Stochastic subgradient method converges at the rate $O(k^{-1/4})$ on weakly convex functions

Damek Davis[*] Dmitriy Drusvyatskiy[†]

Abstract

We prove that the proximal stochastic subgradient method, applied to a weakly convex problem, drives the gradient of the Moreau envelope to zero at the rate $O(k^{-1/4})$. As a consequence, we resolve an open question on the convergence rate of the proximal stochastic gradient method for minimizing the sum of a smooth nonconvex function and a convex proximable function.

1 Introduction

In this work, we consider the optimization problem

$$\min_{x\in\mathbb{R}^d} \ \varphi(x) := g(x) + r(x) \tag{1.1}$$

under the following assumptions on the functional components g and r. Throughout, $r\colon \mathbb{R}^d \to \mathbb{R}\cup\{\infty\}$ is a closed convex function with a computable proximal map

$$\mathrm{prox}_{\alpha r}(x) := \operatorname*{argmin}_{y} \left\{ r(y) + \tfrac{1}{2\alpha}\|y-x\|^2 \right\},$$

while $g\colon \mathbb{R}^d \to \mathbb{R}$ is a ρ-weakly convex function, meaning that the assignment $x \mapsto g(x) + \frac{\rho}{2}\|x\|^2$ is convex. The above assumptions on r are standard in the literature (see e.g. [2,16,19]), while those on g deserve some commentary. The class of weakly convex functions, first introduced in English in [17], is broad. Indeed, it includes all convex functions and smooth functions with Lipschitz continuous gradient. More generally, any function of the form $g = h\circ c$, with h convex and Lipschitz and c a smooth map with Lipschitz Jacobian [7, Lemma 4.2], is weakly convex. Classical literature highlights the importance of weak convexity in optimization [20,21,23], while recent advances in statistical learning and signal processing have further reinvigorated the problem class (1.1). For a recent discussion on the role of weak convexity in large-scale optimization, see for example [5].

[*] School of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14850, USA; `people.orie.cornell.edu/dsd95/`.

[†] Department of Mathematics, U. Washington, Seattle, WA 98195; `www.math.washington.edu/∼ddrusv`.
Research of Drusvyatskiy was supported by the AFOSR YIP award FA9550-15-1-0237 and by the NSF DMS 1651851 and CCF 1740551 awards.

The proximal subgradient method is perhaps the simplest algorithm for the problem (1.1). Given a current iterate x_t, the method repeats the steps

$$\left\{\begin{array}{l} \text{Choose } \zeta_t \in \partial g(x_t) \\ \text{Set } x_{t+1} = \mathrm{prox}_{\alpha_t r}(x_t - \alpha_t \zeta_t) \end{array}\right\},$$

where $\alpha_t > 0$ is an appropriately chosen control sequence. Here, the subdifferential ∂g is meant in a standard variational analytic sense [24, Definition 8.3]; we will recall the precise definition in Section 2. The setting when r is the indicator function of a closed convex set $\mathcal{X}$ reduces the algorithm to the classical projected subgradient method. Indeed, then the proximal map $\mathrm{prox}_{\alpha_t r}(\cdot)$ is simply the nearest point projection $\mathrm{proj}_{\mathcal{X}}(\cdot)$.

The primary goal in nonsmooth nonconvex optimization is the search for stationary points. A point $x \in \mathbb{R}^d$ is called *stationary* for the problem (1.1) if the inclusion $0 \in \partial\varphi(x)$ holds. In "primal terms", these are precisely the points where the directional derivative of φ is nonnegative in every direction [24, Proposition 8.32]:

$$\mathrm{dist}(0; \partial\varphi(x)) = -\inf_{v\colon \|v\|\le 1} \varphi'(x; v). \tag{1.2}$$

It has been known since [17, 18] that the (stochastic) subgradient method with $r = 0$ generates an iterate sequence that subsequentially converges to a stationary point of the problem. A long standing open question in this line of work is to determine the "rate of convergence" of the basic (stochastic) subgradient method and of its proximal extensions.

An immediate difficulty in addressing this question is that it is not a priori clear how to measure the progress of the algorithm. Indeed, neither the functional suboptimality gap, $\varphi(x_t) - \min\varphi$, nor the stationarity measure, $\mathrm{dist}(0; \partial\varphi(x_t))$, necessarily tend to zero along the iterate sequence. Instead, recent literature [5, 7] has identified a different measure of complexity of minimizing weakly convex functions, based on smooth approximations. The key construction we use is the *Moreau envelope*:

$$\varphi_\lambda(x) := \min_y \ \left\{\varphi(y) + \tfrac{1}{2\lambda}\|y - x\|^2\right\},$$

where $\lambda > 0$. Standard results show that as long as $\lambda < \rho^{-1}$, the envelope φ_λ is C^1-smooth with the gradient given by

$$\nabla\varphi_\lambda(x) = \lambda^{-1}(x - \mathrm{prox}_{\lambda\varphi}(x)). \tag{1.3}$$

See for example [22, Theorem 31.5]. Moreover, the norm of the gradient $\|\nabla\varphi_\lambda(x)\|$ has an intuitive interpretation in terms of near-stationarity for the target problem (1.1). Namely, the definition of the Moreau envelope directly implies that for any $x \in \mathbb{R}^d$, the proximal point $\hat{x} := \mathrm{prox}_{\lambda\varphi}(x)$ satisfies

$$\left\{\begin{array}{cl} \|\hat{x} - x\| & = \lambda\|\nabla\varphi_\lambda(x)\|, \\ \varphi(\hat{x}) & \le \varphi(x), \\ \mathrm{dist}(0; \partial\varphi(\hat{x})) & \le \|\nabla\varphi_\lambda(x)\|. \end{array}\right.$$

Thus a small gradient $\|\nabla\varphi_\lambda(x)\|$ implies that x is *near* some point $\hat{x}$ that is *nearly stationary* for (1.1). For a longer discussion of near-stationarity, see [5] or [7, Section 4.1].

