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 $\tilde{O}(k^{-1/2})$.

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 $\varepsilon$-stationary point in $\tilde{O}(\varepsilon^{-1/2})$ iterations. 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 [11], Allen-Zhu investigated the complexity of finding $\varepsilon$-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 $\tilde{O}(\varepsilon^{-2})$.

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.

1In this section to simplify notation, we only show dependence on the accuracy $\varepsilon$ 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

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

(1.1)

where $\mathcal{X} \subseteq \mathbb{R}^d$ is a closed convex set with a computable nearest-point map $\text{proj}_{\mathcal{X}}$, and $g : \mathbb{R}^d \to \mathbb{R}$ 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 $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ 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 $\varepsilon$-stationary points, in the sense of $\text{dist}(0; \partial \varphi(x)) \leq \varepsilon$. Instead we focus on finding approximately stationary points of the Moreau envelope:

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

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

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

(1.2)

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

$$\text{prox}_{\lambda \varphi}(x) := \arg\min_{y \in \mathbb{R}^d} \{ \varphi(y) + \frac{1}{2\lambda} \|y - x\|^2 \}. $$

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 $\hat{x} := \text{prox}_{\lambda \varphi}(x)$ satisfies

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

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 $E\|\nabla \varphi_{\lambda}(x)\| \leq \varepsilon$ 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 $\tilde{O}(\varepsilon^{-2})$ 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 \(\sigma\)-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 \(X\) and a measurable mapping \(G: U \times \Omega \rightarrow \mathbb{R}^d\) satisfying \(E_\xi [G(x, \xi)] \in \partial g(x)\) for all \(x \in U\).

(A3) There is a real \(L \geq 0\) such that the inequality, \(E_\xi [\|G(x, \xi)\|^2] \leq L^2\), holds for all \(x \in X\).

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 \(X\) is bounded by some real \(D > 0\). It was shown in [4, Section 2.1] that the complexity of finding a point \(x\) satisfying \(E_\xi \|\nabla \phi_1/\rho (x)\| \leq \epsilon\) is at most \(O(1) \cdot (L^2 + \epsilon^2) \sqrt{\rho D} \epsilon^2\). We will see here that this complexity can be improved to \(\tilde{O}(L^2 + \rho^2 D^2 \epsilon^2)\) by adapting the technique of [1].

The workhorse 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

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

**Step** \(t = 0, \ldots, T - 2:\)

\[
\begin{align*}
&\text{Sample } \xi_t \sim dP \\
&\text{Set } x_{t+1} = \text{proj}_X \left( x_t - \frac{2}{\mu(t+1)} : G(x_t, \xi_t) \right),
\end{align*}
\]

**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

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

For the time being, let us assume that \(g\) is \(\mu\)-strongly convex on \(X\). 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  
GR$^{\text{sc}}(x_1, \mu, \lambda, T, \mathcal{I}, G)$

| 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 $\hat{x}_{i+1} = \text{PSSM}^{\text{sc}}(\hat{x}_i, \sum_{j=0}^{i} \mu_j, G^{(i)}, T)$ |
| $\mu_{i+1} = \mu \cdot 2^{i+1}$ |
| Define the function and the oracle $\varphi^{(i+1)}(x) := \varphi^{(i)}(x) + \frac{\mu_{i+1}}{2} \|x - \hat{x}_{i+1}\|^2$ and $G^{(i+1)}(x, \xi) := G^{(i)}(x, \xi) + \mu_{i+1}(x - \hat{x}_{i+1})$. |
| 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 $\mu_i$, $\varphi^{(i)}$, and $\hat{x}_i$ be generated by Algorithm 2. Observe that by construction, equality

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

holds for all $i = 1, \ldots, \mathcal{I}$. Consequently, it will be important to relate the Moreau envelope of $\varphi^{(i)}$ to that of $\varphi$. 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. \qed

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

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

4
for some \( z_i \in \mathbb{R}^d \) and \( a_i > 0 \), with \( i = 1, \ldots, I \). Then for any \( \lambda > 0 \), the Moreau envelopes of \( h \) and \( f \) are related by the expression

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

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

**Proof.** By definition of the Moreau envelope, we have

\[
f_{1/\lambda}(x) = \arg\min_y \left\{ h(y) + \sum_{i=1}^{I} \frac{a_i}{2} \| y - z_i \|^2 + \frac{\lambda}{2} \| y - x \|^2 \right\}.
\]

