Ergodic Mirror Descent

Article *in* SIAM Journal on Optimization · May 2011

DOI: 10.1109/Allerton.2011.6120236 · Source: arXiv

4 authors, including:

Mikael Johansson
KTH Royal Institute of Technology
226 PUBLICATIONS **5,960** CITATIONS

SEE PROFILE

Michael Jordan
University of California, Berkeley
636 PUBLICATIONS **74,568** CITATIONS

SEE PROFILE

Ergodic Subgradient Descent

John C. Duchi* Alekh Agarwal* Mikael Johansson† Michael I. Jordan*‡

May 25, 2011

Abstract

We generalize stochastic subgradient descent methods to situations in which we do not receive independent samples from the distribution over which we optimize, but instead receive samples that are 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, 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 with common closed convex domain $\mathcal{X} \subseteq \mathbb{R}^d$. 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

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

We consider solving the following problem:

$$\min_x \; f(x) \quad \text{subject to} \quad x \in \mathcal{X}. \tag{2}$$

Though a wide variety of stochastic optimization methods for solving the problem (2) have been explored in an extensive literature [RM51, PJ92, NB01, NJLS09], most approaches have imposed the restrictive assumption that it is possible to obtain independent and identically distributed samples ξ from the distribution Π. We relax this assumption and instead assume that we receive 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

*Department of Electrical Engineering and Computer Sciences, University of California, Berkeley; Berkeley, CA USA. Email: {jduchi,alekh,jordan}@eecs.berkeley.edu. JCD was supported by an NDSEG fellowship, and AA was supported by a Microsoft Research Fellowship.

†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 distribution Π is unknown—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 space or is a combinatorial space, but it is possible to design Markov chains that converge to the distribution Π [JS96]. 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 [IZ89].

Our approach to solving the problem (2) is related to classical stochastic gradient descent algorithms [RM51, PJ92], where one assumes access to samples ξ from the distribution Π and performs gradient updates using $\nabla F(x; \xi)$. This reduces to the randomized incremental subgradient method of Nedić and Bertsekas [NB01] when Π is concentrated on a set of n points, giving an objective of the form $f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x)$. More generally, our problem belongs to the family of stochastic problems with exogenous correlated noise (see, e.g. [KY03] and the numerous references therein), where the goal is to minimize $\mathbb{E}_{\Pi}[F(x; \xi)]$ as in the objective (2) but we have access only to samples ξ that are not independent over time. Classical results in this setting are asymptotic in nature and generally do not provide finite sample or high-probability convergence guarantees. Our method borrows from standard stochastic subgradient methodology [NY83, NJLS09], but we generalize them 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. In spite of the new setting, we do not modify the standard stochastic gradient algorithms; our algorithm receives samples ξ_t and takes gradient (or mirror descent) steps with respect to the subgradients of $F(x; \xi_t)$. Consequently, as we show more specifically in Section 4, our approach generalizes several recent works on stochastic and non-stochastic optimization, including the randomized incremental subgradient method [NB01] as well as the Markov incremental subgradient method [JRJ09].

The main result of this paper is that performing stochastic gradient steps as described in the previous paragraph results in a provably convergent optimization procedure. The convergence is governed by problem-dependent terms (namely the radius of $\mathcal{X}$ and the Lipschitz constant of the functions F) familiar from previous results on stochastic methods [NB01, Zin03, NJLS09] as well as terms dependent on the rate at which the stochastic process $\xi_1, \xi_2, \ldots$ converges to the stationary distribution. Our two main theorems characterize the convergence rate of Ergodic Mirror Descent in terms of τ_{mix}, which is the time it takes the process ξ_t to converge to the stationary distribution Π (in a sense we make precise later) both in expectation and with high probability. 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.

The remainder of the paper is organized as follows. The next section contains our main assumptions and a description of the algorithm we analyze. Following that, we collect our main technical results in Section 3. We expand on these results in corollaries and examples throughout Section 4, and give complete proofs of all our results in Section 5 and the appendices.

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

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

For a convex function f, we let $\partial f(x) = \{g \in \mathbb{R}^d \mid f(y) \geq f(x) + \langle g, y - x\rangle\}$ 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 $\{1, \ldots, n\}$.

2 Assumptions and Algorithm

We now turn to describing our algorithm and the assumptions underlying it. Our main assumption is on the Lipschitz continuity properties of the functions $F(\cdot; \xi)$.

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

$$|F(x; \xi) - F(y; \xi)| \leq G \|x - y\| \tag{3}$$

for all $x, y \in \mathcal{X}$.

Note that as a consequence of Assumption A, for any $g \in \partial F(x; \xi)$ we have that $\|g\|_* \leq G$ (e.g., [HUL96]), and it is clear that f is also G-Lipschitz.

Our algorithm is a generalization of the stochastic mirror descent algorithm [NY83, BT03, NJLS09], which in turn generalizes gradient descent to elegantly address non-Euclidean geometry. The algorithm is based on a prox-function ψ, which is a differentiable convex function defined on $\mathcal{X}$ that is assumed (w.l.o.g. by scaling) to be strongly convex with respect to the norm $\|\cdot\|$ over $\mathcal{X}$. If we define the Bregman divergence in the usual way, that is, $D_\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla\psi(y), x - y\rangle$, we have

$$D_\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla\psi(y), x - y\rangle \geq \frac{1}{2} \|x - y\|^2 . \tag{4}$$

Note that with the choice $\psi(x) = \frac{1}{2}\|x\|_2^2$, we have $D_\psi(x, y) = \frac{1}{2}\|x - y\|_2^2$, in which case the divergence D_ψ can essentially be viewed as a squared distance between points x and y.

Now we turn to a description of the algorithm. The algorithm is an iterative algorithm that maintains a parameter $x(t) \in \mathcal{X}$, which it updates using stochastic gradient information to form $x(t + 1)$. Specifically, let P^t denote the distribution of the stochastic process P at time t. We assume that we receive a sample $\xi_t \sim P^t$ at each time step t. Given ξ_t, the Ergodic Mirror Descent (EMD) algorithm then computes the update

$$g(t) \in \partial F(x(t); \xi_t), \quad x(t + 1) = \operatorname*{argmin}_{x\in\mathcal{X}} \left\{ \langle g(t), x\rangle + \frac{1}{\alpha(t)} D_\psi(x, x(t)) \right\} . \tag{5}$$

Here $\alpha(t)$ is a (time-dependent) stepsize. Note that the algorithm (5) reduces to projected gradient descent with the choice $\psi(x) = \frac{1}{2}\|x\|_2^2$.