In this paper, we show that under an appropriate choice of the control sequence α_t, the subgradient method will generate a point x satisfying $\|\nabla\varphi_{1/2\rho}(x)\| \leq \varepsilon$ after at most $O(\varepsilon^{-4})$ iterations. A similar guarantee was recently established for the proximally guided projected subgradient method [4]. This scheme proceeds by directly applying the gradient descent method to the Moreau envelope φ_λ, with each proximal point $\mathrm{prox}_{\lambda\varphi}(x)$ approximately evaluated by a convex subgradient method. In contrast, we show here that the basic subgradient method, without any modification or parameter tuning, already satisfies the desired convergence guarantees. This is perhaps surprising, since neither the Moreau envelope $\varphi_\lambda(\cdot)$ nor the proximal map $\mathrm{prox}_{\lambda\varphi}(\cdot)$ explicitly appear in the definition of the subgradient method.

Though our results appear to be new even in this rudimentary deterministic set up, the argument we present applies much more broadly to stochastic proximal subgradient methods, in which only stochastic estimates of ζ_t are available. This is the setting of the paper. In this regard, we improve in two fundamental ways on the results in the seminal papers [9, 10, 25]: first, we allow g to be nonsmooth and second, we do not require the variance of our stochastic estimator for ζ_t to decrease as a function of t. The second contribution removes the well-known "mini-batching" requirements common to [10, 25], while the first significantly expands the class of functions for which the rate of convergence of the stochastic proximal subgradient method is known. It is worthwhile to mention that our techniques crucially rely on convexity of r, while [25] makes no such assumption.

There is an extensive literature on stochastic subgradient methods in convex optimization, which we will not detail here; instead, we refer the interested reader to the seminal works [13, 14]. An in-depth summary of recent work for nonconvex problems appears in [4].

The outline of the paper is as follows. In the Section 2.1, we present a simplified argument for the case in which r is the indicator function of a closed convex set and the stochastic estimator has finite second moment. In this section, we also comment on improved rates in the convex setting. In Section 2.2, we prove convergence of the stochastic proximal subgradient method in full generality. In Section 2.3, we modify the results of the previous section to the case in which g is smooth and the stochastic estimator has finite variance.

2 Convergence guarantees

Henceforth, we assume that the only access to g is through a stochastic subgradient oracle. Formally, we fix a probability space $(\Omega, \mathcal{F}, P)$ and equip $\mathbb{R}^d$ with the Borel σ-algebra. We make the following three standard assumptions:

(A1) It is possible to generate i.i.d. realizations $\xi_1, \xi_2, \ldots \sim dP$.

(A2) There is an open set U containing $\mathrm{dom}\, r$ and a measurable mapping $G \colon U \times \Omega \to \mathbb{R}^d$ satisfying $\mathbb{E}_\xi[G(x,\xi)] \in \partial g(x)$ for all $x \in U$.

(A3) There is a real $L \geq 0$ such that the inequality, $\mathbb{E}_\xi\left[\|G(x,\xi)\|^2\right] \leq L^2$, holds for all $x \in \mathrm{dom}\, r$.

Some comments are in order. First, the symbol $\partial g(x)$ refers to the *subdifferential* of g at x. By definition, this is the set consisting of all vectors $v \in \mathbb{R}^d$ satisfying

$$g(y) \geq g(x) + \langle v, y - x\rangle + o(\|y - x\|) \qquad \text{as } y \to x.$$

Weak convexity automatically guarantees that subgradients of g satisfy the much stronger property [24, Theorem 12.17]:

$$g(y) \geq g(x) + \langle v, y - x \rangle - \frac{\rho}{2}\|y - x\|^2, \qquad \forall x, y \in \mathbb{R}^d,\ v \in \partial g(x). \tag{2.1}$$

One important consequence we will use is the hypo-monotonicity inequality:

$$\langle v - w, x - y \rangle \geq -\rho\|x - y\|^2, \qquad \forall x, y \in \mathbb{R}^d,\ v \in \partial g(x),\ w \in \partial g(y). \tag{2.2}$$

The three assumption (A1), (A2), (A3) are standard in the literature on stochastic subgradient methods. Indeed, assumptions (A1) and (A2) are identical to assumptions (A1) and (A2) in [13], while Assumption (A3) is the same as the assumption listed in [13, Equation (2.5)].

In this work, we investigate the efficiency of the proximal stochastic subgradient method, described in Algorithm 1.

```
Algorithm 1: Proximal stochastic subgradient method
  Data: $x_0 \in \operatorname{dom}(r)$, a sequence $\{\alpha_t\}_{t\geq 0} \subset \mathbb{R}_+$, and iteration count $T$
  Step $t = 0, \ldots, T$:
      $\left\{\begin{array}{l} \text{Sample } \xi_t \sim dP \\ \text{Set } x_{t+1} = \operatorname{prox}_{\alpha_t r}\left(x_t - \alpha_t G(x_t, \xi_t)\right) \end{array}\right\},$
  Sample $t^* \in \{0, \ldots, T\}$ according to the probability distribution $\mathbb{P}(t^* = t) = \frac{\alpha_t}{\sum_{t=0}^T \alpha_t}$.
  Return $x_{t^*}$
```

Henceforth, the symbol $\mathbb{E}_t[\cdot]$ will denote the expectation conditioned on all the realizations $\xi_0, \xi_1, \ldots, \xi_{t-1}$.