(1.3)

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

\[
Q(y) := \frac{\lambda}{2} \| y - x \|^2 + \sum_{i=1}^{I} \frac{a_i}{2} \| y - z_i \|^2.
\]

Lemma 1.2 directly yields the representation \( Q(y) = Q(\bar{x}) + \frac{\lambda + A}{\lambda + A} \| y - \bar{x} \|^2 \). 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

\[
\nabla f_{1/\lambda}(x) = \frac{\lambda}{\lambda + A} \nabla h_{1/(\lambda + A)}(\bar{x}) + \lambda \left( \frac{\lambda}{\lambda + A} - 1 \right) (\bar{x} - x) + \frac{\lambda}{\lambda + A} \sum_{i=1}^{I} a_i (\bar{x} - z_i)
\]

\[
= \frac{\lambda}{\lambda + A} \nabla h_{1/(\lambda + A)}(\bar{x}) + \frac{\lambda}{\lambda + A} \sum_{i=1}^{I} a_i (x - z_i),
\]

as claimed. \( \square \)

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

**Lemma 1.4.** Suppose that for each index \( i = 1, 2, \ldots, I \), 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:

\[
\mathbb{E} \left[ \sum_{i=1}^{I} \mu_i \| x^*_I - \hat{x}_i \|^2 \right] \leq 4 \sum_{i=1}^{I} \sqrt{\delta_i \mu_i},
\]

where \( x^*_I \) is the minimizer of \( \varphi^I \).
Henceforth, set
\[ M_i := \sum_{j=1}^{i} \mu_j \quad \text{and} \quad M := M_T. \]

By convention, we will set \( M_0 = 0 \). Combining Lemmas \ref{lem:1.3} and \ref{lem:1.4}, we arrive at the following basic guarantee of the method.

**Corollary 1.5.** Suppose for \( i = 1, 2, \ldots, I + 1 \), 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:
\[ \mathbb{E}\|\nabla \varphi_1/(\lambda + M)(\bar{x})\| \leq (\lambda + 2M) \sqrt{\frac{2\delta_I + 1}{\mu + M}} + 4 \sum_{i=1}^{T} \sqrt{\delta_i \mu_i}, \]
where \( \bar{x} = \frac{1}{\lambda + M} (\lambda x + \sum_{i=1}^{T} \mu_i \hat{x}_i) \).

**Proof.** Fix an arbitrary point \( x \) and set \( \bar{x} = \frac{1}{\lambda + M} (\lambda x + \sum_{i=1}^{T} \mu_i \hat{x}_i) \). Then Lemma \ref{lem:1.3} along with a triangle inequality, directly implies
\[ \| \nabla \varphi_1/(\lambda + M)(\bar{x}) \| \leq (1 + \frac{M}{\lambda}) \| \nabla \varphi_1/(\lambda + M)(x) \| + \sum_{i=1}^{T} \mu_i \| x - \hat{x}_i \| \]
\[ \quad \leq (1 + \frac{M}{\lambda}) \| \nabla \varphi_1/(\lambda + M)(x) \| + \sum_{i=1}^{T} \mu_i (\| x - x_*^i \| + \| x_*^i - \hat{x}_i \|) \]
\[ \quad \leq (1 + \frac{M}{\lambda}) \| \nabla \varphi_1/(\lambda + M)(x) \| + M \| x - x_*^i \| + \sum_{i=1}^{T} \mu_i \| x_*^i - \hat{x}_i \| \]
\[ \quad \leq (\lambda + 2M) \| x - x_*^i \| + \sum_{i=1}^{T} \mu_i \| x_*^i - \hat{x}_i \|. \]

where the last inequality uses that \( \nabla \varphi_1/(\lambda + M) \) is \( \lambda \)-Lipschitz continuous and \( \nabla \varphi_1/(\lambda + M)(x_*^i) = 0 \) to deduce that \( \| \nabla \varphi_1/(\lambda + M)(x) \| \leq \lambda \| x - x_*^i \| \). Using strong convexity of \( \varphi^T \), we deduce
\[ \| x - x_*^i \|^2 \leq \frac{2}{\mu + M} (\varphi^T(x) - \varphi^T(x_*^i)). \]

Setting \( x = \hat{x}_{i+1} \), taking expectations, and applying Lemma \ref{lem:1.4} completes the proof. \( \square \)

