ERGODIC MIRROR DESCENT

JOHN C. DUCHI*, ALEKH AGARWAL†, MIKAEL JOHANSSON‡, AND MICHAEL I. JORDAN*§

Abstract. We generalize stochastic subgradient descent methods to situations in which we do not receive independent samples from the distribution over which we optimize, instead receiving samples coupled over time. We show that as long as the source of randomness is suitably ergodic—it converges quickly enough to a stationary distribution—the method enjoys strong convergence guarantees, both in expectation and with high probability. This result has implications for stochastic optimization in high-dimensional spaces, peer-to-peer distributed optimization schemes, decision problems with dependent data, and stochastic optimization problems over combinatorial spaces.

1. Introduction. In this paper, we analyze a new algorithm, Ergodic Mirror Descent, for solving a class of stochastic optimization problems. We begin with a statement of the problem. Let $\{F(\cdot;\xi), \xi \in \Xi\}$ be a collection of closed convex functions whose domains contain the common closed convex set $\mathcal{X} \subseteq \mathbb{R}^d$. Let Π be a probability distribution over the statistical sample space Ξ and consider the convex function $f : \mathcal{X} \to \mathbb{R}$ defined by the expectation

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

We study algorithms for solving the following problem:

$$\operatorname*{minimize}_x \; f(x) \quad \text{subject to} \quad x \in \mathcal{X}. \tag{1.2}$$

A wide variety of stochastic optimization methods for solving the problem (1.2) have been explored in an extensive literature [34, 31, 28, 21, 29]. We study procedures that do not assume it is possible to receive samples from the distribution Π, instead receiving samples ξ from a stochastic process P indexed by time t, where the stochastic process P converges to the stationary distribution Π. This is a natural relaxation, because in many circumstances the distribution Π is not even known—for example in statistical applications—and we cannot receive independent samples. In other scenarios, it may be hard to even draw samples from Π efficiently, such as when Ξ is a high-dimensional or combinatorial space, but it is possible [19] to design Markov chains that converge to the distribution Π. Further, in computational applications, it is often unrealistic to assume that one actually has access to a source of independent randomness, so studying the effect of correlation is natural and important [18].

Our approach to solving the problem (1.2) is related to classical stochastic gradient descent algorithms [34, 31], where one assumes access to samples ξ from the distribution Π and performs gradient updates using $\nabla F(x;\xi)$. When Π is concentrated on a set of n points and the functions F are not necessarily differentiable,

[*] Department of Electrical Engineering and Computer Sciences, University of California, Berkeley; Berkeley, CA USA. Email: {jduchi,jordan}@eecs.berkeley.edu. JCD was supported by an NDSEG fellowship. Supported in part by the U.S. Army Research Laboratory and the U.S. Army Research Office under contract/grant number W911NF-11-1-0391.

[†] Microsoft Research, New York, NY USA. Email: alekha@microsoft.com. Supported by Microsoft Research and Google Graduate fellowships.

[‡] School of Electrical Engineering, Royal Institute of Technology (KTH); Stockholm, Sweden. Email: mikael.johansson@ee.kth.se

[§] Department of Statistics, University of California, Berkeley; Berkeley, CA USA.

the objective is of the form $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$ and the incremental subgradient method of Nedić and Bertsekas [28] applies. More generally, our problem belongs to the family of stochastic problems with exogenous correlated noise [21] where the goal is to minimize $\mathbb{E}_\Pi[F(x;\xi)]$ as in the objective (1.2), but we have access only to samples ξ that are not independent over time. Certainly a number of researchers in control, optimization, stochastic approximation, and statistics have studied settings where stochastic data is not i.i.d. (see, for example, the books [21, 36] and the numerous references therein). Nonetheless, classical results in this setting are asymptotic in nature and generally do not provide finite sample or high-probability convergence guarantees; our work provides such results.

Our method borrows from standard stochastic subgradient and stochastic mirror descent methodology [30, 29], but we generalize this work in that we receive samples not from the distribution Π but from an ergodic process $\xi_1, \xi_2, \ldots$ converging to the stationary distribution Π. In spite of the new setting, we do not modify standard stochastic subgradient algorithms; our algorithm receives samples ξ_t and takes mirror descent steps with respect to the subgradients of $F(x;\xi_t)$. Consequently, our approach generalizes several recent works on stochastic and non-stochastic optimization, including the randomized incremental subgradient method [28] as well as the Markov incremental subgradient method [20, 33]. There are a number of applications of this work: in control problems, data is often coupled over time or may come from an autoregressive process [21]; in distributed sensor networks [22], a set of wireless sensors attempt to minimize an objective corresponding to a sequence of correlated measurements; and in statistical problems, data comes from an unknown distribution and may be dependent [39]. See our examples and experiments in § 4 and § 5, as well as the examples in the paper by Ram et al. [33], for other motivating applications.

The main result of this paper is that performing stochastic gradient or mirror descent steps as described in the previous paragraph is a provably convergent optimization procedure. The convergence is governed by problem-dependent quantities (namely the radius of $\mathcal{X}$ and the Lipschitz constant of the functions F) familiar from previous results on stochastic methods [28, 40, 29] and also depends on the rate at which the stochastic process $\xi_1, \xi_2, \ldots$ converges to its stationary distribution. Our three main convergence theorems characterize the convergence rate of Ergodic Mirror Descent in terms of the mixing time τ_{mix} (the time it takes the process ξ_t to converge to the stationary distribution Π, in a sense we make precise later) in expectation, with high probability, and when the mixing times of the process are themselves random. In particular, we show that this rate is $\mathcal{O}\left(\sqrt{\frac{\tau_{\text{mix}}}{T}}\right)$ for a large class of ergodic processes, both in expectation and with high probability. We also give a lower bound that shows that our results are tight: they cannot (in general) be improved by more than numerical constants.

The remainder of the paper is organized as follows. Section 2 contains our main assumptions and a description of the algorithm. Following that, we collect our main technical results in § 3. We expand on these results in example corollaries throughout § 4 and give numerical simulations exploring our algorithms in § 5. We provide complete proofs of all our results in § 6 and the appendices.

Notation. For the reader's convenience, we collect our (standard) notation here. A function f is G-Lipschitz with respect to a norm $\|\cdot\|$ if $|f(x) - f(y)| \le G\|x - y\|$. The dual norm $\|\cdot\|_*$ to a norm $\|\cdot\|$ is defined by $\|z\|_* := \sup_{\|x\|\le 1}\langle z, x\rangle$. A function ψ

is strongly convex with respect to the norm $\|\cdot\|$ over the domain $\mathcal{X}$ if

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

For a convex function f, we let $\partial f(x) = \{g \in \mathbb{R}^d \mid f(y) \ge f(x) + \langle g, y-x\rangle\}$ denote its subdifferential. For a matrix $A \in \mathbb{R}^{n\times m}$, we let $\rho_i(A)$ denote its ith largest singular value, and when $A \in \mathbb{R}^{n\times n}$ is symmetric we let $\lambda_i(A)$ denote its ith largest eigenvalue. The all-ones vector is $\mathbb{1}$, and we denote the transpose of the matrix A by $A^\top$. We let $[n]$ denote the set $\{1,\ldots,n\}$. For functions f and g, we write $f(n) = \mathcal{O}(g(n))$ if there exist $N < \infty$ and $C < \infty$ such that $f(n) \le Cg(n)$ for $n \ge N$, and $f(n) = \Omega(g(n))$ if there exist $N < \infty$ and $c > 0$ such that $f(n) \ge cg(n)$ for $n \ge N$. For a probability measure P and measurable set or event A, $P(A)$ denotes the mass P assigns A.

2. Assumptions and algorithm. We now turn to describing our algorithm and the assumptions underlying it. We begin with a description of the algorithm, which is familiar from the literature on mirror descent algorithms [30, 3]. Specifically, we generalize the stochastic mirror descent algorithm [30, 29], which in turn generalizes gradient descent to elegantly address non-Euclidean geometry. The algorithm is based on a prox-function ψ, a differentiable convex function defined on $\mathcal{X}$ assumed (w.l.o.g. by scaling) to be 1-strongly convex with respect to the norm $\|\cdot\|$ over $\mathcal{X}$. The Bregman divergence D_ψ generated by ψ is defined as

$$D_\psi(x,y) := \psi(x) - \psi(y) - \langle \nabla\psi(y), x-y\rangle \ge \frac{1}{2}\|x-y\|^2. \tag{2.1}$$

We assume $\mathcal{X}$ is compact and that there exists a radius $R < \infty$ such that

$$D_\psi(x,y) \le \frac{1}{2}R^2 \quad \text{for } x,y \in \mathcal{X}. \tag{2.2}$$

The Ergodic Mirror Descent (EMD) algorithm is an iterative algorithm that maintains a parameter $x(t) \in \mathcal{X}$, which it updates using stochastic gradient information to form $x(t+1)$. Let the time-indexed sequence $(\xi_1, \xi_2, \ldots, \xi_t, \ldots)$ represent a draw from P. At time t, given ξ_t, EMD computes the update

$$g(t) \in \partial F(x(t);\xi_t), \quad x(t+1) = \mathop{\mathrm{argmin}}_{x\in\mathcal{X}} \left\{ \langle g(t), x\rangle + \frac{1}{\alpha(t)} D_\psi(x, x(t)) \right\}. \tag{2.3}$$

The initial point $x(1)$ may be selected arbitrarily in $\mathcal{X}$, and here $\alpha(t)$ is a non-increasing (time-dependent) stepsize. The algorithm (2.3) reduces to projected gradient descent with the choice $\psi(x) = \frac{1}{2}\|x\|_2^2$, since then $D_\psi(x,y) = \frac{1}{2}\|x-y\|_2^2$.

Our main assumption on the functions $F(\cdot;\xi)$ regards their continuity and subdifferentiability properties, though we require a bit more notation. Let $\mathsf{G}(x;\xi) \in \partial F(x;\xi)$ denote a fixed and measurable element of the subgradient of $F(\cdot;\xi)$ evaluated at the point x, where (without loss of generality) we assume that in the EMD algorithm (2.3) we have $g(t) = \mathsf{G}(x(t);\xi_t)$. We let $\mathcal{F}_t$ denote the σ-field of the first t random samples, $\xi_1, \ldots, \xi_t$, drawn from the stochastic process P. We make one of the following two assumptions, where in each the norm $\|\cdot\|$ is the norm with respect to which ψ is strongly convex (2.1):

Assumption A (Finite single-step variance). *Let x be measurable with respect to the σ-field $\mathcal{F}_{t-1}$. There exists a constant $G < \infty$ such that with probability 1*

$$\mathbb{E}[\|\mathsf{G}(x;\xi_t)\|_*^2 \mid \mathcal{F}_{t-1}] \le G^2.$$

Assumption B. *For Π-almost every ξ, the functions $F(\cdot;\xi)$ are G-Lipschitz continuous functions with respect to a norm $\|\cdot\|$ over $\mathcal{X}$. That is,*

$$|F(x;\xi) - F(y;\xi)| \le G\,\|x-y\| \quad \textit{for } x,y \in \mathcal{X}.$$

As a consequence of Assumption B, for any $g \in \partial F(x;\xi)$ we have that $\|g\|_* \le G$ (e.g., [17]), and it is clear that the expected function f is also G-Lipschitz. Assumption B implies Assumption A, though Assumption A still guarantees f is G-Lipschitz, and under either assumption, we have

$$\mathbb{E}\left[\|\mathsf{G}(x;\xi)\|_*^2\right] = \mathbb{E}\left[\mathbb{E}\left[\|\mathsf{G}(x;\xi)\|_*^2 \mid \mathcal{F}_{t-1}\right]\right] \le G^2. \tag{2.4}$$

Having described the family of functions $\{F(\cdot;\xi) : \xi \in \Xi\}$, we recall a few definitions from probability theory that are essential to the presentation of our results. We measure the convergence of the stochastic process P using one of two common statistical distances [13]: the Hellinger distance and the total variation distance (our definitions differ by a factor of two from some definitions of these metrics). The total variation distance between probability distributions P and Q defined on a set Ξ, assumed to have densities p and q with respect to an underlying measure μ,[1] is

$$d_{\rm TV}(P,Q) := \int_\Xi |p(\xi) - q(\xi)| d\mu(\xi) = 2\sup_{A\subset\Xi} |P(A) - Q(A)|, \tag{2.5}$$

the supremum taken over measurable subsets of Ξ. The squared Hellinger distance is

$$d_{\rm hel}(P,Q)^2 = \int_\Xi \left(\sqrt{\frac{p(\xi)}{q(\xi)}} - 1\right)^2 q(\xi)d\mu(\xi) = \int_\Xi \left(\sqrt{p(\xi)} - \sqrt{q(\xi)}\right)^2 d\mu(\xi). \tag{2.6}$$

It is a well-known fact [13] that for any probability distributions P and Q,

$$d_{\rm hel}(P,Q)^2 \le d_{\rm TV}(P,Q) \le 2 d_{\rm hel}(P,Q). \tag{2.7}$$

Using the total variation (2.5) and Hellinger (2.6) metrics, we now describe our notion of mixing (convergence) of the stochastic process P. Recall our definition of the σ-field $\mathcal{F}_t = \sigma(\xi_1,\ldots,\xi_t)$. Let $P_{[s]}^t$ denote the distribution of ξ_t conditioned on $\mathcal{F}_s$ (i.e., given the initial samples $\xi_1,\ldots,\xi_s$), so for measurable $A \subset \Xi$ we have $P_{[s]}^t(A) := P(\xi_t \in A \mid \mathcal{F}_s)$. We measure convergence of P to Π in terms of the mixing time of the different $P_{[s]}^t$, defined for the Hellinger and total variation distances as follows. In the definitions, let $p_{[s]}^t$ and π denote the densities of $P_{[s]}^t$ and Π, respectively.

Definition 2.1. *The* total variation mixing time $\tau_{\rm TV}(P_{[s]},\epsilon)$ *of the process P conditioned on the σ-field of the initial s samples $\mathcal{F}_s = \sigma(\xi_1,\ldots,\xi_s)$ is the smallest $t \in \mathbb{N}$ such that $d_{\rm TV}(P_{[s]}^{s+t},\Pi) \le \epsilon$,*

$$\tau_{\rm TV}(P_{[s]},\epsilon) := \inf\left\{t - s : t\in\mathbb{N},\ \int_\Xi \left|p_{[s]}^t(\xi) - \pi(\xi)\right| d\mu(\xi) \le \epsilon\right\}.$$

The Hellinger mixing time $\tau_{\rm hel}(P_{[s]},\epsilon)$ *is the smallest t such that $d_{\rm hel}(P_{[s]}^{s+t},\Pi) \le \epsilon$,*

$$\tau_{\rm hel}(P_{[s]},\epsilon) := \inf\left\{t - s : t\in\mathbb{N},\ \int_\Xi \left(\sqrt{p_{[s]}^t(\xi)} - \sqrt{\pi(\xi)}\right)^2 d\mu(\xi) \le \epsilon^2\right\}.$$

[1] This is no loss of generality, since P and Q are absolutely continuous with respect to $P+Q$.

Put another way, the mixing times $\tau_{\text{TV}}(P_{[s]}, \epsilon)$ and $\tau_{\text{hel}}(P_{[s]}, \epsilon)$ are the number of *additional* steps required until the distribution of ξ_t is close to the stationary distribution Π given the initial s samples $\xi_1, \ldots, \xi_s$.