2.1 Projected stochastic subgradient method

Our analysis of Algorithm 1 is shorter and more transparent when r is the indicator function of a closed, convex set $\mathcal{X}$. This is not surprising, since projected subgradient methods are typically much easier to analyze than their proximal extensions (e.g. [3, 8]). Note that (1.1) then reduces to the constrained problem

$$\min_{x \in \mathcal{X}}\ g(x), \tag{2.3}$$

and the proximal maps $\operatorname{prox}_{\alpha r}(\cdot)$ become the nearest point projection $\operatorname{proj}_{\mathcal{X}}(\cdot)$. Thus throughout Section 2.1, we suppose that Assumptions (A1), (A2), and (A3) hold and that $r(\cdot)$ is the indicator function of a closed convex set $\mathcal{X}$. The following is the main result of this section.

Theorem 2.1 (Stochastic projected subgradient method). *Let x_{t^*} be the point returned by Algorithm 1. Then in terms of any constant $\hat{\rho} > \rho$, the estimate holds:*

$$\mathbb{E}\left[\|\nabla \varphi_{1/\hat{\rho}}(x_{t^*})\|^2\right] \leq \frac{\hat{\rho}}{\hat{\rho} - \rho} \cdot \frac{(\varphi_{1/\hat{\rho}}(x_0) - \min \varphi) + \frac{\hat{\rho}L^2}{2}\sum_{t=0}^T \alpha_t^2}{\sum_{t=0}^T \alpha_t}.$$

Proof. Let x_t denote the points generates by Algorithm 1. For each index t, define $\zeta_t := \mathbb{E}_t[G(x_t, \xi)] \in \partial g(x_t)$ and set $\hat{x}_t := \mathrm{prox}_{\varphi/\hat{\rho}}(x_t)$. We successively deduce

$$\begin{aligned}
\mathbb{E}_t\left[\varphi_{1/\hat{\rho}}(x_{t+1})\right] &\leq \mathbb{E}_t\left[g(\hat{x}_t) + \frac{\hat{\rho}}{2}\|x_{t+1} - \hat{x}_t\|^2\right] &(2.4)\\
&= g(\hat{x}_t) + \frac{\hat{\rho}}{2}\mathbb{E}_t\left[\|\mathrm{proj}_{\mathcal{X}}(x_t - \alpha_t G(x_t, \xi_t)) - \mathrm{proj}_{\mathcal{X}}(\hat{x}_t)\|^2\right]\\
&\leq g(\hat{x}_t) + \frac{\hat{\rho}}{2}\mathbb{E}_t\left[\|(x_t - \hat{x}_t) - \alpha_t G(x_t, \xi_t)\|^2\right] &(2.5)\\
&\leq g(\hat{x}_t) + \frac{\hat{\rho}}{2}\|x_t - \hat{x}_t\|^2 + \hat{\rho}\alpha_t \mathbb{E}_t\left[\langle \hat{x}_t - x_t, G(x_t, \xi_t)\rangle\right] + \frac{\alpha_t^2 \hat{\rho}}{2}L^2\\
&\leq \varphi_{1/\hat{\rho}}(x_t) + \hat{\rho}\alpha_t\langle \hat{x}_t - x_t, \zeta_t\rangle + \frac{\alpha_t^2\hat{\rho}}{2}L^2\\
&\leq \varphi_{1/\hat{\rho}}(x_t) + \hat{\rho}\alpha_t\left(g(\hat{x}_t) - g(x_t) + \frac{\rho}{2}\|x_t - \hat{x}_t\|^2\right) + \frac{\alpha_t^2\hat{\rho}}{2}L^2, &(2.6)
\end{aligned}$$

where (2.4) follows directly from the definition of the proximal map, the inequality (2.5) uses that the projection $\mathrm{proj}_{\mathcal{X}}(\cdot)$ is 1-Lipschitz, and (2.6) follows from weak convexity of g.

Using the law of total expectation to unfold this recursion yields:

$$\mathbb{E}\left[\varphi_{1/\hat{\rho}}(x_{T+1})\right] \leq \varphi_{1/\hat{\rho}}(x_0) + \frac{\hat{\rho}L^2}{2}\sum_{t=0}^{T}\alpha_t^2 - \hat{\rho}\mathbb{E}\sum_{t=0}^{T}\alpha_t\left(g(x_t) - g(\hat{x}_t) - \frac{\rho}{2}\|x_t - \hat{x}_t\|^2\right).$$

Lower-bounding the left-hand side by $\min\varphi$ and rearranging, we obtain the bound:

$$\frac{1}{\sum_{t=0}^{T}\alpha_t}\sum_{t=0}^{T}\alpha_t\mathbb{E}\left[g(x_t) - g(\hat{x}_t) - \frac{\rho}{2}\|x_t - \hat{x}_t\|^2\right] \leq \frac{(\varphi_{1/\hat{\rho}}(x_0) - \min\varphi) + \frac{\hat{\rho}L^2}{2}\sum_{t=0}^{T}\alpha_t^2}{\hat{\rho}\sum_{t=0}^{T}\alpha_t}. \tag{2.7}$$

Notice that the left-hand-side of (2.7) is precisely $\mathbb{E}\left[g(x_{t^*}) - g(\hat{x}_{t^*}) - \frac{\rho}{2}\|x_{t^*} - \hat{x}_{t^*}\|^2\right]$. Next, observe that the function $x \mapsto g(x) + \frac{\hat{\rho}}{2}\|x - x_{t^*}\|^2$ is strongly convex with parameter $\hat{\rho} - \rho$, and therefore

$$\begin{aligned}
&g(x_{t^*}) - g(\hat{x}_{t^*}) - \frac{\rho}{2}\|x_{t^*} - \hat{x}_{t^*}\|^2\\
&= \left(g(x_{t^*}) + \frac{\hat{\rho}}{2}\|x_{t^*} - x_{t^*}\|^2\right) - \left(g(\hat{x}_{t^*}) + \frac{\hat{\rho}}{2}\|x_{t^*} - \hat{x}_{t^*}\|^2\right) + \frac{\hat{\rho} - \rho}{2}\|x_{t^*} - \hat{x}_{t^*}\|^2\\
&\geq (\hat{\rho} - \rho)\|x_{t^*} - \hat{x}_{t^*}\|^2 = \frac{\hat{\rho} - \rho}{\hat{\rho}^2}\|\nabla\varphi_{1/\hat{\rho}}(x_{t^*})\|^2,
\end{aligned}$$

where the last equality follows from (1.3). Using this estimate to lower bound the left-hand-side of (2.7) completes the proof. □

In particular, using the constant stepsize α on the order of $\frac{1}{\sqrt{T+1}}$ yields the following complexity guarantee.