3 Main Results

To state our main results, we need to recall some standard definitions from probability theory (cf. [Bil86]). We measure the convergence of the stochastic process P by convergence in total

variation. The total variation distance between distributions P and Q defined on the same space S, each with densities p and q with respect to an underlying measure μ,[1] is given by

$$d_{\mathrm{TV}}(P,Q) := \sup_{A \subset S} |P(A) - Q(A)| = \frac{1}{2}\int_S |p(s) - q(s)| d\mu(s). \tag{6}$$

Now, define the σ-field $\mathcal{F}_t = \sigma(\xi_1, \ldots, \xi_t)$. Let $P^t_{[s]}$ denote the distribution of ξ_t conditioned on $\mathcal{F}_s$, that is, given the initial samples $\xi_1, \ldots, \xi_s$. We measure convergence in terms of the mixing time of the different $P^t_{[s]}$, defined as follows. In the definition, let $p^t_{[s]}$ and π denote the densities of $P^t_{[s]}$ and Π, respectively.

Definition 3.1. *The* mixing time $\tau_{\mathrm{mix}}(P_{[s]}, \epsilon)$ *of the sampling distribution 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*

$$d_{\mathrm{TV}}(P^{t+s}_{[s]}, \Pi) \le \frac{\epsilon}{2}, \quad \textit{i.e.} \quad \tau_{\mathrm{mix}}(P_{[s]}, \epsilon) := \inf\left\{ t - s \mid t \in \mathbb{N}, \ \int_\Xi \left| p^t_{[s]}(\xi) - \pi(\xi) \right| d\mu(\xi) \le \epsilon \right\}.$$

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

We make the following uniformity assumption on the mixing times of our stochastic process with distribution P.

Assumption B. *The mixing times of the stochastic process (ξ_i) are uniform in the sense that there exists a uniform mixing time $\tau_{\mathrm{mix}}(P, \epsilon) < \infty$ such that with probability 1,*

$$\tau_{\mathrm{mix}}(P, \epsilon) \ge \tau_{\mathrm{mix}}(P_{[s]}, \epsilon)$$

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

Assumption B is a weaker version of the common assumption of ϕ-mixing in the probability literature (e.g. [Mar98, Sam00, Bra05]), but ϕ-mixing requires convergence of the process over the entire tail σ-field $\sigma(\xi_t, \xi_{t+1}, \ldots)$ of the process ξ_t. For example, any finite state-space Markov chain satisfies the above assumption, as do uniformly ergodic Markov chains on general state spaces [MT09]. It is possible to consider weakening this assumption, and we intend to do this in future work.

With the Assumptions A and B in place, we can now give our main convergence results. Our first result gives convergence in expectation of the EMD algorithm (5), with the proof given in Section 5.2.

Theorem 1. *Let Assumptions A and B hold and let $x(t)$ be defined by the EMD update* (5) *with stepsize sequence $\alpha(t)$. In addition assume that $D_\psi(x^*, x(t)) \le R^2/2$ for all t, where x^* is the optimum of $f(x)$. Then*

$$\mathbb{E}\left[\sum_{t=1}^T f(x(t)) - f(x^*) \right] \le \frac{R^2}{2\alpha(T)} + \frac{G^2}{2}\sum_{t=1}^T \alpha(t) + T\epsilon GR + 2\tau_{\mathrm{mix}}(P,\epsilon)G^2 \sum_{t=1}^T \alpha(t) + 2\tau_{\mathrm{mix}}(P,\epsilon)RG,$$

where the expectation is taken with respect to the random samples $\xi_1, \ldots, \xi_T$.

[1] This assumption is without loss, since P and Q are each absolutely continuous with respect to the measure $P+Q$.

As an immediate corollary to the above theorem, obtained by applying Jensen's inequality to the convex function f, we have

Corollary 1. *Define* $\widehat{x}(T) = \frac{1}{T}\sum_{t=1}^{T} x(t)$. *Under the conditions of Theorem 1,*

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

We can also show that the results of Theorem 1 and Corollary 1 hold with high probability. We provide the proof of this theorem in Section 5.3.

Theorem 2. *Under the conditions of Theorem 1, we have with probability at least* $1-\delta$ *that*

$$\sum_{t=1}^{T} f(x(t)) - f(x^*) \leq \frac{R^2}{2\alpha(T)} + \frac{G^2}{2}\sum_{t=1}^{T}\alpha(t) + T\epsilon GR + 2\tau_{\text{mix}}(P,\epsilon)G^2\sum_{t=1}^{T}\alpha(t) + 2\tau_{\text{mix}}(P,\epsilon)RG + 6GR\sqrt{T\tau_{\text{mix}}(P,\epsilon)\log\tau_{\text{mix}}(P,\epsilon)\log\frac{1}{\delta}}.$$

Consequently, by defining $\widehat{x}(T) = \frac{1}{T}\sum_{t=1}^{T} x(t)$, *with probability at least* $1-\delta$

$$f(\widehat{x}(T)) - f(x^*) \leq \frac{R^2}{2T\alpha(T)} + \frac{G^2}{2T}\sum_{t=1}^{T}\alpha(t) + \epsilon GR + \frac{2\tau_{\text{mix}}(P,\epsilon)G^2}{T}\sum_{t=1}^{T}\alpha(t) + \frac{2\tau_{\text{mix}}(P,\epsilon)RG}{T} + 6GR\sqrt{\frac{\tau_{\text{mix}}(P,\epsilon)\log\tau_{\text{mix}}(P,\epsilon)\log\frac{1}{\delta}}{T}}.$$

The rate of convergence that is guaranteed by Theorem 2 is identical to that obtained in Theorem 1 plus an additional term which arises as a result of the control of the deviation of the ergodic process around its expectation. Notably, the high probability guarantee provided by Theorem 2 on the deviation is $\mathcal{O}(\sqrt{\tau_{\text{mix}}(P,\epsilon)/T})$—up to logarithmic factors—and the dominant terms in the convergence rates also appear in the expected bounds in Theorem 1.

The bounds in the above results may appear somewhat complex, so we turn to a slight specialization to build intuition and attain a simplified statement of convergence rates. Theorems 1 and 2, as stated, 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 [MT09]; that is, there exists a constant $\kappa(P)$ such that $\tau_{\text{mix}}(P,\epsilon) \leq \kappa(P)\log(1/\epsilon)$. We have the following corollary for this special case.

