SIAM J. Optim.
Vol. 22, No. 2, pp. 674–701

RANDOMIZED SMOOTHING FOR STOCHASTIC OPTIMIZATION[*]

JOHN C. DUCHI[†], PETER L. BARTLETT[‡], AND MARTIN J. WAINWRIGHT[§]

Abstract. We analyze convergence rates of stochastic optimization algorithms for nonsmooth 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 nonsmooth 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.

Key words. convex programming, nonsmooth optimization, smoothing, stochastic optimization, distributed optimization

AMS subject classifications. 65K05, 90C15, 90C25

DOI. 10.1137/110831659

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 convex real-valued functions, each of whose domains contains 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

$$f(x) := \mathbb{E}\big[F(x;\xi)\big] \;=\; \int_{\Xi} F(x;\xi)dP(\xi). \tag{1.1}$$

We focus on potentially nonsmooth stochastic optimization problems of the form

$$\underset{x\in\mathcal{X}}{\text{minimize}} \;\; \big\{f(x) + \varphi(x)\big\}, \tag{1.2}$$

where $\varphi : \mathcal{X} \to \mathbb{R}$ is a known regularizing function. We assume that φ is closed and convex, but we allow for nondifferentiability so that the framework includes the ℓ_1-norm and related regularizers.

While we do consider effects of the regularizer φ on our optimization procedures, our primary focus is on the properties of the stochastic function f. The problem (1.2) is challenging mainly for two reasons. First, the function f may be nonsmooth. Second, in many cases, f cannot actually be evaluated. When ξ is high-dimensional, the integral (1.1) cannot be efficiently computed, and in statistical learning problems we

[*] Received by the editors April 21, 2011; accepted for publication (in revised form) April 11, 2012; published electronically June 26, 2012. The first and third authors were partially supported by MURI grant N00014-11-1-0688.

http://www.siam.org/journals/siopt/22-2/83165.html

[†] Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720 (jduchi@eecs.berkeley.edu). This author was supported by an NDSEG fellowship.

[‡] Department of Electrical Engineering and Computer Sciences, Department of Statistics, University of California Berkeley, Berkeley, CA 94720. (bartlett@eecs.berkeley.edu) and Mathematical Sciences, Queensland University of Technology, Brisbane QLD 4001, Australia. This author was supported by NSF awards DMS-0830410 and CCF-1115788.

[§] Department of Electrical Engineering and Computer Science, Department of Statistics, University of California Berkeley, Berkeley, CA 94720 (wainwrig@eecs.berkeley.edu).

usually do not even know the distribution P . Thus, throughout this work, we assume only that we have access to a stochastic oracle that allows us to obtain independent and identically distributed (i.i.d.) samples $\xi \sim P$, and we study stochastic gradient procedures for solving the convex program (1.2).

In order to address difficulties associated with nonsmooth objective functions, several researchers have considered techniques for smoothing the objective. Such approaches for deterministic nonsmooth problems are by now well known and include Moreau–Yosida regularization (e.g., [22]), methods based on recession functions [3] and Nesterov's approach using conjugacy and proximal regularization [26]. Several researchers study methods to smooth exact penalties of the form $\max\{0, f(x)\}$ in convex problems, where smoothing is applied to the $\max\{0, \cdot\}$ operator (for instance, 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 are thus impractical in stochastic settings.

Because the convex objective (1.1) cannot actually be evaluated except through stochastic realization of f and its (sub)gradients, we develop an algorithm for solving problem (1.2) based on stochastic subgradient methods. Such methods are classical [29, 12]; in recent work, Juditsky, Nemirovski, and Tauvel [16] and Lan [19] have shown that if f is smooth, meaning that its gradients are Lipschitz continuous, and if the variance of the stochastic gradient estimator is at most σ^2, then the resulting stochastic optimization procedure has convergence rate $\mathcal{O}(\sigma/\sqrt{T})$. 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. Indeed, Dekel et al. [9] exploit this fact to develop asymptotically order-optimal distributed optimization algorithms, as we discuss in what follows.

To the best of our knowledge, there is no work on *nonsmooth* stochastic problems for which a reduction in the variance of the stochastic estimate of the true subgradient gives an improvement in convergence rates. For nonsmooth stochastic optimization, known convergence rates depend 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 f at the point x. In most stochastic optimization procedures, an algorithm updates a parameter x_t after each query of the oracle; we consider the natural extension to the case when the optimizer issues several queries to the stochastic oracle at every iteration.

The starting point for our approach is a convolution-based smoothing technique amenable to nonsmooth stochastic optimization problems. A number of authors (e.g., [17, 31, 18, 37]) have noted that random perturbation of the variable x can be used to transform f into a smooth function. The intuition underlying such approaches is that the convolution of two functions is at least as smooth as the smoothest of the two original functions. In particular, letting μ denote the density of a random variable with respect to Lebesgue measure, consider the smoothed objective function

$$f_\mu(x) := \int_{\mathbb{R}^d} f(x+y)\mu(y)dy = \mathbb{E}[f(x+Z)], \tag{1.3}$$

where Z is a random variable with density μ. Clearly, the function f_μ is convex when f is convex; moreover, since μ is a density with respect to Lebesgue measure, the function f_μ is also guaranteed to be differentiable (e.g., Bertsekas [4]).

We analyze minimization procedures that solve the nonsmooth problem (1.2) by using stochastic gradient samples from the smoothed function (1.3) 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 (1.2) 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 2.1) and, under an additional reasonable tail condition, with high probability (Theorem 2.2). One consequence of our results is that a procedure that queries the nonsmooth 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 the domain $\mathcal{X}$.) 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 section 2, we begin by providing background on some standard techniques for stochastic optimization, noting a few of their deficiencies for our setting. We then describe an algorithm based on the randomized smoothing technique (1.3), and we state our main theorems guaranteeing faster rates of convergence for nonsmooth stochastic problems. In proving these claims, we make frequent use of the analytic properties of randomized smoothing, many of which are collected in Appendix E. In section 3, we discuss applications of our methods and provide experimental results illustrating the merits of our approach. Finally, we provide the proofs of our results in section 4, with certain more technical aspects deferred to the appendices.

Notation. We define $B_p(x,u) = \{y \in \mathbb{R}^d \mid \|x-y\|_p \le u\}$ to be 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 $\mathbb{R}^d$, $A+B = \{x \in \mathbb{R}^d \mid x = y+z, y \in A, z \in B\}$, multiplication of a set A by a scalar α is defined to be $\alpha A := \{\alpha x \mid x \in A\}$, and $\operatorname{aff}(A)$ denotes the affine hull of the set A. We let $\operatorname{supp}\mu := \{x \mid \mu(x) \neq 0\}$ denote the support of a function or distribution μ. We use $\partial f(x)$ to denote the subdifferential set of the convex function f at a point x. Given a norm $\|\cdot\|$, we adopt the shorthand notation $\|\partial f(x)\| = \sup\{\|g\| \mid g \in \partial f(x)\}$. The dual norm $\|\cdot\|_*$ associated with a norm $\|\cdot\|$ is given by $\|z\|_* := \sup_{\|x\|\le 1} \langle z, x\rangle$. A function f is L_0-Lipschitz with respect to the norm $\|\cdot\|$ over $\mathcal{X}$ if

$$|f(x) - f(y)| \le L_0 \|x-y\| \quad \text{for all } x, y \in \mathcal{X}.$$

We note that a convex function f is L_0-Lipschitz if and only if $\sup_{x\in\mathcal{X}} \|\partial f(x)\|_* \le L_0$ (see, e.g., the book [13]). The gradient of f is L_1-Lipschitz continuous with respect to the norm $\|\cdot\|$ over $\mathcal{X}$ if

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

A function ψ is strongly convex with respect to a norm $\|\cdot\|$ over $\mathcal{X}$ if

$$\psi(y) \ge \psi(x) + \langle g, y-x\rangle + \frac{1}{2}\|x-y\|^2 \quad \text{for all } g \in \partial\psi(x) \text{ and } x, y, \in \mathcal{X}.$$

Given a differentiable convex 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 $\|X\|_{\mathrm{Fr}}$ denote its Frobenius norm. The transpose of X is denoted $X^\top$. The notation $\xi \sim P$ indicates that the random variable ξ is drawn from the distribution P, and P-a.e. ξ is shorthand for P-almost every ξ.

