MATHEMATICS OF OPERATIONS RESEARCH
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0364-765X | EISSN 1526-5471 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000

Proximally Guided Stochastic Subgradient Method for Nonsmooth, Nonconvex Problems

Damek Davis
Cornell University, dsd95@cornell.edu, https://people.orie.cornell.edu/dsd95

Benjamin Grimmer
Cornell University, bdg79@cornell.edu

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 validates the use of stochastic subgradient methods in nonsmooth, nonconvex optimization as is common when optimizing neural networks with ReLu activation units.

Key words: Nonsmooth, Nonconvex, Subgradient, Stochastic, Proximal
MSC2000 subject classification: Primary: 65K05,65K10,90C26,90C15,90C30
OR/MS subject classification: Primary: Analysis of Algorithms/Computational complexity, Programming/Nonlinear/Algorithms, Programming/Nondifferentiable,Programming/Stochastic

1. Introduction

In this paper we propose a stochastic subgradient method for finding stationary points of the the following nonsmooth, nonconvex problem

$$\text{minimize}_{x\in X}\ \phi(x) := \mathbb{E}_\xi\left[\phi^\xi(x)\right], \tag{1}$$

where $X \subseteq \mathbb{R}^d$ is a closed convex set and $\phi^\xi : \mathbb{R}^d \to \mathbb{R}$ is a closed function which is ρ-weakly convex on X (i.e., $\phi^\xi + \frac{\rho}{2}\|\cdot\|^2$ is convex) and L-Lipschitz continuous on an convex open set U containing X.[1] The property of ρ-weak convexity is fairly pervasive as the following examples illustrate:

Example: Additive Composite. Suppose that

$$\phi^\xi = f^\xi(x) + g^\xi(x), \tag{2}$$

where the $g^\xi : \mathbb{R}^d \to \mathbb{R}$ is closed convex and $f^\xi(x)$ is C^1 and ∇f^ξ is β-Lipschitz continuous. Then ϕ^ξ is β-weakly convex (see Lemma 2).

[1] For the moment, we make these relatively simple assumptions. See Section 3 for the more general setting.

Example: Convex Composite. Suppose that

$$\phi^\xi = h^\xi \circ c^\xi, \tag{3}$$

where $h^\xi : \mathbb{R}^m \to \mathbb{R}$ is closed convex and L_h Lipschitz continuous and the nonlinear mapping $c^\xi : \mathbb{R}^d \to \mathbb{R}^m$ is C^1 and ∇c^ξ is β-Lipschitz continuous. Then ϕ^ξ is $L_h\beta$-weakly convex (see Lemma 3).

In this paper, we introduce a stochastic projected subgradient method for finding stationary points of (1). The full scheme is presented in Algorithm 1, which depends on a sequence of iteration counters j_t and step sizes α_j. In our convergence analysis, we use a particular selection of these values stated in equations (6) and (7). A more general version of Algorithm 1 is given in Section 3, in which the subgradients of ϕ^ξ are replaced by a suitable stochastic oracle.

Algorithm 1 PGSG: Proximally Guided Stochastic Subgradient Method

Input: $x_0 \in \mathbb{R}^d, \gamma \in (0, 1/\rho), \{j_t\}_{t\in\mathbb{N}} \subseteq \mathbb{N}, \{\alpha_j\}_{j\in\mathbb{N}} \subseteq \mathbb{R}_{++}$

for $t = 0, \ldots, T-1$ **do**
  Set $y = x_t, w = x_t$
  for $j = 0, \ldots, j_t - 2$ **do**
    Sample ξ and choose $\zeta^\xi \in \partial(\phi^\xi + \frac{1}{2\gamma}\|\cdot - x_t\|^2)(y)$ ▷ convex subdifferential
    $y \leftarrow P_X(y - \alpha_j \zeta^\xi)$
    $w \leftarrow \frac{1}{j+2}((j+1)w + y)$ ▷ running average computation
  end for
  $x_{t+1} \leftarrow w$
end for

At a high-level the algorithm is a combination of two well-known methods (see Algorithm 2): (1) an outer loop, governed by the proximal point algorithm [22], and (2) an inner loop wherein the proximal subproblem is inexactly solved through a stochastic projected subgradient method. An exact implementation of the proximal point algorithm constructs the following sequence, starting from an initial point $z_0 \in \mathbb{R}^d$

$$z_{t+1} \in \arg\min_{x\in X} \left\{ \phi(x) + \frac{1}{2\gamma}\|x - z_t\|^2 \right\}.$$

In general, computing z_t exactly is intractable. Thus, we choose to solve each such proximal subproblem inexactly.

Algorithm 2 High-level Interpretation of PGSG

Input: $x_0 \in \mathbb{R}^d, \gamma \in (0, 1/\rho), \{j_t\}_{t\in\mathbb{N}} \subseteq \mathbb{N}$

for $t = 0, \ldots, T-1$ **do**
  Let x_{t+1} = result of applying j_t iterations of the stochastic projected subgradient method to the strongly convex problem $\min_{x\in X} \left\{ \phi(x) + \frac{1}{2\gamma}\|x - x_t\|^2 \right\}$.
end for

An inexact solution of the proximal subproblem may obtained through a stochastic projected subgradient method. Key to this observation is that ρ-weak convexity of ϕ^ξ implies that the function $\phi^\xi + \frac{1}{2\gamma}\|\cdot - z_k\|^2$ is *strongly convex* for all sufficiently small γ. Thus, we can exploit known results in convex optimization which stipulate that for strongly convex problems, the stochastic projected subgradient method achieves ε accuracy after $O(1/\varepsilon)$ iterations. A simple bookkeeping exercise

would then seem to yield an overall complexity statement for Algorithm 1. However, we will see that standard results on subgradient methods cannot be directly applied to the subproblem because the function $\phi^\xi(\cdot)+\frac{1}{2\gamma}\|\cdot-x_t\|^2$ is no longer Lipschitz continuous unless X is assumed to be *bounded*.

