Duchi, John, Bartlett, Peter, & Wainwright, Martin
(2012)
Randomized smoothing for (parallel) stochastic optimization.
In Valcher, E (Ed.) *Proceedings of the 2012 IEEE 51st Annual Conference on Decision and Control (CDC 2012).*
Institute of Electrical and Electronics Engineers Inc., United States of America, pp. 5442-5444.

https://doi.org/10.1109/CDC.2012.6426698

Randomized Smoothing for Stochastic Optimization

John C. Duchi[1] Peter Bartlett[1,2] Martin J. Wainwright[1,2]
jduchi@eecs.berkeley.edu bartlett@eecs.berkeley.edu wainwrig@eecs.berkeley.edu

Department of Electrical Engineering and Computer Sciences[1] Department of Statistics[2]
University of California Berkeley, Berkeley, CA 94720

March 2012

Abstract

We analyze convergence rates of stochastic optimization procedures for non-smooth convex optimization problems. By combining randomized smoothing techniques with accelerated gradient methods, we obtain convergence rates of 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 rates for non-smooth optimization. We give several applications of our results to statistical estimation problems, and provide experimental results that demonstrate the effectiveness of the proposed algorithms. We also describe how a combination of our algorithm with recent work on decentralized optimization yields a distributed stochastic optimization algorithm that is order-optimal.

1 Introduction

In this paper, we develop and analyze randomized smoothing procedures for solving the following class of stochastic optimization problems. Let $\{F(\cdot\,;\xi),\ \xi \in \Xi\}$ be a collection of real-valued functions, each with domain containing the closed convex set $\mathcal{X} \subseteq \mathbb{R}^d$. Letting P be a probability distribution over the index set Ξ, consider the function $f : \mathcal{X} \to \mathbb{R}$ defined via

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

In this paper, we analyze a family of randomized smoothing procedures for solving potentially non-smooth stochastic optimization problems of the form

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

where $\varphi : \mathcal{X} \to \mathbb{R}$ is a known regularizing function. Throughout the paper, we assume that f is convex on its domain $\mathcal{X}$. This condition is satisfied, for instance, if the function $F(\cdot;\xi)$ is convex for P-almost every ξ. We assume that φ is closed and convex, but we allow for non-differentiability so that the framework includes the ℓ_1-norm and related regularizers.

While we will later discuss the effects that $\varphi(x)$ has on our optimization procedures, throughout we will mostly consider the properties of the stochastic function f. Problem (2) is challenging mainly for two reasons. First, the function f may be non-smooth. Second, in many cases, f cannot actually be evaluated. When ξ is high-dimensional, the integral (1) cannot be efficiently computed, and in statistical learning problems we usually do not even know what the distribution P is. Thus, throughout this work, we assume only that we have access to a stochastic oracle that allows us to get i.i.d. samples $\xi \sim P$, and consequently we focus on stochastic gradient procedures for the convex program (2).

To address the first difficulty mentioned above—namely that f may be non-smooth—several researchers have considered techniques for smoothing the objective. Such approaches for deterministic non-smooth problems are by now well-known, and include Moreau-Yosida regularization (e.g. [22]), methods based on recession functions [3]; and a method that uses conjugate and proximal functions [26]. Several works study methods to replace constraints $f(x) \leq 0$ in convex programming problems with exact penalties $\max\{0, f(x)\}$ in the objective, after which smoothing is applied to the $\max\{0, \cdot\}$ operator (e.g., see the paper [8] and references therein). The difficulty of such approaches is that most require quite detailed knowledge of the structure of the function f to be minimized and hence are impractical in stochastic settings.

The second difficulty of solving the convex program (2) is that the function cannot actually be evaluated except through stochastic realizations of f and its (sub)gradients. In this paper, we develop an algorithm for solving problem (2) based on stochastic subgradient methods. Although such methods are classical [30, 11, 28], recent work by Juditsky et al. [15] and Lan [18, 19] has shown that if f is smooth—its gradients are Lipschitz continuous—convergence rates dependent on the variance of the stochastic gradient estimator are achievable. Specifically, if σ^2 is the variance of the gradient estimator, the convergence rate of the resulting stochastic optimization procedure is $\mathcal{O}(\sigma/\sqrt{T})$. Of particular relevance to our study is the following fact: if the oracle (instead of returning just a single estimate) returns m unbiased estimates of the gradient, the variance of the gradient estimator is reduced by a factor of m. Dekel et al. [9] exploit this fact to develop asymptotically order-optimal distributed optimization algorithms, as we discuss in the sequel.

To the best of our knowledge, there is no work on *non-smooth* stochastic problems for which a reduction in the variance of the stochastic estimate of the true subgradient gives an improvement in convergence rates. For non-smooth stochastic optimization, known convergence rates are dependent only on the Lipschitz constant of the functions $F(\cdot; \xi)$ and the number of actual updates performed. Within the oracle model of convex optimization [25], the optimizer has access to a black-box oracle that, given a point $x \in \mathcal{X}$, returns an unbiased estimate of a (sub)gradient of the objective f at the point x. In most stochastic optimization procedures, an algorithm updates a parameter x_t at every iteration by querying the oracle for one stochastic subgradient; we consider the natural extension to the case when the optimizer issues several queries to the stochastic oracle at every iteration.

A convolution-based smoothing technique amenable to non-smooth stochastic optimization problems is the starting point for our approach. A number of authors (e.g., [16, 32, 17, 38]) have noted that particular random perturbations of the variable x transform f into a smooth function. The intuition underlying such approaches is that convolving two functions yields a new function that is at least as smooth as the smoothest of the two original functions. In particular, let μ denote the density of a random variable with respect to Lebesgue measure, and consider the smoothed objective function

$$f_\mu(x) := \int_{\mathbb{R}^d} f(x + y)\mu(y)dy = \mathbb{E}_\mu[f(x + Z)], \tag{3}$$

where Z is a random variable with probability density μ. Clearly, f_μ is convex whenever f is convex; moreover, it is known that if μ is a density with respect to Lebesgue measure, then f_μ is differentiable [4].

We analyze minimization procedures that solve the non-smooth problem (2) by using stochastic gradient samples from the smoothed function (3) with appropriate choice of smoothing density μ. The main contribution of our paper is to 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. Our two main theorems quantify the above statement in terms of expected values (Theorem 1) and, under an additional reasonable tail condition, with high probability (Theorem 2). 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 f and R is the ℓ_2-radius of its domain.) As we discuss in Section 2.4, this convergence rate is optimal up to constant factors. Moreover, this fast rate of convergence has implications for applications in statistical problems, distributed optimization, and other areas, as discussed in Section 3.

The remainder of the paper is organized as follows. In the next section, we review standard techniques for stochastic optimization, noting a few of their deficiencies. After this, we state our algorithm and main theorems achieving faster rates of convergence for non-smooth stochastic problems using the randomized smoothing technique (3). We make strong use of the fine analytic properties of randomized smoothing, and collect several relevant results in Appendix E. In Section 3.1, we outline several applications of the smoothing techniques, which we complement in Section 3.2 with experiments and simulations showing the merits of our new approach. Section 4 contains proofs of our main results, though we defer more technical aspects to the appendices.

Notation: For the reader's convenience, here we specify notation as well as a few definitions. We use $B_p(x, u) = \{y \in \mathbb{R}^d \mid \|x - y\|_p \leq u\}$ to denote the closed p-norm ball of radius u around the point x. Addition of sets A and B is defined as the Minkowski sum in $\mathbb{R}^d$, that is, $A + B = \{x \in \mathbb{R}^d \mid x = y + z, y \in A, z \in B\}$, 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 or distribution μ, we let $\operatorname{supp} \mu := \{x \mid f(x) \neq 0\}$ denote its support. Given a convex function f with domain $\mathcal{X}$, for any $x \in \mathcal{X}$, we use $\partial f(x)$ to denote its subdifferential. We define the shorthand notation $\|\partial f(x)\| = \sup\{\|g\| \mid g \in \partial f(x)\}$ for any norm $\|\cdot\|$. The dual norm $\|\cdot\|_*$ with the norm $\|\cdot\|$ is defined as $\|z\|_* := \sup_{\|x\| \leq 1} \langle z, x\rangle$. A function f is L_0-Lipschitz with respect to the norm $\|\cdot\|$ over $\mathcal{X}$ if

$$|f(x) - f(y)| \leq L_0 \|x - y\|$$

for all $x, y \in \mathcal{X}$. For convex f, it is known [12] that f is L_0-Lipschitz in this sense if and only if $\sup_{x \in \mathcal{X}} \|\partial f(x)\|_* \leq L_0$. We say the gradient of f is L_1-Lipschitz continuous with respect to the norm $\|\cdot\|$ over $\mathcal{X}$ if

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

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

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

Given a convex and differentiable function ψ, the associated Bregman divergence [5] is given by $D_\psi(x, y) := \psi(x) - \psi(y) - \langle \nabla \psi(y), x - y\rangle$. When $X \in \mathbb{R}^{d_1 \times d_2}$ is a matrix, we let $\rho_i(X)$ denote its ith largest singular value, and when $X \in \mathbb{R}^{d \times d}$, we let $\lambda_i(X)$ denote its ith largest eigenvalue by modulus. The transpose of X is denoted $X^\top$. The notation $\xi \sim P$ indicates that ξ is drawn according to the distribution P.

2 Main results and some consequences

In this section, we begin by motivating the algorithm studied in this paper, and then state our main results on its convergence behavior.

2.1 Some background

We focus on stochastic gradient descent methods[1] based on dual averaging schemes [27] for solving the stochastic problem (2). Dual averaging methods are based on a proximal function ψ, which is assumed strongly convex with respect to a norm $\|\cdot\|$. The update scheme of such a method is as follows. Given a point $x_t \in \mathcal{X}$, the algorithm queries a stochastic oracle and receives a random vector g_t such that $\mathbb{E}[g_t] \in \partial f(x_t)$. The algorithm then performs the update

$$x_{t+1} = \operatorname*{argmin}_{x\in\mathcal{X}} \left\{ \sum_{\tau=0}^{t} \langle g_\tau, x\rangle + \frac{1}{\alpha_t}\psi(x) \right\} \tag{4}$$

where $\alpha_t > 0$ is a sequence of stepsizes. Under some mild assumptions, the algorithm is guaranteed to converge for stochastic problems. For instance, suppose that ψ is strongly convex with respect to the norm $\|\cdot\|$, and moreover that $\mathbb{E}[\|g_t\|_*^2] \le L_0^2$ for all t, where we recall that $\|\cdot\|_*$ denotes the dual norm to $\|\cdot\|$. Then, with stepsizes $\alpha_t \propto R/L_0\sqrt{t}$, it is known that the sequence $\{x_t\}_{t=0}^{\infty}$ generated by the updates (4) satisfies

$$\mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=1}^{T} x_t \right)\right] - f(x^*) = \mathcal{O}\left(\frac{L_0\sqrt{\psi(x^*)}}{\sqrt{T}} \right). \tag{5}$$

We refer the reader to papers by Nesterov [27] and Xiao [36] for results of this type.

An unsatisfying aspect of the bound (5) is the absence of any role for the variance of the (sub)gradient estimator g_t. In particular, even if an algorithm is able to obtain $m > 1$ samples of the gradient of f at x_t—thereby giving a significantly more accurate gradient estimate—this result fails to capture the likely improvement of the method. We address this problem by stochastically smoothing the non-smooth objective f and then adapt recent work on so-called "accelerated" gradient methods [19, 34, 36] to achieve variance-based improvements. Accelerated methods work only when the function f is smooth—that is, when it has Lipschitz continuous gradients. Thus, we turn now to developing the tools necessary to stochastically smooth the non-smooth objective (2).

2.2 Description of algorithm

Our algorithm is based on observations of stochastically perturbed gradient information at each iteration, where we slowly decrease the perturbation as the algorithm proceeds. More precisely, our algorithm uses the following scheme. Let $\{u_t\} \subset \mathbb{R}_+$ be a non-increasing sequence of positive real numbers; these quantities control the perturbation size. At iteration t, rather than query the stochastic oracle at the point y_t, the algorithm queries the oracle at m points drawn randomly from some neighborhood around y_t. Specifically, it performs the following three steps:

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

(2) Queries the oracle at the m points $y_t + u_t Z_{i,t}$, $i = 1, 2, \ldots, m$, yielding stochastic gradients

$$g_{i,t} \in \partial F(y_t + u_t Z_{i,t}, \xi_{i,t}), \quad \text{where } \xi_{i,t} \sim P, \text{ for } i = 1, 2, \ldots, m. \tag{6}$$

[1] We note in passing that essentially identical results can also be obtained for methods based on mirror descent [25, 34], though we omit these so as not to overburden the reader.

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

Here and throughout we denote the distribution of the random variable $u_t Z$ by μ_t, and we note that this procedure ensures $\mathbb{E}[g_t \mid y_t] = \nabla f_{\mu_t}(y_t) = \nabla \mathbb{E}[F(y_t + u_t Z; \xi) \mid y_t]$, where f_{μ_t} is the smoothed function (3) and μ_t is the density of u_t.

By combining the sampling scheme (6) with extensions of Tseng's recent work on accelerated gradient methods [34], we can achieve stronger convergence rates for solving the non-smooth objective (2). The update we propose is essentially a smoothed version of the simpler method (4). The method uses three series of points, denoted $\{x_t, y_t, z_t\} \in \mathcal{X}^3$. We use y_t as a "query point", so that at iteration t, the algorithm receives a vector g_t as described in the sampling scheme (6). The three sequences evolve according to a dual-averaging algorithm, which in our case involves three scalars (L_t, θ_t, η_t) to control step sizes. The recursions are as follows:

$$y_t = (1 - \theta_t)x_t + \theta_t z_t \tag{7a}$$

$$z_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \sum_{\tau=0}^{t} \frac{1}{\theta_\tau} \langle g_\tau, x \rangle + \sum_{\tau=0}^{t} \frac{1}{\theta_\tau} \varphi(x) + L_{t+1}\psi(x) + \frac{\eta_{t+1}}{\theta_{t+1}}\psi(x) \right\} \tag{7b}$$

$$x_{t+1} = (1 - \theta_t)x_t + \theta_t z_{t+1}. \tag{7c}$$

In prior work on accelerated schemes for stochastic and non-stochastic optimization [34, 19, 36], the term L_t is set equal to the Lipschitz constant of ∇f; in contrast, our choice of varying L_t allows our smoothing schemes to be oblivious to the number of iterations T. The extra damping term η_t/θ_t provides control over the fluctuations induced by using the random vector g_t as opposed to deterministic subgradient information. As in Tseng's work [34], we assume that $\theta_0 = 1$ and $(1 - \theta_t)/\theta_t^2 = 1/\theta_{t-1}^2$; the latter equality is ensured by setting $\theta_t = 2/(1 + \sqrt{1 + 4/\theta_{t-1}^2})$.