2. Main results and some consequences.

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

2.1. Some background.

We focus on stochastic gradient descent methods[1] based on dual averaging schemes [27] for solving the stochastic problem (1.2). Dual averaging methods are based on a proximal function ψ that is assumed to be strongly convex with respect to a norm $\|\cdot\|$. Given a point $x_t \in \mathcal{X}$, the algorithm queries a stochastic oracle and receives a random vector $g_t \in \mathbb{R}^d$ satisfying the inclusion $\mathbb{E}[g_t \mid x_t, g_1, \dots, g_{t-1}] \in \partial f(x_t)$. The algorithm then performs the update

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \sum_{\tau=0}^{t} \langle g_\tau, x\rangle + \frac{1}{\alpha_t}\psi(x) \right\}, \tag{2.1}$$

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. Then, with stepsizes $\alpha_t \propto R/L_0\sqrt{t}$, the sequence $\{x_t\}_{t=0}^\infty$ generated by the update (2.1) satisfies

$$\mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=1}^{T} x_t \right) \right] - f(x^*) = \mathcal{O}\left(\frac{L_0\sqrt{\psi(x^*)}}{\sqrt{T}} \right). \tag{2.2}$$

We refer the reader to papers by Nesterov [27] and Xiao [35] for results of this type.

An unsatisfying aspect of the bound (2.2) is the absence of any role for the variance of the (sub)gradient estimator g_t. Even if an algorithm is able to obtain $m > 1$ samples of the gradient of f at x_t—giving a more accurate gradient estimate—this result fails to capture the potential improvement of the method. We address this problem by stochastically smoothing the nonsmooth objective f and then adapt recent work on so-called accelerated gradient methods [19, 33, 35], which apply only to smooth functions, to achieve variance-based improvements. With this motivation in mind, we now turn to developing the tools necessary for stochastic smoothing of the nonsmooth objective function (1.2).

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. Consider the following scheme. Let $\{u_t\} \subset \mathbb{R}_+$ be a nonincreasing 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 $\{Z_{i,t}\}_{i=1}^m$ i.i.d. according to the distribution μ.

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

[1] We note in passing that essentially identical results can also be obtained for methods based on mirror descent [25, 33], though we omit these so as not to overburden the reader.

$$(2.3)\qquad g_{i,t} \in \partial F(y_t + u_t Z_{i,t}; \xi_{i,t}), \quad \text{where } \xi_{i,t} \sim P \text{ for } i = 1, 2, \ldots, m.$$

(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 (1.3) and μ_t is the density of u_t.

We combine the sampling scheme (2.3) with extensions of Tseng's recent work on accelerated gradient methods [33] and propose an update that is essentially a smoothed version of the simpler method (2.1). 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 (2.3). 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:

$$(2.4a)\qquad y_t = (1-\theta_t)x_t + \theta_t z_t,$$

$$(2.4b)\qquad z_{t+1} = \operatorname*{argmin}_{x\in\mathcal{X}} \left\{ \sum_{\tau=0}^{t} \frac{1}{\theta_\tau} \langle g_\tau, x\rangle + \sum_{\tau=0}^{t} \frac{1}{\theta_\tau}\varphi(x) + L_{t+1}\psi(x) + \frac{\eta_{t+1}}{\theta_{t+1}}\psi(x) \right\},$$

$$(2.4c)\qquad x_{t+1} = (1-\theta_t)x_t + \theta_t z_{t+1}.$$

In prior work on accelerated schemes for stochastic and nonstochastic optimization [33, 19, 35], 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 [33], 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 (2.3) with accelerated dual-averaging updates (2.4a)–(2.4c). To avoid cluttering the theorem statements, we begin by stating our main assumptions and notation. Whenever we state that a function f is Lipschitz continuous, we mean with respect to the norm $\|\cdot\|$, and we assume that ψ is nonnegative and is strongly convex with respect to the same norm $\|\cdot\|$. Our main assumption ensures that the smoothing operator and smoothed function f_μ are relatively well behaved.

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

Let μ_t denote the density of the random vector $u_t Z$, and define the instantaneous smoothed function $f_{\mu_t} = \int f(x+z)d\mu_t(z)$. 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 nonempty at all sampled points $y_t + u_t Z$. Indeed, since μ is a density with respect to Lebesgue measure on $\operatorname{aff}(\mathcal{X})$, with probability 1 $y_t + u_t Z \in \operatorname{relint}\operatorname{dom} F(\cdot;\xi)$, and thus [13] the subdifferential $\partial F(y_t + u_t Z;\xi) \neq \emptyset$.

In the algorithm (2.4a)–(2.4c), 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 2.1. *Define $u_t = \theta_t u$, use the scalar sequence $L_t = L_1/u_t$, and assume that η_t is nondecreasing. Under Assumption* A, *for any $x^* \in \mathcal{X}$ and $T \geq 4$,*

(2.5)

$$\mathbb{E}[f(x_T)+\varphi(x_T)]-[f(x^*)+\varphi(x^*)] \leq \frac{6L_1\psi(x^*)}{Tu}+\frac{2\eta_T\psi(x^*)}{T}+\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{\eta_t}\mathbb{E}[\|e_t\|_*^2]+\frac{4L_0u}{T},$$

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

Remarks. The convergence rate (2.5) involves the variance $\mathbb{E}[\|e_t\|_*^2]$ explicitly, which we exploit in the corollaries to be stated shortly. In addition, Theorem 2.1 does not require a priori knowledge of the number of iterations T to be performed, thereby rendering it suitable to online and streaming applications. If T is known, a similar result holds for constant smoothing parameter u, as formalized by Theorem 4.4.

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 $s = 0, \ldots, t$, defined in (2.3). In order to achieve high-probability convergence results, a subset of our results involve the following assumption.

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

$$\mathbb{E}[\exp(\|e_t\|_*^2/\sigma^2) \mid \mathcal{F}_{t-1}] \leq \exp(1) \quad \textit{for all } t \in \{1, 2, \ldots\}. \tag{2.6}$$

We refer the reader to Appendix F for more background on sub-Gaussian and subexponential random variables. In past work on smooth optimization, other authors [16, 19, 35] have imposed this type of tail assumption, and we discuss sufficient conditions for the assumption to hold in Corollary 2.6 in the following section.

Theorem 2.2. *In addition to the conditions of Theorem* 2.1, *suppose that $\mathcal{X}$ 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*

$$f(x_T)+\varphi(x_T)-f(x^*)-\varphi(x^*) \leq \frac{6L_1\psi(x^*)}{Tu}+\frac{4L_0u}{T}+\frac{2\eta_T\psi(x^*)}{T}+\sum_{t=0}^{T-1}\frac{\mathbb{E}[\|e_t\|_*^2 \mid \mathcal{F}_{t-1}]}{T\eta_t}$$
$$+\frac{4\sigma^2\max\left\{\log\frac{1}{\delta}, \sqrt{2e(1+\log T)\log\frac{1}{\delta}}\right\}}{\eta T}+\frac{\sigma R\sqrt{\log\frac{1}{\delta}}}{\sqrt{T}}.$$

Remarks. The first four terms in the convergence rate provided by Theorem 2.2 are essentially identical to the rate in expectation stated in Theorem 2.1. There are two additional terms, the first of which decreases at a rate of $1/T$, while the second decreases at a rate of $\sigma/\sqrt{T}$. As discussed in the corollaries to follow, the dependence $\sigma/\sqrt{T}$ on the variance σ^2 is optimal, and an appropriate choice of the sequence η_t in Theorem 2.1 yields identical rates to those in Theorem 2.2.

