Randomized Smoothing for (Parallel) Stochastic Optimization

John C. Duchi JDUCHI@EECS.BERKELEY.EDU
Peter L. Bartlett BARTLETT@EECS.BERKELEY.EDU
Martin J. Wainwright WAINWRIG@EECS.BERKELEY.EDU
University of California, Berkeley, Berkeley, CA 94720

Abstract

By combining randomized smoothing techniques with accelerated gradient methods, we obtain convergence rates for stochastic optimization procedures, both in expectation and with high probability, that have optimal dependence on the variance of the gradient estimates. To the best of our knowledge, these are the first variance-based convergence guarantees for non-smooth optimization. A combination of our techniques with recent work on decentralized optimization yields order-optimal parallel stochastic optimization algorithms. We give applications of our results to several statistical machine learning problems, providing experimental results demonstrating the effectiveness of our algorithms.

1. Introduction

In this paper, we develop and analyze procedures for solving a class of stochastic optimization problems that frequently arise in machine learning and statistics. Formally, consider a collection $\{F(\cdot\,;\xi), \xi \in \Xi\}$ of closed convex functions, each with domain containing the closed convex set $\mathcal{X} \subseteq \mathbb{R}^d$. Let P be a probability distribution over the sample space Ξ and consider the expected convex function $f : \mathcal{X} \to \mathbb{R}$ defined via

$$f(x) := \mathbb{E}\big[F(x;\xi)\big] \;=\; \int_{\Xi} F(x;\xi) dP(\xi). \qquad (1)$$

We focus on potentially non-smooth stochastic optimization problems of the form

$$\min_{x \in \mathcal{X}} \; \big\{f(x) + \varphi(x)\big\}, \qquad (2)$$

where $\varphi : \mathcal{X} \to \mathbb{R}$ is a known regularizing function, which may be non-smooth. The problem (2) has wide applicability in machine learning problems; essentially all empirical risk-minimization procedures fall into the form (2), where the distribution P in the definition (1) is either the empirical distribution over some sample of n datapoints, or it may simply be the (unknown) population distribution of the samples $\xi \in \Xi$.

As a first motivating example, consider support vector machines (SVMs) (Cortes & Vapnik, 1995). In this setting, the loss F and regularizer φ are defined by

$$F(x;\xi) = \left[1 - \langle \xi, x \rangle\right]_+ \quad \text{and} \quad \varphi(x) = \frac{\lambda}{2} \left\|x\right\|_2^2, \qquad (3)$$

where $[\alpha]_+ := \max\{\alpha, 0\}$. Here, the samples take the form $\xi = ba$, where $b \in \{-1, +1\}$ is the label of the data point $a \in \mathbb{R}^d$, and the goal of the learner is to find an $x \in \mathbb{R}^d$ that separates positive b from negative.

More complicated examples include structured prediction (Taskar, 2005), inverse convex or combinatorial optimization (e.g. Ahuja & Orlin, 2001) and inverse reinforcement learning or optimal control (Abbeel, 2008). The learner receives examples of the form (ξ, ν) where ξ is the input to a system (for example, in NLP applications ξ may be a sentence) and $\nu \in \mathcal{V}$ is a target (e.g. the parse tree for the sentence ξ) that belongs to a potentially complicated set $\mathcal{V}$. The goal of the learner is to find parameters x so that $\nu = \operatorname{argmax}_{v \in \mathcal{V}} \langle x, \phi(\xi, v) \rangle$, where ϕ is a feature mapping. Given a loss $\ell(\nu, v)$ measuring the penalty for predicting $v \neq \nu$, the objective $F(x; (\xi, \nu))$ is

$$\max_{v \in \mathcal{V}} \left[\ell(\nu, v) + \langle x, \phi(\xi, v) \rangle - \langle x, \phi(\xi, \nu) \rangle\right]. \qquad (4)$$

These examples highlight one of the two main difficulties in solving the problem (2). The first, which is by now well known (e.g. Nemirovski et al., 2009), is that it is often difficult to compute the integral (1). Indeed, when ξ is high-dimensional, the integral cannot be efficiently computed, and in machine learning problems, we rarely even know the distribution

Appearing in *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.

P. Thus, throughout this work, we assume only that we have access to i.i.d. samples $\xi \sim P$, and consequently we adopt the current method of choice and focus on stochastic gradient procedures for solving the convex program (2) (Nemirovski et al., 2009; Lan, 2012; Duchi & Singer, 2009; Xiao, 2010). In the oracle model we assume, the optimizer issues a query vector x, after which the oracle samples a point ξ i.i.d. according to P and returns a vector $g \in \partial_x F(x;\xi)$. The second difficulty of solving the problem (2), which in this stochastic setting is the main focus of our paper, is that the functions F and expected function f may be non-smooth (i.e. non-differentiable).

When the objective function f is smooth, meaning that it has Lipschitz continuous gradient, recent work by Juditsky et al. (2008) and Lan (2012) has shown that if the variance of a stochastic gradient estimate is at most σ^2 then stochastic optimization procedures may obtain convergence rate $\mathcal{O}(\sigma/\sqrt{T})$. Of particular relevance here is that if instead of receiving single stochastic gradient estimates the algorithm receives m unbiased estimates of the gradient, the variance of the gradient estimator is reduced by a factor of m. Dekel et al. (2011) exploit this fact to develop asymptotically order-optimal distributed optimization algorithms. The dependence on the variance is essential for improvements gained through parallelism; however, to the best of our knowledge there has thus far been no work on *non-smooth* stochastic problems for which a reduction in the variance of the stochastic subgradient estimate gives an improvement in convergence rates.

The starting point for our approach is a convolution-based smoothing technique amenable to non-smooth stochastic optimization problems. Let μ be a density and consider the smoothed objective function

$$f_\mu(x) := \int f(x+y)\mu(y)dy = \mathbb{E}_\mu[f(x+Z)], \qquad (5)$$

where Z is a random variable with density μ. The function f_μ is convex whenever f is convex, and f_μ is guaranteed to be differentiable (e.g. Bertsekas, 1973). The important aspect of the convolution (5) to note is that by Fubini's theorem, we can write

$$f_\mu(x) = \int \mathbb{E}_\mu[F(x+Z;\xi) \mid \xi] dP(\xi), \qquad (6)$$

so that samples of subgradients $g \in \partial F(x+Z;\xi)$ for $Z \sim \mu$ and $\xi \sim P$ are unbiased gradient estimates of $f_\mu(x)$. By adding random perturbations, we do not assume we know anything about the function F, and the perturbations allow us to automatically smooth even complex F for which finding a smooth proxy is difficult (e.g. the structured prediction problem (4)).

The main contribution of our paper is to develop algorithms for non-smooth stochastic optimization whose convergence rate depends on the variance σ^2 of the stochastic (sub)gradient estimate. In particular, we show that the ability to issue several queries to the stochastic oracle for the original objective (2) can give faster rates of convergence than a simple stochastic oracle (to our knowledge, this is the first such result for non-smooth optimization). Our theorems quantify the above statement in terms of expected values (Theorem 1) and, under an additional reasonable tail condition, with high probability (Theorem 2). In addition, we give extensions to the strongly convex case in Theorem 3. One consequence of our results is that a procedure that queries the non-smooth stochastic oracle for m subgradients at iteration t achieves rate of convergence $\mathcal{O}(RL_0/\sqrt{Tm})$ in expectation and with high probability. (Here L_0 is the Lipschitz constant of the function and R is the ℓ_2-radius of its domain.) This convergence rate is optimal up to constant factors, and our algorithms have applications in statistics, distributed optimization, and machine learning.

Notation For a parameter $p \in [1,\infty]$, we define the ℓ_p ball $B_p(x,u) := \{y \mid \|x-y\|_p \le u\}$. Addition of sets A and B is defined as the Minkowski sum $A+B = \{x \in \mathbb{R}^d \mid x = y+z, y \in A, z \in B\}$, and multiplication of a set A by a scalar α is defined to be $\alpha A := \{\alpha x \mid x \in A\}$. For any function f, we let $\operatorname{supp} f := \{x \mid f(x) \neq 0\}$ denote its support. Given a convex function f we use $\partial f(x)$ to denote its subdifferential at the point x. We define the shorthand notation $\|\partial f(x)\| = \sup\{\|g\| \mid g \in \partial f(x)\}$. The dual norm $\|\cdot\|_*$ of the norm $\|\cdot\|$ is defined as $\|z\|_* := \sup_{\|x\| \le 1} \langle z, x\rangle$. A function f is L_0-Lipschitz with respect to the norm $\|\cdot\|$ over $\mathcal{X}$ if $|f(x)-f(y)| \le L_0 \|x-y\|$ for all $x, y \in \mathcal{X}$. The gradient of f is L_1-Lipschitz continuous with respect to the norm $\|\cdot\|$ over $\mathcal{X}$ if

$$\|\nabla f(x) - \nabla f(y)\|_* \le L_1 \|x-y\| \quad \text{for } x, y \in \mathcal{X}.$$

A function ψ is strongly convex with respect to a norm $\|\cdot\|$ over $\mathcal{X}$ if for all $x, y \in \mathcal{X}$,

$$\psi(y) \ge \psi(x) + \langle \nabla\psi(x), y-x\rangle + \frac{1}{2}\|x-y\|^2.$$

Given a convex and differentiable function ψ, the associated Bregman divergence between x and y is $D_\psi(x,y) := \psi(x) - \psi(y) - \langle \nabla\psi(y), x-y\rangle$. We write drawing ξ from the distribution P as $\xi \sim P$.

2. Algorithms and Main Results

We begin by describing our base algorithm, which builds off of Tseng's (2008) work on accelerated gradient methods. The method generates three sequences of

points, denoted $\{x_t, y_t, z_t\} \in \mathcal{X}^3$. The algorithm also requires a non-increasing sequence of smoothing parameters $\{u_t\} \subset \mathbb{R}$ to control the perturbation and—as is standard (Tseng, 2008; Lan, 2012; Xiao, 2010)—uses a proximal function ψ strongly convex with respect to the norm $\|\cdot\|$ to regularize the points. At iteration t, the algorithm computes stochastic gradients at m points drawn from a neighborhood around y_t:

(i) Draw random variables $\{Z_{i,t}\}_{i=1}^m$ i.i.d. according to the distribution μ.

(ii) Compute m stochastic (sub)gradients at the points $y_t + u_t Z_{i,t}$ for $i = 1, 2, \ldots, m$:

$$g_{i,t} \in \partial F(y_t + u_t Z_{i,t}, \xi_{i,t}), \quad \text{where } \xi_{i,t} \sim P, \text{ for } i = 1, 2, \ldots, m. \tag{7}$$

(iii) Compute the average $g_t = \frac{1}{m}\sum_{i=1}^m g_{i,t}$.

With the definition of the gradient sampling scheme, we can give our update scheme. We require scalars L_1 and η to control step sizes, and as in Tseng's work assume the sequence $\{\theta_t\} \subset \mathbb{R}$ satisfies $\theta_0 = 1$ and $\theta_t = 2/(1 + \sqrt{1 + 4/\theta_{t-1}^2})$. Starting from an initial point $x_0 \in \mathcal{X}$, we define the update recursions

$$y_t = (1 - \theta_t)x_t + \theta_t z_t \tag{8a}$$

$$z_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \sum_{\tau=0}^t \frac{1}{\theta_\tau} \langle g_\tau, x \rangle + \sum_{\tau=0}^t \frac{1}{\theta_\tau}\varphi(x) + \left[\frac{L_1}{u_t} + \frac{\eta\sqrt{t+1}}{\theta_{t+1}}\right] D_\psi(x, x_0) \right\} \tag{8b}$$

$$x_{t+1} = (1 - \theta_t)x_t + \theta_t z_{t+1}. \tag{8c}$$

In prior work on accelerated schemes for stochastic and non-stochastic optimization (Tseng, 2008; Lan, 2012; Xiao, 2010), the term L_1 is set equal to the Lipschitz constant of ∇f; we use L_1/u_t to allow varying amounts of smoothness due to the shrinking sequence of u_t; the damping term $\eta\sqrt{t}/\theta_t$ provides control over the fluctuations induced by the random vector g_t.

2.1. Convergence Rates for Convex Objectives

We now state our main results on the convergence rate of the randomized smoothing procedure (7) with accelerated dual averaging updates (8a)–(8c), providing proofs in the appendices. We begin with our main assumption, which ensures that the smoothing operator and smoothed function f_μ are relatively well-behaved.