2.3 Convergence rates

We now state our two main results on the convergence rate of the randomized smoothing procedure (6) with accelerated dual averaging updates (7a)–(7c). So as to avoid cluttering the theorem statements, we begin by stating our main assumptions and notation. When we state that a function f is Lipschitz continuous, we mean with respect to the norm $\|\cdot\|$, whose dual norm we denote $\|\cdot\|_*$, and we assume that ψ is nonnegative and strongly convex with respect to $\|\cdot\|$. Our main assumption ensures that the smoothing operator and smoothed function f_μ are relatively well-behaved.

Assumption A (Smoothing properties)**.** *The random variable Z is zero-mean with density μ (with respect to Lebesgue measure on the affine hull $\operatorname{aff}(\mathcal{X})$ of $\mathcal{X}$), and there are constants L_0 and L_1 such that for all $u > 0$, $\mathbb{E}[f(x+uZ)] \le f(x) + L_0 u$, and $\mathbb{E}[f(x+uZ)]$ has $\frac{L_1}{u}$-Lipschitz continuous gradient with respect to the norm $\|\cdot\|$. For P-almost every $\xi \in \Xi$, we have $\operatorname{dom} F(\cdot;\xi) \supseteq u_0 \operatorname{supp}\mu + \mathcal{X}$.*

Let μ_t denote the density of the random vector $u_t Z$ and define the instantaneous smoothed function $f_{\mu_t} = \int f(x + z) d\mu_t(z)$. As discussed in the introduction, the function f_{μ_t} is guaranteed to be smooth whenever μ (and hence μ_t) is a density with respect to Lebesgue measure, so Assumption A ensures that f_{μ_t} is uniformly close to f and not too "jagged." Many smoothing distributions, including Gaussians and uniform distributions on norm balls, satisfy Assumption A (see Appendix E); we use such examples in the corollaries to follow. The containment of $u_0 \operatorname{supp}\mu + \mathcal{X}$ in $\operatorname{dom} F(\cdot;\xi)$ guarantees that the subdifferential $\partial F(\cdot;\xi)$ is non-empty at all sampled points $y_t + u_t Z$.

Indeed, since μ is a density with respect to Lebesgue measure on $\text{aff}(\mathcal{X})$, with probability one $y_t + u_t Z \in \text{relint}\,\text{dom}\, F(\cdot;\xi)$ and thus [12] the subdifferential $\partial F(y_t + u_t Z; \xi) \neq \emptyset$. There are many smoothing distributions μ, including standard Gaussian and uniform distributions on norm balls, for which Assumption A holds (see Appendix E), and we use such examples in the corollaries to follow.

In the algorithm (7a)–(7c), we set L_t to be an upper bound on the Lipschitz constant $\frac{L_1}{u_t}$ of the gradient of $\mathbb{E}[f(x + u_t Z)]$; this choice ensures good convergence properties of the algorithm. The following is the first of our main theorems.

Theorem 1. *Define $u_t = \theta_t u$, use the scalar sequence $L_t = L_1/u_t$, and assume that η_t is non-decreasing. Under Assumption A, for any $x^* \in \mathcal{X}$ and $T \geq 4$,*

$$\mathbb{E}[f(x_T) + \varphi(x_T)] - [f(x^*) + \varphi(x^*)] \leq \frac{6L_1\psi(x^*)}{Tu} + \frac{2\eta_T\psi(x^*)}{T} + \frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{\eta_t}\mathbb{E}\left[\,\|e_t\|_*^2\,\right] + \frac{4L_0 u}{T}, \quad (8)$$

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

Remarks: Note that the convergence rate (8) involves the variance $\mathbb{E}[\|e_t\|_*^2]$ explicitly. We exploit this fact in the corollaries to be stated shortly. In addition, note that Theorem 1 does not require a priori knowledge of the number of iterations T to be performed, which renders it suitable to online and streaming applications. If such knowledge is available, then it is possible to give a similar result using the smoothing parameter $u_t \equiv u$ for all t; such a result is stated as Theorem 3 in Section 4.

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 $e_t := \nabla f_{\mu_t}(y_t) - g_t$. In particular, let $\mathcal{F}_t$ denote the σ-field of the random variables $g_{i,s}$, $i = 1, \ldots, m$ and $s = 0, \ldots, t$. In order to achieve high-probability convergence results, a subset of our results involve the following assumption.

Assumption B (Sub-Gausian errors)**.** *The error is* $(\|\cdot\|_*, \sigma)$ sub-Gaussian *for some $\sigma > 0$, meaning that with probability 1*

$$\mathbb{E}[\exp(\|e_t\|_*^2/\sigma^2) \mid \mathcal{F}_{t-1}] \leq \exp(1) \quad \textit{for all } t \in \{1, 2, 3, \ldots\}. \quad (9)$$

We refer the reader to Appendix F for more background on sub-Gaussian and sub-exponential random variables. In past work on smooth optimization, other authors [15, 19, 36] have imposed this type of tail assumption, and we discuss sufficient conditions for the assumption to hold in Corollary 4 in the following section.

Theorem 2. *In addition to the conditions of Theorem 1, assume $\mathcal{X}$ is compact with $\|x - x^*\| \leq R$ for all $x \in \mathcal{X}$ and that Assumption B holds. Then with probability at least $1 - 2\delta$, the algorithm with step size $\eta_t = \eta\sqrt{t+1}$ satisfies*

$$f(x_T) + \varphi(x_T) - [f(x^*) + \varphi(x^*)] \leq \frac{6L_1\psi(x^*)}{Tu} + \frac{4L_0 u}{T} + \frac{4\eta_T\psi(x^*)}{T+1} + \theta_{T-1}\sum_{t=0}^{T-1}\frac{1}{2\eta_t}\mathbb{E}[\|e_t\|_*^2] + \frac{4\sigma^2 \max\left\{\log\frac{1}{\delta}, \sqrt{2e(\log T + 1)\log\frac{1}{\delta}}\right\}}{\eta T} + \frac{\sigma R\sqrt{\log\frac{1}{\delta}}}{\sqrt{T}}.$$

Remarks: The first four terms in the convergence rate Theorem 2 gives are essentially identical to the expected rate of Theorem 1. The first of the additional terms decreases at a rate of $1/T$, while the second decreases at a rate of $\sigma/\sqrt{T}$. As we discuss in the Corollaries that follow, the dependence $\sigma/\sqrt{T}$ on the variance σ^2 is optimal, and an appropriate choice of the sequence η_t in Theorem 1 yields the same rates to constant factors.

2.4 Some consequences

The corollaries of the above theorems—and the consequential optimality guarantees of the algorithm above—are our main focus for the remainder of this section. Specifically, we show concrete convergence bounds for algorithms using different choices of the smoothing distribution μ. For each corollary, we make the assumption that $x^* \in \mathcal{X}$ satisfies $\psi(x^*) \le R^2$, but is otherwise arbitrary, that the iteration number $T \ge 4$, and that $u_t = u\theta_t$.

We begin with a corollary that provides bounds when the smoothing distribution μ is uniform on the ℓ_2-ball. The conditions on F in the corollary hold, for example, when $F(\cdot;\xi)$ is L_0-Lipschitz with respect to the ℓ_2-norm for P-a.e. sample of ξ.

Corollary 1. *Let μ be uniform on $B_2(0,1)$ and assume $\mathbb{E}[\|\partial F(x;\xi)\|_2^2] \le L_0^2$ for $x \in \mathcal{X} + B_2(0,u)$, where we set $u = Rd^{1/4}$. With the step size choices $\eta_t = L_0\sqrt{t+1}/R\sqrt{m}$ and $L_t = L_0\sqrt{d}/u_t$,*

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

The following corollary shows that similar convergence rates are attained when smoothing with the normal distribution.

Corollary 2. *Let μ be the d-dimensional normal distribution with zero-mean and identity covariance I and assume that $F(\cdot;\xi)$ is L_0-Lipschitz with respect to the ℓ_2-norm for P-a.e. ξ. With smoothing parameter $u = Rd^{-1/4}$ and step sizes $\eta_t = L_0\sqrt{t+1}/R\sqrt{m}$ and $L_t = L_0/u_t$, we have*

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

We remark here (deferring deeper discussion to Lemma 10) that the dimension dependence of $d^{1/4}$ on the $1/T$ term in the previous corollaries cannot be improved by more than a constant factor. Essentially, functions f exist whose smoothed version f_μ cannot have both Lipschitz continuous gradient and be uniformly close to f in a dimension-independent sense, at least for the uniform or normal distributions.

The advantage of using normal random variables—as opposed to Z uniform on $B_2(0,u)$—is that no normalization of Z is necessary, though there are more stringent requirements on f. The lack of normalization is a useful property in very high dimensional scenarios, such as statistical natural language processing [23]. Similarly, we can sample Z from an ℓ_∞ ball, which, like $B_2(0,u)$, is still compact, but gives slightly looser bounds than sampling from $B_2(0,u)$. Nonetheless, it is much easier to sample from $B_\infty(0,u)$ in high dimensional settings, especially sparse data scenarios such as NLP where only a few coordinates of the random variable Z are needed.

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 [25]. By using different distributions μ for the random perturbations $Z_{i,t}$ in (6), 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 optimization performance than the ℓ_2-ℓ_2 pair above.

Corollary 3. *Let μ be the uniform density on $B_\infty(0,1)$ 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 $\xi \in \Xi$, where we set $u = R\sqrt{d \log d}$. Use the proximal function $\psi(x) = \frac{1}{2(p-1)} \|x\|_p^2$ for $p = 1 + 1/\log d$ and set $\eta_t = L_0\sqrt{t+1}/R\sqrt{m \log d}$ and $L_t = L_0/u_t$. There is a universal constant C such that*

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

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 when taking advantage of non-Euclidean geometry. Nonetheless, for large T, the $1/\sqrt{Tm}$ terms dominate the convergence rates, and Corollary 3 can be an improvement.

Our final corollary specializes the high probability convergence result in Theorem 2 by showing that the error is sub-Gaussian (9) under the assumptions in the corollary. 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 smoothing distribution μ uniform on $B_2(0,1)$, smoothing parameter $u = Rd^{1/4}$, damping parameter $\eta_t = L_0\sqrt{t+1}/R\sqrt{m}$, and instantaneous Lipschitz estimate $L_t = L_0\sqrt{d}/u_t$, with probability at least $1-\delta$,*

$$\begin{aligned}&f(x_T) + \varphi(x_T) - f(x^*) - \varphi(x^*) \\ &= \mathcal{O}\left(\frac{L_0 R d^{1/4}}{T} + \frac{L_0 R}{\sqrt{Tm}} + \frac{L_0 R\sqrt{\log\frac{1}{\delta}}}{\sqrt{Tm}} + \frac{L_0 R \max\{\log\frac{1}{\delta}, \log T\}}{T\sqrt{m}}\right).\end{aligned}$$

Remarks: We make two remarks about the above corollaries. The first is that if one abandons the requirement that the optimization procedure be an "anytime" algorithm—always able to return a result—it is possible to give similar results by using a fixed setting of $u_t \equiv u$ throughout. In particular, using Theorem 3 in Section 4.4 we can use $u_t = u/T$ to get essentially the same results as Corollaries 1–3. As a side benefit, it is then easier to satisfy the Lipschitz condition that $\mathbb{E}[\|\partial F(x;\xi)\|^2] \le L_0^2$ for $x \in \mathcal{X} + \operatorname{supp}\mu$. Our second observation is that Theorem 1 and the corollaries appear to require a very specific setting of the constant L_t to achieve fast rates. However, the algorithm is in fact robust to mis-specification of L_t, since the instantaneous smoothness constant L_t is dominated by the stochastic damping term η_t in the algorithm. Indeed, since η_t grows proportionally to $\sqrt{t}$ for each corollary, we always have $L_t = L_1/u_t = L_1/\theta_t u = \mathcal{O}(\eta_t/\sqrt{t}\theta_t)$; that is, L_t is order $\sqrt{t}$ smaller than η_t/θ_t, so setting L_t incorrectly up to order $\sqrt{t}$ has essentially negligible effect.

We can show the bounds in the theorems above are tight, that is, unimprovable up to constant factors, by exploiting known lower bounds [25, 1] for stochastic optimization problems. We re-state some of these results here. For instance, let $\mathcal{X} = \{x \in \mathbb{R}^d \mid \|x\|_2 \le R_2\}$, and consider all convex functions f that are $L_{0,2}$-Lipschitz with respect to the ℓ_2-norm. Assume that the stochastic oracle, when queried at a point x, returns a vector g whose expectation is in $\partial f(x)$ with $\mathbb{E}[\|g\|_2^2] \le L_{0,2}^2$. Then for *any* method that outputs a point $x_T \in \mathcal{X}$ after T queries of the oracle, we have the lower bound

$$\sup_f \left\{ \mathbb{E}[f(x_T)] - \min_{x \in \mathcal{X}} f(x) \right\} = \Omega\left(\frac{L_{0,2} R_2}{\sqrt{T}} \right),$$

where the supremum is taken over convex f that are $L_{0,2}$-Lipschitz with respect to the ℓ_2-norm [1, Section 3.1]. Similar bounds hold for problems with non-Euclidean geometry [1]; in particular, consider convex f that are $L_{0,\infty}$-Lipschitz with respect to the ℓ_1-norm, that is, $|f(x) - f(y)| \le L_{0,\infty} \|x - y\|_1$. Then setting $\mathcal{X} = \{x \in \mathbb{R}^d \mid \|x\|_1 \le R_1\}$, we have $B_\infty(0, R_1/d) \subset B_1(0, R_1)$ and thus

$$\sup_f \left\{ \mathbb{E}[f(x_T)] - \min_{x \in \mathcal{X}} f(x) \right\} = \Omega\left(\frac{L_{0,\infty} R_1}{\sqrt{T}} \right).$$

In either geometry, no method can have optimization error smaller than $\mathcal{O}(LR/\sqrt{T})$ after T queries of the stochastic oracle.

Let us compare the above lower bounds to the convergence rates in Corollaries 1 through 3. Examining the bound in Corollaries 1 and 2, we see that the dominant terms are order $L_0 R/\sqrt{Tm}$ so long as $m \le T/\sqrt{d}$. Since our method issues Tm queries to the oracle, this is optimal. Similarly, the strategy of sampling uniformly from the ℓ_∞-ball in Corollary 3 is optimal up to factors logarithmic in the dimension. In contrast to other optimization procedures, however, our algorithm performs an update to the parameter x_t only after every m queries to the oracle; as we show in the next section, this is beneficial in several applications.

3 Applications and experimental results

In this section, we describe some applications of our results, and then give experimental results that illustrate our theoretical predictions.