The following assumption, which makes the mixing times of the stochastic process P uniform, is our main probabilistic assumption.

Assumption C. *The mixing times of the stochastic process $\{\xi_i\}$ are uniform in the sense that there exist uniform mixing times $\tau_{\text{TV}}(P, \epsilon), \tau_{\text{hel}}(P, \epsilon) < \infty$ such that with probability* 1,

$$\tau_{\text{TV}}(P, \epsilon) \geq \tau_{\text{TV}}(P_{[s]}, \epsilon) \quad \textit{and} \quad \tau_{\text{hel}}(P, \epsilon) \geq \tau_{\text{hel}}(P_{[s]}, \epsilon)$$

for all $\epsilon > 0$ and $s \in \mathbb{N}$.

Assumption C is a weaker version of the common assumption of ϕ-mixing in the probability literature (e.g. [9]); ϕ-mixing requires convergence of the process over the entire "future" σ-field $\sigma(\xi_t, \xi_{t+1}, \ldots)$ of the process ξ_t. Any finite state-space time-homeogeneous Markov chain satisfies the above assumption, as do uniformly ergodic Markov chains on general state spaces [26].

We remark that the definition 2.1 of mixing time does not assume that the distributions $P_{[s]}$ are time-homogeneous. Indeed, Assumption C requires only that there exists a uniform upper bound on the mixing times. We can weaken Assumption C to allow randomness in the probability distributions $P^t_{[s]}$ themselves, that is, conditional on $\mathcal{F}_s$, the mixing time $\tau_{\text{TV}}(P_{[s]}, \epsilon)$ is an $\mathcal{F}_s$-measurable random variable. Our weakened probabilistic assumption is

Assumption D. *The mixing times of the stochastic process $\{\xi_i\}$ are stochastically uniform in the sense that there exists a uniform mixing time $\tau_{\text{TV}}(P, \epsilon) < \infty$, continuous from the right as a function of ϵ, such that for all $\epsilon > 0$, $s \in \mathbb{N}$, and $c \in \mathbb{R}$*

$$P\big(\tau_{\text{TV}}(P_{[s]}, \epsilon) \geq \tau_{\text{TV}}(P, \epsilon) + \kappa c\big) \leq \exp(-c).$$

Assumption D allows us to provide convergence guarantees for a much wider range of processes, such as auto-regressive processes, than permitted by Assumption C.

3. Main results.

With our assumptions in place, we can now give our main results. We begin with three general theorems that guarantee the convergence of the EMD algorithm in expectation and with high probability. The second part of the section shows that our analysis is sharp—unimprovable by more than numerical constant factors—by giving an information-theoretic lower bound on the convergence rate of any optimization procedure receiving non-i.i.d. samples from P.

3.1. Convergence guarantees.

Our first result gives convergence in expectation of the EMD algorithm (2.3); we provide the proof in § 6.2.

Theorem 3.1. *Let Assumption C hold and let $x(t)$ be defined by the EMD update* (2.3) *with non-increasing stepsize sequence $\{\alpha(t)\}$. Let $x^\star \in \mathcal{X}$ be arbitrary and let* (2.2) *hold. If Assumption A holds, then for any $\epsilon > 0$,*

$$\mathbb{E}\Big[\sum_{t=1}^T (f(x(t)) - f(x^\star))\Big] \\ \leq \frac{R^2}{2\alpha(T)} + \frac{G^2}{2}\sum_{t=1}^T \alpha(t) + 3T\epsilon GR + (\tau_{\text{hel}}(P, \epsilon) - 1)\Big[G^2 \sum_{t=1}^T \alpha(t) + RG\Big],$$

while if Assumption B holds, then for any $\epsilon > 0$,

$$\mathbb{E}\biggl[\sum_{t=1}^{T}(f(x(t)) - f(x^\star))\biggr]$$

$$\leq \frac{R^2}{2\alpha(T)} + \frac{G^2}{2}\sum_{t=1}^{T}\alpha(t) + T\epsilon GR + (\tau_{\rm TV}(P,\epsilon) - 1)\biggl[G^2\sum_{t=1}^{T}\alpha(t) + RG\biggr].$$

The expectation in both bounds is taken with respect to the samples $\xi_1, \ldots, \xi_T$.

We obtain an immediate corollary to Theorem 3.1 by applying Jensen's inequality to the convex function f:

Corollary 3.2. *Define $\widehat{x}(T) = \frac{1}{T}\sum_{t=1}^{T} x(t)$ and let the conditions of Theorem 3.1 hold. If Assumption A holds, then for any $\epsilon > 0$*

$$\mathbb{E}[f(\widehat{x}(T)) - f(x^\star)] \leq \frac{R^2}{2\alpha(T)T} + \frac{G^2}{2T}\sum_{t=1}^{T}\alpha(t) + 3\epsilon GR + \frac{\tau_{\rm hel}(P,\epsilon) - 1}{T}\biggl[G^2\sum_{t=1}^{T}\alpha(t) + RG\biggr].$$

If Assumption B holds, then for any $\epsilon > 0$

$$\mathbb{E}[f(\widehat{x}(T)) - f(x^\star)] \leq \frac{R^2}{2\alpha(T)T} + \frac{G^2}{2T}\sum_{t=1}^{T}\alpha(t) + \epsilon GR + \frac{\tau_{\rm TV}(P,\epsilon) - 1}{T}\biggl[G^2\sum_{t=1}^{T}\alpha(t) + RG\biggr].$$

Corollary 3.2 shows that so long as the stepsize sequence $\alpha(t)$ is non-increasing and satisfies the asymptotic conditions $T\alpha(T) \to \infty$ and $(1/T)\sum_{t=1}^{T}\alpha(t) \to 0$, the EMD method converges. We can also provide similar high-probability convergence guarantees:

Theorem 3.3. *Let the conditions of Theorem 3.1 and Assumption B hold. Let $\delta \in (0,1)$ and define the average $\widehat{x}(T) = \frac{1}{T}\sum_{t=1}^{T} x(t)$. With probability at least $1 - \delta$, for $\epsilon > 0$ such that $\tau_{\rm TV}(P,\epsilon) \leq T/2$,*

$$f(\widehat{x}(T)) - f(x^\star) \leq \frac{R^2}{2T\alpha(T)} + \frac{G^2}{2T}\sum_{t=1}^{T}\alpha(t) + \frac{\tau_{\rm TV}(P,\epsilon) - 1}{T}\biggl[G^2\sum_{t=1}^{T}\alpha(t) + GR\biggr]$$
$$+ \epsilon GR + 4GR\sqrt{\frac{\tau_{\rm TV}(P,\epsilon)\log\frac{\tau_{\rm TV}(P,\epsilon)}{\delta}}{T}}.$$

We provide the proof of this theorem in § 6.3. Note that the rate of convergence in Theorem 3.3 is identical to that obtained in Theorem 3.1 plus an additional term that arises as a result of the control of the deviation of the ergodic process around its expectation. The additional $\log\frac{1}{\delta}$-dependent term arises from the application of martingale concentration inequalities [2], which requires some care because the process $\{\xi_t\}$ is coupled over time. Nonetheless, as we discuss briefly following Corollary 3.5—and as made clear by our lower bound in Theorem 3.7—the additional terms introduce a factor of at most $\sqrt{\log\tau_{\rm TV}(P,\epsilon)}$ to the bounds. That is, the dominant terms in the convergence rates (modulo logarithmic factors) also appear in the expected bounds in Theorem 3.1.

The last of our convergence theorems extends the previous two to the case when the stochastic process is not uniformly mixing, but has mixing properties that may depend on its state. We provide the proof of Theorem 3.4 in § 6.4.

Theorem 3.4. *Let the conditions of Theorem 3.3 hold, except that we replace the uniform mixing assumption C with the probabilistic mixing assumption D. Let* $\delta \in (0,1)$*. In the notation of Assumption D, define*

$$\tau(\epsilon, \delta) := \tau_{\rm TV}(P, \epsilon) + \kappa \left(\log \frac{2}{\delta} + 2 \log(T) \right).$$

With probability at least $1 - \delta$*, for any* $x^\star \in \mathcal{X}$*,*

$$f(\widehat{x}(T)) - f(x^\star) \le \inf_{\epsilon > 0} \left\{ \frac{R^2}{2T\alpha(T)} + \frac{G^2}{2T} \sum_{t=1}^T \alpha(t) + \frac{\tau(\epsilon,\delta) - 1}{T} \left[G^2 \sum_{t=1}^T \alpha(t) + GR \right] \right.$$
$$\left. + \epsilon GR + 4GR \sqrt{\frac{\tau(\epsilon,\delta) \log \frac{\tau(\epsilon,\delta)}{\delta}}{T}} \right\}.$$

In § 4.2 we give two applications of Theorem 3.4 (to estimation in autoregressive processes and a fault-tolerant distributed optimization scheme) that show how it significantly increases the range of applicability of our framework.

We now turn to a slight specialization of our bounds to build intuition and attain a simplified statement of convergence rates. Theorems 3.1, 3.3, and 3.4 hold for essentially any ergodic process that converges to the stationary distribution Π. For a large class of processes, the convergence of the distributions P^t to the stationary distribution Π is uniform and at a geometric rate [26]: there exist constants κ_1 and κ_2 such that $\tau_{\rm TV}(P, \epsilon) \le \kappa_1 \log(\kappa_2/\epsilon)$. We have the following corollary for this special case; we only present the version yielding expected convergence rates, as the high-probability corollary is similar. In addition, by the fact (2.7) relating $d_{\rm hel}$ to $d_{\rm TV}$, if the process P satisfies $\tau_{\rm TV}(P, \epsilon) \le \kappa_1 \log(\kappa_2/\epsilon)$, then there exist constants κ_1' and κ_2' such that $\tau_{\rm hel}(P, \epsilon) \le \kappa_1' \log(\kappa_2'/\epsilon)$. Thus we only state the corollary for total variation mixing and under Assumption B; an analogous result holds under Assumption A for mixing with respect to the Hellinger distance.

Corollary 3.5. *Under the conditions of Theorem 3.1, assume in addition that* $\tau_{\rm TV}(P, \epsilon) \le \kappa_1 \log(\kappa_2/\epsilon)$ *and let Assumption B hold. The EMD update* (2.3) *with stepsize* $\alpha(t) = \alpha/\sqrt{t}$ *satisfies*

$$\mathbb{E}\left[f(\widehat{x}(T)) - f(x^\star)\right] \le \frac{R^2}{2\alpha\sqrt{T}} + \frac{2\alpha G^2}{\sqrt{T}} \left(\kappa_1 \log \frac{\kappa_2}{\epsilon} \right) + \epsilon GR + \frac{RG\kappa_1 \log \frac{\kappa_2}{\epsilon}}{T}.$$

Proof. Using the definition $\alpha(t) = \alpha/\sqrt{t}$ and the integral bound

$$\sum_{t=1}^T \frac{1}{\sqrt{t}} \le 1 + \int_1^T t^{-1/2} dt = 2\sqrt{T} - 1 < 2\sqrt{T}, \tag{3.1}$$

we have $\sum_{t=1}^T \alpha(t) \le 2\alpha\sqrt{T}$. The corollary now follows from Theorem 3.1. □

We can obtain a simplified convergence rate with appropriate choice of the stepsize multiplier α and mixing parameter ϵ: choosing $\alpha = R/(G\sqrt{\kappa_1 \log(\kappa_2 T)})$ and $\epsilon = T^{-1/2}$ reduces the corollary to

$$\mathbb{E}[f(\widehat{x}(T)) - f(x^\star)] = \mathcal{O}\left(\frac{RG\sqrt{\kappa_1 \log(\kappa_2 T)}}{\sqrt{T}} \right). \tag{3.2}$$

More generally, using the stepsize $\alpha(t) = \alpha/\sqrt{t}$ and the same argument as in Corollary 3.5 gives

$$\mathbb{E}[f(\widehat{x}(T)) - f(x^\star)] \le \inf_{\epsilon > 0} \left\{ \frac{R^2}{2\alpha\sqrt{T}} + \frac{2\alpha G^2}{\sqrt{T}} \tau_{\rm TV}(P, \epsilon) + \epsilon G R + \frac{RG(\tau_{\rm TV}(P, \epsilon) - 1)}{T} \right\}. \tag{3.3}$$

Again choosing $\epsilon = T^{-1/2}$ and defining the shorthand $\tau_{\rm mix} = \tau_{\rm TV}(P, T^{-1/2})$, by choosing $\alpha = R/(G\sqrt{\tau_{\rm mix}})$, we see the bound (3.3) implies that

$$\mathbb{E}[f(\widehat{x}(T)) - f(x^\star)] \le \frac{5RG}{2} \cdot \frac{\sqrt{\tau_{\rm mix}}}{\sqrt{T}} + \frac{RG}{\sqrt{T}} + \frac{RG(\tau_{\rm mix} - 1)}{T}. \tag{3.4}$$

In the classical setting [29] of i.i.d. samples, $\xi_t \sim \Pi$, stochastic gradient descent and its mirror descent generalizations attain convergence rates of $\mathcal{O}(RG/\sqrt{T})$. Since $\tau_{\rm TV}(P, 0) = \tau_{\rm hel}(P, 0) = 1$ for an i.i.d. process, the rate (3.3) shows that our results subsume existing results for i.i.d. noise. Moreover, they are sharp in the i.i.d. case, that is, unimprovable by more than a numerical constant factor [30, 1].

In addition, we note that the conclusions of Corollary 3.5 (and the bound (3.3)) hold—modulo an additional $\log \tau_{\rm TV}(P, \epsilon)$—with high probability. We may also note that replacing ϵGR with $3\epsilon GR$ and $\tau_{\rm TV}$ with $\tau_{\rm hel}$ in the bound (3.3) yields a guarantee under Assumption A. Further, the step-size choice $\alpha(t) = \alpha/\sqrt{t}$ is robust—in a way similarly noted by Nemirovski et al. [29]—for quickly mixing ergodic processes. Indeed, using the inequalities (3.3) and (3.4), we see that setting the multiplier $\alpha = \gamma R/(G\sqrt{\tau_{\rm mix}})$ yields $\mathbb{E}[f(\widehat{x}(T)) - f(x^*)] = \mathcal{O}(\max\{\gamma, \gamma^{-1}\} RG\sqrt{\tau_{\rm mix}}/\sqrt{T})$, so misspecification of α by a constant γ leads to a penalty in convergence that scales at worst linearly in $\max\{\gamma^{-1}, \gamma\}$. In classical stochastic approximation settings [34, 36, 21], one usually chooses step size sequence $\alpha(t) = \mathcal{O}(t^{-m})$ for $m \in (.5, 1]$; in our case, such choices may yield sub-optimal rates because we study convergence of the averaged parameter $\widehat{x}(T)$ rather than the final parameter $x(t)$. Nonetheless, averaging is known to yield robustness in i.i.d. settings [31, 29], and moreover gives unimprovable convergence rates in many cases (see § 3.2 as well as references [30, 1]). We provide some evidence of this robustness in numerical simulations in § 5, and we see generally that EMD has qualitative convergence behavior similar to stochastic mirror descent for a broad class of ergodic processes.

Before continuing, we make two final remarks. First, none of our main theorems assume Markovianity or even homogeneity of the stochastic process P; all that is needed is that the mixing time $\tau_{\rm TV}$ (or $\tau_{\rm hel}$) exists, or even that it exists only with some reasonably high probability. Previous work similar to ours [33, 20] assumes Markovianity (see also our discussion concluding § 4.2). Further, general ergodic processes do not always enjoy the geometric mixing assumed in Corollary 3.5, satisfying either Assumption D's probabilistic mixing condition or simply mixing more slowly. In § 4.2, we present examples of such probabilistically mixing processes on general state spaces, while the bound (3.3) suggests an approach to attain convergence for more slowly mixing processes (see § 4.3).