Assumption A (Smoothing)**.** *The random variable Z is zero-mean with density μ and there are constants L_0 and L_1 such that for $u > 0$, $\mathbb{E}[f(x+uZ)] \le f(x) + L_0 u$, and $\mathbb{E}[f(x+uZ)]$ has $\frac{L_1}{u}$-Lipschitz continuous gradient.*

Since the function $f_\mu = f * \mu$ is smooth, Assumption A ensures that f_μ is close to f but not too "jagged." We elaborate on conditions under which Assumption A holds after stating our first two theorems:

Theorem 1. *Define $u_t = \theta_t u$ and μ_t to be the density of $u_t Z$. For any $x^* \in \mathcal{X}$*

$$\mathbb{E}[f(x_T) + \varphi(x_T)] - [f(x^*) + \varphi(x^*)] \le \frac{6L_1\psi(x^*)}{Tu} + \frac{2\eta\psi(x^*)}{\sqrt{T}} + \frac{1}{\eta T}\sum_{t=1}^T \frac{1}{\sqrt{t}}\mathbb{E}\left\|e_t\right\|_*^2 + \frac{4L_0 u}{T},$$

where $e_t := \nabla f_{\mu_t}(y_t) - g_t$ is the error in the gradient estimate.

The preceding result, which provides convergence in expectation, can be extended to bounds that hold with high probability under suitable tail conditions on the error $e_t := \nabla f_\mu(y_t) - g_t$. In particular, let $\mathcal{F}_t$ denote the σ-field of the random variables $g_{i,s}$, $i = 1, \ldots, m$ and $s = 0, \ldots, t$. To obtain high-probability results, we make the following assumption.

Assumption B (Sub-Gaussian errors)**.** *The error is $(\|\cdot\|_*, \sigma)$* sub-Gaussian*, meaning that there exists a constant $\sigma > 0$ such that*

$$\mathbb{E}[\exp(\|e_t\|_*^2/\sigma^2) \mid \mathcal{F}_{t-1}] \le \exp(1) \quad \text{for } t \in \mathbb{N}. \tag{9}$$

Past work on smooth stochastic optimization (Juditsky et al., 2008; Lan, 2012; Xiao, 2010) has imposed this type of tail assumption, and Corollary 4 gives sufficient conditions for the assumption to hold.

Theorem 2. *In addition to the conditions of Theorem 1, suppose that $\mathcal{X}$ is compact with $\|x - x^*\| \le R$ for all $x \in \mathcal{X}$ and that Assumption B holds. With probability at least $1 - 2\delta$,*

$$f(x_T) + \varphi(x_T) - [f(x^*) + \varphi(x^*)] \le \frac{6L_1\psi(x^*)}{Tu} + \frac{4L_0 u}{T} + \frac{4\eta\psi(x^*)}{\sqrt{T}} + \frac{1}{\eta T}\sum_{t=1}^T \frac{1}{\sqrt{t}}\mathbb{E}[\|e_t\|_*^2] + \frac{4\sigma^2 \max\left\{\log\frac{1}{\delta}, \sqrt{3e\log(T)\log\frac{1}{\delta}}\right\}}{\eta T} + \frac{\sigma R\sqrt{\log\frac{1}{\delta}}}{\sqrt{T}}.$$

We now give two corollaries that help to explain the bounds Theorem 1 provides. In each corollary we assume that the point $x^* \in \mathcal{X}$ satisfies $\psi(x^*) \le R^2$, and we use the proximal function $\psi(x) = \frac{1}{2}\|x\|_2^2$.

Corollary 1. *Let μ be uniform on $\{z \mid \|z\|_2 \le u\}$ and assume $\mathbb{E}[\|\partial F(x;\xi)\|_2^2] \le L_0^2$. If we set $u = Rd^{1/4}$ and $\eta = L_0/R\sqrt{m}$, then*

$$\mathbb{E}[f(x_T)+\varphi(x_T)]-[f(x^*)+\varphi(x^*)] \le \frac{10L_0 R d^{1/4}}{T} + \frac{5L_0 R}{\sqrt{Tm}}.$$

Corollary 2. *Let μ be the d-dimensional normal distribution with covariance $u^2 I$ and assume $F(\cdot;\xi)$ is L_0-Lipschitz for $\xi \in \Xi$. With smoothing $u = Rd^{-1/4}$,*

$$\mathbb{E}[f(x_T)+\varphi(x_T)]-[f(x^*)+\varphi(x^*)] \le \frac{10L_0Rd^{1/4}}{T}+\frac{5L_0R}{\sqrt{Tm}}.$$

There are several objectives $f+\varphi$ with domains $\mathcal{X}$ for which the natural geometry is non-Euclidean, which motivates the mirror descent family of algorithms (Nemirovski & Yudin, 1983). By using different distributions μ for the random perturbations $Z_{i,t}$ in (7), we can take advantage of non-Euclidean geometry. Here we give an example that is quite useful for problems in which the optimizer x^* is sparse; for example, the optimization set $\mathcal{X}$ may be a simplex or ℓ_1-ball, or $\varphi(x) = \lambda \|x\|_1$. The idea in this corollary is that we achieve a pair of dual norms that may give better performance than the ℓ_2-ℓ_2 pair above.

Corollary 3. *Let μ be the uniform density on $B_\infty(0,u)$ and assume that $F(\cdot;\xi)$ is L_0-Lipschitz continuous with respect to the ℓ_1-norm over $\mathcal{X}+B_\infty(0,u)$ for $\xi \in \Xi$. Use the prox function $\psi(x) = \frac{1}{2(p-1)}\|x\|_p^2$ for $p = 1+1/\log d$, set $u = R\sqrt{d\log d}$ and $\eta = L_0/R\sqrt{m\log d}$. Then*

$$\mathbb{E}[f(x_T)+\varphi(x_T)] - [f(x^*)+\varphi(x^*)] = \mathcal{O}\left(\frac{L_0\|x^*\|_1\sqrt{d\log d}}{T} + \frac{L_0\|x^*\|_1\log d}{\sqrt{Tm}}\right).$$

The dimension dependence of $\sqrt{d\log d}$ on the leading $1/T$ term in the corollary is weaker than the $d^{1/4}$ dependence in the earlier corollaries, so for very large m the corollary is not as strong as one desires for problems with non-Euclidean geometry. But for large T, the $1/\sqrt{Tm}$ terms dominate the convergence rates.

Inspection of the above corollaries shows that we have achieved our goal: we have convergence rates that provably improve with the "mini-batch" (or gradient sample) size m. Additionally, Agarwal et al. (2012) give lower bounds for stochastic optimization that show these rates are optimal: they cannot be improved by more than constant factors.

Our final corollary specializes the high probability convergence result in Theorem 2 by showing that the error is sub-Gaussian (9). We state the corollary for problems with Euclidean geometry, but it is clear that similar results hold for non-Euclidean geometry as above.

Corollary 4. *Assume that $F(\cdot,\xi)$ is L_0-Lipschitz with respect to the ℓ_2-norm. Let $\psi(x) = \frac{1}{2}\|x\|_2^2$ and assume that $\mathcal{X}$ is compact with $\|x-x^*\|_2 \le R$ for $x, x^* \in \mathcal{X}$. Using a smoothing distribution μ uniform on $B_2(0,u)$, smoothing parameter $u = Rd^{1/4}$, and damping parameter $\eta = L_0/R\sqrt{m}$, with probability at least $1-\delta$,*

$$f(x_T)+\varphi(x_T) - f(x^*) - \varphi(x^*) = \mathcal{O}\bigg(\frac{L_0Rd^{1/4}}{T} + \cdots \frac{L_0R}{\sqrt{Tm}} + \frac{L_0R\sqrt{\log\frac{1}{\delta}}}{\sqrt{Tm}} + \frac{L_0R\max\{\log\frac{1}{\delta}, \log T\}}{T\sqrt{m}}\bigg).$$

Algorithm 1 Epoch-based stochastic gradient

Input: strong convexity parameter λ, Lipschitz parameter L_1, smoothing u, damping η
for $i = 1$ to k **do**
  Set $\eta(i) = \eta \cdot 2^i$ and $u(i) = u \cdot 2^{-i}$
  Set $t(i) = \max\{12\eta(i)/\lambda, 4\sqrt{L_1/u(i)\lambda}\}$
  Run updates (8a)–(8c) for $t = t(i)$ timesteps with
  - Damping parameter $\eta(i)$
  - Gradients g_t computed via the procedure (7) with smoothing parameter $u(i)$
  - Initial points $x_0 = x(i-1)$ and $z_0 = x(i-1)$

  Set $x(i)$ to be the output of the previous step.
end for
Output: $x(k)$

2.2. Strongly Convex Optimization

When the objective $f+\varphi$ is strongly convex—for example, when the regularizer $\varphi(x) = \frac{\lambda}{2}\|x\|_2^2$—it is possible to attain rates of convergence faster than $1/\sqrt{T}$ (Hazan & Kale, 2011; Ghadimi & Lan, 2010). Following the idea of Hazan and Kale, we apply the algorithm (8a)–(8c) in a series of epochs. Let $x(i)$ denote the output of epoch i. Each epoch i consists of running the accelerated algorithm (8a)–(8c) for $t(i)$ iterations, except we replace $D_\psi(x, x_0)$ with $D_\psi(x, x(i-1))$. In addition to Assumption A, we make

Assumption C (Strong convexity)**.** *The function $f+\varphi$ is λ-strongly convex over $\mathcal{X}$: for any $x, y \in \mathcal{X}$ and any $f'(x) \in \partial f(x)$ and $\varphi'(x) \in \partial\varphi(x)$,*

$$f(y)+\varphi(y) \ge f(x)+\varphi(x)+\langle f'(x)+\varphi'(x), y-x\rangle + \frac{\lambda}{2}\|y-x\|_2^2.$$

Now we describe the parameters and execution of the algorithm. We perform the updates (7) and (8a)–(8c) with the damping parameter $\eta(i)$ and smoothing magnitude $u_t = u(i)$ fixed within each epoch, and set them to $\eta(i) = 2^i \cdot \eta$ and $u(i) = 2^{-i} \cdot u$ within epoch i. We set number of iterations $t(i)$ in epoch i to be

$$t(i) = \max\left\{4\sqrt{\frac{L_1}{u(i)\lambda}}, \frac{12\eta(i)}{\lambda}\right\}.$$

See Algorithm 1 for pseudocode. We obtain

Theorem 3. *Let $x(k)$ be the output of the epoch-based Algorithm 1 after k epochs, let σ^2 be a bound on the variance of the stochastic gradient estimate, and let $\mathsf{M} \geq f(x(0)) + \varphi(x(0)) - f(x^*) - \varphi(x^*)$. Then after $\log_2 \frac{\mathsf{M}}{\epsilon}$ epochs of Algorithm 1,*

$$\mathbb{E}[f(x(k)) + \varphi(x(k))] - [f(x^*) + \varphi(x^*)] \leq \epsilon + 5 \cdot \frac{\epsilon}{\mathsf{M}} \left[\frac{\sigma^2}{2\eta} + L_0 u\right]$$

and the number of updates (8a)–(8c) is bounded by

$$10\sqrt{\frac{L_1 \mathsf{M}}{u\lambda\epsilon}} + 24\frac{\mathsf{M}\eta}{\lambda\epsilon}.$$

To translate Theorem 3 into a slightly more intuitive bound, we set $\eta = \sigma^2/2\mathsf{M}$ and $u = \mathsf{M}/L_0$. These choices imply that after at most

$$10\sqrt{\frac{L_0 L_1}{\lambda\epsilon}} + 12\frac{\sigma^2}{\lambda\epsilon} \tag{10}$$

updates (8a)–(8c), we obtain optimization error 11ϵ.

We can apply these bounds to the SVM (3) and structured prediction (4) problems described in the introduction. The next corollary makes clear that computing m stochastic gradients in parallel (while employing our randomized perturbation techniques) yields linear speedup as a function of the batch size m:

Corollary 5. *Let the regularizer $\varphi(x) = \frac{\lambda}{2}\|x\|_2^2$. Assume that $\|\xi\|_2 \leq L_0$ (in the SVM case) or $\|\phi(\xi, \nu)\|_2 \leq L_0$ (for the structured prediction problem). Then with m subgradient evaluations and μ either the d-dimensional normal or uniform on $B_2(0, u)$, Algorithm 1 outputs an $\widehat{x}$ with*

$$\mathbb{E}[f(\widehat{x}) + \varphi(\widehat{x})] - [f(x^*) + \varphi(x^*)] = \mathcal{O}(\epsilon)$$

after at most

$$\mathcal{O}\left(\frac{L_0 d^{1/4}}{\sqrt{\lambda\epsilon}} + \frac{L_0}{\lambda m \epsilon}\right)$$

iterations of the method (8a)–(8c).

