}{4S}\min\{\varepsilon,\gamma^2\varepsilon\},\end{aligned}$$

which proves the second bound and completes the proof. □

We now state the convergence guarantees for Algorithm 3.

THEOREM 3.7. *Let $x_0\in\mathcal{X}$ and $S=\log_2(2/\Lambda)$. Then $x^*=2\text{PGSG}(x_0,G,\gamma,J,S)$ returned by Algorithm 3 is an (ε,Λ)-solution. The total number of stochastic oracle evaluations called by Algorithm 3 is equal to*

$$S\cdot(j_t\cdot T+J)=O\left(\frac{\log_2(1/\Lambda)\Delta L^2}{\varepsilon^2}+\frac{\log_2(1/\Lambda)L^2}{\varepsilon\Lambda}\right). \tag{15}$$

Proof. By Lemma 3.2, it suffices to show that

$$\mathbb{P}\left(\|x^*-\hat{x}^*\|^2\le\min\{\varepsilon,\gamma^2\varepsilon\}\right)\ge 1-\Lambda.$$

To that end, note that

$$\begin{aligned}\|x^*-\hat{x}^*\|^2 &= \|(x_{R^{\bar{s}}}-\hat{x}_{R^{\bar{s}}})\|^2\\ &\le 2\|x_{R^{\bar{s}}}-\tilde{x}_{R^{\bar{s}}}\|^2+2\|\tilde{x}_{R^{\bar{s}}}-\hat{x}_{R^{\bar{s}}}\|^2\\ &\le 2\min_{s=1,\ldots,S}\|x_{R^s}-\tilde{x}_{R^s}\|^2+2\max_{s=1,\ldots,S}\|\tilde{x}_{R^s}-\hat{x}_{R^s}\|^2.\end{aligned}$$

Therefore, we have

$$\begin{aligned}&\mathbb{P}\left(\|x^*-\hat{x}^*\|^2\ge\min\{\varepsilon,\gamma^2\varepsilon\}\right)\\ &\le\mathbb{P}\left\{\min_{s=1,\ldots,S}\|x_{R^s}-\tilde{x}_{R^s}\|^2\ge\frac{1}{2}\min\{\varepsilon,\gamma^2\varepsilon\}\right)\\ &\quad+\mathbb{P}\left(\max_{s=1,\ldots,S}\|\tilde{x}_{R^s}-\hat{x}_{R^s}\|^2\ge\frac{1}{2}\min\{\varepsilon,\gamma^2\varepsilon\}\right).\end{aligned}$$

Notice that by Markov's inequality, independence, and Proposition 3.6, we have

$$\mathbb{P}\left(2\min_{s=1,\ldots,S}\|x_{R^s}-\tilde{x}_{R^s}\|^2\ge\frac{1}{2}\min\{\varepsilon,\gamma^2\varepsilon\}\right)\le 2^{-S}\le\frac{\Lambda}{2}.$$

On the other hand, by Markov's inequality, a union bound, and Proposition 3.6, we have

$$\mathbb{P}\left(2 \max_{s=1,\ldots,S} \|\tilde{x}_{R^s} - \hat{x}_{R^s}\|^2 \geq \frac{1}{2}\min\{\varepsilon, \gamma^2\varepsilon\}\right) \leq \frac{\Lambda}{2},$$

which shows that x^* is an (ε, Λ)-solution. ∎

When the second term in (15) is dominating, the bound (15) is $\log_2(2/\Lambda)/\varepsilon\Lambda$ times smaller than the bound (13) obtained by the PGSG algorithm.

3.3. PGSG with unknown weak convexity constant. Algorithm 2 requires that the parameters ε, L, ρ, and Δ are known. In practice, computing bounds on L, ρ, and Δ may be nontrivial. In this section we show that a simple strategy—letting j_t tend to infinity and γ_t tend to zero—results in a sublinear convergence rate without knowledge of any problem parameters. We formalize this procedure in Algorithm 4 using the following parameters: fix a hyper-parameter $0 < \beta < 1$, and define

$$\gamma_t := (t+1)^{-\beta}, \tag{16}$$

$$j_t := t + 44, \tag{17}$$

$$\alpha_{t,j} := \frac{4\gamma_t}{j + 1 + \frac{288}{j+1}}. \tag{18}$$

The algorithm now follows.