3.1 Some applications

The first application of our results is to parallel computation and distributed optimization. Imagine that instead of querying the stochastic oracle serially, we can issue queries and aggregate the resulting stochastic gradients in parallel. In particular, assume that we have a procedure in which the m queries of the stochastic oracle occur concurrently. Then Corollaries 1–4 imply that in the same amount of time required to perform T queries and updates of the stochastic gradient oracle serially, achieving an optimization error of $\mathcal{O}(1/\sqrt{T})$, the parallel implementation can process Tm queries and consequently has optimization error $\mathcal{O}(1/\sqrt{Tm})$.

We now briefly describe two possibilities for a distributed implementation of the above. The simplest architecture is a master-worker architecture, in which one master maintains the parameters (x_t, y_t, z_t), and each of m workers has access to an uncorrelated stochastic oracle for P and the smoothing distribution μ. The master broadcasts the point y_t to the workers, which sample $\xi_i \sim P$ and $Z_i \sim \mu$, returning sample gradients to the master. In a tree-structured network, broadcast and aggregation require at most $\mathcal{O}(\log m)$ steps; the relative speedup over a serial implementation is $\mathcal{O}(m/\log m)$. In recent work, Dekel et al. [9] give a series of reductions showing how to distribute variance-based stochastic algorithms and achieve an asymptotically optimal convergence rate. The algorithm given here, as specified by equations (6) and (7a)–(7c), can be exploited within their framework to achieve an $\mathcal{O}(m)$ improvement in convergence rate over a serial implementation. More precisely, whereas achieving optimization error ϵ requires $\mathcal{O}(1/\epsilon^2)$ iterations for a centralized algorithm, the distributed adaptation requires only $\mathcal{O}(1/(m\epsilon^2))$ iterations. Such an improvement is possible as a consequence of the variance reduction techniques we have described.

A second application of interest involves problems where the set $\mathcal{X}$ and/or the function φ are complicated, so that calculating the proximal update (7b) becomes expensive. The proximal update may be distilled to computing

$$\min_{x\in\mathcal{X}} \left\{ \langle g, x\rangle + \psi(x) \right\} \quad \text{or} \quad \min_{x\in\mathcal{X}} \left\{ \langle g, x\rangle + \psi(x) + \varphi(x) \right\}. \tag{10}$$

In such cases, it may be beneficial to accumulate gradients by querying the stochastic oracle several times in each iteration, using the averaged subgradient in the update (7b), and thus solve only one proximal sub-problem for a collection of samples.

Let us consider some concrete examples. In statistical applications involving the estimation of covariance matrices, the domain $\mathcal{X}$ is constrained in the positive semidefinite cone $\{X \in \mathbb{S}_n \mid X \succeq 0\}$; other applications involve additional nuclear-norm constraints of the form $\mathcal{X} = \{X \in \mathbb{R}^{d_1\times d_2} \mid \sum_{j=1}^{\min\{d_1,d_2\}} \rho_j(X) \le C\}$. Examples of such problems include covariance matrix estimation, matrix completion, and model identification in vector autoregressive processes (see the paper [24] and references therein for further discussion). Another example is the problem of metric learning [37, 33], in which the learner is given a set of n points $\{a_1, \ldots, a_n\} \subset \mathbb{R}^d$ and a matrix $B \in \mathbb{R}^{n\times n}$ indicating which points are close together in an unknown metric. The goal is to estimate a positive semidefinite matrix $X \succeq 0$ such that $\langle (a_i - a_j), X(a_i - a_j)\rangle$ is small when a_i and a_j belong to the same class or are close, while $\langle (a_i - a_j), X(a_i - a_j)\rangle$ is large when a_i and a_j belong to different classes. It is desirable that the matrix X have low rank, which allows the statistician to discover structure or guarantee performance on unseen data. As a concrete illustration, suppose that we are given a matrix $B \in \{-1, 1\}^{n\times n}$, where $b_{ij} = 1$ if a_i and a_j belong to the same class, and $b_{ij} = -1$ otherwise. In this case, one possible optimization-based estimator involves solving the non-smooth program

$$\min_{X,x} \frac{1}{\binom{n}{2}} \sum_{i<j} \Big[1 + b_{ij}(\mathrm{tr}(X(a_i - a_j)(a_i - a_j)^\top) + x)\Big]_+ \quad \text{s.t.} \;\; X \succeq 0, \;\; \mathrm{tr}(X) \le C. \tag{11}$$

Now let us consider the cost of computing the projection update (10) for the metric learning problem (11). When $\psi(X) = \frac{1}{2}\|X\|_{\mathrm{Fr}}^2$, the update (10) reduces for appropriate choice of V to

$$\min_X \frac{1}{2}\|X - V\|_{\mathrm{Fr}}^2 \quad \text{subject to} \quad X \succeq 0, \;\; \mathrm{tr}(X) \le C.$$

(As a side-note, it is possible to generalize this update to Schatten p-norms [10].) The above problem is equivalent to projecting the eigenvalues of V to the simplex $\{x \in \mathbb{R}^d \mid x \succeq 0, \langle \mathbb{1}, x\rangle \le C\}$, which after an $\mathcal{O}(d^3)$ eigen-decomposition requires time $\mathcal{O}(d)$ [6]. To see the benefits of the randomized perturbation and averaging technique (6) over standard stochastic gradient descent (4), consider that the cost of querying a stochastic oracle for the objective (11) for one sample pair (i, j) requires time $\mathcal{O}(d^2)$. Thus, m queries require $\mathcal{O}(md^2)$ computation, and each update requires

$\mathcal{O}(d^3)$. So we see that after $Tmd^2 + Td^3$ units of computation, our randomized perturbation method has optimization error $\mathcal{O}(1/\sqrt{Tm})$, while standard stochastic gradient requires Tmd^3 units of computation. In short, for $m \approx d$ the randomized smoothing technique (6) uses a factor $\mathcal{O}(d)$ less computation than standard stochastic gradient; we give experiments corroborating this in Section 3.2.2.

3.2 Experimental results

We now describe some experimental results that confirm the sharpness of our theoretical predictions. The first experiment explores the benefit of using multiple samples m when estimating the gradient $\nabla f(y_t)$ as in the averaging step (6). The second experiment studies the actual amount of time required to solve a statistical metric learning problem, as described in the objective (11) above. The third investigates whether the smoothing technique is essential to algorithms solving non-smooth stochastic problems—that is, whether the smoothing is only a proof device or whether it is necessary to achieve good performance.

3.2.1 Iteration Complexity of Reduced Variance Estimators

In this experiment, we consider the number of iterations of the accelerated method (7a)–(7c) required to achieve an ϵ-optimal solution to the problem (2). To understand how the iteration scales with the number m of gradient samples, it can be useful to consider our results in terms of the number of iterations $T(\epsilon, m)$ required to achieve optimization error ϵ for the optimization procedure when using m gradient samples in the averaging step (6). Specifically, we define

$$T(\epsilon, m) = \inf \Big\{ t \in \mathbb{N} \mid f(x_t) - \min_{x \in \mathcal{X}} f(x^*) \le \epsilon \Big\}.$$

We focus on the algorithm analyzed in Corollary 1, which uses uniform sampling of the ℓ_2-ball. The theorem implies there should be two regimes of convergence: one when the number m of samples is small, so that the $L_0R/\sqrt{Tm}$ term is dominant, and the other when m is large, so the $L_0Rd^{1/4}/T$ term is dominant. By inverting the first term, we see that for small m, $T(\epsilon, m) = \mathcal{O}(L_0^2R^2/m\epsilon^2)$, while the second gives $T(\epsilon, m) = \mathcal{O}(L_0Rd^{1/4}/\epsilon)$. In particular, our theory predicts

$$T(\epsilon, m) = \mathcal{O}\left(\max\left\{ \frac{L_0^2R^2}{m\epsilon^2}, \frac{L_0Rd^{1/4}}{\epsilon} \right\} \right). \tag{12}$$

To assess the accuracy of the prediction (12), we consider a robust linear regression problem, commonly studied in system identification and robust statistics [29, 14]. Specifically, we have a matrix $A \in \mathbb{R}^{n \times d}$ and vector $b \in \mathbb{R}^n$, and seek to minimize

$$f(x) = \frac{1}{n} \left\| Ax - b \right\|_1 = \frac{1}{n} \sum_{i=1}^{n} | \langle a_i, x \rangle - b_i |, \tag{13}$$

where $a_i \in \mathbb{R}^d$ denotes a transposed row of A. It is clear that the problem (13) is non-smooth. The stochastic oracle in this experiment, when queried at a point x, chooses an index $i \in [n]$ uniformly at random and returns $\text{sign}(\langle a_i, x \rangle - b_i)a_i$.

To perform our experiments, we generate $n = 1000$ points in dimensions $d \in \{50 \cdot 2^i\}_{i=0}^5$, each with fixed norm $\left\| a_i \right\|_2 = L_0$, and then assign values b_i by computing $\langle a_i, w \rangle$ for a random vector

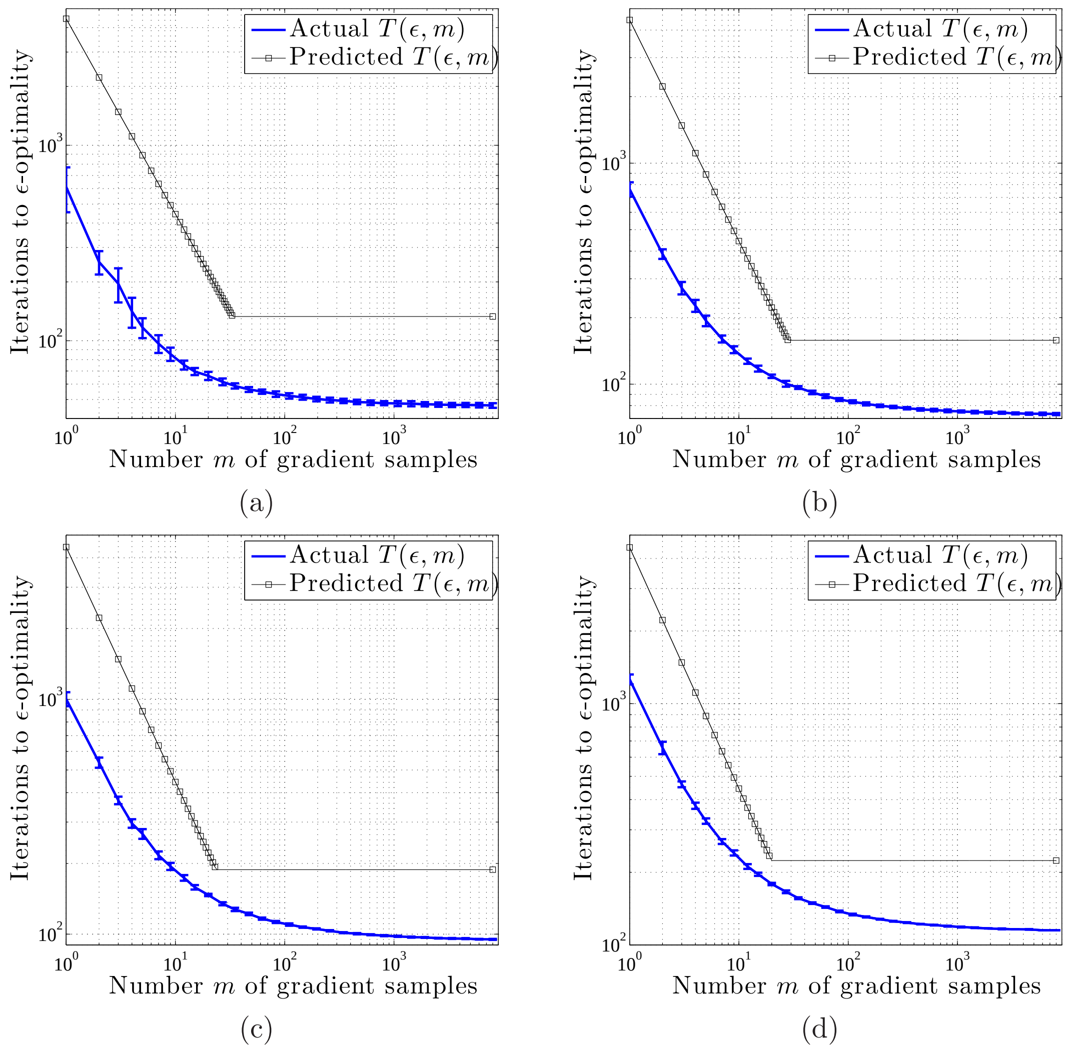

Figure 1. The number of iterations $T(\epsilon, m)$ to achieve an ϵ-optimal solution for the problem (13) as a function of the number of samples m used in the gradient estimate (6). The prediction (12) is the square black line in each plot, and each plot shows results for different dimensions d: (a) $d = 50$, (b) $d = 100$, (c) $d = 200$, (d) $d = 400$

w (adding normally distributed noise with variance 0.1). We estimate the quantity $T(\epsilon, m)$ for solving the robust regression problem (13) for several values of m and d. Figure 1 shows results for dimensions $d \in \{50, 100, 200, 400\}$, averaged over 20 experiments for each choice of dimension d. (Other settings of d exhibited similar behavior.) Each panel in the figure shows—on a log-log scale—the experimental average $T(\epsilon, m)$ and the theoretical prediction (12). The decrease in $T(\epsilon, m)$ is nearly linear for smaller numbers of samples m; for larger m, the effect is quite diminished. We present numerical results in Table 1 that allow us to roughly estimate the number m at which increasing the batch size in the gradient estimate (6) gives no further improvement. For each dimension d, Table 1 indeed shows that from $m = 1$ to 5, the iteration count $T(\epsilon, m)$ decreases linearly, halving again when we reach $m \approx 20$, but between $m = 100$ and 1000 there is at most

	m	1	2	3	5	20	100	1000	10000
$d = 50$	Mean	612.2	252.7	195.9	116.7	66.1	52.2	47.7	46.6
	Std	158.29	34.67	38.87	13.63	3.18	1.66	1.42	1.28
$d = 100$	Mean	762.5	388.3	272.4	193.6	108.6	83.3	75.3	73.3
	Std	56.70	19.50	17.59	10.65	1.91	1.27	0.78	0.78
$d = 200$	Mean	1002.7	537.8	371.1	267.2	146.8	109.8	97.9	95.0
	Std	68.64	26.94	13.75	12.70	1.66	1.25	0.54	0.45
$d = 400$	Mean	1261.9	656.2	463.2	326.1	178.8	133.6	118.6	115.0
	Std	60.17	38.59	12.97	8.36	2.04	1.02	0.49	0.00
$d = 800$	Mean	1477.1	783.9	557.9	388.3	215.3	160.6	142.0	137.4
	Std	44.29	24.87	12.30	9.49	2.90	0.66	0.00	0.49
$d = 1600$	Mean	1609.5	862.5	632.0	448.9	251.5	191.1	169.4	164.0
	Std	42.83	30.55	12.73	8.17	2.73	0.30	0.49	0.00

Table 1. The number of iterations $T(\epsilon, m)$ to achieve ϵ-accuracy for the regression problem (13) as a function of number of gradient samples m used in the gradient estimate (6) and the dimension d. Each box in the table shows the mean and standard deviation of $T(\epsilon, m)$ measured over 20 trials.

an 11% difference in $T(\epsilon, m)$, while between $m = 1000$ and $m = 10000$ the decrease amounts to at most 3%. The good qualitative match between the iteration complexity predicted by our theory and the actual performance of the methods is clear.