3.2. Lower bounds and optimality guarantees. Our final main result concerns the optimality of the results we have presented. Informally, the theorem states that our results are unimprovable by more than numerical constant factors, though making this formal requires additional notation. In the stochastic gradient oracle model of convex optimization [30, 1], a method $\mathcal{M}$ issues queries of the form $x \in \mathcal{X}$ to an oracle that returns noisy function and gradient information. In our setting, the oracle is represented by the pair $\theta = (P, \mathsf{G})$, and when the oracle is queried at a point x at time t (i.e., this is the tth query θ has received), it draws a sample ξ_t according to the distribution $P(\cdot \mid \xi_1, \ldots, \xi_{t-1})$ and returns $\mathsf{G}(x, \xi_t) \in \mathbb{R}^d$. The method issues a sequence of queries $x(1), \ldots, x(t)$ to the oracle and may use $\{\mathsf{G}(x(1), \xi_1), \ldots, \mathsf{G}(x(t), \xi_t)\}$ to devise a new query point $x(t+1)$. For an oracle θ, we define the error of the method $\mathcal{M}$ on a function f after T queries of the oracle as

$$\epsilon_T(\mathcal{M}, f, \mathcal{X}, \theta) = f(\widehat{x}) - \inf_{x \in \mathcal{X}} f(x), \tag{3.5}$$

where $\widehat{x}$ denotes the method $\mathcal{M}$'s estimate of the minimizer of f after seeing the T samples $\{\mathsf{G}(x(1), \xi_1), \ldots, \mathsf{G}(x(T), \xi_T)\}$. The quantity (3.5) is random, so we measure accuracy in terms of the expected value $\mathbb{E}_\theta[\epsilon_T(\mathcal{M}, f, \mathcal{X}, \theta)]$, where the expectation is taken with respect to the randomness in θ.

Now we define a natural collection of stochastic oracles for our dependent setting.

Definition 3.6. *For f convex, $\tau \in \mathbb{N}$, $G \in (0, \infty)$, and $p \in [1, \infty]$, the* admissible oracle set $\Theta\left(f, \tau, G, p\right)$ *is the set of oracles $\theta = (P, \mathsf{G})$ for which there exists a probability distribution Π on ξ such that*

$$\begin{aligned} &\left\|\mathsf{G}(x;\xi)\right\|_p \le G \text{ for } x \in \mathcal{X} \text{ and } \xi \in \Xi, \quad \mathbb{E}_\Pi[\mathsf{G}(x;\xi)] \in \partial f(x) \text{ for } x \in \mathcal{X}, \\ &\text{and } \; d_{\rm TV}\left(P_{[t]}^{t+\tau}, \Pi\right) = 0 \text{ for all } t \in \mathbb{N} \text{ with probability } 1. \end{aligned}$$

The set $\Theta\left(f, \tau, G, p\right)$ is the collection of oracles $\theta = (P, \mathsf{G})$ for which the distribution P has stationary distribution Π, mixing time bounded by τ, and returns ℓ_p-norm bounded stochastic subgradients of the function f. The condition $\left\|\mathsf{G}(x;\xi)\right\|_p \le G$ guarantees that Assumptions A and B hold, while $d_{\rm TV}(P_{[t]}^{t+\tau}, \Pi) = 0$ satisfies Assumption C. With Definition 3.6, for any collection $\mathcal{C}$ of convex functions f, we can define the minimax error over distributions with mixing times bounded by τ as

$$\epsilon_T^*(\mathcal{C}, \mathcal{X}, \tau, G, p) := \inf_{\mathcal{M}} \sup_{f \in \mathcal{C}} \sup_{\theta \in \Theta(f, \tau, G, p)} \mathbb{E}_\theta\left[\epsilon_T(\mathcal{M}, f, \mathcal{X}, \theta)\right]. \tag{3.6}$$

We have the following theorem on this minimax error (see § 6.5 for a proof).

Theorem 3.7. *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a convex set containing the ℓ_∞ ball of radius r for some $r > 0$. Let $1/p + 1/q = 1$ and $p \ge 1$ and let the set $\mathcal{C}$ consist of convex functions that are G-Lipschitz continuous with respect to the ℓ_q-norm over the set $\mathcal{X}$. For $p \in [1, 2]$ and for any $\tau \in \mathbb{N}$, the minimax oracle complexity* (3.6) *satisfies*

$$\epsilon_T^*(\mathcal{C}, \mathcal{X}, \tau, G, p) = \Omega\left(Gr\sqrt{d}\sqrt{\frac{\tau}{T}}\right). \tag{3.7a}$$

For $p \in [2, \infty]$ and for any $\tau \in \mathbb{N}$, the minimax oracle complexity (3.6) *satisfies*

$$\epsilon_T^*(\mathcal{C}, \mathcal{X}, \tau, G, p) = \Omega\left(Grd^{\frac{1}{q}}\sqrt{\frac{\tau}{T}}\right). \tag{3.7b}$$

We make a few brief comments on the implications of Theorem 3.7. First, the dependence on τ and T in the bounds of $\sqrt{\tau/T}$ matches that of the upper bound (3.4). In addition, following the discussion of Agarwal et al. [1, Section III.A and Appendix C], we can see that the dependence of the bounds (3.7a) and (3.7b) on the quantities

r, G, and the dimension d are optimal (to within logarithmic factors). In brief, the bound (3.7a) is achieved by taking $\psi(x) = \frac{1}{2}\left\|x\right\|_2^2$ in the definition of the proximal function for the EMD algorithm, while the bound (3.7b) is achieved by taking $\psi(x) = \frac{1}{2}\left\|x\right\|_q^2$ for $q = 1 + 1/\log(d)$ (see also [4, 3, Section 5]). Summarizing, we find that Theorems 3.1–3.4 are unimprovable by more than numerical constants, and the EMD algorithm (2.3) attains the minimax optimal rate of convergence.

4. Examples and Consequences.

We now collect several consequences of the convergence rates of Theorems 3.1, 3.3, and 3.4 to provide insight and illustrate applications of the theoretical statements. We begin with a concrete example and move toward more abstract principles, completing the section with finite sample and asymptotic convergence guarantees for more slowly mixing ergodic processes. Most of the results are new or improve over previously known bounds, and we provide a few additional examples in the extended version of this paper [14].

4.1. Peer-to-peer optimization and Markov incremental gradient descent.

The Markov incremental gradient descent (MIGD) procedure due to Johansson et al. [20] is a generalization of Nedić and Bertsekas's randomized incremental subgradient method [28], which Ram et al. [33] further analyze. The motivation for MIGD comes from a distributed optimization algorithm using a simple (locally computable) peer-to-peer communication scheme. We assume we have n processors or computers, each with a convex function $f_i : \mathcal{X} \to \mathbb{R}$, and the goal is to minimize

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) \quad \text{subject to} \quad x \in \mathcal{X}. \tag{4.1}$$

The procedure works as follows. The current set of parameters $x(t) \in \mathcal{X}$ is passed among the processors in the network, where a token $i(t) \in [n]$ indicates the processor holding $x(t)$ at iteration t. At iteration t, the algorithm computes the update

$$g(t) \in \partial f_{i(t)}(x(t)), \qquad x(t+1) = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \langle g(t), x \rangle + \frac{1}{\alpha(t)} D_\psi(x, x(t)) \right\},$$

after which the token $i(t)$ moves to a new processor. This update is a generalization of the papers [20, 33], which assume $\psi(x) = \frac{1}{2}\left\|x\right\|_2^2$. Slightly more generally, the local functions may be defined as expectations, $f_i(x) = \mathbb{E}_{\Pi_i}[F(x;\xi)]$, for a local distribution Π_i. At iteration t, a sample $\xi_{t,i(t)}$ is drawn from the local distribution $\Pi_{i(t)}$ and the algorithm computes the update

$$g(t) \in \partial F(x(t); \xi_{t,i(t)}), \quad x(t+1) = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \langle g(t), x \rangle + \frac{1}{\alpha(t)} D_\psi(x, x(t)) \right\}. \tag{4.2}$$

We view the token $i(t)$ as evolving according to a Markov chain with doubly-stochastic transition matrix P, so its stationary distribution is the uniform distribution. In this case,

$$P(i(t) = j \mid i(t-1) = i) = P_{ij}.$$

The total variation distance of the stochastic process initialized at $i(0) = i$ from the true (uniform) distribution is $\left\|P^t e_i - \mathbb{1}/n\right\|_1$, where e_i denotes the ith standard basis vector. In addition, since P is doubly stochastic, we have $P\mathbb{1} = \mathbb{1}$ and thus

$$\begin{aligned}\left\|P^t e_i - \mathbb{1}/n\right\|_1 \le \sqrt{n}\left\|P^t e_i - \mathbb{1}/n\right\|_2 &= \sqrt{n}\left\|P^t(e_i - \mathbb{1})\right\|_2 \\ &\le \sqrt{n}\rho_2(P)^t \left\|e_i - \mathbb{1}/n\right\|_2 \le \sqrt{n}\rho_2(P)^t,\end{aligned}$$

where $\rho_2(P)$ denotes the second singular value of the matrix P. From this spectral bound on the total variation distance, we see that if $t \ge \frac{\frac{1}{2}\log(Tn)}{\log \rho_2(P)^{-1}}$ we have $\|P^t e_i - \mathbb{1}/n\|_1 \le \frac{1}{\sqrt{T}}$. In addition, recalling the sandwich inequalities (2.7), we have

$$d_{\rm hel}(P^t e_i, \mathbb{1}/n) \le \sqrt{d_{\rm TV}(P^t e_i, \mathbb{1}/n)} \le n^{1/4}\rho_2(P)^{t/2}$$

so $d_{\rm hel}(P^t e_i, \mathbb{1}/n) \le 1/\sqrt{T}$ when $t \ge \frac{\log(Tn)}{\log \rho_2(P)^{-1}}$. In the notation of Assumption C,

$$\tau_{\rm TV}(P, T^{-1/2}) \le \frac{\log(Tn)}{2\log \rho_2(P)^{-1}} \le \frac{\log(Tn)}{2(1-\rho_2(P))} \quad \text{and} \quad \tau_{\rm hel}(P, T^{-1/2}) \le \frac{\log(Tn)}{1-\rho_2(P)}. \tag{4.3}$$

(Since $\log \rho^{-1} \approx 1 - \rho$ for $\rho \approx 1$, using $1 - \rho$ is no significant loss in our applications.) Consequently, we have the following result, similar to Corollary 3.5.

Corollary 4.1. *Let $x(t)$ evolve according to the Markov incremental descent update (4.2), where $i(t)$ evolves via the doubly stochastic transition matrix P and $\alpha(t) = \alpha/\sqrt{t}$. Define $\widehat{x}(T) = \frac{1}{T}\sum_{t=1}^T x(t)$ and $\tau_{\rm mix} = \sqrt{\log(Tn)}/\sqrt{1-\rho_2(P)}$. Choose stepsize multiplier $\alpha = R/G\sqrt{\tau_{\rm mix}}$. If for each distribution Π_i we have $\mathbb{E}_{\Pi_i}[\|\mathsf{G}(x;\xi)\|_*^2] \le G^2$, then*

$$\mathbb{E}[f(\widehat{x}(T))] - f(x^\star) \le \frac{5RG}{2} \cdot \frac{\sqrt{\tau_{\rm mix}}}{\sqrt{T}} + \frac{5RG}{\sqrt{T}} + \frac{RG}{T} \cdot \tau_{\rm mix}. \tag{4.4}$$

Let $\delta \in (0,1)$ and assume $\tau_{\rm mix} \le T/2$. If for each i and Π_i-almost every ξ we have $\|\mathsf{G}(x;\xi)\|_* \le G$, then with probability at least $1-\delta$

$$f(\widehat{x}(T)) - f(x^\star) \le \frac{5RG}{2} \cdot \frac{\sqrt{\tau_{\rm mix}}}{\sqrt{T}} + \frac{2RG}{\sqrt{T}} + \frac{RG}{T} \cdot \tau_{\rm mix} + \frac{3RG}{\sqrt{T}}\sqrt{\tau_{\rm mix}\log\frac{\tau_{\rm mix}}{\delta}}.$$

Proof. The proof is a consequence of Theorems 3.1 and 3.3 and Corollary 3.5. We use the uniform bound (4.3) on the mixing time of the random walk, in Hellinger or total variation distance, and the result follows via algebra. □

Corollary 4.1 gives convergence rates sharper and somewhat more powerful than those in the original Markov incremental gradient descent papers [20, 33]. First, our results allow us to use mirror descent updates, thus applying to problems having non-Euclidean geometry; it is by now well known that this is essential for obtaining efficient methods for high-dimensional problems [30, 4, 3]. Secondly, because we base our convergence analysis on mixing time rather than return times, we can give sharp high-probability convergence guarantees. Finally, our convergence rates are often tighter. Ram et al. [33] do not appear to give finite sample convergence rates, and as discussed by Duchi et al. [15], Johansson et al. [20] show that MIGD—with optimal choice of their algorithm parameters—has convergence rate $\mathcal{O}(RG \max_i \sqrt{\frac{n\Gamma_{ii}}{T}})$, where Γ is the return time matrix given by $\Gamma = (I - P + \mathbb{1}\mathbb{1}^\top/n)^{-1}$. When P is symmetric (as in [20, Lemma 1]), the eigenvalues of Γ are 1 and $1/(1-\lambda_i(P))$ for $i > 1$, and

$$n \max_{i \in [n]} \Gamma_{ii} \ge \operatorname{tr}(\Gamma) = 1 + \sum_{i=2}^n \frac{1}{1-\lambda_i(P)} > \frac{1}{1-\rho_2(P)}.$$

Thus, up to logarithmic factors, the bound (4.4) from Corollary 4.1 is never weaker. For well-connected graphs, the bound is substantially stronger; for example, a random walk on an expander graph has constant spectral gap [10], so $(1-\rho_2(P))^{-1} = \mathcal{O}(1)$, while the previous bound is $n \max_{i\in[n]} \Gamma_{ii} = \Omega(n)$.

4.2. Probabilistically mixing processes. We now turn to two examples to show the broader applicability of the EMD algorithm guaranteed by Theorem 3.4. Our first example generalizes the Markov incremental gradient method of § 4.1 to allow random communication matrices P, while our second considers optimization problems where the data comes from a (potentially nonlinear) autoregressive moving average (ARMA) process. For both examples, we require a conversion from expected convergence of the total variation distance $d_{\rm TV}(P_{[t]}^{t+\tau}, \Pi)$ as $\tau \to \infty$ to the probabilistic bound in Assumption D. To that end, we prove the following lemma in Appendix D.