Algorithm 4 Parameter-free proximally guided stochastic subgradient method PFPGSG(y_0, G, T, β).

```
Input: $x_0 \in \mathcal{X}$, maximum iterations $T \in \mathbb{N}$, hyper-parameter $0 < \beta < 1$.
1: Define the sequence $\{\gamma_t\}$ as in (16).
2: Define the step-size sequence $\{\alpha_{t,j}\}$ as in (18).
3: Define the maximum inner-loop iteration $\{j_t\}$ as in (17).
4: for $t = 0, \ldots, T-2$ do
5:     $x_{t+1} = \mathrm{PSSM}(x_t, G, \gamma_t, \{\alpha_{t,0}, \alpha_{t,1}, \alpha_{t,2}, \ldots\}, j_t)$
6: end for
Output: $x_R$, where $R$ is sampled with probability $\mathbb{P}(R = t) \propto \gamma_t$ from $\{0, \ldots, T-1\}$.
```

In the following, we establish convergence guarantees for the parameter-free variant of PGSG. The proof splits the analysis of PFPGSG into two parts. In the first part, $\gamma_t \geq 1/\rho$. In this setting, the analysis of the previous section does not apply. Thus, we show that the iterates do not wander very far. In the second part, $\gamma_t \leq 1/\rho$, and an argument similar to the one presented in Theorem 3.3 applies. Combining these results then leads to the theorem. To that end, we address the first part now.

LEMMA 3.8. *Let $T_0 = \lceil (2\rho)^{1/\beta} \rceil$. Then*

$$\mathbb{E}\left[F(x_{T_0})\right] \leq F(x_0) + L^2 T_0 \log(T_0 + 125).$$

Proof. By Proposition 3.5, as $\alpha_j < 2\gamma_t$, we have

$$\mathbb{E}_{T_0}\|x_{t+1} - x_t\| \leq L \sum_{j=0}^{j_t - 1} \alpha_j$$

for all $t = 0, \ldots, T_0 - 1$. Therefore

$$\begin{aligned}
\mathbb{E}_{T_0} F(x_{T_0}) &\leq F(x_0) + L\mathbb{E}_{T_0}\|x_{T_0} - x_0\| \\
&\leq F(x_0) + L\sum_{t=0}^{T_0-1}\mathbb{E}_{T_0}\|x_{t+1} - x_t\| \\
&\leq F(x_0) + L^2\sum_{t=0}^{T_0-1}\sum_{j=0}^{j_t-1}\alpha_{t,j} \\
&\leq F(x_0) + L^2\sum_{t=0}^{T_0-1}\sum_{j=0}^{j_t-1}\frac{4}{j+1} \\
&\leq F(x_0) + L^2 T_0 \log(T_0 + 125), \qquad (19)
\end{aligned}$$

as desired. □

We now address the second part of the argument, and with it, deduce the following theorem. At first glance, the presented rate appears to be better than the rate obtained by Algorithm 2, which requires knowledge of ρ. However, it is not because the factor $\gamma_R^{-2} = (R+1)^{2\beta}$ is no longer a constant. Instead, the convergence rate of Algorithm 4 is of the order of $O(T^{1-\beta})$ in the worst case.

THEOREM 3.9 (convergence of parameter-free PGSG). *Let $T_0 = \lceil (2\rho)^{1/\beta} \rceil$ and consider any $T \in \mathbb{N}$. Let $x_R = \mathrm{PFPGSG}(x_0, G, T, \beta)$. Define the quantity*

$$\mathcal{C}_{T,\{j_t\}} := \frac{8(1+\beta)}{(T+1)^{1+\beta}}\left(F(x_0) - \inf F + (144C + T_0\log(T_0+125) + \tfrac{T_0}{2})L^2\right),$$

where $C := \sum_{t=T_0}^{\infty} t^{-1-\beta} < \infty$. Then $\mathbb{E}\|x_R - \hat{x}_R\|^2 \leq \mathcal{C}_{T,\{j_t\}}$. Consequently, we have the following bound:

$$\mathbb{E}\left[\mathrm{dist}(x_R, \{x \mid \mathrm{dist}(0, \partial F(x))^2 \leq (R+1)^{2\beta}\mathcal{C}_{T,\{j_t\}}\})^2\right] \leq \mathcal{C}_{T,\{j_t\}}.$$

