SIAM J. OPTIM.
Vol. 29, No. 2, pp. 1350–1365

CONVERGENCE RATES FOR DETERMINISTIC AND STOCHASTIC SUBGRADIENT METHODS WITHOUT LIPSCHITZ CONTINUITY[*]

BENJAMIN GRIMMER[†]

Abstract. We extend the classic convergence rate theory for subgradient methods to apply to non-Lipschitz functions. For the deterministic projected subgradient method, we present a global $O(1/\sqrt{T})$ convergence rate for any convex function which is locally Lipschitz around its minimizers. This approach is based on Shor's classic subgradient analysis and implies generalizations of the standard convergence rates for gradient descent on functions with Lipschitz or Hölder continuous gradients. Further, we show a $O(1/\sqrt{T})$ convergence rate for the stochastic projected subgradient method on convex functions with at most quadratic growth, which improves to $O(1/T)$ under either strong convexity or a weaker quadratic lower bound condition.

Key words. convex optimization, subgradient method, convergence, non-Lipschitz optimization

AMS subject classifications. 90C25, 90C52, 65K15

DOI. 10.1137/18M117306X

1. Introduction.

We consider the nonsmooth, convex optimization problem given by

$$\min_{x \in Q} f(x)$$

for some lower semicontinuous convex function $f\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ and closed convex feasible region Q. We assume Q lies in the domain of f and that this problem has a nonempty set of minimizers X^* (with minimum value denoted by f^*). Further, we assume that orthogonal projection onto Q, which we denote by $P_Q(\cdot)$, is computationally tractable.

Since f may be nondifferentiable, we weaken the notion of gradients to subgradients. The set of all subgradients at some $x \in Q$ (referred to as the subdifferential) is denoted by

$$\partial f(x) = \{g \in \mathbb{R}^d \mid \left(\forall y \in \mathbb{R}^d\right)\ f(y) \geq f(x) + g^T(y - x)\}.$$

We consider solving this problem via a (potentially stochastic) projected subgradient method. These methods have received much attention lately due to their simplicity and scalability; see [2, 16], as well as [9, 10, 11, 14, 17] for a sample of more recent works.

Deterministic and stochastic subgradient methods differ in the type of oracle used to access the subdifferential of f. For deterministic methods, we consider an oracle $g(x)$, which returns an arbitrary subgradient at x. For stochastic methods, we utilize a weaker, random oracle $g(x; \xi)$, which is an unbiased estimator of a subgradient

[*]Received by the editors February 27, 2018; accepted for publication (in revised form) March 19, 2019; published electronically May 14, 2019.

http://www.siam.org/journals/siopt/29-2/M117306.html

Funding: This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under grant DGE-1650441. This work was done while the author was visiting the Simons Institute for the Theory of Computing. It was partially supported by the DIMACS/Simons Collaboration on Bridging Continuous and Discrete Optimization through NSF grant CCF-1740425.

[†]Operations Research and Information Engineering, Cornell University, Ithaca, NY 14850-3711 (bdg79@cornell.edu).

(i.e., $\mathbb{E}_{\xi\sim D}\ g(x;\xi) \in \partial f(x)$ for some easily sampled distribution D). One of the earliest works motivating stochastic gradient methods was Robbins and Monro [19]. An overview of the early work analyzing stochastic subgradient methods in optimization is given by Shor [22, section 2.4] and the references therein.

We analyze two classic subgradient methods, differing in their step-size policy. Let $\|\cdot\|$ denote the Euclidean norm on $\mathbb{R}^d$. Given a deterministic oracle, we consider the normalized subgradient method

$$x_{k+1} := P_Q\left(x_k - \alpha_k \frac{g(x_k)}{\|g(x_k)\|}\right) \tag{1.1}$$

for some positive sequence $(\alpha_k)_{k=0}^T$. Note that since $\|g(x_k)\| = 0$ only if x_k minimizes f, this iteration is well defined until a minimizer is found. Given a stochastic oracle, we consider the method

$$x_{k+1} := P_Q\left(x_k - \alpha_k g(x_k;\xi_k)\right) \tag{1.2}$$

for some positive sequence $(\alpha_k)_{k=0}^T$ and an i.i.d. sample sequence $\xi_k \sim D$.

The standard convergence bounds for these methods assume all $x \in Q$ satisfy $\|g(x)\| \leq L$ or $\mathbb{E}_\xi\|g(x;\xi)\|^2 \leq L^2$ for some constant $L > 0$. Then after $T > 0$ iterations, a point is found with objective gap (in expectation for (1.2)) bounded by

$$f(x) - f^* \leq O\left(\frac{L\|x_0 - x^*\|}{\sqrt{T}}\right) \tag{1.3}$$

for any $x^* \in X^*$ under reasonable selection of the step-size sequence $(\alpha_k)_{k=0}^T$.

The bound $\|g(x)\| \leq L$ for all $x \in Q$ is implied by f being L-Lipschitz continuous on some open convex set U containing Q (which is often the assumption made). Uniformly bounding subgradients restricts the classic convergence rates to functions with at most linear growth (at rate L). When Q is bounded, one can invoke a compactness argument to produce a uniform Lipschitz constant. However, such an approach may lead to a large constant heavily dependent on the size of Q (and, frankly, lacks the elegance that such a fundamental method deserves).

In stark contrast to these limitations, early in the development of subgradient methods Shor [21] observed that the normalized subgradient method (1.1) enjoys some form of convergence guarantee for any convex function with a nonempty set of minimizers. Shor showed, for any minimizer $x^* \in X^*$, that there exists some iterate $k \leq T$ for which either $x_k \in X^*$ or

$$\left(\frac{g(x_k)}{\|g(x_k)\|}\right)^T (x_k - x^*) \leq O\left(\frac{\|x_0 - x^*\|}{\sqrt{T}}\right)$$

under reasonable selection of the step-size sequence $(\alpha_k)_{k=0}^T$. Thus, for any convex function, the subgradient method has convergence in terms of this inner product value (which convexity implies is always nonnegative). This quantity can be interpreted as the distance from the hyperplane $\{x \mid g(x_k)^T(x - x_k) = 0\}$ to x^*. By driving this distance to zero via proper selection of $(\alpha_k)_{k=0}^T$, Shor characterized the asymptotic convergence of (1.1).

There is a substantial discrepancy in generality between the standard convergence bound (1.3) and Shor's result. In this paper, we address this for both deterministic and stochastic subgradient methods. In the remainder of this section we formally state our generalized convergence rate bounds. For the deterministic case our bounds follow directly from Shor's result, while the stochastic case requires an alternative approach. Then in section 2 we apply these bounds to a few common problem classes outside the scope of uniform Lipschitz continuity. Finally, our convergence analysis is presented in section 3 and an extension of our model is discussed in section 4.

1.1. Extended deterministic convergence bounds.

Shor's convergence guarantees for general convex functions will serve as the basis of our objective gap convergence rates for the subgradient method (1.1) without assuming uniform Lipschitz continuity. Formally, Shor [21] showed the following general guarantee for any sequence of step sizes $(\alpha_k)_{k=0}^T$ (for completeness, an elementary proof is provided in section 3).

