Complexity of finding near-stationary points of convex functions stochastically

Damek Davis* Dmitriy Drusvyatskiy†

Abstract

In the recent paper [3], it was shown that the stochastic subgradient method applied to a weakly convex problem, drives the gradient of the Moreau envelope to zero at the rate $O(k^{-1/4})$. In this supplementary note, we present a stochastic subgradient method for minimizing a convex function, with the improved rate $\widetilde{O}(k^{-1/2})$.

1 Introduction

Efficiency of algorithms for minimizing smooth convex functions is typically judged by the rate at which the function values decrease along the iterate sequence. A different measure of performance, which has received some attention lately, is the magnitude of the gradient. In the short note [12], Nesterov showed that performing two rounds of a fast-gradient method on a slightly regularized problem yields an ε-stationary point in $\widetilde{O}(\varepsilon^{-1/2})$ iterations.[1] This rate is in sharp contrast to the blackbox optimal complexity of $O(\varepsilon^{-2})$ in smooth nonconvex optimization [2], trivially achieved by gradient descent. An important consequence is that the prevalent intuition – smooth convex optimization is easier than its nonconvex counterpart – attains a very precise mathematical justification. In the recent work [1], Allen-Zhu investigated the complexity of finding ε-stationary points in the setting when only stochastic estimates of the gradient are available. In this context, Nesterov's strategy paired with a stochastic gradient method (SG) only yields an algorithm with complexity $O(\varepsilon^{-2.5})$. Consequently, the author introduced a new technique based on running SG for logarithmically many rounds, which enjoys the near-optimal efficiency $\widetilde{O}(\varepsilon^{-2})$.

In this short technical note, we address a similar line of questions for nonsmooth convex optimization. Clearly, there is a caveat: in nonsmooth optimization, it is impossible to find points with small subgradients, within a first-order oracle model. Instead, we focus on

*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.

[1] In this section to simplify notation, we only show dependence on the accuracy ε and suppress all dependence on the initialization and Lipschitz constants.

the gradients of an implicitly defined smooth approximation of the function, the Moreau envelope.

Throughout, we consider the optimization problem

$$\min_{x\in\mathcal{X}}\ g(x), \tag{1.1}$$

where $\mathcal{X} \subseteq \mathbb{R}^d$ is a closed convex set with a computable nearest-point map $\mathrm{proj}_{\mathcal{X}}$, and $g\colon \mathbb{R}^d \to \mathbb{R}$ a Lipschitz convex function. Henceforth, we assume that the only access to g is through a stochastic subgradient oracle; see Section 1.1 for a precise definition. It will be useful to abstract away the constraint set $\mathcal{X}$ and define $\varphi\colon \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ to be equal to g on $\mathcal{X}$ and $+\infty$ off $\mathcal{X}$. Thus the target problem (1.1) is equivalent to $\min_{x\in\mathbb{R}^d}\varphi(x)$. In this generality, there are no efficient algorithms within the first-order oracle model that can find ε-stationary points, in the sense of $\mathrm{dist}(0;\partial\varphi(x)) \le \varepsilon$. Instead we focus on finding approximately stationary points of the Moreau envelope:

$$\varphi_\lambda(x) = \min_{y\in\mathbb{R}^d}\ \{\varphi(y) + \tfrac{1}{2\lambda}\|y-x\|^2\}.$$

It is well-known that $\varphi_\lambda(\cdot)$ is C^1-smooth for any $\lambda > 0$, with gradient

$$\nabla\varphi_\lambda(x) = \lambda^{-1}(x - \mathrm{prox}_{\lambda\varphi}(x)), \tag{1.2}$$

where $\mathrm{prox}_{\lambda\varphi}(x)$ is the proximal point

$$\mathrm{prox}_{\lambda\varphi}(x) := \operatorname*{argmin}_{y\in\mathbb{R}^d}\ \left\{\varphi(y) + \tfrac{1}{2\lambda}\|y-x\|^2\right\}.$$

When g is smooth, the norm of the gradient $\|\nabla\varphi_\lambda(x)\|$ is proportional to the norm of the prox-gradient (e.g. [5], [6, Theorem 3.5]), commonly used in convergence analysis of proximal gradient methods [7, 13]. In the broader nonsmooth setting, the quantity $\|\nabla\varphi_\lambda(x)\|$ nonetheless has an appealing 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