Proof. Suppose that $t \geq T_0$ and notice this ensures $\gamma_t \in (0, 1/\rho)$. Following an argument nearly identical to the proof of Theorem 3.3, we find that for all $t \geq T_0$, we have

$$\mathbb{E}_t\left[F\left(x_{t+1}\right)\right] \leq F(x_t) + \frac{72L^2}{\mu_t(j_t+1)} - \left(\frac{\mu_t}{2} - \frac{30}{\gamma_t^4\mu_t^3 j_t(j_t+1)}\right)\|x_t - \hat{x}_t\|^2,$$

where $\mu_t = \gamma_t^{-1} - \rho$ and $\mathbb{E}_t[\cdot]$ denotes the expectation conditioned on $x_1, \ldots, x_t$. We now show that the coefficient of $-\|x_t - \hat{x}_t\|^2$ is greater than or equal to $\mu_t/4$. Indeed, it suffices to show that $j_t \geq 12/(1-\gamma_t\rho)^2$. To that end, note that

$$\gamma_t \leq (\lceil (2\rho)^{1/\beta} \rceil + 1)^{-\beta} \leq 1/(2\rho).$$

Therefore, $1 - \gamma_t\rho \geq 1/2$, which leads to the claimed inequality: $12/(1-\gamma_t\rho)^2 \leq 44 \leq j_t$.

Using the lower bound $\mu_t \geq 1/(2\gamma_t)$ (which follows because $t \geq T_0$), we thus find that

$$\begin{aligned}
\sum_{t=T_0}^{T-1}\frac{1}{8\gamma_t}\mathbb{E}\left[\|x_t - \hat{x}_t\|^2\right] &\leq \mathbb{E}\left[F(x_{T_0}) - \inf F\right] + \sum_{t=T_0}^{T-1}\frac{144\gamma_t L^2}{(j_t+1)} \\
&\leq F(x_0) - \inf F + \sum_{t=T_0}^{T-1}\frac{144\gamma_t L^2}{(j_t+1)} + L^2 T_0\log(T_0+125).
\end{aligned}$$

We would like to extend the sum on the left-hand side of the previous inequality to all t between 0 and $T-1$. To that end, we bound the excess terms:

$$\sum_{t=0}^{T_0-1} \frac{1}{8\gamma_t} \mathbb{E}\left[\|x_t - \hat{x}_t\|^2\right] \leq \sum_{t=0}^{T_0-1} \frac{\gamma_t L^2}{2} \leq \frac{T_0 L^2}{2}.$$

Therefore, using the bounds $\sum_{t=T_0}^{\infty} \gamma_t/(j_t+1) \leq \sum_{t=T_0}^{\infty} t^{-1-\beta} = C < \infty$ and $\sum_{t=0}^{T-1} \gamma_t^{-1} \geq \int_{-1}^{T-1} (t+1)^\beta dt = T^{1+\beta}/(1+\beta)$, we have

$$\begin{aligned} &\mathbb{E}\left[\|x_R - \hat{x}_R\|^2\right] \\ &= \frac{1}{\sum_{t=0}^{T-1} \gamma_t^{-1}} \sum_{t=0}^{T-1} \frac{1}{\gamma_t} \mathbb{E}\left[\|x_t - \hat{x}_t\|^2\right] \\ &\quad \leq \frac{8(1+\beta)}{(T+1)^{1+\beta}} \left(F(x_0) - \inf F + 144CL^2 + L^2 T_0 \log(T_0 + 125) + \frac{T_0 L^2}{2} \right), \end{aligned}$$

as desired. To complete the proof, apply Lemma 3.2. □

We remark that this convergence rate can be directly utilized to give a complexity bound on computing an ε-expected stationary point. Completing T outer iterations requires $O(T^2)$ oracle evaluations since j_t is selected according to (17). Then observing that $\mathcal{C}_{T,\{j_t\}} \leq \varepsilon$ if $T \geq O(\varepsilon^{-1/(1+\beta)})$, we must find an ε-expected stationary point after at most $O(\varepsilon^{-2/(1+\beta)})$ oracle evaluations.