THEOREM 1.1 (Shor's hyperplane distance convergence). *Consider any convex f and fix some $x^* \in X^*$. Then for any positive sequence $(\alpha_k)_{k=0}^T$, there exists some iterate $k \leq T$ of the iteration* (1.1) *for which either $x_k \in X^*$ or*

$$\left(\frac{g(x_k)}{\|g(x_k)\|}\right)^T (x_k - x^*) \leq \frac{\|x_0 - x^*\|^2 + \sum_{k=0}^T \alpha_k^2}{2\sum_{k=0}^T \alpha_k}. \tag{1.4}$$

Two simple step-size selections. The classic objective gap convergence of the subgradient method follows as a simple consequence of this. Indeed, f being convex and L-Lipschitz continuous on an open set containing Q (which implies $\|g(x_k)\| \leq L$) together with (1.4) imply

$$\min_{k=0,\ldots,T} \left\{\frac{f(x_k) - f^*}{L}\right\} \leq \frac{\|x_0 - x^*\|^2 + \sum_{k=0}^T \alpha_k^2}{2\sum_{k=0}^T \alpha_k}.$$

Given either an upper bound[1] $R \geq \|x_0 - x^*\|$ or a target accuracy $\epsilon > 0$, a convergence rate follows for either of the two following choices of the step-size sequence $(\alpha_k)_{k=0}^T$. Taking $\alpha_k = R/\sqrt{T+1}$ produces

$$\min_{k=0,\ldots,T} \{f(x_k) - f^*\} \leq \frac{LR}{\sqrt{T+1}}.$$

Alternatively, taking $\alpha_k = \epsilon/L$ yields

$$T \geq \left(\frac{L\|x_0 - x^*\|}{\epsilon}\right)^2 \implies \min_{k=0,\ldots,T} \{f(x_k) - f^*\} \leq \epsilon.$$

Here Lipschitz continuity enabled us to convert a bound on "hyperplane distance to a minimizer" into a bound on the objective gap. Our extended convergence bounds for the deterministic subgradient method follow from observing that more general assumptions than uniform Lipschitz continuity suffice to provide such a conversion. In particular, we assume there is an upper bound on f of the form

$$f(x) - f^* \leq \mathcal{D}(\|x - x^*\|) \quad (\forall x \in \mathbb{R}^d) \tag{1.5}$$

for some fixed $x^* \in X^*$ and nondecreasing nonnegative function $\mathcal{D}\colon \mathbb{R}_+ \to \mathbb{R}_+ \cup \{\infty\}$. In this case, we show the following objective gap convergence guarantee.

[1] Note that an upper bound R can often be produced when Q is simple and bounded or when f possesses some structural property like strong convexity.

THEOREM 1.2 (extended deterministic rate). *Consider any convex f satisfying* (1.5). *Then for any positive sequence $(\alpha_k)_{k=0}^T$, the iteration* (1.1) *satisfies*

$$\min_{k=0,\ldots,T} \{f(x_k) - f^*\} \leq \mathcal{D}\left(\frac{\|x_0 - x^*\|^2 + \sum_{k=0}^T \alpha_k^2}{2\sum_{k=0}^T \alpha_k}\right).$$

Two simple step-size selections. Suppose either an upper bound $R \geq \|x_0 - x^*\|$ or a target accuracy $\epsilon > 0$ is known. Under the constant step size $\alpha_k = R/\sqrt{T+1}$, the iteration (1.1) satisfies

$$\min_{k=0,\ldots,T} \{f(x_k) - f^*\} \leq \mathcal{D}\left(\frac{R}{\sqrt{T+1}}\right).$$

Under the constant step size $\alpha_k = \mathcal{D}^{-1}(\epsilon)$, the iteration (1.1) satisfies

$$T \geq \left(\frac{\|x_0 - x^*\|}{\mathcal{D}^{-1}(\epsilon)}\right)^2 \implies \min_{k=0,\ldots,T} \{f(x_k) - f^*\} \leq \epsilon,$$

where $\mathcal{D}^{-1}(\epsilon) = \inf\{t \mid \mathcal{D}(t) \geq \epsilon\}$.

Note that any L-Lipschitz continuous function on $\mathbb{R}^d$ satisfies this growth bound with $\mathcal{D}(t) = Lt$. Thus we immediately recover the standard $L\|x_0 - x^*\|/\sqrt{T}$ convergence rate for unconstrained problems. Similarly, any L-Lipschitz continuous function on an open neighborhood of Q satisfies this growth bound with

$$\mathcal{D}(t) = \begin{cases} Lt & \text{if } t \leq \delta, \\ +\infty & \text{otherwise} \end{cases}$$

for some $\delta > 0$. From this, we recover the standard $L\|x_0 - x^*\|/\sqrt{T}$ rate for constrained problems provided $T \geq \|x_0 - x^*\|^2/\delta^2$.

Using growth bounds allows us to apply our convergence guarantees to many problems outside the scope of uniform Lipschitz continuity. Theorem 1.2 also implies the classic convergence rate for gradient descent on differentiable functions with an L-Lipschitz continuous gradient of $O(L\|x_0 - x^*\|^2/T)$ [16]. Any such function has growth bounded by $\mathcal{D}(t) = Lt^2/2$ on $Q = \mathbb{R}^d$ (see Lemma 2.1). Then a convergence rate immediately follows from Theorem 1.2 (for simplicity, we consider a constant step size given an upper bound $R \geq \|x_0 - x^*\|$).

COROLLARY 1.3 (generalizing gradient descent's convergence). *Consider any convex function f satisfying* (1.5) *with $\mathcal{D}(t) = Lt^2/2$. Then under the constant step size $\alpha_k = R/\sqrt{T+1}$, the iteration* (1.1) *satisfies*

$$\min_{k=0,\ldots,T} \{f(x_k) - f^*\} \leq \frac{LR^2}{2(T+1)}.$$

Thus a convergence rate of $O(LR^2/T)$ can be attained without any mention of smoothness or differentiability. In section 2, we provide a similar growth bound and thus objective gap convergence for any function with a Hölder continuous gradient, which also parallels the standard rate for gradient descent. In general, for any problem with $\lim_{t\to 0^+} \mathcal{D}(t)/t = 0$, Theorem 1.2 produces convergence at a rate of $o(1/\sqrt{T})$.

Suppose that $\mathcal{D}(t)/t$ is finite in some neighborhood of 0 (as is the case for any f that is locally Lipschitz around x^*). Then simple limiting arguments yield the following eventual convergence rate of (1.1) based on Theorem 1.2: for any $\epsilon > 0$, there exists $T_0 > 0$ such that all $T > T_0$ have

$$\min_{k=0\ldots T} \{f(x_k) - f^*\} \leq \mathcal{D}\left(\frac{R}{\sqrt{T+1}}\right) \leq \left(\limsup_{t\to 0^+} \frac{\mathcal{D}(t)}{t} + \epsilon\right) \frac{R}{\sqrt{T+1}}.$$

As a result, the asymptotic convergence rate of (1.1) is determined entirely by the rate of growth of f around its minimizers, and conversely, steepness far from optimality plays no role in the asymptotic behavior.

1.2. Extended stochastic convergence bounds.

Now we turn our attention to giving more general convergence bounds for the stochastic subgradient method. This is harder as we can no longer leverage Shor's result, since normalizing stochastic subgradients may introduce bias or may not be well defined if $g(x_k;\xi) = 0$. As a consequence, we need a different approach to generalizing the standard stochastic assumptions.

We begin by reviewing the standard convergence results for this method. The following convergence guarantee is immediate from the analysis given in [22, section 2.4] and is well known in the literature.