Corollary 2.2 (Complexity guarantee)**.** *Fix an index $T > 0$ and set the constant steplength $\alpha = \frac{\gamma}{\sqrt{T+1}}$ for some real $\gamma > 0$. Then the point x_{t^*} returned by Algorithm 1 satisfies:*

$$\mathbb{E}\left[\|\nabla \varphi_{1/(2\rho)}(x_{t^*})\|^2\right] \leq 2 \cdot \frac{\left(\varphi_{1/(2\rho)}(x_0) - \min \varphi\right) + \rho L^2 \gamma^2}{\gamma\sqrt{T+1}}. \tag{2.8}$$

Proof. This follows immediately from Theorem 2.1 by setting $\hat{\rho} = 2\rho$. □

Let us look closer at the guarantee of Corollary 2.2 by minimizing out in γ. Namely, suppose we have available some real $R > 0$ satisfying $R \geq \varphi_{1/(2\rho)}(x_0) - \min \varphi$. We deduce from (2.8) the estimate, $\mathbb{E}\left[\|\nabla \varphi_{1/(2\rho)}(x_{t^*})\|^2\right] \leq 2 \cdot \frac{R + \rho L^2\gamma^2}{\gamma\sqrt{T+1}}$. Minimizing the right-hand side in γ yields the choice $\gamma = \sqrt{\frac{R}{\rho L^2}}$ and therefore the guarantee

$$\mathbb{E}\left[\|\nabla \varphi_{1/(2\rho)}(x_{t^*})\|^2\right] \leq 4 \cdot \sqrt{\frac{\rho R L^2}{T+1}}. \tag{2.9}$$

In particular, suppose that g is L-Lipschitz and the diameter of $\mathcal{X}$ is bounded by some $D > 0$. Then we may set $R := \min\{\rho D^2, DL\}$, where the first term follows from the definition of the Moreau envelope and the second follows from Lipschitz continuity. Then the number of subgradient evaluations required to find a point x satisfying $\mathbb{E}\|\nabla \varphi_{1/(2\rho)}(x)\| \leq \varepsilon$ is at most

$$\left\lceil 16 \cdot \frac{(\rho L D)^2 \cdot \min\left\{1, \frac{L}{\rho D}\right\}}{\varepsilon^4} \right\rceil. \tag{2.10}$$

This complexity in ε matches the guarantees of the stochastic gradient method for finding an ε-stationary point of a smooth function [9, Corollary 2.2].

It is intriguing to ask if the complexity (2.10) can be improved when g is a convex function. The answer, unsurprisingly, is yes. Since g is convex, here and for the rest of the section, we will let the constant $\rho > 0$ be arbitrary. As a first attempt, one may follow the observation of Nesterov [15] for smooth minimization. The idea is that the right-hand-side of the complexity bound (2.9) dependence on the initial gap $\varphi(x_0) - \min \varphi$. We can make this quantity as small as we wish by a separate subgradient method. Namely, we may simply run a subgradient method for T iterations to decrease the gap $\varphi(x_0) - \min \varphi$ to $R := LD/\sqrt{T+1}$; see for example [11, Proposition 5.5] for the this basic guarantee. Then we run another round of a subgradient method for T iterations using the optimal choice $\gamma := \sqrt{\frac{R}{\rho L^2}}$. A quick computation shows that the resulting scheme will find a point x satisfying $\mathbb{E}\|\nabla \varphi_{1/(2\rho)}(x)\| \leq \varepsilon$ after at most $O(1) \cdot \frac{L^2(\rho D)^{2/3}}{\varepsilon^{8/3}}$ iterations.

This complexity can be improved slightly by first regularizing the problem. We will only outline the procedure here, since the details are standard and easy to verify. Define the function $\widehat{\varphi} := \varphi + \frac{\mu}{2}\|\cdot - x_{\mathrm{c}}\|^2$, for some $\mu > 0$ and arbitrary $x_{\mathrm{c}} \in \mathcal{X}$. We will apply optimization algorithms to $\widehat{\varphi}$ instead of φ, and therefore we must relate their Moreau envelopes. Fixing an arbitrary $\lambda > 0$, it is straightforward to verify the following equality by completing the square in the Moreau envelope:

$$\widehat{\varphi}_{1/\lambda}(x) = \varphi_{1/(\lambda+\mu)}\left(\tfrac{\mu}{\mu+\lambda}x_{\mathrm{c}} + \tfrac{\lambda}{\mu+\lambda}x\right) + \tfrac{\lambda\mu}{2(\mu+\lambda)}\|x - x_{\mathrm{c}}\|^2.$$

Differentiating in x yields the bound

$$\left\|\nabla\varphi_{1/(\lambda+\mu)}\left(\tfrac{\mu}{\mu+\lambda}x_{\mathrm{c}}+\tfrac{\lambda}{\mu+\lambda}x\right)\right\| \leq \tfrac{\lambda+\mu}{\lambda}\|\nabla\widehat{\varphi}_{1/\lambda}(x)\| + \mu D. \tag{2.11}$$