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

Thus a small gradient $\|\nabla\varphi_\lambda(x)\|$ implies that x is *near* some point $\hat{x}$ that is *nearly stationary* for (1.1). The recent paper [3] notes that following Nesterov's strategy of running two rounds of the projected stochastic subgradient method on a quadratically regularized problem, will find a point x satisfying $\mathbb{E}\|\nabla\varphi_\lambda(x)\| \le \varepsilon$ after at most $O(\varepsilon^{-2.5})$ iterations. This is in sharp contrast to the complexity $O(\varepsilon^{-4})$ for minimizing functions that are only weakly convex — the main result of [3]. Notice the parallel here to the smooth setting. In this short note, we show that the gradual regularization technique of Allen-Zhu [1], along with averaging of the iterates, improves the complexity to $\widetilde{O}(\varepsilon^{-2})$ in complete analogy to the smooth setting.

1.1 Convergence Guarantees

Let us first make precise the notion of a stochastic subgradient oracle. To this end, 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 $\mathcal{X}$ 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 $x \in \mathcal{X}$.

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 [11], while Assumption (A3) is the same as the assumption listed in [11, Equation (2.5)].

Henceforth, we fix an arbitrary constant $\rho > 0$ and assume that diameter of $\mathcal{X}$ is bounded by some real $D > 0$. It was shown in [4, Section 2.1] that the complexity of finding a point x satisfying $\mathbb{E}\|\nabla\varphi_{1/\rho}(x)\| \leq \varepsilon$ is at most $O(1)\cdot\frac{(L^2+\varepsilon^2)\sqrt{\rho D}}{\varepsilon^{2.5}}$. We will see here that this complexity can be improved to $\widetilde{O}\left(\frac{L^2+\rho^2D^2}{\varepsilon^2}\right)$ by adapting the technique of [1].

The work horse of the strategy is the subgradient method for minimizing strongly convex functions [8–10,14]. For the sake of concreteness, we summarize in Algorithm 1 the stochastic subgradient method taken from [10].

Algorithm 1: Projected stochastic subgradient method for strongly convex functions $\mathrm{PSSM}^{\mathrm{sc}}(x_0, \mu, G, T)$

Data: $x_0 \in \mathcal{X}$, strong convexity constant $\mu > 0$ on $\mathcal{X}$, maximum iterations $T \in \mathbb{N}$, stochastic subgradient oracle G.

Step $t = 0, \ldots, T-2$:

$$\left\{\begin{array}{l}\text{Sample } \xi_t \sim dP \\ \text{Set } x_{t+1} = \mathrm{proj}_{\mathcal{X}}\left(x_t - \frac{2}{\mu(t+1)}\cdot G(x_t,\xi_t)\right)\end{array}\right\},$$

Return: $\bar{x} = \frac{2}{T(T+1)}\sum_{t=0}^{T-1}(t+1)x_t$.

The following is the basic convergence guarantee of Algorithm 1, proved in [10].

Theorem 1.1. *The point $\bar{x}$ returned by Algorithm 1 satisfies the estimate*

$$\mathbb{E}\left[\varphi(\bar{x}) - \min\varphi\right] \leq \frac{2L^2}{\mu(T+1)}.$$

For the time being, let us assume that g is μ-strongly convex on $\mathcal{X}$. Later, we will add a small quadratic to g to ensure this to be the case. The algorithm we consider follows an inner outer construction, proposed in [1]. We will fix the number of inner iterations $T \in \mathbb{N}$. and the number of outer iterations $\mathcal{I} \in \mathbb{N}$. We set $\varphi^{(0)} = \varphi$ and for each $i = 1, \ldots, \mathcal{I}$ define the quadratic perturbations

$$\varphi^{(i+1)}(x) := \varphi^{(i)}(x) + \mu 2^{i-1}\|x - \hat{x}_{i+1}\|^2.$$

Each center $\hat{x}_{i+1}$ is obtained by running T iterations of the subgradient method Algorithm 1 on $\varphi^{(i)}$. We record the resulting procedure in Algorithm 2. We emphasize that this algorithm is identical to the method in [1], with the only difference being the stochastic subgradient method used in the inner loop.

Algorithm 2: Gradual regularization for strongly convex problems $\mathrm{GR}^{\mathrm{sc}}(x_1, \mu, \lambda, T, \mathcal{I}, G)$

Data: Initial point $x_1 \in \mathcal{X}$, strong convexity constant $\mu > 0$, an averaging parameter $\lambda > 0$, inner iterations $T \in \mathbb{N}$, outer iterations $\mathcal{I} \in \mathbb{N}$, stochastic oracle $G(\cdot,\cdot)$.