Let us now determine \( \delta_i > 0 \) by invoking Theorem \ref{thm:1.4} for each function \( \varphi^{(i)} \). Observe
\[ \mathbb{E}_x \| G^{(i)}(x, \xi) \|^2 \leq 2(L^2 + D^2 M_i^2). \]
Thus Theorem \ref{thm:1.4} guarantees the estimates:
\[ \mathbb{E}[\varphi^{(i-1)}(\hat{x}_i) - \min \varphi^{(i-1)}] \leq \frac{4(L^2 + D^2 M_{i-1}^2)}{(T + 1)(\mu + M_{i-1})}. \]
Hence for $i = 1, \ldots, I$, we may set $\delta_i$ to be the right-hand side of (1.4). Applying Corollary 1.5, we therefore deduce

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

\[
\leq \frac{1}{\sqrt{T+1}} \left( (\lambda + 2M) \sqrt{\frac{8(L^2 + D^2 M^2)}{(\mu + M)^2}} + 4 \sum_{i=1}^{I} \sqrt{\frac{4(L^2 + D^2 M_i^2)}{(\mu + M_{i-1}) \mu_i}} \right).
\]

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

\[
\frac{\mu_i}{\mu + M_{i-1}} \leq \frac{\mu_i}{\mu + \mu_{i-1}} = \frac{2^i}{1 + 2^{i-1}} \leq 2.
\]

Hence, continuing (1.5), we conclude

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

In particular, by setting $I = \log_2(1 + \frac{\lambda}{2\mu})$, 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 $\mu$-strongly convex on $\mathcal{X}$ and we set $I = \log_2(1 + \frac{\lambda}{2\mu})$ for some $\lambda > 0$. Then $\bar{x}$ returned by Algorithm 2 satisfies

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

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 $I \in \mathbb{N}$, stochastic oracle $G(\cdot, \cdot)$.

Set $\tilde{\varphi}(x) := \varphi(x) + \frac{\mu}{2} \| x - x_c \|^2$, $\tilde{G}(x, \xi) = G(x, \xi) + \mu(x - x_c)$, $x_0 = x_c$.

Set $\bar{x} = \text{GR}^\alpha(x_c, \mu, \lambda/2, T, I, \tilde{G})$

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 $\varepsilon \leq 2\rho D$. Set $\mu := \frac{\varepsilon}{2D}$, $\lambda := 2\rho - \frac{\varepsilon}{2D}$, and $I = \log_2\left( \frac{3}{4} + \frac{\rho D}{\varepsilon} \right)$. Then for any $T > 0$, Algorithm 3 returns a point $\bar{z}$ satisfying:

\[
\mathbb{E} \| \nabla \varphi_{1/(2\rho)}(\bar{z}) \| \leq \frac{(28\sqrt{2} \cdot \log_2\left( \frac{3}{4} + \frac{\rho D}{\varepsilon} \right)) \cdot \sqrt{2L^2 + 3\rho^2 D^2}}{\sqrt{T+1}} + \frac{\varepsilon}{2}
\]
Setting the right hand side to $\varepsilon$ and solving for $T$, we deduce that it suffices to make

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

calls to $\text{proj}_X$ and to the stochastic subgradient oracle in order to find a point $\bar{z} \in X$ satisfying $E\|\nabla \varphi_1/(2\rho)(\bar{z})\| \leq \varepsilon$.

Proof. Lemma 1.3 guarantees the bound

$$\left\| \nabla \varphi_1/(\lambda + \mu) \left( \frac{\mu}{\mu + \lambda} x_c + \frac{1}{\mu + \lambda} \bar{x} \right) \right\| \leq \frac{\lambda + \mu}{\lambda} \| \nabla \hat{\varphi}_1/\lambda(\bar{x}) \| + \mu D.$$  

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

$$E \left\| \nabla \varphi_1/(2\rho)(\bar{z}) \right\| \leq \frac{\lambda + \mu}{\lambda} \left( 14\sqrt{2} \log_2 \left( 1 + \frac{\lambda}{4\mu} \right) \right) \sqrt{\frac{2(L^2 + D^2 \mu^2) + \frac{5}{4} D^2 \lambda^2}{T+1}} + \frac{\varepsilon}{2}.$$  

Some elementary simplifications yield the result. \hfill \square

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 Conference on International Conference on Machine Learning*, ICML’12, pages 1571–1578, USA, 2012.