2.3. Robustness and Distributed Computing

In this final expository section, we give a few remarks on the robustness of our algorithms to the choice of stepsizes, and we describe techniques that allow parallelization. For concreteness, we focus our remarks on the strongly convex Algorithm 1, though essentially identical results hold for Section 2.1's procedures.

At first glance, Algorithm 1 appears to require knowledge of many of the parameters of the stochastic optimization problem (2). However, a closer inspection of the convergence rates guaranteed by Theorem 3 shows that the algorithm is robust to mis-specification of the values. Indeed, the smoothness Lipschitz constant L_1 appears in the algorithm only via $L_1/u(i)\lambda$, and the strong convexity parameter λ appears in $\eta(i)/\lambda$ and $L_1/u(i)\lambda$. In both cases, we can write L_1 and λ as a function of the other inputs u and η: indeed, set $C_u = L_1/\lambda u$ and $C_\eta = \eta/\lambda$. The optimization error in Theorem 3 has linear dependence on η^{-1}, and the number of iterations (10) also has linear dependence on $C_\eta = \eta/\lambda$, so the algorithm is robust to (and the global rate of convergence does not change with) mis-specification of η and λ. Similarly, Theorem 3's optimization error is linear in $L_0 u$, and the iteration bound (10) grows as $\sqrt{C_u}$, or is linear in $u^{-1/2}$. As a consequence, the epoch based algorithm—similar to Nemirovski et al.'s 2009 study of stochastic gradient descent—is robust to mis-specification of its input parameters. See also our simulations in Section 3.

Now we turn to exploiting parallel computation to achieve variance reductions, which yield convergence rate improvements for our algorithms. As motivation, note that Corollary 5 makes clear that the convergence rate of Algorithm 1 to optimization accuracy ϵ linearly improves with the number of samples m, at least until $m \geq \frac{L_0}{d^{1/4}\sqrt{\lambda\epsilon}}$. In our distributed setting, we assume we have n processors and take as our objective

$$f(x) := \frac{1}{n}\sum_{i=1}^n f_i(x) \quad \text{for} \quad f_i(x) := \mathbb{E}_{P_i}[F(x; \xi)]. \tag{11}$$

The distributed variant of Algorithm 1 is as follows: a centralized processor executes Algorithm 1 while gradient computation is distributed among several processors, each of which computes some number m of local gradients that are then averaged across the network. We replace the gradient sampling (7) at iteration t with the following: each processor i draws m i.i.d. samples $Z_{i,j} \sim \mu$, $j = 1, \ldots, m$, then the sampling scheme (7) is replaced with each processor i querying its local oracle at the m points $y_t + Z_{i,j}$, yielding stochastic gradients

$$g_{i,j} \in \partial F(y_t + Z_{i,j}; \xi_{i,j}) \quad \text{where} \quad \xi_{i,j} \stackrel{i.i.d.}{\sim} P_i \quad \text{for } j \in \{1, \ldots, m\}. \tag{12}$$

The network then computes $g_t = \frac{1}{nm}\sum_{i=1}^n \sum_{j=1}^m g_{i,j}$.

We may analyze the run-time and convergence rate of Algorithm 1 with the sampling scheme (12) using the technique developed by Dekel et al. (2011). Assume

one unit of time is required to sample and compute a single gradient vector as in (12), and let $c(n)$ be the amount of time required to achieve consensus in an n-node network. Then

Corollary 6. *Under the conditions of Theorem 3,*

$$\mathcal{O}\left(\max\left\{(m+c(n))\cdot\sqrt{\frac{L_0 L_1}{\lambda\epsilon}}, \frac{m+c(n)}{m}\cdot\frac{\sigma^2}{n\lambda\epsilon}\right\}\right)$$

units of time of the procedure (12) are sufficient for

$$\mathbb{E}[f(\widehat{x})+\varphi(\widehat{x})] - [f(x^*)+\varphi(x^*)] \leq \epsilon.$$

Corollary 6 gives us our desired result: order-optimal distributed optimization—i.e. linear speedup in the number n of processors—so long as $c(n)$ is at most the order of m.

3. Experimental Results

Though the results we have presented thus far are essentially theoretically unimprovable (Agarwal et al., 2012), it is important to understand their practical performance. In particular, we would like to understand the robustness properties of the algorithms—whether they still perform well under mis-specification of problem parameters—and how they compare to state-of-the-art stochastic optimization algorithms.

Robustness: We begin by studying the robustness of Algorithm 1 as a function of the parameters L_1, u, and η: the Lipschitz continuity constant of the gradient ∇f_μ of the smoothed objective, the amount of perturbation, and the damping stepsize. (For lack of space, we omit robustness results for our other algorithms.) As we discuss following Theorem 3, it is no loss of generality to view estimating the constant L_1 as choosing $1/u$, whence our robustness analysis reduces to studying the effects of the choices for u and η.

For our experiment, we generate $n = 1000$ vectors $a_i \in \{-1, 0, 1\}^d$ with $d = 200$ dimensions, each entry set to 0 with probability $\frac{1}{2}$ and uniformly $\{\pm 1\}$ with probability $\frac{1}{2}$. We set $b_i = \text{sign}(\langle a_i, w\rangle)$ for a random vector $w \in \mathbb{R}^d$ distributed as $N(0, I_{d\times d})$ and flip the signs of 10% of the b_i, and solve the SVM problem

$$\underset{x\in\mathbb{R}^d}{\text{minimize}} \quad \frac{1}{n}\sum_{i=1}^{n}\left[1 - b_i\langle a_i, x\rangle\right]_+ + \frac{\lambda}{2}\|x\|_2^2 \qquad (13)$$

with $\lambda = .1$. By construction, the Lipschitz constant $L_0 \approx 10$, so that we can estimate the values u and η Corollary 5 specifies. We run Algorithm 1 for 2000 iterations of the updates (8a)–(8c), using $m = 5$ samples. We perform 50 such experiments.

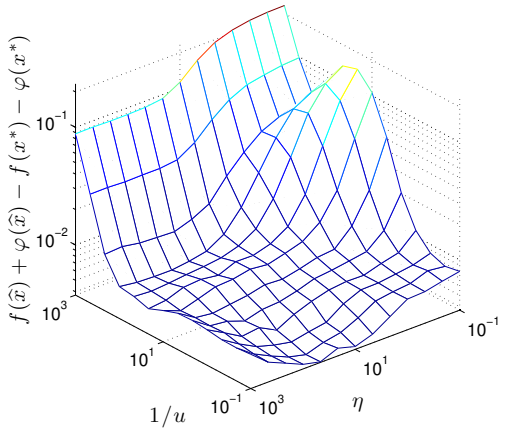

Figure 1. Optimality gap of the vector $\widehat{x}$ output by Algorithm 1 for the SVM problem (13) plotted against the inverse smoothing parameter $1/u$ and damping stepsize η.

In Figure 1, we plot the average optimality gap of the point $\widehat{x}$ Algorithm 1 selects. From the figure, we see that the performance of the method is nearly identical—achieving optimization accuracy better than 10^{-2} after 2000 iterations—so long as $(\eta, 1/u) \in [1, 1000] \times [.1, 100]$. The method suffers some performance degradation if η is too small, that is, $\eta \leq 1$ or so, or u is too small, that is, $1/u \geq 100$. Even when these extreme settings of u or η occur, however, the method still has optimization accuracy on the order of 10^{-1}. The breakdown points in the figure are intuitive: if η is too small, the damping in the proximal gradient update (8b) will not overcome the stochasticity of the vectors g_t; if u is too small, the perturbed function $\mathbb{E}[F(x + uZ; \xi)]$ is nearly non-smooth.

Metric Learning: Now we turn to showing the performance benefits of parallelization and the randomized perturbation methods. We begin with experiments based on a metric learning problem (Xing et al., 2003). For points $i, j = 1, \ldots, n$ we are given a vector $a_i \in \mathbb{R}^d$, and a measure $b_{ij} \geq 0$ of the similarity between the vectors a_i and a_j. (Here $b_{ij} = 0$ means that a_i and a_j are the same.) The goal is to learn a matrix X such that $\langle (a_i - a_j), X(a_i - a_j)\rangle \approx b_{ij}$. One method for doing so is to minimize the objective

$$f(X) = \frac{1}{\binom{n}{2}}\sum_{i\neq j}\left|\text{tr}\left(X(a_i - a_j)(a_i - a_j)^\top\right) - b_{ij}\right|$$
$$\text{subject to} \quad \text{tr}(X) \leq C, \; X \succeq 0. \qquad (14)$$

A stochastic gradient for this problem is simple: given a matrix X, choose a pair (i, j) uniformly at random, then compute the subgradient

$$\text{sign}\left[\langle (a_i - a_j), X(a_i - a_j)\rangle - b_{ij}\right](a_i - a_j)(a_i - a_j)^\top.$$

We solve ten random instances of the metric learning problem (14) with $d = 100$ and $n = 2000$, yielding an

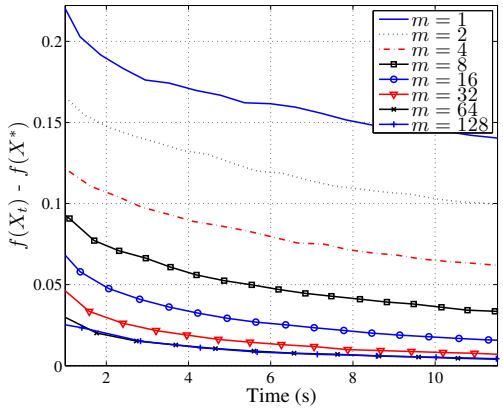

Figure 2. Optimization error in the metric learning problem (14) versus time in seconds. Each line indicates error when using m samples in the gradient estimate (7).

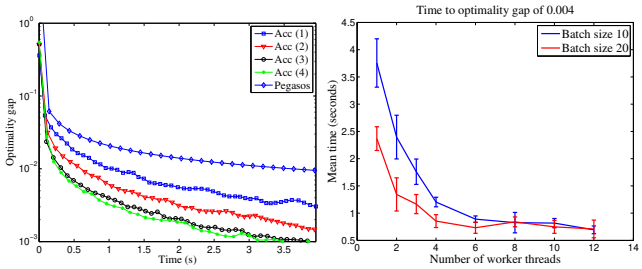

Figure 3. Left: Optimality gap of Pegasos and the accelerated strongly convex methods with $\{1, \ldots, 4\}$ threads versus time. Right: time to achieve optimality gap $\epsilon = .004$ for accelerated methods versus number of threads.

objective with $\approx 2 \cdot 10^6$ terms. In Figure 2, we plot the optimality gap $f(X_t) - \inf_{X^* \in \mathcal{X}} f(X^*)$ as a function of computation time. We plot several lines, each of which captures the performance of the algorithm using a different number m of samples in the smoothing step (7). Receiving more samples gives improvements in convergence rate as a function of time. Our theory also predicts that for $m \geq d$, there should be no improvement in time taken to minimize the objective. Figure 2 suggests this is correct: the plots for $m = 64$ and $m = 128$ are essentially indistinguishable.

Support Vector Machines: For this experiment, we investigate solving an SVM problem (3) using the Reuters RCV1 dataset (Lewis et al., 2004), which consists of 800,000 training examples for binary classification tasks involving prediction of the topic of a news document from the words it contains. We compare Algorithm 1 to the state-of-the-art SVM solver Pegasos (Shalev-Shwartz et al., 2007). In the left plot of Fig. 3 we plot the optimality gap $f(x_t) + \varphi(x_t) - f(x^*) - \varphi(x^*)$ as a function of computation time for Pegasos and our perturbation method with 1, 2, 3, and 4 worker threads (denoted Acc (i)). In the right plot, we show the time in seconds required to achieve an $\epsilon = .004$ optimality gap for Algorithm 1 as a function of the number of threads computing stochastic subgradient estimates. The blue (top) line corresponds to each worker thread using a batch of size $m = 10$ to estimate a stochastic gradient, which is further averaged, while the red (lower) line corresponds to each worker using batches of size 20. We see improvement of approximately $1/n$ for n workers, as we expect.