3.2.2 Metric Learning

Our second set of experiments apply to instances of metric learning. The data we receive consists of a set of vectors $a_i \in \mathbb{R}^d$ and measures $b_{ij} \geq 0$ 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 statistical goal is to learn a matrix X—inducing a pseudo-norm via $\|a\|_X^2 := \langle a, Xa \rangle$—such that $\langle (a_i - a_j), X(a_i - a_j) \rangle \approx b_{ij}$. Consequently, we solve the regression-like problem

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

The stochastic oracle for this problem is simple: given a query matrix X, the oracle chooses a pair (i, j) uniformly at random, then returns the subgradient

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

We solve ten random problems with dimension $d = 100$ and $n = 2000$, giving an objective with $4 \cdot 10^6$ terms and 5050 parameters. We plot experimental results in Fig. 2 showing 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 (6). It is clear that performing stochastic optimization is more viable for this problem than a non-stochastic method, as even computing the objective requires $\mathcal{O}(n^2 d^2)$ operations. As predicted by our theory and discussion in Sec. 3, it is clear that receiving more samples m 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 actual time taken to minimize the objective; the plot in Fig. 2 suggests that this too is correct, as the plots for $m = 64$ and $m = 128$ are essentially indistinguishable.

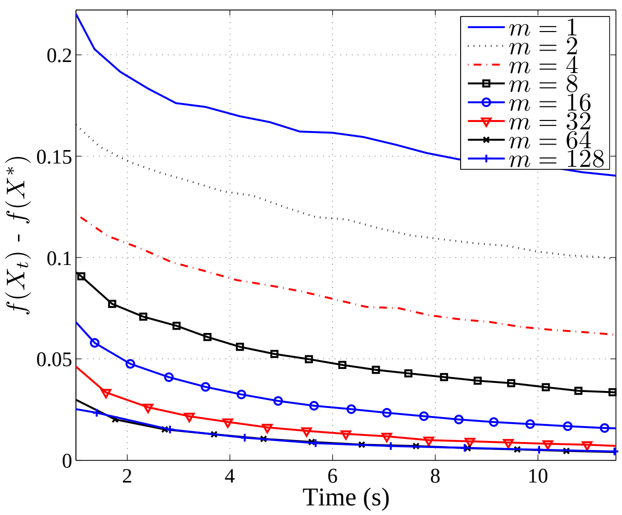

Figure 2. Optimization error $f(X_t) - \inf_{X^* \in \mathcal{X}} f(X^*)$ in the metric learning problem of Sec. 3.2.2 as a function of time in seconds. Each line indicates optimization error over time for a particular number of samples m in the gradient estimate (6); we set $m = 2^i$ for $i = \{1, \ldots, 7\}$.

3.2.3 Necessity of randomized smoothing

A reasonable question is whether the extra sophistication of the random smoothing (6) is necessary. Can receiving more samples m from the stochastic oracle—all evaluated at the same point—give the same benefit to the simple dual averaging method (4)? We do not know the full answer to this question, though we give an experiment here that suggests that the answer is negative, in that smoothing does give demonstrable improvement.

For this experiment, we use the objective

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} \|x - a_i\|_1 , \tag{14}$$

where the $a_i \in \{-1, +1\}^d$, and each component j of the vector a_i is sampled independently from $\{-1, 1\}$ and equal to 1 with probability $1/\sqrt{j}$. Even as $n \uparrow \infty$, the function f remains non-smooth, since the a_i belong to a discrete set and each value of a_i occurs with positive probability. As in Sec. 3.2.1, we compute $T(\epsilon, m)$ to be the number of iterations required to achieve an ϵ-optimal solution to the objective (14). We compare two algorithms that use m queries of the stochastic gradient oracle, which when queried at a point x chooses an index $i \in [n]$ uniformly at random and returns $\text{sign}(x - a_i) \in \partial \|x - a_i\|_1$. The first algorithm is the dual averaging algorithm (4), where g_t is the average of m queries to the stochastic oracle at the current iterate x_t. The second is the accelerated method (7a)–(7c) with the randomized averaging (6). We plot the results in Fig. 3. We plot the best stepsize sequence α_t for the update (4) of several we tested to make comparison as favorable as possible for simple mirror descent. It is clear that while there is moderate improvement for the non-smooth method when the number of samples m grows, and both methods are (unsurprisingly) essentially indistinguishable for $m = 1$, the smoothed sampling strategy has much better iteration complexity as m grows.

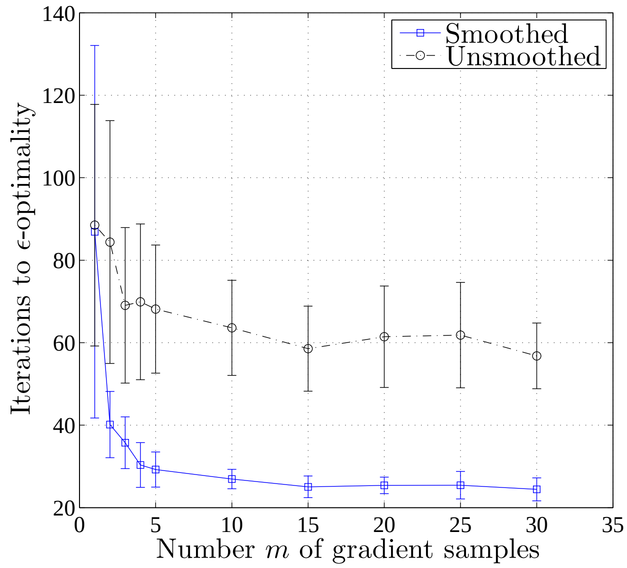

Figure 3. The number of iterations $T(\epsilon, m)$ to achieve an ϵ-optimal solution to (14) for simple mirror descent and the smoothed gradient method.

4 Proofs

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 Theorems 3 or 1. The second step is to realize that with the sampling procedure (6), the variance $\mathbb{E}\,\|e_t\|_*^2$ 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 (3), so we refer frequently to several lemmas stated in Appendix E.

4.1 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 $g_{i,t} \in \partial F(y_t + u_t Z_i; \xi_i)$, where the random variables Z_i are distributed uniformly on the ball $B_2(0, 1)$. From Lemma 8 in Appendix E, the variance of g_t as an estimate of $\nabla f_{\mu_t}(y_t)$ satisfies

$$\sigma^2 := \mathbb{E}[\|e_t\|_2^2] = \mathbb{E}[\|g_t - \nabla f_{\mu_t}(y_t)\|_2^2] \le \frac{L_0^2}{m}. \tag{15}$$

Further, for Z distributed uniformly on $B_2(0, 1)$, we have the bound

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

and moreover, the function $x \mapsto \mathbb{E}_\mu[f(x + uZ)]$ 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

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

where we have used the bound (15).

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

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

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

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

Substituting 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 E.

4.2 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 L_0/u-Lipschitz continuous gradient with respect to the ℓ_1-norm, and moreover it satisfies the upper bound $f_\mu(x) \le f(x) + \frac{L_0du}{2}$. 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 for any u the error satisfies

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

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 $\|g_i\|_\infty \le L_0$ and $\|\nabla f_\mu(x)\|_\infty \le L_0$, 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 Lemma 14 (see Appendix F) with $X = g - \nabla f_\mu(x)$ and $\sigma^2 = 4L_0^2/m$ yields the claim (17).

Now, as in the proof of Corollary 1, we can apply Theorem 1. Recall that $\frac{1}{2(p-1)}\|x\|_p^2$ is strongly convex over $\mathbb{R}^d$ with respect to the ℓ_p-norm for any $p \in (1,2]$ [25]. 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 $\|x^*\|_p^2\log d \le \|x^*\|_1^2\log d$. 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 $L \le L_0 e$, where L_0 is the Lipschitz constant with respect to the ℓ_1 norm, we apply Theorem 1 and the variance bound (17) to obtain 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 u, η_t, and $R \le \|x^*\|_1 \sqrt{\log d}$ and using bound (16) completes the proof.

4.3 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 does not appear to be well-known, so we provide a proof in Appendix A. In stating it, we recall the definition of the sigma field $\mathcal{F}_t$ from Assumption B.

Lemma 1. *In the notation of Theorem 2, suppose that $F(\cdot;\xi)$ is L_0-Lipschitz continuous with respect to the norm $\|\cdot\|$ over $\mathcal{X} + u_0 \operatorname{supp}\mu$ for P-a.e. ξ. Then*

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

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

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

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

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

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

4.4 Proof of Theorem 1

This proof is more involved than that of the above corollaries. In particular, we build on techniques used in the work of Tseng [34], Lan [19], and Xiao [36]. The changing smoothness of the stochastic objective—which comes from changing the shape parameter of the sampling distribution Z in the averaging step (6)—adds some challenge. Essentially, the idea of the proof is to let μ_t be the density of $u_t Z$ and define $f_{\mu_t}(x) := \mathbb{E}_\mu[f(x + u_t Z)]$, where u_t is the non-increasing sequence of shape parameters in the averaging scheme (6). We show via Jensen's inequality that $f(x) \le f_{\mu_t}(x) \le f_{\mu_{t-1}}(x)$ for all t, which is intuitive because the variance of the sampling scheme is decreasing. Then we apply a suitable modification of the accelerated gradient method [34] to the sequence of functions f_{μ_t} decreasing to f, and by allowing u_t to decrease appropriately we achieve our result. At the end of this section, we state a third result (Theorem 3), which gives an alternative setting for u given a priori knowledge of the number of iterations.

We begin by stating two technical lemmas:

Lemma 2. *Let f_{μ_t} be a sequence of functions such that f_{μ_t} has L_t-Lipschitz continuous gradients with respect to the norm $\|\cdot\|$ and assume that $f_{\mu_t}(x) \le f_{\mu_{t-1}}(x)$ for any $x \in \mathcal{X}$. Let the sequence $\{x_t, y_t, z_t\}$ be generated according to the updates (7a)–(7c), and define the error term $e_t = \nabla f_{\mu_t}(y_t) - g_t$. Then for any $x^* \in \mathcal{X}$,*

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

See Appendix B 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 [34]; the first follows by a straightforward induction.

We now proceed with the proof. Recalling $f_{\mu_t}(x) = \mathbb{E}[f(x + u_t Z)]$, let us verify that $f_{\mu_t}(x) \le f_{\mu_{t-1}}(x)$ for any x and t so we can apply Lemma 2. Since $u_t \le u_{t-1}$, we may define a random variable $U \in \{0, 1\}$ such that $\mathbb{P}(U = 1) = \frac{u_t}{u_{t-1}} \in [0, 1]$. Then

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

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

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

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

which by the definition of u_t is in turn bounded by

$$\frac{1}{\theta_{T-1}^2}[f_{\mu_{t-1}}(x_T) + \varphi(x_T)] - \sum_{t=0}^{T-1} \frac{1}{\theta_t}[f_{\mu_t}(x^*) + \varphi(x^*)] + T L_0 u. \tag{18}$$

Now we simply apply Lemma 2 to the bound (18), which gives us

$$\frac{1}{\theta_{T-1}^2}\left[f(x_T)+\varphi(x_T)-f(x^*)-\varphi(x^*)\right]$$
$$\leq L_T\psi(x^*)+\frac{\eta_T}{\theta_T}\psi(x^*)+\sum_{t=0}^{T-1}\frac{1}{2\theta_t\eta_t}\|e_t\|_*^2+\sum_{t=0}^{T-1}\frac{1}{\theta_t}\langle e_t, z_t-x^*\rangle+TL_0u. \tag{19}$$

The non-probabilistic bound (19) 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 (19).

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 $\nabla f_{\mu_t}(y_t)$ by construction, we have $\mathbb{E}[g_t \mid \mathcal{F}_{t-1}] = \nabla f_{\mu_t}(y_t)$ and

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

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

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

Furthermore, we have $\theta_{t+1} \leq \theta_t$, so by multiplying both sides of our bound (19) by θ_{T-1}^2 and taking expectations over the random vectors g_t,

$$\begin{aligned}
&\mathbb{E}[f(x_T)+\varphi(x_T)]-[f(x^*)+\varphi(x^*)]\\
&\leq \theta_{T-1}^2 L_T\psi(x^*)+\theta_{T-1}\eta_T\psi(x^*)+\theta_{T-1}\sum_{t=0}^{T-1}\frac{1}{2\eta_t}\mathbb{E}\,\|e_t\|_*^2+\theta_{T-1}\sum_{t=0}^{T-1}\mathbb{E}[\langle e_t, z_t-x^*\rangle]+\theta_{T-1}^2TL_0u\\
&\leq \frac{6L_1\psi(x^*)}{Tu}+\frac{2\eta_T\psi(x^*)}{T}+\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{\eta_t}\mathbb{E}\,\|e_t\|_*^2+\frac{4L_0u}{T},
\end{aligned}$$

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

As promised, we conclude with a theorem using a fixed setting of the smoothing parameter u_t.

Theorem 3. *Suppose that $u_t \equiv u$ for all t and set $L_t \equiv L_1/u$. With the remaining conditions as in Theorem 1, then for any $x^* \in \mathcal{X}$, we have*

$$\mathbb{E}[f(x_T)+\varphi(x_T)]-[f(x^*)+\varphi(x^*)] \leq \frac{4L_1\psi(x^*)}{T^2u}+\frac{2\eta_T\psi(x^*)}{T}+\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{\eta_t}\mathbb{E}\big[\,\|e_t\|_*^2\,\big]+L_0u,$$

where $e_t := \nabla f_\mu(y_t) - g_t$.

Proof The proof is brief. If we fix $u_t \equiv u$ for all t, then the bound (19) holds with the last term TL_0u replaced by $\theta_{T-1}^2 L_0u$, which we see by invoking Lemma 3. The remainder of the proof follows unchanged, with $L_t \equiv L_1$ for all t. □

It is clear that by setting $u \propto 1/T$, the rates achieved by Theorem 1 and Theorem 3 are identical to constant factors.