We seek to avoid boundedness assumptions on X. In order to do so, we warm start each iteration of the stochastic projected subgradient method with the current iterate x_t, as shown in Algorithm 1. We eventually succeed in showing that such a strategy circumvents the need to bound X, which yields the **primary contribution** of this paper: an overall complexity statement for Algorithm 1.

THEOREM 1 **(Informal Statement of Convergence Theorem).** *Let $\varepsilon>0$. For all $t\in\mathbb{N}$, let j_t, respectively α_t, be chosen as in (6), respectively (7). Then after $T=O(1/\varepsilon)$ outer iterations of Algorithm 1, there exists a point $z\in\mathbb{R}^d$ and a $t\in[T]$ such that*

$$\frac{1}{\gamma}(x_t-z)\in\partial_F\left(\phi+\delta_X\right)(z) \qquad \text{and} \qquad \mathbb{E}_\xi\left[\left\|\frac{1}{\gamma}(x_t-z)\right\|^2\right]=\widetilde{O}\left(\varepsilon\right),$$

where δ_X is the $\{0,\infty\}$-valued indicator function of X. Moreover, x_t is constructed using just $\widetilde{O}(1/\varepsilon^2)$ subgradient evaluations.

This theorem appears to be the first convergence rate analysis of a stochastic subgradient method for nonsmooth, nonconvex problems. See Theorem 2 for the formal statement of the result. Notice that the proof gives the existence of a nearby point z that is nearly stationary, which is inline with standard methods for measuring stationarity in nonsmooth nonconvex problems [10, 8]. We use $\widetilde{O}$ to hide a $\log^2(1/\varepsilon)$ term. At least one power of these logs can be removed by modifying the simple argument presented in [16]. Our choice of measuring the square in Theorem 1 is, perhaps, slightly misleading. We choose this measure for easy comparison with seminal results in the existing literature [13, 14, 24].

1.1. Two Instantiations of the Proposed Algorithm It is informative to instantiate Algorithm 1 for the additive and convex composite problem classes. Since we already know the weak convexity parameter of each class, to apply Algorithm 1, we need only compute the *convex subdifferentials* of $\phi^\xi(\cdot)+\frac{1}{2\gamma}\|\cdot-x_t\|^2$. To that end, we show in Section 2.3 the following two identities for problems (2) and (3): for all $x,y\in\mathbb{R}^d$, we have

$$\partial\left(f^\xi+g^\xi+\frac{1}{2\gamma}\|\cdot-x\|^2\right)(y)=\nabla f^\xi(y)+\partial g^\xi(y)+\frac{1}{\gamma}(y-x);$$
$$\partial\left(h^\xi\circ c^\xi(\cdot)+\frac{1}{2\gamma}\|\cdot-x\|^2\right)(y)=\nabla c^\xi(y)^T\partial h^\xi(c^\xi(y))+\frac{1}{\gamma}(y-x),$$

where ∂ denotes the convex subdifferentiation operator [2].

1.2. Related Work

Stochastic Gradient Methods. Surprisingly, the rate presented in the above informal theorem matches known rates for smooth stochastic gradient methods in nonconvex optimization [13]. There, the standard stochastic gradient method may be used without modification.

Stochastic Proximal-Gradient Methods. For additive composite problems of the form

$$\text{minimize } \left\{\mathbb{E}_\xi\left[f^\xi(x)\right]+g(x)\right\},$$

all known stochastic schemes are stochastic proximal-gradient methods, which require, at every iteration, an evaluation of the mapping $\mathbf{prox}_g(y)=\arg\min\{g(x)+(1/2)\|x-y\|^2\}$, which is a costly operation in general. In the case that such proximal computations are computationally cheap,

such methods achieve the same complexity as Theorem 1, with the caveat that every iteration t requires constructing a batch I_t of gradients, with $|I_t| = \theta(1/\varepsilon)$, and forming the average $\zeta_t = (1/|I_t|)\sum_{\xi\in I_t}\nabla f^\xi(x_t)$ [14]. Such proximal gradient methods have also been extended to allow g that are arbitrary closed prox-bounded functions [24], a setting which we do not recover in the proposed method.

However, it can be the case that evaluating the proximal mapping of g is substantially more expensive than simply computing its subgradient. For example, if $g = \|\cdot\|_2$ is the spectral norm on $\mathbb{R}^{n\times n}$, then its proximal mapping is computed through a singular value decomposition which has complexity $O(n^3)$. In contrast, subgradients may be computed from a single maximal eigenvector, which has complexity $O(n^2)$.

Another advantage of the proposed approach over stochastic proximal-gradient methods, is that multiple nonsmooth functions may be present in the objective function ϕ. The same is not true for stochastic proximal-gradient methods because even if two functions g_1 and g_2 have simple proximal operators, the proximal operator of the sum $g = g_1 + g_2$ can be quite complex. Similarly, the proximal operator of an expectation $\mathbb{E}_\xi[g^\xi]$ is could be intractable.

Stochastic Methods for Convex Composite. Recently two methods have been proposed for finding stationary points of the convex composite problem in which $\phi^\xi = h^\xi \circ c^\xi$ [11]. The first proposed method adapts the Prox-Linear algorithm [4, 5, 8, 3, 10, 17, 12] to the stochastic setting: given x_t, sample ξ and form x_{t+1} as the solution to the convex optimization problem:

$$x_{k+1} = \arg\min_{x\in X}\left\{h^\xi(c^\xi(x_t) + \nabla c^\xi(x_t)(x - x_t)) + \frac{1}{2\gamma_t}\|x - x_t\|^2\right\}, \tag{4}$$

where γ_t is a stepsize decreasing roughly at rate $1/\sqrt{k}$. The second proposed method is a straightforward application of the stochastic projected subgradient method [18]. The authors show that both methods almost surely converge to stationary points, but no rates of convergence are given. In this paper, we do not provide a convergence rate for the prox-linear algorithm (4), but we do provide a convergence rate for Algorithm 1, which is a slight modification of stochastic projected subgradient method analyzed in [11]. When each prox-linear subproblem is costly, stochastic subgradient methods are computationally preferrable to prox-linear methods because subgradients of $h^\xi \circ c^\xi$ are readily computed $\partial(h^\xi \circ c^\xi)(x) := \nabla c^\xi(x)^T \partial h^\xi(c^\xi(x))$.