Structured Prediction: Our final experiment is to learn feature weights for a probabilistic parsing task using a hypergraph parser. Hypergraph parsers (Klein & Manning, 2002) convert parsing tasks—which require assigning the productions in a probabilistic context-free grammar (PCFG) to a sentence—to finding maximum-weight paths in hypergraphs. To learn the weights on the features, we minimize a loss of the form (4), where the datum ξ is a sentence and the label ν is a parse tree, which corresponds to a weighted path between a sentence node and an initial sentential production node in the hypergraph associated with the PCFG. We only sketch our setup here. The important conditions we note are that the feature function ϕ decomposes across edges in the hypergraph, and we use standard lexical features (Taskar et al., 2004). If we let v be a $\{0, 1\}$ matrix with entries for each edge in a hypergraph, where $v_{j,k} = 1$ means edge (j, k) is selected, then labels ν are paths in the hypergraph and our loss is the Hamming loss: $\ell(v, \nu) = \sum_{j,k} 1_{(v_{j,k} \neq \nu_{j,k})}$. This loss decomposes across edges of the hypergraph, meaning the objective

$$\max_{v \in \mathcal{V}} \left[\ell(v, \nu) + \langle x, \phi(\xi, v)\rangle - \langle x, \phi(\xi, \nu)\rangle\right]_+ =$$
$$\max_{v \in \mathcal{V}} \sum_{j,k} \left(1_{(v_{j,k} \neq \nu_{j,k})} + \langle x, \phi_{j,k}(\xi, v_{j,k}) - \phi_{j,k}(\xi, \nu_{j,k})\rangle\right)$$

is computable in time cubic in the length of the sentence ξ (Klein & Manning, 2002). (Here we have used $\phi_{j,k}$ to indicate the features associated with the edge (j, k) in the hypergraph). The hypergraph representation we use is quite large: each word in a sentence generates some 200^2 different possible productions in the corresponding context free grammar, each of which requires thousands of features, yielding billions of weights; moreover, each hypergraph requires approximately 200kB of memory to store.

In Figure 4, we plot the results of 10 experiments for minimizing the structured prediction loss (4) on the Wall Street Journal portion of the Penn Treebank (Marcus et al., 1994). We use ℓ_2-regularization $\varphi(x) = (\lambda/2) \|x\|_2^2$ with multiplier $\lambda = .25$. We plot

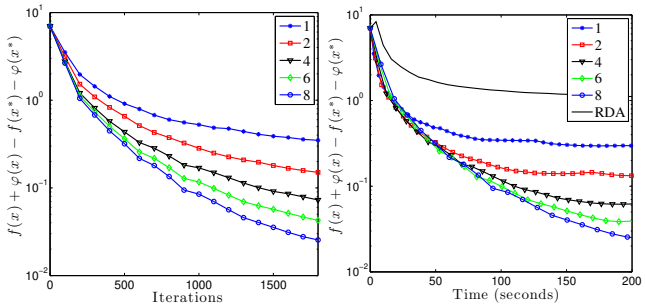

Figure 4. Optimality gaps for hypergraph-based parsing task (structured prediction). The legend for each plot gives the number of threads used to compute stochastic subgradients. Left: optimality gap versus number of iterations. Right: optimality gap versus computation time.

Algorithm 1's optimality gap as a function of the number of iterations (left) and the amount of time (right) required by the method. The right plot in Figure 4 also shows the performance of the state-of-the-art regularized dual averaging (RDA) algorithm (Xiao, 2010).

The left plot evidences a striking benefit in the number of iterations performed by the method as the number of threads increases. As the right plot shows, it is not trivial to translate this into improvements in actual running time. This is a consequence of the memory overhead engendered by multiple threads accessing the same memory as well as synchronization among the threads. Nonetheless, as the amount of time increases, the benefit of using multiple threads—and thus reducing the variance of the stochastic gradient estimate via the averaged gradient (12)—is clear. We also see the perhaps surprising result that even when using a single thread, the perturbation-based accelerated method yields improved performance over prior algorithms.

References

Abbeel, P. *Apprenticeship Learning and Reinforcement Learning with Application to Robotic Control*. PhD thesis, Stanford University, 2008.

Agarwal, Alekh, Bartlett, Peter L., Ravikumar, Pradeep, and Wainwright, Martin J. Information-theoretic lower bounds on the oracle complexity of convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012.

Ahuja, R. and Orlin, J. Inverse optimization. *Operations Research*, 49(5):771–783, 2001.

Bertsekas, Dmitri P. Stochastic optimization problems with nondifferentiable cost functionals. *Journal of Optimization Theory and Applications*, 12(2):218–231, 1973.

Cortes, C. and Vapnik, V. Support-vector networks. *Machine Learning*, 20(3):273–297, September 1995.

Dekel, O., Gilad-Bachrach, R., Shamir, O., and Xiao, L. Optimal distributed online prediction. In *Proceedings of the 28th International Conference on Machine Learning*, 2011.

Duchi, J. C. and Singer, Y. Efficient online and batch learning using forward-backward splitting. *Journal of Machine Learning Research*, 10:2873–2898, 2009.

Ghadimi, S. and Lan, G. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization. 2010.

Hazan, Elad and Kale, Satyen. An optimal algorithm for stochastic strongly convex optimization. In *Proceedings of the Twenty Fourth Annual Conference on Computational Learning Theory*, 2011. URL http://arxiv.org/abs/1006.2425.

Juditsky, Anatoli, Nemirovski, Arkadi, and Tauvel, Claire. Solving variational inequalities with the stochastic mirror-prox algorithm. *arXiv:0809.0815 [math.OC]*, 2008. URL http://arxiv.org/abs/0809.0815.

Klein, D. and Manning, C. Parsing and hypergraphs. In *New Developments in Parsing Technology*. Kluwer Academic, 2002.

Lan, Guanghui. An optimal method for stochastic composite optimization. *Mathematical Programming, Series A*, 133(1–2):365–397, 2012.

Lewis, D., Yang, Y., Rose, T., and Li, F. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19:313–330, 1994.

Nemirovski, A. and Yudin, D. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Shalev-Shwartz, S., Singer, Y., and Srebro, N. Pegasos: Primal estimated sub-gradient solver for SVM. In *Proceedings of the 24th International Conference on Machine Learning*, 2007.

Taskar, B. Klein, D. Collins, M. Koller, D. and Manning, C. Max-margin parsing. In *Empirical Methods in Natural Language Processing*, 2004.

Taskar, Ben. *Learning Structured Prediction Models: A Large Margin Approach*. PhD thesis, Stanford University, 2005.

Tseng, P. On accelerated proximal gradient methods for convex-concave optimization. 2008. URL http://www.mit.edu/~dimitrib/PTseng/papers/apgm.pdf.

Xiao, Lin. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11:2543–2596, 2010.

Xing, E., Ng, A.Y., Jordan, M., and Russell, S. Distance metric learning, with application to clustering with side-information. In *Advances in Neural Information Processing Systems 16*, 2003.

Yousefian, F., Nedic, A., and Shanbhag, U. Convex nondifferentiable stochastic optimization: a local randomized smoothing technique. In *Proceedings of IEEE American*

Control Conference, pp. 4875–4880, 2010.

Appendix: Proofs

A. Proofs for Non-Strongly Convex Optimization

In this section, we provide the proofs of Theorems 1 and 2, as well as Corollaries 1 through 4. We begin with the proofs of the corollaries, after which we give the full proofs of the theorems. In both cases, we defer some of the more technical lemmas to appendices.

The general technique for the proof of each corollary is as follows. First, we recognize that the randomly smoothed function $f_\mu(x) = \mathbb{E}f(x+Z)$ for $Z \sim \mu$ has Lipschitz continuous gradients and is uniformly close to the original non-smooth function f. This allows us to apply Theorem 1. The second step is to realize that with the sampling procedure (7), the variance $\mathbb{E}\left\|e_t\right\|_*^2$ decreases at a rate of approximately $1/m$, the number of gradient samples. Choosing the stepsizes appropriately in the theorems then completes the proofs. Proofs of these corollaries require relatively tight control of the smoothness properties of the smoothing convolution (5), so we refer frequently to lemmas stated in Appendix C.

A.1. Proof of Corollaries 1 and 2

We begin by proving Corollary 1. Recall the averaged quantity $g_t = \frac{1}{m}\sum_{i=1}^m g_{i,t}$, and that $g_{i,t} \in \partial F(y_t + Z_i; \xi_i)$, where the random variables Z_i are distributed uniformly on the ball $B_2(0,u)$.

From Lemma 8 in Appendix C, the variance of g_t as an estimate of $\nabla f_\mu(y_t)$ satisfies

$$\sigma^2 := \mathbb{E}\left\|e_t\right\|_2^2 = \mathbb{E}\left\|g_t - \nabla f_\mu(y_t)\right\|_2^2 \le \frac{L_0^2}{m}. \quad (15)$$

Further, for Z distributed uniformly on $B_2(0,u)$, we have the bound

$$f(x) \le \mathbb{E}[f(x+Z)] \le f(x) + L_0 u,$$

and moreover, the function f_μ has $L_0\sqrt{d}/u$-Lipschitz continuous gradient. Thus, applying Lemma 8 and Theorem 1 with the setting $L_t = L_0\sqrt{d}/u\theta_t$, we obtain

$$\begin{aligned}&\mathbb{E}[f(x_T) + \varphi(x_T)] - [f(x^*) + \varphi(x^*)] \\ &\le \frac{6L_0R^2\sqrt{d}}{Tu} + \frac{2\eta_T R^2}{T} + \frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{\eta_t}\cdot\frac{L_0^2}{m} + \frac{4L_0u}{T},\end{aligned}$$

where we have used the bound (15).

Now recall that $\eta_t = L_0\sqrt{t+1}/R\sqrt{m}$ by construction. Coupled with the inequality

$$\sum_{t=1}^T \frac{1}{\sqrt{t}} \le 1 + \int_1^T \frac{1}{\sqrt{t}}dt = 1 + 2(\sqrt{T}-1) \le 2\sqrt{T}, \quad (16)$$

we use that $2\sqrt{T+1}/T + 2/\sqrt{T} \le 5/\sqrt{T}$ to obtain

$$\begin{aligned}&\mathbb{E}[f(x_T) + \varphi(x_T)] - [f(x^*) + \varphi(x^*)] \\ &\le \frac{6L_0R^2\sqrt{d}}{Tu} + \frac{5L_0R}{\sqrt{Tm}} + \frac{4L_0u}{T}.\end{aligned}$$

Plugging the specified setting of $u = Rd^{1/4}$ completes the proof.

The proof of Corollary 2 is essentially identical, differing only in the setting of $u = Rd^{-1/4}$ and the application of Lemma 9 instead of Lemma 8 in Appendix C.

A.2. Proof of Corollary 3

Under the stated conditions of the corollary, Lemma 6 implies that when μ is uniform on $B_\infty(0,u)$, then the function $f_\mu(x) := \mathbb{E}_\mu f(x+Z)$ has a L_0/u-Lipschitz continuous gradient with respect to the ℓ_1-norm, and moreover it satisfies the upper bound $f_\mu(x) \le f(x) + \frac{L_0 du}{2}$.

Now, fix $x \in \mathcal{X}$ and let $g_i \in \partial F(x+Z_i;\xi_i)$, with $g = \frac{1}{m}\sum_{i=1}^m g_i$. We claim that the error satisfies

$$\mathbb{E}\left[\left\|g - \nabla f_\mu(x)\right\|_\infty^2\right] \le C\,\frac{L_0^2\log d}{m} \quad (17)$$

for some universal constant C. Indeed Lemma 6 shows that $\mathbb{E}[g] = \nabla f_\mu(x)$; moreover, component j of the random vector g_i is an unbiased estimator of the jth component of $\nabla f_\mu(x)$. Since $\left\|g_i\right\|_\infty \le L_0$ and $\left\|\nabla f_\mu(x)\right\|_\infty \le L_0$, the vector $g_i - \nabla f_\mu(x)$ is a d-dimensional random vector whose components are sub-Gaussian with sub-Gaussian parameter $4L_0^2$. Conditional on x, the g_i are independent, so $g - \nabla f_\mu(x)$ has sub-Gaussian components with parameter at most $4L_0^2/m$. Applying standard sub-Gaussian tail bounds to the ℓ_∞-norm bounded vectors $g_i - \nabla f_\mu(x)$ (we omit the proof) yields the claim (17).