4.5 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*

$$\mathbb{P}\left[\theta_{T-1}^2 \sum_{t=0}^{T-1} \frac{1}{\theta_t} \langle e_t, z_t - x^* \rangle \ge \epsilon\right] \le \exp\left(-\frac{T\epsilon^2}{R^2\sigma^2}\right). \tag{20}$$

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

$$\theta_{T-1}^2 \sum_{t=0}^{T-1} \frac{1}{\theta_t} \langle e_t, z_t - x^* \rangle \le R\sigma \sqrt{\frac{\log \frac{1}{\delta}}{T}}. \tag{21}$$

Lemma 5. *In the notation of Theorem 2 and under Assumption B, we have*

$$\log \mathbb{P}\left[\sum_{t=0}^{T-1} \frac{1}{2\eta_t} \|e_t\|_*^2 \ge \sum_{t=0}^{T-1} \frac{1}{2\eta_t} \mathbb{E}[\|e_t\|_*^2] + \epsilon\right] \le \max\left\{-\frac{\epsilon^2}{32e\sigma^4 \sum_{t=0}^{T-1} \frac{1}{\eta_t^2}}, -\frac{\eta_0}{4\sigma^2}\epsilon\right\}. \tag{22}$$

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

$$\sum_{t=0}^{T-1} \frac{1}{2\eta_t} \|e_t\|_*^2 \le \sum_{t=0}^{T-1} \frac{1}{2\eta_t} \mathbb{E}[\|e_t\|_*^2] + \frac{4\sigma^2}{\eta} \max\left\{\log \frac{1}{\delta}, \sqrt{2e(\log T + 1) \log \frac{1}{\delta}}\right\}. \tag{23}$$

See Appendices C and D, respectively, for the proofs of these two lemmas.

Equipped with these lemmas, we now prove Theorem 2. Let us recall the deterministic bound (19) from the proof of Theorem 1:

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

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

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

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

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

The terms remaining to control are deterministic, and were bounded previously in the proof of Theorem 1; in particular, we have

$$\theta_{T-1}^2 L_T \le \frac{6L_1}{Tu}, \quad \frac{\theta_{T-1}^2 \eta_T}{\theta_T} \le \frac{4\eta_T}{T+1}, \quad \text{and} \quad \theta_{T-1}^2 T L_0 u \le \frac{4L_0 u}{T+1}.$$

Combining the above bounds completes the proof.

5 Discussion

In this paper, we have analyzed smoothing strategies for stochastic non-smooth optimization. We have developed methods that are provably optimal in the stochastic oracle model of optimization complexity, and given—to our knowledge—the first variance reduction techniques for non-smooth stochastic optimization. We think that at least two obvious questions remain. The first, to which we have alluded earlier, is whether the randomized smoothing is necessary to achieve such optimal rates of convergence. The second question is whether dimension-independent smoothing techniques are possible, that is, whether the d-dependent factors in the bounds in Corollaries 1–4 are necessary. Answering this question would require study of different smoothing distributions, as the dimension dependence for our choices of μ is tight. We have outlined several applications for which smoothing techniques give provable improvement over standard methods. Our experiments also show qualitatively good agreement with the theoretical predictions we have developed.

Acknowledgments

We thank Alekh Agarwal and Ohad Shamir for discussions about variance reduction in stochastic optimization and Steven Evans for some useful pointers to smoothing operators. We also acknowledge the anonymous reviewers for their helpful feedback and thoughtful comments. JCD was supported by a fellowship from the National Defense Science and Engineering Graduate Fellowship Program (NDSEG). PLB gratefully acknowledges the support of the NSF under award DMS-0830410. In addition, MJW and JCD were partially supported by grant number AFOSR-09NL184.

A Proof of Lemma 1

The proof of this lemma requires several auxiliary results on sub-Gaussian and sub-exponential random variables, which we collect and prove in Appendix F.

Define $X_i = \nabla f_\mu(x_t) - g_{i,t}$ and $S_m = \sum_{i=1}^m X_i$, so $\frac{1}{m} S_m = \nabla f_\mu(x_t) - \frac{1}{m}\sum_{i=1}^m g_{i,t}$. Note that conditioned on $\mathcal{F}_{t-1}$, the X_i are independent, so that for $L = 2L_0$, we have $\|X_i\|_* \le L$, and we can apply Lemma 16 from Appendix F. In particular, we see that $\left\|\frac{1}{m}S_m\right\|_* - \mathbb{E}\left\|\frac{1}{m}S_m\right\|_*$ is sub-Gaussian with parameter at most $4L^2/m$. Consequently, we can apply Lemma 13 from Appendix F so as to obtain

$$\mathbb{E}\exp\left(\frac{sm\left\|\frac{1}{m}S_m\right\|_*^2}{8L^2}\right) \le \frac{1}{\sqrt{1-s}}\exp\left(\frac{m(\mathbb{E}\left\|\frac{1}{m}S_m\right\|_*)^2}{8L^2}\,\frac{s}{1-s}\right).$$

Moreover, we can weaken the sub-Gaussian parameter $4L^2/m$ with $\max\{\mathbb{E}\left\|\frac{1}{m}S_m\right\|_*^2, 4L^2/m\}$:

$$\mathbb{E}\left[\exp(\lambda(\|S_m/m\|_* - \mathbb{E}\|S_m/m\|_*))\right] \le \exp\left(\frac{\lambda^2 \max\{4L^2/m, \mathbb{E}\left\|\frac{1}{m}S_m\right\|_*^2\}}{2}\right).$$

Recalling that for any random variable X, Jensen's inequality gives $(\mathbb{E}X)^2 \le \mathbb{E}X^2$, we have

$$\begin{aligned}\mathbb{E}\exp\left(\frac{s\left\|\frac{1}{m}S_m\right\|_*^2}{2\max\{\mathbb{E}\left\|\frac{1}{m}S_m\right\|_*^2,\frac{4}{m}L^2\}}\right) &\le \frac{1}{\sqrt{1-s}}\exp\left(\frac{\mathbb{E}\left\|\frac{1}{m}S_m\right\|_*^2}{2\max\{\mathbb{E}\left\|\frac{1}{m}S_m\right\|_*^2,\frac{4}{m}L^2\}}\ \frac{s}{1-s}\right)\\ &\le \frac{1}{\sqrt{1-s}}\exp\left(\frac{1}{2}\cdot\frac{s}{1-s}\right).\end{aligned}$$

By taking $s=\frac{1}{2}$, we get

$$\frac{1}{\sqrt{1-s}}\exp\left(\frac{1}{2}\ \frac{s}{1-s}\right) = \sqrt{2}\exp\left(\frac{1}{2}\right) \le \exp(1).$$

Replacing L with $2L_0$ completes the proof.

B Proof of Lemma 2

Define the linearized version of the cumulative objective

$$\ell_t(z) := \sum_{\tau=0}^{t}\frac{1}{\theta_\tau}[f_{\mu_\tau}(y_\tau)+\langle g_\tau, z-y_\tau\rangle+\varphi(z)], \tag{24}$$

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

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

By the smoothness of f_{μ_t}, we have

$$\underbrace{f_{\mu_t}(x_{t+1})+\varphi(x_{t+1})}_{\phi_t(x_{t+1})} \le f_{\mu_t}(y_t)+\langle\nabla f_{\mu_t}(y_t), x_{t+1}-y_t\rangle+\frac{L_t}{2}\left\|x_{t+1}-y_t\right\|^2+\varphi(x_{t+1}).$$

From the definition (7a)–(7c) of the triple (x_t, y_t, z_t), we obtain

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

Finally, by convexity of the regularizer φ, we conclude

$$\begin{aligned}\phi_t(x_{t+1}) \le \theta_t&\left[f_{\mu_t}(y_t)+\langle\nabla f_{\mu_t}(y_t), z_{t+1}-y_t\rangle+\varphi(z_{t+1})+\frac{L_t\theta_t}{2}\left\|z_{t+1}-z_t\right\|^2\right]\\ &+(1-\theta_t)[f_{\mu_t}(y_t)+\langle\nabla f_{\mu_t}(y_t), x_t-y_t\rangle+\varphi(x_t)].\end{aligned} \tag{25}$$

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

$$D_\psi(x,y) \ = \ \psi(x)-\psi(y)-\langle\nabla\psi(y), x-y\rangle \ \ge \ \frac{1}{2}\left\|x-y\right\|^2. \tag{26}$$

On the other hand, by the convexity of f_{μ_t}, we have

$$f_{\mu_t}(y_t)+\langle\nabla f_{\mu_t}(y_t), x_t-y_t\rangle \le f_{\mu_t}(x_t). \tag{27}$$

Substituting inequalities (26) and (27) into the upper bound (25) and simplifying yields

$$\phi_t(x_{t+1}) \le \theta_t \left[f_{\mu_t}(y_t) + \langle \nabla f_{\mu_t}(y_t), z_{t+1} - y_t\rangle + \varphi(z_{t+1}) + L_t\theta_t D_\psi(z_{t+1}, z_t)\right] + (1-\theta_t)[f_{\mu_t}(x_t) + \varphi(x_t)].$$

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

$$\begin{aligned}
&\phi_t(x_{t+1}) \\
&\le \theta_t \left[f_{\mu_t}(y_t) + \langle g_t, z_{t+1} - y_t\rangle + \varphi(z_{t+1}) + L_t\theta_t D_\psi(z_{t+1}, z_t)\right] \\
&\qquad + (1-\theta_t)[f_{\mu_t}(x_t) + \varphi(x_t)] + \theta_t \langle e_t, z_{t+1} - y_t\rangle \\
&= \theta_t^2 \left[\ell_t(z_{t+1}) - \ell_{t-1}(z_{t+1}) + L_t D_\psi(z_{t+1}, z_t)\right] + (1-\theta_t)[f_{\mu_t}(x_t) + \varphi(x_t)] + \theta_t \langle e_t, z_{t+1} - y_t\rangle . \quad (28)
\end{aligned}$$

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 $\left\langle g + \frac{1}{\alpha_t}\nabla\psi(z_t), x - z_t\right\rangle \ge 0$. Thus, first-order convexity gives

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

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

$$\ell_t(z_{t+1}) - \ell_{t-1}(z_{t+1}) \le \ell_t(z_{t+1}) - \ell_{t-1}(z_t) - \frac{1}{\alpha_t}\psi(z_t) + \frac{1}{\alpha_t}\psi(z_{t+1}) - \frac{1}{\alpha_t}D_\psi(z_{t+1}, z_t).$$

Combining this inequality with the bound (28) and using the definition $\alpha_t^{-1} = L_t + \eta_t/\theta_t$, we find

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

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 $f_{\mu_t} \le f_{\mu_{t-1}}$ by construction, so we obtain

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

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 $f_{\mu_t} \le f_{\mu_{t-1}}$. By applying the three steps above successively to $[f_{\mu_{t-1}}(x_t) + \varphi(x_t)]/\theta_{t-1}^2$, then to $[f_{\mu_{t-2}}(x_{t-1}) + \varphi(x_{t-1})]/\theta_{t-2}^2$, and so on until $t = 0$, we find

$$\frac{1}{\theta_t^2}[f_{\mu_t}(x_{t+1}) + \varphi(x_{t+1})] \le \frac{1-\theta_0}{\theta_0^2}[f_{\mu_0}(x_0) + \varphi(x_0)] + \ell_t(z_{t+1}) + \frac{1}{\alpha_{t+1}}\psi(z_{t+1}) - \frac{1}{\alpha_0}\psi(z_0) \\ - \sum_{\tau=0}^{t} \frac{\eta_\tau}{\theta_\tau} D_\psi(z_{\tau+1}, z_\tau) + \sum_{\tau=0}^{t} \frac{1}{\theta_\tau} \langle e_\tau, z_{\tau+1} - y_\tau \rangle - \ell_{-1}(z_0).$$

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

$$\frac{1}{\theta_t^2}[f_{\mu_t}(x_{t+1}) + \varphi(x_{t+1})] \le \ell_t(x^*) + \frac{1}{\alpha_{t+1}}\psi(x^*) - \sum_{\tau=0}^{t} \frac{\eta_\tau}{\theta_\tau} D_\psi(z_{\tau+1}, z_\tau) + \sum_{\tau=0}^{t} \frac{1}{\theta_\tau} \langle e_\tau, z_{\tau+1} - y_\tau \rangle .$$

Recalling the definition (24) of ℓ_t and noting that $f_{\mu_t}(y_t) + \langle \nabla f_{\mu_t}(y_t), x - y_t \rangle \le f_{\mu_t}(x)$ by convexity, we expand ℓ_t and have

$$\begin{aligned}
&\frac{1}{\theta_t^2}[f_{\mu_t}(x_{t+1}) + \varphi(x_{t+1})] \\
&\le \sum_{\tau=0}^{t} \frac{1}{\theta_\tau}[f_{\mu_\tau}(y_\tau) + \langle g_\tau, x^* - y_\tau \rangle + \varphi(x^*)] + \frac{1}{\alpha_{t+1}}\psi(x^*) - \sum_{\tau=0}^{t} \frac{\eta_\tau}{\theta_\tau} D_\psi(z_{\tau+1}, z_\tau) + \sum_{\tau=0}^{t} \frac{1}{\theta_\tau} \langle e_\tau, z_{\tau+1} - y_t \rangle \\
&= \sum_{\tau=0}^{t} \frac{1}{\theta_\tau}[f_{\mu_\tau}(y_\tau) + \langle \nabla f_{\mu_\tau}(y_\tau), x^* - y_\tau \rangle + \varphi(x^*)] + \frac{1}{\alpha_{t+1}}\psi(x^*) - \sum_{\tau=0}^{t} \frac{\eta_\tau}{\theta_\tau} D_\psi(z_{\tau+1}, z_\tau) + \sum_{\tau=0}^{t} \frac{1}{\theta_\tau} \langle e_\tau, z_{\tau+1} - x^* \rangle \\
&\le \sum_{\tau=0}^{t} \frac{1}{\theta_\tau}[f_{\mu_\tau}(x^*) + \varphi(x^*)] + \frac{1}{\alpha_{t+1}}\psi(x^*) - \sum_{\tau=0}^{t} \frac{\eta_\tau}{\theta_\tau} D_\psi(z_{\tau+1}, z_\tau) + \sum_{\tau=0}^{t} \frac{1}{\theta_\tau} \langle e_\tau, z_{\tau+1} - x^* \rangle . \qquad (29)
\end{aligned}$$

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

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

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 (29) gives the statement of the lemma.