Corollary 2 (Geometric mixing)**.** *Under the conditions of Theorem 1, assume further that* $\tau_{\text{mix}}(P,\epsilon) \leq \kappa(P)\log(1/\epsilon)$. *Then the EMD update* (5) *with stepsize* $\alpha(t) = \alpha/\sqrt{t}$ *satisfies*

$$\mathbb{E}\left[f(\widehat{x}(T)) - f(x^*)\right] \leq \frac{R^2}{2\alpha\sqrt{T}} + \frac{\alpha G^2}{\sqrt{T}}\Big(4\kappa(P)\log\frac{1}{\epsilon} + 1\Big) + \epsilon GR + \frac{2\kappa(P)\log\frac{1}{\epsilon}RG}{T},$$

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

$$f(\widehat{x}(T)) - f(x^*) \leq \frac{R^2}{2\alpha\sqrt{T}} + \frac{\alpha G^2}{\sqrt{T}}\Big(4\kappa(P)\log\frac{1}{\epsilon} + 1\Big) + \epsilon GR + \frac{2\kappa(P)\log\frac{1}{\epsilon}RG}{T} + 6GR\sqrt{\frac{\kappa(P)\log\frac{1}{\epsilon}\log(\kappa(P)\log\frac{1}{\epsilon})\log\frac{1}{\delta}}{T}}.$$

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

$$\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \le 1 + \int_1^T t^{-1/2} dt = 2\sqrt{T} - 1 < 2\sqrt{T}, \tag{7}$$

we have $\sum_{t=1}^{T} \alpha(t) \le 2\alpha\sqrt{T}$. The bounds in the corollary now follow from Theorems 1 and 2. □

We make a few remarks on the corollary. First, the $\kappa(P)\log(1/\epsilon)/T$ term is of smaller order than the other terms in the bounds, so we can essentially ignore it. Secondly, an appropriate choice of the mixing parameter ϵ and stepsize multiplier α yields a simplified convergence rate. In particular, choosing $\epsilon = T^{-1/2}$ and $\alpha = R/(G\sqrt{\kappa(P)\log T}$ in the corollary yields

$$f(\widehat{x}(T)) - f(x^*) = \mathcal{O}\left(\frac{RG\sqrt{\tau_{\mathrm{mix}}(P, T^{-1/2})}}{\sqrt{T}}\right) = \mathcal{O}\left(\frac{RG\sqrt{\kappa(P)\log T}}{\sqrt{T}}\right) \tag{8}$$

both in expectation and (modulo an additional $\log(\kappa(P)\log T)$ factor) with high probability. In the classical setting [NJLS09] of i.i.d. samples $\xi \sim \Pi$, stochastic gradient descent and its mirror descent generalizations attain convergence rates of $\mathcal{O}(RG/\sqrt{T})$. Since $\tau_{\mathrm{mix}}(P, 0) = 1$ for an i.i.d. process, the rate (8) shows that our results subsume existing results for i.i.d. noise. In addition, the step-size choice $\alpha(t) = \alpha/\sqrt{t}$ is robust—in a way similarly noted by Nemirovski et al. [NJLS09]—for quickly mixing ergodic processes: mis-specification of α leads to a penalty in convergence that is essentially linear in $\max\{\alpha^{-1}, \alpha\}\sqrt{\log(T)}$. Thus, up to logarithmic factors, EMD has qualitative convergence behavior similar to stochastic mirror descent, but for a much broader class of ergodic processes. General ergodic processes do not always enjoy the geometric mixing assumed in Corollary 2, and we present an example of a more slowly (polynomially) mixing process in Section 4.4 in the sequel. We give finite sample convergence results for slower mixing processes there as well, which unsurprisingly are slower than the $\mathcal{O}(T^{-1/2})$ rate established in the corollary.

4 Examples and Consequences

We now collect several examples using the convergence rates of Theorems 1 and 2 to provide insight into the theoretical statements. We begin with a concrete example and move toward more abstract principles in the following three, completing the section with finite sample rates and asymptotic convergence guarantees for more slowly mixing ergodic processes. Most of the results are new or improve over previously known bounds.

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

The Markov incremental gradient descent (MIGD) procedure due to Johansson et al. [JRJ09] is a generalization of Nedić and Bertsekas's randomized incremental subgradient method [NB01]. The motivation for the algorithm comes from a distributed optimization algorithm using a simple (locally computable) peer-to-peer communication scheme. In this setting, 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

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

The procedure works as follows. A token $i(t)$ moves among the processors in the network along with the current set of parameters $x(t) \in \mathcal{X}$, with processor $i(t) \in [n]$ being chosen at iteration t. The algorithm then computes the update

$$g(t) \in \partial f_{i(t)}(x(t)), \quad x(t+1) = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \langle g(t), x \rangle + \frac{1}{\alpha(t)} D_\psi(x, x(t)) \right\}. \tag{9}$$

(This is a strict generalization of [JRJ09], who assume $\psi(x) = \frac{1}{2}\|x\|_2^2$.)

Abstractly, we view the token as evolving according to a Markov chain with doubly-stochastic transition matrix P so that its stationary distribution is the uniform distribution. In this case,

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

With this setup, we have that $\xi_t = i(t)$, and the total variation distance of the stochastic process initialized at $i(0) = i$ is $\frac{1}{2}\left\|P^t e_i - \mathbb{1}/n\right\|_1$, 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

$$\left\|P^t e_i - \mathbb{1}/n\right\|_1 \le \sqrt{n}\left\|P^t e_i - \mathbb{1}/n\right\|_2 = \sqrt{n}\left\|P^t(e_i - \mathbb{1})\right\|_2 \le \sqrt{n}\rho_2(P)^t \left\|e_i - \mathbb{1}/n\right\|_2 \le \sqrt{n}\rho_2(P)^t, \tag{10}$$

where $\rho_2(P)$ denotes the second singular value of the matrix P. From the spectral bound (10) on the total variation distance, we see that if $t \ge \frac{\frac{1}{2}\log(Tn)}{\log \rho_2(P)^{-1}}$, we have $\left\|P^t e_i - \mathbb{1}/n\right\|_1 \le \frac{1}{\sqrt{T}}$. In particular, in the notation of Assumption B,

$$\tau_{\mathrm{mix}}(P, T^{-1/2}) \le \frac{\log(Tn)}{2\log \rho_2(P)^{-1}} \le \frac{\log(Tn)}{2(1-\rho_2(P))}. \tag{11}$$

The second inequality uses the concavity of the log, which shows that $-\log \rho_2(P) \ge 1 - \rho_2(P)$. Consequently, we have the following result, similar to Corollary 2.