Now, as in the proof of Corollary 1, we can apply Theorem 1. Recall that $\frac{1}{2(p-1)}\left\|x\right\|_p^2$ is strongly convex over $\mathbb{R}^d$ with respect to the ℓ_p-norm for any

$p \in (1, 2]$ (e.g. Xiao, 2010). Thus, with the choice $\psi(x) = \frac{1}{2(p-1)} \|x\|_p^2$ for $p = 1 + 1/\log d$, it is clear that the squared radius R^2 of the set $\mathcal{X}$ is order $\|x^*\|_p^2 \log d \le \|x^*\|_1^2 \log d$. Essentially, all that remains is to relate the Lipschitz constant L_0 with respect to the ℓ_1 norm to that for the ℓ_p norm. Let q be conjugate to p, that is, $1/q + 1/p = 1$. Under the assumptions of the theorem, we have $q = 1 + \log d$. For any $g \in \mathbb{R}^d$, we have $\|g\|_q \le d^{1/q} \|g\|_\infty$. Of course, $d^{1/(\log d+1)} \le d^{1/(\log d)} = \exp(1)$, so $\|g\|_q \le e \|g\|_\infty$.

Having shown that the Lipschitz constant L for the ℓ_p norm satisfies $L \le L_0 \exp(1)$, where L_0 is the Lipschitz constant with respect to the ℓ_1 norm, we simply apply Theorem 1 and the variance bound (17) to get the result. Specifically, Theorem 1 implies

$$\mathbb{E}[f(x_T) + \varphi(x_T)] - [f(x^*) + \varphi(x^*)] \le \frac{6L_0R^2}{Tu} + \frac{2\eta_T R^2}{T} + \frac{C}{T}\sum_{t=0}^{T-1} \frac{1}{\eta_t} \cdot \frac{L_0^2 \log d}{m} + \frac{4L_0 du}{2T}.$$

Plugging in the values for u, η_t, and $R \le \|x^*\|_1 \sqrt{\log d}$ and using bound (16) completes the proof.

A.3. Proof of Corollary 4

The proof of this corollary requires an auxiliary result showing that Assumption B holds under the stated conditions. The following result (whose proof we omit) can be shown using a Doob martingale construction with norm-bounded random vectors. In stating it, we recall the definition of the sigma field $\mathcal{F}_t$ from Assumption B.

Lemma 1. *Using the notation of Theorem 2, suppose that $F(\cdot; \xi)$ is L_0-Lipschitz continuous with respect to the norm $\|\cdot\|$ over $\mathcal{X} + \operatorname{supp} \mu$ for P-a.e. ξ. Then*

$$\mathbb{E}\left[\exp\left(\frac{\|e_t\|_*^2}{\sigma^2}\right) \mid \mathcal{F}_{t-1}\right] \le \exp(1),$$
$$\text{where } \sigma^2 := 2\max\left\{\mathbb{E}[\|e_t\|_*^2 \mid \mathcal{F}_{t-1}], \frac{16L_0^2}{m}\right\}.$$

Using this lemma, we now prove Corollary 4. When μ is the uniform distribution on $B_2(0, u)$, Lemma 8 from Appendix C implies that ∇f_μ is Lipschitz with constant $L_1 = L_0\sqrt{d}/u$. As discussed previously, Lemma 1 ensures that the error e_t satisfies Assumption B. Noting the inequality

$$\max\{\log(1/\delta), \sqrt{\log T \log(1/\delta)}\} \le \max\{\log(1/\delta), \log T\}$$

and combining the bound in Theorem 2 with Lemma 1, we see that with probability at least $1 - 2\delta$

$$f(x_T) + \varphi(x_T) - f(x^*) - \varphi(x^*) \le \frac{6L_0R^2\sqrt{d}}{Tu} + \frac{4L_0u}{T} + \frac{4R^2\eta}{\sqrt{T+1}} + \frac{2L_0^2}{m\sqrt{T}\eta} + C\frac{L_0^2 \max\left\{\log\frac{1}{\delta}, \log(T+1)\right\}}{(T+1)m\eta} + \frac{L_0R\sqrt{\log\frac{1}{\delta}}}{\sqrt{Tm}}$$

for a universal constant C. Plugging in $\eta = L_0/R\sqrt{m}$ and $u = Rd^{1/4}$ gives the desired result.

A.4. Proof of Theorem 1

This proof is more involved than that of the above corollaries. In We build on techniques used in the work of Tseng (2008), Lan (2012), and Xiao (2010). The changing smoothness of the stochastic objective—which comes from changing the shape parameter of the sampling distribution Z in the averaging step (7)—adds some challenge. Essentially, the idea of the proof is to define $f_t(x) := \mathbb{E}_\mu f(x + u_t Z)$, where u_t is the nonincreasing sequence of shape parameters in the averaging scheme (7). We show that $f_t(x) \le f_{t-1}(x)$ for all t, which is intuitive because the variance of the sampling scheme is decreasing, while Jensen's inequality tells us that $f(x) \le f_t(x)$. Then we apply the (stochastic) accelerated gradient method (Tseng, 2008; Xiao, 2010) to the sequence of functions f_t decreasing to f, and by allowing u_t to decrease appropriately we achieve our result. In the proof, we frequently use L_t as shorthand for the quantity L_1/u_t. We also simply assume that $\psi(x) = D_\psi(x, x_0)$ for notational convenience.

We begin by stating two technical lemmas:

Lemma 2. *Let f_t be a sequence of functions such that f_t has L_t-Lipschitz continuous gradients with respect to the norm $\|\cdot\|$ and assume that $f_t(x) \le f_{t-1}(x)$ for any $x \in \mathcal{X}$. Let the sequence $\{x_t, y_t, z_t\}$ be generated according to the updates (8a)–(8c), and define the error term $e_t = \nabla f_t(y_t) - g_t$. Then for any $x^* \in \mathcal{X}$,*

$$\frac{1}{\theta_t^2}[f_t(x_{t+1}) + \varphi(x_{t+1})] \le \sum_{\tau=0}^{t} \frac{1}{\theta_\tau}[f_\tau(x^*) + \varphi(x^*)] + \left(L_{t+1} + \frac{\eta_{t+1}}{\theta_{t+1}}\right)\psi(x^*) + \sum_{\tau=0}^{t} \frac{1}{2\theta_\tau\eta_\tau}\|e_t\|_*^2 + \sum_{\tau=0}^{t} \frac{1}{\theta_\tau}\langle e_\tau, z_\tau - x^*\rangle.$$

See Appendix A.6 for the proof of this claim.

Lemma 3. *Let the sequence θ_t satisfy $\frac{1-\theta_t}{\theta_t^2} = \frac{1}{\theta_{t-1}^2}$ and $\theta_0 = 1$. Then $\theta_t \le \frac{2}{t+2}$, and $\sum_{\tau=0}^{t} \frac{1}{\theta_\tau} = \frac{1}{\theta_t^2}$.*

The second statement was proved by Tseng (2008); the first follows by a straightforward induction.

We now proceed with the proof of the theorem. Defining $f_t(x) := \mathbb{E}[f(x + u_t Z)]$, let us first verify that $f_t(x) \le f_{t-1}(x)$ for any $x \in \mathcal{X}$ and t so that Lemma 2 can be applied. Since $u_t \le u_{t-1}$, we may define a random variable U with support on $\{0, 1\}$ such that $\mathbb{P}(U = 1) = \frac{u_t}{u_{t-1}} \in [0, 1]$. Then

$$\begin{aligned} f_t(x) &= \mathbb{E}[f(x + u_t Z)] = \mathbb{E}\big[f\big(x + u_{t-1} Z \mathbb{E}[U]\big)\big] \\ &\le \mathbb{P}[U = 1]\, \mathbb{E}[f(x + u_{t-1} Z)] + \mathbb{P}[U = 0]\, f(x), \end{aligned}$$

where the inequality follows from Jensen's inequality. By a second application of Jensen's inequality, we have $f(x) = f(x + u_{t-1}\mathbb{E}Z) \le \mathbb{E}f(x + u_{t-1}Z) = f_{t-1}(x)$. Combined with the previous inequality, we conclude that $f_t(x) \le \mathbb{E}[f(x + u_{t-1}Z)] = f_{t-1}(x)$ as claimed. Consequently, we have verified that the function f_t satisfies the assumptions of Lemma 2 where ∇f_t has Lipschitz parameter $L_t = L_1/u_t$ and error term $e_t = \nabla f_t(y_t) - g_t$. We apply the lemma momentarily.

Using Assumption A that $f(x) \ge \mathbb{E}[f(x + u_t Z)] - L_0 u_t = f_t(x) - L_0 u_t$ for all $x \in \mathcal{X}$, Lemma 3 implies

$$\begin{aligned} &\frac{1}{\theta_{T-1}^2}[f(x_T) + \varphi(x_T)] - \frac{1}{\theta_{T-1}^2}[f(x^*) + \varphi(x^*)] \\ &\quad = \frac{1}{\theta_{T-1}^2}[f(x_T) + \varphi(x_T)] - \sum_{t=0}^{T-1} \frac{1}{\theta_t}[f(x^*) + \varphi(x^*)] \\ &\quad \le \frac{1}{\theta_{T-1}^2}[f_{T-1}(x_T) + \varphi(x_T)] \qquad (18) \\ &\qquad - \sum_{t=0}^{T-1} \frac{1}{\theta_t}[f_t(x^*) + \varphi(x^*)] + \sum_{t=0}^{T-1} \frac{L_0 u_t}{\theta_t}, \end{aligned}$$

which by the definition of u_t is in turn bounded by

$$\frac{1}{\theta_{T-1}^2}[f_{T-1}(x_T)+\varphi(x_T)]-\sum_{t=0}^{T-1} \frac{1}{\theta_t}[f_t(x^*)+\varphi(x^*)]+TL_0u. \qquad (19)$$

Now we apply Lemma 2 to the bound (19), which gives

$$\begin{aligned} &\frac{1}{\theta_{T-1}^2}\,[f(x_T) + \varphi(x_T) - f(x^*) - \varphi(x^*)] \\ &\quad \le L_T \psi(x^*) + \frac{\eta_T}{\theta_T}\psi(x^*) \sum_{t=0}^{T-1} \frac{1}{2\theta_t \eta_t} \|e_t\|_*^2 \\ &\qquad + \sum_{t=0}^{T-1} \frac{1}{\theta_t} \langle e_t, z_t - x^* \rangle + TL_0 u. \qquad (20) \end{aligned}$$

The non-probabilistic bound (20) is the key to the remainder of this proof, as well as the starting point for the proof of Theorem 2 in the next section. What remains is to take expectations in the bound (20).

Recall the filtration of σ-fields $\mathcal{F}_t$ so that $x_t, y_t, z_t \in \mathcal{F}_{t-1}$, that is, $\mathcal{F}_t$ contains the randomness in the stochastic oracle to time t. Since g_t is an unbiased estimator of $\nabla f_t(y_t)$ by construction, we have $\mathbb{E}[g_t \mid \mathcal{F}_{t-1}] = \nabla f_t(y_t)$ and

$$\begin{aligned} \mathbb{E}[\langle e_t, z_t - x^* \rangle] &= \mathbb{E}\big[\mathbb{E}[\langle e_t, z_t - x^* \rangle \mid \mathcal{F}_{t-1}]\big] \\ &= \mathbb{E}\big[\, \langle \mathbb{E}[e_t \mid \mathcal{F}_{t-1}], z_t - x^* \rangle \,\big] \;=\; 0, \end{aligned}$$

where we have used the fact that z_t is measurable with respect to $\mathcal{F}_{t-1}$. Now, recall from Lemma 3 that $\theta_t \le \frac{2}{2+t}$ and that $(1 - \theta_t)/\theta_t^2 = 1/\theta_{t-1}^2$. Thus

$$\frac{\theta_{t-1}^2}{\theta_t^2} = \frac{1}{1 - \theta_t} \le \frac{1}{1 - \frac{2}{2+t}} = \frac{2+t}{t} \le \frac{3}{2} \qquad \text{for } t \ge 4.$$

Furthermore, we have $\theta_{t+1} \le \theta_t$, so by multiplying both sides of our bound (20) by θ_{T-1}^2 and taking expectations over the random vectors g_t,

$$\begin{aligned} &\mathbb{E}[f(x_T) + \varphi(x_T)] - [f(x^*) + \varphi(x^*)] \\ &\le \theta_{T-1}^2 L_T \psi(x^*) + \theta_{T-1}\eta_T \psi(x^*) + \theta_{T-1}^2 T L_0 u \\ &\quad + \theta_{T-1} \sum_{t=0}^{T-1} \frac{1}{2\eta_t} \mathbb{E}\,\|e_t\|_*^2 + \theta_{T-1} \sum_{t=0}^{T-1} \mathbb{E}[\langle e_t, z_t - x^* \rangle] \\ &\le \frac{6L_1 \psi(x^*)}{Tu} + \frac{2\eta_T \psi(x^*)}{T} + \frac{1}{T}\sum_{t=0}^{T-1} \frac{1}{\eta_t} \mathbb{E}\,\|e_t\|_*^2 + \frac{4L_0 u}{T}, \end{aligned}$$

where we used the fact that $L_T = L_1/u_T = L_1/\theta_T u$. This completes the proof of Theorem 1.