C Proof of Lemma 4

Consider the sequence $\sum_{t=0}^{T-1} \frac{1}{\theta_t}\langle e_t, z_t - x^* \rangle$. Since $\mathcal{X}$ is compact and $\|z_t - x^*\| \le R$, we have $\langle e_t, z_t - x^* \rangle \le \|e_t\|_* R$. Further, $\mathbb{E}[\langle e_t, z_t - x^* \rangle \mid \mathcal{F}_{t-1}] = 0$, so that $\frac{1}{\theta_t}\langle e_t, z_t - x^* \rangle$ is a martingale difference sequence. Further, by setting $c_t = R\sigma/\theta_t$, we have

$$\mathbb{E}\left[\exp\left(\frac{\langle e_t, z_t - x^* \rangle^2}{c_t^2\theta_t^2}\right) \mid \mathcal{F}_{t-1}\right] \le \mathbb{E}\left[\exp\left(\frac{\|e_t\|_*^2 R^2}{c_t^2\theta_t^2}\right) \mid \mathcal{F}_{t-1}\right] = \mathbb{E}\left[\exp\left(\frac{\|e_t\|_*^2}{\sigma^2}\right) \mid \mathcal{F}_{t-1}\right] \le \exp(1)$$

by Assumption B. By applying Lemma 18 from Appendix F, we conclude that $\frac{1}{\theta_t}\langle e_t, z_t - x^*\rangle$ is (conditionally) sub-Gaussian with parameter $\sigma_t^2 \le 4R^2\sigma^2/3\theta_t^2$. Applying the Azuma-Hoeffding inequality (see Eq. (41), Appendix F) yields

$$\mathbb{P}\Big[\sum_{t=0}^{T-1}\frac{1}{\theta_t}\langle e_t, z_t - x^*\rangle \ge w\Big] \le \exp\Big(-\frac{3w^2}{8R^2\sigma^2\sum_{t=0}^{T-1}\frac{1}{\theta_t^2}}\Big).$$

Setting $w = \epsilon/\theta_{T-1}$ yields that

$$\mathbb{P}\Big[\theta_{T-1}\sum_{t=0}^{T-1}\frac{1}{\theta_t}\langle e_t, z_t - x^*\rangle \ge \epsilon\Big] \le \exp\Big(-\frac{3\epsilon^2}{8R^2\sigma^2\sum_{t=0}^{T-1}\frac{\theta_{T-1}^2}{\theta_t^2}}\Big).$$

Noting that $\theta_{T-1} \le \theta_t$ for any $t < T$, we have $R^2\sigma^2\sum_{t=0}^{T-1}\frac{\theta_{T-1}^2}{\theta_t^2} \le R^2\sigma^2\sum_{t=0}^{T-1}1 = R^2\sigma^2 T$, dividing ϵ again by θ_{T-1}, and recalling that $\theta_{T-1} \le \frac{2}{T+1}$, we have

$$\mathbb{P}\left[\theta_{T-1}^2\sum_{t=0}^{T-1}\frac{1}{\theta_t}\langle e_t, z_t - x^*\rangle \ge \epsilon\right] \le \exp\left(-\frac{12(T+1)\epsilon^2}{8R^2\sigma^2}\right) \le \exp\left(-\frac{3T\epsilon^2}{2R^2\sigma^2}\right),$$

as claimed (20). The second claim (21) follows by setting $\delta = \exp(-\frac{3T\epsilon^2}{2R^2\sigma^2})$, and then solving to obtain $\epsilon^2 = \frac{2R^2\sigma^2}{3T}\log\frac{1}{\delta}$.

D Proof of Lemma 5

Again, recall the σ-fields $\mathcal{F}_t$ defined prior to Assumption B. Define the random variables

$$X_t := \frac{1}{2\eta_t}\|e_t\|_*^2 - \frac{1}{2\eta_t}\mathbb{E}[\|e_t\|_*^2 \mid \mathcal{F}_{t-1}].$$

As an intermediate step, we claim that for $\lambda \le \eta_t/2\sigma^2$, the following bound holds:

$$\mathbb{E}[\exp(\lambda X_t) \mid \mathcal{F}_{t-1}] = \mathbb{E}\left[\exp\left(\frac{\lambda}{2\eta_t}(\|e_t\|_*^2 - \mathbb{E}[\|e_t\|_*^2 \mid \mathcal{F}_{t-1}])\right) \mid \mathcal{F}_{t-1}\right] \le \exp\left(\frac{8e}{\eta_t^2}\lambda^2\sigma^4\right). \qquad (30)$$

For now, we proceed with the proof, returning to establish this intermediate claim later.

The bound (30) implies that X_t is sub-exponential with parameters $\Lambda_t = \eta_t/2\sigma^2$ and $\tau_t^2 \le 16e\sigma^4/\eta_t^2$. Since $\eta_t = \eta\sqrt{t+1}$, it is clear that $\min_t\{\Lambda_t\} = \Lambda_0 = \eta_0/2\sigma^2$. By defining $C^2 = \sum_{t=0}^{T-1}\tau_t^2$, we can apply Theorem I.5.1 from the book [7, pg. 26] to conclude that

$$\mathbb{P}\left(\sum_{t=0}^{T-1}X_t \ge \epsilon\right) \le \begin{cases}\exp\left(-\frac{\epsilon^2}{2C^2}\right) & \text{for } 0 \le \epsilon \le \Lambda_0 C^2\\ \exp\left(-\frac{\Lambda_0\epsilon}{2}\right) & \text{otherwise, i.e. } \epsilon > \Lambda_0 C^2,\end{cases} \qquad (31)$$

which yields the first claim in Lemma 5.

The second statement involves inverting the bound for the different regimes of ϵ. Before proving the bound, we note that for $\epsilon = \Lambda_0 C^2$, we have $\exp(-\epsilon^2/2C^2) = \exp(-\Lambda\epsilon/2)$, so we can invert each of the exp terms to solve for ϵ and take the maximum of the bounds. We begin with ϵ in the regime closest to zero, recalling that $\eta_t = \eta\sqrt{t+1}$. We see that

$$C^2 \le \frac{16e\sigma^4}{\eta^2} \sum_{t=0}^{T-1} \frac{1}{t+1} \le \frac{16e\sigma^4}{\eta^2} \log(T+1).$$

Thus, inverting the bound $\delta = \exp(-\epsilon^2/2C^2)$, we get $\epsilon = \sqrt{2C^2 \log \frac{1}{\delta}}$, or that

$$\sum_{t=0}^{T-1} \frac{1}{2\eta_t} \|e_t\|_*^2 \le \sum_{t=0}^{T-1} \frac{1}{2\eta_t} \mathbb{E}[\|e_t\|_*]^2 + 4\sqrt{2e}\frac{\sigma^2}{\eta}\sqrt{\log\frac{1}{\delta}\log(T+1)}$$

with probability at least $1-\delta$. In the large ϵ regime, we solve $\delta = \exp(-\eta\epsilon/4\sigma^2)$ or $\epsilon = \frac{4\sigma^2}{\eta}\log\frac{1}{\delta}$, which gives that

$$\sum_{t=0}^{T-1} \frac{1}{2\eta_t} \|e_t\|_*^2 \le \sum_{t=0}^{T-1} \frac{1}{2\eta_t} \mathbb{E}\|e_t\|_*^2 + \frac{4\sigma^2}{\eta}\log\frac{1}{\delta}$$

with probability at least $1-\delta$, by the bound (31).

We now return to prove the intermediate claim (30). Let $X := \|e_t\|_*$. By assumption, we have $\mathbb{E}\exp(X^2/\sigma^2) \le \exp(1)$, so for $\lambda \in [0,1]$ we see

$$\mathbb{P}(X^2/\sigma^2 > \epsilon) \le \mathbb{E}[\exp(\lambda X^2/\sigma^2)]\exp(-\lambda\epsilon) \le \exp(\lambda - \lambda\epsilon)$$

and replacing ϵ with $1+\epsilon$ we have $\mathbb{P}(X^2 > \sigma^2 + \epsilon\sigma^2) \le \exp(-\epsilon)$. If $\epsilon\sigma^2 \ge \sigma^2 - \mathbb{E}X^2$, then $\sigma^2 - \mathbb{E}X^2 + \epsilon\sigma^2 \le 2\epsilon\sigma^2$ so

$$\mathbb{P}(X^2 > \mathbb{E}X^2 + 2\epsilon\sigma^2) \le \mathbb{P}(X^2 > \sigma^2 + \epsilon\sigma^2) \le \exp(-\epsilon),$$

while for $\epsilon\sigma^2 < \sigma^2 - \mathbb{E}X^2$, we clearly have $\mathbb{P}(X^2 - \mathbb{E}X^2 > \epsilon\sigma^2) \le 1 \le \exp(1)\exp(-\epsilon)$ since $\epsilon \le 1$. In either case, we have

$$\mathbb{P}(X^2 - \mathbb{E}X^2 > \epsilon) \le \exp(1)\exp\left(-\frac{\epsilon}{2\sigma^2}\right).$$

For the opposite concentration inequality, we see that

$$\mathbb{P}((\mathbb{E}X^2 - X^2)/\sigma^2 > \epsilon) \le \mathbb{E}[\exp(\lambda\mathbb{E}X^2/\sigma^2)\exp(-\lambda X^2/\sigma^2)]\exp(-\lambda\epsilon) \le \exp(\lambda - \lambda\epsilon)$$

or $\mathbb{P}(X^2 - \mathbb{E}X^2 < -\sigma^2\epsilon) \le \exp(1)\exp(-\epsilon)$. Using the union bound, we have

$$\mathbb{P}(|X^2 - \mathbb{E}X^2| > \epsilon) \le 2\exp(1)\exp\left(-\frac{\epsilon}{2\sigma^2}\right). \tag{32}$$

Now we apply Lemma 17 to the bound (32) to see that $\|e_t\|_*^2 - \mathbb{E}[\|e_t\|_*^2] = X^2 - \mathbb{E}[X^2]$ is sub-exponential with parameters $\Lambda \ge \sigma^2$ and $\tau^2 \le 32e\sigma^4$.

E Properties of randomized smoothing

In this section, we discuss the analytic properties of the smoothed function f_μ from the convolution (3). 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 [4, 31]). By Fubini's theorem, we have

$$f_\mu(x) = \int_{\mathbb{R}^d}\int_{\Xi} F(x+y;\xi)dP(\xi)\mu(y)dy = \int_{\Xi}\int_{\mathbb{R}^d} F(x+y;\xi)\mu(y)dydP(\xi) = \int_{\Xi} F_\mu(x;\xi)dP(\xi).$$

Here $F_\mu(x;\xi) = (F(\cdot;\xi) * \mu(-\cdot))(x)$. We begin with the observation that since μ is a density with respect to Lebesgue measure, the function f_μ is in fact differentiable [4]. 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.

E.1 Statements of smoothing lemmas

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. We prove the following set of smoothness lemmas at the end of this section.

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 exists a function f 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*

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

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. give (i)–(iii) of the following lemma [38].

Lemma 8 (Yousefian, Nedić, Shanbhag)**.** *Let f_μ be defined as in (3) 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) is tight—cannot be improved by more than a constant factor—simultaneously.

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. The following lemma is similar to a result of Lakshmanan and de Farias [17, Lemma 3.3], but they consider functions Lipschitz-continuous with respect to the ℓ_∞-norm, i.e. $|f(x) - f(y)| \le L\|x-y\|_\infty$, which is too stringent for our purposes, and we carefully quantify the dependence on the dimension of the underlying problem.

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.

Our final lemma illustrates the sharpness of the bounds we have proved for functions that are Lipschitz with respect to the ℓ_2-norm. Specifically, we show that at least for the normal and uniform distributions, it is impossible to get more favorable tradeoffs between the uniform approximation error of the smoothed function f_μ and the Lipschitz continuity of ∇f_μ. We begin with the following definition of our two types of error, then give the lemma:

$$E_U(f) := \inf \left\{ L \in \mathbb{R} \mid \sup_{x \in \mathcal{X}} |f(x) - f_\mu(x)| \le L \right\} \tag{33}$$

$$E_\nabla(f) := \inf \left\{ L \in \mathbb{R} \mid \|\nabla f_\mu(x) - \nabla f_\mu(y)\|_2 \le L \|x - y\|_2 \ \forall\, x, y \in \mathcal{X} \right\} \tag{34}$$

Lemma 10. *For μ equal to either the uniform distribution on $B_2(0, u)$ or $N(0, u^2 I_{d\times d})$, there exists an L_0-Lipschitz continuous function f and dimension independent constant $c > 0$ such that*

$$E_U(f) E_\nabla(f) \ge c L_0^2 \sqrt{d}.$$

Remark The importance of the above bound is made clear by inspecting the convergence guarantee of Theorem 1. The terms L_1 and L_0 in the bound (8) can be replaced with $E_\nabla(f)$ and $E_U(f)$, respectively. Minimizing over u, we see that the leading term in the convergence guarantee (8) is of order $\frac{\sqrt{E_\nabla(f) E_U(f) \psi(x^*)}}{T} \ge \frac{c L_0 d^{1/4} \sqrt{\psi(x^*)}}{T}$. In particular, this result shows that our analysis of the dimension dependence of the randomized smoothing in Lemmas 8 and 9 is sharp and cannot be improved by more than a constant factor (see also Corollaries 1 and 2).

E.2 Proof of smoothing lemmas

The following technical lemma is a building block for our results; we provide a proof in Sec. E.2.5.

Lemma 11. *Let f be convex and L_0-Lipschitz continuous with respect to a norm $\|\cdot\|$ over the domain $\operatorname{supp}\mu + \mathcal{X}$. Let Z be distributed according to the distribution μ. Then*

$$\|\nabla f_\mu(x) - \nabla f_\mu(y)\|_* = \|\mathbb{E}[\nabla f(x + Z) + \nabla f(y + Z)]\|_* \le L_0 \int |\mu(z - x) - \mu(z - y))| dz. \tag{35}$$

If the norm $\|\cdot\|$ is the ℓ_2-norm and the density $\mu(z)$ is rotationally symmetric and non-increasing as a function of $\|z\|_2$, the bound (35) is tight; specifically, it is attained by the function

$$f(x) = L_0 \left| \left\langle \frac{y}{\|y\|_2}, x \right\rangle - \frac{1}{2} \right|.$$

E.2.1 Proof of Lemma 6

Throughout, we let $Z \sim \mu$, where μ is the uniform density on $B_\infty(0, u)$), and $h_u(x)$ denote the (shifted) Huber loss

$$h_u(x) = \begin{cases} \frac{x^2}{2u} + \frac{u}{2} & \text{for } x \in [-u, u] \\ |x| & \text{otherwise.} \end{cases} \tag{36}$$

Now we prove each of the parts of the lemma in turn.

(i) Since $\mathbb{E}[Z] = 0$, Jensen's inequality shows $f(x) = f(x + \mathbb{E}[Z]) \le \mathbb{E}[f(x + Z)] = f_\mu(x)$, by definition of f_μ. Now recall the definition of $\|\partial f(x)\| = \sup\{\|g\| \mid g \in \partial f(x)\}$ from the introduction. To get the upper uniform bound, note first that by assumption, f is L_0-Lipschitz continuous over $\mathcal{X} + B_\infty(0, u)$, since by assumption

$$\|\partial f(x)\|_\infty \le \mathbb{E}[\|\partial F(x;\xi)\|_\infty] \le \sqrt{\mathbb{E}[\|\partial F(x;\xi)\|_\infty^2]} \le L_0,$$

again using Jensen's inequality. Thus f is L_0-Lipschitz with respect to the ℓ_1-norm,

$$f_\mu(x) = \mathbb{E}[f(x + Z)] \le \mathbb{E}[f(x)] + L_0\mathbb{E}[\|Z\|_1] = f(x) + \frac{dL_0u}{2}.$$

To see that the estimate is tight, note that for $f(x) = \|x\|_1$, we have $f_\mu(x) = \sum_{i=1}^d h_u(x_i)$, where h_u is the shifted Huber loss (36), and $f_\mu(0) = du/2$, while $f(0) = 0$.