2.4. Some consequences. We now turn to various corollaries of the above theorems and the consequential optimality guarantees of the algorithm. More precisely, we establish concrete convergence bounds for algorithms using different choices of the smoothing distribution μ. For each corollary, we impose the assumptions that the point $x^* \in \mathcal{X}$ satisfies $\psi(x^*) \leq R^2$, the iteration number $T \geq 4$, and $u_t = u\theta_t$.

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 2.3. *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 step sizes $\eta_t = L_0\sqrt{t+1}/R\sqrt{m}$ 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 2.4. *Let μ be the d-dimensional normal distribution with zero mean and identity covariance I and assume $F(\cdot;\xi)$ 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 note here (deferring deeper discussion to Lemma E.4) 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 versions 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 (NLP) [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]. 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 point here is that an alternative pair of dual norms may give better optimization performance than the ℓ_2-ℓ_2 pair above.

COROLLARY 2.5. *Let μ be uniform 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 $\eta_t = \sqrt{t+1}/R\sqrt{m\log d}$ and $L_t = L_0/u_t$. There is a universal constant C such that*

$$\begin{aligned}\mathbb{E}[f(x_T) + \varphi(x_T)] - [f(x^*) + \varphi(x^*)] &\le C\frac{L_0R\sqrt{d}}{T} + C\frac{L_0R\sqrt{\log d}}{\sqrt{Tm}}\\ &= \mathcal{O}\left(\frac{L_0\|x^*\|_1\sqrt{d\log d}}{T} + \frac{L_0\|x^*\|_1\log d}{\sqrt{Tm}}\right).\end{aligned}$$

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 might desire when applied to non-Euclidean geometries. Nonetheless, for large T the $1/\sqrt{Tm}$ terms dominate the convergence rates, and Corollary 2.5 can be an improvement.

Our final corollary specializes the high probability convergence result in Theorem 2.2 by showing that the error is sub-Gaussian (2.6) 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 such as that above.

COROLLARY 2.6. *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 the smoothing distribution μ uniform on $B_2(0,1)$ and parameters u, η_t, and L_t identical to those in Corollary* 2.3*, with probability at least $1-\delta$,*

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

Remarks. Let us pause to make some remarks concerning the corollaries given above. First, if one abandons the requirement that the optimization procedure be an "any time" algorithm, meaning that it is able to return a result at any iteration, it is possible to obtain essentially the same results as Corollaries 2.3–2.5 by choosing a fixed setting $u_t = u/T$ (see Theorem 4.4 in section 4.4). 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 $x \in \mathcal{X} + u_0 \operatorname{supp}\mu$. Our second observation is that Theorem 2.1 and the corollaries appear to require a very specific setting of the constant L_t to achieve fast rates. However, the algorithm is robust to misspecification 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 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 an essentially negligible effect.

We can show that the bounds in the theorems above are tight, meaning unimprovable up to constant factors, by exploiting known lower bounds [25, 1] for stochastic optimization problems. For instance, let us set $\mathcal{X} = \{x \in \mathbb{R}^d \mid \|x\|_2 \le R_2\}$ and consider the class of 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 for which $\mathbb{E}[g] \in \partial f(x)$ and $\mathbb{E}[\|g\|_2^2] \le L_{0,2}^2$. 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 $L_{0,2}$-Lipschitz convex f (see section 3.1 of the paper [1]). Moreover, similar bounds hold for problems with non-Euclidean geometry [1]. For instance, let us consider convex functions f that are $L_{0,\infty}$-Lipschitz with respect to the ℓ_1-norm, meaning that $|f(x) - f(y)| \le L_{0,\infty}\|x-y\|_1$. If we define $\mathcal{X} = \{x \in \mathbb{R}^d \mid \|x\|_1 \le R_1\}$, we then 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 2.3–2.5. Examining the bound in Corollaries 2.3 and 2.4, we see that the dominant terms are on the order of $L_0R/\sqrt{Tm}$ so long as $m \leq T/\sqrt{d}$. Since our method issues Tm queries to the oracle, this is optimal. Similarly, the strategy of sampling uniformly from the ℓ_∞-ball in Corollary 2.5 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 applications of our results and give experiments 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 2.3–2.6 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 (2.3) and (2.4a)–(2.4c), 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 (2.4b) becomes expensive. The proximal update may be distilled to computing

$$\min_{x\in\mathcal{X}} \left\{ \langle g, x\rangle + \psi(x) \right\} \quad \text{or} \quad \min_{x\in\mathcal{X}} \left\{ \langle g, x\rangle + \psi(x) + \varphi(x) \right\}. \tag{3.1}$$

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 (2.4b), and thus solve only one proximal subproblem 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 $\mathcal{X} = \{X \in \mathbb{R}^{d_1\times d_2} \mid \sum_{j=1}^{\min\{d_1,d_2\}} \rho_j(X) \leq C\}$. 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 [36, 32], 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 nonsmooth program

$$(3.2)\quad \min_{X,x} \frac{1}{\binom{n}{2}} \sum_{i<j} \left[1 + b_{ij}(\mathrm{tr}(X(a_i - a_j)(a_i - a_j)^\top) + x)\right]_+ \text{ s.t. } X \succeq 0,\ \ \mathrm{tr}(X) \le C.$$

Now let us consider the cost of computing the projection update (3.1) for the metric learning problem (3.2). When $\psi(X) = \frac{1}{2}\|X\|_{\mathrm{Fr}}^2$, the update (3.1) 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 [11].) 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 $\mathcal{O}(d^3)$ eigendecomposition requires time $\mathcal{O}(d)$ [6]. To see the benefits of the randomized perturbation and averaging technique (2.3) over standard stochastic gradient descent (2.1), consider that the cost of querying a stochastic oracle for the objective (3.2) 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 the standard stochastic gradient method requires Tmd^3 units of computation. In short, for $m \approx d$ the randomized smoothing technique (2.3) 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 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 (2.3). The second experiment studies the actual amount of time required to solve a statistical metric learning problem, as described in the objective (3.2) above.

3.2.1. Iteration complexity of reduced variance estimators. In this experiment, we consider the number of iterations of the accelerated method (2.4a)–(2.4c) necessary to achieve an ϵ-optimal solution to the problem (1.2). To understand how the iteration scales with the number m of gradient samples, we consider our results in terms of the number of iterations

$$T(\epsilon, m) := \inf\left\{t \in \{1, 2, \ldots\} \mid f(x_t) - \min_{x \in \mathcal{X}} f(x^*) \le \epsilon\right\}$$

required to achieve optimization error ϵ when using m gradient samples in the averaging step (2.3). We focus on the algorithm analyzed in Corollary 2.3, which uses