A.5. Proof of Theorem 2

An examination of the proof of Theorem 1 shows that to control the probability of deviation from the expected convergence rate, we need to control two terms: the squared error sequence $\sum_{t=0}^{T-1} \frac{1}{2\eta_t} \|e_t\|_*^2$ and the sequence $\sum_{t=0}^{T-1} \frac{1}{\theta_t} \langle e_t, z_t - x^* \rangle$. The next two lemmas handle these terms.

Lemma 4. *Let $\mathcal{X}$ be compact with $\|x - x^*\| \le R$ for all $x \in \mathcal{X}$. Under Assumption B, we have*

$$\mathbb{P}\left[\theta_{T-1}^2 \sum_{t=0}^{T-1} \frac{1}{\theta_t} \langle e_t, z_t - x^* \rangle \ge \epsilon\right] \le \exp\left(-\frac{T\epsilon^2}{R^2\sigma^2}\right). \qquad (21)$$

Consequently, with probability at least $1 - \delta$,

$$\theta_{T-1}^2 \sum_{t=0}^{T-1} \frac{1}{\theta_t} \langle e_t, z_t - x^* \rangle \le R\sigma\sqrt{\frac{\log \frac{1}{\delta}}{T}}. \qquad (22)$$

Lemma 5. *In the notation of Theorem 2 and under*

Assumption B, we have

$$\log \mathbb{P}\left[\sum_{t=0}^{T-1} \frac{1}{2\eta_t}\|e_t\|_*^2 \geq \sum_{t=0}^{T-1} \frac{1}{2\eta_t}\mathbb{E}[\|e_t\|_*^2] + \epsilon\right] \leq \max\left\{-\frac{\epsilon^2}{32e\sigma^4 \sum_{t=0}^{T-1} \frac{1}{\eta_t^2}}, -\frac{\eta_0}{4\sigma^2}\epsilon\right\}. \quad (23)$$

Consequently, with probability at least $1-\delta$,

$$\sum_{t=0}^{T-1} \frac{1}{2\eta_t}\|e_t\|_*^2 \leq \sum_{t=0}^{T-1} \frac{1}{2\eta_t}\mathbb{E}[\|e_t\|_*^2] + \frac{4\sigma^2}{\eta}\max\left\{\log\frac{1}{\delta}, \sqrt{2e\log(T+1)\log\frac{1}{\delta}}\right\}. \quad (24)$$

The proofs of these probabilistic lemmas are technical and build off of concentration results for sums of random variables similar to those used by Nemirovski et al. (2009); we omit them as they are somewhat lengthy and require several auxilary statements on concentration of sub-Gaussian and sub-exponential random variables. (We provide proofs in the journal version of this paper.)

Equipped with these lemmas, we now prove Theorem 2. Let us recall the deterministic bound (20) from the proof of Theorem 1:

$$\frac{1}{\theta_{T-1}^2}[f(x_T)+\varphi(x_T)-f(x^*)-\varphi(x^*)] \leq L_T\psi(x^*) + \frac{\eta_T}{\theta_T}\psi(x^*) + \sum_{t=0}^{T-1}\frac{1}{2\theta_t\eta_t}\|e_t\|_*^2 + \sum_{t=0}^{T-1}\frac{1}{\theta_t}\langle e_t, z_t - x^*\rangle + TL_0u.$$

Noting that $\theta_{T-1} \leq \theta_t$ for $t \in \{0,\dots,T-1\}$, Lemma 5 implies that with probability at least $1-\delta$

$$\theta_{T-1}\sum_{t=0}^{T-1}\frac{1}{2\theta_t\eta_t}\|e_t\|_*^2 \leq \sum_{t=0}^{T-1}\frac{1}{2\eta_t}\mathbb{E}[\|e_t\|_*^2] + \frac{4\sigma^2}{\eta}\max\left\{\log(1/\delta), \sqrt{2e\log(T+1)\log(1/\delta)}\right\}.$$

Applying Lemma 4, we see that with probability at least $1-\delta$

$$\theta_{T-1}^2\sum_{t=0}^{T-1}\frac{1}{\theta_t}\langle e_t, z_t - x^*\rangle \leq \frac{R\sigma\sqrt{\log\frac{1}{\delta}}}{\sqrt{T}}.$$

The terms remaining to control are deterministic, and were bounded previously in the proof of Theorem 1; in particular, we have $\theta_{T-1}^2 L_T \leq \frac{6L_1}{Tu}$, $\frac{\theta_{T-1}^2\eta_T}{\theta_T} \leq \frac{4\eta_T}{T+1}$, and $\theta_{T-1}^2 TL_0u \leq \frac{4L_0u}{T+1}$. Combining the above bounds completes the proof.

A.6. Proof of Lemma 2

Define the linearized version of the cumulative objective

$$\ell_t(z) := \sum_{\tau=0}^{t}\frac{1}{\theta_\tau}[f_\tau(y_\tau) + \langle g_\tau, z - y_\tau\rangle + \varphi(z)], \quad (25)$$

and let $\ell_{-1}(z)$ denote the indicator function of $\mathcal{X}$. For conciseness, we adopt the shorthand notation

$$\alpha_t^{-1} = L_t + \eta_t/\theta_t \qquad \text{and} \qquad \phi_t(x) = f_t(x) + \varphi(x).$$

By the smoothness of f_t, we have

$$\underbrace{f_t(x_{t+1}) + \varphi(x_{t+1})}_{\phi_t(x_{t+1})} \leq f_t(y_t) + \langle \nabla f_t(y_t), x_{t+1} - y_t\rangle + \frac{L_t}{2}\|x_{t+1} - y_t\|^2 + \varphi(x_{t+1}).$$

From the definition (8a)–(8c) of the triple (x_t, y_t, z_t), we obtain

$$\phi_t(x_{t+1}) \leq f_t(y_t) + \langle \nabla f_t(y_t), \theta_t z_{t+1} + (1-\theta_t)x_t\rangle + \frac{L_t}{2}\|\theta_t z_{t+1} - \theta_t z_t\|^2 + \varphi(\theta_t z_{t+1} + (1-\theta_t)x_t).$$

Finally, by convexity of the regularizer φ, we conclude

$$\phi_t(x_{t+1}) \leq \theta_t\Big[f_t(y_t) + \langle \nabla f_t(y_t), z_{t+1} - y_t\rangle + \varphi(z_{t+1}) + \frac{L_t\theta_t}{2}\|z_{t+1} - z_t\|^2\Big] + (1-\theta_t)[f_t(y_t) + \langle \nabla f_t(y_t), x_t - y_t\rangle + \varphi(x_t)]. \quad (26)$$

By the strong convexity of ψ, it is clear that we have the lower bound

$$D_\psi(x,y) \geq \frac{1}{2}\|x-y\|^2. \quad (27)$$

On the other hand, by the convexity of f_t, we have

$$f_t(y_t) + \langle \nabla f_t(y_t), x_t - y_t\rangle \leq f_t(x_t). \quad (28)$$

Substituting inequalities (27) and (28) into the upper bound (26) and simplifying yields

$$\phi_t(x_{t+1}) \leq \theta_t\big[f_t(y_t) + \langle \nabla f_t(y_t), z_{t+1} - y_t\rangle + \varphi(z_{t+1}) + L_t\theta_t D_\psi(z_{t+1}, z_t)\big] + (1-\theta_t)[f_t(x_t) + \varphi(x_t)].$$

We now re-write this upper bound in terms of the error $e_t = \nabla f_t(y_t) - g_t$. In particular,

$$\begin{aligned}\phi_t(x_{t+1}) &\leq \theta_t\big[f_t(y_t) + \langle g_t, z_{t+1} - y_t\rangle + \varphi(z_{t+1}) + L_t\theta_t D_\psi(z_{t+1}, z_t)\big] + (1-\theta_t)[f_t(x_t) + \varphi(x_t)] + \theta_t\langle e_t, z_{t+1} - y_t\rangle \\ &= \theta_t^2[\ell_t(z_{t+1}) - \ell_{t-1}(z_{t+1}) + L_t D_\psi(z_{t+1}, z_t)] + (1-\theta_t)[f_t(x_t) + \varphi(x_t)] + \theta_t\langle e_t, z_{t+1} - y_t\rangle \quad (29)\end{aligned}$$

Using the fact that z_t minimizes $\ell_{t-1}(x)+\frac{1}{\alpha_t}\psi(x)$, the first order conditions for optimality imply that for all $g \in \partial\ell_{t-1}(z_t)$, we have $\left\langle g+\frac{1}{\alpha_t}\nabla\psi(z_t), x-z_t\right\rangle \geq 0$. Thus, first-order convexity gives

$$\begin{aligned}\ell_{t-1}(x)-\ell_{t-1}(z_t) &\geq \langle g, x-z_t\rangle \geq -\frac{1}{\alpha_t}\langle\nabla\psi(z_t), x-z_t\rangle \\ &= \frac{1}{\alpha_t}\psi(z_t)-\frac{1}{\alpha}\psi(x)+\frac{1}{\alpha_t}D_\psi(x,z_t).\end{aligned}$$

Adding $\ell_t(z_{t+1})$ to both sides of the above and substituting $x=z_{t+1}$, we conclude

$$\begin{aligned}\ell_t(z_{t+1})-\ell_{t-1}(z_{t+1}) &\leq \ell_t(z_{t+1})-\ell_{t-1}(z_t) \\ &\quad -\frac{1}{\alpha_t}\psi(z_t)+\frac{1}{\alpha_t}\psi(z_{t+1})-\frac{1}{\alpha_t}D_\psi(z_{t+1},z_t).\end{aligned}$$

Combining this inequality with the bound (29) and using the definition $\alpha_t^{-1}=L_t+\eta_t/\theta_t$, we find

$$\begin{aligned}&f_t(x_{t+1})+\varphi(x_{t+1}) \\ &\leq \theta_t^2\Big[\ell_t(z_{t+1})-\ell_t(z_t)-\frac{1}{\alpha_t}\psi(z_t) \\ &\qquad +\frac{1}{\alpha_t}\psi(z_{t+1})-\frac{\eta_t}{\theta_t}D_\psi(z_{t+1},z_t)\Big] \\ &\quad +(1-\theta_t)[f_t(x_t)+\varphi(x_t)]+\theta_t\langle e_t, z_{t+1}-y_t\rangle \\ &\leq \theta_t^2\Big[\ell_t(z_{t+1})-\ell_t(z_t)-\frac{1}{\alpha_t}\psi(z_t) \\ &\qquad +\frac{1}{\alpha_{t+1}}\psi(z_{t+1})-\frac{\eta_t}{\theta_t}D_\psi(z_{t+1},z_t)\Big] \\ &\quad +(1-\theta_t)[f_t(x_t)+\varphi(x_t)]+\theta_t\langle e_t, z_{t+1}-y_t\rangle\end{aligned}$$

since α_t^{-1} is non-decreasing. We now divide both sides by θ_t^2 and unwrap the recursion. Recall that $(1-\theta_t)/\theta_t^2=1/\theta_{t-1}^2$ and $f_t\leq f_{t-1}$ by construction, so we obtain

$$\begin{aligned}&\frac{1}{\theta_t^2}[f_t(x_{t+1})+\varphi(x_{t+1})] \\ &\leq \frac{1-\theta_t}{\theta_t^2}[f_t(x_t)+\varphi(x_t)]+\ell_t(z_{t+1})-\ell_t(z_t)-\frac{1}{\alpha_t}\psi(z_t) \\ &\quad +\frac{1}{\alpha_{t+1}}\psi(z_{t+1})-\frac{\eta_t}{\theta_t}D_\psi(z_{t+1},z_t)+\frac{1}{\theta_t}\langle e_t, z_{t+1}-y_t\rangle \\ &\overset{(i)}{=} \frac{1}{\theta_{t-1}^2}[f_t(x_t)+\varphi(x_t)]+\ell_t(z_{t+1})-\ell_t(z_t)-\frac{1}{\alpha_t}\psi(z_t) \\ &\quad +\frac{1}{\alpha_{t+1}}\psi(z_{t+1})-\frac{\eta_t}{\theta_t}D_\psi(z_{t+1},z_t)+\frac{1}{\theta_t}\langle e_t, z_{t+1}-y_t\rangle \\ &\overset{(ii)}{\leq} \frac{1}{\theta_{t-1}^2}[f_{t-1}(x_t)+\varphi(x_t)]+\ell_t(z_{t+1})-\ell_t(z_t)-\frac{1}{\alpha_t}\psi(z_t) \\ &\quad +\frac{1}{\alpha_{t+1}}\psi(z_{t+1})-\frac{\eta_t}{\theta_t}D_\psi(z_{t+1},z_t)+\frac{1}{\theta_t}\langle e_t, z_{t+1}-y_t\rangle.\end{aligned}$$