4. Experimental results.

In this section we address the population version of the robust real phase retrieval problem: fix a vector $\bar{x} \in \mathbb{R}^d$ and define

$$F(x) := \mathbb{E}_{a,\delta,\xi}\left[|\langle a, x\rangle^2 - (\langle a, \bar{x}\rangle^2 + \delta \cdot \xi)|\right], \tag{20}$$

where a, δ, and ξ are independent random variables satisfying the following assumptions:

(B1) a is a zero mean standard Gaussian random variable in $\mathbb{R}^d$,
(B2) δ is a $\{0,1\}$-random variable with $P(\delta = 1) = 0.25$,
(B3) ξ is a zero mean Laplace random variable with scale parameter 1.

In this setting, it is possible to show that the only minimizers of $F(x)$ are $\pm\bar{x}$ [10, Lemma B.8]. In Lemma B.1, we show that this function is 2-weakly convex.

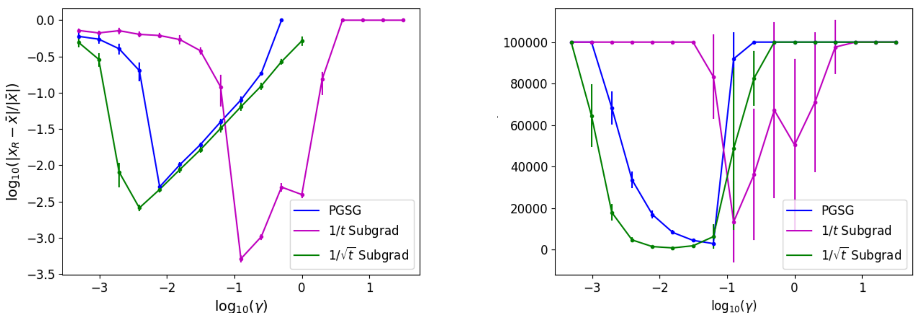
Implementation. Each step of PGSG and the stochastic subgradient method requires access to a subgradient of a random function of the form

$$f(x, a, \delta, \xi) = |\langle a, x\rangle^2 - (\langle a, \bar{x}\rangle^2 + \delta \cdot \xi)|.$$

We choose the selection operator

$$G(x, a, \delta, \xi) = 2\langle a, x\rangle a \cdot \operatorname{sign}(\langle a, x\rangle^2 - (\langle a, \bar{x}\rangle^2 + \delta \cdot \xi)) \in \partial_x f(x, a, \delta, \xi).$$

It is a straightforward exercise to show that G satisfies Assumption A on any bounded set $\mathcal{X}$. For our purposes we choose $\mathcal{X}$ to be a closed ball with a large radius, $r = 10^6$. In our experiments, we never had to explicitly enforce this constraint.



(a) (b)

FIG. 1. *Performance of PGSG and the subgradient method for values of* γ *averaged over* 50 *trials. Error bars are included to show one standard deviation. Plot* (a) *shows the relative distance to a minimizer after* 25000 *subgradient evaluations. Plot* (b) *shows the number of subgradient evaluations needed until the relative distance* 0.05 *to a minimizer.*

Experiment 1: Sensitivity to step size. In the first experiment we compare the performance of PGSG to the stochastic subgradient method, which possessed no complexity guarantees at the time of writing this manuscript. In the stochastic subgradient method, we chose step sizes of the form $\gamma/(t+10)^\beta$ for varying $\gamma > 0$ and $\beta \in \{1/2, 1\}$. For PGSG, we chose varying values of $\gamma > 0$ and then set α_j by (9), $j_t = 250$, and $\mu = 1/2\gamma$. Figure 1 shows the result of running these two methods to solve robust real phase retrieval problems with $d = 50$.

