Convergence Rates for Deterministic and Stochastic Subgradient Methods Without Lipschitz Continuity

Benjamin Grimmer*

Abstract

We generalize the classic convergence rate theory for subgradient methods to apply to non-Lipschitz functions via a new measure of steepness. For the deterministic projected subgradient method, we derive a global $O(1/\sqrt{T})$ convergence rate for any function with at most exponential growth. Our approach 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 functions with at most quadratic growth, which improves to $O(1/T)$ under strong convexity.

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^*. Further, we assume orthogonal projection onto Q is computationally tractable (which we denote by $P_Q(\cdot)$).

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 [1, 10], as well as [5, 6, 7, 9, 11] for a sample of more recent works.

*bdg79@cornell.edu; Cornell University, Ithaca NY

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 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.

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 (i.e., $\mathbb{E}_{\xi\sim D}\, g(x;\xi) \in \partial f(x)$ for some easily sampled distribution D).

We analyze two classic subgradient methods, differing in their step size policy. Given a deterministic oracle, we consider the following normalized subgradient method

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

for some positive sequence α_k. 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 following method

$$x_{k+1} := P_Q(x_k - \alpha_k g(x_k;\xi_k)), \tag{2}$$

for some positive sequence α_k and i.i.d. sample sequence $\xi_k \sim D$.

Let $D_0 := \mathbf{dist}(x_0, X^*)$ denote the Euclidean distance from the initial iterate to a minimizer, and $\delta(x) := f(x) - \min_{x'\in Q} f(x')$ denote the objective gap at a point $x \in Q$. The following three theorems establish the classic objective gap convergence rates of (1) and (2).

Theorem 1 (Classic Deterministic Rate)**.** *Consider any convex function f and subgradient oracle satisfying $\|g(x)\| \le L$ for all $x \in Q$. Then for any positive sequence α_k, the iteration* (1) *satisfies*

$$\min_{k=0\ldots T}\{\delta(x_k)\} \le L\frac{D_0^2 + \sum_{k=0}^T \alpha_k^2}{2\sum_{k=0}^T \alpha_k}.$$

For example, under the constant step size $\alpha_k = D_0/\sqrt{T+1}$, the iteration (1) *satisfies*

$$\min_{k=0\ldots T}\{\delta(x_k)\} \le \frac{LD_0}{\sqrt{T+1}}.$$

Theorem 2 (Classic Stochastic Rate)**.** *Consider any convex function f and stochastic subgradient oracle satisfying $\mathbb{E}_\xi\|g(x;\xi)\|^2 \le L^2$ for all $x \in Q$. Then for any positive sequence α_k, the iteration* (2) *satisfies*

$$\mathbb{E}_{\xi_{0\ldots T}}\left[\delta\left(\frac{\sum_{k=0}^T \alpha_k x_k}{\sum_{k=0}^T \alpha_k}\right)\right] \le \frac{D_0^2 + L^2\sum_{k=0}^T \alpha_k^2}{2\sum_{k=0}^T \alpha_k}.$$

For example, under the constant step size $\alpha_k = D_0/L\sqrt{T+1}$, the iteration (2) *satisfies*

$$\mathbb{E}_{\xi_{0\ldots T}}\left[\delta\left(\frac{1}{T+1}\sum_{k=0}^T x_k\right)\right] \le \frac{LD_0}{\sqrt{T+1}}.$$

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

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

If this holds for some $\mu > 0$, the convergence of (2) can be improved to $O(1/T)$ [5, 6, 11]. Below, we present one such bound from [6].

Theorem 3 (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 decreasing sequence of step sizes $\alpha_k = 2/\mu(k+2)$, the iteration (2) satisfies*

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

Remarks on the Generality of Theorems 1-3. The assumed subgradient norm 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). This assumption restricts the classic convergence results 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 introduce humongous constants heavily dependent on the size of Q (and frankly, lacks the elegance that such a fundamental method deserves).

We also 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.

Recently, Renegar [12] 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 [4] showed 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 [7] proposes 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 gives convergence guarantees based on a relative-continuity constant instead of a uniform Lipschitz constant.

Although the works of Renegar [12], Grimmer [4], and Lu [7] give convergence rates for specialized subgradient methods without assuming Lipschitz continuity, its unclear what guarantees the classic subgradient methods (1) and (2) have for non-Lipschitz problems. In this paper, we propose a generalization of Lipschitz continuity, which greatly extends the applicability and quality of convergence rate bounds for these classic methods.