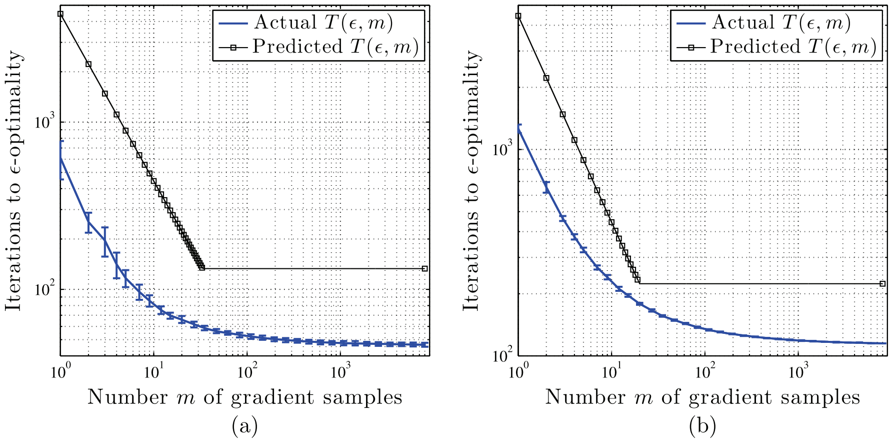

FIG. 3.1. *The number of iterations $T(\epsilon, m)$ to achieve the ϵ-optimal solution for the problem* (3.4) *as a function of the number of samples m used in the gradient estimate* (2.3)*. The prediction* (3.3) *is the square black line in each plot; plot* (a) *shows results for dimension $d = 50$ and* (b) *for $d = 400$.*

uniform sampling of the ℓ_2-ball. The corollary implies there should be two regimes of convergence—one where the $L_0 R/\sqrt{Tm}$ term is dominant, and the other where the number of samples m is so large that the $L_0 R d^{1/4}/T$ term is dominant. By inverting the first term, we see that for small m, $T(\epsilon, m) = \mathcal{O}(L_0^2 R^2/m\epsilon^2)$, while the second gives $T(\epsilon, m) = \mathcal{O}(L_0 R d^{1/4}/\epsilon)$. In particular, our theory predicts that

$$T(\epsilon, m) = \mathcal{O}\left(\max\left\{\frac{L_0^2 R^2}{m\epsilon^2}, \frac{L_0 R d^{1/4}}{\epsilon}\right\}\right). \tag{3.3}$$

In order to assess the accuracy of this prediction, we consider a robust linear regression problem commonly studied in system identification and robust statistics [28, 15]. Specifically, given a matrix $A \in \mathbb{R}^{n \times d}$ and a vector $b \in \mathbb{R}^n$, the goal is to minimize the nonsmooth objective function

$$f(x) = \frac{1}{n}\left\|Ax - b\right\|_1 = \frac{1}{n}\sum_{i=1}^{n} |\left\langle a_i, x\right\rangle - b_i|, \tag{3.4}$$

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

In performing our experiments, we generated $n = 1000$ points in dimensions $d \in \{50, 100, 200, 400, 800, 1600\}$, each with fixed norm $\left\|a_i\right\|_2 = L_0$, and then assigned values b_i by computing $\langle a_i, w\rangle$ for a random vector w (adding normally distributed noise with variance 0.1). We estimated the quantity $T(\epsilon, m)$ for solving the robust regression problem (3.4) for several values of m and d. Figure 3.1 shows results for dimensions $d \in \{50, 400\}$ 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 (3.3). 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 3.1 that allow us to roughly estimate the number m at which increasing the batch size in the

Table 3.1

The number of iterations $T(\epsilon, m)$ to achieve ϵ-accuracy for the regression problem (3.4) as a function of number of gradient samples m used in the gradient estimate (2.3) and the dimension d. Each box in the table shows the mean and standard deviations of $T(\epsilon, m)$ measured over 20 trials.

	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

gradient estimate (2.3) gives no further improvement. For each dimension d, Table 3.1 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 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 were based on instances of the metric learning problem. For each $i, j = 1, \ldots, n$, we are given a vector $a_i \in \mathbb{R}^d$ and a measure $b_{ij} \geq 0$ of the similarity between the vectors a_i and a_j. (Here $b_{ij} = 0$ means that a_i and a_j are the same.) The statistical goal is to learn a matrix X—inducing a pseudonorm via $\|a\|_X^2 := \langle a, Xa \rangle$—such that $\langle (a_i - a_j), X(a_i - a_j) \rangle \approx b_{ij}$. One method for doing so is to minimize the objective

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

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

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

We solve ten random problems with dimension $d = 100$ and $n = 2000$, giving an objective with $4 \cdot 10^6$ terms and 5050 parameters. Performing stochastic optimization is more viable for this problem than a nonstochastic method, as even computing the objective requires $\mathcal{O}(n^2 d^2)$ operations. We plot experimental results in Figure 3.2 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 (2.3). As predicted by our theory and discussion in section 3, 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 Figure 3.2 suggests that this too is correct, as the plots for $m = 64$ and $m = 128$ are essentially indistinguishable.

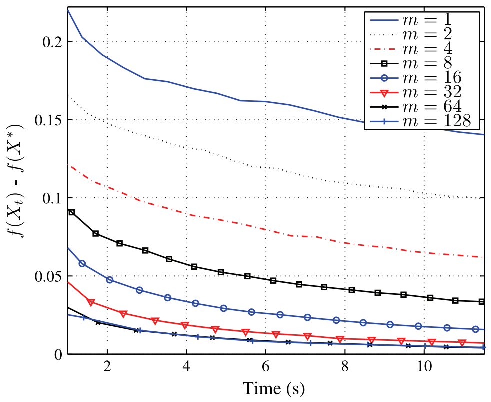

FIG. 3.2. *Optimization error* $f(X_t) - \inf_{X^* \in \mathcal{X}} f(X^*)$ *in the metric learning problem of section* 3.2.2 *as a function of time in seconds. Each line indicates the optimization error over time for a particular number of samples* m *in the gradient estimate* (2.3)*; we set* $m = 2^i$ *for* $i = \{1, \ldots, 7\}$.

4. Proofs. In this section, we provide the proofs of Theorems 2.1 and 2.2 as well as of Corollaries 2.3–2.6. 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 the appendices.

The general technique for the proof of each corollary is as follows. First, we note that the randomly smoothed function $f_\mu(x) = \mathbb{E}[f(x + Z)]$ has Lipschitz continuous gradient, and it is uniformly close to the original nonsmooth function f. This fact allows us to apply Theorem 2.1. The second step is to realize that with the sampling procedure (2.3), the variance $\mathbb{E}[\|e_t\|_*^2]$ decreases by a factor of approximately 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 (1.3), and so we refer frequently to lemmas stated in Appendix E.

4.1. Proofs of Corollaries 2.3 and 2.4. We begin by proving Corollary 2.3. 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 E.2 in Appendix E, the variance of g_t as an estimate of $\nabla f_{\mu_t}(y_t)$ satisfies

$$\sigma^2 := \mathbb{E}[\|e_t\|_2^2] = \mathbb{E}[\|g_t - \nabla f_{\mu_t}(y_t)\|_2^2] \le \frac{L_0^2}{m}. \tag{4.1}$$

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 $x \mapsto \mathbb{E}[f(x + uZ)]$ has $L_0\sqrt{d}/u$-Lipschitz continuous gradient. Thus, applying Lemma E.2 and Theorem 2.1 with the setting $L_t = L_0\sqrt{d}/u\theta_t$, we obtain

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

where we have used the bound (4.1).

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

$$\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 1 + \int_1^T \frac{1}{\sqrt{t}} dt = 1 + 2(\sqrt{T}-1) \leq 2\sqrt{T}, \tag{4.2}$$

we use the bound $2\sqrt{T+1}/T + 2/\sqrt{T} \leq 5/\sqrt{T}$ to obtain

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

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

The proof of Corollary 2.4 is essentially identical, differing only in the setting of $u = Rd^{-1/4}$ and the application of Lemma E.3 instead of Lemma E.2 in Appendix E.