The equality (i) follows since $(1-\theta_t)/\theta_t^2=1/\theta_{t-1}^2$, while the inequality (ii) is a consequence of the fact that $f_t\leq f_{t-1}$. By applying the three steps above successively to $[f_{t-1}(x_t)+\varphi(x_t)]/\theta_{t-1}^2$, then to $[f_{t-2}(x_{t-1})+\varphi(x_{t-1})]/\theta_{t-2}^2$, and so on until $t=0$, we find

$$\begin{aligned}&\frac{1}{\theta_t^2}[f_t(x_{t+1})+\varphi(x_{t+1})] \\ &\leq \frac{1-\theta_0}{\theta_0^2}[f_0(x_0)+\varphi(x_0)]+\ell_t(z_{t+1})+\frac{1}{\alpha_{t+1}}\psi(z_{t+1}) \\ &\qquad -\sum_{\tau=0}^{t}\frac{\eta_\tau}{\theta_\tau}D_\psi(z_{\tau+1},z_\tau)+\sum_{\tau=0}^{t}\frac{1}{\theta_\tau}\langle e_\tau, z_{\tau+1}-y_\tau\rangle \\ &\qquad -\ell_{-1}(z_0)-\frac{1}{\alpha_0}\psi(z_0).\end{aligned}$$

By construction, $\theta_0=1$, we have $\ell_{-1}(z_0)=0$, and z_{t+1} minimizes $\ell_t(x)+\frac{1}{\alpha_{t+1}}\psi(x)$ over $\mathcal{X}$. Thus, for any $x^*\in\mathcal{X}$, we have

$$\begin{aligned}\frac{1}{\theta_t^2}[f_t(x_{t+1})+\varphi(x_{t+1})] &\leq \ell_t(x^*)+\frac{1}{\alpha_{t+1}}\psi(x^*) \\ &\quad -\sum_{\tau=0}^{t}\frac{\eta_\tau}{\theta_\tau}D_\psi(z_{\tau+1},z_\tau)+\sum_{\tau=0}^{t}\frac{1}{\theta_\tau}\langle e_\tau, z_{\tau+1}-y_\tau\rangle.\end{aligned}$$

Recalling the definition (25) of ℓ_t and noting that $f_t(y_t)+\langle\nabla f_t(y_t), x-y_t\rangle\leq f_t(x)$ by convexity, we expand ℓ_t and have

$$\begin{aligned}&\frac{1}{\theta_t^2}[f_t(x_{t+1})+\varphi(x_{t+1})] \\ &\leq \sum_{\tau=0}^{t}\frac{1}{\theta_\tau}[f_\tau(y_\tau)+\langle g_\tau, x^*-y_\tau\rangle+\varphi(x^*)]+\frac{1}{\alpha_{t+1}}\psi(x^*) \\ &\qquad -\sum_{\tau=0}^{t}\frac{\eta_\tau}{\theta_\tau}D_\psi(z_{\tau+1},z_\tau)+\sum_{\tau=0}^{t}\frac{1}{\theta_\tau}\langle e_\tau, z_{\tau+1}-y_t\rangle \\ &= \sum_{\tau=0}^{t}\frac{1}{\theta_\tau}[f_\tau(y_\tau)+\langle\nabla f_\tau(y_\tau), x^*-y_\tau\rangle+\varphi(x^*)] \\ &\qquad +\frac{1}{\alpha_{t+1}}\psi(x^*) \\ &\qquad -\sum_{\tau=0}^{t}\frac{\eta_\tau}{\theta_\tau}D_\psi(z_{\tau+1},z_\tau)+\sum_{\tau=0}^{t}\frac{1}{\theta_\tau}\langle e_\tau, z_{\tau+1}-x^*\rangle \\ &\leq \sum_{\tau=0}^{t}\frac{1}{\theta_\tau}[f_\tau(x^*)+\varphi(x^*)]+\frac{1}{\alpha_{t+1}}\psi(x^*) \\ &\qquad -\sum_{\tau=0}^{t}\frac{\eta_\tau}{\theta_\tau}D_\psi(z_{\tau+1},z_\tau)+\sum_{\tau=0}^{t}\frac{1}{\theta_\tau}\langle e_\tau, z_{\tau+1}-x^*\rangle.\end{aligned} \tag{30}$$

Now we use the Fenchel-Young inequality applied to the conjugates $\frac{1}{2}\|\cdot\|^2$ and $\frac{1}{2}\|\cdot\|_*^2$, which gives

$$\begin{aligned}\langle e_t, z_{t+1}-x^*\rangle &= \langle e_t, z_t-x^*\rangle+\langle e_t, z_{t+1}-z_t\rangle \\ &\leq \langle e_t, z_t-x^*\rangle+\frac{1}{2\eta_t}\|e_t\|_*^2+\frac{\eta_t}{2}\|z_t-z_{t+1}\|^2.\end{aligned}$$

In particular,

$$-\frac{\eta_t}{\theta_t}D_\psi(z_{t+1},z_t)+\frac{1}{\theta_t}\langle e_t, z_{t+1}-x^*\rangle \le \frac{1}{2\eta_t\theta_t}\|e_t\|_*^2+\frac{1}{\theta_t}\langle e_t, z_t-x^*\rangle .$$

Using this inequality and rearranging (30) gives the statement of the lemma.

B. Proof of Theorem 3

In this section, we provide the promised proof of Theorem 1. Our proof is based on the following proposition, which shows the exponential decrease of the optimization error as a function of the number of epochs.

Proposition 1. *Let $x(k)$ denote the output of Algorithm 1 and let Assumptions A and C. In addition, let $\mathsf{M}\ge f(x(0))+\varphi(x(0))-f(x^*)-\varphi(x^*)$ denote an upper bound on the initial optimality gap. Then*

$$\mathbb{E}[f(x(k))+\varphi(x(k))]-[f(x^*)+\varphi(x^*)] \le 2^{-k}\mathsf{M}+5\cdot 2^{-k}\left[\frac{\sigma^2}{2\eta}+L_0u\right]. \tag{31}$$

Before proving Proposition 1, we give the proof of Theorem 3. To that end, we compute the number of iterations required to achieve a particular ϵ-accuracy for the optimization error (31). We first note that by choosing $k=\log_2\frac{\mathsf{M}}{\epsilon}$, we can replace the expected optimality gap (31) with

$$\frac{\epsilon}{\mathsf{M}}\mathsf{M}+5\cdot\frac{\epsilon}{\mathsf{M}}\left[\frac{\sigma^2}{2\eta}+L_0u\right]=\epsilon+5\cdot\frac{\epsilon}{\mathsf{M}}\left[\frac{\sigma^2}{2\eta}+L_0u\right].$$

What remains is to compute the number of iterations of the three-series updates (8a)–(8c). To that end, we compute the sum of $t(i)$ as chosen across all the epochs of Algorithm 1. Using that $\max\{a,b\}\le a+b$ for $a,b\ge 0$, we have

$$\begin{aligned}\sum_{i=1}^k t(i) &\le 4\sum_{i=1}^k\sqrt{\frac{L_1}{u\lambda}}(\sqrt{2})^i+12\sum_{i=1}^k\frac{2^i\eta}{\lambda}\\ &= 4(\sqrt{2})^k\sqrt{\frac{L_1}{u\lambda}}\sum_{i=1}^k(\sqrt{2})^{-i}+\frac{12\eta}{\lambda}2^k\sum_{i=1}^k 2^{i-k}\\ &\le \frac{4}{\sqrt{2}}\sqrt{\frac{L_1}{u\lambda}}\cdot\frac{1}{1-\sqrt{2}/2}(\sqrt{2})^k+\frac{24\eta}{\lambda}2^k.\end{aligned}$$

Plugging in the choice of $k=\log_2(\mathsf{M}/\epsilon)$, we conclude that

$$\sum_{i=1}^k t(i)\le\frac{4}{\sqrt{2}-1}\sqrt{\frac{L_1}{u\lambda}}\cdot\sqrt{\frac{\mathsf{M}}{\epsilon}}+\frac{24\eta}{\lambda}\cdot\frac{\mathsf{M}}{\epsilon},$$

which is the content of Theorem 1.

Proof of Proposition 1 We begin by recasting the convergence guarantee of Lemma 2 in the necessary epoch-based notation. Let $x_\tau(i)$ denote the value of x_τ in epoch i, and similarly for $z_\tau(i)$, $g_\tau(i)$, and so on. Let $f_i(x)=\mathbb{E}_\mu[f(x+u(i)Z)]$ denote the mollified function during epoch i, and define the error of the gradient estimate $g_\tau(i)$ at iteration τ of epoch i to be $e_\tau(i)=\nabla f_i(y_\tau(i))-g_\tau(i)$. Since $\psi(\cdot)$ is 1-strongly convex with respect to the norm $\|\cdot\|$, the output $x(i)$ of iteration i of Algorithm 1 satisfies

$$\begin{aligned}&f_i(x(i))+\varphi(x(i))-[f_i(x^*)+\varphi(x^*)]\\ &\le\theta_{t(i)-1}^2\left(\frac{L_1}{u(i)}+\frac{\eta(i)}{\theta_{t(i)}}\right)D_\psi(x^*,x(i-1))\\ &\quad+\theta_{t(i)-1}^2\sum_{\tau=0}^{t(i)-1}\frac{1}{2\theta_\tau\eta(i)}\|e_\tau(i)\|_*^2\\ &\quad+\theta_{t(i)-1}^2\sum_{\tau=0}^{t(i)-1}\frac{1}{\theta_\tau}\langle e_\tau(i),z_\tau(i)-x^*\rangle .\end{aligned}\tag{32}$$

Define the error factors

$$\begin{aligned}E(i):=&\sum_{\tau=0}^{t(i)-1}\frac{1}{2\theta_\tau\eta(i)}\|e_\tau(i)\|_*^2\\ &+\sum_{\tau=0}^{t(i)-1}\frac{1}{\theta_\tau}\langle e_\tau(i),z_\tau(i)-x^*\rangle ,\end{aligned}\tag{33}$$

and apply the smoothing assumption A to the single-epoch bound (32) to see that

$$\begin{aligned}&f(x(i))+\varphi(x(i))-[f(x^*)+\varphi(x^*)]\\ &\le\theta_{t(i)-1}^2\left(\frac{L_1}{u(i)}+\frac{\eta(i)}{\theta_{t(i)}}\right)D_\psi(x^*,x(i-1))\\ &\quad+\theta_{t(i)-1}^2E(i)+L_0u(i).\end{aligned}\tag{34}$$

Note that by our strong-convexity assumption C, for any $x\in\mathcal{X}$ we have

$$D_\psi(x^*,x)\le\frac{1}{\lambda}[f(x)+\varphi(x)-f(x^*)-\varphi(x^*)].$$

Define $\phi(x)=f(x)+\varphi(x)$ for notational convenience. Then we can replace the upper bound (34) with

$$\begin{aligned}&\frac{\theta_{t(i)-1}^2}{\lambda}\left(\frac{L_1}{u(i)}+\frac{\eta(i)}{\theta_{t(i)}}\right)[\phi(x(i-1))-\phi(x^*)]\\ &\quad+\theta_{t(i)-1}^2E(i)+L_0u(i).\end{aligned}$$