1.1 Our Contributions

We propose the following generalization of an absolute bound on subgradient norm.

Definition 4. *Consider any nonnegative valued function $\mathcal{L}\colon \mathbb{R}_+ \to \mathbb{R}_+$.*
We say a subgradient oracle is $\boldsymbol{\mathcal{L}(t)}$**-steep** *on Q if $\|g(x)\| \leq \mathcal{L}(\delta(x))$ for all $x \in Q$.*
Similarly, a stochastic oracle is $\boldsymbol{\mathcal{L}(t)}$**-steep** *on Q if $\mathbb{E}_\xi\|g(x;\xi)\|^2 \leq \mathcal{L}(\delta(x))^2$ for all $x \in Q$.*

We say a function is $\mathcal{L}(t)$-steep if every subgradient oracle is $\mathcal{L}(t)$-steep. This definition allows subgradients to be large when the objective gap is large as well (where the exact relation between these is governed by $\mathcal{L}(t)$). Note when $\mathcal{L}(t)$ is a constant function, steepness

is identical to the classic model. In this case, our convergence rates stated below exactly match their classic counterparts. In Section 2, we discuss a number of examples of steepness and applications of our bounds. First, consider the deterministic subgradient method (1).

Theorem 5 (Extended Deterministic Rate)**.** *Consider any convex function f and $\mathcal{L}(t)$-steep subgradient oracle on Q. Then for any positive sequence α_k, the iteration* (1) *satisfies*

$$\min_{k=0\dots T}\{\delta(x_k)\} \le \sup\left\{ t \mid \frac{t}{\mathcal{L}(t)} \le \frac{D_0^2 + \sum_{k=0}^T \alpha_k^2}{2\sum_{k=0}^T \alpha_k} \right\}.$$

For example, under the constant step size $\alpha_k = D_0/\sqrt{T+1}$, the iteration (1) *satisfies*

$$\min_{k=0\dots T}\{\delta(x_k)\} \le \sup\left\{ t \mid \frac{t}{\mathcal{L}(t)} \le \frac{D_0}{\sqrt{T+1}} \right\}.$$

Remarks on the Generality of Theorem 5. When $\mathcal{L}(t)$ is at most linear (i.e., there exists some $L_1, L_0 \ge 0$ such that $\mathcal{L}(t) \le L_1 t + L_0$), the supremum above approaches 0 as $T \to \infty$ and provides a meaningful rate of convergence. For reasonably simple $\mathcal{L}(t)$, this supremum has a closed form. However, this bound may be vacuous when $\mathcal{L}(t)$ is superlinear since having $\liminf_{t\to\infty} t/\mathcal{L}(t) = 0$ implies the supremum above equals $+\infty$.

Having an $L_1 t + L_0$-steep oracle can be viewed as allowing functions with at most exponential growth (see Proposition 14). Intuitively, this is reasonable as such steepness is roughly a differential inequality of the form $f'(x) \le L_1 f(x) + L_0$, which has a classic exponential solution. This is a large improvement on the linear growth required by the classic theory. In Section 4, we discuss how more general convergence rates can be given.

Provided $\mathcal{L}(t)$ is at most linear, simple limiting arguments give the following eventual convergence rate of (1) based on Theorem 5: For any $\epsilon > 0$, there exists $T_0 > 0$, such that all $T > T_0$ have

$$\sup\left\{ t \mid \frac{t}{\mathcal{L}(t)} \le \frac{D_0}{\sqrt{T+1}} \right\} \le \frac{(\limsup_{t\to 0^+} \mathcal{L}(t) + \epsilon) D_0}{\sqrt{T+1}}.$$

As a result, the asymptotic convergence rate of (1) is determined entirely by the size of subgradients around the set of minimizers, and conversely, steepness far from optimality plays no role in the asymptotic behavior.

As a surprising consequence of Theorem 5, we recover the classic convergence rate for gradient descent on differentiable functions with an L-Lipschitz continuous gradient of $O(LD_0^2/T)$ [10]. Any such function is $\sqrt{2Lt}$-steep on $\mathbb{R}^d$ (see Lemma 9). Then a convergence rate immediately follows from Theorem 5 (for simplicity, we consider constant step size).