Thus, supposing $\varepsilon \leq 2\rho D$, we may set $\mu = \frac{\varepsilon}{2D}$ and $\lambda = 2\rho - \frac{\varepsilon}{2D}$, obtaining the estimate

$$\left\|\nabla\varphi_{1/(2\rho)}\left(\tfrac{\mu}{\mu+\lambda}x_{\mathrm{c}}+\tfrac{\lambda}{\mu+\lambda}x\right)\right\| \leq 2\|\nabla\widehat{\varphi}_{1/\lambda}(x)\| + \tfrac{\varepsilon}{2}.$$

Hence, if we find a point x satisfying $\mathbb{E}\|\nabla\widehat{\varphi}_{1/\lambda}(x)\| \leq \frac{\varepsilon}{4}$, then the convex combination $z := \frac{\mu}{\mu+\lambda}x_{\mathrm{c}} + \frac{\lambda}{\mu+\lambda}x$ would satisfy $\mathbb{E}\|\nabla\varphi_{1/(2\rho)}(z)\| \leq \varepsilon$, as desired. Let us now apply the two-stage procedure on the strongly convex function $\widehat{\varphi}$. We first apply the projected stochastic subgradient method [12] for T iterations yielding the estimate $R := \frac{4(L^2+\mu^2D^2)}{\mu(T+1)} = \frac{2D(4L^2+\varepsilon^2)}{\varepsilon(T+1)}$. Then we apply the subgradient method (Algorithm 1) for T iterations on $\widehat{\varphi}$ with an optimal step-size γ. Solving for T in terms of ε, a quick computation shows that the resulting scheme will find a point z satisfying $\mathbb{E}\|\nabla\varphi_{1/(2\rho)}(z)\| \leq \varepsilon$ after at most $O(1)\cdot\frac{(L^2+\varepsilon^2)\sqrt{\rho D}}{\varepsilon^{2.5}}$ iterations. By following a completely different technique, introduced by Allen-Zhu [1] for smooth stochastic minimization, this complexity can be even further improved to $\widetilde{O}\left(\frac{(L^2+\rho^2D^2)\log^3(\frac{\rho D}{\varepsilon})}{\varepsilon^2}\right)$ by running logarithmically many rounds of the subgradient method. Since this procedure and its analysis is somewhat technical and is independent of the rest of the material, we have placed it in a supplementary text that can be found at `www.math.washington.edu/~ddrusv/sms.pdf`

2.2 Proximal stochastic subgradient method

We next move on to convergence guarantees of Algorithm 1 in full generality – the main result of this work. To this end, in this section, in addition to assumptions (A1), (A2), and (A3) we will also assume that g is L-Lipschitz.

We break up the analysis of Algorithm 1 into two lemmas. Henceforth, fix a real $\hat{\rho} > \rho$. Let x_t be the iterates produced by Algorithm 1 and let $\xi_t \sim dP$ be the i.i.d. realizations used. For each index t, define $\zeta_t := \mathbb{E}_t[G(x_t,\xi)] \in \partial g(x_t)$ and set $\hat{x}_t := \mathrm{prox}_{\varphi/\hat{\rho}}(x_t)$. Observe that by the optimality conditions of the proximal map, there exists a vector $\hat{\zeta}_t \in \partial g(\hat{x}_t)$ satisfying $\hat{\rho}(x_t - \hat{x}_t) \in \partial r(\hat{x}_t) + \hat{\zeta}_t$. The following lemma realizes $\hat{x}_t$ as a proximal point of r.

Lemma 2.3. *For each index $t \geq 0$, equality holds:*

$$\hat{x}_t = \mathrm{prox}_{\alpha_t r}\left(\alpha_t\hat{\rho}x_t - \alpha_t\hat{\zeta}_t + (1-\alpha_t\hat{\rho})\hat{x}_t\right).$$

Proof. By the definition of $\hat{\zeta}_t$, we have

$$\begin{aligned}\alpha_t\hat{\rho}(x_t-\hat{x}_t) \in \alpha_t\partial r(\hat{x}_t) + \alpha_t\hat{\zeta}_t &\iff \alpha_t\hat{\rho}x_t - \alpha_t\hat{\zeta}_t + (1-\alpha_t\hat{\rho})\hat{x}_t \in \hat{x}_t + \alpha_t\partial r(\hat{x}_t)\\ &\iff \hat{x}_t = \mathrm{prox}_{\alpha_t r}(\alpha_t\hat{\rho}x_t - \alpha_t\hat{\zeta}_t + (1-\alpha_t\hat{\rho})\hat{x}_t),\end{aligned}$$

where the last equivalence follows from the optimality conditions for the proximal subproblem. This completes the proof. □

The next lemma establishes a crucial descent property for the iterates.

Lemma 2.4. *Suppose $\hat{\rho} \in (\rho, 2\rho]$ and we have $\alpha_t \leq \hat{\rho}^{-1}$ for all indices $t \geq 0$. Then the inequality holds:*

$$\mathbb{E}_t\|x_{t+1} - \hat{x}_t\|^2 \leq \|x_t - \hat{x}_t\|^2 + 2\alpha_t^2 L^2 - 2\alpha_t(\hat{\rho} - \rho)\|x_t - \hat{x}_t\|^2.$$