To simplify, we make the definitions of the multiplication factors $M(i)$ and $\pi_k(i)$ as

$$M(i):=\frac{\theta_{t(i)-1}^2}{\lambda}\left(\frac{L_1}{u(i)}+\frac{\eta(i)}{\theta_{t(i)}}\right)\text{ and }\pi_k(i):=\prod_{j=i+1}^k M(j).\tag{35}$$

Then recursively applying the bound (34), the definitions (33), (35), and that of the joint function ϕ imply

$$\begin{aligned}
&f(x(k)) + \varphi(x(k)) - f(x^*) - \varphi(x^*) \\
&\le M(k)[f(x(k-1)) + \varphi(x(k-1)) - f(x^*) - \varphi(x^*)] \\
&\quad + \theta_{t(k)-1}^2 E(k) + L_0 u(k) \\
&\le M(k)\Big(M(k-1)\left[\phi(x(k-2)) - \phi(x^*)\right] \\
&\quad + \theta_{t(k-1)-1}^2 E(k-1) + L_0 u(k-1)\Big) \\
&\quad + \theta_{t(k)-1}^2 E(k) + L_0 u(k) \\
&\le \left(\prod_{i=1}^k M(i)\right)[f(x(0)) + \varphi(x(0)) - f(x^*) - \varphi(x^*)] \\
&\quad + \sum_{i=1}^k \left(\prod_{j=i+1}^k M(j)\right)\left[\theta_{t(i)-1}^2 E(i) + L_0 u(i)\right] \\
&\le \left(\prod_{i=1}^k M(i)\right)\mathsf{M} + \sum_{i=1}^k \pi_k(i)\left[\theta_{t(i)-1}^2 E(i) + L_0 u(i)\right]
\end{aligned} \tag{36}$$

where in the last line (36) we recalled the definition $\mathsf{M} \ge f(x(0)) + \varphi(x(0)) - [f(x^*) + \varphi(x^*)]$. So if we can choose $\eta(i)$ and $t(i)$ so that $M(i) < \frac{1}{2}$, we will have bounds that decrease geometrically in the number of epochs k, so very few epochs are necessary.

To that end, choose the iteration count specified in Algorithm 1:

$$t(i) = \max\left\{4\sqrt{\frac{L_1}{u(i)\lambda}}, \frac{12\eta(i)}{\lambda}\right\}.$$

Noting that

$$\frac{\theta_{t-1}^2}{\theta_t} = \frac{\theta_t^2}{(1-\theta_t)\theta_t} = \frac{\theta_t}{1-\theta_t} \le \frac{\frac{2}{2+t}}{1-\frac{2}{2+t}} = \frac{2}{t},$$

this choice of $t(i)$ yields the bound

$$\begin{aligned}
M(i) &= \frac{\theta_{t(i)-1}^2}{\lambda}\left(\frac{L_1}{u(i)} + \frac{\eta(i)}{\theta_{t(i)}}\right) \\
&\le \frac{4L_1}{\lambda u(i) t(i)^2} + \frac{2\eta(i)}{\lambda t(i)} \le \frac{1}{4} + \frac{2}{12} = \frac{5}{12} < \frac{1}{2}
\end{aligned}$$

on the multipliers $M(i)$. So after k epochs, the bound (36) tells us that

$$\begin{aligned}
&f(x(k)) + \varphi(x(k)) - f(x^*) - \varphi(x^*) \\
&\le 2^{-k}\mathsf{M} + \sum_{i=1}^k \pi_k(i)\theta_{t(i)-1}^2 E(i) + L_0 \sum_{i=1}^k \pi_k(i) u(i).
\end{aligned}$$

Now we take expectations of the error terms $E(i)$. Recalling that $\mathbb{E}[e_\tau(i) \mid \mathcal{F}_{\tau-1}(i)] = 0$, since the draws of Z and ξ are independent in each iteration, we have

$$\mathbb{E}[E(i)] \le \sum_{\tau=0}^{t(i)-1} \frac{\sigma^2}{2\theta_\tau \eta(i)} = \frac{\sigma^2}{2^{i+1}\theta_{t(i)-1}^2 \eta}$$

by the definition of the recursion for the θ_t (recall Lemma 3). Consequently, we use the choices $\eta(i) = \eta \cdot 2^i$ and $u(i) = u \cdot 2^{-i}$ in Algorithm 1, and we have

$$\begin{aligned}
&\mathbb{E}[f(x(k)) + \varphi(x(k)) - f(x^*) - \varphi(x^*)] \\
&\le 2^{-k}\mathsf{M} + \frac{\sigma^2}{2\eta}\sum_{i=1}^k \pi_k(i) 2^{-i} + L_0 u \sum_{i=1}^k \pi_k(i) 2^{-i} \\
&\le 2^{-k}\mathsf{M} + \left[\frac{\sigma^2}{2\eta} + L_0 u\right]\sum_{i=1}^k \left(\frac{5}{12}\right)^{k-i} 2^{-i} \\
&= 2^{-k}\mathsf{M} + 2^{-k}\left[\frac{\sigma^2}{2\eta} + L_0 u\right]\sum_{i=1}^k \left(\frac{5}{6}\right)^{k-i} \\
&\le 2^{-k}\mathsf{M} + 5 \cdot 2^{-k}\left[\frac{\sigma^2}{2\eta} + L_0 u\right].
\end{aligned}$$

The bound above is evidently our desired result (31). □

C. Smoothing Properties

In this section, we discuss the analytic properties of the smoothed function f_μ from the convolution (5). We simply state the results, preferring to avoid lengthening the already lengthy theoretical treatment, and referring the reader to Yousefian et al. (2010) for one example proof, excepting the sharpness argument (the other proofs are similar). We assume throughout that functions are sufficiently integrable without bothering with measurability conditions (since $F(\cdot;\xi)$ is convex, this is no real loss of generality (Bertsekas, 1973)). By Fubini's theorem, we have

$$\begin{aligned}
f_\mu(x) &= \int_\Xi \int_{\mathbb{R}^d} F(y;\xi)\mu(x-y)dydP(\xi) \\
&= \int_\Xi F_\mu(x;\xi)dP(\xi).
\end{aligned}$$

Here $F_\mu(x;\xi) = (F(\cdot;\xi) * \mu)(x)$. We begin with the observation that since μ is a density with respect to Lebesgue measure, the function f_μ is in fact differentiable (Bertsekas, 1973). So we have already made our problem somewhat smoother, as it is now differentiable; for the remainder, we consider finer properties of the smoothing operation. In particular, we will show that under suitable conditions on μ, $F(\cdot;\xi)$, and P, the

function f_μ is uniformly close to f over $\mathcal{X}$ and ∇f_μ is Lipschitz continuous.

A remark on notation before proceeding: since f is convex, it is almost-everywhere differentiable, and we can abuse notation and take its gradient inside of integrals and expectations with respect to Lebesgue measure. Similarly, $F(\cdot;\xi)$ is almost everywhere differentiable with respect to Lebesgue measure, so we use the same abuse of notation for F and write $\nabla F(x+Z;\xi)$, which exists with probability 1.

Lemma 6. *Let μ be the uniform density on the ℓ_∞-ball of radius u. Assume that $\mathbb{E}[\|\partial F(x;\xi)\|_\infty^2] \le L_0^2$ for all $x \in \mathcal{X} + B_\infty(0,u)$ Then*

(i) $f(x) \le f_\mu(x) \le f(x) + \frac{L_0 d}{2} u$

(ii) f_μ *is L_0-Lipschitz with respect to the ℓ_1-norm over $\mathcal{X}$.*

(iii) f_μ *is continuously differentiable; moreover, its gradient is $\frac{L_0}{u}$-Lipschitz continuous with respect to the ℓ_1-norm.*

(iv) Let $Z \sim \mu$. Then $\mathbb{E}[\nabla F(x+Z;\xi)] = \nabla f_\mu(x)$ and $\mathbb{E}[\|\nabla f_\mu(x) - \nabla F(x+Z;\xi)\|_\infty^2] \le 4L_0^2$.

There exist functions for which each of the estimates (i)–(iii) are tight simultaneously, and (iv) is tight at least to a factor of $1/4$.

Remark: Note that the hypothesis of this lemma is satisfied if for any fixed $\xi \in \Xi$, the function $F(\cdot;\xi)$ is L_0-Lipschitz with respect to the ℓ_1-norm.

The following lemma provides bounds for uniform smoothing of functions Lipschitz with respect to the ℓ_2-norm while sampling from an ℓ_∞-ball.

Lemma 7. *Let μ be the uniform density on $B_\infty(0,u)$ and assume that $\mathbb{E}[\|\partial F(x;\xi)\|_2^2] \le L_0^2$ for $x \in \mathcal{X} + B_\infty(0,u)$. Then*

(i) The function f satisfies the upper bound $f(x) \le f_\mu(x) \le f(x) + L_0 u\sqrt{d}$

(ii) The function f_μ is L_0-Lipschitz over $\mathcal{X}$.

(iii) The function f_μ is continuously differentiable; moreover, its gradient is $\frac{2\sqrt{d}L_0}{u}$ Lipschitz continuous.

(iv) For random variables $Z \sim \mu$ and $\xi \sim P$, we have

$$\mathbb{E}[\nabla F(x+Z;\xi)] = \nabla f_\mu(x)$$

and

$$\mathbb{E}[\|\nabla f_\mu(x) - \nabla F(x+Z;\xi)\|_2^2] \le L_0^2.$$

The latter estimate is tight.

A similar lemma can be proved when μ is the density of the uniform distribution on $B_2(0,u)$. In this case, Yousefian et al. (2010) give (i)–(iii) of the following lemma.

Lemma 8 (Yousefian, Nedić, Shanbhag)**.** *Let f_μ be defined as in (5) where μ is the uniform density on the ℓ_2-ball of radius u. Assume that $\mathbb{E}[\|\partial F(x;\xi)\|_2^2] \le L_0^2$ for $x \in \mathcal{X} + B_2(0,u)$. Then*

(i) $f(x) \le f_\mu(x) \le f(x) + L_0 u$

(ii) f_μ *is L_0-Lipschitz over $\mathcal{X}$.*

(iii) f_μ *is continuously differentiable; moreover, its gradient is $\frac{L_0\sqrt{d}}{u}$-Lipschitz continuous.*

(iv) Let $Z \sim \mu$. Then $\mathbb{E}[\nabla F(x+Z;\xi)] = \nabla f_\mu(x)$, and $\mathbb{E}[\|\nabla f_\mu(x) - \nabla F(x+Z;\xi)\|_2^2] \le L_0^2$.

In addition, there exists a function f for which each of the bounds (i)–(iv) cannot be improved by more than a constant factor.

Lastly, for situations in which $F(\cdot;\xi)$ is L_0-Lipschitz with respect to the ℓ_2-norm over all of $\mathbb{R}^d$ and for P-a.e. ξ, we can use the normal distribution to perform smoothing of the expected function f. In the next lemma, we study continuity properties with respect to the ℓ_2-norm.

Lemma 9. *Let μ be the $N(0, u^2 I_{d\times d})$ distribution. Assume that $F(\cdot;\xi)$ is L_0-Lipschitz with respect to the ℓ_2-norm—that is*

$$\sup\{\|g\|_2 \mid g \in \partial F(x;\xi), x \in \mathcal{X}\} \le L_0 \qquad \text{for } P\text{-a.e. } \xi.$$

Then the following properties hold:

(i) $f(x) \le f_\mu(x) \le f(x) + L_0 u\sqrt{d}$

(ii) f_μ *is L_0-Lipschitz with respect to the ℓ_2 norm*

(iii) f_μ *is continuously differentiable; moreover, its gradient is $\frac{L_0}{u}$-Lipschitz continuous with respect to the ℓ_2-norm.*

(iv) Let $Z \sim \mu$. Then $\mathbb{E}[\nabla F(x+Z;\xi)] = \nabla f_\mu(x)$, and $\mathbb{E}[\|\nabla f_\mu(x) - \nabla F(x+Z;\xi)\|_2^2] \le L_0^2$.

In addition, there exists a function f for which each of the bounds (i)–(iv) cannot be improved by more than a constant factor.