Finally, we mention that the convergence proof presented in [11] is complex, being based on the highly nontrivial theory of nonconvex subdifferential inclusions. We believe there is a benefit to having a simple proof of convergence, albeit for a slightly different algorithm, which is what we provide in this paper.

Inexact Proximal Point in Nonconvex Optimization. The idea of using inexact the proximal point method to guide a nonconvex optimization algorithm to stationary points is not new, and can be found in the recent work [21]. The results presented in this paper, however, do not follow from the catalyst framework, as they exploit linearly convergent algorithms for solving the proximal subproblems. In contrast, there are no linearly convergent stochastic subgradient algorithms capable of inexactly minimizing proximal point step.

Subgradient Methods for Weakly Convex Problems. This paper is not the first to consider subgradient methods under the weak convexity assumption. For example, the early work [20] proves subsequential convergence of the (non projected) subgradient method for weakly convex *deterministic* optimization problems. However, no rates of convergence were given in that work.

Algorithm	**Problem Class**	**Global Complexity**	**Comparable Regime**
Gradient Descent [19]	Smooth	$O(m/\varepsilon)$	$\varepsilon = \Omega(1/m)$
SAGA/SVRG [15, 1]	Add. Composite	$O(m + m^{2/3}/\varepsilon)$	$\varepsilon = \Omega(1/m^{2/3})$
Inexact Prox-Linear [9]	Convex Composite	$O(m/\varepsilon + \min\{\sqrt{m}/\varepsilon^{3/2}, 1/\varepsilon^2\})$	$\varepsilon = \Omega(1/m)$

TABLE 1. Convergence rates for existing algorithms that minimizing finite sums. The comparable regime is the space of ε for which the global complexity reduces to $O(1/\varepsilon^2)$. See Section 1.3 for more details.

1.3. Rates for Finite Sums in Low-Accuracy Regime Suppose that

$$\phi := \frac{1}{m}\sum_{i=1}^{m} \phi_i(x),$$

is a finite sum. For this problem, the rate presented in Theorem 1 is independent of m. For large m it is interesting to compare the convergence rates of the proposed method with known algorithms for minimizing finite sums. Since there are methods which take advantage of the finite sum structure, Algorithm 1 is not asymptotically better than such methods. However, in the so-called low-accuracy regimes in which $\varepsilon = \Omega(1/m^\alpha)$ (for some $\alpha > 0$), the proposed method can perform on par with existing methods for minimizing finite sums. Moreover, to achieve accuracy $O(1/\sqrt{m})$, a single pass through the dataset suffices. We stress that for extremely large-scale problems, say $m = 10^9$, $1/m^\alpha$ is extremely small. We present the results of this comparison in Table 1

2. Notation and Basic Results Most of the notation and concepts we use in this paper can be found in [23]. Our main probabilistic assumption is that we work in a probability space $(\Omega, \mathcal{F}, P)$ and $\mathbb{R}^d$ is equipped with the Borel σ-algebra.

2.1. Subdifferentials and Normal Cones We denote $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$. Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be a proper closed function. We use three different subdifferentials in this paper. First we let

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

denote the *convex subdifferential* of f. Second we let

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

denote the *Fréchet subdifferential* of f. Third, we let

$$\partial_L f(x) = \{v \in \mathbb{R}^d \mid (\exists \{x_i\}, \{v_i\}) \text{ such that } v_i \in \partial_F f(x_i) \text{ and } (x_i, v_i, f(x_i)) \to (x, v, f(x))\},$$

denote the *limiting subdifferential* of f.

DEFINITION 1 (FIST-ORDER STATIONARY CONDITION). Let ϕ and X be as in the Problem (1). A point $x \in X$ is first-order stationary if

$$0 \in \partial_F \left(\phi(x) + \delta_X\right)(x),$$

where δ_X is the $\{0, \infty\}$-valued indicator function of X.
The above is a necessary condition for optimality Problem (1).

2.2. General Weakly Convex Functions The following calculation, which is based on [7, Theorem 3.1] and standard results in variational analysis, enables us to compute the subdifferentials of ϕ^ξ.

PROPOSITION 1 **(Subgradients of weakly convex functions).** *Suppose that f is ρ-weakly convex and Locally Lipschitz on a convex open set U. Then for all $x \in U$, we have*

$$\partial\left(f + \frac{\rho}{2}\|\cdot\|^2\right)(x) = \partial_F f(x) + \rho x = \partial_L f(x) + \rho x$$

In particular, f is Clarke regular [6, Definition 2.3.4] on U, i.e., $\partial_F f(x)$ and $\partial_L f(x)$ coincide and are nonempty for all $x \in U$. Moreover, the following are equivalent

1. *f is ρ-weakly convex on U. That is, $f + \frac{\rho}{2}\|x\|^2$ is convex on U.*
2. *For any $x, y \in U$ with $v \in \partial_F f(x)$, we have*

$$f(y) \geq f(x) + \langle v, y - x\rangle - \frac{\rho}{2}\|y - x\|^2. \tag{5}$$

3. *For all $x, y \in U$ and $\alpha \in [0,1]$, we have*

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

The following result is trivial, but useful.

LEMMA 1. *Let $U \subseteq \mathbb{R}^d$ be an open convex set. Suppose that f_1, respectively, f_2 is ρ_1, respectively, ρ_2, weakly convex on U. Then $f_1 + f_2$ is $\rho_1 + \rho_2$ weakly convex.*

We now present a continuous analogue of this fact.

PROPOSITION 2 **(Expectations of Weakly Convex Functions).** *Assume the setting of problem 1. Then ϕ is ρ weakly convex.*

Proof. Each $\phi^\xi + \frac{\rho}{2}\|x\|^2$ is convex on U so $\phi(x) + \frac{\rho}{2}\|x\|^2 = \mathbb{E}_\xi\left[\phi^\xi(x) + \frac{\rho}{2}\|x\|^2\right]$ is convex, as well. □

Finally, we specialize some basic results regarding subdifferentials of expectations to the current setting.