Corollary 3. *Let $x(t)$ evolve according to the Markov incremental descent update (9), where $i(t)$ evolves via the transition matrix P and $\alpha(t) = \alpha/\sqrt{t}$. Define $\widehat{x}(T) = \frac{1}{T}\sum_{t=1}^T x(t)$ and choose stepsize multiplier $\alpha = \frac{R(1-\rho_2(P))}{G\sqrt{\log(Tn)}}$. Then*

$$\mathbb{E}[f(\widehat{x}(T))] - f(x^*) \le 5\frac{RG}{\sqrt{T}} \cdot \frac{\sqrt{\log(Tn)}}{\sqrt{1-\rho_2(P)}} + \frac{RG}{T} \cdot \frac{\log(Tn)}{1-\rho_2(P)} \tag{12}$$

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

$$f(\widehat{x}(T)) - f(x^*) \le 5\frac{RG}{\sqrt{T}} \cdot \frac{\sqrt{\log(Tn)}}{\sqrt{1-\rho_2(P)}} + \frac{RG}{T} \cdot \frac{\log(Tn)}{1-\rho_2(P)} + 4GR\sqrt{\frac{\log\frac{1}{\delta}}{T}} \cdot \sqrt{\frac{\log(Tn)\log(\frac{\log(Tn)}{1-\rho_2(P)})}{1-\rho_2(P)}}$$

Proof The proof is a straightforward consequence of Theorems 1 and 2 and Corollary 2. We use the uniform bound (11) on the mixing time of the random walk, and the result follows from simple algebra. □

Examining the results in Corollary 3, we see that they are sharper and more powerful than the analysis in the original Markov incremental gradient descent paper [JRJ09]. First, our results allow us to use mirror descent updates, thus applying to problems having non-Euclidean geometry. Secondly, because we base our convergence analysis on mixing time rather than return times, we can give sharp high-probability convergence guarantees. Thirdly, our analysis does not require a fixed setting of the stepsize $\alpha(t)$ for all times (though our results also hold under such a setting), which gives us an "anytime" algorithm: the procedure does not need to know the number of iterations T that it will run.

Finally, our expected convergence rates are tighter. As also discussed by Duchi et al. [DAW11], Johansson et al. show that MIGD—under the assumption that T is known and α is optimally chosen—has convergence rate $\mathcal{O}(RG \max_i \sqrt{\frac{n\Gamma_{ii}}{T}})$, where Γ is the return time matrix given by $\Gamma = (I - P + \mathbb{1}\mathbb{1}^\top/n)^{-1}$. Since Johansson et al. assume P is symmetric, the eigenvalues of Γ are 1 and $1/(1-\lambda_i(P))$ for $i > 1$, and

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

Thus, up to logarithmic factors, the bound (12) from Corollary 3 is never weaker. For well-connected graphs, the bound is substantially stronger; for example, random walks on expander graphs (e.g., [Chu98]) have constant spectral gaps, 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 Optimization over combinatorial spaces

For our second example, we retain the general form of the objective (1), but we assume that Ξ is a combinatorial space from which it is difficult to obtain uniform samples but for which we can construct a Markov chain that converges quickly to the uniform distribution over Ξ. See Jerrum and Sinclair [JS96] for an overview of such problems.

As our concrete motivating example, consider the statistical problem of learning a ranking function for web searches. The statistician receives information in the form of a user's clicks on particular search results, which impose a partial order on the results (since only a few are clicked on). We would like the resulting ranking function to be oblivious to the order of the remaining results, which leads us to define Ξ to be the set of all total orders of the search results consistent with the partial order imposed by the user. Certainly the set Ξ is exponentially large; it is also challenging to draw a uniform sample from it.

Though sampling is challenging, it is possible to develop a rapidly-mixing Markov chain whose stationary distribution is uniform on Ξ. Specifically, Karzanov and Khachiyan [KK91] develop the following Markov chain. Let $\mathcal{P}$ be a partial order on the set $[n]$, whose elements are of the form $i \prec j$ for $i, j \in [n]$. The states of the Markov chain are permutations σ of $[n]$ respecting the partial order $\mathcal{P}$, and the Markov chain transitions between permutations σ and σ' by randomly selecting a pair $i, j \in [n]$, then swapping their orders if this is consistent with the partial order $\mathcal{P}$. Wilson [Wil04] showed that the mixing time of this Markov chain—in the uniform sense of Assumption B—is bounded by

$$\tau_{\text{mix}}(P, \epsilon) \leq \frac{4}{\pi^2} n^3 \log \frac{n}{\epsilon}. \tag{13}$$

Similar results hold for sampling from other combinatorial spaces [JS96].

As a consequence of the bound (13) on the mixing time of the Karzanov-Khachiyan Markov chain to a uniform sample from the set of permutations consistent with the partial order $\mathcal{P}$, Theorem 2 gives the following result. We denote the set of permutations σ consistent with the partial order $\mathcal{P}$ by $\sigma \in \mathcal{P}$, and the objective (1) thus has the form

$$f(x) := \frac{1}{\mathrm{card}(\sigma \in \mathcal{P})} \sum_{\sigma \in \mathcal{P}} F(x; \sigma).$$

We have

Corollary 4. *Let $x(t)$ evolve according to the EMD update (5), where the sample space is the set of permutations $\{\sigma\}$ consistent with the partial order $\mathcal{P}$ over $[n]$. Define $\widehat{x}(T) = \frac{1}{T}\sum_{t=1}^{T} x(t)$. Under Assumption A with $\alpha(t) = \alpha/\sqrt{t}$ and the appropriate choice of step size multiplier α,*

$$f(\widehat{x}(T)) - f(x^*) \leq 4\frac{GR}{\sqrt{T}} \cdot n^{3/2} \log(Tn) + \frac{GR}{T} \cdot n^3 \log(Tn) + 6GR\sqrt{\frac{\log\frac{1}{\delta}}{T}} \cdot n^{3/2}\sqrt{3\log(Tn)(\log[n\log(Tn)])}$$

with probability at least $1 - \delta$.

4.3 Markov chains on general state spaces

In general, the statistical sample space Ξ may be uncountable or continuous, in which case standard (finite-dimensional) Markov chain theory does not apply. Such situations commonly arise, for example, in physical simulations of natural phenomena or autoregressive processes [MT09], as well as in many statistical learning applications, such as Monte Carlo-sampling based variants of the expectation maximization (EM) algorithm [WT90]. We assume now that we have a Markov chain over the sample space Ξ that is *uniformly ergodic*, which is defined as follows [MT09, Chapter 16]. Let $P^t(\xi, \cdot)$ denote the distribution of the tth sample ξ_t from the Markov chain given that the initial sample $\xi_0 = \xi$. Then P is uniformly (or geometrically) ergodic if there exist $\rho \in [0, 1)$ and $M < \infty$ such that $d_{\mathrm{TV}}(P^t(\xi, \cdot), \Pi) \leq M\rho^t$ for all $\xi \in \Xi$. Sufficient conditions for a Markov chain on a general state space to be uniformly ergodic include Doeblin's condition [MT09, Theorem 16.2.3], which is often satisfied if the space Ξ is compact.