Lemma 4.2. *Let $\mathbb{E}[d_{\rm TV}(P_{[t]}^{t+\tau}, \Pi)] \le K\rho^\tau$ for all $\tau \in \mathbb{N}$, where $K \ge 1$ and $\rho \in (0, 1)$. Define*

$$\tau_{\rm TV}(P, \epsilon) := \left\lceil \frac{\log \frac{1}{\epsilon}}{|\log \rho|} + \frac{\log K}{|\log \rho|} \right\rceil + 1 \quad \text{and} \quad \kappa := \frac{1}{|\log \rho|}.$$

For any $\epsilon \in (0, 1]$ and $c \in \mathbb{R}$,

$$P\left(\tau_{\rm TV}(P_{[t]}, \epsilon) \ge \tau_{\rm TV}(P, \epsilon) + \kappa c\right) \le \exp(-c).$$

We begin with the analysis of the random version of the Markov incremental gradient descent (MIGD) procedure. As before, a token $i(t)$ moves among the processors in a network of n nodes, but now the transition matrix P governing the token is random. At time t, the transition probability $P(i(t) = j \mid i(t-1) = i) = P_{ij}(t)$, where $\{P(t)\}$ is an i.i.d. sequence of doubly stochastic matrices. Let Δ_n denote the probability simplex in $\mathbb{R}^n$ and $u(0) \in \Delta_n$ be arbitrary. Define the sequence $u(t+1) = P(t)u(t)$, so $u(t)$ is the distribution of $i(t)$ if the token has initial distribution $u(0)$. As shown by Boyd et al. [7] and further studied by Duchi et al. [15], we obtain

$$\mathbb{E}\left[\|u(t) - \mathbb{1}/n\|_1\right] \le \sqrt{n}\,\|u(0)\|_2^2\,\lambda_2(\mathbb{E}[P(1)^\top P(1)])^t \le \sqrt{n}\lambda_2(\mathbb{E}[P(1)^\top P(1)])^t. \quad (4.5)$$

Notably, with $\rho = \lambda_2(\mathbb{E}[P(1)^\top P(1)]) < 1$ and $K = \sqrt{n}$, the estimate (4.5) satisfies the conditions of Lemma 4.2, since $d_{\rm TV}(P^t, \Pi) = \|u(t) - \mathbb{1}/n\|_1$. Generally, $\mathbb{E}[P(1)^\top P(1)]$ has much smaller second eigenvalue than any of the random matrices $P(t)$ (indeed, it may be the case that $\lambda_2(P(t)) = 1$ with probability 1, as in randomized gossip [7]). Using (4.5), if we define $\lambda_2 = \lambda_2(\mathbb{E}[P(1)^\top P(1)])$, we may take

$$\tau_{\rm TV}(P, \epsilon) \le \frac{\log \frac{n}{\epsilon}}{1 - \lambda_2} \quad \text{and} \quad \kappa \le \frac{1}{1 - \lambda_2}$$

in Lemma 4.2. Applying Theorem 3.4 we obtain the following corollary.

Corollary 4.3. *Let the conditions of Theorem 3.4 hold, and in the notation of the previous paragraph, define $\lambda_2 := \lambda_2(\mathbb{E}[P(1)^\top P(1)])$. Fix $\delta \in (0, 1]$. With stepsize choice $\alpha(t) = \alpha/\sqrt{t}$, there is a constant $C \le 4$ such that with probability at least $1 - \delta$*

$$f(\widehat{x}(T)) - f(x^\star) \le \inf_{\epsilon > 0} C \cdot \left(\frac{R^2}{\alpha\sqrt{T}} + \frac{\log \frac{n}{\epsilon} + \log \frac{T}{\delta}}{1 - \lambda_2} \cdot \frac{G^2 \alpha}{\sqrt{T}} + \epsilon G R \right.$$
$$\left. + \frac{GR}{\sqrt{T}} \sqrt{\frac{\log \frac{Tn}{\epsilon\delta} \log(\frac{1}{\delta} \log \frac{Tn}{\epsilon\delta} / (1 - \lambda_2))}{1 - \lambda_2}} \right).$$

As an example of the applicability of this approach, suppose that in the network of communicating agents used in MIGD, each communication link fails with a probability

$\gamma \in (0,1)$, independently of the other links. Let P denote the transition matrix used by the MIGD algorithm without network failures. Then (under suitable conditions on the network topology; see [15] for details)

$$\lambda_2(\mathbb{E}[P(1)^\top P(1)]) \le \gamma + (1-\gamma)\lambda_2(P).$$

Applying Corollary 4.3 and taking $\epsilon = 1/T$ and $\delta = 1/T^2$, we obtain (ignoring doubly logarithmic factors) that there is a universal constant C such that with probability at least $1 - T^{-2}$

$$f(\widehat{x}(T)) - f(x^\star) \le C \cdot \left(\frac{R^2}{\alpha\sqrt{T}} + \frac{\log(Tn)}{(1-\gamma)(1-\lambda_2(P))} \cdot \frac{G^2\alpha}{\sqrt{T}} \right).$$

Roughly, we see the intuitive result that as the failure probability γ increases to 1, the convergence rate of the algorithm suffers; for γ bounded away from 1, we suffer only constant factor losses over the rates in Corollary 4.1.

As another example of the applicability of Theorem 3.4, we look to problems where the statistical sample space Ξ is uncountable. In such scenarios, standard (finite-dimensional) Markov chain theory does not apply. Uncountable spaces commonly arise, for example, in physical simulations of natural phenomena or autoregressive processes [26], control problems [21], as well as in statistical learning applications, such as Monte Carlo-sampling based variants of the expectation maximization (EM) algorithm [38]. To apply results based on Assumption C, however, requires *uniform ergodicity* [26, Chapter 16] of the Markov chain. Uniform ergodicity is difficult to verify and often requires conditions essentially equivalent to compactness of Ξ.

Theorem 3.4 allows us to avoid such difficulties. For concreteness, we focus on autoregressive moving average (ARMA) processes, common models for control problems and statistical time series. In general, an ARMA process is defined by the recursion

$$\xi_{t+1} = A(\xi_t) + \Sigma(\xi_t)W_t, \tag{4.6}$$

where $A : \mathbb{R}^d \to \mathbb{R}^d$ and $\Sigma : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ are measurable, the innovations $W_t \in \mathbb{R}^d$ are i.i.d. and $\mathrm{Cov}(W_t)$ exists. When $A(z) = Az$, that is, A is identified with a matrix $A \in \mathbb{R}^{d\times d}$, and $\Sigma(z)$ is a constant matrix Σ, we recover the standard linear ARMA model. The convergence of such processes is area of recent research (e.g., [26, 27, 23]), but we focus particularly on the paper of Liebscher [23]. As a consequence of Liebscher's Theorem 2, we obtain that if $A(\xi) = A\xi + h(\xi)$, where $h(\xi) = o(\|\xi\|)$ as $\|\xi\| \to \infty$, the matrix A satisfies $\rho_1(A) < 1$, and $\Sigma(\xi) \equiv \Sigma$ is a fixed matrix, then there exist constants $M \ge 0$ and $\rho \in (0,1)$ such that for all $t, \tau \in \mathbb{N}$

$$\mathbb{E}\left[d_{\mathrm{TV}}(P_{[t]}^{t+\tau}, \Pi)\right] \le M\rho^\tau \quad \text{whenever} \quad \mathbb{E}[\|\xi_0\|] < \infty.$$

Here Π is the stationary distribution of the ARMA process (4.6).

In particular, for any ARMA process (4.6) satisfying the conditions, Lemma 4.2 guarantees that Assumption D holds. We thus have the following corollary (it appears challenging to obtain sharp constants [23, 26], so we leave many unspecified).

Corollary 4.4. *Let the stochastic process P be the nonlinear ARMA process*

$$\xi_{t+1} = A\xi_t + h(\xi_t) + \Sigma W_t,$$

where the singular value $\rho_1(A) < 1$, $h(\xi) = o(\|\xi\|)$ as $\|\xi\| \to \infty$, and $\mathbb{E}[\|W_t\|_2^2] < \infty$. Let Assumption B hold and $\delta \in (0,1)$. Then there exist constants $M \ge 1$, $\rho \in (0,1)$, *and a universal constant $C \leq 4$ such that with probability at least $1 - \delta$*

$$f(\widehat{x}(T)) - f(x^\star) \leq \inf_{\epsilon > 0} C \cdot \left(\frac{R^2}{\alpha\sqrt{T}} + \frac{\log \frac{MT}{\epsilon\delta}}{1-\rho} \cdot \frac{G^2\alpha}{\sqrt{T}} + \epsilon GR \right. \\ \left. + \frac{GR}{\sqrt{T}} \sqrt{\frac{\log \frac{MT}{\epsilon\delta} \log(\frac{1}{\delta} \log \frac{MT}{\epsilon\delta}/(1-\rho))}{1-\rho}} \right).$$

Having provided Corollaries 4.3 and 4.4, we can now somewhat more concretely contrast our results with those of Ram et al. [33]. Ram et al.'s results (essentially) apply when the set Ξ is finite, as they define their objective $f(x) = \sum_{i=1}^n f_i(x)$ for functions f_i; the ARMA example does not satisfy this property. In addition, Ram et al. assume in the MIGD case that the network of agents $\{1, \ldots, n\}$ is strongly connected over time: for any t, if one defines $E(t) = \{(i,j) : P(t)_{ij} > 0\}$, there exists a finite $t' \in \mathbb{N}$ such that $\cup_{s=t}^{t'} E(s)$ defines a strongly connected graph. This assumption need not hold for our analysis and fails for the examples motivating Corollary 4.3.

4.3. Slowly mixing processes. Many ergodic processes do not enjoy the fast convergence rates of the previous three examples. Thus we turn to a brief discussion of more slowly mixing processes, which culminates in a result (Corollary 4.5) establishing asymptotic convergence of EMD for any ergodic process satisfying Assumption C.

In general, attaining optimal rates of convergence for slowly mixing processes requires knowledge of the mixing rate of P. Choosing an incorrect rate—that is, setting $\alpha(t) \propto t^{-m}$ for incorrect m—can lead to substantially slower convergence. In contrast, as noted in § 3, our other bounds are robust to mis-specification of the step size so long as the ergodic process P mixes suitably quickly and we can choose $\alpha(t) \propto t^{-1/2}$. There is a simple technique we can use to demonstrate that the stepsize choice $\alpha(t) = \alpha/\sqrt{t}$ provably yields convergence, both in expectation and with high probability, even for slowly mixing processes. To be specific, note that the bound in Corollary 3.2 guarantees that for $\widehat{x}(T) = \frac{1}{T}\sum_{t=1}^T x(t)$, if we choose $\alpha(t) = \alpha/\sqrt{t}$ then

$$\mathbb{E}[f(\widehat{x}(T))] - f(x^\star) \leq \frac{R^2}{2\alpha\sqrt{T}} + \frac{G^2\alpha}{\sqrt{T}} + 3\epsilon GR + \frac{2\tau_{\rm mix}(P,\epsilon)G^2\alpha}{\sqrt{T}} + \frac{\tau_{\rm mix}(P,\epsilon)RG}{T}, \quad (4.7)$$

where $\tau_{\rm mix}$ denotes either the Hellinger or total variation mixing time. The convergence guarantee (4.7) holds regardless of our choice of ϵ, so we can choose ϵ minimizing the right-hand side. That is (setting $\alpha = R/G$ for notational convenience),

$$\mathbb{E}[f(\widehat{x}(T))] - f(x^\star) \leq \frac{3GR}{2\sqrt{T}} + \inf_{\epsilon \geq 0} \left\{ 3\epsilon GR + \frac{2\tau_{\rm mix}(P,\epsilon)GR}{\sqrt{T}} + \frac{\tau_{\rm mix}(P,\epsilon)GR}{T} \right\}.$$

For any fixed $\epsilon > 0$, the term inside the infimum decreases to $4\epsilon GR$ as $T \uparrow \infty$, so the infimal term decreases to zero as $T \uparrow \infty$. High probability convergence follows similarly by using Theorem 3.3, since for any $\delta_T > 0$ we have

$$f(\widehat{x}(T)) - f(x^\star) \leq \frac{3GR}{2\sqrt{T}} + \inf_{\epsilon \geq 0} \left\{ \epsilon GR + \frac{2\tau_{\rm TV}(P,\epsilon)GR}{\sqrt{T}} + \frac{\tau_{\rm TV}(P,\epsilon)GR}{T} \right. \\ \left. + \frac{4GR}{\sqrt{T}} \sqrt{\tau_{\rm TV}(P,\epsilon) \log \frac{\tau_{\rm TV}(P,\epsilon)}{\delta_T}} \right\} \quad (4.8)$$

with probability at least $1 - \delta_T$. We obtain the following corollary:

COROLLARY 4.5. *Define $\widehat{x}(T) = \frac{1}{T}\sum_{t=1}^{T} x(t)$. Under the conditions of Theorem 3.3, the stepsize sequence $\alpha(t) = \alpha/\sqrt{t}$ for any $\alpha > 0$ yields $f(\widehat{x}(T)) \to f(x^\star)$ as $T \to \infty$ both in expectation and with probability 1.*

Proof. Fix $\gamma > 0$ and let E_T denote the event that $f(\widehat{x}(T)) - f(x^\star) > \gamma$. We use the Borel-Cantelli lemma [6] to argue that E_T occurs for only a finite number of T with probability one. Take the sequence $\delta_T = 1/T^2$ (any sequence for which $\log(1/\delta_T)/T \downarrow 0$ as $T \to \infty$ and $\sum_{T=1}^{\infty} \delta_T < \infty$ will suffice) and choose some T_0 such that the right-hand side of the bound (4.8) is less than γ. Then we have

$$\sum_{T=1}^{\infty} P(f(\widehat{x}(T)) - f(x^\star) > \gamma) = \sum_{T=1}^{\infty} P(E_T) \le T_0 + \sum_{T=T_0+1}^{\infty} P(E_T) \le T_0 + \sum_{T=1}^{\infty} \delta_T < \infty.$$

For any $\gamma > 0$, we have $P(f(\widehat{x}(T)) - f(x^\star) > \gamma \text{ i.o.}) = 0$. □

5. Numerical results.

In this section, we present simulation experiments that further investigate the behavior of the EMD algorithm (2.3). Though Theorem 3.7 guarantees that our rates are essentially unimprovable, it is interesting to compare our method with other natural well-known procedures. We would also like to understand the benefits of the mirror descent approach for problems in which the natural geometry is non-Euclidean as well as the robustness properties of the algorithm.

5.1. Sampling strategies.

For our first experiment, we study the performance of the EMD algorithm on a robust system identification task [32], where we assume the data is generated by an autoregressive process. More precisely, our data generation mechanism is as follows. For each experiment, we set the matrix A to be a sub-diagonal matrix (all entries are 0 except those on the sub-diagonal), where $A_{i,i-1}$ is drawn uniformly from $[.8, .99]$. We then draw a vector u uniformly from surface of the d-dimensional ℓ_2-ball of radius $R = 5$. The data comes in pairs $(\xi_t^1, \xi_t^2) \in \mathbb{R}^d \times \mathbb{R}$ with $d = 50$ and is generated as follows:

$$\xi_t^1 = A\xi_{t-1}^1 + e_1 W_t, \quad \xi_t^2 = \langle u, \xi_t^1 \rangle + E_t, \tag{5.1}$$

where e_1 is the first standard basis vector, W_t are i.i.d. samples from $N(0,1)$, and E_t are i.i.d. bi-exponential random variables with variance 1. Polyak and Tsypkin [32] suggest the method of least-moduli for the system identification task, setting

$$F(x; (\xi^1, \xi^2)) = \left|\langle x, \xi^1 \rangle - \xi^2\right|,$$

which is optimal (in a minimax sense) when little is known about the noise distribution [32]. Our minimization problem is

$$\operatorname*{minimize}_x \; f(x) = \mathbb{E}_\Pi\left[\left|\langle x, \xi^1 \rangle - \xi^2\right|\right] \quad \text{subject to} \quad \|x\|_2 \le R, \tag{5.2}$$

where Π is the stationary distribution of the AR model (5.1) and we take $R = 5$.