Theorem 1.4 (classic stochastic rate). *Consider any convex function f and stochastic subgradient oracle satisfying $\mathbb{E}_\xi\|g(x;\xi)\|^2 \leq L^2$ for all $x \in Q$. Fix some $x^* \in X^*$. Then for any positive sequence $(\alpha_k)_{k=0}^T$, the iteration (1.2) satisfies*

$$\mathbb{E}_{\xi_{0,\ldots,T}}\left[f\left(\frac{\sum_{k=0}^T \alpha_k x_k}{\sum_{k=0}^T \alpha_k}\right) - f^*\right] \leq \frac{\|x_0 - x^*\|^2 + L^2\sum_{k=0}^T \alpha_k^2}{2\sum_{k=0}^T \alpha_k}.$$

Two simple step-size selections. Similar to the deterministic setting, given either an upper bound $R \geq \|x_0 - x^*\|$ or a target accuracy $\epsilon > 0$, simple constant step sizes can be analyzed. Under the selection $\alpha_k = R/(L\sqrt{T+1})$, the iteration (1.2) satisfies

$$\mathbb{E}_{\xi_{0,\ldots,T}}\left[f\left(\frac{1}{T+1}\sum_{k=0}^T x_k\right) - f^*\right] \leq \frac{LR}{\sqrt{T+1}}.$$

Under the selection $\alpha_k = \epsilon/L^2$, the iteration (1.2) satisfies

$$T \geq \left(\frac{L\|x_0 - x^*\|}{\epsilon}\right)^2 \implies \mathbb{E}_{\xi_{0,\ldots,T}}\left[f\left(\frac{1}{T+1}\sum_{k=0}^T x_k\right) - f^*\right] \leq \epsilon.$$

We say f is μ-strongly convex on Q for some $\mu > 0$ if for every $x \in Q$ and $g \in \partial f(x)$,

$$f(y) \geq f(x) + g^T(y-x) + \frac{\mu}{2}\|y-x\|^2 \quad (\forall y \in Q).$$

Under this condition, the convergence of (1.2) can be improved to $O(1/T)$ [9, 10, 17]. Below, we present one such bound from [10].

Theorem 1.5 (classic strongly convex stochastic rate). *Consider any μ-strongly convex function f and stochastic subgradient oracle satisfying $\mathbb{E}_\xi\|g(x;\xi)\|^2 \leq L^2$ for all $x \in Q$. Then for the sequence of step sizes $\alpha_k = 2/\mu(k+2)$, the iteration (1.2) satisfies*

$$\mathbb{E}_{\xi_{0,\ldots,T}}\left[f\left(\frac{2}{(T+1)(T+2)}\sum_{k=0}^T (k+1)x_k\right) - f^*\right] \leq \frac{2L^2}{\mu(T+2)}.$$

We remark that Lipschitz continuity and strong convexity are fundamentally at odds. Lipschitz continuity allows at most linear growth while strong convexity requires quadratic growth. The only way both can occur is when Q is bounded.

The standard analysis assumes that $\mathbb{E}_\xi\|g(x;\xi)\|^2$ is uniformly bounded by some $L^2 > 0$. We generalize this by allowing the expectation to be larger when the objective gap at x is large as well. In particular, we assume a bound of the form

$$\mathbb{E}_\xi\|g(x;\xi)\|^2 \leq L_0^2 + L_1(f(x) - f^*) \tag{1.6}$$

for some constants $L_0, L_1 \geq 0$. When L_1 equals zero, this is exactly the classic model. When L_1 is positive, this model allows functions with up to quadratic growth. (To see this, suppose the subgradient oracle is deterministic. Then (1.6) corresponds to a differential inequality of the form $\|\nabla f(x)\| \leq \sqrt{L_1(f(x) - f^*) + L_0^2}$, which has a simple quadratic solution. This interpretation is formalized in section 2.4.)

The additional generality allowed by (1.6) is important for two reasons. First, it allows us to consider many classic problems which fundamentally have quadratic growth (for example, any quadratically regularized problem, like training a support vector machine, which is considered in section 2.3). Secondly, this model allows us to avoid the inherent conflict in Theorem 1.5 between Lipschitz continuity and strong convexity since a function can globally satisfy both (1.6) and strong convexity.

Based on this generalization of Lipschitz continuity, we have the following guarantees for convex and strongly convex problems.

THEOREM 1.6 (extended stochastic rate). *Consider any convex function f and stochastic subgradient oracle satisfying* (1.6). *Fix some $x^* \in X^*$. Then for any positive sequence $(\alpha_k)_{k=0}^T$ with $L_1\alpha_k < 2$ for all $k = 0, \dots, T$, the iteration* (1.2) *satisfies*

$$\mathbb{E}_{\xi_{0,\dots,T}}\left[f\left(\frac{\sum_{k=0}^T \alpha_k(2 - L_1\alpha_k)x_k}{\sum_{k=0}^T \alpha_k(2 - L_1\alpha_k)}\right) - f^*\right] \leq \frac{\|x_0 - x^*\|^2 + L_0^2\sum_{k=0}^T \alpha_k^2}{\sum_{k=0}^T \alpha_k(2 - L_1\alpha_k)}.$$

Two simple step-size selections. Given either an upper bound $R \geq \|x_0 - x^*\|$ or a target accuracy $\epsilon > 0$, we present bounds for two simple constant step sizes. Under the selection $\alpha_k = R/L_0\sqrt{T+1}$, the iteration (1.2) satisfies

$$\mathbb{E}_{\xi_{0,\dots,T}}\left[f\left(\frac{1}{T+1}\sum_{k=0}^T x_k\right) - f^*\right] \leq \frac{L_0R}{\sqrt{T+1}} + \frac{L_1R^2}{T+1},$$

provided $T \geq (RL_1/L_0)^2$. Under the selection $\alpha_k = \epsilon/(2L_0^2)$, the iteration (1.2) satisfies

$$T \geq \left(\frac{L_0\|x_0 - x^*\|}{\epsilon}\right)^2 \implies \mathbb{E}_{\xi_{0,\dots,T}}\left[f\left(\frac{1}{T+1}\sum_{k=0}^T x_k\right) - f^*\right] \leq \epsilon,$$

provided $\epsilon \leq 2L_0^2/L_1$.

THEOREM 1.7 (extended strongly convex stochastic rate[2]). *Consider any μ-strongly convex function f and stochastic subgradient oracle satisfying* (1.6). *Fix*

[2] A predecessor of Theorem 1.7 was given by Davis and Grimmer in Proposition 3.2 of [5], where an $O(\log(T)/T)$ convergence rate was shown for certain non-Lipschitz strongly convex problems.

some $x^* \in X^*$. Then for the sequence of step sizes

$$\alpha_k = \frac{2}{\mu(k+2) + \frac{L_1^2}{\mu(k+1)}},$$

the iteration (1.2) *satisfies*

$$\mathbb{E}_{\xi_{0,\dots,T}}\left[f\left(\frac{\sum_{k=0}^T (k+1)(2-L_1\alpha_k)x_k}{\sum_{k=0}^T (k+1)(2-L_1\alpha_k)}\right) - f^*\right] \le \frac{2L_0^2(T+1) + L_1^2\|x_0 - x^*\|^2/2}{\mu\sum_{k=0}^T (k+1)(2-L_1\alpha_k)}.$$

The following simpler averaging yields a bound weakened roughly by a factor of two:

$$\mathbb{E}_{\xi_{0,\dots,T}}\left[f\left(\frac{2}{(T+1)(T+2)}\sum_{k=0}^T (k+1)x_k\right) - f^*\right] \le \frac{4L_0^2}{\mu(T+2)} + \frac{L_1^2\|x_0 - x^*\|^2}{\mu(T+1)(T+2)}.$$

We remark that one important insight given by Theorem 1.7 comes from its dependence on the initial point x_0. The rate only depends on the initial iterate in the second term above, which decays at a rate of $O(1/T^2)$. This shows the choice of the initial point has a relatively small impact on the asymptotic guarantees. Note that the standard analysis of this method (see Theorem 1.5) does not give any insight into the dependence on x_0 and instead uses the implicit bound on $\|x_0 - x^*\|$ given by strong convexity and Lipschitz continuity.