(ii) We now prove that f_μ is L_0-Lipschitz with respect to $\|\cdot\|_1$. Under the stated conditions, we have $\partial f(x) = \mathbb{E}[\partial F(x;\xi)]$, which shows that $\|\partial f(x)\|_\infty^2 \le \mathbb{E}[\|\partial F(x;\xi)\|_\infty^2] \le L_0^2$. Thus, we obtain the upper bound

$$\|\nabla f_\mu(x)\|_\infty = \|\mathbb{E}[\nabla f(x + Z)]\|_\infty \le \mathbb{E}[\|\nabla f(x + Z)\|_\infty] \le L_0.$$

Tightness follows again by considering $f(x) = \|x\|_1$, where $L_0 = 1$.

(iii) Recall that differentiability is directly implied by earlier work of Bertsekas [4]. Since f is a.e.-differentiable, we have $\nabla f_\mu(x) = \mathbb{E}[\nabla f(x + Z)]$ for Z uniform on $[-u, u]^d$. We now establish Lipschitz continuity of $\nabla f_\mu(x)$.

For a fixed pair $x, y \in \mathcal{X} + B_\infty(0, u)$, we have from Lemma 11

$$\|\mathbb{E}[\nabla f(x + Z)] - \mathbb{E}[\nabla f(y + Z)]\|_\infty \le L_0 \cdot \frac{1}{(2u)^d}\lambda\big(B_\infty(x, u)\Delta B_\infty(y, u)\big),$$

where λ denotes Lebesgue measure and Δ denotes the symmetric set-difference. By a straightforward geometric calculation, we see that

$$\lambda\left(B_\infty(x, u)\Delta B_\infty(y, u)\right) = 2\left((2u)^d - \prod_{i=1}^d \left[2u - |x_i - y_i|\right]_+\right). \tag{37}$$

To control the volume term (37) and complete the proof, we need an auxiliary lemma (which we prove at the end of this subsection).

Lemma 12. *Let $a \in \mathbb{R}_+^d$ and $u \in \mathbb{R}_+$. Then $\prod_{i=1}^d [u - a_i]_+ \ge u^d - \|a\|_1 u^{d-1}$.*

The volume (37) is easy to control using Lemma 12. Indeed, we have

$$\frac{1}{2}\lambda\left(B_\infty(x, u)\Delta B_\infty(y, u)\right) \le (2u)^d - (2u)^d + \|x - y\|_1 (2u)^{d-1},$$

which implies the desired result, that is, that

$$\|\mathbb{E}[\nabla f(x + Z)] - \mathbb{E}\nabla[f(y + Z)]\|_\infty \le \frac{L_0\|x - y\|_1}{u}.$$

To see the tightness claimed in the proposition, consider as usual $f(x) = \|x\|_1$ and let e_i denote the ith standard basis vector. Then $L_0 = 1$, $\nabla f_\mu(0) = 0$, $\nabla f_\mu(ue_i) = e_i$, and $\|\nabla f_\mu(0) - \nabla f_\mu(ue_i)\|_\infty = 1 = \frac{L_0}{u}\|0 - ue_i\|_1$.

(iv) The equality $\mathbb{E}[\nabla F(x+Z;\xi)] = \nabla f_\mu(x)$ follows from Fubini's theorem. The second statement is simply a consequence of the triangle inequality. Finally, the tightness follows from the following one-dimensional example. Let $f(x) = L_0|x|$ for $x \in \mathbb{R}$ and $L_0 > 0$. Then $f_\mu(x)$ is L_0 times the Huber loss $h_u(x)$ defined earlier, and $f_\mu'(0) = 0$. Thus for Z uniform on $[-u, u]$,

$$\mathbb{E}(f_\mu'(0) - f'(Z))^2 = \mathbb{E}[L_0^2 \operatorname{sign}(Z)^2] = L_0^2,$$

which is the Lipschitz constant of f.

Proof of Lemma 12 We begin by noting that the statement of the lemma trivially holds whenever $\|a\|_1 \geq u$, as the right hand side of the inequality is then non-positive. Now, fix some $c < u$, and consider the problem

$$\min_a \prod_{i=1}^d (u - a_i)_+ \quad \text{s.t.} \quad a \succeq 0, \|a\|_1 \leq c. \tag{38}$$

We show that the minimum is achieved when one index is set to $a_i = c$ and the rest to 0. Indeed, suppose for the sake of contradiction that $\tilde{a}$ is the solution to (38) but that there are indices i, j with $a_i \geq a_j > 0$, that is, at least two non-zero indices. By taking a logarithm, it is clear that minimizing the objective (38) is equivalent to minimizing $\sum_{i=1}^d \log(u - a_i)$. Taking the derivative of $\log(u - a_i)$ for i and j, we see that

$$\frac{\partial}{\partial a_i} \log(u - a_i) = \frac{-1}{u - a_i} \leq \frac{-1}{u - a_j} = \frac{\partial}{\partial a_j} \log(u - a_j).$$

Since $\frac{-1}{u-a}$ is decreasing function of a, increasing a_i slightly and decreasing a_j slightly causes $\log(u - a_i)$ to decrease faster than $\log(u - a_j)$ increases, thus decreasing the overall objective. This is the desired contradiction. □

E.2.2 Proof of Lemma 7

The proof of this lemma is nearly identical to the proof of Lemma 6, though we replace $\|\cdot\|_\infty$ norms with $\|\cdot\|_2$. We prove each of the statements in turn, and throughout let Z denote a variable distributed uniformly on $B_\infty(0, u)$.

(i) Jensen's inequality implies that $f(x) = f(x + \mathbb{E}[Z]) \leq \mathbb{E}[f(x + Z)] = f_\mu(x)$. For the upper bound on f_μ, use the Lipschitz continuity of f and Jensen's inequality to see that

$$f_\mu(x) \leq f(x) + L_0 \mathbb{E}[\|Z\|_2] \leq f(x) + L_0 \sqrt{\mathbb{E}[\|Z\|_2^2]} = f(x) + L_0 \sqrt{\frac{du^2}{3}}.$$

(ii) As earlier, since $\mathbb{E}[\nabla f(x+Z)] = \nabla f_\mu(x)$, we have $\|\mathbb{E}[\nabla f(x + Z)]\|_2 \leq \mathbb{E}[\|\nabla f(x + Z)\|_2] \leq L_0$.

(iii) Using the same sequence of steps as in the proof of part (iii) in Lemma 6, we see that

$$\begin{aligned}\|\nabla f_\mu(x) - \nabla f_\mu(y)\|_2 &\leq \frac{1}{(2u)^d} L_0 \lambda\left(B_\infty(x, u) \Delta B_\infty(y, u)\right) \\ &\leq \frac{2}{(2u)^d} L_0 (2u)^{d-1} \|x - y\|_1 \leq \frac{L_0\sqrt{d}}{u} \|x - y\|_2 .\end{aligned}$$

(iv) As in the proof of Lemma 6, Fubini's theorem implies the first part of the statement, while the second part is a consequence of the fact that

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

by the assumptions on F. Tightness follows from considering the one dimensional function $f(x) = |x|$ as earlier.

E.2.3 Proof of Lemma 9

Throughout this proof, we use Z to denote a random variable distributed as $N(0, u^2 I)$.

(i) As in the earlier lemmas, Jensen's inequality gives $f(x) = f(x + \mathbb{E}Z) \le \mathbb{E}f(x+Z) = f_\mu(x)$. Our assumption on $\partial F(\cdot;\xi)$ implies that f is L_0-Lipschitz, so

$$f_\mu(x) = \mathbb{E}[f(x+Z)] \le \mathbb{E}[f(x)] + L_0\mathbb{E}[\|Z\|_2] \le f(x) + L_0\sqrt{\mathbb{E}[\|Z\|_2^2]} = f(x) + L_0 u\sqrt{d}.$$

(ii) This proof is analogous to that of part (ii) of Lemmas 6 and 7. The tightness of the Lipschitz constant can be verified by taking $f(x) = \langle v, x\rangle$ for $v \in \mathbb{R}^d$, in which case $f_\mu(x) = f(x)$, and both have gradient v.

(iii) Now we show that ∇f_μ is Lipschitz continuous. Indeed, applying Lemma 11 we have

$$\|\nabla f_\mu(x) - \nabla f_\mu(y)\|_2 \le L_0 \underbrace{\int |\mu(z-x) - \mu(z-y)|dz.}_{I_2} \tag{39}$$

What remains is to control the integral term (39), denoted I_2.

In order to do so, we follow a technique used by Lakshmanan and Pucci de Farias [17]. Since μ satisfies $\mu(z-x) \ge \mu(z-y)$ if and only if $\|z-x\|_2 \ge \|z-y\|_2$, we have

$$I_2 = \int |\mu(z-x) - \mu(z-y)|dz = 2\int_{z:\|z-x\|_2 \le \|z-y\|_2} (\mu(z-x) - \mu(z-y))dz.$$

By making the change of variable $w = z - x$ for the $\mu(z-x)$ term in I_2 and $w = z - y$ for $\mu(z-y)$, we rewrite I_2 as

$$\begin{aligned} I_2 &= 2\int_{w:\|w\|_2 \le \|w-(x-y)\|_2} \mu(w)dw - 2\int_{w:\|w\|_2 \ge \|w-(x-y)\|_2} \mu(w)dw \\ &= 2\mathbb{P}_\mu(\|Z\|_2 \le \|Z - (x-y)\|_2) - 2\mathbb{P}_\mu(\|Z\|_2 \ge \|Z-(x-y)\|_2) \end{aligned}$$

where $\mathbb{P}_\mu$ denotes probability according to the density μ. Squaring the terms inside the probability bounds, we note that

$$\begin{aligned} \mathbb{P}_\mu\left(\|Z\|_2^2 \le \|Z-(x-y)\|_2^2\right) &= \mathbb{P}_\mu\left(2\langle Z, x-y\rangle \le \|x-y\|_2^2\right) \\ &= \mathbb{P}_\mu\left(2\left\langle Z, \frac{x-y}{\|x-y\|_2}\right\rangle \le \|x-y\|_2\right) \end{aligned}$$

Since $(x-y)/\|x-y\|_2$ has norm 1 and $Z \sim N(0, u^2 I)$ is rotationally invariant, the random variable $W = \left\langle Z, \frac{x-y}{\|x-y\|_2} \right\rangle$ has distribution $N(0, u^2)$. Consequently, we have

$$\begin{aligned}
\frac{I_2}{2} &= \mathbb{P}\left(W \le \|x-y\|_2/2\right) - \mathbb{P}\left(W \ge \|x-y\|_2/2\right) \\
&= \int_{-\infty}^{\|x-y\|_2/2} \frac{1}{\sqrt{2\pi u^2}} \exp(-w^2/(2u^2))dw - \int_{\|x-y\|_2/2}^{\infty} \frac{1}{\sqrt{2\pi u^2}} \exp(-w^2/(2u^2))dw \\
&\le \frac{1}{u\sqrt{2\pi}} \|x-y\|_2,
\end{aligned}$$

where we have exploited symmetry and the inequality $\exp(-w^2) \le 1$. Combining this bound with the earlier inequality (39), we have

$$\|\nabla f_\mu(x) - \nabla f_\mu(y)\|_2 \le \frac{2L_0}{u\sqrt{2\pi}} \|x-y\|_2 \le \frac{L_0}{u} \|x-y\|_2.$$

(iv) The proof of the variance bound is completely identical to that for Lemma 7.

That each of the bounds above is tight is a consequence of Lemma 10.

E.2.4 Proof of Lemma 10

Throughout this proof, c will denote a dimension independent constant and may change from line to line and inequality to inequality. We will show the result holds by considering a convex combination of "difficult" functions, in this case $f_1(x) = L_0 \|x\|_2$ and $f_2(x) = L_0 |\langle x, y/\|y\|_2\rangle - 1/2|$, and choosing $f = \frac{1}{2}f_1 + \frac{1}{2}f_2$. Our first step in the proof will be to control E_U.

By definition of the constant E_U in Eq. (33), for any convex f_1 and f_2 we have $E_U(\frac{1}{2}f_1 + \frac{1}{2}f_2) \ge \frac{1}{2}\max\{E_U(f_1), E_U(f_2)\}$. Thus for $Z \sim N(0, u^2 I_{d\times d})$ we have $\mathbb{E}[f_1(Z)] \ge cL_0 u\sqrt{d}$, i.e. $E_U(f) \ge cL_0 u\sqrt{d}$, and for Z uniform on $B_2(0,u)$, we have $\mathbb{E}[f_1(Z)] \ge cL_0 u$, i.e. $E_U(f) \ge cL_0 u$.

Turning to control of E_∇, we note that for any random variable Z rotationally symmetric about the origin, symmetry implies that

$$\mathbb{E}[\nabla f_1(Z+y)] = L_0 \mathbb{E}\left[\frac{Z+y}{\|Z+y\|_2}\right] = a_z y$$

where $a_z > 0$ is a constant dependent on Z. Thus we have

$$\mathbb{E}[\nabla f_1(Z)] - \mathbb{E}[\nabla f_1(Z+y)] + \mathbb{E}[\nabla f_2(Z)] - \mathbb{E}[\nabla f_2(Z+y)] = 0 - a_z y - L_0 \frac{y}{\|y\|_2} \int |\mu(z) - \mu(z-y)| dz$$

from Lemma 11. As a consequence (since $a_z y$ is parallel to $y/\|y\|_2$), we see that

$$E_\nabla\left(\frac{1}{2}f_1 + \frac{1}{2}f_2\right) \ge \frac{1}{2} L_0 \int |\mu(z) - \mu(z-y)| dz.$$

So what remains is to lower bound $\int |\mu(z) - \mu(z-y)| dz$ for the uniform and normal distributions. As we saw in the proof of Lemma 9, for the normal distribution

$$\int |\mu(z) - \mu(z-y)| dz = \frac{1}{u\sqrt{2\pi}} \int_{-\|y\|_2/2}^{\|y\|_2/2} \exp(-w^2/(2u^2))dw = \frac{1}{u\sqrt{2\pi}} \|y\|_2 + \mathcal{O}\left(\frac{\|y\|_2^2}{u}\right).$$

By taking small enough $\|y\|_2$, we achieve the inequality $E_{\nabla}\left(\frac{1}{2}f_1 + \frac{1}{2}f_2\right) \geq c\frac{L_0}{u}$ when $Z \sim N(0, u^2 I_{d\times d})$.