4.2. Proof of Corollary 2.5. Under the conditions of the corollary, Lemma E.1 implies that when μ is uniform on $B_\infty(0,u)$, then the function $f_\mu(x) := \mathbb{E}[f(x+Z)]$ has L_0/u-Lipschitz continuous gradient with respect to the ℓ_1-norm, and moreover it satisfies the upper bound $f_\mu(x) \leq f(x) + \frac{L_0du}{2}$. 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

$$\mathbb{E}\big[\|g - \nabla f_\mu(x)\|_\infty^2\big] \leq C\,\frac{L_0^2 \log d}{m} \tag{4.3}$$

for some universal constant C. Indeed, Lemma E.1 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 $\|g_i\|_\infty \leq L_0$ and $\|\nabla f_\mu(x)\|_\infty \leq L_0$, the vector $g_i - \nabla f_\mu(x)$ is a d-dimensional random vector whose components are sub-Gaussian with parameter $4L_0^2$. Conditional on x, the g_i are independent, and so $g - \nabla f_\mu(x)$ has sub-Gaussian components with parameter at most $4L_0^2/m$. Applying Lemma F.3 from Appendix F with $X = g - \nabla f_\mu(x)$ and $\sigma^2 = 4L_0^2/m$ yields the claim (4.3).

Now, as in the proof of Corollary 2.3, we can apply Theorem 2.1. Recall [25] that $\frac{1}{2(p-1)}\|x\|_p^2$ is strongly convex over $\mathbb{R}^d$ with respect to the ℓ_p-norm for $p \in (1,2]$. 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 $\|x^*\|_p^2 \log d \leq \|x^*\|_1^2 \log d$. 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 $\|g\|_q \leq d^{1/q}\|g\|_\infty$. Of course, $d^{1/(\log d + 1)} \leq d^{1/(\log d)} = \exp(1)$, and so $\|g\|_q \leq e\|g\|_\infty$.

Having shown that the Lipschitz constant L for the ℓ_p-norm satisfies $L \leq L_0 e$, where L_0 is the Lipschitz constant with respect to the ℓ_1-norm, we apply Theorem 2.1 and the variance bound (4.3) to obtain the result. Specifically, Theorem 2.1 implies

$$\mathbb{E}[f(x_T)+\varphi(x_T)] - [f(x^*)+\varphi(x^*)] \leq \frac{6L_0R^2}{Tu} + \frac{2\eta_T R^2}{T} + \frac{C}{T}\sum_{t=0}^{T-1}\frac{1}{\eta_t}\cdot\frac{L_0^2\log d}{m} + \frac{4L_0du}{2T}.$$

Plugging in u, η_t, and $R \leq \|x^*\|_1\sqrt{\log d}$ and using bound (4.2) completes the proof.

4.3. Proof of Corollary 2.6. 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 4.1. *In the notation of Theorem* 2.2, *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 2.6. When μ is the uniform distribution on $B_2(0,u)$, Lemma E.2 from Appendix E implies that ∇f_μ is Lipschitz with constant $L_1 = L_0\sqrt{d}/u$. Lemma 4.1 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 2.2 with Lemma 4.1, 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 2.1. This proof is more involved than those of the above corollaries. In particular, we build on techniques used in the work of Tseng [33], Lan [19], and Xiao [35]. The changing smoothness of the stochastic objective—which comes from changing the shape parameter of the sampling distribution Z in the averaging step (2.3)—adds some challenge. Essentially, the idea of the proof is to let μ_t be the density of u_tZ and define $f_{\mu_t}(x) := \mathbb{E}[f(x+u_tZ)]$, where u_t is the nonincreasing sequence of shape parameters in the averaging scheme (2.3). 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 [33] 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 4.4), which gives an alternative setting for u given a priori knowledge of the number of iterations.

We begin by stating two technical lemmas.

LEMMA 4.2. *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* (2.4a)–(2.4c), *and define the error term $e_t = \nabla f_{\mu_t}(y_t) - g_t$. Then for any $x^* \in \mathcal{X}$,*

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

See Appendix B for the proof of this claim.

LEMMA 4.3. *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 $\sum_{\tau=0}^{t}\frac{1}{\theta_\tau} = \frac{1}{\theta_t^2}$.*

Tseng [33] proves the second statement; the first follows by 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 4.2. 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 4.2 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 4.3 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 $u_t = \theta_t u$ is in turn bounded by

$$\frac{1}{\theta_{T-1}^2}[f_{\mu_{T-1}}(x_T) + \varphi(x_T)] - \sum_{t=0}^{T-1} \frac{1}{\theta_t}[f_{\mu_t}(x^*) + \varphi(x^*)] + T L_0 u. \tag{4.4}$$

Now we apply Lemma 4.2 to the bound (4.4), which gives us

$$\begin{aligned} &\frac{1}{\theta_{T-1}^2}\,[f(x_T) + \varphi(x_T) - f(x^*) - \varphi(x^*)] \\ &\le L_T \psi(x^*) + \frac{\eta_T}{\theta_T}\psi(x^*) + \sum_{t=0}^{T-1} \frac{1}{2\theta_t \eta_t}\,\|e_t\|_*^2 + \sum_{t=0}^{T-1} \frac{1}{\theta_t}\,\langle e_t, z_t - x^* \rangle + T L_0 u. \end{aligned} \tag{4.5}$$

The nonprobabilistic bound (4.5) is the key to the remainder of this proof, as well as the starting point for the proof of Theorem 2.2 in the next section. What remains here is to take expectations in the bound (4.5).

Recall the filtration of σ-fields $\mathcal{F}_t$, which satisfy $x_t, y_t, z_t \in \mathcal{F}_{t-1}$; 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 are measurable with respect to $\mathcal{F}_{t-1}$. Now, recall from Lemma 4.3 that $\theta_t \le \frac{2}{2+t}$ and that $(1 - \theta_t)/\theta_t^2 = 1/\theta_{t-1}^2$. Thus

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

Furthermore, we have $\theta_{t+1} \le \theta_t$, so by multiplying both sides of our bound (4.5) by θ_{T-1}^2 and taking expectations over the random vectors g_t,

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

where we used the fact that $L_T = L_1/u_T = L_1/\theta_T u$. This completes the proof of Theorem 2.1. □

We conclude with a theorem using a fixed setting of the smoothing parameter u_t. It is clear that by setting $u \propto 1/T$, the rates achieved by Theorems 2.1 and 4.4 are identical up to constant factors.

THEOREM 4.4. *Suppose that $u_t \equiv u$ for all t and set $L_t \equiv L_1/u$. With the remaining conditions as in Theorem* 2.1*, then for any $x^* \in \mathcal{X}$, we have*

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

Proof. If we fix $u_t \equiv u$ for all t, then the bound (4.5) holds with the last term TL_0u replaced by $\theta_{T-1}^2 L_0 u$, which we see by invoking Lemma 4.3. The remainder of the proof follows unchanged, with $L_t \equiv L_1$ for all t. □

4.5. Proof of Theorem 2.2. An examination of the proof of Theorem 2.1 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 4.5. *Let $\mathcal{X}$ satisfy $\|x - x^*\| \le R$ for all $x \in \mathcal{X}$. Under Assumption* B*,*

$$\mathbb{P}\bigg[\theta_{T-1}^2\sum_{t=0}^{T-1}\frac{1}{\theta_t}\langle e_t, z_t - x^*\rangle \ge \epsilon\bigg] \le \exp\bigg(-\frac{T\epsilon^2}{R^2\sigma^2}\bigg). \tag{4.6}$$

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

$$\theta_{T-1}^2\sum_{t=0}^{T-1}\frac{1}{\theta_t}\langle e_t, z_t - x^*\rangle \le R\sigma\sqrt{\frac{\log\frac{1}{\delta}}{T}}. \tag{4.7}$$