Set $\varphi^{(0)} = \varphi$, $G^{(0)} = G$, $\hat{x}_0 = x_0$, $\mu_0 = \mu$.

Step $i = 0, \ldots, \mathcal{I}$:

Set $\hat{x}_{i+1} = \mathrm{PSSM}^{\mathrm{sc}}(\hat{x}_i, \sum_{j=0}^{i} \mu_j, G^{(i)}, T)$

$\mu_{i+1} = \mu \cdot 2^{i+1}$

Define the function and the oracle

$$\varphi^{(i+1)}(x) := \varphi^{(i)}(x) + \frac{\mu_{i+1}}{2}\|x - \hat{x}_{i+1}\|^2 \quad \text{and} \quad G^{(i+1)}(x,\xi) := G^{(i)}(x,\xi) + \mu_{i+1}(x - \hat{x}_{i+1}).$$

Return: $\bar{x} = \frac{1}{\lambda + \sum_{i=1}^{\mathcal{I}} \mu_i}(\lambda \hat{x}_{\mathcal{I}+1} + \sum_{i=1}^{\mathcal{I}} \mu_i \hat{x}_i)$.

Henceforth, let μ_i, $\varphi^{(i)}$, and $\hat{x}_i$ be generated by Algorithm 2. Observe that by construction, equality

$$\varphi^{(i)}(x) = \varphi(x) + \sum_{j=1}^{i} \frac{\mu_i}{2}\|x - \hat{x}_i\|^2,$$

holds for all $i = 1, \ldots, \mathcal{I}$. Consequently, it will be important to relate the Moreau envelope of $\varphi^{(i)}$ to that of φ. This is the content of the following two elementary lemmas.

Lemma 1.2 (Completing the square)**.** *Fix a set of points $z_i \in \mathbb{R}^d$ and real $a_i > 0$, for $i = 1, \ldots, \mathcal{I}$. Define the convex quadratic*

$$Q(y) = \sum_{i=1}^{\mathcal{I}} \frac{a_i}{2}\|y - z_i\|^2.$$

Then equality holds:

$$Q(y) = Q(\bar{z}) + \frac{\sum_{i=1}^{\mathcal{I}} a_i}{2}\|y - \bar{z}\|^2,$$

where $\bar{z} = \frac{1}{\sum_{i=1}^{\mathcal{I}} a_i}\sum_{i=1}^{\mathcal{I}} a_i z_i$ is the centroid.

Proof. Taking the derivative shows that $Q(\cdot)$ is minimized at $\bar{z}$. The result follows. $\square$

Lemma 1.3 (Moreau envelope of the regularization)**.** *Consider a function $h\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ and define the quadratic perturbation*

$$f(x) = h(x) + \sum_{i=1}^{\mathcal{I}} \frac{a_i}{2}\|x - z_i\|^2,$$

for some $z_i \in \mathbb{R}^d$ *and* $a_i > 0$*, with* $i = 1, \ldots, \mathcal{I}$*. Then for any* $\lambda > 0$*, the Moreau envelopes of* h *and* f *are related by the expression*

$$\nabla f_{1/\lambda}(x) = \tfrac{\lambda}{\lambda+A}\left(\nabla h_{1/(\lambda+A)}(\bar{x}) + \sum_{i=1}^{\mathcal{I}} a_i(x - z_i)\right),$$

where we define $A := \sum_{i=1}^{\mathcal{I}} a_i$ *and* $\bar{x} := \frac{1}{\lambda+A}\left(\lambda x + \sum_{i=1}^{\mathcal{I}} a_i z_i\right)$ *is the centroid.*

Proof. By definition of the Moreau envelope, we have

$$f_{1/\lambda}(x) = \operatorname*{argmin}_{y}\left\{h(y) + \sum_{i=1}^{\mathcal{I}} \tfrac{a_i}{2}\|y - z_i\|^2 + \tfrac{\lambda}{2}\|y - x\|^2\right\}. \tag{1.3}$$

We next complete the square in the quadratic term. Namely define the convex quadratic:

$$Q(y) := \tfrac{\lambda}{2}\|y - x\|^2 + \sum_{i=1}^{\mathcal{I}} \tfrac{a_i}{2}\|y - z_i\|^2.$$