Given the definition of uniform ergodicity for general state space chains, it is clear that if

$$t \geq \frac{\log\frac{2M}{\epsilon}}{\log\rho^{-1}} \quad \text{then} \quad \sup_{\xi \in \Xi}\left[d_{\mathrm{TV}}(P^t(\xi, \cdot), \Pi)\right] \leq \frac{\epsilon}{2}.$$

In particular, we have $\tau_{\mathrm{mix}}(P, \epsilon) \leq \frac{\log(2M/\epsilon)}{-\log\rho} \leq \frac{\log(2M/\epsilon)}{1-\rho}$, and $\tau_{\mathrm{mix}}(P, T^{-1}) \leq \frac{\log(2MT)}{1-\rho}$ in the notation of Assumption B. Consequently, we have the following corollary.

Corollary 5. *Let $x(t)$ evolve according to the EMD update (5), where the sampling distribution P is a uniformly ergodic Markov chain with parameters $M < \infty$ and $\rho \in [0, 1)$. Define $\widehat{x}(T) = \frac{1}{T}\sum_{t=1}^{T} x(t)$. Under Assumption A and with $\alpha(t) = R\sqrt{1-\rho}/(G\sqrt{2\log(2MT)}\sqrt{t})$,*

$$f(\widehat{x}(T)) - f(x^*) \leq 5\frac{GR}{\sqrt{T}} \cdot \frac{\sqrt{\log(2MT)}}{\sqrt{1-\rho}} + 3\frac{GR}{T} \cdot \frac{\log(2MT)}{1-\rho} + 6GR\sqrt{\frac{\log\frac{1}{\delta}}{T}} \cdot \sqrt{\frac{\log(2MT)\log\frac{\log(2MT)}{1-\rho}}{1-\rho}}$$

with probability at least $1 - \delta$.

4.4 Slowly mixing processes

As mentioned in our earlier discussion of our main results, 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 will culminate in a result (Corollary 7) establishing asymptotic convergence of EMD for any ergodic process satisfying Assumption B.

Our starting point is an example of a continuous state space Markov chain that exhibits a mixing rate of the form (w.l.o.g. let $M \geq 1$ and $\beta \geq 0$)

$$\tau_{\text{mix}}(P, \epsilon) \leq M\epsilon^{-\beta}. \tag{14}$$

As an example, we consider a Metropolis-Hastings sampler [RC04] with the stationary distribution Π, assumed (for simplicity) to have a density π. The Metropolis-Hastings sampler uses a Markov chain Q as a "proposal" distribution, where $Q(\xi_t, \cdot)$ denotes the distribution of ξ_{t+1} conditioned on ξ_t, and $Q(\xi_t, \cdot)$ is assumed to have density $q(\xi_t, \cdot)$. The Markov chain constructed from Q and Π transitions from a point ξ_1 to ξ_2 as follows: first, the procedure samples ξ according to $Q(\xi_1, \cdot)$; second, the sample is accepted and ξ_2 is set to ξ with probability $\min\{\frac{\pi(\xi_2)q(\xi_2,\xi_1)}{\pi(\xi_1)q(\xi_1,\xi_2)}, 1\}$, otherwise $\xi_2 = \xi_1$. Metropolis-Hastings algorithms are the backbone for a large family of MCMC sampling procedures [RC04]. In the case that Q generates independent samples—that is, $q(\xi, \cdot) \equiv q(\cdot)$ for all ξ—then the associated Markov chain is uniformly ergodic (as in the previous section) only when the ratio $q(\xi)/\pi(\xi)$ is bounded away from zero over the sample space Ξ [MT09, Chapter 20].

When such a lower bound fails to exist, the proposal and stationary distributions are ill-matched and the mixing time can be sub-geometric, taking the form (14). Jarner and Roberts [JR02] give an example where Π is uniform on $[0, 1]$ and the density $q(x) = (r + 1)x^r$ for some $r > 0$. For this case, they show a polynomial mixing rate (14) with $\beta = 1/r$; other examples of similar rates include particular random walks on $[0, \infty)$ or queuing processes in continuous time.

We now state a corollary of our main results when the mixing time takes the form (14).

Corollary 6 (Sub-geometric mixing)**.** *Let $x(t)$ evolve according to the EMD update* (5)*, where the sampling distribution P is a polynomially mixing Markov chain with $\tau_{\text{mix}}(P, \epsilon) \leq M\epsilon^{-\beta}$. Assume that $T \geq (R/G)^2$. Under Assumption A and with $\alpha(t) \equiv \frac{R}{G}T^{-(\beta+1)/(\beta+2)}$,*

$$\mathbb{E}[f(\widehat{x}(T)]) - f(x^*) \leq \frac{3eGRM^{\frac{1}{\beta+1}}}{T^{\frac{1}{2+\beta}}}.$$

The stepsize choice $\alpha(t) = R/(G\sqrt{t})$ gives that

$$\mathbb{E}[f(\widehat{x}(T))] - f(x^*) \leq \frac{3GR}{2\sqrt{T}} + \frac{eGR(9M)^{\frac{1}{\beta+1}}}{T^{\frac{1}{2\beta+2}}}.$$

Proof By applying the bound in Corollary 1, we see that the expected convergence rate for the fixed setting of $\alpha(t) \equiv \alpha$ in the statement of the corollary is

$$\frac{R^2}{2T\alpha} + \frac{G^2}{2}\alpha + \epsilon GR + 2M\epsilon^{-\beta}G^2\alpha + \frac{M\epsilon^{-\beta}RG}{T} \leq \frac{R^2}{2T\alpha} + \frac{G^2}{2}\alpha + \epsilon GR + 3M\epsilon^{-\beta}G^2\alpha,$$

using the assumption that $T \geq (R/G)^2$. We can choose ϵ arbitrarily, so set $\epsilon = (3\beta GM\alpha/R)^{1/(1+\beta)}$. Using the proposed stepsize $\alpha(t) = (R/G)T^{-(\beta+1)/(\beta+2)}$, we find that the above is equal to

$$\frac{R^2}{2T\alpha} + \frac{G^2}{2}\alpha + \left(1 + \beta^{-\frac{\beta}{1+\beta}}\right)\alpha^{\frac{1}{1+\beta}}(3M)^{\frac{1}{1+\beta}}G^{\frac{2+\beta}{1+\beta}}R^{\frac{\beta}{1+\beta}} \leq \frac{GR}{2T^{\frac{1}{2+\beta}}} + \frac{GR}{2T^{\frac{\beta+1}{\beta+2}}} + \frac{GRe(3M)^{\frac{1}{1+\beta}}}{T^{\frac{1}{2+\beta}}},$$

where we use $1 + \beta^{-\beta/(1+\beta)} \le e$. Noting that $\beta \ge 0$ yields the first statement of the corollary.