LEMMA 4.6. *In the notation of Theorem* 2.2 *and under Assumption* B*, we have*

$$\log\mathbb{P}\bigg[\sum_{t=0}^{T-1}\frac{\|e_t\|_*^2}{2\eta_t} \ge \sum_{t=0}^{T-1}\frac{\mathbb{E}[\|e_t\|_*^2 \mid \mathcal{F}_{t-1}]}{2\eta_t} + \epsilon\bigg] \le \max\bigg\{-\frac{\epsilon^2}{32e\sigma^4\sum_{t=0}^{T-1}\frac{1}{\eta_t^2}}, -\frac{\eta_0}{4\sigma^2}\epsilon\bigg\}. \tag{4.8}$$

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

$$\sum_{t=0}^{T-1}\frac{\|e_t\|_*^2}{2\eta_t} \le \sum_{t=0}^{T-1}\frac{\mathbb{E}[\|e_t\|_*^2 \mid \mathcal{F}_{t-1}]}{2\eta_t} + \frac{4\sigma^2}{\eta}\max\bigg\{\log\frac{1}{\delta}, \sqrt{2e(1+\log T)\log\frac{1}{\delta}}\bigg\}. \tag{4.9}$$

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

Equipped with Lemmas 4.5 and 4.6, we now prove Theorem 2.2. Let us recall the deterministic bound (4.5) from the proof of Theorem 2.1:

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

Since $\theta_{T-1}\leq\theta_t$ for $t\in\{0,\ldots,T-1\}$, Lemmas 4.5 and 4.6 combined with a union bound imply that with probability at least $1-2\delta$

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

The terms remaining to control are deterministic, and as in Theorem 2.1 we have

$$\theta_{T-1}^2L_T\leq\frac{6L_1}{Tu},\quad\frac{\theta_{T-1}^2\eta_T}{\theta_T}\leq\frac{2\eta_T}{T},\quad\text{and}\quad\theta_{T-1}^2TL_0u\leq\frac{4L_0u}{T+1}.$$

Combining the above bounds completes the proof. □

5. Discussion. In this paper, we have developed and analyzed smoothing strategies for stochastic nonsmooth optimization that are provably optimal in the stochastic oracle model of optimization complexity, and we have given—to the best of our knowledge—the first variance reduction techniques for nonsmooth stochastic optimization. We think that at least two obvious questions remain. The first 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 2.3–2.6 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.

Appendix A. Proof of Lemma 4.1. The proof of this lemma requires several auxiliary results on sub-Gaussian and subexponential random variables, which we collect and prove in Appendix F. For notational convenience, we take expectations $\mathbb{E}$ conditional on $\mathcal{F}_{t-1}$ without mention.

For each $i=1,\ldots,m$, define the random variable $X_i=\nabla f_{\mu_t}(y_t)-g_{i,t}$, and define the sum $S_m=\sum_{i=1}^m X_i$. With these definitions, we have the convenient relation $\frac{1}{m}S_m=\nabla f_{\mu_t}(y_t)-\frac{1}{m}\sum_{i=1}^m g_{i,t}$. Conditioned on $\mathcal{F}_{t-1}$, the X_i are independent, and we have $\left\|X_i\right\|_*\leq L:=2L_0$. Consequently, by applying Lemma F.5 from Appendix F, we conclude that the random variable $\|\frac{1}{m}S_m\|_*-\mathbb{E}[\|\frac{1}{m}S_m\|_*]$ is sub-Gaussian with parameter at most $4L^2/m$. Applying Lemma F.2 from Appendix F yields

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

Since $4L^2/m \le \max\{\mathbb{E}\left[\|\frac{1}{m}S_m\|_*^2\right], 4L^2/m\}$, we conclude that

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

For any random variable X, Jensen's inequality implies that $(\mathbb{E}X)^2 \le \mathbb{E}X^2$, so that

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

Taking $s = \frac{1}{2}$ yields the upper bound

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

which completes the proof.

Appendix B. Proof of Lemma 4.2. Define the linearized version of the cumulative objective

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

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

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

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 (2.4a)–(2.4c) of the triple (x_t, y_t, z_t), we obtain

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

Finally, by convexity of the regularizer φ, we conclude that

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

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

$$D_\psi(x, y) \; = \; \psi(x) - \psi(y) - \langle\nabla\psi(y), x - y\rangle \; \ge \; \frac{1}{2}\left\|x - y\right\|^2. \tag{B.3}$$

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

$$f_{\mu_t}(y_t) + \langle \nabla f_{\mu_t}(y_t), x_t - y_t \rangle \le f_{\mu_t}(x_t). \tag{B.4}$$

Substituting inequalities (B.3) and (B.4) into the bound (B.2) and simplifying yields

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

We now rewrite this upper bound in terms of the error $e_t = \nabla f_{\mu_t}(y_t) - g_t$:

$$\begin{aligned}\phi_t(x_{t+1}) &\le \theta_t \left[f_{\mu_t}(y_t) + \langle g_t, z_{t+1} - y_t\rangle + \varphi(z_{t+1}) + L_t\theta_t D_\psi(z_{t+1}, z_t)\right] \\ &\qquad + (1-\theta_t)[f_{\mu_t}(x_t) + \varphi(x_t)] + \theta_t \langle e_t, z_{t+1} - y_t\rangle \\ &= \theta_t^2 \left[\ell_t(z_{t+1}) - \ell_{t-1}(z_{t+1}) + L_t D_\psi(z_{t+1}, z_t)\right] \\ &\qquad + (1-\theta_t)[f_{\mu_t}(x_t) + \varphi(x_t)] + \theta_t \langle e_t, z_{t+1} - y_t\rangle . \end{aligned} \tag{B.5}$$

The first order convexity conditions for optimality imply that for some $g \in \partial\ell_{t-1}(z_t)$ and all $x \in \mathcal{X}$, we have $\langle g + \frac{1}{\alpha_t}\nabla\psi(z_t), x - z_t\rangle \ge 0$ since z_t minimizes $\ell_{t-1}(x) + \frac{1}{\alpha_t}\psi(x)$. Thus, first-order convexity gives

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

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

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

Combining this inequality with the bound (B.5) and the definition $\alpha_t^{-1} = L_t + \eta_t/\theta_t$,

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

since α_t^{-1} is nondecreasing. We now divide both sides by θ_t^2 and unwrap the recursion. By construction $(1-\theta_t)/\theta_t^2 = 1/\theta_{t-1}^2$ and $f_{\mu_t} \le f_{\mu_{t-1}}$, so we obtain

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

The second inequality follows by combination of the facts that $(1-\theta_t)/\theta_t^2 = 1/\theta_{t-1}^2$ and $f_{\mu_t} \le f_{\mu_{t-1}}$. By applying the two 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