Proof. We successively deduce

$$\begin{aligned}
&\mathbb{E}_t\|x_{t+1} - \hat{x}_t\|^2 \\
&= \mathbb{E}_t\|\mathrm{prox}_{\alpha_t r}(x_t - \alpha_t G(x_t, \xi_t)) - \mathrm{prox}_{\alpha_t r}(\alpha_t\hat{\rho}x_t - \alpha_t\hat{\zeta}_t + (1 - \alpha_t\hat{\rho})\hat{x}_t)\|^2 && (2.12)\\
&\leq \mathbb{E}_t\|x_t - \alpha_t G(x_t, \xi_t) - (\alpha_t\hat{\rho}x_t - \alpha_t\hat{\zeta}_t + (1 - \alpha_t\hat{\rho})\hat{x}_t)\|^2 && (2.13)\\
&= \mathbb{E}_t\|(1 - \alpha_t\hat{\rho})(x_t - \hat{x}_t) - \alpha_t(G(x_t, \xi_t) - \hat{\zeta}_t)\|^2 && (2.14)\\
&= (1 - \alpha_t\hat{\rho})^2\|x_t - \hat{x}_t\|^2 - 2(1 - \alpha_t\hat{\rho})\alpha_t\mathbb{E}_t\left[\langle x_t - \hat{x}_t, G(x_t, \xi_t) - \hat{\zeta}_t\rangle\right] + \alpha_t^2\mathbb{E}_t\|G(x_t, \xi_t) - \hat{\zeta}_t\|^2 \\
&= (1 - \alpha_t\hat{\rho})^2\|x_t - \hat{x}_t\|^2 - 2(1 - \alpha_t\hat{\rho})\alpha_t\langle x_t - \hat{x}_t, \zeta_t - \hat{\zeta}_t\rangle + \alpha_t^2\mathbb{E}_t\|G(x_t, \xi_t) - \hat{\zeta}_t\|^2 \\
&\leq (1 - \alpha_t\hat{\rho})^2\|x_t - \hat{x}_t\|^2 + 2(1 - \alpha_t\hat{\rho})\alpha_t\rho\|x_t - \hat{x}_t\|^2 + 2\alpha_t^2 L^2 && (2.15)\\
&= \|x_t - \hat{x}_t\|^2 + 2\alpha_t^2 L^2 - (2\alpha_t(\hat{\rho} - \rho) + \alpha_t^2\hat{\rho}(2\rho - \hat{\rho}))\|x_t - \hat{x}_t\|^2,
\end{aligned}$$

where (2.12) follows from Lemma 2.3, the inequality (2.13) uses that $\mathrm{prox}_{\alpha_t r}(\cdot)$ is 1-Lipschitz [24, Proposition 12.19], and (2.15) follows from the inequality (2.2). The result now follows from the assumed inequality $\hat{\rho} \leq 2\rho$. □

With Lemma 2.4 proved, we can now establish convergence guarantees of Algorithm 1 in full generality.

Theorem 2.5 (Stochastic proximal subgradient method)**.** *Fix a real $\hat{\rho} \in (\rho, 2\rho]$ and a stepsize sequence $\alpha_t \in (0, \hat{\rho}^{-1}]$. Then the point x_{t^*} returned by Algorithm 1 satisfies:*

$$\mathbb{E}\left[\|\nabla\varphi_{1/\hat{\rho}}(x_{t^*})\|^2\right] \leq \frac{\hat{\rho}}{\hat{\rho} - \rho} \cdot \frac{(\varphi_{1/\hat{\rho}}(x_0) - \min\varphi) + \hat{\rho}L^2\sum_{t=0}^T \alpha_t^2}{\sum_{t=0}^T \alpha_t}.$$

Proof. We successively observe

$$\begin{aligned}
\mathbb{E}_t\left[\varphi_{1/\hat{\rho}}(x_{t+1})\right] &\leq \mathbb{E}_t\left[\varphi(\hat{x}_t) + \frac{\hat{\rho}}{2}\|x_{t+1} - \hat{x}_t\|^2\right] \\
&\leq \varphi(\hat{x}_t) + \frac{\hat{\rho}}{2}\left[\|x_t - \hat{x}_t\|^2 + 2\alpha_t^2 L^2 - 2\alpha_t(\hat{\rho} - \rho)\|x_t - \hat{x}_t\|^2\right] \\
&= \varphi_{1/\hat{\rho}}(x_t) + \hat{\rho}\left[\alpha_t^2 L^2 - \alpha_t(\hat{\rho} - \rho)\|x_t - \hat{x}_t\|^2\right],
\end{aligned}$$

where the first inequality follows directly from the definition of the proximal map and the second follows from Lemma 2.4.

Using the law of total expectation to unfold this recursion yields:

$$\mathbb{E}\left[\varphi_{1/\hat{\rho}}(x_{T+1})\right] \leq \varphi_{1/\hat{\rho}}(x_0) + \hat{\rho}L^2\sum_{t=0}^T \alpha_t^2 - \hat{\rho}(\hat{\rho} - \rho)\mathbb{E}\sum_{t=0}^T \alpha_t\|x_t - \hat{x}_t\|^2.$$

Next using the inequality $\varphi_{1/\hat{\rho}}(x_{T+1}) \geq \min \varphi$ and rearranging, we obtain the bound:

$$\frac{1}{\sum_{t=0}^{T} \alpha_t} \sum_{t=0}^{T} \alpha_t \mathbb{E}\left[\|x_t - \hat{x}_t\|^2\right] \leq \frac{(\varphi_{1/\hat{\rho}}(x_0) - \min \varphi) + \hat{\rho} L^2 \sum_{t=0}^{T} \alpha_t^2}{\hat{\rho}(\hat{\rho} - \rho) \sum_{t=0}^{T} \alpha_t}. \tag{2.16}$$

To complete the proof, observe that the left-hand-side is exactly $\mathbb{E}\left[\|x_{t^*} - \hat{x}_{t^*}\|^2\right]$, while from from (1.3) we have the equality $\|x_{t^*} - \hat{x}_{t^*}\|^2 = (1/\hat{\rho}^2)\|\nabla \varphi_{1/\hat{\rho}}(x_{t^*})\|^2$. □

In particular, using the constant stepsize α on the order of $\frac{1}{\sqrt{T+1}}$ yields the following complexity guarantee.