Corollary 6 (Generalizing Gradient Descent's Convergence)**.** *Consider any convex function f and $\sqrt{2Lt}$-steep subgradient oracle. Then under the constant step size $\alpha_k = D_0/\sqrt{T+1}$, the iteration* (1) *satisfies*

$$\min_{k=0\dots T}\{\delta(x_k)\} \le \frac{2LD_0^2}{T+1}.$$

Thus a convergence rate of $O(LD_0^2/T)$ can be attained without any mention of smoothness or differentiability. Instead, the essential property is that gradients (or subgradients) have norm go to zero sufficiently fast when approaching the minimum function value. In Section 2, we bound the steepness of any function with a Hölder continuous gradient (an extension of Lipschitz continuity) and state the resulting convergence bound. In general, for any at most linear $\mathcal{L}(t)$ with $\lim_{t\to 0^+} \mathcal{L}(t) = 0$, Theorem 5 gives convergence at a rate of $o(1/\sqrt{T})$.

Now we consider the stochastic subgradient method defined by (2). Here we limit our analysis to $\sqrt{L_1 t + L_0^2}$-steepness for some $L_1, L_0 \geq 0$ (note the classic model restricts to $L_1 = 0$). We have the following guarantees for convex and strongly convex problems.

Theorem 7 (Extended Stochastic Rate)**.** *Consider any convex function f and $\sqrt{L_1 t + L_0^2}$-steep stochastic subgradient oracle on Q. Then for any positive sequence α_k with $L_1\alpha_k < 2$, the iteration* (2) *satisfies*

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

For example, under the constant step size $\alpha_k = D_0/L_0\sqrt{T+1}$, the iteration (2) *satisfies*

$$\mathbb{E}_{\xi_{0\ldots T}}\left[\delta\left(\frac{1}{T+1}\sum_{k=0}^T x_k\right)\right] \leq \frac{L_0 D_0}{\sqrt{T+1}} \cdot \frac{2}{2 - L_1\alpha_k},$$

provided T is large enough to have $L_1\alpha_k < 2$.

Theorem 8 (Extended Strongly Convex Stochastic Rate[1])**.** *Consider any μ-strongly convex function f and $\sqrt{L_1 t + L_0^2}$-steep stochastic subgradient oracle on Q. Then for the decreasing sequence of step sizes*

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

the iteration (2) *satisfies*

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

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

$$\mathbb{E}_{\xi_{0\ldots T}}\left[\delta\left(\frac{2}{(T+1)(T+2)}\sum_{k=0}^T (k+1)x_k\right)\right] \leq \frac{4L_0^2}{\mu(T+2)} + \frac{L_1^2 D_0^2}{\mu(T+1)(T+2)}.$$

Remarks on the Generality of Theorems 7 and 8. Having a deterministic $\sqrt{L_1 t + L_0^2}$-steep oracle can be viewed as allowing functions with at most quadratic growth (see Proposition 14). Intuitively, this is reasonable since the corresponding differential inequality $f'(x) \leq \sqrt{L_1 f(x) + L_0^2}$ has a simple quadratic solution. As a result, we avoid the inherent conflict in Theorem 3 between Lipschitz continuity and strong convexity since a function can globally have both square root steepness and strong convexity. In Section 4, we show a weaker condition than strong convexity is sufficient to ensure a $O(1/T)$ rate.

[1] A predecessor of Theorem 8 was given by Davis and Grimmer in Proposition 3.2 of [2], where a $O(\log(T)/T)$ convergence rate was shown for certain non-Lipschitz strongly convex problems.

Function	$\mathcal{L}(t)$-Steepness	Function	$\mathcal{L}(t)$-Steepness
$\|x\|$	1	$\|x\|^2$	$\sqrt{4t}$
$\|x\| + \|x\|^2$	$\sqrt{4t+1}$	$\max\{\|x\|^2, 1\}$	$\sqrt{4t+4}$
$\exp(\|x\|)$	$t+1$	$\exp(\|x\|^2)$	$2(t+1)\sqrt{\log(t+1)}$

Table 1: Numerous simple convex functions and bounds on their steepness on $\mathbb{R}^d$ are given. Each steepness bound is pointwise as small as possible. All of the examples given, except $\exp(\|x\|^2)$, have at most linear steepness, and thus fall within the scope of Theorem 5.

2 Examples and Applications of Steepness

In this section, we show several examples of problems that are steep with respect to simple functions $\mathcal{L}(t)$, as well as provide an alternative characterization of steepness in terms of upper bounds on function growth. To establish a baseline for understanding steepness, Table 1 gives a variety of simple functions and corresponding steepness bounds.