Like the choice of γ and consequently α_j, the choice of j_t is also important for the practical performance of PGSG. Condition (8) used in the analysis of PGSG is often overly conservative and leads to worse performance in practice. In our experiments, we chose the constant $j_t = 250$ to balance the quality of the solution to the proximal subproblems with the total number of approximately solved subproblems.

Experiment 2: Mean and variance of solution estimates. Unlike the subgradient method, PGSG provides an easily computed estimate of the of how close x_R is to a nearly stationary point; see the discussion surrounding Lemma 11 and Remark 1. For PGSG, 2PGSG, and PFPGSG, this is given by $\gamma^{-1}\|x_R - x_{R+1}\|$, $\gamma^{-1}\|x_{R^s} - \tilde{x}_{R^s}\|$, and $\gamma_R^{-1}\|x_R - x_{R+1}\|$, respectively. Proposition 3.5 shows these estimates are close to $\gamma^{-1}\|x_R - \hat{x}_R\|$ in expectation, which, according to Lemma 11, is a natural measure of stationarity. Using these stationarity measures, we analyze the numerical performance of the three algorithms proposed in this manuscript.

Based on the results of Experiment 1, we set $\gamma = 2^{-6}$ for the PGSG and 2PGSG algorithms. We furthermore set α_j according to (9) and let $\mu = 1/2\gamma$. For both methods, we consider two different selections for the number of inner iterations $j_t \in \{10^3, 10^4\}$. These choices determine the level of stationarity reached by the algorithm. For 2PGSG, we fix $S = 5$ and $J = 5T$. For PFPGSG, we set $\beta = 1/2$, $\gamma_t = (t+1)^{-\beta}/10$ (which differs from (16) by a factor of ten), j_t as in (17), and α_j as in (18).

Table 1 lists the mean and variance of the stationarity measures averaged over 50 trials. Each subcolumn shows the performance of the target algorithm as the computational budget increases. We find that with $j_t = 10^3$, both PGSG and 2PGSG quickly converge to a region of stationarity and then do not improve. With $j_t = 10^4$, both of these methods reach a level of stationarity an order of magnitude smaller than with the

TABLE 1
Estimated stationarity level for each of the proposed algorithms averaged over 50 *trails.*

Oracle calls		PGSG $j_t = 1000$	PGSG $j_t = 10000$	2PGSG $j_t = 1000$	2PGSG $j_t = 10000$	PFPGSG
		$d = 50$				
100000	mean	1.538	10.02	1.099	12.46	2.877
	var.	0.0380	1.683	0.0153	5.871	0.178
500000	mean	1.492	0.2043	1.024	8.406	1.615
	var.	0.0542	9.27e−4	0.0119	0.669	0.0421
2500000	mean	1.575	0.2083	1.034	0.1331	0.847
	var.	0.0600	7.53e−4	0.0152	2.562e−4	0.0128
		$d = 100$				
100000	mean	3.632	17.12	2.703	23.04	6.625
	var.	0.137	3.287	0.0544	6.117	0.361
500000	mean	3.579	3.678	2.534	11.83	3.815
	var.	0.145	22.35	0.0430	0.891	0.145
2500000	mean	3.622	0.540	2.564	0.365	2.121
	var.	0.127	2.71e−3	0.0468	1.01e−3	0.0380
		$d = 500$				
100000	mean	27.67	76.86	24.32	100.7	41.65
	var.	8.843	16.95	2.465	20.31	6.471
500000	mean	25.53	23.52	17.13	42.25	25.59
	var.	1.519	1.474	0.341	4.772	1.946
2500000	mean	25.64	4.759	17.10	3.519	15.09
	var.	1.236	0.0454	0.374	0.0118	0.452
		$d = 1000$				
100000	mean	64.73	156.5	34.36	199.5	59.25
	var.	14.37	49.09	2.388	53.92	167.2
500000	mean	55.97	40.99	33.61	86.97	54.48
	var.	3.426	2.091	0.890	9.233	71.38
2500000	mean	55.27	11.88	33.40	9.008	33.86
	var.	4.854	0.119	0.634	0.055	1.350

choice $j_t = 10^3$. Under sufficiently large computational budget (2500000 stochastic subgradient evaluations), the variance of the stationarity reported by 2PGSG is consistently lower than that of PGSG as expected from Theorem 3.7. Finally, we note that the performance of PFPGSG is similar to PGSG in most regimes.