We use this experiment to investigate two issues. In addition to studying the performance of the EMD algorithm in minimizing the expected objective (5.2), we compare EMD to a natural alternative. In many engineering applications it is possible to generate samples from a distribution P that converges to Π, in which case a natural algorithm is to use the so-called "multiple replications" approach (e.g., [16]). In this approach, one specifies initial conditions of the stochastic process P, then simulates it for some number k of steps, and obtains a sample ξ according to the

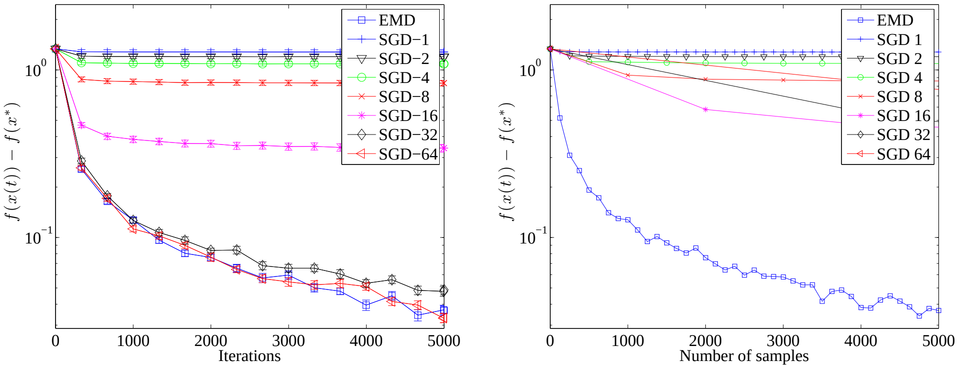

FIG. 5.1. *Performance of the EMD algorithm* (2.3) *on a robust system identification task where data is generated according to an autoregressive process.*

marginal distribution P^k, which (hopefully) is close to Π. Repeating this, one can obtain multiple independent samples ξ from P^k, then use standard algorithms and analyses for independent data.[2] A difficulty with this approach—which we see in our experiments—is that the mixing time of the process P may be unknown, and if P^k does not converge precisely to Π for any finite $k \in \mathbb{N}$, then any algorithm using such samples will be biased even in the limit of infinite gradient steps.

As a natural representative from the multiple-replications family of algorithms, we use the classical stochastic gradient descent (SGD) algorithm (in the form studied by Nemirovski et al. [29]). To generate each sample for SGD, we begin with the point $\xi_1^1 = 0$ and perform k of steps of the procedure (5.1), using $\xi_k^{\{1,2\}}$ to compute subgradients for SGD. For EMD, we use the proximal function $\psi(x) = \frac{1}{2}\|x\|_2^2$, which yields the direct analogue of stochastic gradient descent. To measure the objective value $f(x)$, we generate an independent fixed sample of size $N = 10^5$ from the process (5.1), using $f(x) \approx \frac{1}{N}\sum_{i=1}^N F(x;\xi_i)$. For each algorithm, to choose the stepsize multiplier $\alpha \propto R/G$, we estimate G by taking 100 samples ξ_t^1 and computing the empirical average of $\|\xi_t^1\|_2^2$. For EMD, we deliberately underestimate the mixing time by the constant 1 (other estimates of the mixing time yielded similar performance).

In Figure 5.1, we show the convergence behavior (as a function of number of samples) for the EMD algorithm compared with the behavior of the stochastic gradient method for different numbers k of initial simulation steps before obtaining the sample ξ used in each iteration of SGD. The line in each plot corresponding to SGD-k shows the convergence of stochastic gradient descent as a function of number of iterations when k initial samples are used for each independent sample ξ. The left plot in Figure 5.1 makes clear that if the mixing time is underestimated, the multiple-replications approach fails. As demonstrated by our theory, however, EMD still guarantees convergence even with poor stepsize choices (see also our experiments in the next section). For large enough mixing time estimate k, the multiple-replication stochastic gradient method and the EMD method have comparable performance in terms of optimization error as a function of number of gradient steps. The right plot in Figure 5.1 shows the convergence behavior of the competing methods as a function of the number of

[2] This approach is inapplicable when the data ξ_t comes from a real (unsimulated) source, such as in streaming, online optimization, or statistical applications, though the EMD algorithm still applies.

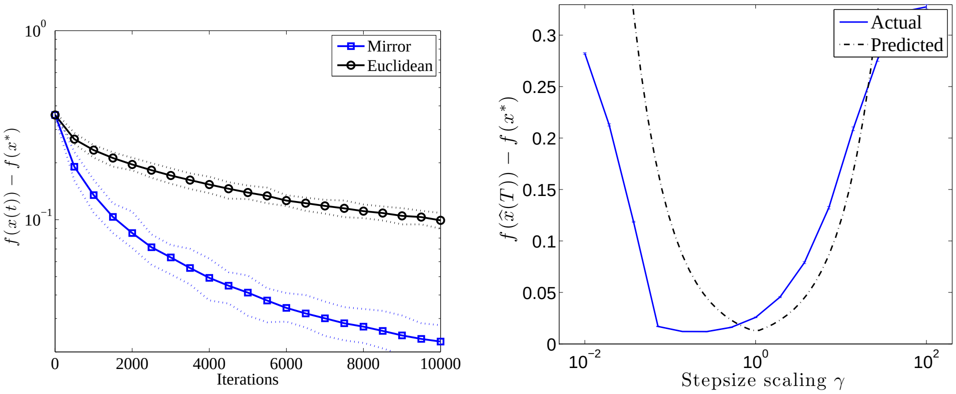

FIG. 5.2. *Left: optimization error on a statistical machine learning task of the Euclidean variant of the EMD algorithm* (2.3) *versus that of the ℓ_q-norm variant with $\psi(x) = \frac{1}{2}\|x\|_q^2$, $q = 1 + 1/\log d$, plotted against number of iterations. Right: robustness of the EMD algorithm* (2.3) *to modifications in the choice of stepsize.*

samples of the stochastic process (5.1). From this plot, it becomes clear that using each sample sequentially as in EMD—rather than attempting to draw independent samples at each iteration—is the more computationally efficient approach.

5.2. Robustness and non-Euclidean geometry. In our second numerical experiment, we study an important problem that takes motivation from distributed statistical machine learning problems: the support vector machine problem [11], where the samples $\xi \in \mathbb{R}^d$ and the instantaneous objective is

$$F(x;\xi) = [1 - \langle \xi, x\rangle]_+ .$$

We study the performance of the EMD algorithm for the distributed Markov incremental mirror descent framework in § 4.1. In the notation of § 4.1, we simulate $n = 50$ "processors," and for each we draw a sample of $m = 50$ samples according to the following process. Before performing any sampling, we set u to be a random vector from $\{x \in \mathbb{R}^d : \|x\|_1 \le R\}$, where $R = 5$ and $d = 500$. To generate the ith data sample, we draw a vector $a_i \in \mathbb{R}^d$ with entries $a_{i,j} \in \{-1, 1\}$ each with probability $\frac{1}{2}$, and set $b_i = \mathrm{sign}(\langle a_i, u\rangle)$. With probability .05, we flip the sign of b_i (this makes the problem slightly more difficult, as no vector x will perfectly satisfy $b_i = \mathrm{sign}(\langle a_i, x\rangle)$), and regardless we set $\xi_i = b_i a_i$. We thus generate a total of $N = nm = 2500$ samples, and set the ith objective in the distributed minimization problem (4.1) to be

$$f_i(x) = \frac{1}{m}\sum_{k=m(i-1)+1}^{mi} F(x;\xi_k) = \mathbb{E}_{\Pi_i}[F(x;\xi)] = \mathbb{E}_{\Pi_i}\left[[1 - \langle \xi, x\rangle]_+\right], \qquad (5.3)$$

where Π_i denotes the uniform distribution over the ith block of m samples. Our algorithm to minimize $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$ is the Markov analogue (4.2) of the general EMD algorithm (2.3). We minimize $f(x)$ over $\{x : \|x\|_1 \le R\}$ offline using standard LP software to obtain the optimal value $f(x^\star)$ of the problem.

We use the objectives (5.3) to (i) understand the effectiveness of allowing non-Euclidean proximal functions ψ in the update (2.3) and (ii) study the robustness of the EMD algorithm (2.3) to stepsize selection. We begin with the first goal. As noted by Ben-Tal et al. [4], the choice $\psi(x) = \frac{1}{2}\left\|x\right\|_q^2$ with $q = 1 + 1/\log(d)$ yields a nearly optimal dependence on dimension in non-Euclidean gradient methods. Let $\tau_{\rm mix}$ denote the mixing time of the Markov chain (for Hellinger or total variation distance). Applying Corollary 4.1 and the analysis of Ben-Tal et al. with this choice of proximal function and $\alpha = R/\sqrt{\log(d)\tau_{\rm mix}}$ yields

$$\mathbb{E}[f(\widehat{x}(T))] - \inf_{x \in \mathcal{X}} f(x) = \mathcal{O}\left(\frac{R\sqrt{\tau_{\rm mix}\log d}}{\sqrt{T}}\right),$$

since $\left\|\partial_x F(x;\xi)\right\|_\infty \le \left\|\xi\right\|_\infty = 1$ by our sampling of the vectors $a_i \in \{-1,1\}^d$, and R is the radius of $\mathcal{X}$ in ℓ_1-norm. Compared to the Euclidean variant [20, 33] with $\psi(x) = \frac{1}{2}\left\|x\right\|_2^2$, whose convergence rate also follows from Corollary 4.1, this is an improvement of $\sqrt{d/\log d}$, since $\left\|\partial_x F(x;\xi)\right\|_2$ can be as large as $\sqrt{d}$.

We plot the results of 50 simulations of the distributed minimization problem in the left plot of Figure 5.2. For our underlying network topology, we use a 4-connected cycle (each node in the cycle is connected to its 4 neighbors on the right and left) and $n = 50$ nodes. The line of blue squares is the mirror-descent approach with $\psi(x) = \frac{1}{2}\left\|x\right\|_q^2$ with $q = 1 + 1/\log(d)$ (we use $d = 500$), while the black line of circles denotes the Euclidean variant with $\psi(x) = \frac{1}{2}\left\|x\right\|_2^2$. The dotted lines below and above each plot give the 5th and 95th percentiles, respectively, of the optimization error across all simulations. For each algorithm, we use the optimal step size setting $\alpha(t)$ predicted by our theory (recall Corollary 4.1). It is clear that the non-Euclidean variant enjoys better performance, as our theory (and previous work on the dimension dependence of mirror descent [30, 29, 4, 3]) suggests.

The final simulation we perform is on the same problem, but we investigate the robustness of the EMD algorithm to mis-specified stepsizes. We take the stepsize α^* predicted by our theory (Corollary 4.1), and use $\alpha(t) = \gamma\alpha^*/\sqrt{t}$ for values of γ uniformly logarithmically spaced from $\gamma = 10^{-2}$ to $\gamma = 10^2$. The plot on the right side of Figure 5.2 shows the mean optimality gap of $\widehat{x}(T)$ after $T = 10000$ iterations for different values of γ, along with standard deviations, across 50 experiments. The black dotted line shows the predicted optimality gap as a function of the mis-specification (recall our discussion on robustness following Corollary 3.5). The EMD algorithm is certainly affected by mis-specification of the initial stepsize, though for a range of values of roughly $\gamma = 10^{-1}$ to $\gamma = 10$, the performance degradation does not appear extraordinary. In addition, our experiments show that our theoretical predictions appear to capture the empirical behavior of the method quite well.

6. Analysis. In this section, we analyze the convergence of the EMD algorithm from Section 2. Our first subsection lays the groundwork, gives necessary notation, and provides a few optimization-based results. The second subsection contains the proofs of results on expected rates of convergence, while the third subsection shows how to achieve convergence guarantees with high probability. The fourth subsection shows the convergence of the EMD method under probabilistic (random) mixing times, while the final subsection proves the order-optimality of the EMD method.

6.1. Definitions, assumptions, and optimization-based results. To state our results formally, we begin by giving a few standard definitions and collecting a few consequences of Assumptions A and B that make our proofs cleaner. Recall the measurable selection G, where $\mathsf{G}(x;\xi) \in \partial_x F(x;\xi)$ represents a fixed and measurable element of the subgradient of $F(\cdot;\xi)$ evaluated at x, and the EMD algorithm (2.3) has $g(t) = \mathsf{G}(x(t); \xi_t)$. By our assumptions on F, for any distribution Q for which the expectations below are defined, expectation and subdifferentiation commute [35, 5]:

$$f_Q(x) := \mathbb{E}_Q[F(x; \xi)] = \int_\Xi F(x; \xi) dQ(\xi) \quad \text{then} \quad \partial f_Q(x) = \mathbb{E}_Q[\partial F(x; \xi)].$$

In particular, $\mathbb{E}_\Pi[\partial F(x; \xi)] = \partial f(x)$ and $\mathbb{E}_\Pi[\mathsf{G}(x; \xi)] \in \partial f(x)$. In addition, the compactness assumption that $D_\psi(x^\star, x(t)) \le \frac{1}{2}R^2$ for all t coupled with the strong convexity of ψ implies

$$\|x(t) - x^\star\|^2 \le 2D_\psi(x^\star, x(t)) \le R^2 \quad \text{so} \quad \|x(t) - x^\star\| \le R. \tag{6.1}$$

We now provide two relatively standard optimization-theoretic results that make our proofs substantially easier. To make the presentation self-contained, we give proofs of these results in Appendix A. The two lemmas are essentially present in earlier work [30, 3], but our stochastic setting requires a bit of care.

Lemma 6.1. *Let $x(t)$ be defined by the EMD update* (2.3). *For any $\tau \in \mathbb{N}$ and any $x^\star \in \mathcal{X}$,*

$$\sum_{t=\tau+1}^{T} F(x(t); \xi_t) - F(x^\star; \xi_t) \le \frac{1}{2\alpha(T)} R^2 + \sum_{t=\tau+1}^{T} \frac{\alpha(t)}{2} \|g(t)\|_*^2 .$$

Lemma 6.2. *Let $x(t)$ be generated according to the EMD algorithm* (2.3). *Then*

$$\|x(t) - x(t+1)\| \le \alpha(t) \|g(t)\|_* .$$

6.2. Expected convergence rates. Now that we have established the relevant optimization-based results and setup in Section 6.1, the proof of Theorem 3.1 requires that we understand the impact of the ergodic sequence $\xi_1, \xi_2, \ldots$ on the EMD procedure. The key equality that allows us to prove Theorems 3.1 and 3.3 is the following: for any $\tau \ge 0$,

$$\begin{aligned}\sum_{t=1}^{T} f(x(t)) - f(x^\star) &= \sum_{t=1}^{T-\tau} f(x(t)) - f(x^\star) - F(x(t); \xi_{t+\tau}) + F(x^\star; \xi_{t+\tau}) \\ &\quad + \sum_{t=1}^{T-\tau} F(x(t); \xi_{t+\tau}) - F(x(t+\tau); \xi_{t+\tau}) \\ &\quad + \sum_{t=\tau+1}^{T} F(x(t); \xi_t) - F(x^\star; \xi_t) + \sum_{t=T-\tau+1}^{T} f(x(t)) - f(x^\star).\end{aligned} \tag{6.2}$$

We may set $\tau = 0$ in the expression (6.2), taking expectations and applying Lemma 6.1, to recover the known convergence rates [29] for the stochastic gradient method with independent samples. However, the essential idea that the expansion (6.2) allows us to implement is that for large enough τ, the sample $\xi_{t+\tau}$ is nearly "independent" of the parameters $x(t)$, since the stochastic process P is mixing. By allowing $\tau > 0$, we can bound the four sums (6.2) using a combination of Lemmas 6.1 and 6.2, then apply the mixing properties of the stochastic process P to show that $F(x(t); \xi_{t+\tau})$ is a nearly unbiased estimate of $f(x(t))$:

$$\mathbb{E}[f(x(t)) - F(x(t); \xi_{t+\tau})] \approx 0.$$

We formalize this intuition with two lemmas, whose proofs we provide in Appendix B.