2.1 Smooth Optimization

The standard analysis of gradient descent in 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 $x, y \in \mathbb{R}^d$

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

Note this is exactly Lipschitz continuity of the gradient when $v = 1$. Below, we show the steepness of any such function admits a simple description.

Lemma 9. *Every $f \in C^1$ with a (L, v)-Hölder continuous gradient on $\mathbb{R}^d$ is*

$$\left(\frac{v+1}{v} L^{1/v} t\right)^{v/(v+1)} \text{-steep on } \mathbb{R}^d$$

Proof. The following upper bound holds for each $x \in \mathbb{R}^d$: For all $y \in \mathbb{R}^d$,

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

Selecting $y = x - \alpha \nabla f(x)$ for $\alpha = (\|\nabla f(x)\|^{1-v}/L)^{1/v}$ minimizes this upper bound. Therefore

$$\min_{x' \in \mathbb{R}^d} f(x') \leq f(y) \leq f(x) - \left(1 - \frac{1}{v+1}\right) \frac{\|\nabla f(x)\|^{(v+1)/v}}{L^{1/v}}.$$

Rearranging this inequality gives the claimed bound. □

This lemma with $v = 1$ implies any function with an L-Lipschitz gradient is $\sqrt{2Lt}$-steep. Then Theorem 5 gives our generalization of the classic gradient descent convergence rate claimed in Corollary 6. Further, for any function with a Hölderian gradient, we find the following $O(1/T^{(v+1)/2})$ convergence rate.

Corollary 10 (Generalizing Hölderian Gradient Descent's Convergence)**.** *Consider any convex function f and $\left(\frac{v+1}{v}L^{1/v}t\right)^{v/(v+1)}$-steep subgradient oracle. Then under the constant step size $\alpha_k = D_0/\sqrt{T+1}$, the iteration (1) satisfies*

$$\min_{k=0\dots T}\{\delta(x_k)\} \leq \left(\frac{v+1}{v}\right)^v \frac{LD_0^{v+1}}{(T+1)^{(v+1)/2}}.$$

2.2 Additive Composite Optimization

Often problems arise where the objective is to minimize a sum of smooth and nonsmooth functions. We consider the following 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 $h(x)$ 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 $h(x)$ 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 each iteration, which is given by $\partial f(x) = \nabla\Phi(x) + \partial h(x)$. The classic convergence theory fails to give any guarantees for this problem since f may be non-Lipschitz. In contrast, we find this problem class has a simple steepness bound from which guarantees for the classic subgradient method directly follow.

Lemma 11. *For any oracle $g_h(x) \in \partial h(x)$, the oracle $\nabla\Phi(x) + g_h(x) \in \partial f(x)$ is*

$$\left(\frac{v+1}{v}L_\Phi^{1/v}t\right)^{v/(v+1)} + 2L_h\text{-steep on } \mathbb{R}^d.$$

Proof. Consider any $x^* \in X^*$ and $g^* := -\nabla\Phi(x^*) \in \partial h(x^*)$. 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 $f(x^*) = l(x^*)$. Further, since $l(x)$ has a (L_Φ, v)-Hölder continuous gradient, Lemma 9 bounds the size of $\nabla l(x)$ for any $x \in \mathbb{R}^d$ as

$$\|\nabla\Phi(x) + g^*\| \leq \left(\frac{v+1}{v}L_\Phi^{1/v}(l(x) - l(x^*))\right)^{v/(v+1)} \leq \left(\frac{v+1}{v}L_\Phi^{1/v}(f(x) - f(x^*))\right)^{v/(v+1)}.$$

The Lipschitz continuity of h implies $\|g^* - g_h(x)\| \leq 2L_h$, and so the triangle inequality completes the proof □

One could plug $\mathcal{L}(t) = \left(\frac{v+1}{v} L_\Phi^{1/v} t\right)^{v/(v+1)} + 2L_h$ into Theorem 5 and evaluate the supremum to produce a convergence guarantee. For ease of presentation, we weaken our steepness bound to the following, which may be up to a factor of two larger,

$$\mathcal{L}(t) = 2\max\left\{\left(\frac{v+1}{v} L_\Phi^{1/v} t\right)^{v/(v+1)},\ 2L_h\right\}.$$

Then Theorem 5 immediately gives the following $O(1/\sqrt{T})$ convergence rate (for simplicity, we state the bound for constant step size).