1.3. Related works.

Recently, Renegar [18] introduced a novel framework that allows first-order methods to be applied to general (non-Lipschitz) convex optimization problems via a radial transformation. Based on this framework, Grimmer [8] showed that a simple radial subgradient method has convergence paralleling the classic $O(1/\sqrt{T})$ rate without assuming Lipschitz continuity. This algorithm is applied to a transformed version of the original problem and replaces orthogonal projection by a line search at each iteration.

Lu [11] analyzes an interesting subgradient-type method (which is a variation of mirror descent) for non-Lipschitz problems that is customized for a particular problem via a reference function. This approach produces convergence guarantees for both deterministic and stochastic problems based on a relative-continuity constant instead of a uniform Lipschitz constant.

Although the works of Renegar [18], Grimmer [8], and Lu [11] provide convergence rates for specialized subgradient methods without assuming Lipschitz continuity, objective gap guarantees for the classic subgradient methods (1.1) and (1.2), such as the ones in the present paper, have been missing prior to our work.

2. Applications of our extended convergence bounds.

In this section, we apply our convergence bounds to a variety of problems outside the scope of the traditional theory based on uniform Lipschitz constants.

2.1. Smooth optimization.

The standard analysis of gradient descent for smooth optimization assumes the gradient of the objective function is uniformly Lipschitz continuous, or more generally, uniformly Hölder continuous. A differentiable function f has (L, v)-Hölder continuous gradient on $\mathbb{R}^d$ for some $L > 0$ and $v \in (0, 1]$ if, for all $x, y \in \mathbb{R}^d$,

$$\|\nabla f(y) - \nabla f(x)\| \le L\|y - x\|^v.$$

Note this is exactly Lipschitz continuity of the gradient when $v = 1$. Below, we state a simple bound on the growth $\mathcal{D}(t)$ of any such function.

Lemma 2.1. *Consider any $f \in C^1$ with a (L, v)-Hölder continuous gradient on $\mathbb{R}^d$ and any minimizer $x^* \in X^*$. Then*

$$f(x) - f(x^*) \le \frac{L}{v+1}\|x - x^*\|^{v+1} \quad (\forall x \in \mathbb{R}^d).$$

Proof. Since $\nabla f(x^*) = 0$, the bound follows directly as

$$\begin{aligned} f(x) &= f(x^*) + \int_0^1 \nabla f(x^* + t(x - x^*))^T (x - x^*)\ dt \\ &\le f(x^*) + \nabla f(x^*)^T (x - x^*) + \int_0^1 Lt^v \|x - x^*\|^{v+1}\ dt \\ &= f(x^*) + \frac{L}{v+1}\|x - x^*\|^{v+1}. \end{aligned}$$ □

This lemma with $v = 1$ implies any function with an L-Lipschitz gradient has growth bounded by $\mathcal{D}(t) = Lt^2/2$. Then Theorem 1.2 produces our generalization of the classic gradient descent convergence rate claimed in Corollary 1.3. Further, for any function with a Hölderian gradient, Theorem 1.2 gives an $O(1/T^{(v+1)/2})$ convergence rate. The following Corollary generalizes this fact giving a convergence rate for any (potentially nondifferentiable) function with upper bound $\mathcal{D}(t) = Lt^{v+1}/(v + 1)$.

Corollary 2.2 (generalizing Hölder gradient descent's convergence). *Consider any convex function f satisfying* (1.5) *with $\mathcal{D}(t) = Lt^{v+1}/(v + 1)$. Then under the constant step size $\alpha_k = R/\sqrt{T+1}$, the iteration* (1.1) *satisfies*

$$\min_{k=0,\dots,T} \{f(x_k) - f^*\} \le \frac{LR^{v+1}}{(v + 1)(T + 1)^{(v+1)/2}}.$$

2.2. Additive composite optimization. Often problems arise where the objective is to minimize a sum of smooth and nonsmooth functions. We consider the general formulation of this problem

$$\min_{x\in\mathbb{R}^d} f(x) := \Phi(x) + h(x)$$

for any differentiable convex function Φ with (L_Φ, v)-Hölderian gradient and any L_h-Lipschitz continuous convex function h. Such problems occur when regularizing smooth optimization problems, where h would be the sum of one or more nonsmooth regularizers (for example, $\|\cdot\|_1$ to induce sparsity).

Additive composite problems can be solved by prox-gradient or splitting methods, which solve a subproblem based on h each iteration. However, this limits these methods to problems where h is relatively simple. The subgradient method avoids this limitation by only requiring the computation of a subgradient of f at each iteration, with the subdifferential being given by $\partial f(x) = \nabla\Phi(x) + \partial h(x)$. The classic convergence theory fails to provide any guarantees for this problem since f may be non-Lipschitz. In contrast, we show this problem class has a simple growth bound from which guarantees for the classic subgradient method directly follow.

Lemma 2.3. *Consider any $\Phi \in C^1$ with a (L_Φ, v)-Hölder continuous gradient on $\mathbb{R}^d$, any L_h-Lipschitz continuous h on $\mathbb{R}^d$, and any minimizer $x^* \in X^*$. Then*

$$f(x) - f(x^*) \le \frac{L_\Phi}{v+1}\|x - x^*\|^{v+1} + 2L_h\|x - x^*\| \quad (\forall x \in \mathbb{R}^d).$$

Proof. From the first-order optimality conditions of f, we know $g^* := -\nabla\Phi(x^*) \in \partial h(x^*)$. Define the following lower bound on $f(x)$:

$$l(x) := \Phi(x) + h(x^*) + g^{*T}(x - x^*).$$

Notice that $f(x)$ and $l(x)$ both minimize at x^* with $f(x^*) = l(x^*)$. Since $l(x)$ has an (L_Φ, v)-Hölder continuous gradient, Lemma 2.1 implies, for any $x \in \mathbb{R}^d$,

$$l(x) - l(x^*) \leq \frac{L_\Phi}{v+1}\|x - x^*\|^{v+1}.$$

The Lipschitz continuity of h implies

$$l(x) = \Phi(x) + h(x^*) + g^{*T}(x - x^*) \geq \Phi(x) + (h(x) - L_h\|x - x^*\|) - L_h\|x - x^*\|.$$

Combining these two inequalities completes the proof. □

Plugging $\mathcal{D}(t) = L_\Phi t^{v+1}/(v+1) + 2L_h t$ into Theorem 1.2 immediately results in the following $O(1/\sqrt{T})$ convergence rate (for simplicity, we state the bound for constant step size).