To show that the bound in the lemma is sharp for the case of the uniform distribution on $B_2(0,u)$, we slightly modify the proof of Lemma 2 in [38]. In particular, by using a Taylor expansion instead of first-order convexity in inequality (11) of [38], it is not difficult to show that

$$\int |\mu(z) - \mu(z-y)| dz = \kappa \frac{d!!}{(d-1)!!} \frac{\|y\|_2}{u} + \mathcal{O}\left(\frac{d\, \|y\|_2^2}{u^2}\right),$$

where $\kappa = 2/\pi$ if d is even and 1 otherwise. Since $d!!/(d-1)!! = \Theta(\sqrt{d})$, we have proved that for small enough $\|y\|_2$, there is a constant c such that $\int |\mu(z) - \mu(z-y)| dz \geq c\sqrt{d}\, \|y\|_2 / u$.

E.2.5 Proof of Lemma 11

Without loss of generality, we assume that $x = 0$ (a linear change of variables allows this). Let $g : \mathbb{R}^d \to \mathbb{R}^d$ be a vector-valued function such that $\|g(z)\|_* \leq L_0$ for all $z \in \{y\} + \operatorname{supp} \mu$. Then

$$\begin{aligned}
\mathbb{E}[g(Z) - g(y+Z)] &= \int g(z)\mu(z)dz - \int g(y+z)\mu(z)dz \\
&= \int g(z)\mu(z)dz - \int g(z)\mu(z-y)dz \\
&= \int_{I_>} g(z)[\mu(z) - \mu(z-y)]dz - \int_{I_<} g(z)[\mu(z-y) - \mu(z)]dz \qquad (40)
\end{aligned}$$

where $I_> = \{z \in \mathbb{R}^d \mid \mu(z) > \mu(z-y)\}$ and $I_< = \{z \in \mathbb{R}^d \mid \mu(z) < \mu(z-y)\}$. It is now clear that when we take norms we have

$$\begin{aligned}
\|\mathbb{E}g(Z) - g(y+Z)\|_* &\leq \sup_{z \in I_> \cup I_<} \|g(z)\|_* \left| \int_{I_>} [u(z) - u(z-y)]dz + \int_{I_<} [u(z-y) - u(z)]dz \right| \\
&\leq L_0 \left| \int_{I_>} \mu(z) - \mu(z-y)dz + \int_{I_<} \mu(z-y) - \mu(z)dz \right| \\
&= L_0 \int |\mu(z) - \mu(z-y)| dz.
\end{aligned}$$

Taking $g(z)$ to be an arbitrary element of $\partial f(z)$ completes the proof of the bound (35).

To see that the result is tight when μ is rotationally symmetric and the norm $\|\cdot\| = \|\cdot\|_2$, we note the following. From the equality (40), we see that $\|\mathbb{E}[g(Z) - g(y+Z)]\|_2$ is maximized by choosing $g(z) = v$ for $z \in I_>$ and $g(z) = -v$ for $z \in I_<$ for any v such that $\|v\|_2 = L_0$. Since μ is rotationally symmetric and non-increasing in $\|z\|_2$,

$$\begin{aligned}
I_> &= \left\{z \in \mathbb{R}^d \mid \mu(z) > \mu(z-y)\right\} = \left\{z \in \mathbb{R}^d \mid \|z\|_2^2 < \|z-y\|_2^2\right\} = \left\{z \in \mathbb{R}^d \mid \langle z, y\rangle < \frac{1}{2}\|y\|_2^2\right\} \\
I_< &= \left\{z \in \mathbb{R}^d \mid \mu(z) < \mu(z-y)\right\} = \left\{z \in \mathbb{R}^d \mid \|z\|_2^2 > \|z-y\|_2^2\right\} = \left\{z \in \mathbb{R}^d \mid \langle z, y\rangle > \frac{1}{2}\|y\|_2^2\right\}.
\end{aligned}$$

So all we need do is find a function f for which there exists v with $\|v\|_2 = L_0$, and such that $\partial f(x) = \{v\}$ for $x \in I_>$ and $\partial f(x) = \{-v\}$ for $x \in I_<$. By inspection, the function f defined in the statement of the lemma satisfies these two desiderata for $v = L_0 \frac{y}{\|y\|_2}$.

F Sub-Gaussian and sub-exponential tail bounds

For reference purposes, we state here some standard definitions and facts about sub-Gaussian and sub-exponential random variables (see the books [7, 21, 35] for further details).

F.1 Sub-Gaussian variables

This class of random variables is characterized by a quadratic upper bound on the moment generating function:

Definition F.1. *A zero-mean random variable X is called* sub-Gaussian with parameter σ^2 *if* $\mathbb{E}\exp(\lambda X) \leq \exp(\sigma^2\lambda^2/2)$ *for all* $\lambda \in \mathbb{R}$.

Remarks: If X_i, $i = 1, \ldots, n$ are independent sub-Gaussian with parameter σ^2, it follows from this definition that $\frac{1}{n}\sum_{i=1}^n X_i$ is sub-Gaussian with parameter σ^2/n. Moreover, it is well-known that any zero-mean random variable X satisfying $|X| \leq C$ is sub-Gaussian with parameter $\sigma^2 \leq C^2$.

Lemma 13 (Buldygin and Kozachenko [7], Lemma 1.6)**.** *Let $X - \mathbb{E}X$ be sub-Gaussian with parameter σ^2. Then for $s \in [0, 1]$,*

$$\mathbb{E}\exp\left(\frac{sX^2}{2\sigma^2}\right) \leq \frac{1}{\sqrt{1-s}}\exp\left(\frac{(\mathbb{E}X)^2}{2\sigma^2} \cdot \frac{s}{1-s}\right).$$

The maximum of d sub-Gaussian random variables grows logarithmically in d, as shown by the following result:

Lemma 14. *Let $X \in \mathbb{R}^d$ be a random vector with sub-Gaussian components, each with parameter at most σ^2. Then $\mathbb{E}\|X\|_\infty^2 \leq \max\{6\sigma^2 \log d, 2\sigma^2\}$.*

Using the definition of sub-Gaussianity, the result can be proved by a combination of union bounds and Chernoff's inequality (see van der Vaart and Wellner [35, Lemma 2.2.2] or Buldygin and Kozachenko [7, Chapter II] for details).

The following martingale-based bound for variables with conditionally sub-Gaussian behavior is essentially standard [2, 13, 7].

Lemma 15 (Azuma-Hoeffding)**.** *Let X_i be a martingale difference sequence adapted to the filtration $\mathcal{F}_i$, and assume that each X_i is conditionally sub-Gaussian with parameter σ_i^2, meaning that $\mathbb{E}[\exp(\lambda X_i) \mid \mathcal{F}_{i-1}] \leq \exp(\lambda^2\sigma_i^2/2)$. Then for all $\epsilon > 0$,*

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n X_i \geq \epsilon\right] \leq \exp\left(-\frac{n\epsilon^2}{2\sum_{i=1}^n \sigma_i^2/n}\right). \tag{41}$$

The next lemma uses martingale techniques to establish the sub-Gaussianity of a normed sum:

Lemma 16. *Let $X_1, \ldots, X_n$ be independent random vectors with $\|X_i\| \leq L$ for all i. Define $S_n = \sum_{i=1}^n X_i$. Then $\|S_n\| - \mathbb{E}\|S_n\|$ is sub-Gaussian with parameter at most $4nL^2$.*

Proof The proof follows from the realization that when $\|X_i\| \le L$, the quantity $\|S_n\| - \mathbb{E}\|S_n\|$ can be controlled using single-dimensional martingale techniques [21, Chapter 6]. We construct the Doob martingale for the sequence X_i. Let $\mathcal{F}_i$ be the σ-field of $X_1, \ldots, X_i$ and define the real-valued random variables $Z_i = \mathbb{E}[\|S_n\| \mid \mathcal{F}_i] - \mathbb{E}[\|S_n\| \mid \mathcal{F}_{i-1}]$, where $\mathcal{F}_0$ is the trivial σ-field. Let $S_{n\setminus i} = \sum_{j \ne i} X_j$. Then $\mathbb{E}[Z_i \mid \mathcal{F}_{i-1}] = 0$ and

$$\begin{aligned}
|Z_i| &= |\mathbb{E}[\|S_n\| \mid \mathcal{F}_{i-1}] - \mathbb{E}[\|S_n\| \mid \mathcal{F}_i]| \\
&\le \left|\mathbb{E}\left[\|S_{n\setminus i}\| \mid \mathcal{F}_{i-1}\right] - \mathbb{E}\left[\|S_{n\setminus i}\| \mid \mathcal{F}_i\right]\right| + \mathbb{E}[\|X_i\| \mid \mathcal{F}_{i-1}] + \mathbb{E}[\|X_i\| \mid \mathcal{F}_i] \\
&= \|X_i\| + \mathbb{E}[\|X_i\|] \le 2L
\end{aligned}$$

since X_j is independent of $\mathcal{F}_{i-1}$ for $j \ge i$. Thus Z_i defines a bounded martingale difference sequence, and $\sum_{i=1}^n Z_i = \|S_n\| - \mathbb{E}[\|S_n\|]$. Since $|Z_i| \le 2L$, the Z_i are conditionally sub-Gaussian with parameter at most $4L^2$. Thus $\sum_{i=1}^n Z_i$ is sub-Gaussian with parameter at most $4nL^2$. □

F.2 Sub-exponential random variables

A slightly less restrictive tail condition defines the class of sub-exponential random variables:

Definition F.2. *A zero-mean random variable X is* sub-exponential *with parameters (Λ, τ) if*

$$\mathbb{E}[\exp(\lambda X)] \le \exp\left(\frac{\lambda^2 \tau^2}{2}\right) \quad \text{for all } |\lambda| \le \Lambda.$$

The following lemma provides an equivalent characterization of sub-exponential variable via a tail bound:

Lemma 17. *Let X be a zero-mean random variable. If there are constants $a, \alpha > 0$ such that*

$$\mathbb{P}(|X| \ge t) \le a \exp(-\alpha t) \quad \text{for all } t > 0$$

then X is sub-exponential with parameters $\Lambda = \alpha/2$ and $\tau^2 = 4a/\alpha^2$.

The proof of the lemma follows from a Taylor expansion of $\exp(\cdot)$ and the identity $\mathbb{E}[|X|^k] = \int_0^\infty \mathbb{P}(|X|^k \ge t)dt$ (for similar results, see Buldygin and Kozachenko [7, Chapter I.3]).

Lastly, any random variable whose square is sub-exponential is sub-Gaussian, as shown by the following result:

Lemma 18 (Lan, Nemirovski, Shapiro [20], Lemma 6)**.** *Let X be a zero-mean random variable satisfying the moment generating inequality $\mathbb{E}[\exp(X^2/\sigma^2)] \le \exp(1)$. Then X is sub-Gaussian with parameter at most $3/2\sigma^2$.*

References

[1] A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright. Information-theoretic lower bounds on the oracle complexity of convex optimization. *IEEE Transactions on Information Theory*, 2012. To appear.

[2] K. Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 68:357–367, 1967.

[3] A. Ben-Tal and M. Teboulle. A smoothing technique for nondifferentiable optimization problems. In *Optimization*, Lecture Notes in Mathematics 1405, pages 1–11. Springer Verlag, 1989.

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

[5] L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7:200–217, 1967.

[6] P. Brucker. An $O(n)$ algorithm for quadratic knapsack problems. *Operations Research Letters*, 3(3):163–166, 1984.

[7] V. Buldygin and Y. Kozachenko. *Metric Characterization of Random Variables and Random Processes*, volume 188 of *Translations of Mathematical Monographs*. American Mathematical Society, 2000.

[8] C. Chen and O. L. Mangasarian. A class of smoothing functions for nonlinear and mixed complementarity problems. *Computational Optimization and Applications*, 5:97–138, 1996.

[9] O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13:165–202, 2012.

[10] J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *Proceedings of the Twenty Third Annual Conference on Computational Learning Theory*, 2010.

[11] Y. M. Ermoliev. On the stochastic quasi-gradient method and stochastic quasi-Feyer sequences. *Kibernetika*, 2:72–83, 1969.

[12] J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I*. Springer, 1996.

[13] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, Mar. 1963.

[14] P. J. Huber. *Robust Statistics*. John Wiley and Sons, New York, 1981.

[15] A. Juditsky, A. Nemirovski, and C. Tauvel. Solving variational inequalities with the stochastic mirror-prox algorithm. URL http://arxiv.org/abs/0809.0815, 2008.

[16] V. Katkovnik and Y. Kulchitsky. Convergence of a class of random search algorithms. *Automation and Remote Control*, 33(8):1321–1326, 1972.

[17] H. Lakshmanan and D. P. de Farias. Decentralized resource allocation in dynamic networks of agents. *SIAM Journal on Optimization*, 19(2):911–940, 2008.

[18] G. Lan. *Convex Optimization under Inexact First-order Information*. PhD thesis, Georgia Institute of Technology, 2009.

[19] G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 2010. Online first. URL http://www.ise.ufl.edu/glan/papers/OPT_SA4.pdf.

[20] G. Lan, A. Nemirovski, and A. Shapiro. Validation analysis of robust stochastic approximation method. *Mathematical Programming*, 2010. Online first. URL http://www.ise.ufl.edu/glan/papers/SA-test-May05.pdf.

[21] M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Springer, 1991.

[22] C. Lemaréchal and C. Sagastizábal. Practical aspects of the Moreau-Yosida regularization: theoretical preliminaries. *SIAM Journal on Optimization*, 7(2):367–385, 1997.

[23] C. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[24] S. Negahban and M. J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Annals of Statistics*, 39(2):1069–1097, 2011.

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

[26] Y. Nesterov. Smooth minimization of nonsmooth functions. *Mathematical Programming*, 103:127–152, 2005.

[27] Y. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):261–283, 2009.

[28] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

[29] B. T. Polyak and J. Tsypkin. Robust identification. *Automatica*, 16:53–63, 1980.

[30] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

[31] R. T. Rockafellar and R. J. B. Wets. *Variational Analysis*. Springer, New York, 1998.

[32] R. Y. Rubinstein. *Simulation and the Monte Carlo Method*. Wiley, 1981.

[33] S. Shalev-Shwartz, Y. Singer, and A. Ng. Online and batch learning of pseudo-metrics. In *Proceedings of the Twenty-First International Conference on Machine Learning*, 2004.

[34] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. URL http://www.math.washington.edu/~tseng/papers/apgm.pdf, 2008.

[35] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, 1996.

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

[37] E. Xing, A. Ng, M. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In *Advances in Neural Information Processing Systems 15*, 2003.

[38] F. Yousefian, A. Nedić, and U. V. Shanbhag. On stochastic gradient and subgradient methods with adaptive steplength sequences. *Automatica*, 48:56–67, 2012.