Corollary 12 (Additive Composite Convergence)**.** *For any deterministic subgradient oracle $\nabla\Phi(x) + g_h(x)$, under the constant step size $\alpha_k = D_0/\sqrt{T+1}$, the iteration* (1) *satisfies*

$$\min_{k=0\ldots T}\{\delta(x_k)\} \le \max\left\{2^{v+1}\left(\frac{v+1}{v}\right)^v \frac{L_\Phi D_0^{v+1}}{(T+1)^{(v+1)/2}},\ \frac{4L_h D_0}{\sqrt{T+1}}\right\}.$$

Up to small factors, the first term in the above maximum matches the convergence rate on functions with Hölderian gradient like Φ (see Corollary 10) and the second term matches the convergence rate on Lipschitz continuous functions like h (see Theorem 1). 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 result 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 $\mathbb{E}_\xi\ g_h(x;\xi) \in \partial h(x)$ satisfying $\mathbb{E}_\xi\|g_h(x;\xi)\|^2 \le L^2$. 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 Lipschitz continuity. Simple arguments give a steepness bound for this problem and the following $O(1/T)$ convergence rate.

Corollary 13 (Quadratically Regularized Convergence)**.** *For the decreasing step sizes*

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

and stochastic subgradient oracle $g_h(x;\xi) + \lambda x$, the iteration (2) *satisfies*

$$\mathbb{E}_{\xi_{0\ldots T}}\left[\delta\left(\frac{2}{(T+1)(T+2)}\sum_{k=0}^{T}(k+1)x_k\right)\right] \le \frac{24L^2}{\lambda(T+2)} + \frac{36\lambda D_0^2}{(T+1)(T+2)}.$$

Proof. Consider any $x^* \in X^*$ and $g^* := -\lambda x^* \in \partial h(x^*)$. Since $(\lambda/2)\|x\|^2$ has λ-Lipschitz gradient, the same argument used in Lemma 11 shows $\|\lambda x + g^*\| \leq \sqrt{2\lambda\delta(x)}$. Applying Jensen's inequality and the assumed subgradient bound implies

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

Thus our stochastic oracle is $\sqrt{6\lambda t + 6L^2}$-steep. Noting f is λ-strongly convex, our bound follows from Theorem 8. □

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 with feature vector $w_i \in \mathbb{R}^d$ labeled $y_i \in \{-1, 1\}$. 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 has $L^2 = \frac{1}{n}\sum_{i=1}^n \|w_i\|^2$.

Much work has previously been done to give guarantees for SVMs. If one adds the constraint that x lies in some large ball Q (which will then be projected onto each iteration), the classic strongly convex rate can be applied [13]. A similar approach utilized in [6] 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). The specialized mirror descent method proposed by Lu [7] gives 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 [3] and [15], respectively, which both avoid needing to assume subgradient bounds.

2.4 Alternative Characterizations of Deterministic Steepness

Here we give an alternative interpretation of bounding the size of subgradients, either absolutely or with steepness on some convex open set $U \subseteq \mathbb{R}^d$. First we consider the classic model. Suppose a convex function f has $\|g\| \leq L$ for all $x \in U$ and $g \in \partial f(x)$. This is equivalent to f being L-Lipschitz continuous on U and can be restated as the following linear upper bound holding for each $x \in U$

$$\delta(y) \leq \delta(x) + L\|y - x\| \quad (\forall y \in U),$$

where $\delta(x) := f(x) - \inf_{x'\in U} f(x')$.

This characterization shows the limitation to linear growth of the classic model (i.e., constant steepness). In the following proposition, we give similar upper bound characterizations for linear and square root steepness, which can be seen as allowing up to exponential and quadratic growth, respectively.

Proposition 14. *A convex function f is $L_1 t + L_0$-steep on some open convex $U \subseteq \mathbb{R}^d$ if and only if the following exponential upper bound holds for each $x \in U$*

$$\delta(y) \leq \left(\delta(x) + \frac{L_0}{L_1}\right) \exp(L_1 \|y - x\|) - \frac{L_0}{L_1} \quad (\forall y \in U).$$

Similarly, a convex function f is $\sqrt{L_1 t + L_0^2}$-steep on some open convex $U \subseteq \mathbb{R}^d$ if and only if the following quadratic upper bound holds for each $x \in U$