Corollary 2.4 (additive composite convergence). *Consider the deterministic subgradient oracle* $\nabla\Phi(x) + g_h(x)$*. Then under the constant step size* $\alpha_k = R/\sqrt{T+1}$*, the iteration* (1.1) *satisfies*

$$\min_{k=0,\dots,T}\{f(x_k) - f^*\} \leq \frac{L_\Phi R^{v+1}}{(v+1)(T+1)^{(v+1)/2}} + \frac{2L_h R}{\sqrt{T+1}}.$$

The first term in this rate exactly matches the convergence rate on functions, such as Φ, with Hölderian gradient (see Corollary 2.2). Further, up to a factor of two, the second term matches the convergence rate on Lipschitz continuous functions, such as h (see (1.3)). Thus the subgradient method on $\Phi(x) + h(x)$ has convergence guarantees no worse than those of the subgradient method on $\Phi(x)$ or $h(x)$ separately.

2.3. Quadratically regularized stochastic optimization. Another common class of optimization problems results from adding a quadratic regularization term $(\lambda/2)\|x\|^2$ to the objective function for some parameter $\lambda > 0$. Consider solving

$$\min_{x\in\mathbb{R}^d} f(x) := h(x) + \frac{\lambda}{2}\|x\|^2$$

for any Lipschitz continuous convex function h. Suppose we have a stochastic subgradient oracle for h denoted by $g_h(x;\xi)$, which satisfies $\mathbb{E}_\xi g_h(x;\xi) \in \partial h(x)$ and $\mathbb{E}_\xi\|g_h(x;\xi)\|^2 \leq L^2$. Although the function h and its stochastic oracle meet the necessary conditions for the classic theory to be applied, the addition of a quadratic term violates uniform Lipschitz continuity. Nonetheless, simple arguments yield a subgradient norm bound like (1.6) and the following $O(1/T)$ convergence rate.

Corollary 2.5 (quadratically regularized convergence). *Consider the step sizes*

$$\alpha_k = \frac{2}{\lambda(k+2) + \frac{36\lambda}{k+1}}$$

and stochastic subgradient oracle $g_h(x;\xi) + \lambda x$*. Fix some* $x^* \in X^*$*. The iteration* (1.2) *satisfies*

$$\mathbb{E}_{\xi_0,\dots,T}\left[f\left(\frac{2}{(T+1)(T+2)}\sum_{k=0}^{T}(k+1)x_k\right) - f^*\right] \leq \frac{24L^2}{\lambda(T+2)} + \frac{36\lambda\|x_0 - x^*\|^2}{(T+1)(T+2)}.$$

Proof. Consider any $x^* \in X^*$ and $g^* := -\lambda x^* \in \partial h(x^*)$ (this inclusion follows from the first-order optimality conditions for x^*). From the assumed stochastic subgradient norm bound $\mathbb{E}_\xi \|g_h(x;\xi)\|^2 \leq L^2$, all subgradients of h must have norm bounded by L. This follows since each differentiable point has $\|\nabla f(x)\|^2 \leq \mathbb{E}_\xi \|g_h(x;\xi)\|^2 \leq L^2$ and the subdifferential at nondifferentiable points is given by the convex hull of nearby gradients: $\partial f(x) = \text{conv}\{\lim \nabla f(z_k) \mid \lim z_k \to x,\ z_k \in Q\}$, where Q is the set of differentiable points near x (see [4] for a proof of this characterization). Then the expected norm squared of the stochastic subgradient $g_h(x;\xi) + \lambda x$ is bounded by

$$\begin{aligned}\mathbb{E}_\xi \|g_h(x;\xi) + \lambda x\|^2 &= \mathbb{E}_\xi \|g_h(x;\xi) - g^* + g^* + \lambda x\|^2 \\ &\leq 3\mathbb{E}_\xi \|g_h(x;\xi)\|^2 + 3\|g^*\|^2 + 3\|g^* + \lambda x\|^2 \\ &\leq 6L^2 + 3\|g^* + \lambda x\|^2 \\ &\leq 6L^2 + 6\lambda(f(x) - f(x^*)),\end{aligned}$$

where the first inequality uses Jensen's inequality, the second inequality uses the subgradient norm bound, and the third inequality uses the λ-strong convexity of f. From this, our bound follows by Theorem 1.7. □

One common example of a problem of the form $h(x) + (\lambda/2)\|x\|^2$ is training a support vector machine (SVM). Suppose one has n data points each with a feature vector $w_i \in \mathbb{R}^d$ and label $y_i \in \{-1, 1\}$. Then one trains a model $x \in \mathbb{R}^d$ for some parameter $\lambda > 0$ by solving

$$\min_{x \in \mathbb{R}^d} f(x) := \frac{1}{n}\sum_{i=1}^{n} \max\{0, 1 - y_i w_i^T x\} + \frac{\lambda}{2}\|x\|^2.$$

Here, a stochastic subgradient oracle can be given by selecting a summand $i \in [n]$ uniformly at random and then setting

$$g_h(x;i) = \begin{cases} -y_i w_i & \text{if } 1 - y_i w_i^T x \geq 0, \\ 0 & \text{otherwise,} \end{cases}$$

which satisfies $\mathbb{E}_i \|g_h(x,i)\|^2 \leq \frac{1}{n}\sum_{i=1}^{n} \|w_i\|^2$.

Much work has previously been done solving problems of the form $h(x) + \frac{\lambda}{2}\|x\|^2$ and SVMs in particular. If one adds the constraint that x lies in some large ball Q (which will then be projected onto at each iteration), the classic strongly convex rate can be applied [20] as the objective function will be Lipschitz on Q. A similar approach utilized in [10] is to show that, in expectation, all of the iterates of a stochastic subgradient method lie in a large ball (provided the initial iterate does). We remark that the resulting guarantees only apply for $x_0 \in Q$ and utilize a constant L dependent on the size of Q. Corollary 2.5 avoids these issues, giving a convergence bound for any choice of $x_0 \in \mathbb{R}^d$.

The specialized mirror descent method proposed by Lu [11] produces convergence guarantees for SVMs at a rate of $O(1/\sqrt{T})$ without needing a bounding ball. Splitting methods and quasi-Newton methods capable of solving this problem are given in [7] and [23], respectively, which both avoid needing to assume subgradient bounds.

2.4. Interpreting (1.6) as a quadratic growth upper bound. Here we provide an alternative interpretation of bounding the size of subgradients by (1.6) on some convex open set $U \subseteq \mathbb{R}^d$ for deterministic subgradient oracles. In particular, suppose all $x \in U$ have

$$\|g(x)\|^2 \leq L_0^2 + L_1\Big(f(x) - \inf_{x'\in U} f(x')\Big). \tag{2.1}$$

First consider the classic model where $L_1 = 0$. This is equivalent to f being L_0-Lipschitz continuous on U and can be restated as the following upper bound holding for each $x \in U$:

$$f(y) \leq f(x) + L_0\|y-x\| \quad (\forall y \in U)\,.$$

This characterization shows the limitation to linear growth of the classic model. In the following proposition, we present an upper bound characterization when $L_1 > 0$, which can be viewed as allowing up to quadratic growth.