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

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

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

Recalling the definition (B.1) of ℓ_t and noting that the first-order conditions for convexity imply that $f_{\mu_t}(y_t) + \langle\nabla f_{\mu_t}(y_t), x-y_t\rangle \le f_{\mu_t}(x)$, we expand ℓ_t and have

$$\begin{aligned}\frac{1}{\theta_t^2}[f_{\mu_t}(x_{t+1})+\varphi(x_{t+1})] &\le \sum_{\tau=0}^{t}\frac{1}{\theta_\tau}[f_{\mu_\tau}(y_\tau)+\langle g_\tau, x^*-y_\tau\rangle + \varphi(x^*)] + \frac{1}{\alpha_{t+1}}\psi(x^*) \\ &\quad - \sum_{\tau=0}^{t}\frac{\eta_\tau}{\theta_\tau}D_\psi(z_{\tau+1},z_\tau) + \sum_{\tau=0}^{t}\frac{1}{\theta_\tau}\langle e_\tau, z_{\tau+1}-y_t\rangle \\ &= \sum_{\tau=0}^{t}\frac{1}{\theta_\tau}[f_{\mu_\tau}(y_\tau)+\langle\nabla f_{\mu_\tau}(y_\tau), x^*-y_\tau\rangle + \varphi(x^*)] + \frac{1}{\alpha_{t+1}}\psi(x^*) \\ &\quad - \sum_{\tau=0}^{t}\frac{\eta_\tau}{\theta_\tau}D_\psi(z_{\tau+1},z_\tau) + \sum_{\tau=0}^{t}\frac{1}{\theta_\tau}\langle e_\tau, z_{\tau+1}-x^*\rangle \\ &\le \sum_{\tau=0}^{t}\frac{1}{\theta_\tau}[f_{\mu_\tau}(x^*)+\varphi(x^*)] + \frac{1}{\alpha_{t+1}}\psi(x^*) \\ &\quad - \sum_{\tau=0}^{t}\frac{\eta_\tau}{\theta_\tau}D_\psi(z_{\tau+1},z_\tau) + \sum_{\tau=0}^{t}\frac{1}{\theta_\tau}\langle e_\tau, z_{\tau+1}-x^*\rangle.\end{aligned} \tag{B.6}$$

Now we apply the Fenchel inequality to the conjugates $\frac{1}{2}\|\cdot\|^2$ and $\frac{1}{2}\|\cdot\|_*^2$, yielding

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

In particular,

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

Using this inequality and rearranging (B.6) proves the lemma.

Appendix C. Proof of Lemma 4.5. Consider the sum $\sum_{t=0}^{T-1} \frac{1}{\theta_t} \langle e_t, z_t - x^* \rangle$. 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 $\frac{1}{\theta_t} \langle e_t, z_t - x^* \rangle$ is a martingale difference sequence. By setting $c_t = R\sigma/\theta_t$, we have by Assumption B that

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

By applying Lemma F.8 from Appendix F, we conclude that $\frac{1}{\theta_t} \langle e_t, z_t - x^* \rangle$ is (conditionally) sub-Gaussian with parameter $\sigma_t^2 \le 4R^2\sigma^2/3\theta_t^2$, and applying the Azuma–Hoeffding inequality (see (F.1) in Appendix F) yields

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

for $w \ge 0$. Setting $w = \epsilon/\theta_{T-1}$ yields that

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

Since $\theta_{T-1} \le \theta_t$ for $t < T$, we have $R^2\sigma^2 \sum_{t=0}^{T-1} \frac{\theta_{T-1}^2}{\theta_t^2} \le R^2\sigma^2 \sum_{t=0}^{T-1} 1 = R^2\sigma^2 T$, and dividing ϵ again by θ_{T-1} while 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 in (4.6). The second claim (4.7) follows by setting $\delta = \exp(-\frac{3T\epsilon^2}{2R^2\sigma^2})$.

Appendix D. Proof of Lemma 4.6. 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:

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

For now, we proceed with the proof, returning to establish the claim (D.1) later.

The bound (D.1) implies that X_t is subexponential 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] to conclude that

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

which yields the first claim in Lemma 4.6.

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 obtain $\epsilon = \sqrt{2C^2\log\frac{1}{\delta}}$ or that

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

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

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

with probability at least $1-\delta$, by the bound (D.2).

We now return to prove the intermediate claim (D.1). Let $X := \left\|e_t\right\|_*$. For notational convenience in this paragraph, we take all probabilities and expectations conditional on $\mathcal{F}_{t-1}$. By assumption $\mathbb{E}[\exp(X^2/\sigma^2)] \le \exp(1)$, so for $\lambda \in [0,1]$

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

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 $\sigma^2 - \mathbb{E}[X^2] + \epsilon\sigma^2 \le 2\epsilon\sigma^2$, and 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

$$\text{(D.3)} \qquad \mathbb{P}(|X^2 - \mathbb{E}[X^2]| > \epsilon) \le 2\exp(1)\exp\left(-\frac{\epsilon}{2\sigma^2}\right).$$

Now we apply Lemma F.7 to the bound (D.3) to see that $\left\|e_t\right\|_*^2 - \mathbb{E}[\left\|e_t\right\|_*^2 \mid \mathcal{F}_{t-1}]$ is subexponential with parameters $\Lambda \ge \sigma^2$ and $\tau^2 \le 32e\sigma^4$.

Appendix E. Properties of randomized smoothing. In this section, we discuss the analytic properties of the smoothed function f_μ from the convolution (1.3).

We assume throughout that functions are sufficiently integrable without bothering with measurability conditions (since $F(\cdot;\xi)$ is convex, this entails no real loss of generality [4, 30]). By Fubini's theorem, we have

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

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.

We 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, and similarly for $F(\cdot;\xi)$. That is, we write $\nabla F(x+Z;\xi)$, which exists with probability 1 (see also [4]). We give proofs of the following set of smoothing lemmas in the full version of this paper [10].

LEMMA E.1. *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 $x \in \mathrm{int}(\mathcal{X} + B_\infty(0,u))$ Then the following hold:*

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

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

There exists a function f for which each of the estimates (i)–(iii) *is tight simultaneously, and* (iv) *is tight at least to a factor of* 1/4.

Remarks. 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.

A similar lemma can be proved when μ is the density of the uniform distribution on $B_2(0,u)$. In this case, Yousefian, Nedić, and Shanbhag [37] give (i)–(iii) of the following lemma (though the tightness of the bounds is new).

LEMMA E.2 (Yousefian, Nedić, and Shanbhag [37]). *Let f_μ be defined as in* (1.3), *where μ is the uniform density on the ℓ_2-ball of radius u. Assume $\mathbb{E}[\|\partial F(x;\xi)\|_2^2] \le L_0^2$ for $x \in \mathrm{int}(\mathcal{X} + B_2(0,u))$. Then the following hold:*

(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) *For random variables $Z \sim \mu$ and $\xi \sim P$,*

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

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.*

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 [18, 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 E.3. *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 for P-*a.e. *ξ. 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) *For random variables $Z \sim \mu$ and $\xi \sim P$,*

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

In addition, there exists a function f for which each of the bounds (i)–(iv) *is tight (to within a constant factor) simultaneously.*

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 obtain 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 (uniform and gradient) and then give the lemma:

$$\text{(E.1)} \qquad E_U(f) := \inf\left\{L \in \mathbb{R} \mid \sup_{x\in\mathcal{X}} |f(x) - f_\mu(x)| \le L\right\},$$

$$\text{(E.2)} \qquad E_\nabla(f) := \inf\left\{L \in \mathbb{R} \mid \|\nabla f_\mu(x) - \nabla f_\mu(y)\|_2 \le L \|x - y\|_2 \ \forall\, x, y \in \mathcal{X}\right\}.$$

Lemma E.4. *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*

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

Remarks. Inspecting the convergence guarantee of Theorem 2.1 makes the importance of the above bound clear. The terms L_1 and L_0 in the bound (2.5) 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 (2.5) 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 E.2 and E.3 is sharp and cannot be improved by more than a constant factor (see also Corollaries 2.3 and 2.4).

Appendix F. Sub-Gaussian and subexponential tail bounds. For reference purposes, we state here some standard definitions and facts about sub-Gaussian and subexponential random variables (see the books [7, 21, 34] for further details).

F.1. Sub-Gaussian variables. This class of random variables is characterized by a quadratic upper bound on the cumulant generating function.

Definition F.1. *A zero-mean random variable X is called* sub-Gaussian with parameter σ^2 *if* $\mathbb{E}[\exp(\lambda X)] \le \exp(\sigma^2 \lambda^2/2)$ *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 $|X| \le C$ is sub-Gaussian with parameter $\sigma^2 \le C^2$.