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

Proof. First we prove the forward direction of both claims. 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 $h(t) = \delta(x + tv)$ denote the restriction of δ to this line. Notice that $h(0) = \delta(x)$ and $h(\|y - x\|) = \delta(y)$. The convexity of h implies it is differentiable almost everywhere in the interval $[0, \|y - x\|]$. Thus h satisfies the following, for almost every $t \in [0, \|y - x\|]$,

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

For any $L_1 t + L_0$-steep function, this gives the differential inequality of

$$|h'(t)| \leq L_1 h(t) + L_0.$$

Standard calculus arguments imply $h(t) \leq (h(0) + L_0/L_1)\exp(L_1 t) - L_0/L_1$, which is equivalent to our claimed upper bound at $t = \|y - x\|$.

For any $\sqrt{L_1 t + L_0^2}$-steep function, this gives the differential inequality of

$$|h'(t)| \leq \sqrt{L_1 h(t) + L_0^2}.$$

Standard calculus arguments imply $h(t) \leq h(0) + \frac{L_1}{4}t^2 + t\sqrt{L_1 h(0) + L_0^2}$, which is equivalent to our claimed upper bound at $t = \|y - x\|$.

Now we prove the reverse direction of both claims. Let $u_x(y)$ denote either of our upper bounds given by some $x \in U$. Further, let D_v the directional derivative operator in some unit direction $v \in \mathbb{R}^d$. Then for any subgradient $g(x) \in \partial f(x)$,

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

where the first inequality uses the definition of D_v and the second uses the fact that u_x upper bounds δ. A simple calculation shows our first upper bound has $D_v u_x(x) \leq L_1 \delta(x) + L_0$ and our second upper bound has $D_v u_x(x) \leq \sqrt{L_1 \delta(x) + L_0^2}$. Then both of our steepness bounds follow by taking $v = g(x)/\|g(x)\|$. □

3 Convergence Proofs

Each of our extended convergence theorems follows from essentially the same proof as its classic counterpart. Only minor modification is needed to replace L-Lipschitz continuity by $\mathcal{L}(t)$-steepness. The central inequality in analyzing subgradient methods is the following.

Lemma 15. *Consider any μ-strongly convex function f. For any $x, y \in Q$ and $\alpha > 0$,*

$$\mathbb{E}_\xi \|P_Q(x - \alpha g(x;\xi)) - y\|^2 \leq (1 - \mu\alpha)\|x - y\|^2 - 2\alpha(f(x) - f(y)) + \alpha^2 \mathbb{E}_\xi \|g(x;\xi)\|^2.$$

Note that this holds for any convex function at $\mu = 0$.

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

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

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

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

Applying the definition of strong convexity completes the proof. □

Let $D_k = \mathbb{E}_{\xi_{0\ldots T}} \mathbf{dist}(x_k, X^*)$ denote the expected distance from each iterate to the set of minimizers. Each of our proofs follows the same general outline: use Lemma 15 to set up a telescoping inequality on D_k, then sum the telescope.

3.1 Proof of Theorem 5

From Lemma 15 with $\mu = 0$, $x = x_k$, $y = P_{X^*}(x_k)$, and $\alpha = \alpha_k / \|g(x_k)\|$, it follows that

$$D_{k+1}^2 \leq D_k^2 - \frac{2\alpha_k \delta(x_k)}{\|g(x_k)\|} + \alpha_k^2 \leq D_k^2 - \frac{2\alpha_k \delta(x_k)}{\mathcal{L}(\delta(x_k))} + \alpha_k^2,$$

where the second inequality uses $\mathcal{L}(t)$-steepness of $g(x)$. Inductively applying this implies

$$0 \leq D_{T+1}^2 \leq D_0^2 - \sum_{k=0}^{T} \frac{2\alpha_k \delta(x_k)}{\mathcal{L}(\delta(x_k))} + \sum_{k=0}^{T} \alpha_k^2.$$

Thus

$$\min_{k=0\ldots T} \left\{ \frac{\delta(x_k)}{\mathcal{L}(\delta(x_k))} \right\} \leq \frac{D_0^2 + \sum_{k=0}^{T} \alpha_k^2}{2\sum_{k=0}^{T} \alpha_k}.$$