PROPOSITION 2.6. *A convex function f satisfies* (2.1) *on some open convex $U \subseteq \mathbb{R}^d$ if and only if the following quadratic upper bound holds for each $x \in U$:*

$$f(y) \leq f(x) + \frac{L_1}{4}\|y-x\|^2 + \|y-x\|\sqrt{L_1\Big(f(x) - \inf_{x'\in U} f(x')\Big) + L_0^2} \quad (\forall y \in U)\,.$$

Proof. First we prove the forward direction. Consider any $x, y \in U$ and subgradient oracle $g(\cdot)$. Let $v = (y-x)/\|y-x\|$ denote the unit direction from x to y, and $h(t) = f(x+tv) - \inf_{x'\in U} f(x')$ denote the restriction of f to this line shifted to have nonnegative value. Notice that $h(0) = f(x) - \inf_{x'\in U} f(x')$ and $h(\|y-x\|) = f(y) - \inf_{x'\in U} f(x')$. The convexity of h implies it is differentiable almost everywhere in the interval $[0, \|y-x\|]$. Thus h satisfies the following for almost every $t \in [0, \|y-x\|]$:

$$|h'(t)| = |v^T g(x+tv)| \leq \|g(x+tv)\|.$$

This produces the differential inequality of $|h'(t)| \leq \sqrt{L_1 h(t) + L_0^2}$. Note that the unique solution to the ordinary differential equation $y'(t) = \sqrt{L_1 y(t) + L_0^2}$ with initial condition $y(0) = h(0)$ is $y(t) = h(0) + \frac{L_1}{4}t^2 + t\sqrt{L_1 h(0) + L_0^2}$. Then the claimed bound will follow from showing $h(t) \leq y(t)$ at $t = \|y-x\|$. This inequality must be the case for all $t \geq 0$, as otherwise some $t \geq 0$ must have $h(t) = y(t)$ and $\limsup_{t'\to t} h'(t') > y'(t)$, which implies

$$\limsup_{t'\to t} h'(t') > \sqrt{L_1 y(t) + L_0^2} = \sqrt{L_1 h(t) + L_0^2},$$

contradicting our premise.

Now we prove the reverse direction. Denote the upper bound given by some $x \in U$ as

$$u_x(y) := f(x) + \frac{L_1}{4}\|y-x\|^2 + \|y-x\|\sqrt{L_1\Big(f(x) - \inf_{x'\in U} f(x')\Big) + L_0^2}.$$

Further, let D_v denote the directional derivative operator in some unit direction $v \in \mathbb{R}^d$. Then for any subgradient $g \in \partial f(x)$,

$$v^T g \leq D_v f(x) \leq D_v u_x(x),$$

where the first inequality uses the definition of D_v and the second uses that u_x upper bounds f. Direct calculation shows $D_v u_x(x) \leq \sqrt{L_1(f(x) - \inf_{x'\in U} f(x')) + L_0^2}$. Then our subgradient bound follows by taking $v = g/\|g\|$. ∎

3. Convergence proofs.

Each of our extended convergence theorems follows from essentially the same proof as its classic counterpart. The central inequality in analyzing subgradient methods is the following.

LEMMA 3.1. *Consider any convex function f. For any $x, y \in Q$ and $\alpha > 0$,*

$$\mathbb{E}_\xi \|P_Q(x - \alpha g(x;\xi)) - y\|^2 \le \|x - y\|^2 - 2\alpha(\mathbb{E}_\xi\ g(x;\xi))^T(x - y) + \alpha^2 \mathbb{E}_\xi \|g(x;\xi)\|^2.$$

Proof. Since orthogonal projection onto a convex set is nonexpansive, we have

$$\begin{aligned}\|P_Q(x - \alpha g(x;\xi)) - y\|^2 &\le \|x - \alpha g(x;\xi) - y\|^2 \\ &= \|x - y\|^2 - 2\alpha g(x;\xi)^T(x - y) + \alpha^2 \|g(x;\xi)\|^2.\end{aligned}$$

Taking the expectation over $\xi \sim D$ yields

$$\mathbb{E}_\xi \|P_Q(x - \alpha g(x;\xi)) - y\|^2 \le \|x - y\|^2 - 2\alpha(\mathbb{E}_\xi\ g(x;\xi))^T(x - y) + \alpha^2 \mathbb{E}_\xi \|g(x;\xi)\|^2. \square$$

Let $D_k^2 = \mathbb{E}_{\xi_{0,\dots,T}} \|x_k - x^*\|^2$ denote the expected distance squared from each iterate to the minimizer x^*. Each of our proofs follows the same general outline: use Lemma 3.1 to set up a telescoping inequality on D_k^2, then sum the telescope. We begin by proving Shor's convergence result as its derivation is short and informative.

3.1. Proof of Shor's theorem (Theorem 1.1).

From Lemma 3.1 with $x = x_k$, $y = x^*$, and $\alpha = \alpha_k / \|g(x_k)\|$, it follows that

$$D_{k+1}^2 \le D_k^2 - \frac{2\alpha_k g(x_k)^T(x_k - x^*)}{\|g(x_k)\|} + \alpha_k^2.$$

Inductively applying this implies

$$0 \le D_{T+1}^2 \le D_0^2 - \sum_{k=0}^{T} 2\alpha_k \frac{g(x_k)^T(x_k - x^*)}{\|g(x_k)\|} + \sum_{k=0}^{T} \alpha_k^2.$$

Thus

$$\min_{k=0,\dots,T} \left\{ \frac{g(x_k)^T(x_k - x^*)}{\|g(x_k)\|} \right\} \le \frac{D_0^2 + \sum_{k=0}^T \alpha_k^2}{2\sum_{k=0}^T \alpha_k},$$

completing the proof. $\square$

3.2. Proof of Theorem 1.2.

This follows directly from Theorem 1.1. Note the result trivially holds if some iterate $0 \le k \le T$ satisfies $x_k \in X^*$. Suppose x_k satisfies the inequality in Theorem 1.1. Let y be the closest point in $\{x \mid g(x_k)^T(x - x_k) = 0\}$ to x^*. Then our assumed growth bound implies

$$f(y) - f^* \le \mathcal{D}(\|y - x^*\|) = \mathcal{D}\left(\frac{g_k^T(x_k - x^*)}{\|g_k\|}\right) \le \mathcal{D}\left(\frac{D_0^2 + \sum_{k=0}^T \alpha_k^2}{2\sum_{k=0}^T \alpha_k}\right).$$

The convexity of f implies $f(x_k) \le f(y)$, completing the proof. $\square$

3.3. Proof of Theorem 1.6.

From Lemma 3.1 with $x = x_k$, $y = x^*$, and $\alpha = \alpha_k$, it follows that

$$\begin{aligned}D_{k+1}^2 &\le D_k^2 - \mathbb{E}_{\xi_{0,\dots,T}}\left[2\alpha_k(\mathbb{E}_\xi g(x_k;\xi_k))^T(x_k - x^*)\right] + \alpha_k^2 \mathbb{E}_{\xi_{0,\dots,T}} \|g(x_k,\xi_k)\|^2 \\ &\le D_k^2 - \mathbb{E}_{\xi_{0,\dots,T}}\left[(2\alpha_k - L_1\alpha_k^2)(f(x_k) - f^*)\right] + L_0^2\alpha_k^2,\end{aligned}$$

where the second inequality uses the convexity of f and the bound on $\mathbb{E}_\xi\|g(x;\xi)\|^2$. Inductively applying this implies

$$0 \leq D_{T+1}^2 \leq D_0^2 - \mathbb{E}_{\xi_{0,\dots,T}}\left[\sum_{k=0}^{T}(2\alpha_k - L_1\alpha_k^2)(f(x_k)-f^*)\right] + L_0^2\sum_{k=0}^{T}\alpha_k^2.$$