PROPOSITION 3 **(Measurable Selections).** *Assume the setting of Problem 1 (allowing for weak convexity to be replaced by Clarke regularity). Let $G : U \times \Omega \to \mathbb{R}^d$ be a measurable selection of $\partial_F \phi^\xi$, i.e., for all $(x, \xi) \in U \times \Omega$, we have $G(x;\xi) \in \partial_F \phi^\xi(x)$. Then for all $x \in U$, we have*

$$\mathbb{E}_\xi\left[G(x;\xi)\right] \in \partial_F \phi(x).$$

Proof. Given that each ϕ^ξ is Clarke regular, this follows immediately from [6, Definition 2.7.2]. □

2.3. Two Classes of Weakly Convex Functions The next lemma describes the weak convexity parameter and subdifferential formula for additive composite functions.

LEMMA 2 **(Weak Convexity and Subdifferentials of Additive Composite Functions).** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be C^1 and let ∇f be β-Lipschitz continuous. Let $g : \mathbb{R}^d \to C^1$ be closed, proper, and convex and Lipschitz continuous. Then*

1. *The sum $f + g$ is β-weakly convex.*
2. *For all $x, y \in \mathbb{R}^d$ and $\gamma \in [0, 1/\beta]$, we have*

$$\partial\left(f + g + \frac{1}{2\gamma}\|\cdot - x\|^2\right)(y) = \nabla f(y) + \partial g(y) + \frac{1}{\gamma}(y - x).$$

Proof. Part 1. By Lemma 1 it suffices to prove that f is β-weakly convex. This follows immediately by a rearrangement of the descent lemma [19]:

$$\left(\forall x, y \in \mathbb{R}^d\right) \quad f^\xi(x) \le f^\xi(y) + \langle \nabla f^\xi(x), x - y\rangle + \frac{\beta}{2}\|x-y\|^2.$$

Part 2. For convex functions the Fréchet and convex subdifferentials agree. Moreover, because f is C^1 and g is Lipschitz continuous, the sum rule for differentiation holds. Therefore,

$$\partial\left(f + g + \frac{1}{2\gamma}\|\cdot - x\|^2\right)(y) = \partial_F\left(f + g + \frac{1}{2\gamma}\|\cdot - x\|^2\right)(y) = \nabla f(y) + \partial g(y) + \frac{1}{\gamma}(y-x). \qquad \square$$

The next lemma describes the weak convexity parameter and subdifferential formula for convex composite functions.

LEMMA 3 **(Weak Convexity and Subdifferentials of Convex Composite Functions)**. *Let $h : \mathbb{R}^d \to \mathbb{R}$ be closed, convex, and L_h-Lipschitz continuous. Let $c : \mathbb{R}^d \to \mathbb{R}^m$ be C^1 and let ∇c be β-Lipschitz continuous. Then*

1. *The function $h \circ c$ is $L_h\beta$-weakly convex.*
2. *For all $x, y \in \mathbb{R}^d$ and $\gamma \in [0, 1/(L_h\beta)]$, we have*

$$\partial\left(h \circ c + \frac{1}{2\gamma}\|\cdot - x\|^2\right)(y) = \nabla c(y)^T \partial h(c(y)) + \frac{1}{\gamma}(y-x).$$

Proof. Part 1. See [9].

Part 2. For convex functions, the Fréchet and convex subdifferentials agree. As h is Lipschitz and $h \circ c$ is locally Lipschitz at y, the sum and chain rules hold:

$$\begin{aligned}\partial\left(h \circ c + \frac{1}{2\gamma}\|\cdot - x\|^2\right)(y) = \partial_F\left(h \circ c + \frac{1}{2\gamma}\|\cdot - x\|^2\right)(y) &= \partial_F(h\circ c)(y) + \frac{1}{\gamma}(y-x) \\ &= \nabla c(y)^T \partial h(c(y)) + \frac{1}{\gamma}(y-x). \qquad \square\end{aligned}$$

See [7] for more examples of weakly convex functions.

3. Convergence Theory In this section we develop the convergence rate presented in Theorem 1. We need the following basic assumptions.

(A1) It is possible to generate i.i.d. samples $\xi_1, \xi_2, \ldots$ of realizations of ξ.
(A2) There exists a measurable mapping $G : U \times \Omega \to \mathbb{R}^d$ such that $\mathbb{E}_\xi[G(x,\xi)] \in \partial_F \phi(x)$.
(A3) There exists $L \ge 0$ such that for all $x \in X$, we have $\mathbb{E}_\xi\left[\|G(x,\xi)\|^2\right] \le L^2$.
(A4) The function ϕ is ρ-weakly convex on X.

On the Generality of (A1)-(A4) The above assumptions are more general than those listed in the statement of Problem 1. Notice that ρ-weak convexity of ϕ^ξ is a sufficient, but not necessary condition to guarantee weak convexity of ϕ. By Proposition 3, any measurable selection G of $\partial_F \phi^\xi$ satisfies (A2), as long as all ϕ^ξ are Clarke-regular. In that case, if each ϕ^ξ is L-Lipschitz continuous, (A3) is automatically satisfied.

In formulating proofs of the main statements, it is useful to have global and local iteration indices on each variable. We reformulate Algorithm 1 in terms of these explicit iteration counts and our subgradient oracle in Algorithm 3.