With the step size choice $\alpha(t) = \alpha/\sqrt{t}$ with multiplier $\alpha = R/G$, we can apply Theorem 1, along with the bound (7) in the proof of Corollary 2, to see that with probability at least $1 - \delta$,

$$f(\widehat{x}(T)) - f(x^*) \le \frac{3RG}{2\sqrt{T}} + \epsilon GR + \frac{8\tau_{\mathrm{mix}}(P,\epsilon)GR}{\sqrt{T}} + \frac{2\tau_{\mathrm{mix}}(P,\epsilon)GR}{T} + 6GR\sqrt{\frac{\log\frac{1}{\delta}}{T}}. \tag{15}$$

Noting that $2/T + 8/\sqrt{T} \le 9/\sqrt{T}$, we turn to bounding

$$\epsilon GR + \frac{9\tau_{\mathrm{mix}}(P,\epsilon)GR}{\sqrt{T}} \le \epsilon GR + \epsilon^{-\beta}\frac{9MGR}{\sqrt{T}}. \tag{16}$$

Since ϵ does not enter into the algorithm at all, we are free to minimize over ϵ, and taking derivatives we see that we must solve

$$GR - \beta\epsilon^{-\beta-1}\frac{9MGR}{\sqrt{T}} = 0 \quad \text{or} \quad \epsilon = \left(\frac{9M\beta}{\sqrt{T}}\right)^{\frac{1}{\beta+1}}.$$

Since $\beta^{1/(\beta+1)} \le e/2$ and $\beta^{-\beta/(\beta+1)} \le e/2$, the above choice of ϵ in the bound (16) yields

$$\inf_{\epsilon}\left\{\epsilon GR + \frac{9\tau_{\mathrm{mix}}(P,\epsilon)GR}{\sqrt{T}}\right\} \le eGR(9M)^{\frac{1}{\beta+1}} \cdot T^{\frac{-1}{2\beta+2}}.$$

By inspection, this inequality and the convergence guarantee (15) give the second statement of the corollary. □

A weakness of the above bound is that the sharper rate of convergence requires knowledge of the mixing rate of P, and choosing the polynomial incorrectly can lead to significantly slower convergence. In contrast, as noted in Section 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 $\alpha(t) \propto t^{-1/2}$. Nonetheless, Corollary 6 gives a finite sample convergence rate whose dependence on the slower mixing of the ergodic process is clear. In addition, the proof of Corollary 6 exhibits 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. To be specific, note that the bound in Corollary 1 guarantees that for $\widehat{x}(T) = \frac{1}{T}\sum_{t=1}^{T} x(t)$, if we choose $\alpha(t) = \alpha/\sqrt{t}$ then

$$\mathbb{E}[f(\widehat{x}(T))] - f(x^*) \le \frac{R^2}{2\alpha\sqrt{T}} + \frac{G^2\alpha}{\sqrt{T}} + \epsilon GR + \frac{4\tau_{\mathrm{mix}}(P,\epsilon)G^2\alpha}{\sqrt{T}} + \frac{\tau_{\mathrm{mix}}(P,\epsilon)RG}{T}. \tag{17}$$

The convergence guarantee (17) holds regardless of our choice of ϵ, so we can choose ϵ minimizing the above. That is (setting $\alpha = R/G$ for notational convenience),

$$\mathbb{E}[f(\widehat{x}(T))] - f(x^*) \le \frac{3GR}{2\sqrt{T}} + \inf_{\epsilon \ge 0}\left\{\epsilon GR + \frac{4\tau_{\mathrm{mix}}(P,\epsilon)GR}{\sqrt{T}} + \frac{\tau_{\mathrm{mix}}(P,\epsilon)GR}{T}\right\}.$$

For any fixed $\epsilon > 0$, the term inside the infimum decreases to ϵGR as $T \uparrow \infty$, so the infimal term decreases to zero as $T \uparrow \infty$. High probability convergence follows similarly by using Theorem 2,

since for any $\delta_T > 0$ we have

$$f(\widehat{x}(T)) - f(x^*) \le \frac{3GR}{2\sqrt{T}} + \inf_{\epsilon \ge 0} \left\{ \epsilon GR + \frac{4\tau_{\text{mix}}(P,\epsilon)GR}{\sqrt{T}} + \frac{\tau_{\text{mix}}(P,\epsilon)GR}{T} \right.$$
$$\left. + 6GR\sqrt{\frac{\log \frac{1}{\delta_T}}{T}} \cdot \sqrt{\tau_{\text{mix}}(P,\epsilon) \log \tau_{\text{mix}}(P,\epsilon)} \right\} \tag{18}$$

with probability at least $1 - \delta_T$. Fix an arbitrary $\gamma > 0$ and let E_T denote the event that $f(\widehat{x}(T)) - f(x^*) > \gamma$. We will use the Borel Cantelli lemma [Bil86] 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 (18) is less than γ. Then we have

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

As a consequence, we see that for any $\gamma > 0$, $\mathbb{P}(f(\widehat{x}(T)) - f(x^*) > \gamma \text{ i.o.}) = 0$, and thus

Corollary 7. *Define $\widehat{x}(T) = \frac{1}{T}\sum_{t=1}^{T} x(t)$. Under the conditions of Theorem 1, the stepsize sequence $\alpha(t) = \alpha/\sqrt{t}$ for any $\alpha > 0$ yields $f(\widehat{x}(T)) \to f(x^*)$ as $T \to \infty$ both in expectation and with probability 1.*

5 Analysis

In this section, we analyze the convergence of the EMD algorithm from Section 2. Our first subsection lays the groundwork, gives some necessary notation, and provides a few optimization-based results that we build on. 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.

5.1 Definitions and optimization-based results

To state our results formally, we give a few standard definitions. Let $\mathsf{G}(x;\xi) \in \partial F(x;\xi)$ represent a fixed and measurable element of the subgradient of $F(\cdot;\xi)$ evaluated at x, where we make the assumption (without loss) that in the EMD algorithm (5), $g(t) = \mathsf{G}(x(t);\xi_t)$. By our assumptions on F, for any distribution Q for which the below expectations are defined, expectation and subdifferentiation commute [RW82]; that is, if we define

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

In particular, $\mathbb{E}_\Pi[\partial F(x;\xi)] = \partial f(x)$ and $\mathbb{E}_\Pi[\mathsf{G}(x;\xi)] \in \partial f(x)$. We thus use the following notation for the errors in stochastic gradient evaluation for the method (5):

$$f'(x(t)) := \mathbb{E}_\Pi[\mathsf{G}(x(t);\xi)] \in \partial f(x(t)) \quad \text{and} \quad e(t) := f'(x(t)) - g(t). \tag{19}$$

To make the presentation self-contained, we provide proofs of standard results from optimization theory as lemmas in Appendix A, but we state the results here. We begin with a standard single-step lemma, which is essentially present in earlier work [NY83, BT03], but we state and prove it for completeness and because we must allow errors in subgradient computations.