Corollary 2.6 (Complexity guarantee)**.** *Fix a constant $\gamma \in (0, \frac{1}{2\rho}]$ and an index $T > 0$, and set the constant steplength $\alpha = \frac{\gamma}{\sqrt{T+1}}$. Then the point x_{t^*} returned by Algorithm 1 satisfies:*

$$\mathbb{E}\left[\|\nabla \varphi_{1/(2\rho)}(x_{t^*})\|^2\right] \leq 2 \cdot \frac{\left(\varphi_{1/(2\rho)}(x_0) - \min \varphi\right) + \rho L^2 \gamma^2}{\gamma \sqrt{T+1}}.$$

Proof. This follows immediately from Theorem 2.5 by setting $\hat{\rho} = 2\rho$. □

2.3 Proximal stochastic gradient for smooth minimization

Let us now look at the consequences of our results in the setting when g is C^1-smooth with ρ-Lipschitz gradient. Note, that then g is automatically ρ-weakly convex. In this smooth setting, it is common to replace assumption (A3) with the finite variance condition:

$\overline{(A3)}$ There is a real $\sigma \geq 0$ such that the inequality, $\mathbb{E}_\xi\left[\|G(x,\xi) - \nabla g(x)\|^2\right] \leq \sigma^2$, holds for all $x \in \operatorname{dom} r$.

Henceforth, let us therefore assume that g is C^1-smooth with ρ-Lipschitz gradient and Assumptions (A1), (A2), and $\overline{\text{(A3)}}$ hold.

All of the results in Section 2.2 can be easily modified to apply to this setting. In particular, Lemma 2.3 holds verbatim, while Lemma 2.4 extends as follows.

Lemma 2.7. *Fix a real $\hat{\rho} > \rho$ and a sequence $\alpha_t \in (0, \hat{\rho}^{-1}]$. Then the inequality holds:*

$$\mathbb{E}_t\|x_{t+1} - \hat{x}_t\|^2 \leq \|x_t - \hat{x}_t\|^2 + \alpha_t^2\sigma^2 - \alpha_t(\hat{\rho} - \rho)\|x_t - \hat{x}_t\|^2.$$

Proof. By the same argument as in Lemma 2.7, we arrive at the inequality (2.14) with $\hat{\zeta}_t = \nabla g(\hat{x}_t)$. Adding and subtracting $\nabla g(x_t)$, we successively deduce

$$\begin{aligned}
&\mathbb{E}_t\|x_{t+1} - \hat{x}_t\|^2 \\
&= \mathbb{E}_t\|(1 - \alpha_t\hat{\rho})(x_t - \hat{x}_t) - \alpha_t(G(x_t,\xi_t) - \nabla g(\hat{x}_t))\|^2 \\
&= \mathbb{E}_t\|(1 - \alpha_t\hat{\rho})(x_t - \hat{x}_t) - \alpha_t(\nabla g(x_t) - \nabla g(\hat{x}_t)) - \alpha_t(G(x_t,\xi_t) - \nabla g(x_t))\|^2 \\
&= \|(1 - \alpha_t\hat{\rho})(x_t - \hat{x}_t) - \alpha_t(\nabla g(x_t) - \nabla g(\hat{x}_t))\|^2 + \alpha_t^2\mathbb{E}_t\|G(x_t,\xi_t) - \nabla g(x_t)\|^2 && (2.17)\\
&\leq (1 - \alpha_t\hat{\rho})^2\|x_t - \hat{x}_t\|^2 - 2(1 - \alpha_t\hat{\rho})\alpha_t\langle x_t - \hat{x}_t, \nabla g(x_t) - \nabla g(\hat{x}_t)\rangle \\
&\qquad + \alpha_t^2\|\nabla g(x_t) - \nabla g(\hat{x}_t)\|^2 + \alpha_t^2\sigma^2 && (2.18)\\
&= (1 - \alpha_t\hat{\rho})^2\|x_t - \hat{x}_t\|^2 + 2(1 - \alpha_t\hat{\rho})\alpha_t\rho\|x_t - \hat{x}_t\|^2 + \rho^2\alpha_t^2\|x_t - \hat{x}_t\|^2 + \alpha_t^2\sigma^2 && (2.19)\\
&= \|x_t - \hat{x}_t\|^2 + \alpha_t^2\sigma^2 - (2\alpha_t(\hat{\rho} - \rho) + \alpha_t^2\hat{\rho}(2\rho - \hat{\rho}) - \rho^2\alpha_t^2)\|x_t - \hat{x}_t\|^2, \\
&= \|x_t - \hat{x}_t\|^2 + \alpha_t^2\sigma^2 - \alpha_t(\hat{\rho} - \rho)(2 - \alpha_t(\hat{\rho} - \rho))\|x_t - \hat{x}_t\|^2,
\end{aligned}$$

where (2.17) follows from assumption (A2), namely $\mathbb{E}_t G(x_t, \xi_t) = \nabla g(x_t)$, inequality (2.18) follows by expanding the square and using assumption $\overline{(A3)}$, and inequality (2.19) follows from (2.2) and Lipschitz continuity of ∇g. The assumption $\hat{\rho} \geq \rho$ guarantees $2 - \alpha_t(\hat{\rho} - \rho) \geq 1$. The result follows. □

We can now state the convergence guarantees of the proximal stochastic gradient method. The proof is completely analogous to that of Theorem 2.5, with Lemma 2.7 playing the role of Lemma 2.4.