Algorithm 3 PGSG: Proximally Guided Stochastic Subgradient Method

Input: $x_0 \in \mathbb{R}^d, \rho > 0, \gamma \in (0, 1/\rho), j_t$ as in (6), α_j as in (7)

for $t = 0, \ldots, T-1$ **do**
  Set $y_{t,0} = x_t$
  for $j = 0, \ldots, j_t - 2$ **do**
    Sample $\xi_{t,j}$ and set $\zeta^{\xi_{t,j}} = G(y_{t,j}, \xi_{t,j}) + \frac{1}{\gamma}(y_{t,j} - x_t)$
    $y_{t,j+1} \leftarrow P_X(y_{t,j} - \alpha_j \zeta^{\xi_{t,j}})$
  end for
  $x_{t+1} \leftarrow \frac{1}{j_t} \sum_{j=0}^{j_t - 1} y_{t,j}$
end for

To fully specify our algorithm, we need to specify the sequence of inner iterations counts j_t and the sequence of step sizes used in the inner iteration α_j. Given the user specified $\gamma \in (0, 1/\rho)$, we choose the following values for these sequences:

$$j_t := t + \max\left\{1, \left\lceil \left(\frac{4}{1-\gamma\rho} + 1\right) \frac{36}{1-\rho\gamma} \log\left(\left(\frac{4}{1-\gamma\rho} + 1\right) \frac{36}{1-\rho\gamma}\right) \right\rceil \right\} = \theta(t) \tag{6}$$

$$\alpha_j := \frac{2}{(\gamma^{-1} - \rho)(j+1) + 12/(\gamma - \gamma^2\rho)} = \theta(j^{-1}). \tag{7}$$

With these choices in hand, we can now state our convergence theorem: Let the sample history of our algorithm be denoted by $\Xi = (\xi_0, \ldots, \xi_T)$, where $\xi_t = (\xi_{t,0}, \ldots, \xi_{t,j_t})$ is the sample history of the tth inner iteration. For all $t \in \mathbb{N}$, define

$$z_{t+1} = \arg\min_{x \in X} \{\phi(x) + \frac{1}{2\gamma}\|x - x_t\|^2\}. \tag{8}$$

Then we can give the following convergence guarantee to a stationary point in expectation.

THEOREM 2 **(Convergence Theorem)**. *By outer iteration $T \in \mathbb{N}$, some x_t with $0 \le t \le T-1$ satisfies*

$$\mathbb{E}_\Xi\left[\|z_{t+1} - x_t\|^2\right] \le \frac{6\gamma}{T}\left(\phi(x_0) - \mathbb{E}_\Xi[\phi(x_T)] + \left(\frac{4}{1-\gamma\rho} + 1\right)\frac{3L^2}{\gamma^{-1} - \rho}\log^2(T)\right)$$

and

$$\mathbb{E}_\Xi\left[\mathbf{dist}(\partial_F(\phi + \delta_X)(z_{t+1}), 0)^2\right] \le \frac{6}{T\gamma}\left(\phi(x_0) - \mathbb{E}_\Xi[\phi(x_T)] + \left(\frac{4}{1-\gamma\rho} + 1\right)\frac{3L^2}{\gamma^{-1} - \rho}\log^2(T)\right).$$

Thus after $T = \tilde{O}(\epsilon^{-1})$ outer iterations, an expected ϵ-stationary point will be found. The total number of gradient evaluations required to compute this point is bounded by

$$\sum_{t=0}^{T-1} j_t = \frac{T(T-1)}{2} + \max\left\{1, \left\lceil \left(\frac{4}{1-\gamma\rho} + 1\right) \frac{36}{1-\rho\gamma} \log\left(\left(\frac{4}{1-\gamma\rho} + 1\right) \frac{36}{1-\rho\gamma}\right) \right\rceil \right\} T = \tilde{O}(\epsilon^{-2}).$$

To prove this theorem, we first analyze the convergence of the inner loop of our algorithm for any $t \ge 0$. The inner iteration can be viewed as applying a stochastic projected subgradient method to the function $\phi(x) + \frac{1}{2\gamma}\|x - x_t\|^2$ with x_t as its initial iterate. For convenience, we denote this regularized function in the tth outer iteration by $f_t(x) := \phi(x) + \frac{1}{2\gamma}\|x - x_t\|^2$. Then the inner loop has the following guarantee, where z_{t+1} is the unique minimizer of $f_t(x)$ over X.

PROPOSITION 4 **(Inner Iteration Convergence).** *For any $t \geq 0$,*

$$\mathbb{E}_{\xi_t}\left[f_t\left(\frac{1}{j_t}\sum_{j=0}^{j_t-1} y_{t,j}\right) - f_t(z_{t+1}) \mid \xi_0, \ldots, \xi_{t-1}\right] \leq \frac{3L^2}{\gamma^{-1}-\rho}\frac{\log(j_t+1)}{j_t} + \left(\frac{3}{\gamma-\rho\gamma^2}\frac{\log(j_t+1)}{j_t} + \frac{3}{\gamma-\rho\gamma^2}\frac{1}{j_t}\right)\|x_t - z_{t+1}\|^2.$$

Further,

$$\mathbb{E}_{\xi_t}\left[\left\|\frac{1}{j_t}\sum_{j=0}^{j_t-1} y_{t,j} - z_{t+1}\right\|^2 \mid \xi_0, \ldots, \xi_{t-1}\right] \leq \frac{6L^2}{(\gamma^{-1}-\rho)^2}\frac{\log(j_t+1)}{j_t} + \left(\frac{6}{\gamma^2(\gamma^{-1}-\rho)^2}\frac{\log(j_t+1)}{j_t} + \frac{6}{\gamma^2(\gamma^{-1}-\rho)^2}\frac{1}{j_t}\right)\|x_t - z_{t+1}\|^2. \tag{9}$$

Proof. Recall that the weak convexity of ϕ^ξ implies that f_t is $\mu := \gamma^{-1} - \rho$ strongly convex. From the nonexpansiveness of P_X, every $j \in \mathbb{N}$ satisfies

$$\begin{aligned}\|y_{t,j+1} - z_{t+1}\|^2 &\leq \|y_{t,j} - \alpha_j \zeta^{\xi_{t,j}} - z_{t+1}\|^2 \\ &\leq \|y_{t,j} - z_{t+1}\|^2 + \alpha_j^2\|\zeta^{\xi_{t,j}}\|^2 - 2\alpha_j\langle y_{t,j} - z_{t+1}, \zeta^{\xi_{t,j}}\rangle.\end{aligned}$$