Appendix A. Trimmed estimation.

PROPOSITION A.1. *Suppose that* $f_1, \ldots, f_n$ *are convex, L-Lipschitz continuous functions on* $\mathbb{R}^d$. *Then the objective*

$$F(w, x) = \begin{cases} \frac{1}{n}\sum_{i=1}^n w_i f_i(x) & \text{if } w_i \in [0,1] \text{ and } \sum_{i=1}^n w_i = h, \\ \infty & \text{otherwise} \end{cases}$$

is L-weakly convex.

Proof. We argue using Proposition 2.1. Let $(w, x), (\tilde{w}, \tilde{x}) \in \operatorname{dom} F$ and let $\lambda \in [0, 1]$. Then

$$\begin{aligned} &F((1-\lambda)(w,x) + \lambda(\tilde{w}, \tilde{x})) \\ &\quad = \frac{1}{n}\sum_{i=1}^n ((1-\lambda)w_i + \lambda\tilde{w}_i) f_i((1-\lambda)x + \lambda\tilde{x}) \\ &\quad = \frac{1}{n}\sum_{i=1}^n (1-\lambda)w_i f_i((1-\lambda)x + \lambda\tilde{x}) + \frac{1}{n}\sum_{i=1}^n \lambda\tilde{w}_i f_i((1-\lambda)x + \lambda\tilde{x}) \end{aligned}$$

$$
\begin{aligned}
&\le \frac{1}{n}\sum_{i=1}^{n}(1-\lambda)w_i((1-\lambda)f_i(x)+\lambda f_i(\tilde{x})) + \frac{1}{n}\sum_{i=1}^{n}\lambda\tilde{w}_i((1-\lambda)f_i(x)+\lambda f_i(\tilde{x}))\\
&= \frac{1}{n}\sum_{i=1}^{n}(1-\lambda)w_i f_i(x) + \frac{1}{n}\sum_{i=1}^{n}\lambda(1-\lambda)w_i(f_i(\tilde{x})-f_i(x))\\
&\qquad + \frac{1}{n}\sum_{i=1}^{n}\lambda\tilde{w}_i f_i(\tilde{w}_i) + \frac{1}{n}\sum_{i=1}^{n}\lambda(1-\lambda)\tilde{w}_i(f_i(x)-f_i(\tilde{x}))\\
&= (1-\lambda)F(w,x)+\lambda F(\tilde{w},\tilde{x}) + \frac{1}{n}\lambda(1-\lambda)\sum_{i=1}^{n}(\tilde{w}_i-w_i)(f_i(x)-f_i(\tilde{x}))\\
&\le (1-\lambda)F(w,x)+\lambda F(\tilde{w},\tilde{x}) + \frac{\lambda(1-\lambda)L}{2}\|w-\tilde{w}\|^2 + \frac{\lambda(1-\lambda)L}{2}\|x-\tilde{x}\|^2,
\end{aligned}
$$

as desired. □

Appendix B. Weak convexity of robust phase retrieval.

LEMMA B.1. *The robust phase retrieval loss defined in* (20) *is* 2*-weakly convex.*

Proof. For all $x, y, a \in \mathbb{R}^d$, we have

$$\langle a, \lambda x + (1-\lambda)y\rangle^2 = \lambda\langle a, x\rangle^2 + (1-\lambda)\langle a, y\rangle^2 - \lambda(1-\lambda)\langle a, y-x\rangle^2.$$

Thus, we have

$$
\begin{aligned}
&F(\lambda x + (1-\lambda)y)\\
&= \mathbb{E}_{a,\delta,\xi}\left[|\langle a, \lambda x + (1-\lambda)y\rangle^2 - (\langle a, \bar{x}\rangle^2 + \delta\cdot\xi)|\right]\\
&\le \lambda F(x) + (1-\lambda)F(y) + \lambda(1-\lambda)\mathbb{E}_a\left[\langle a, y-x\rangle^2\right]\\
&= \lambda F(x) + (1-\lambda)F(y) + \lambda(1-\lambda)\|x-y\|^2.
\end{aligned}
$$

Therefore, by Proposition 2.1, F is 2-weakly convex. □

Acknowledgments. We thank Dmitriy Drusvyatskiy, George Lan, and the two anonymous reviewers for helpful comments.