The convexity of f yields

$$\mathbb{E}_{\xi_{0,\dots,T}}\left[f\left(\frac{\sum_{k=0}^T \alpha_k(2-L_1\alpha_k)x_k}{\sum_{k=0}^T \alpha_k(2-L_1\alpha_k)}\right) - f^*\right] \leq \frac{D_0^2 + L_0^2\sum_{k=0}^T \alpha_k^2}{\sum_{k=0}^T \alpha_k(2-L_1\alpha_k)},$$

completing the proof. □

3.4. Proof of Theorem 1.7. Our proof follows the style of [10]. Observe that our choice of step size α_k satisfies the following pair of conditions: first, note that it is a solution to the recurrence

$$(k+1)\alpha_k^{-1} = (k+2)(\alpha_{k+1}^{-1} - \mu); \tag{3.1}$$

second, note that $L_1\alpha_k \leq 1$ for all $k \geq 0$ since

$$L_1\alpha_k = \frac{2\mu(k+2)L_1}{(\mu(k+2))^2 + \frac{k+2}{k+1}L_1^2} \leq \frac{2\mu(k+2)L_1}{(\mu(k+2))^2 + L_1^2} \leq 1. \tag{3.2}$$

From Lemma 3.1 with $x = x_k$, $y = x^*$, and $\alpha = \alpha_k$, it follows that

$$\begin{aligned} D_{k+1}^2 &\leq D_k^2 - \mathbb{E}_{\xi_{0,\dots,T}}\left[2\alpha_k(\mathbb{E}_\xi g(x_k;\xi))^T(x_k - x^*)\right] + \alpha_k^2\mathbb{E}_{\xi_{0,\dots,T}}\|g(x_k,\xi_k)\|^2 \\ &\leq (1-\mu\alpha_k)D_k^2 - \mathbb{E}_{\xi_{0,\dots,T}}\left[(2\alpha_k - L_1\alpha_k^2)(f(x_k)-f^*)\right] + L_0^2\alpha_k^2, \end{aligned}$$

where the second inequality uses the strong convexity of f and the assumed bound on $\mathbb{E}_\xi\|g(x;\xi)\|^2$. Multiplying by $(k+1)/\alpha_k$ and invoking (3.2) yields

$$\begin{aligned} (k+1)\alpha_k^{-1}D_{k+1}^2 \leq &(k+1)(\alpha_k^{-1} - \mu)D_k^2 \\ &- \mathbb{E}_{\xi_{0,\dots,T}}\left[(k+1)(2-L_1\alpha_k)(f(x_k)-f^*)\right] + L_0^2(k+1)\alpha_k. \end{aligned}$$

Notice that this inequality telescopes due to (3.1). Inductively applying this implies

$$0 \leq (\alpha_0^{-1} - \mu)D_0^2 - \mathbb{E}_{\xi_{0,\dots,T}}\left[\sum_{k=0}^{T}(k+1)(2-L_1\alpha_k)(f(x_k)-f^*)\right] + L_0^2\sum_{k=0}^{T}(k+1)\alpha_k.$$

Since $\sum_{k=0}^T (k+1)\alpha_k \leq 2(T+1)/\mu$ and $\alpha_0^{-1} - \mu = L_1^2/2\mu$, we have

$$\mathbb{E}_{\xi_{0,\dots,T}}\left[\sum_{k=0}^{T}(k+1)(2-L_1\alpha_k)(f(x_k)-f^*)\right] \leq \frac{L_1^2D_0^2}{2\mu} + \frac{2L_0^2(T+1)}{\mu}.$$

Observe that the coefficients of each $f(x_k) - f^*$ above are positive due to (3.2). Then the convexity of f yields our first convergence bound. From (3.2), we know $2 - L_1\alpha_k \geq 1$ for all $k \geq 0$. Then the previous inequality can be weakened to

$$\mathbb{E}_{\xi_{0,\dots,T}}\left[\sum_{k=0}^{T}(k+1)(f(x_k)-f^*)\right] \leq \frac{L_1^2D_0^2}{2\mu} + \frac{2L_0^2(T+1)}{\mu}.$$

The convexity of f yields our second convergence bound. □

4. Improved convergence without strong convexity.

The idea of utilizing growth lower bounds to improve convergence guarantees is far from new. Nemirovskii and Nesterov [15] showed simple variations of standard first-order methods can give optimal convergence rates for convex problems satisfying the following Hölder growth bound (referred to therein as a "strict minimum condition") for all $x \in \mathbb{R}^d$: $f(x)-f^* \geq \mu\|x - x^*\|^{v+1}$, where x^* is the unique minimizer of f. Later, the work of Burke and Ferris [3] studied (potentially nonconvex) problems with *weak sharp minima*, that is, for some $S \subseteq \mathbb{R}^d$ and $y \in S$, all x near S have $f(x) - f(y) \geq \mu\mathbf{dist}(x, S)$. They show this condition often holds and enables improved convergence guarantees, sometimes ensuring finite convergence.

Many recent works have considered weakening the assumption of strong convexity while maintaining the standard improvements in convergence rate for smooth optimization problems (see, for example, [1, 6, 12, 13]). Instead the weaker condition of requiring quadratic growth away from the set of minimizers suffices. We demonstrate that this weakening of strong convexity is also sufficient for (1.2) to have a convergence rate of $O(1/T)$.

A function f has μ-quadratic growth for some $\mu > 0$ if all $x \in Q$ satisfy

$$f(x) \geq f^* + \frac{\mu}{2}\mathbf{dist}(x, X^*)^2.$$

The proof of Theorem 1.7 only uses strong convexity once for the following inequality:

$$g(x_k)^T(x_k - x^*) \geq f(x_k) - f^* + \frac{\mu}{2}\|x_k - x^*\|^2.$$

Having μ-quadratic growth suffices to produce a similar inequality, weakened by a factor of $1/2$:

$$g(x_k)^T(x_k - P_{X^*}(x_k)) \geq f(x_k) - f^* \geq \frac{1}{2}\left(f(x_k) - f^*\right) + \frac{\mu}{4}\mathbf{dist}(x_k, X^*)^2.$$

Then simple modifications of the proof of Theorem 1.7 yield the following convergence rate.

THEOREM 4.1. *Consider any convex function f with μ-quadratic growth and a stochastic subgradient oracle satisfying* (1.6). *Then for the sequence of step sizes*

$$\alpha_k = \frac{4}{\mu(k+2) + \frac{4L_1^2}{\mu(k+1)}},$$

the iteration (1.2) *satisfies*

$$\mathbb{E}_{\xi_{0,\dots,T}}\left[f\left(\frac{\sum_{k=0}^{T}(k+1)(1-L_1\alpha_k)x_k}{\sum_{k=0}^{T}(k+1)(1-L_1\alpha_k)}\right) - f^*\right] \leq \frac{4L_0^2(T+1) + L_1^2\mathbf{dist}(x_0, X^*)^2}{\mu\sum_{k=0}^{T}(k+1)(1-L_1\alpha_k)}.$$

Proof. Observe that our choice of step size α_k satisfies the following pair of conditions. First, note that it is a solution to the recurrence

$$(k+1)\alpha_k^{-1} = (k+2)(\alpha_{k+1}^{-1} - \mu/2). \tag{4.1}$$

Second, note that $L_1\alpha_k < 1$ for all $k \geq 0$. This follows as

$$L_1\alpha_k = \frac{4\mu(k+2)L_1}{(\mu(k+2))^2 + 4\frac{k+2}{k+1}L_1^2} \leq \frac{4\mu(k+2)L_1}{(\mu(k+2))^2 + (2L_1)^2} \leq 1, \tag{4.2}$$

where the first inequality is strict if $L_1 > 0$ and the second inequality is strict if $L_1 = 0$.