Taking the expectation with the inner iteration sample history ξ_t yields

$$\begin{aligned}\mathbb{E}_{\xi_t}\|y_{t,j+1} - z_{t+1}\|^2 &\leq \mathbb{E}_{\xi_t}\|y_{t,j} - z_{t+1}\|^2 + \alpha_j^2\mathbb{E}_{\xi_t}\|\zeta^{\xi_{t,j}}\|^2 - 2\alpha_j\langle y_{t,j} - z_{t+1}, \mathbb{E}_{\xi_t}\zeta^{\xi_{t,j}}\rangle \\ &\leq \mathbb{E}_{\xi_t}\|y_{t,j} - z_{t+1}\|^2 + \alpha_j^2\mathbb{E}_{\xi_t}\|\zeta^{\xi_{t,j}}\|^2 - 2\alpha_j\left(\mathbb{E}_{\xi_t} f_t(y_{t,j}) - f_t(z_{t+1}) + \frac{\mu}{2}\mathbb{E}_{\xi_t}\|y_{t,j} - z_{t+1}\|^2\right),\end{aligned}$$

where the second inequality uses the strong convexity of f_t. From the triangle inequality and Lipschitz continuity of ϕ, we know the following

$$\begin{aligned}\mathbb{E}_{\xi_t}\|\zeta^{\xi_{t,j}}\|^2 &= \mathbb{E}_{\xi_t}\|G(y_{t,j}, \xi_{t,j}) + \frac{1}{\gamma}(y_{t,j} - x_t)\|^2 \\ &= \mathbb{E}_{\xi_t}\|G(y_{t,j}, \xi_{t,j}) + \frac{1}{\gamma}(y_{t,j} - z_{t+1} + z_{t+1} - y_0)\|^2 \\ &\leq 3L^2 + \mathbb{E}_{\xi_t}\frac{3}{\gamma^2}\|y_{t,j} - z_{t+1}\|^2 + \frac{3}{\gamma^2}\|x_t - z_{t+1}\|^2.\end{aligned}$$

Combining this with our previous inequality yields

$$2\alpha_j\left(\mathbb{E}_{\xi_t} f_t(y_{t,j}) - f_t(z_{t+1})\right) \leq 3L^2\alpha_j^2 + \frac{3\alpha_j^2}{\gamma^2}\|x_t - z_{t+1}\|^2 + \left(1 - \mu\alpha_j + \frac{3\alpha_j^2}{\gamma^2}\right)\mathbb{E}_{\xi_t}\|y_{t,j} - z_{t+1}\|^2 - \mathbb{E}_{\xi_t}\|y_{t,j+1} - z_{t+1}\|^2.$$

Since $\alpha_j \leq (\gamma - \gamma^2\rho)/6$, (see Lemma 4), we have

$$2\alpha_j\left(\mathbb{E}_{\xi_t} f_t(y_{t,j}) - f_t(z_{t+1})\right) \leq 3L^2\alpha_j^2 + \frac{3\alpha_j^2}{\gamma^2}\|x_t - z_{t+1}\|^2 + (1 - \mu\alpha_j/2)\mathbb{E}_{\xi_t}\|y_{t,j} - z_{t+1}\|^2 - \mathbb{E}_{\xi_t}\|y_{t,j+1} - z_{t+1}\|^2.$$

Thus

$$\mathbb{E}_{\xi_t} f_t(y_{t,j}) - f_t(z_{t+1}) \leq \frac{3L^2\alpha_j}{2} + \frac{3\alpha_j}{2\gamma^2}\|x_t - z_{t+1}\|^2 + \frac{\alpha_j^{-1} - \mu/2}{2}\mathbb{E}_{\xi_t}\|y_{t,j} - z_{t+1}\|^2 - \frac{\alpha_j^{-1}}{2}\mathbb{E}_{\xi_t}\|y_{t,j+1} - z_{t+1}\|^2.$$

Plugging in our choice of α_k yields

$$\mathbb{E}_{\xi_t} f_t(y_{t,j}) - f_t(z_{t+1}) \leq \frac{3L^2}{\mu(j+1)+12/\gamma^2\mu} + \frac{3}{\gamma^2\mu(j+1)+12/\mu}\|x_t - z_{t+1}\|^2 + \frac{\mu j + 12/\gamma^2\mu}{4}\mathbb{E}_{\xi_t}\|y_{t,j} - z_{t+1}\|^2 - \frac{\mu(j+1)+12/\gamma^2\mu}{4}\mathbb{E}_{\xi_t}\|y_{t,j+1} - z_{t+1}\|^2.$$

Averaging this inequality over $j = 0$ to $j = j_t - 1$ gives

$$\mathbb{E}_{\xi_t} f_t\left(\frac{1}{j_t}\sum_{j=0}^{j_t-1} y_{t,j}\right) - f_t(z_{t+1}) \leq \frac{1}{j_t}\sum_{j=0}^{j_t-1}\frac{3L^2}{\mu(j+1)+12/\gamma^2\mu} + \frac{1}{j_t}\sum_{j=0}^{j_t-1}\frac{3}{\gamma^2\mu(j+1)+12/\mu}\|x_t - z_{t+1}\|^2 + \frac{1}{j_t}\frac{3}{\gamma^2\mu}\|x_t - z_{t+1}\|^2.$$

Then using a Reimann sum approximation, we arrive at the bound of

$$\mathbb{E}_{\xi_t} f_t\left(\frac{1}{j_t}\sum_{j=0}^{j_t-1} y_{t,j}\right) - f_t(z_{t+1}) \leq \frac{3L^2}{\mu}\frac{\log(j_t+1)}{j_t} + \left(\frac{3}{\gamma^2\mu}\frac{\log(j_t+1)}{j_t} + \frac{3}{\gamma^2\mu}\frac{1}{j_t}\right)\|x_t - z_{t+1}\|^2.$$

Recalling that $\mu = \gamma^{-1} - \rho$, we arrive at our claimed bound. Our distance bound follows as a direct consequence of the strong convexity of f_t. □

Proposition 4 quantifies how our inner iteration computes an inexact solution to the proximal subproblem. From this understanding of the inner iteration, we can directly give a convergence bound for the entire algorithm, proving our main theorem.