Corollary 2.8 (Stochastic proximal gradient method for smooth minimization)**.** *Fix a real $\hat{\rho} > \rho$ and a stepsize sequence $\alpha_t \in (0, \hat{\rho}^{-1}]$. Then the point x_{t^*} returned by Algorithm 1 satisfies:*

$$\mathbb{E}\left[\|\nabla \varphi_{1/\hat{\rho}}(x_{t^*})\|^2\right] \leq \frac{2\hat{\rho}}{\hat{\rho} - \rho} \cdot \frac{(\varphi_{1/\hat{\rho}}(x_0) - \min \varphi) + \frac{\hat{\rho}\sigma^2}{2} \sum_{t=0}^{T} \alpha_t^2}{\sum_{t=0}^{T} \alpha_t}.$$

In particular, setting $\alpha = \frac{\gamma}{\sqrt{T+1}}$ for some real $\gamma \in (0, \frac{1}{2\rho}]$ yields the guarantee

$$\mathbb{E}\left[\|\nabla \varphi_{1/(2\rho)}(x_{t^*})\|^2\right] \leq 4 \cdot \frac{(\varphi_{1/\hat{\rho}}(x_0) - \min \varphi) + \rho\sigma^2\gamma^2}{\gamma\sqrt{T+1}}.$$

It is worth noting that in this smooth setting, the norm $\|\nabla \varphi_\lambda(x)\|$ is closely related to the magnitude of the *prox-gradient mapping*

$$\mathcal{G}_\lambda(x) = \lambda^{-1}\left(x - \mathrm{prox}_{\lambda r}(x - \lambda \nabla g(x))\right).$$

This stationarity measure typically appears when analyzing proximal gradient methods (e.g. [10, 16]). It is straightforward to see that this measure is proportional to the norm of the gradient of the Moreau envelope, the quantity we have been using [6, Theorem 3.5]:

$$(1 - \rho\lambda)\|\mathcal{G}_\lambda(x)\| \leq \|\nabla \varphi_\lambda(x)\| \leq (1 + \rho\lambda)\|\mathcal{G}_\lambda(x)\| \qquad \text{for all } x \in \mathbb{R}^d.$$

Hence, the convergence guarantees of Corollary 2.8 can be immediately translated in terms of $\|\mathcal{G}_{1/(2\rho)}(x_t^*)\|$, allowing for a direct comparison with previous results.

References

[1] Z. Allen-Zhu. How to make gradients small stochastically. *Preprint arXiv:1801.02982 (version 1)*, 2018.

[2] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

[3] J.Y.B. Cruz. On proximal subgradient splitting method for minimizing the sum of two nonsmooth convex functions. *Set-Valued Var. Anal.*, 25(2):245–263, 2017.

[4] D. Davis and B. Grimmer. Proximally guided stochastic method for nonsmooth, nonconvex problems. *Preprint arXiv:1707.03505*, 2017.

[5] D. Drusvyatskiy. The proximal point method revisited. *To appear in SIAG/OPT Views and News, arXiv:1712.06038*, 2018.

[6] D. Drusvyatskiy and A.S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *To appear in Math. Oper. Res., arXiv:1602.06661*, 2016.

[7] D. Drusvyatskiy and C. Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Preprint arXiv:1605.00125*, 2016.

[8] J. Duchi and Y. Singer. Efficient online and batch learning using forward backward splitting. *J. Mach. Learn. Res.*, 10:2899–2934, 2009.

[9] S. Ghadimi and G. Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM J. Optim.*, 23(4):2341–2368, 2013.

[10] S. Ghadimi, G. Lan, and H. Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Math. Program.*, 155(1):267–305, 2016.

[11] A. Juditsky and A. Nemirovski. First order methods for nonsmooth convex large-scale optimization, I: General purpose methods. In S. Sra, S. Nowozin, and S.J. Write, editors, *Optimization for Machine Learning*, chapter 1, pages 266–290. MIT Press, 2011.

[12] S. Lacoste-Julien, M.W. Schmidt, and F.R. Bach. A simpler approach to obtaining an $O(1/t)$ convergence rate for the projected stochastic subgradient method. *CoRR*, abs/1212.2002, 2012.

[13] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, 19(4):1574–1609, 2009.

[14] A.S. Nemirovsky and D.B. Yudin. *Problem complexity and method efficiency in optimization.* A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1983.

[15] Y. Nesterov. How to make the gradients small. *OPTIMA, MPS Newsletter*, (88):10–11, 2012.

[16] Yu. Nesterov. Gradient methods for minimizing composite functions. *Math. Program.*, 140(1, Ser. B):125–161, 2013.

[17] E. A. Nurminskii. The quasigradient method for the solving of the nonlinear programming problems. *Cybernetics*, 9(1):145–150, Jan 1973.

[18] E. A. Nurminskii. Minimization of nondifferentiable functions in the presence of noise. *Cybernetics*, 10(4):619–621, Jul 1974.

[19] N. Parikh and S. Boyd. Proximal algorithms. *Found. Trends Optim.*, 1(3):127–239, January 2014.

[20] R.A. Poliquin and R.T. Rockafellar. Amenable functions in optimization. In *Nonsmooth optimization: methods and applications (Erice, 1991)*, pages 338–353. Gordon and Breach, Montreux, 1992.

[21] R.A. Poliquin and R.T. Rockafellar. Prox-regular functions in variational analysis. *Trans. Amer. Math. Soc.*, 348:1805–1838, 1996.

[22] R.T. Rockafellar. *Convex Analysis.* Princeton University Press, 1970.

[23] R.T. Rockafellar. Favorable classes of Lipschitz-continuous functions in subgradient optimization. In *Progress in nondifferentiable optimization*, volume 8 of *IIASA Collaborative Proc. Ser. CP-82*, pages 125–143. Int. Inst. Appl. Sys. Anal., Laxenburg, 1982.

[24] R.T. Rockafellar and R.J-B. Wets. *Variational Analysis*. Grundlehren der mathematischen Wissenschaften, Vol 317, Springer, Berlin, 1998.

[25] Y. Xu and W. Yin. Block stochastic gradient iteration for convex and nonconvex optimization. *SIAM J. Optim.*, 25(3):1686–1716, 2015.