Lemma 8 ([BT03])**.** *Let $x(t)$ evolve according to the EMD algorithm* (5)*. For any $x^* \in \mathcal{X}$,*

$$f(x(t)) - f(x^*) \le \frac{1}{\alpha(t)} D_\psi(x^*, x(t)) - \frac{1}{\alpha(t)} D_\psi(x^*, x(t+1)) + \frac{\alpha(t)}{2} \|g(t)\|_*^2 + \langle e(t), x(t) - x^* \rangle$$

The next proposition gives a non-probabilistic convergence guarantee for the EMD algorithm. The proposition is similar to arguments in the analysis of stochastic mirror descent [NY83, BT03], and we provide its proof in Appendix A.

Proposition 1. *Let $x(t)$ evolve according to the EMD algorithm* (5)*. Under the assumptions of Theorem 1, for any $x^* \in \mathcal{X}$,*

$$\sum_{t=1}^{T} f(x(t)) - f(x^*) \le \frac{R^2}{2\alpha(T)} + \frac{G^2}{2} \sum_{t=1}^{T} \alpha(t) + \sum_{t=1}^{T} \langle e(t), x(t) - x^* \rangle .$$

Finally, we give two useful non-probabilistic results. First, the compactness assumption that $D_\psi(x^*, x(t)) \le \frac{1}{2} R^2$ for all t coupled with the strong convexity of ψ implies

$$\|x(t) - x^*\|^2 \le 2 D_\psi(x^*, x(t)) \le R^2 \quad \text{so} \quad \|x(t) - x^*\| \le R. \tag{20}$$

The following simple lemma controls the differences between $x(t)$ and $x(t+1)$.

Lemma 9. *Let $x(t)$ be generated according to the EMD algorithm* (5)*. Then*

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

5.2 Expected convergence rates

In this subsection, we analyze the convergence of the algorithm (5), which culminates in the proof of Theorem 1. The main result that allows us to prove the theorem is contained in Proposition 2, which controls the expectations of error (19) in the subgradient estimators $\partial F(x(t); \xi_t)$ versus $\partial f(x(t))$. Having established the relevant optimization-based results in Section 5.1, we can now turn to understanding the impact of the ergodic sequence $\xi_1, \xi_2, \ldots$ on the EMD procedure. In the i.i.d. noise setting—standard stochastic mirror descent—the error terms $\langle e(t), x(t) - x^* \rangle$ have zero expectation, so Proposition 1 immediately yields an expected convergence rate. To use Proposition 1 to prove Theorem 1, what remains is to control the expectation $\mathbb{E}[\langle e(t), x(t) - x^* \rangle]$. For a general ergodic process, the preceding expectation is rarely zero, so we use the mixing properties of P in the remaining analysis to control the error.

The essential idea in the next proposition is to note that the errors $e(t)$ are nearly "independent" of parameters $x(s)$ for s far enough in the past. Specifically, the decorrelation occurs when $t - s \ge \tau_{\text{mix}}(P, \epsilon)$, since conditioned on $\xi_1, \ldots, \xi_s$, the distribution of ξ_t is close to the stationary distribution Π for large t. We then know that $\mathbb{E}[e(t) \mid \xi_1, \ldots, \xi_s] \approx 0$, and we can easily control the distance between $x(t)$ and $x(t-s)$ so that $\langle e(t), x(t) - x^* \rangle$ is small. We now formalize this intuition.

Proposition 2. *Let $x(t)$ be defined by the update* (5) *with $\alpha(t)$ and let R be the radius of the set $\mathcal{X}$ with respect to D_ψ, that is, $D_\psi(x^*, x(t)) \le \frac{1}{2} R^2$ for all t. Then*

$$\mathbb{E}\left[\sum_{t=1}^{T} \langle e(t), x(t) - x^* \rangle \right] \le \epsilon G R T + 2\tau_{\text{mix}}(P, \epsilon) G^2 \sum_{t=\tau_{\text{mix}}(P,\epsilon)+1}^{T} \alpha(t - \tau_{\text{mix}}(P, \epsilon)) + 2\tau_{\text{mix}}(P, \epsilon) R G.$$

Proof Let τ_{mix} be a shorthand for $\tau_{\mathrm{mix}}(P, \epsilon)$. We need to control the expectation of the error terms $\langle e(t), x(t) - x^* \rangle$ in Proposition 1. We begin by expanding the error terms to take advantage of the weakened dependence for $e(t)$ and $x(s)$ for s far enough from t. We have

$$\langle e(t), x(t) - x^* \rangle = \langle e(t), x(t - \tau_{\mathrm{mix}}) - x^* \rangle + \langle e(t), x(t) - x(t - \tau_{\mathrm{mix}}) \rangle \,. \tag{21}$$

We define the sigma field $\mathcal{F}_t = \sigma(\xi_1, \ldots, \xi_t)$. Taking expectations, we see that

$$\begin{aligned} \mathbb{E}\left[\langle e(t), x(t - \tau_{\mathrm{mix}}) - x^* \rangle\right] &= \mathbb{E}\mathbb{E}\left[\langle e(t), x(t - \tau_{\mathrm{mix}}) - x^* \rangle \mid \mathcal{F}_{t-\tau_{\mathrm{mix}}}\right] \\ &= \mathbb{E}\left(\langle \mathbb{E}[e(t) \mid \mathcal{F}_{t-\tau_{\mathrm{mix}}}], x(t - \tau_{\mathrm{mix}}) - x^* \rangle\right), \end{aligned} \tag{22}$$

since $x(t - \tau_{\mathrm{mix}})$ is measurable with respect to $\mathcal{F}_{t-\tau_{\mathrm{mix}}}$. We focus on showing that $\mathbb{E}[e(t) \mid \mathcal{F}_{t-\tau_{\mathrm{mix}}}]$ is small. Indeed, recall our definition of the subgradient selection $\mathsf{G}(x; \xi) \in F(x; \xi)$, which gave $g(t) = \mathsf{G}(x(t); \xi_t)$ and $f'(x(t)) = \mathbb{E}_\Pi[\mathsf{G}(x(t); \xi)] \in \partial f(x(t))$. Also recall that Π and $P^t_{[s]}$ are without loss assumed to have densities π and $p^t_{[s]}$ with respect to an underlying measure μ, so that

$$\begin{aligned} \mathbb{E}[e(t) \mid \mathcal{F}_{t-\tau_{\mathrm{mix}}}] &= f'(x(t)) - \int_\Xi \mathsf{G}(x(t); \xi) dP^t_{[t-\tau_{\mathrm{mix}}]}(\xi) \\ &= \int_\Xi \mathsf{G}(x(t); \xi) d\Pi(\xi) - \int_\Xi \mathsf{G}(x(t); \xi) dP^t_{[t-\tau_{\mathrm{mix}}]}(\xi) \\ &= \int_\Xi \mathsf{G}(x(t); \xi) \left[\pi(\xi) - p^t_{[t-\tau_{\mathrm{mix}}]}(\xi)\right] d\mu(\xi). \end{aligned} \tag{23}$$