LEMMA 6.3. *Let x be $\mathcal{F}_t$-measurable and $\tau \geq 1$. If Assumption A holds,*

$$\left|\mathbb{E}\left[f(x) - f(x^\star) - F(x;\xi_{t+\tau}) + F(x^\star;\xi_{t+\tau}) \mid \mathcal{F}_t\right]\right| \leq 3GR \cdot d_{\rm hel}\left(P_{[t]}^{t+\tau}, \Pi\right).$$

If Assumption B holds,

$$\left|\mathbb{E}\left[f(x) - f(x^\star) - F(x;\xi_{t+\tau}) + F(x^\star;\xi_{t+\tau}) \mid \mathcal{F}_t\right]\right| \leq GR \cdot d_{\rm TV}\left(P_{[t]}^{t+\tau}, \Pi\right).$$

The next lemma applies a type of stability argument, showing that function values evaluated at $x(t)$ and $x(t+\tau)$ cannot be too far apart.

LEMMA 6.4. *Let $\tau \geq 0$ and $\alpha(t)$ be non-increasing. If Assumption A holds, then*

$$\mathbb{E}[F(x(t);\xi_{t+\tau}) - F(x(t+\tau);\xi_{t+\tau}) \mid \mathcal{F}_{t-1}] \leq \tau\alpha(t)G^2.$$

If Assumption B holds, then

$$F(x(t);\xi_{t+\tau}) - F(x(t+\tau);\xi_{t+\tau}) \leq \tau\alpha(t)G^2.$$

We can now apply Lemmas 6.1–6.4 to give the promised proof of Theorem 3.1.

Proof of Theorem 3.1. The equality (6.2) is non-probabilistic, so all we need to complete the proof is to take expectations, applying the preceding lemmas. First, we map τ to $\tau - 1$ in the previous results, which will make our analysis cleaner. Throughout this proof, the quantity $d(\cdot, \Pi)$ will denote $3d_{\rm hel}(\cdot, \Pi)$ when we apply Assumption A and will denote $d_{\rm TV}(\cdot, \Pi)$ when using Assumption B, as the proof is identical in either case. We control the expectation of each of the four sums (6.2) in turn. First, we apply Lemma 6.3 to see that

$$\sum_{t=1}^{T-\tau+1} \mathbb{E}[f(x(t)) - f(x^\star) - F(x(t);\xi_{t+\tau-1}) + F(x^\star;\xi_{t+\tau-1})] \leq GR \sum_{t=0}^{T-\tau+1} \mathbb{E}[d(P_{[t]}^{t+\tau}, \Pi)].$$

The second of the four sums (6.2) requires Lemma 6.4, which yields

$$\sum_{t=1}^{T-\tau+1} \mathbb{E}[F(x(t);\xi_{t+\tau-1}) - F(x(t+\tau-1);\xi_{t+\tau-1})] \leq (\tau-1)G^2 \sum_{t=1}^{T-\tau+1} \alpha(t).$$

Lemma 6.1 controls the third term in the series (6.2), and taking expectations gives $\mathbb{E}[\|g(t)\|_*^2] \leq G^2$. The final term in the sum (6.2) is bounded by $(\tau-1)RG$ when either of the Lipschitz assumptions A or B hold. Summing our four bounds, we obtain that for any $\tau \geq 1$,

$$\mathbb{E}\left[\sum_{t=1}^T f(x(t)) - f(x^\star)\right] \leq GR \sum_{t=1}^{T-\tau+1} \mathbb{E}\left[d\left(P_{[t-1]}^{t+\tau-1}, \Pi\right)\right] + \frac{R^2}{2\alpha(T)} + \frac{G^2}{2}\sum_{t=\tau}^T \alpha(t)$$
$$+ (\tau-1)G^2 \sum_{t=1}^{T-\tau+1} \alpha(t) + (\tau-1)RG. \qquad (6.3)$$

Assumption C states that there exists a uniform mixing time $\tau_{\rm mix}(P, \epsilon)$ (for both total variation and Hellinger mixing) such that $d(P_{[t-1]}^{t+\tau-1}, \Pi) \leq \epsilon$. Applying the definition of $\tau_{\rm mix}$ for Hellinger or total variation mixing completes the proof. □

6.3. High-probability convergence. In this section, we complement the convergence bounds in Section 6.2 with high-probability statements. We use martingale theory to show that the bound of Theorem 3.1 holds with high probability. We begin from the same starting point as the proof of Theorem 3.1—with the expansion (6.2)—but now we show that the random sum

$$\sum_{t=1}^{T-\tau+1} f(x(t)) - f(x^\star) - F(x(t);\xi_{t+\tau-1}) + F(x^\star;\xi_{t+\tau-1}) \tag{6.4}$$

is small with high probability. Intuitively, this follows because given the initial $t-\tau$ samples $\xi_1,\ldots,\xi_{t-\tau}$, the tth sample ξ_t is almost a sample from the stationary distribution Π. With this in mind, we can show that an appropriately subsampled version of the above sequence behaves approximately as a martingale, and we can then apply Azuma's inequality [2] to derive high-probability guarantees on the sum (6.4).

PROPOSITION 6.5. *Let Assumption B hold and $\delta \in (0,1)$. With probability at least $1-\delta$, for $\tau \in \mathbb{N}$ with $\tau \in [1, T/2]$,*

$$\sum_{t=1}^{T-\tau+1} \left[f(x(t)) - F(x(t);\xi_{t+\tau-1}) + F(x^\star;\xi_{t+\tau-1}) - f(x^\star)\right]$$
$$\leq 4GR\sqrt{T\tau \log\frac{\tau}{\delta}} + GR\sum_{t=1}^{T} d_{\text{TV}}\left(P_{[t-\tau]}^t, \Pi\right).$$

We provide a proof in Appendix C and can now prove Theorem 3.3.

Proof of Theorem 3.3. The proof is a combination of the proofs of previous results. Starting from the expansion (6.2), we use Lemma 6.4 to see that

$$\sum_{t=1}^{T-\tau+1} F(x(t);\xi_{t+\tau-1}) - F(x(t+\tau-1);\xi_{t+\tau-1}) \leq (\tau-1)G^2\sum_{t=1}^{T-\tau+1}\alpha(t),$$

and applying the G-Lipschitz continuity of the functions $F(\cdot;\xi)$ and compactness of $\mathcal{X}$ we obtain

$$\sum_{t=T-\tau+2}^{T} f(x(t)) - f(x^\star) \leq (\tau-1)GR.$$

In addition, the convergence guarantee in Lemma 6.1 guarantees that

$$\sum_{t=\tau}^{T} F(x(t);\xi_t) - F(x^\star;\xi_t) \leq \frac{1}{2\alpha(T)}R^2 + \frac{G^2}{2}\sum_{t=1}^{T}\alpha(t).$$

Combining these bounds, we can replace the equality (6.2) with the bound

$$\sum_{t=1}^{T} f(x(t)) - f(x^\star) \leq \frac{1}{2\alpha(T)}R^2 + \frac{G^2}{2}\sum_{t=1}^{T}\alpha(t) + (\tau-1)\left[GR + G^2\sum_{t=1}^{T}\alpha(t)\right] \tag{6.5}$$
$$+ \sum_{t=1}^{T-\tau+1}\left[f(x(t)) - F(x(t);\xi_{t+\tau-1}) + F(x^\star;\xi_{t+\tau-1}) - f(x^\star)\right],$$

which holds for any $\tau \geq 1$. What remains is to replace the last term in the non-probabilistic bound (6.5) with the upper bound in Proposition 6.5, which holds with probability $1-\delta$, and then to replace τ with $\tau_{\text{TV}}(P,\epsilon)$, which guarantees the inequality $d_{\text{TV}}(P_{[t-\tau]}^t, \Pi) \leq \epsilon$. □

6.4. Random mixing. In this section, we give the proof of Theorem 3.4. The proof is similar to that of Theorem 3.3, but we need an auxiliary lemma that allows us to guarantee that the mixing times are bounded uniformly for all times and for all desired accuracies of mixing ϵ. See Appendix D for the proof of the lemma.

LEMMA 6.6. *Let Assumption D hold and $\delta \in (0, 1)$. With probability at least $1 - \delta$,*

$$\max_{s \in \{1,\ldots,T\}} \sup_{\{\epsilon \,:\, \tau_{\rm TV}(P,\epsilon) \le T\}} \left(\tau_{\rm TV}(P_{[s]}, \epsilon) - \tau_{\rm TV}(P, \epsilon)\right) \le \kappa \left(\log \frac{1}{\delta} + 2\log(T)\right).$$

Rewriting Lemma 6.6 slightly, we may define $\tau = \tau_{\rm TV}(P, \epsilon) + \kappa(\log \frac{1}{\delta} + 2\log(T))$, and we find that with probability at least $1 - \delta$,

$$d_{\rm TV}\left(P_{[s]}^{s+\tau}, \Pi\right) \le \epsilon \tag{6.6}$$

for all $s \in \{1, \ldots, T\}$ and for all $\epsilon > 0$ with $\tau_{\rm TV}(P, \epsilon) \le T$. This leads us to

Proof of Theorem 3.4. All that is different in the proof of this theorem from that of Theorem 3.3 is that in the penultimate inequality (6.5), when we apply Proposition 6.5, we no longer have the guarantee that $d_{\rm TV}(P_{[t-\tau]}^t, \Pi) \le \epsilon$ for all t. To that end, let ϵ be such that $\tau_{\rm TV}(P, \epsilon) \le T$. Apply Lemma 6.6 and its consequence (6.6), which states that if we take $\tau = \tau_{\rm TV}(P, \epsilon) + \kappa(\log \frac{1}{\delta} + 2\log(T))$, then we obtain that $d_{\rm TV}(P_{[t-\tau]}^t, \Pi) \le \epsilon$ with probability at least $1 - \delta$. If $\tau_{\rm TV}(P, \epsilon) > T$, the bound in the theorem holds vacuously, so we may extend the result to all $\epsilon > 0$. □

6.5. Lower bounds on optimization accuracy. Our proof of Theorem 3.7 mirrors the proof of Theorem 1 in the paper by Agarwal et al. [1], so we are somewhat terse in our description and proof. The intuition in the proof is that if the stochastic process P returns a sample from the stationary distribution Π every τ timesteps, otherwise returning a sample identical to the previous one, then the convergence rate of any algorithm should be a factor of τ slower than if it could receive independent samples from Π. Mesterharm [25] employs a similar approach to give a lower bound on the performance of online learning algorithms. More formally, by using an identical construction to [1, Section IV.A], we may reduce the problem of minimization of a function $f : \mathbb{R}^d \to \mathbb{R}$ to that of identifying the bias of d coins. To that end, let $\mathcal{V} \subset \{-1, 1\}^d$ be a packing of the d-dimensional hypercube such that $\nu, \nu' \in \mathcal{V}$ with $\nu \ne \nu'$ satisfy $\|\nu - \nu'\|_1 \ge d/2$; it is a classical fact [24] that there is such a set with cardinality $|\mathcal{V}| \ge (2e^{-\frac{1}{2}})^{d/2}$.

Now for a fixed $\tau \in \mathbb{N}$, consider the following sequential sampling procedure, which generates a set of pairs of random vectors $\{(U_t, Y_t)\}_{t=1}^{\infty}$. Choose a vector $\nu \in \mathcal{V}$ uniformly at random and let $\delta \in (0, 1/4]$. Let P_ν denote the distribution (conditional on ν) that corresponds to the following: for each t, construct samples according to

(a) If $(t-1) \mod \tau \ne 0$, take $U_t = U_{t-1}$ and $Y_t = Y_{t-1}$.

(b) Otherwise, pick a uniformly random subset $U_t \subset \{1, \ldots, d\}$ of size $|U_t| = m$, then
  (i) For each $i \in U_t$, construct a random variable C_i such that $C_i = 1$ with probability $\frac{1}{2} + \nu_i\delta$ and $C_i = -1$ with probability $\frac{1}{2} - \nu_i\delta$.
  (ii) Construct the vector $Y_t \in \{-1, 1\}^d$ such that $Y_{t,i} = C_i$ if $i \in U_t$, and otherwise $Y_{t,i}$ is uniform Bernoulli, that is, if $i \notin U_t$ then $Y_{t,i} = 1$ with probability $\frac{1}{2}$ and $Y_{t,i} = -1$ with probability $\frac{1}{2}$.

This sampling procedure yields a sequence $\xi_t = (U_t, Y_t)$, where if Π_ν is the distribution of a pair (U, Y) such that $U \subset \{1, \ldots, d\}$ is chosen uniformly at random with size

$|U| = m$ and Y is sampled according to the steps (i)–(ii) above, then Π_ν is the stationary distribution of P_ν. Moreover, we see that $d_{\rm TV}(P_{[t]}^{t+\tau+k}, \Pi) = 0$ for any $k \ge 0$ and any t, since the distribution P_ν corresponds to receiving an independent sample (U, Y) from Π_ν every τ steps.