Let $D_k^2 = \mathbb{E}_{\xi_{0,\dots,T}}\mathbf{dist}(x_k, X^*)^2$ denote the expected distance squared from each iterate to the set of minimizers X^*. From Lemma 3.1 with $x = x_k$, $y = P_{X^*}(x_k)$, and $\alpha = \alpha_k$, it follows that

$$\begin{aligned} D_{k+1}^2 &\le D_k^2 - \mathbb{E}_{\xi_{0,\dots,T}}\left[2\alpha_k(\mathbb{E}_\xi g(x_k;\xi))^T(x_k - P_{X^*}(x_k))\right] + \alpha_k^2\mathbb{E}_{\xi_{0,\dots,T}}\|g(x_k,\xi_k)\|^2 \\ &\le (1-\mu\alpha_k/2)D_k^2 - \mathbb{E}_{\xi_{0,\dots,T}}\left[(\alpha_k - L_1\alpha_k^2)(f(x_k)-f^*)\right] + L_0^2\alpha_k^2, \end{aligned}$$

where the second inequality uses the quadratic growth of f and the assumed bound on $\mathbb{E}_\xi\|g(x;\xi)\|^2$. Multiplying by $(k+1)/\alpha_k$ and invoking (4.2) yields

$$\begin{aligned} (k+1)\alpha_k^{-1}D_{k+1}^2 \le (k+1)(\alpha_k^{-1} - \mu/2)D_k^2 \\ - \mathbb{E}_{\xi_{0,\dots,T}}\left[(k+1)(1-L_1\alpha_k)(f(x_k)-f^*)\right] + L_0^2(k+1)\alpha_k. \end{aligned}$$

Notice that this inequality telescopes due to (4.1). Inductively applying this implies

$$0 \le (\alpha_0^{-1} - \mu/2)D_0^2 - \mathbb{E}_{\xi_{0,\dots,T}}\left[\sum_{k=0}^T (k+1)(1-L_1\alpha_k)(f(x_k)-f^*)\right] + L_0^2\sum_{k=0}^T(k+1)\alpha_k.$$

Since $\sum_{k=0}^T(k+1)\alpha_k \le 4(T+1)/\mu$ and $\alpha_0^{-1} - \mu/2 = L_1^2/\mu$, we have

$$\mathbb{E}_{\xi_{0,\dots,T}}\left[\sum_{k=0}^T(k+1)(1-L_1\alpha_k)(f(x_k)-f^*)\right] \le \frac{L_1^2D_0^2}{\mu} + \frac{4L_0^2(T+1)}{\mu}.$$

Observe that the coefficients of each $f(x_k) - f^*$ above are positive due to (4.2). Then the convexity of f completes the proof. □

Observe that this convergence rate is on the order of $O(1/T)$. To see this, we need to show the sum $\sum_{k=0}^T(k+1)(1-L_1\alpha_k)$ is at least $\Omega(T^2)$, which follows as

$$\begin{aligned} \sum_{k=0}^T(k+1)(1-L_1\alpha_k) &= \sum_{k=0}^T(k+1) - \sum_{k=0}^T(k+1)L_1\alpha_k \\ &\ge \frac{(T+1)(T+2)}{2} - \sum_{k=0}^T\frac{4L_1}{\mu} \\ &= \frac{(T+1)(T+2)}{2} - (T+1)\frac{4L_1}{\mu} \\ &= \frac{(T+1)(T+2-8L_1/\mu)}{2}. \end{aligned}$$

Acknowledgments. The author thanks Jim Renegar for providing feedback on an early draft of this work. The author also thanks the associate editor and two anonymous reviewers for their helpful comments.

REFERENCES

[1] J. Bolte, T. P. Nguyen, J. Peypouquet, and B. W. Suter, *From error bounds to the complexity of first-order descent methods for convex functions*, Math. Program., 165 (2017), pp. 471–507, https://doi.org/10.1007/s10107-016-1091-6.

[2] S. Bubeck, *Convex optimization: Algorithms and complexity*, Found. Trends Mach. Learn., 8 (2015), pp. 231–357.

[3] J. Burke and M. Ferris, *Weak sharp minima in mathematical programming*, SIAM J. Control Optim., 31 (1993), pp. 1340–1359, https://doi.org/10.1137/0331063.

[4] F. Clarke, *Optimization and Nonsmooth Analysis*, Classics Appl. Math., SIAM, Philadelphia, PA, 1990.

[5] D. Davis and B. Grimmer, *Proximally Guided Stochastic Subgradient Method for Nonsmooth, Nonconvex Problems*, preprint, https://arxiv.org/abs/1707.03505, 2017.

[6] D. Drusvyatskiy and A. S. Lewis, *Error bounds, quadratic growth, and linear convergence of proximal methods*, Math. Oper. Res., 43 (2018), pp. 919–948.

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

[8] B. Grimmer, *Radial subgradient method*, SIAM J. Optim., 28 (2018), pp. 459–469, https://doi.org/10.1137/17M1122980.

[9] E. Hazan and S. Kale, *Beyond the regret minimization barrier: Optimal algorithms for stochastic strongly-convex optimization*, J. Mach. Learn. Res., 15 (2014), pp. 2489–2512.

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

[11] H. Lu, *"Relative-Continuity" for Non-Lipschitz Non-Smooth Convex Optimization Using Stochastic (or Deterministic) Mirror Descent*, preprint, https://arxiv.org/abs/1710.04718, 2017.

[12] Z.-Q. Luo and P. Tseng, *Error bounds and convergence analysis of feasible descent methods: A general approach*, Ann. Oper. Res., 46 (1993), pp. 157–178, https://doi.org/10.1007/BF02096261.

[13] I. Necoara, Y. Nesterov, and F. Glineur, *Linear convergence of first order methods for non-strongly convex optimization*, Math. Program., 175 (2018), pp. 69–107, https://doi.org/10.1007/s10107-018-1232-1.

[14] A. Nedić and S. Lee, *On stochastic subgradient mirror-descent algorithm with weighted averaging*, SIAM J. Optim., 24 (2014), pp. 84–107, https://doi.org/10.1137/120894464.

[15] A. Nemirovskii and Y. Nesterov, *Optimal methods of smooth convex minimization*, USSR Comput. Math. Math. Phys., 25 (1985), pp. 21–30.

[16] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, 1st ed., Springer, New York, 2004.

[17] 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, Omnipress, Madison, WI, 2012, pp. 1571–1578.

[18] J. Renegar, *"Efficient" subgradient methods for general convex optimization*, SIAM J. Optim., 26 (2016), pp. 2649–2676, https://doi.org/10.1137/15M1027371.

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

[20] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter, *Pegasos: Primal estimated sub-gradient solver for SVM*, Math. Program., 127 (2011), pp. 3–30.

[21] N. Z. Shor, *Minimization methods for non-differentiable functions*, in Minimization Methods for Non-Differentiable Functions, Springer Ser. Comput. Math. 3, Springer, Berlin, 1985, pp. 22–47.

[22] N. Z. Shor, *Subgradient and ϵ-subgradient methods*, in Nondifferentiable Optimization and Polynomial Problems, Nonconvex Optim. Appl. 24, Springer, Boston, MA, 1998, pp. 35–70.

[23] J. Yu, S. Vishwanathan, S. Günter, and N. N. Schraudolph, *A quasi-Newton approach to nonsmooth convex optimization problems in machine learning*, J. Mach. Learn. Res., 11 (2010), pp. 1145–1200.