Lemma 1.2 directly yields the representation $Q(y) = Q(\bar{x}) + \frac{\lambda+A}{2}\|y - \bar{x}\|^2$. Combining with (1.3), we deduce

$$f_{1/\lambda}(x) = h_{1/(\lambda+A)}(\bar{x}) + Q(\bar{x}).$$

Differentiating in x yields the equalities

$$\begin{aligned}\nabla f_{1/\lambda}(x) &= \tfrac{\lambda}{\lambda+A}\nabla h_{1/(\lambda+A)}(\bar{x}) + \lambda\left(\tfrac{\lambda}{\lambda+A} - 1\right)(\bar{x} - x) + \tfrac{\lambda}{\lambda+A}\sum_{i=1}^{\mathcal{I}} a_i(\bar{x} - z_i)\\ &= \tfrac{\lambda}{\lambda+A}\nabla h_{1/(\lambda+A)}(\bar{x}) + \tfrac{\lambda}{\lambda+A}\sum_{i=1}^{\mathcal{I}} a_i(x - z_i),\end{aligned}$$

as claimed. □

The following is the key estimate from [1, Claim 8.3].

Lemma 1.4. *Suppose that for each index* $i = 1, 2, \ldots, \mathcal{I}$*, the vectors* $\hat{x}_i$ *satisfy*

$$\mathbb{E}[\varphi^{(i-1)}(\hat{x}_i) - \min \varphi^{(i-1)}] \le \delta_i.$$

Then the inequality holds:

$$\mathbb{E}\left[\sum_{i=1}^{\mathcal{I}} \mu_i\|x^*_{\mathcal{I}} - \hat{x}_i\|\right] \le 4\sum_{i=1}^{\mathcal{I}} \sqrt{\delta_i \mu_i},$$

where $x^*_{\mathcal{I}}$ *is the minimizer of* $\varphi^{\mathcal{I}}$*.*

Henceforth, set

$$M_i := \sum_{j=1}^{i} \mu_j \qquad \text{and} \qquad M := M_{\mathcal{I}}.$$

By convention, we will set $M_0 = 0$. Combining Lemmas 1.3 and 1.4, we arrive at the following basic guarantee of the method.

Corollary 1.5. *Suppose for $i = 1, 2, \ldots, \mathcal{I}+1$, the vectors $\hat{x}_i$ satisfy*

$$\mathbb{E}[\varphi^{(i-1)}(\hat{x}_i) - \min \varphi^{(i-1)}] \leq \delta_i.$$

Then the inequality holds:

$$\mathbb{E}\|\nabla \varphi_{1/(\lambda+M)}(\bar{x})\| \leq (\lambda + 2M)\sqrt{\frac{2\delta_{\mathcal{I}+1}}{\mu + M}} + 4\sum_{i=1}^{\mathcal{I}} \sqrt{\delta_i \mu_i},$$

where $\bar{x} = \frac{1}{\lambda+M}(\lambda \hat{x}_{\mathcal{I}+1} + \sum_{i=1}^{\mathcal{I}} \mu_i \hat{x}_i)$.

Proof. Fix an arbitrary point x and set $\bar{x} = \frac{1}{\lambda+M}(\lambda x + \sum_{i=1}^{\mathcal{I}} \hat{x}_i)$. Then Lemma 1.3, along with a triangle inequality, directly implies

$$\begin{aligned}
\|\nabla \varphi_{1/(\lambda+M)}(\bar{x})\| &\leq \left(1 + \tfrac{M}{\lambda}\right) \|\nabla \varphi^{(\mathcal{I})}_{1/\lambda}(x)\| + \sum_{i=1}^{\mathcal{I}} \mu_i \|x - \hat{x}_i\| \\
&\leq \left(1 + \tfrac{M}{\lambda}\right) \|\nabla \varphi^{(\mathcal{I})}_{1/\lambda}(x)\| + \sum_{i=1}^{\mathcal{I}} \mu_i (\|x - x^*_{\mathcal{I}}\| + \|x^*_{\mathcal{I}} - \hat{x}_i\|) \\
&\leq \left(1 + \tfrac{M}{\lambda}\right) \|\nabla \varphi^{(\mathcal{I})}_{1/\lambda}(x)\| + M\|x - x^*_{\mathcal{I}}\| + \sum_{i=1}^{\mathcal{I}} \mu_i \|x^*_{\mathcal{I}} - \hat{x}_i\| \\
&\leq (\lambda + 2M)\|x - x^*_{\mathcal{I}}\| + \sum_{i=1}^{\mathcal{I}} \mu_i \|x^*_{\mathcal{I}} - \hat{x}_i\|.
\end{aligned}$$