Let $I(X;Y)$ denote the mutual information between random variables X and Y and let $H(X)$ denote the (Shannon) entropy of X. By inspection of Agarwal et al.'s proof [1, Lemma 3], since Π_ν is the stationary distribution of P_ν, a tight enough bound on the mutual information $I((U_1, Y_1), \ldots, (U_T, Y_T); \nu)$ proves Theorem 3.7. Hence, we provide the following lemma:

LEMMA 6.7. *Let the sequence $\xi_t = (U_t, Y_t)$ be generated according to the steps* (a)–(b) *above. Then for $\delta \in (0, 1/4]$,*

$$I\left((U_1, Y_1), \ldots, (U_T, Y_T); \nu\right) \le 16 \left\lceil \frac{T}{\tau} \right\rceil m\delta^2.$$

Proof. Our sampling model (a)–(b) sets blocks of size τ to be equal, that is, $(U_1, Y_1) = \cdots = (U_\tau, Y_\tau)$, $(U_{\tau+1}, Y_{\tau+1}) = \cdots = (U_{2\tau}, Y_{2\tau})$, and so on, whereas different blocks are independent given the variable ν. We thus see that by the definitions of mutual information, conditional entropy, and that entropy is sub-additive [12],

$$\begin{aligned}
&I\left((U_1, Y_1), \ldots, (U_T, Y_T); \nu\right) \\
&= H((U_1, Y_1), \ldots, (U_T, Y_T)) - H((U_1, Y_1), \ldots, (U_T, Y_T) \mid \nu) \\
&= H((U_1, Y_1), \ldots, (U_T, Y_T)) - \sum_{k=1}^{\lceil T/\tau \rceil} H\left((U_{(k-1)\tau+1}, Y_{(k-1)\tau+1}), \ldots, (U_{k\tau}, Y_{k\tau}) \mid \nu\right) \\
&\le \sum_{k=1}^{\lceil T/\tau \rceil} \Big[H\left(((U_{(k-1)\tau+1}, Y_{(k-1)\tau+1}), \ldots, (U_{k\tau}, Y_{k\tau})\right) \\
&\qquad\qquad - H\left((U_{(k-1)\tau+1}, Y_{(k-1)\tau+1}), \ldots, (U_{k\tau}, Y_{k\tau}) \mid \nu\right) \Big] \\
&= \sum_{k=1}^{\lceil T/\tau \rceil} I\left((U_{(k-1)\tau+1}, Y_{(k-1)\tau+1}), \ldots, (U_{k\tau}, Y_{k\tau}); \nu\right) = \sum_{k=1}^{\lceil T/\tau \rceil} I\left((U_{k\tau}, Y_{k\tau}); \nu\right). \quad (6.7)
\end{aligned}$$

In the last line we have used that within the same block of size τ, all (U_t, Y_t) pairs are equal. Now, using the bound (6.7), we apply an identical derivation as that given in the proof of Agarwal et al.'s Lemma 3 (following Eq. (25) there). For any fixed k we have $I((U_{k\tau}, Y_{k\tau}); \nu) \le 16m\delta^2$, which completes the proof of the lemma. □

Proof of Theorem 3.7. Use Agarwal et al.'s construction (see Eq. (16) in Section IV.A of [1]) of a "difficult" subclass of functions, then in the proof of Theorem 1 from [1], replace their coin-flipping oracle with the steps (a)–(b) and applications of their Lemma 3 with Lemma 6.7 above. □

7. Conclusions. In this paper, we have shown that stochastic subgradient and mirror descent approaches extend in an elegant way to situations in which we have no access to i.i.d. samples from the desired distribution. In spite of this difficulty, we are able to achieve reasonably fast rates of convergence for the ergodic mirror descent algorithm—the natural extension of stochastic mirror descent—under reasonable assumptions on the ergodicity of the stochastic process $\{\xi_t\}$ that generates the samples. We gave several examples showing the strengths and uses of our new analysis, and believe that there are many more. In the extended version of this paper [14], we give additional applications to optimization over combinatorial spaces where Markov chain Monte Carlo samplers can be designed to sample efficiently from the space [19]. In addition, our results give a relatively clean and simple way to derive finite sample rates of convergence for statistical estimators with dependent data without requiring the full machinery of empirical process theory (e.g., [39]). Though we have provided lower bounds showing that our analysis is tight to numerical constants, it may be possible to sharpen our results for interesting special cases, such as when the distribution of the stochastic process $\{\xi_t\}$ has nice enough Markovianity properties. We leave such questions to future work.

Acknowledgments. We thank Lester Mackey for several interesting questions he posed that helped lead to this work. In addition, we thank the three anonymous reviewers and the editor for many insightful comments and suggestions.

Appendix A. Proofs of Optimization Results.

Proof of Lemma 6.1. The proof of the lemma begins by controlling the amount of progress made by one step of the EMD method, then summing the resulting bound. By the first-order convexity inequality and definition of the subgradient $g(t)$, we have

$$\begin{aligned} F(x(t);\xi_t) - F(x^*;\xi_t) &\le \langle g(t), x(t) - x^* \rangle \\ &= \langle g(t), x(t+1) - x^* \rangle + \langle g(t), x(t+1) - x(t) \rangle . \end{aligned} \tag{A.1}$$

For $y \in \mathcal{X}$, the first-order optimality conditions for $x(t+1)$ in the update (2.3) imply

$$\langle \alpha(t) g(t) + \nabla\psi(x(t+1)) - \nabla\psi(x(t)), y - x(t+1) \rangle \ge 0.$$

In particular, we can take $y = x^*$ in this bound to find

$$\alpha(t) \langle g(t), x(t+1) - x^* \rangle \le \langle \nabla\psi(x(t+1)) - \nabla\psi(x(t)), x^* - x(t+1) \rangle . \tag{A.2}$$

Now we use the definition of the Bregman divergence D_ψ, to obtain

$$\begin{aligned} &\langle \nabla\psi(x(t+1)) - \nabla\psi(x(t)), x^* - x(t+1) \rangle \\ &\qquad = D_\psi(x^*, x(t)) - D_\psi(x^*, x(t+1)) - D_\psi(x(t+1), x(t)). \end{aligned}$$

Combining this result with the expanded gradient term (A.1) and the the first-order convexity inequality (A.2), we get

$$\begin{aligned} F(x(t);\xi_t) - F(x^*;\xi_t) &\le \frac{1}{\alpha(t)} D_\psi(x^*, x(t)) - \frac{1}{\alpha(t)} D_\psi(x^*, x(t+1)) \\ &\qquad - \frac{1}{\alpha(t)} D_\psi(x(t+1), x(t)) + \langle g(t), x(t+1) - x(t) \rangle \\ &\overset{(i)}{\le} \frac{1}{\alpha(t)} D_\psi(x^*, x(t)) - \frac{1}{\alpha(t)} D_\psi(x^*, x(t+1)) - \frac{1}{\alpha(t)} D_\psi(x(t+1), x(t)) \\ &\qquad + \frac{\alpha(t)}{2} \|g(t)\|_*^2 + \frac{1}{2\alpha(t)} \|x(t+1) - x(t)\|^2 \\ &\overset{(ii)}{\le} \frac{1}{\alpha(t)} D_\psi(x^*, x(t)) - \frac{1}{\alpha(t)} D_\psi(x^*, x(t+1)) + \frac{\alpha(t)}{2} \|g(t)\|_*^2 . \end{aligned}$$

The inequality (i) is a consequence of the Fenchel-Young inequality applied to the conjugates $\frac{1}{2}\|\cdot\|^2$ and $\frac{1}{2}\|\cdot\|_*^2$ (see, e.g., [8, Example 3.27]), while the inequality (ii) follows by the strong convexity of ψ, which gives $D_\psi(x(t+1), x(t)) \ge \frac{1}{2}\|x(t+1) - x(t)\|^2$.

Summing the final inequality, we obtain

$$\sum_{t=\tau+1}^{T} [F(x(t);\xi_t) - F(x^*;\xi_t)] \\ \leq \sum_{t=\tau+1}^{T} \frac{1}{\alpha(t)} [D_\psi(x^*, x(t)) - D_\psi(x^*, x(t+1))] + \sum_{t=\tau+1}^{T} \frac{\alpha(t)}{2} \|g(t)\|_*^2 .$$

Using the compactness assumption that $D_\psi(x^*, x) \leq \frac{1}{2}R^2$ for all $x \in \mathcal{X}$, we have

$$\sum_{t=\tau+1}^{T} \frac{1}{\alpha(t)} [D_\psi(x^*, x(t)) - D_\psi(x^*, x(t+1))] \\ \leq \sum_{t=\tau+2}^{T} D_\psi(x^*, x(t)) \left[\frac{1}{\alpha(t)} - \frac{1}{\alpha(t-1)}\right] + \frac{1}{\alpha(\tau+1)} D_\psi(x^*, x(\tau+1)) \\ \leq \frac{R^2}{2} \sum_{t=\tau+2}^{T} \left[\frac{1}{\alpha(t)} - \frac{1}{\alpha(t-1)}\right] + \frac{1}{2\alpha(\tau+1)} R^2 = \frac{R^2}{2\alpha(T)},$$

where for the last inequality we used that the stepsizes $\alpha(t)$ are non-increasing. ☐

Proof of Lemma 6.2. By the first-order condition for the optimality of $x(t+1)$ for the update (2.3), we have

$$\langle \alpha(t)g(t) + \nabla\psi(x(t+1)) - \nabla\psi(x(t)), x(t) - x(t+1)\rangle \geq 0.$$

Rewriting, we have

$$\langle \nabla\psi(x(t)) - \nabla\psi(x(t+1)), x(t) - x(t+1)\rangle \leq \alpha(t) \langle g(t), x(t) - x(t+1)\rangle \\ \leq \alpha(t) \|g(t)\|_* \|x(t) - x(t+1)\|$$

using Hölder's inequality. Simple algebra shows that

$$D_\psi(x(t), x(t+1)) + D_\psi(x(t+1), x(t)) = \langle \nabla\psi(x(t)) - \nabla\psi(x(t+1)), x(t) - x(t+1)\rangle ,$$

and by the assumed strong convexity of ψ, we see

$$\|x(t) - x(t+1)\|^2 \leq D_\psi(x(t+1), x(t)) + D_\psi(x(t), x(t+1)) \\ \leq \alpha(t) \|g(t)\|_* \|x(t) - x(t+1)\| .$$

Dividing by $\|x(t) - x(t+1)\|$ gives the desired result. ☐

Appendix B. Mixing and expected function values.

Proof of Lemma 6.3. Since $x \in \mathcal{F}_t$, we may integrate only against ξ when taking expectations, which yields

$$\mathbb{E}\left[f(x) - f(x^\star) - F(x;\xi_{t+\tau}) + F(x^\star;\xi_{t+\tau}) \mid \mathcal{F}_t\right] \\ = \int (F(x;\xi) - F(x^\star;\xi)) d\Pi(\xi) - \int (F(x;\xi) - F(x^\star;\xi)) dP_{[t]}^{t+\tau}(\xi).$$

Since we assume $P_{[s]}^t$ and Π have densities $p_{[s]}^t$ and π with respect to a measure μ, this difference becomes $\int (F(x;\xi) - F(x^\star;\xi))(\pi(\xi) - p_{[t]}^{t+\tau}(\xi)) d\mu(\xi)$. Setting $p = p_{[t]}^{t+\tau}$

for shorthand, we obtain

$$\begin{aligned}
&\left|\int (F(x;\xi) - F(x^\star;\xi))(\pi(\xi) - p(\xi))d\mu(\xi)\right| \le \int |F(x;\xi) - F(x^\star;\xi)|\,|p(\xi) - \pi(\xi)|\,d\mu(\xi) \\
&= \int |F(x;\xi) - F(x^\star;\xi)| \left(\sqrt{\pi(\xi)} + \sqrt{p(\xi)}\right)\left|\sqrt{\pi(\xi)} - \sqrt{p(\xi)}\right| d\mu(\xi) \\
&\le \sqrt{\int (F(x;\xi) - F(x^\star;\xi))^2 \left(\sqrt{\pi(\xi)} + \sqrt{p(\xi)}\right)^2 d\mu(\xi) \int \left(\sqrt{p(\xi)} - \sqrt{\pi(\xi)}\right)^2 d\mu(\xi)} \\
&= \sqrt{\int (F(x;\xi) - F(x^\star;\xi))^2 \left(\sqrt{\pi(\xi)} + \sqrt{p(\xi)}\right)^2 d\mu(\xi)}\, d_{\rm hel}(P_{[t]}^{t+\tau}, \Pi)
\end{aligned}$$

by Hölder's inequality. Applying the inequality $(a+b)^2 \le 2a^2 + 2b^2$, valid for $a, b \in \mathbb{R}$, we obtain the further bound

$$\begin{aligned}
&\left(2\int (F(x;\xi) - F(x^\star;\xi))^2 (\pi(\xi) + p(\xi))d\mu(\xi)\right)^{\frac{1}{2}} d_{\rm hel}(P_{[t]}^{t+\tau}, \Pi) = \qquad \text{(B.1)} \\
&\sqrt{2}\left(\mathbb{E}_\Pi[(F(x;\xi) - F(x^\star;\xi))^2] + \mathbb{E}[(F(x;\xi_{t+\tau}) - F(x^\star;\xi_{t+\tau}))^2 \mid \mathcal{F}_t]\right)^{\frac{1}{2}} d_{\rm hel}(P_{[t]}^{t+\tau}, \Pi).
\end{aligned}$$

To control the expectation terms in the bound (B.1), we now use Assumption A. By the (P-almost sure) convexity of the function $x \mapsto F(x;\xi)$, we observe that

$$F(x;\xi) - F(x^\star;\xi) \le \langle \mathsf{G}(x;\xi), x - x^\star\rangle \quad \text{and} \quad F(x^\star;\xi) - F(x;\xi) \le \langle \mathsf{G}(x^\star;\xi), x^\star - x\rangle.$$

Combining these two inequalities, we see that

$$\begin{aligned}
(F(x;\xi) - F(x^\star;\xi))^2 &\le \max\left\{\langle \mathsf{G}(x;\xi), x - x^\star\rangle^2, \langle \mathsf{G}(x^\star;\xi), x^\star - x\rangle^2\right\} \\
&\le \max\left\{\|\mathsf{G}(x;\xi)\|_*^2, \|\mathsf{G}(x^\star;\xi)\|_*^2\right\}\|x - x^\star\|^2 \\
&\le \max\left\{\|\mathsf{G}(x;\xi)\|_*^2, \|\mathsf{G}(x^\star;\xi)\|_*^2\right\} R^2,
\end{aligned}$$

where the last inequality uses our compactness assumption (2.2). Now we invoke Assumption A combined with the above inequality to obtain the further bound

$$\begin{aligned}
&\mathbb{E}\left[(F(x;\xi_{t+\tau}) - F(x^\star;\xi_{t+\tau}))^2 \mid \mathcal{F}_t\right] \\
&\qquad \le R^2\mathbb{E}\left[\|\mathsf{G}(x;\xi_{t+\tau})\|_*^2 + \|\mathsf{G}(x^\star;\xi_{t+\tau})\|_*^2 \mid \mathcal{F}_t\right] \le 2G^2R^2.
\end{aligned}$$

An analogous argument yields the same bound for the expectation under the stationary distribution, so based on our earlier bound (B.1) we have

$$\begin{aligned}
&\left|\int (F(x;\xi) - F(x^\star;\xi))\left(d\Pi(\xi) - dP_{[t]}^{t+\tau}(\xi)\right)\right| \\
&\le \int |F(x;\xi) - F(x^\star;\xi)|\left|d\Pi(\xi) - dP_{[t]}^{t+\tau}(\xi)\right| \le \sqrt{8G^2R^2}\, d_{\rm hel}\left(P_{[t]}^{t+\tau}, \Pi\right).
\end{aligned}$$

This completes the proof of the first statement of the lemma.