REFERENCES

[1] E. Abbe, A. S. Bandeira, A. Bracher, and A. Singer, *Decoding binary node labels from censored edge measurements: Phase transition and efficient recovery*, IEEE Trans. Netw. Sci. Eng., 1 (2014), pp. 10–22.

[2] E. Abbe, A. S. Bandeira, and G. Hall, *Exact recovery in the stochastic block model*, IEEE Trans. Inf. Theory, 62 (2016), pp. 471–487.

[3] A. Aravkin and D. Davis, *A SMART Stochastic Algorithm for Nonconvex Optimization with Applications to Robust Machine Learning*, preprint, https://arxiv.org/abs/1610.01101, 2016.

[4] J. V. Burke, *Descent methods for composite nondifferentiable optimization problems*, Math. Program., 33 (1985), pp. 260–279.

[5] J. V. Burke, *Second order necessary and sufficient conditions for convex composite NDO*, Math. Program., 38 (1987), pp. 287–302.

[6] J. V. Burke and M. C. Ferris, *A Gauss–Newton method for convex composite optimization*, Math. Program., 71 (1995), pp. 179–194.

[7] F. Clarke, *Optimization and Nonsmooth Analysis*, SIAM, Philadelphia, PA, 1990.

[8] A. Daniilidis and J. Malick, *Filling the gap between lower-c1 and lower-c2 functions*, J. Convex Anal., 12 (2005), pp. 315–329.

[9] D. Davis and D. Drusvyatskiy, *Stochastic model-based minimization of weakly convex functions*, SIAM J. Optim., 29 (2019), pp. 207–239, https://doi.org/10.1137/18M1178244.

[10] D. Davis, D. Drusvyatskiy, and C. Paquette, *The Nonsmooth Landscape of Phase Retrieval*, preprint, https://arxiv.org/abs/1711.03247, 2017.

[11] D. Drusvyatskiy, A. D. Ioffe, and A. S. Lewis, *Nonsmooth Optimization using Taylor-like Models: Error Bounds, Convergence, and Termination Criteria*, preprint, https://arxiv.org/abs/1610.03446, 2016.

[12] D. Drusvyatskiy and C. Kempton, *An Accelerated Algorithm for Minimizing Convex Compositions*, Preprint, Department of Mathematics, University of Washington, 2016; available at https://sites.math.washington.edu/~ddrusv/nonconv_paper.pdf.

[13] D. Drusvyatskiy and A. S. Lewis, *Error bounds, quadratic growth, and linear convergence of proximal methods*, Math. Oper. Res., 43 (2018), pp. 919–948, https://doi.org/10.1287/moor.2017.0889.

[14] J. Duchi and F. Ruan, *Stochastic methods for composite and weakly convex optimization problems*, SIAM J. Optim., 28 (2018), pp. 3229–3259,

[15] Y. M. Ermolév and V. I. Norkin, *Stochastic generalized gradient method for nonconvex nonsmooth stochastic optimization*, Cybernet. Systems Anal., 34 (1998), pp. 196–215.

[16] Y. M. Ermoliev and V. I. Norkin, *Solution of nonconvex nonsmooth stochastic optimization problems*, Cybernet. Systems Anal., 39 (2003), pp. 701–715.

[17] C. Fang, C. J. Li, Z. Lin, and T. Zhang, *SPIDER: Near-optimal non-convex optimization via stochastic path integrated differential estimator*, in Adv. Neural Inf. Process. Syst. 31, Curran Associates, Red Hook, NY, 2018, pp. 689–699.