where the last inequality uses that $\nabla \varphi^{(\mathcal{I})}_{1/\lambda}$ is λ-Lipschitz continuous and $\nabla \varphi^{(\mathcal{I})}_{1/\lambda}(x^*_{\mathcal{I}}) = 0$ to deduce that $\|\nabla \varphi^{(\mathcal{I})}_{1/\lambda}(x)\| \leq \lambda \|x - x^*_{\mathcal{I}}\|$. Using strong convexity of $\varphi^{\mathcal{I}}$, we deduce

$$\|x - x^*_{\mathcal{I}}\|^2 \leq \tfrac{2}{\mu+M}(\varphi^{(\mathcal{I})}(x) - \varphi^{(\mathcal{I})}(x^*_{\mathcal{I}})).$$

Setting $x = \hat{x}_{\mathcal{I}+1}$, taking expectations, and applying Lemma 1.4 completes the proof. □

Let us now determine $\delta_i > 0$ by invoking Theorem 1.1 for each function $\varphi^{(i)}$. Observe

$$\mathbb{E}_\xi \|G^{(i)}(x, \xi)\|^2 \leq 2(L^2 + D^2 M_i^2).$$

Thus Theorem 1.1 guarantees the estimates:

$$\mathbb{E}[\varphi^{(i-1)}(\hat{x}_i) - \min \varphi^{(i-1)}] \leq \frac{4(L^2 + D^2 M_{i-1}^2)}{(T+1)(\mu + M_{i-1})}, \tag{1.4}$$

Hence for $i = 1, \dots, \mathcal{I}$, we may set δ_i to be the right-hand side of (1.4). Applying Corollary 1.5, we therefore deduce

$$\begin{aligned}\mathbb{E}\|\nabla\varphi_{1/(\lambda+M)}(\bar{x})\| &\le (\lambda+2M)\sqrt{\frac{2\delta_{\mathcal{I}+1}}{\mu+M}} + 4\sum_{i=1}^{\mathcal{I}}\sqrt{\delta_i\mu_i} \\ &\le \frac{1}{\sqrt{T+1}}\left((\lambda+2M)\sqrt{\frac{8(L^2+D^2M^2)}{(\mu+M)^2}} + 4\sum_{i=1}^{\mathcal{I}}\sqrt{\frac{4(L^2+D^2M_{i-1}^2)}{(\mu+M_{i-1})}\cdot\mu_i}\right).\end{aligned} \tag{1.5}$$

Clearly we have $\frac{\mu_1}{\mu} = 2$, while for all $i > 1$, we also obtain

$$\frac{\mu_i}{\mu+M_{i-1}} \le \frac{\mu_i}{\mu+\mu_{i-1}} = \frac{2^i}{1+2^{i-1}} \le 2.$$

Hence, continuing (1.5), we conclude

$$\mathbb{E}\|\nabla\varphi_{1/(\lambda+M)}(\bar{x})\| \le \frac{1}{\sqrt{T+1}}\left(\sqrt{8}\cdot(\lambda+2M)\sqrt{\left(\tfrac{L}{M}\right)^2+D^2} + 8\sqrt{2}\cdot|\mathcal{I}|\cdot\sqrt{L^2+D^2M^2}\right)$$

In particular, by setting $\mathcal{I} = \log_2(1+\frac{\lambda}{2\mu})$, we may ensure $M = \lambda$. For simplicity, we assume the former is an integer. Thus we have proved the following key result.

Theorem 1.6 (Convergence on strongly convex functions)**.** *Suppose g is μ-strongly convex on $\mathcal{X}$ and we set $\mathcal{I} = \log_2(1+\frac{\lambda}{2\mu})$ for some $\lambda > 0$. Then $\bar{x}$ returned by Algorithm 2 satisfies*

$$\mathbb{E}\|\nabla\varphi_{1/(2\lambda)}(\bar{x})\| \le \frac{\left(14\sqrt{2}\cdot\log_2(1+\frac{\lambda}{2\mu})\right)\cdot\sqrt{L^2+D^2\lambda^2}}{\sqrt{T+1}}$$

When g is not strongly convex, we can simply add a small quadratic to the function and run Algorithm 2. For ease of reference, we record the full procedure in Algorithm 3