Proof of Theorem 2. The definition of the proximal mapping implies

$$\phi(z_{t+1}) \leq \phi(x_t) - \frac{1}{2\gamma}\|z_{t+1} - x_t\|^2. \tag{10}$$

Observe that

$$\begin{aligned}\|z_{t+1} - x_t\|^2 - \|x_{t+1} - x_t\|^2 &= (\|z_{t+1} - x_t\| + \|x_{t+1} - x_t\|)(\|z_{t+1} - x_t\| - \|x_{t+1} - x_t\|)\\ &\leq (2\|z_{t+1} - x_t\| + \|x_{t+1} - z_{t+1}\|)\|x_{t+1} - z_{t+1}\|\\ &= 2\|z_{t+1} - x_t\|\|x_{t+1} - z_{t+1}\| + \|x_{t+1} - z_{t+1}\|^2\\ &\leq \frac{1}{3}\|z_{t+1} - x_t\|^2 + 4\|x_{t+1} - z_{t+1}\|^2.\end{aligned}$$

For convenience, let $E_t = \frac{3L^2}{\mu}\frac{\log(j_t+1)}{j_t} + \left(\frac{3}{\gamma^2\mu}\frac{\log(j_t+1)}{j_t} + \frac{3}{\gamma^2\mu}\frac{1}{j_t}\right)\|x_t - z_{t+1}\|^2$ denote the inexactness bound derived in Proposition 4. Then we know

$$\begin{aligned}\mathbb{E}_{\xi_t}\phi(x_{t+1}) &\leq \phi(z_{t+1}) + \frac{1}{2\gamma}\left(\frac{1}{3}\|z_{t+1} - x_t\|^2 + 4\mathbb{E}_{\xi_t}\|x_{t+1} - z_{t+1}\|^2\right) + E_t\\ &\leq \phi(z_{t+1}) + \frac{1}{6\gamma}\|z_{t+1} - x_t\|^2 + \frac{4}{1-\gamma\rho}E_t + E_t,\end{aligned}$$

where we used (9) in the last inequality. Combining this result with (10) implies

$$\mathbb{E}_{\xi_t}\phi(x_{t+1}) \leq \phi(x_t) - \frac{1}{2\gamma}\|z_{t+1} - x_t\|^2 + \frac{1}{6\gamma}\|z_{t+1} - x_t\|^2 + \left(\frac{4}{1-\gamma\rho} + 1\right)E_t.$$

Plugging in the value of E_t, we have

$$\mathbb{E}_{\xi_t}\phi(x_{t+1}) \leq \phi(x_t) - \left(\frac{1}{3\gamma} - \left(\frac{4}{1-\gamma\rho}+1\right)\left(\frac{3}{\gamma-\rho\gamma^2}\frac{\log(j_t+1)}{j_t} + \frac{3}{\gamma-\rho\gamma^2}\frac{1}{j_t}\right)\right)\|z_{t+1} - x_t\|^2 + \left(\frac{4}{1-\gamma\rho}+1\right)\frac{3L^2}{\gamma^{-1}-\rho}\frac{\log(j_t+1)}{j_t}.$$

We have selected j_t large enough to that the coefficient of $\|z_{t+1}-x_t\|^2$ is positive (see Lemma 4). In particular, we have

$$\mathbb{E}_{\xi_t}\phi(x_{t+1}) \le \phi(x_t) - \frac{1}{6\gamma}\|z_{t+1}-x_t\|^2 + \left(\frac{4}{1-\gamma\rho}+1\right)\frac{3L^2}{\gamma^{-1}-\rho}\frac{\log(j_t+1)}{j_t}.$$

Inductively applying this inequality from $t=0$ to $t=T-1$, we have

$$\mathbb{E}_{\Xi}\phi(x_T) \le \phi(x_0) - \sum_{t=0}^{T-1}\frac{1}{6\gamma}\mathbb{E}_{\Xi}\|z_{t+1}-x_t\|^2 + \sum_{t=0}^{T-1}\left(\frac{4}{1-\gamma\rho}+1\right)\frac{3L^2}{\gamma^{-1}-\rho}\frac{\log(j_t+1)}{j_t}.$$

Hence

$$\frac{1}{6\gamma}\sum_{t=0}^{T-1}\mathbb{E}_{\Xi}\|z_{t+1}-x_t\|^2 \le \phi(x_0) - \mathbb{E}_{\Xi}\phi(x_T) + \left(\frac{4}{1-\gamma\rho}+1\right)\frac{3L^2}{\gamma^{-1}-\rho}\sum_{t=0}^{T-1}\frac{\log(j_t+1)}{j_t}.$$

Using a Reimann sum approximation, we have

$$\frac{1}{6\gamma}\sum_{t=0}^{T-1}\mathbb{E}_{\Xi}\|z_{t+1}-x_t\|^2 \le \phi(x_0) - \mathbb{E}_{\Xi}\phi(x_T) + \left(\frac{4}{1-\gamma\rho}+1\right)\frac{3L^2}{\gamma^{-1}-\rho}\log^2(T).$$

Thus we have the desired bound of

$$\min_{0\le t\le T-1}\{\mathbb{E}_{\Xi}\|z_{t+1}-x_t\|^2\} \le \frac{6\gamma}{T}\left(\phi(x_0) - \mathbb{E}_{\Xi}\phi(x_T) + \left(\frac{4}{1-\gamma\rho}+1\right)\frac{3L^2}{\gamma^{-1}-\rho}\log^2(T)\right).$$

Finally, based on the definition of z_{t+1} in (8), we have, by the sum [23, Exercise 8.8] and Fermat [23, Theorem 10.1] rules for the Fréchet subdifferential, $(1/\gamma)(x_t - z_{t+1}) \in \partial_F(\phi+\delta_X)(z_{t+1})$, so the subdifferential distance bounds immediately follow. □

4. Conclusion In this paper we introduced a stochastic projected subgradient method for finding stationary points of expectations of weakly convex functions—a wide class of functions which includes additive and convex composite minimization problems. We provided a simple proof that, under mild assumptions, the algorithm converges at the same rate as the stochastic gradient method for smooth minimization problems. We showed that for extremely large-scale finite sum problems, the proposed stochastic subgradient method performs on par with specialized optimization methods, while, in some cases, being notably simpler to implement.

Acknowledgments. We thank Prof. Dmitriy Drusvyatskiy for early comments on this work.