[18] M. Fickus, D. Mixon, A. Nelson, and Y. Wang, *Phase retrieval from very few measurements*, Linear Algebra Appl., 449 (2014), pp. 475–499, https://doi.org/10.1016/j.laa.2014.02.011.

[19] R. Fletcher, *A Model Algorithm for Composite Nondifferentiable Optimization Problems*, Springer, Berlin, 1982, pp. 67–76.

[20] S. Ghadimi and G. Lan, *Stochastic first- and zeroth-order methods for nonconvex stochastic programming*, SIAM J. Optim., 23 (2013), pp. 2341–2368.

[21] S. Ghadimi, G. Lan, and H. Zhang, *Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization*, Math. Program., 155 (2016), pp. 267–305.

[22] W. Hare and C. Sagastizábal, *Computing proximal points of nonconvex functions*, Math. Program., 116 (2009), pp. 221–258, https://doi.org/10.1007/s10107-007-0124-6.

[23] W. Hare and C. Sagastizábal, *A redistributed proximal bundle method for nonconvex optimization*, SIAM J. Optim., 20 (2010), pp. 2442–2473, https://doi.org/10.1137/090754595.

[24] S. Lacoste-Julien, M. Schmidt, and F. Bach, *A Simpler Approach to Obtaining an $O(1/t)$ Convergence Rate for the Projected Stochastic Subgradient Method*, preprint, https://arxiv.org/abs/1212.2002, 2012.

[25] A. S. Lewis and S. J. Wright, *A proximal method for composite minimization*, Math. Program., 158 (2016), pp. 501–546.

[26] X. Lian, M. Wang, and J. Liu, *Finite-sum composition optimization via variance reduced gradient descent*, in Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Proc. Mach. Learn. Res. 54, 2017, pp. 1159–1167; available at http://proceedings.mlr.press/v54/.

[27] M. Menickelly and S. M. Wild, *Robust learning of trimmed estimators via manifold sampling*, in Modern Trends in Nonconvex Optimization for Machine Learning (ICML 2018 Workshop), 2018; available at https://sites.google.com/view/icml2018nonconvex/papers.

[28] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM J. Optim., 19 (2009), pp. 1574–1609.

[29] A. Nemirovsky and D. Yudin, *Problem Complexity and Method Efficiency in Optimization*, Wiley-Interscience Ser. Discrete Math., Interscience, New York, 1983.

[30] E. A. Nurminski, *On ε-Subgradient Methods of Non-Differentiable Optimization*, Springer, Berlin, 1979, pp. 187–195.

[31] C. Paquette, H. Lin, D. Drusvyatskiy, J. Mairal, and Z. Harchaoui, *Catalyst for gradient-based non-convex optimization*, in Proceedings of the International Conference on Artificial Intelligence and Statistics, Proc. Mach. Learn. Res. 84, 2018, pp. 613–622; available at http://proceedings.mlr.press/v84/.

[32] H. Robbins and S. Monro, *A stochastic approximation method*, Ann. Math. Statist., 22 (1951), pp. 400–407.

[33] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.

[34] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 2015.

[35] R. T. Rockafellar and R. J. B. Wets, *Variational Analysis*, Grundlehren Math. Wiss. 317, Springer, Berlin, 1998.

[36] P. J. Rousseeuw, *Multivariate Estimation with High Breakdown Point*, Math. Statist. Appl.,

8 (1985), pp. 283–297.

[37] A. Ruszczynski, *A linearization method for nonsmooth stochastic programming problems*, Math. Oper. Res., 12 (1987), pp. 32–49.

[38] M. Wang, E. X. Fang, and H. Liu, *Stochastic compositional gradient descent: Algorithms for minimizing compositions of expected-value functions*, Math. Program., 161 (2017), pp. 419–449.

[39] M. Wang, J. Liu, and E. Fang, *Accelerating stochastic composition optimization*, in Adv. Neural Inf. Process. Syst. 29, Curran Associates, Red Hook, NY, 2016, pp. 1714–1722.

[40] Y. Xu and W. Yin, *Block stochastic gradient iteration for convex and nonconvex optimization*, SIAM J. Optim., 25 (2015), pp. 1686–1716.