Applying the sup-inverse of $t/\mathcal{L}(t)$ completes the proof. □

3.2 Proof of Theorem 7

From Lemma 15 with $\mu = 0$, $x = x_k$, $y = P_{X^*}(x_k)$, and $\alpha = \alpha_k$, it follows that

$$\begin{aligned} D_{k+1}^2 &\leq D_k^2 - \mathbb{E}_{\xi_{0\ldots T}}[2\alpha_k \delta(x_k)] + \alpha_k^2 \mathbb{E}_{\xi_{0\ldots T}} \|g(x_k, \xi_k)\|^2 \\ &\leq D_k^2 - \mathbb{E}_{\xi_{0\ldots T}}\big[(2\alpha_k - L_1\alpha_k^2)\delta(x_k)\big] + L_0^2\alpha_k^2, \end{aligned}$$

where the second inequality uses the steepness of $g(x;\xi)$. Inductively applying this implies

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

The convexity of f gives

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

completing the proof. □

3.3 Proof of Theorem 8

Our proof follows the style of [6]. 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

$$(k+1)\alpha_k^{-1} = (k+2)(\alpha_{k+1}^{-1} - \mu). \tag{3}$$

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

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

From Lemma 15 with $x = x_k$, $y = P_{X^*}(x_k)$, and $\alpha = \alpha_k$, it follows that

$$\begin{aligned} D_{k+1}^2 &\leq (1 - \mu\alpha_k)D_k^2 - \mathbb{E}_{\xi_{0\ldots T}}[2\alpha_k \delta(x_k)] + \alpha_k^2 \mathbb{E}_{\xi_{0\ldots T}} \|g(x_k, \xi_k)\|^2 \\ &\leq (1 - \mu\alpha_k)D_k^2 - \mathbb{E}_{\xi_{0\ldots T}}\big[(2\alpha_k - L_1\alpha_k^2)\delta(x_k)\big] + L_0^2\alpha_k^2, \end{aligned}$$

where the second inequality uses the steepness of $g(x;\xi)$. Multiplying by $(k+1)/\alpha_k$ yields

$$(k+1)\alpha_k^{-1}D_{k+1}^2 \leq (k+1)(\alpha_k^{-1} - \mu)D_k^2 - \mathbb{E}_{\xi_{0\ldots T}}[(k+1)(2 - L_1\alpha_k)\delta(x_k)] + L_0^2(k+1)\alpha_k.$$

Notice that this inequality telescopes due to (3). Inductively applying this implies

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

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

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

Observe that the coefficients of each $\delta(x_k)$ above are positive due to (4). Then the convexity of f gives our first convergence bound. From (4), we know $2 - L_1\alpha_k \geq 1$ for all $k \geq 0$. Then the previous inequality can be weakened to

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

The convexity of f gives our second convergence bound. □

4 Extensions of our Convergence Rates

4.1 Convergence Beyond Exponential Growth

Early in the development of subgradient methods, Shor [14] observed that the normalized subgradient method (1) enjoys some form of convergence guarantee for any convex function with a nonempty set of minimizers. Shor used the same elementary argument underlying Theorem 5 to show for any minimizer $x^* \in X^*$: either some $k = 0 \dots T$ has $x_k \in X^*$ or

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

Thus, for any convex function, the subgradient method has convergence in terms of this inner product (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^*.

To turn this into an objective gap convergence rate for general convex problems, one needs to convert having small "subgradient hyperplane distance to a minimizer" into having small objective gap. The immediate convergence theorem based on this idea is the following.

Theorem 16. *Consider any convex f and subgradient oracle. Fix some $x^* \in X^*$. If*

$$\left(\frac{g(x)}{\|g(x)\|}\right)^T(x - x^*) \leq \epsilon \implies \delta(x) \leq v(\epsilon) \quad (\forall x \in Q, \epsilon > 0)$$

for some function $v\colon \mathbb{R}_+ \to \mathbb{R}_+ \cup \{\infty\}$, then the iteration (1) satisfies

$$\min_{k=0\dots T}\{\delta(x_k)\} \leq v\left(\frac{\|x_0 - x^*\|^2 + \sum_{k=0}^{T}\alpha_k^2}{2\sum_{k=0}^{T}\alpha_k}\right).$$