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

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

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

LEMMA F.3. *Let $X \in \mathbb{R}^d$ be a random vector with sub-Gaussian components, each with parameter at most σ^2. Then $\mathbb{E}[\|X\|_\infty^2] \le \max\{6\sigma^2 \log d, 2\sigma^2\}$.*

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 [34, 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 of the Azuma–Hoeffding type (e.g., [2, 14, 7]).

LEMMA F.4. *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 $\mathbb{E}[\exp(\lambda X_i) \mid \mathcal{F}_{i-1}] \le \exp(\lambda^2\sigma_i^2/2)$. Then for all $\epsilon > 0$,*

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

We use martingale techniques to establish the sub-Gaussianity of a normed sum.

LEMMA F.5. *Let $X_1, \ldots, X_n$ be independent random vectors with $\|X_i\| \le L$ 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. Since $\|X_i\| \le L$, the quantity $\|S_n\| - \mathbb{E}[\|S_n\|]$ can be controlled using techniques for bounded martingales [21, Chapter 6]. We begin by constructing the Doob martingale associated with the sequence $\{X_i\}$. Let $\mathcal{F}_0$ be the trivial σ-field, and for $i \ge 1$, let $\mathcal{F}_i$ be the σ-field defined by the random variables $X_1, \ldots, X_i$. Define the real-valued random variables $Z_i = \mathbb{E}[\|S_n\| \mid \mathcal{F}_i] - \mathbb{E}[\|S_n\| \mid \mathcal{F}_{i-1}]$, and note that $\mathbb{E}[Z_i \mid \mathcal{F}_{i-1}] = 0$ by construction. Defining the quantity $S_{n\setminus i} = \sum_{j \ne i} X_j$, we have

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

where we have exploited the fact that X_j is independent of $\mathcal{F}_{i-1}$ for $j \ge i$. Consequently, the variables Z_i define a bounded martingale difference sequence. More precisely, since $|Z_i| \le 2L$, the Z_i are conditionally sub-Gaussian with parameter at most $4L^2$. Thus, the sum $\sum_{i=1}^n Z_i = \|S_n\| - \mathbb{E}[\|S_n\|]$ is sub-Gaussian with parameter at most $4nL^2$, as claimed. □

F.2. Subexponential random variables. A slightly less restrictive tail condition defines the class of subexponential random variables.

Definition F.6. *A zero-mean random variable X is* subexponential *with parameters (Λ, τ) if*

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

The following lemma provides an equivalent characterization of subexponential variable via a tail bound.

Lemma F.7. *Let X be a zero-mean random variable. If there are constants $a, \alpha > 0$ such that*

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

then X is subexponential 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 $\mathbb{E}[|X|^k] = \int_0^\infty \mathbb{P}(|X|^k \geq t)dt$ (for similar results, see Buldygin and Kozachenko [7, Chapter I.3]).

Finally, any random variable whose square is subexponential is sub-Gaussian, as shown by the following result.

Lemma F.8. (Lan, Nemirovski, and Shapiro [20, Lemma 2]) *Let X be a zero-mean random variable satisfying the moment generating inequality $\mathbb{E}[\exp(X^2/\sigma^2)] \leq \exp(1)$. Then X is sub-Gaussian with parameter at most $3/2\sigma^2$.*

Acknowledgments. We thank Alekh Agarwal and Ohad Shamir for discussions about variance reduction in stochastic optimization and Steven Evans for some useful pointers on smoothing operators. We also acknowledge the anonymous reviewers and the editor for their helpful feedback and thoughtful comments.

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 Trans. Inform. Theory, 58 (2012), pp. 3235–3249.

[2] K. Azuma, *Weighted sums of certain dependent random variables*, Tohoku Math. J., 68 (1967), pp. 357–367.

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

[4] D. P. Bertsekas, *Stochastic optimization problems with nondifferentiable cost functionals*, J. Optim. Theory Appl., 12 (1973), pp. 218–231.

[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 Comput. Math. Math. Phys., 7 (1967), pp. 200–217.

[6] P. Brucker, *An $O(n)$ algorithm for quadratic knapsack problems*, Oper. Res. Lett., 3 (1984), pp. 163–166.

[7] V. Buldygin and Y. Kozachenko, *Metric Characterization of Random Variables and Random Processes*, Transl. Math. Monogr. 188, AMS, Providence, RI, 2000.

[8] C. Chen and O. L. Mangasarian, *A class of smoothing functions for nonlinear and mixed complementarity problems*, Comput. Optim. Appl., 5 (1996), pp. 97–138.

[9] O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao, *Optimal distributed online prediction using mini-batches*, J. Mach. Learn. Res., 13 (2012), pp. 165–202.

[10] J. C. Duchi, P. L. Bartlett, and M. J. Wainwright, *Randomized Smoothing for Stochastic Optimization*, preprint, http://arxiv.org/abs/1103.4296, 2011.

[11] 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.

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

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

[14] W. Hoeffding, *Probability inequalities for sums of bounded random variables*, J. Amer. Statist. Assoc., 58 (1963), pp. 13–30.

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

[16] A. Juditsky, A. Nemirovski, and C. Tauvel, *Solving Variational Inequalities with the Stochastic Mirror-Prox Algorithm*, preprint, http://arxiv.org/abs/0809.0815, 2008.

[17] V. Katkovnik and Y. Kulchitsky, *Convergence of a class of random search algorithms*, Automat. Remote Control, 33 (1972), pp. 1321–1326.

[18] H. Lakshmanan and D. P. de Farias, *Decentralized resource allocation in dynamic networks of agents*, SIAM J. Optim., 19 (2008), pp. 911–940.

[19] G. Lan, *An optimal method for stochastic composite optimization*, Math. Program., 133 (2012), pp. 365–397.

[20] G. Lan, A. Nemirovski, and A. Shapiro, *Validation analysis of robust stochastic approximation method*, Math. Program., to appear; also available online from http://www.ise.ufl.edu/glan/files/2011/12/Validate-SA-rev-Oct28-2010-final.pdf.

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

[22] C. Lemaréchal and C. Sagastizábal, *Practical aspects of the Moreau–Yosida regularization: Theoretical preliminaries*, SIAM J. Optim., 7 (1997), pp. 367–385.

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

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

[25] A. Nemirovski and D. Yudin, *Problem complexity and method efficiency in optimization*, Wiley, New York, 1983.

[26] Y. Nesterov, *Smooth minimization of nonsmooth functions*, Math. Program., 103 (2005), pp. 127–152.

[27] Y. Nesterov, *Primal-dual subgradient methods for convex problems*, Math. Program., 120 (2009), pp. 261–283.

[28] B. T. Polyak and J. Tsypkin, *Robust identification*, Automatica J. IFAC, 16 (1980), pp. 53–63.

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

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

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

[32] 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.

[33] P. Tseng, *On Accelerated Proximal Gradient Methods for Convex-Concave Optimization*, http://www.math.washington.edu/∼tseng/papers/apgm.pdf, 2008.

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

[35] L. Xiao, *Dual averaging methods for regularized stochastic learning and online optimization*, J. Mach. Learn. Res., 11 (2010), pp. 2543–2596.

[36] E. Xing, A. Ng, M. Jordan, and S. Russell, *Distance Metric Learning, with Application to Clustering with Side-Information*, Adv. Neural Inf. Process. Syst. 15, MIT Press, Cambridge, MA, 2003.

[37] F. Yousefian, A. Nedić, and U. V. Shanbhag, *On stochastic gradient and subgradient methods with adaptive steplength sequences*, Automatica J. IFAC, 48 (2012), pp. 56–67.