The second statement is simpler: apply Assumption B to obtain

$$\left|\int (F(x;\xi) - F(x^\star;\xi))(\pi(\xi) - p(\xi))d\mu(\xi)\right| \le GR\int |p(\xi) - \pi(\xi)|d\mu(\xi).$$

Observing that the above bound is equal to $GRd_{\text{TV}}\left(P_{[t]}^{t+\tau},\Pi\right)$ completes the proof. □

Proof of Lemma 6.4. For any x measurable with respect to the σ-field $\mathcal{F}_s$, we can define the function $h_{[s]}(x) = \mathbb{E}[F(x;\xi_{s+1}) \mid \mathcal{F}_s]$. Assumption A implies that $h_{[s]}$ is a G-Lipschitz continuous function so long as its argument is $\mathcal{F}_s$-measurable, that is, $|h_{[s]}(x) - h_{[s]}(y)| \le G\|x-y\|$ for $x, y \in \mathcal{F}_s$. In turn, this implies that

$$\begin{aligned}
&\mathbb{E}[F(x(t);\xi_{t+\tau}) - F(x(t+\tau);\xi_{t+\tau}) \mid \mathcal{F}_{t-1}] \\
&\quad= \sum_{s=t}^{t+\tau-1} \mathbb{E}[F(x(s);\xi_{t+\tau}) - F(x(s+1);\xi_{t+\tau}) \mid \mathcal{F}_{t-1}] \\
&\quad= \sum_{s=t}^{t+\tau-1} \mathbb{E}\left[\mathbb{E}[F(x(s);\xi_{t+\tau}) - F(x(s+1);\xi_{t+\tau}) \mid \mathcal{F}_{t+\tau-1}] \mid \mathcal{F}_{t-1}\right] \\
&\quad\le \sum_{s=t}^{t+\tau-1} \mathbb{E}\left[G\|x(s) - x(s+1)\| \mid \mathcal{F}_{t-1}\right]
\end{aligned}$$

since $x(s)$ is $\mathcal{F}_{t+\tau-1}$-measurable for $s \le t+\tau$. Now we apply Lemma 6.2, which shows that $\|x(s) - x(s+1)\| \le \alpha(s)\|g(s)\|_*$, and we have the further inequality

$$\mathbb{E}[F(x(t);\xi_{t+\tau}) - F(x(t+\tau);\xi_{t+\tau}) \mid \mathcal{F}_{t-1}] \le \sum_{s=t}^{t+\tau-1} G\alpha(s)\mathbb{E}[\|g(s)\|_* \mid \mathcal{F}_{t-1}].$$

Applying Jensen's inequality and Assumption A, we see that

$$\mathbb{E}[\|g(s)\|_* \mid \mathcal{F}_{t-1}] \le \sqrt{\mathbb{E}[\mathbb{E}[\|g(s)\|_*^2 \mid \mathcal{F}_{s-1}] \mid \mathcal{F}_{t-1}]} \le \sqrt{G^2} = G.$$

In conclusion, we have the first statement of the lemma:

$$\mathbb{E}[F(x(t);\xi_{t+\tau}) - F(x(t+\tau);\xi_{t+\tau}) \mid \mathcal{F}_{t-1}] \le G^2 \sum_{s=t}^{t+\tau-1} \alpha(s) \le G^2\tau\alpha(t),$$

since the sequence $\alpha(t)$ is non-increasing. The proof of the second statement is entirely similar, but we do not need to apply conditional expectations. □

Appendix C. Martingale concentration.

Proof of Proposition 6.5. We construct a family of τ different martingales from the summation in the statement of the proposition, each of which we control with high probability. Applying a union bound gives us control on the deviation of the entire series. We begin by defining the random variables

$$Z_t := f(x(t-\tau+1)) - F(x(t-\tau+1);\xi_t) + F(x^\star;\xi_t) - f(x^\star),$$

noting that

$$\sum_{t=\tau}^{T} Z_t = \sum_{t=1}^{T-\tau+1} \left[f(x(t)) - F(x(t);\xi_{t+\tau-1}) + F(x^\star;\xi_{t+\tau-1}) - f(x^\star)\right].$$

By defining the filtration of σ-fields $\mathcal{A}_i^j = \mathcal{F}_{\tau i+j}$ for $j = 1, \ldots, \tau$, we can construct a set of Doob martingales $\{X_1^j, X_2^j, \ldots\}$ for $j = 1, \ldots, \tau$ by making the definition

$$\begin{aligned}
X_i^j &:= Z_{\tau i+j} - \mathbb{E}[Z_{\tau i+j} \mid \mathcal{A}_{i-1}^j] = Z_{\tau i+j} - \mathbb{E}[Z_{\tau i+j} \mid \mathcal{F}_{\tau(i-1)+j}] \\
&= f(x(\tau(i-1)+j+1)) - F(x(\tau(i-1)+j+1);\xi_{\tau i+j}) \\
&\qquad + F(x^\star;\xi_{\tau i+j}) - f(x^\star) - \mathbb{E}[Z_t \mid \mathcal{F}_{\tau(i-1)+j}].
\end{aligned}$$

By inspection, X_i^j is measurable with respect to the σ-field $\mathcal{A}_i^j$, and $\mathbb{E}[X_i^j \mid \mathcal{A}_{i-1}^j] = 0$. So, for each j, the sequence $\{X_i^j : i = 1, 2, \ldots\}$ is a martingale difference sequence adapted to the filtration $\{\mathcal{A}_i^j : i = 1, 2, \ldots\}$. Define the index set $\mathcal{I}(j)$ to be the indices $\{1, \ldots, \lfloor T/\tau \rfloor + 1\}$ for $j \le T - \tau \lfloor T/\tau \rfloor$ and $\{1, \ldots, \lfloor T/\tau \rfloor\}$ otherwise. With the definition of X_i^j and the indices $\mathcal{I}(j)$, we see that

$$\sum_{t=\tau}^{T} Z_t = \sum_{j=1}^{\tau} \sum_{i \in \mathcal{I}(j)} X_i^j + \sum_{t=\tau}^{T} \mathbb{E}[Z_t \mid \mathcal{F}_{t-\tau}] = \sum_{j=1}^{\tau} \sum_{i=1}^{|\mathcal{I}(j)|} X_i^j + \sum_{t=\tau}^{T} \mathbb{E}[Z_t \mid \mathcal{F}_{t-\tau}]. \quad \text{(C.1)}$$

Now we note the following important fact: by the compactness assumption (6.1) and Assumption B, the $\mathcal{F}_{\tau(i-1)+j}$-measurability of $f(x(\tau(i-1)+j+1))$ implies

$$|X_i^j| = \left| Z_{\tau i + j} - \mathbb{E}[Z_{\tau i + j} \mid \mathcal{F}_{\tau(i-1)+j}] \right| \le 2GR.$$

This bound, coupled with the representation (C.1), shows that $\sum_{t=\tau}^{T} Z_t$ is a sum of τ different bounded-difference martingales plus a sum of conditional expectations that we will bound later. To control the martingale portion of the sum (C.1), we apply the triangle inequality, a union bound, and Azuma's inequality [2] to find

$$P\left(\sum_{j=1}^{\tau} \sum_{i \in \mathcal{I}(j)} X_i^j > \gamma \right) \le \sum_{j=1}^{\tau} P\left(\sum_{i \in \mathcal{I}(j)} X_i^j > \frac{\gamma}{\tau} \right) \le \sum_{j=1}^{\tau} \exp\left(-\frac{\gamma^2}{16 G^2 R^2 \tau T} \right),$$

since there are fewer than $2T/\tau$ terms in each of the sums X_i^j (by our assumption that $T/2 \ge \tau$). Substituting $\gamma = 4GR\sqrt{T\tau \log(\tau/\delta)}$, we find

$$P\left(\sum_{j=1}^{\tau} \sum_{i \in \mathcal{I}(j)} X_i^j > 4GR\sqrt{T\tau \log \frac{\tau}{\delta}} \right) \le \delta.$$

To bound the final term $\mathbb{E}[Z_t \mid \mathcal{F}_{t-\tau}]$ in the sum (C.1), we recall from Lemma 6.3 that

$$|\mathbb{E}[Z_t \mid \mathcal{F}_{t-\tau}]| \le GR \cdot d_{\text{TV}}\left(P_{[t-\tau]}^{t}, \Pi \right).$$

Summing this bound completes the proof. $\square$

Appendix D. Probabilistic Mixing.

Proof of Lemma 4.2. Using the definitions in the statement of the lemma, take

$$\tau = \lfloor \tau_{\text{TV}}(P, \epsilon) + \kappa c \rfloor \ge \frac{\log \frac{1}{\epsilon}}{|\log \gamma|} + \frac{\log K}{|\log \gamma|} + \frac{c}{|\log \gamma|},$$

which implies by Markov's inequality that

$$\mathbb{P}\left(d_{\text{TV}}\left(P_{[t]}^{t+\tau}, \Pi \right) \ge \epsilon \right) \le \frac{K\gamma^\tau}{\epsilon} \le \frac{K \exp(-\log \frac{1}{\epsilon}) \exp(-\log K)}{\epsilon} \exp(-c) = \exp(-c)$$

since $\gamma^{a/|\log \gamma|} = \exp(-a)$ for $0 < \gamma < 1$. Noting that

$$\mathbb{P}\left(\tau_{\text{TV}}(P_{[t]}, \epsilon) > \tau \right) \le \mathbb{P}\left(d_{\text{TV}}\left(P_{[t]}^{t+\tau}, \Pi \right) > \epsilon \right)$$

for any $\tau \in \mathbb{N}$ completes the proof. $\square$

Proof of Lemma 6.6. We use a covering number argument, which is common in uniform concentration inequalities in probability theory (e.g., [37]). For each $t \in \{1, \ldots, T\}$, define

$$\epsilon_t := \inf \{\epsilon > 0 : \tau_{\rm TV}(P, \epsilon) \le t\}.$$

By the right-continuity of $\epsilon \mapsto \tau_{\rm TV}(P, \epsilon)$, we have $\tau_{\rm TV}(P, \epsilon_t) \le t$ but $\tau_{\rm TV}(P, \epsilon_t - \delta) > t$ for any $\delta > 0$. As a consequence, we see that for some $\epsilon \ge \epsilon_T$ to exist satisfying $\tau_{\rm TV}(P_{[s]}, \epsilon) > \tau_{\rm TV}(P, \epsilon) + c$, it must be the case that

$$\tau_{\rm TV}(P_{[s]}, \epsilon_t) - \tau_{\rm TV}(P, \epsilon_t) > c$$

for some ϵ_t, where $t \in \{1, \ldots, T\}$. That is, we have

$$P\left(\tau_{\rm TV}(P_{[s]}, \epsilon) > \tau_{\rm TV}(P, \epsilon) + c \text{ for some } s \in \{1, \ldots, T\} \text{ and } \epsilon \ge \epsilon_T\right) \\ \le P\left(\max_{t,s \le T} \left[\tau_{\rm TV}(P_{[s]}, \epsilon_t) - \tau_{\rm TV}(P, \epsilon_t)\right] > c\right).$$

Applying a union bound and Assumption D, we thus see that for any $c \ge 0$,

$$P\left(\max_{s \le T} \sup_{\epsilon \ge \epsilon_T} \left(\tau_{\rm TV}(P_{[s]}, \epsilon) - \tau_{\rm TV}(P, \epsilon)\right) > c\right) \\ \le T^2 \max_{t,s \le T} P\left(\tau_{\rm TV}(P_{[s]}, \epsilon_t) > \tau_{\rm TV}(P, \epsilon) + c\right) \le T^2 \exp(-c/\kappa).$$

Setting the final equation equal to δ and solving, we obtain $c = \kappa[\log(1/\delta) + 2\log(T)]$, which is equivalent to the statement of the lemma. $\square$

REFERENCES

[1] A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright, *Information-theoretic lower bounds on the oracle complexity of convex optimization*, IEEE Transactions on Information Theory, 58 (2012), pp. 3235–3249.

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

[3] A. Beck and M. Teboulle, *Mirror descent and nonlinear projected subgradient methods for convex optimization*, Operations Research Letters, 31 (2003), pp. 167–175.

[4] A. Ben-Tal, T. Margalit, and A. Nemirovski, *The ordered subsets mirror descent optimization method with applications to tomography*, SIAM Journal on Optimization, 12 (2001), pp. 79–108.

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

[6] P. Billingsley, *Probability and Measure*, Wiley, Second ed., 1986.

[7] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, *Randomized gossip algorithms*, IEEE Transactions on Information Theory, 52 (2006), pp. 2508–2530.

[8] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[9] R. C. Bradley, *Basic properties of strong mixing conditions. a survey and some open questions*, Probability Surveys, 2 (2005), pp. 107–144.

[10] F. R. K. Chung, *Spectral Graph Theory*, AMS, 1998.

[11] C. Cortes and V. Vapnik, *Support-vector networks*, Machine Learning, 20 (1995), pp. 273–297.

[12] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, 1991.

[13] I. Csiszár, *Information-type measures of difference of probability distributions and indirect observation*, Studia Scientifica Mathematica Hungary, 2 (1967), pp. 299–318.

[14] J. C. Duchi, A. Agarwal, M. Johansson, and M. Jordan, *Ergodic mirror descent*. URL `http://arxiv.org/abs/1105.4681`, 2011.

[15] J. C. Duchi, A. Agarwal, and M. J. Wainwright, *Dual averaging for distributed optimization: convergence analysis and network scaling*, IEEE Transactions on Automatic Control, 57 (2012), pp. 592–606.

[16] A. Gelman and D. B. Rubin, *Inference from iterative simulation using multiple sequences*, Statistical Science, 7 (1992), pp. 457–472.

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

[18] R. Impagliazzo and D. Zuckerman, *How to recycle random bits*, in 30th Annual Symposium on Foundations of Computer Science, 1989, pp. 248–253.

[19] M. Jerrum and A. Sinclair, *The Markov chain Monte Carlo method: an approach to approximate counting and integration*, in Approximation Algorithms for NP-hard Problems, D. S. Hochbaum, ed., PWS Publishing, 1996.

[20] B. Johansson, M. Rabi, and M. Johansson, *A randomized incremental subgradient method for distributed optimization in networked systems*, SIAM Journal on Optimization, 20 (2009), pp. 1157–1170.

[21] H. J. Kushner and G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, Springer, Second ed., 2003.

[22] V. Lesser, C. Ortiz, and M. Tambe, eds., *Distributed Sensor Networks: A Multiagent Perspective*, vol. 9, Kluwer Academic Publishers, 2003.

[23] E. Liebscher, *Towards a unified approach for proving geometric ergodicity and mixing properties of nonlinear autoregressive processes*, Journal of Time Series Analysis, 26 (2005), pp. 669–689.

[24] J. Matousek, *Lectures on Discrete Geometry*, Springer, 2002.

[25] C. Mesterharm, *On-line learning with delayed feedback*, in Algorithmic Learning Theory, 2005, pp. 399–413.

[26] S. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, Cambridge University Press, Second ed., 2009.

[27] A. Mokkadem, *Mixing properties of ARMA processes*, Stochastic Processes and their Applications, 29 (1988), pp. 309–315.

[28] A. Nedić and D. P. Bertsekas, *Incremental subgradient methods for nondifferentiable optimization*, SIAM Journal on Optimization, 12 (2001), pp. 109–138.

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

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

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

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

[33] S. S. Ram, A. Nedić, and V. V. Veeravalli, *Incremental stochastic subgradient algorithms for convex optimization*, SIAM Journal on Optimization, 20 (2009), pp. 691–717.

[34] H. Robbins and S. Monro, *A stochastic approximation method*, Annals of Mathematical Statistics, 22 (1951), pp. 400–407.

[35] R. T. Rockafellar and R. J. B. Wets, *On the interchange of subdifferentiation and conditional expectation for convex functionals*, Stochastics: An International Journal of Probability and Stochastic Processes, 7 (1982), pp. 173–182.

[36] J. C. Spall, *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*, Wiley, 2003.

[37] V. N. Vapnik and A. Y. Chervonenkis, *On the uniform convergence of relative frequencies of events to their probabilities*, Theory of Probability and its applications, XVI (1971), pp. 264–280.

[38] G. Wei and M. A. Tanner, *A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms*, Journal of the American Statistical Association, 85 (1990), pp. 699–704.

[39] B. Yu, *Rates of convergence for empirical processes of stationary mixing sequences*, Annals of Probability, 22 (1994), pp. 94–116.

[40] M. Zinkevich, *Online convex programming and generalized infinitesimal gradient ascent*, in Proceedings of the Twentieth International Conference on Machine Learning, 2003.