The primary difficulty in applying the above theorem to a particular problem is in identifying a function $v\colon \mathbb{R}_+ \to \mathbb{R}_+ \cup \{\infty\}$ where the necessary implication holds. However, this approach can circumvent the limitation of Theorem 5 to having at most exponential growth. For example, $f(x) = \exp(\|x\|^2)$ satisfies this implication with $v(\epsilon) = \exp(\epsilon^2)$. Theorem 5 can be viewed as a one particular way to construct a suitable v: For any $\mathcal{L}(t)$-steep oracle,

$$\left(\frac{g(x)}{\|g(x)\|}\right)^T(x - x^*) \leq \epsilon \implies \frac{\delta(x)}{\mathcal{L}(\delta(x))} \leq \epsilon \implies \delta(x) \leq \sup\left\{t \mid \frac{t}{\mathcal{L}(t)} \leq \epsilon\right\}.$$

4.2 Improved Convergence Without Strong Convexity

Strong convexity is stronger than necessary to achieve many of the standard improvements in convergence rate for smooth optimization problems [8]. Instead the weaker condition of requiring quadratic growth away from the set of minimizer suffices. We find that this weaker condition is also sufficient for (2) to have a convergence rate of $O(1/T)$.

We say a function f has μ-quadratic growth if all $y \in Q$ satisfy

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

The proof of Theorem 8 only uses strong convexity once for the following inequality:

$$g(x_k)^T(x_k - P_{X^*}(x_k)) \geq f(x_k) - \min_{x \in Q} f(x) + \frac{\mu}{2}\mathbf{dist}(x_k, X^*)^2.$$

Having μ-quadratic growth is sufficient to give this inequality, weakened by a factor of $1/2$:

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

Then simple modifications of the proof of Theorem 8 give a $O(1/T)$ convergence rate.

Theorem 17. *Consider any convex function f with μ-quadratic growth and $\sqrt{L_1 t + L_0^2}$-steep stochastic subgradient oracle on Q. Then for the decreasing sequence of step sizes*

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

the iteration (2) *satisfies*

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

Acknowledgments. The author thanks Jim Renegar for providing feedback on an early draft of this work.

References

[1] Sébastien Bubeck. Convex Optimization: Algorithms and Complexity. *Found. Trends Mach. Learn.*, 8(3-4):231–357, November 2015.

[2] Damek Davis and Benjamin Grimmer. Proximally Guided Stochastic Subgradient Method for Nonsmooth, Nonconvex Problems. *ArXiv e-prints*, 1707.03505, July 2017.

[3] John Duchi and Yoram Singer. Efficient Online and Batch Learning Using Forward Backward Splitting. *J. Mach. Learn. Res.*, 10:2899–2934, December 2009.

[4] Benjamin Grimmer. Radial Subgradient Method. *To appear in SIAM Journal on Optimization.*

[5] Elad Hazan and Satyen Kale. Beyond the Regret Minimization Barrier: Optimal Algorithms for Stochastic Strongly-convex Optimization. *J. Mach. Learn. Res.*, 15(1):2489–2512, January 2014.

[6] Simon Lacoste-Julien, Mark Schmidt, and Francis Bach. A simpler approach to obtaining an O(1/t) convergence rate for the projected stochastic subgradient method. *ArXiv e-prints*, 1212.2002, December 2012.

[7] Haihao Lu. "Relative-Continuity" for Non-Lipschitz Non-Smooth Convex Optimization using Stochastic (or Deterministic) Mirror Descent. *ArXiv e-prints*, 1710.04718, October 2017.

[8] Ion Necoara, Yurii Nesterov, and Francois Glineur. Linear convergence of first order methods for non-strongly convex optimization. *To appear in Mathematical Programming.*

[9] Angelia Nedi and Soomin Lee. On stochastic subgradient mirror-descent algorithm with weighted averaging. *SIAM Journal on Optimization*, 24(1):84–107, 2014.

[10] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course.* Springer Publishing Company, Incorporated, 1 edition, 2004.

[11] Alexander Rakhlin, Ohad Shamir, and Karthik 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. Omnipress.

[12] James Renegar. "Efficient" Subgradient Methods for General Convex Optimization. *SIAM Journal on Optimization*, 26(4):2649–2676, 2016.

[13] Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: primal estimated sub-gradient solver for svm. *Mathematical Programming*, 127(1):3–30, Mar 2011.

[14] Naun Zuselevich Shor. *Minimization Methods for Non-Differentiable Functions*, page 23. Springer Berlin Heidelberg, Berlin, Heidelberg, 1985.

[15] Jin Yu, S.V.N. Vishwanathan, Simon Günter, and Nicol N. Schraudolph. A Quasi-Newton Approach to Nonsmooth Convex Optimization Problems in Machine Learning. *J. Mach. Learn. Res.*, 11:1145–1200, March 2010.