Appendix A: Proofs of Auxiliary Facts.

LEMMA 4. *For any $t \ge 0$, we have*

$$\frac{1}{6\gamma} \ge \left(\frac{4}{1-\gamma\rho}+1\right)\left(\frac{3}{\gamma-\rho\gamma^2}\frac{\log(j_t+1)}{j_t} + \frac{3}{\gamma-\rho\gamma^2}\frac{1}{j_t}\right)$$

For any $j \in \mathbb{N}$, we have

$$\alpha_j \le (\gamma-\gamma^2\rho)/6.$$

Proof. Letting $a = \left(\frac{4}{1-\gamma\rho}+1\right)\frac{18}{1-\rho\gamma}$, we can state our first inequality as

$$j_t \ge a(\log(j_t)+1).$$

As a simple exercise, it can verified that this is satisfied by any $j_t \ge 2a\log(2a)$. Then, because $j_t \ge t + 2a\log(2a)$ for all $t \ge 0$ have, we can conclude this bound always holds.

Our second bound is easily verified by direct algebra:

$$\alpha_j = \frac{2}{(\gamma^{-1}-\rho)(j+1)+12/(\gamma-\gamma^2\rho)} \le \frac{2}{12/(\gamma-\gamma^2\rho)} = (\gamma-\gamma^2\rho)/6. \qquad \square$$

References

[1] Aravkin A, Davis D (2016) A smart stochastic algorithm for nonconvex optimization with applications to robust machine learning. *arXiv preprint arXiv:1610.01101* .

[2] Bauschke HH, Combettes PL (2011) *Convex Analysis and Monotone Operator Theory in Hilbert Spaces* (Springer Science & Business Media).

[3] Burke JV (1985) Descent methods for composite nondifferentiable optimization problems. *Mathematical Programming* 33(3):260–279, URL http://dx.doi.org/10.1007/BF01584377.

[4] Burke JV (1987) Second order necessary and sufficient conditions for convex composite ndo. *Mathematical Programming* 38(3):287–302, ISSN 1436-4646, URL http://dx.doi.org/10.1007/BF02592016.

[5] Burke JV, Ferris MC (1995) A gauss—newton method for convex composite optimization. *Mathematical Programming* 71(2):179–194, ISSN 1436-4646, URL http://dx.doi.org/10.1007/BF01585997.

[6] Clarke F (1990) *Optimization and Nonsmooth Analysis* (Society for Industrial and Applied Mathematics), URL http://dx.doi.org/10.1137/1.9781611971309.

[7] Daniilidis A, Malick J (2005) Filling the gap between lower-c1 and lower-c2 functions. *Journal of Convex Analysis* 12(2):315–329.

[8] Drusvyatskiy D, Ioffe AD, Lewis AS (2016) Nonsmooth optimization using taylor-like models: error bounds, convergence, and termination criteria. *arXiv preprint arXiv:1610.03446* .

[9] Drusvyatskiy D, Kempton C (2016) An accelerated algorithm for minimizing convex compositions. *arXiv preprint arXiv:1605.00125* .

[10] Drusvyatskiy D, Lewis AS (2016) Error bounds, quadratic growth, and linear convergence of proximal methods. *arXiv preprint arXiv:1602.06661* .

[11] Duchi J, Ruan F (2017) Stochastic methods for composite optimization problems. *arXiv preprint arXiv:1703.08570* .

[12] Fletcher R (1982) *A model algorithm for composite nondifferentiable optimization problems*, 67–76 (Berlin, Heidelberg: Springer Berlin Heidelberg), ISBN 978-3-642-00815-3, URL http://dx.doi.org/10.1007/BFb0120959.

[13] Ghadimi S, Lan G (2013) Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization* 23(4):2341–2368, URL http://dx.doi.org/10.1137/120880811.

[14] Ghadimi S, Lan G, Zhang H (2016) Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming* 155(1):267–305, ISSN 1436-4646, URL http://dx.doi.org/10.1007/s10107-014-0846-1.

[15] J Reddi S, Sra S, Poczos B, Smola AJ (2016) Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. Lee DD, Sugiyama M, Luxburg UV, Guyon I, Garnett R, eds., *Advances in Neural Information Processing Systems 29*, 1145–1153 (Curran Associates, Inc.), URL http://papers.nips.cc/paper/6116-proximal-stochastic-methods-for-nonsmooth-nonconvex-finite-sum-opt

[16] Lacoste-Julien S, Schmidt M, Bach F (2012) A simpler approach to obtaining an O(1/t) convergence rate for the projected stochastic subgradient method. *ArXiv e-prints* .

[17] Lewis AS, Wright SJ (2016) A proximal method for composite minimization. *Mathematical Programming* 158(1):501–546, ISSN 1436-4646, URL http://dx.doi.org/10.1007/s10107-015-0943-9.

[18] Nemirovski A, Juditsky A, Lan G, Shapiro A (2009) Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization* 19(4):1574–1609, URL http://dx.doi.org/10.1137/070704277.

[19] Nesterov Y (2013) *Introductory lectures on convex optimization: A basic course*, volume 87 (Springer Science & Business Media).

[20] Nurminski EA (1979) *On ε-subgradient methods of non-differentiable optimization*, 187–195 (Berlin, Heidelberg: Springer Berlin Heidelberg), ISBN 978-3-540-35232-7, URL http://dx.doi.org/10.1007/BFb0002654.

[21] Paquette C, Lin H, Drusvyatskiy D, Mairal J, Harchaoui Z (2017) Catalyst acceleration for gradient-based non-convex optimization. *arXiv preprint arXiv:1703.10993* .

[22] Rockafellar RT (1976) Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization* 14(5):877–898, URL http://dx.doi.org/10.1137/0314056.

[23] Rockafellar RT, Wets RJB (1998) *Variational Analysis*, volume 317 (Springer), URL http://dx.doi.org/10.1007/978-3-642-02431-3.

[24] Xu Y, Yin W (2015) Block stochastic gradient iteration for convex and nonconvex optimization. *SIAM Journal on Optimization* 25(3):1686–1716, URL http://dx.doi.org/10.1137/140983938.