Algorithm 3: Gradual regularization for non strongly convex problems
Data: Initial point $x_c \in \mathcal{X}$, regularization parameter $\mu > 0$, an averaging parameter $\lambda > 0$, inner iterations $T \in \mathbb{N}$, outer iterations $\mathcal{I} \in \mathbb{N}$, stochastic oracle $G(\cdot,\cdot)$. Set $\widehat{\varphi}(x) := \varphi(x) + \frac{\mu}{2}\|x-x_c\|^2$, $\widehat{G}(x,\xi) = G(x,\xi) + \mu(x-x_c)$, $x_0 = x_c$. Set $\bar{x} = \mathrm{GR}^{\mathrm{sc}}(x_c, \mu, \lambda/2, T, \mathcal{I}, \widehat{G})$ **Return:** $\bar{z} = \frac{\mu}{\mu+\lambda}x_c + \frac{\lambda}{\mu+\lambda}\bar{x}$.

Our main theorem now follows.

Theorem 1.7 (Convergence on convex functions after regularization)**.** *Let $\rho > 0$ be a fixed constant, and suppose we are given a target accuracy $\varepsilon \le 2\rho D$. Set $\mu := \frac{\varepsilon}{2D}$, $\lambda := 2\rho - \frac{\varepsilon}{2D}$, and $\mathcal{I} = \log_2(\frac{3}{4}+\frac{\rho D}{\varepsilon})$. Then for any $T > 0$, Algorithm 3 returns a point $\bar{z}$ satisfying:*

$$\mathbb{E}\|\nabla\varphi_{1/(2\rho)}(\bar{z})\| \le \frac{\left(28\sqrt{2}\cdot\log_2(\frac{3}{4}+\frac{\rho D}{\varepsilon})\right)\cdot\sqrt{2L^2+3\rho^2D^2}}{\sqrt{T+1}} + \frac{\varepsilon}{2}$$

Setting the right hand side to ε and solving for T, we deduce that it suffices to make

$$O\left(\frac{\log^3(\frac{\rho D}{\varepsilon})(L^2+\rho^2 D^2)}{\varepsilon^2}\right)$$

calls to $\mathrm{proj}_{\mathcal{X}}$ *and to the stochastic subgradient oracle in order to find a point $\bar{z} \in \mathcal{X}$ satisfying $\mathbb{E}\|\nabla\varphi_{1/(2\rho)}(\bar{z})\| \leq \varepsilon$.*

Proof. Lemma 1.3 guarantees the bound

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

Applying Theorem 1.6 with λ replaced by $\frac{1}{2}\lambda$ and L replaced by $2(L^2+D^2\mu^2)$, we obtain

$$\mathbb{E}\left\|\nabla\varphi_{1/(2\rho)}(\bar{z})\right\| \leq \tfrac{\lambda+\mu}{\lambda}\frac{\left(14\sqrt{2}\cdot\log_2\left(1+\frac{\lambda}{4\mu}\right)\right)\cdot\sqrt{2(L^2+D^2\mu^2)+\frac{1}{4}D^2\lambda^2}}{\sqrt{T+1}} + \frac{\varepsilon}{2}.$$

Some elementary simplifications yield the result. □

References

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

[2] Y. Carmon, J.C. Duchi, O. Hinder, and A. Sidford. Lower bounds for finding stationary points I. *arXiv:1710.11606*, 2017.

[3] D. Davis and D. Drusvyatskiy. Stochastic subgradient method converges at the rate $O(k^{-1/4})$ on weakly convex functions. *arXiv:1802.02988*, 2018.

[4] D. Davis and D. Drusvyatskiy. Stochastic subgradient method converges at the rate $O(k^{-1/4})$ on weakly convex functions. *arXiv:1802.02988*, 2018.

[5] Y. Dong. An extension of Luque's growth condition. *Appl. Math. Lett.*, 22(9):1390–1393, 2009.

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

[8] E. Hazan and S. Kale. Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization. In S.M. Kakade and U. von Luxburg, editors, *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proc. of Machine Learning Res.*, pages 421–436, Budapest, Hungary, 09–11 Jun 2011. PMLR.

[9] A. Juditsky and Y. Nesterov. Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization. *Stoch. Syst.*, 4(1):44–80, 2014.

[10] 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. *arXiv:1212.2002*, 2012.

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

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

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

[14] A. Rakhlin, O. Shamir, and K. Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pages 1571–1578, USA, 2012.