In order to control the size of the expectation (22), we apply Hölder's inequality to see

$$\mathbb{E}\left[\langle e(t), x(t - \tau_{\mathrm{mix}}) - x^* \rangle\right] \le \mathbb{E}\left\|\mathbb{E}[e(t) \mid \mathcal{F}_{t-\tau_{\mathrm{mix}}}]\right\|_* \left\|x(t - \tau_{\mathrm{mix}}) - x(t)\right\|,$$

so that it suffices to control the norm of the conditional expectation of $e(t)$. To this end, we apply the triangle inequality, the Lipschitz assumption that $\|\mathsf{G}(x; \xi)\|_* \le G$, and the uniform mixing assumption to the equality (23) to obtain

$$\begin{aligned} \left\|\mathbb{E}[e(t) \mid \mathcal{F}_{t-\tau_{\mathrm{mix}}}]\right\|_* &\le \int_\Xi \left\|\mathsf{G}(x; \xi)\right\|_* \left|\pi(\xi) - p^t_{[t-\tau_{\mathrm{mix}}]}(\xi)\right| d\mu(\xi) \\ &\le G \int_\Xi \left|\pi(\xi) - p^t_{[t-\tau_{\mathrm{mix}}]}(\xi)\right| d\mu(\xi) = 2G \cdot d_{\mathrm{TV}}(\Pi, P^t_{[t-\tau_{\mathrm{mix}}]}) \le G\epsilon. \end{aligned} \tag{24}$$

What remains is to control the $\langle e(t), x(t) - x(t - \tau_{\mathrm{mix}}) \rangle$ terms from the expansion (21), which are bounded by $\|e(t)\|_* \|x(t) - x(t - \tau_{\mathrm{mix}})\|$. Applying Lemma 9 to $\|x(t) - x(t - \tau_{\mathrm{mix}})\|$, we have

$$\left\|x(t) - x(t - \tau_{\mathrm{mix}})\right\| \le \sum_{s=t-\tau_{\mathrm{mix}}}^{t-1} \left\|x(s) - x(s+1)\right\| \le \sum_{s=t-\tau_{\mathrm{mix}}}^{t-1} \alpha(s) \left\|g(s)\right\|_* \le G\tau_{\mathrm{mix}} \cdot \alpha(t - \tau_{\mathrm{mix}}),$$

where for the last inequality we use $\|g(s)\|_* \le G$ (the Lipschitz assumption on the functions F) and the fact that $\alpha(t)$ is decreasing in t. We combine this result with (21) and (24) to see that for $t > \tau_{\mathrm{mix}}$

$$\mathbb{E}\left[\langle e(t), x(t) - x^* \rangle\right] \le \epsilon G \mathbb{E}\left\|x(t - \tau_{\mathrm{mix}}) - x^*\right\| + \alpha(t - \tau_{\mathrm{mix}}) \mathbb{E}[\|e(t)\|_*] \tau_{\mathrm{mix}} G.$$

Using the compactness assumption on $\mathcal{X}$, we have $\|x(t-\tau_{\text{mix}}) - x^*\| \leq R$ (see the bound (20)), and consequently for $t > \tau_{\text{mix}}$

$$\mathbb{E}[\langle e(t), x(t) - x^* \rangle] \leq \epsilon GR + \alpha(t - \tau_{\text{mix}})\mathbb{E}[\|e(t)\|_*]\tau_{\text{mix}}G.$$

For $t \leq \tau_{\text{mix}}$, we have $\langle e(t), x(t) - x^* \rangle \leq 2GR$, since $\|e(t)\|_* \leq 2G$ by assumption. Adding the bounds on the error terms $\langle e(t), x(t) - x^* \rangle$ for $t \leq \tau_{\text{mix}}$ and $t > \tau_{\text{mix}}$ proves the proposition. □

Proof of Theorem 1 The bound in Proposition 1 is non-probabilistic, so all we need to complete the proof is to take expectations, applying Proposition 2. To that end, we note that $\|e(t)\|_* \leq 2G$ by the Lipschitz assumption A. Combining the bounds from the two propositions yields

$$\mathbb{E}\sum_{t=1}^{T} f(x(t)) - f(x^*) \leq \frac{R^2}{2\alpha(T)} + \frac{G^2}{2}\sum_{t=1}^{T}\alpha(t) + T\epsilon GR + 2\tau_{\text{mix}}(P,\epsilon)G^2\sum_{t=1}^{T}\alpha(t) + 2\tau_{\text{mix}}(P,\epsilon)RG,$$

where we use the fact that the sum of $\alpha(t)$ from $t = \tau_{\text{mix}}(P,\epsilon)+1$ to T is bounded by the sum over all T. □

5.3 High-probability convergence

In this section, we complement the convergence bounds in Section 5.2 with high-probability statements. We use martingale theory to show that the bound of Theorem 1 holds with high probability. Specifically, we use similar techniques to those we employed in the proof of Proposition 2 to argue that $\langle e(t), x(t-\tau_{\text{mix}}) - x^* \rangle$ is small. Intuitively, this follows because given the initial $t-\tau_{\text{mix}}$ samples $\xi_1, \ldots, \xi_{t-\tau_{\text{mix}}}$, 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 sequence $\langle e(t), x(t-\tau_{\text{mix}}) - x^* \rangle$ behaves approximately as a martingale, and we can then apply Azuma's inequality [Azu67] to derive high probability guarantees on $\sum_{t=1}^{T}\langle e(t), x(t) - x^* \rangle$.

Proposition 3. *Use the same conditions as those in Proposition 2. With probability at least* $1-\delta$,

$$\sum_{t=1}^{T}\langle e(t), x(t) - x^* \rangle$$
$$\leq \epsilon GRT + 2\tau_{\text{mix}}(P,\epsilon)G^2\sum_{t=1}^{T}\alpha(t) + 2\tau_{\text{mix}}(P,\epsilon)GR + 6GR\sqrt{T\tau_{\text{mix}}(P,\epsilon)\log\tau_{\text{mix}}(P,\epsilon)\log\frac{1}{\delta}}.$$

Proof The proof is somewhat similar to that of Proposition 2. We use τ as a shorthand for $\tau_{\text{mix}}(P,\epsilon)$. We begin by rewriting the random error sequence as

$$\sum_{t=1}^{T}\langle e(t), x(t) - x^* \rangle = \sum_{t=\tau+1}^{T}\langle e(t), x(t) - x^* \rangle + \sum_{t=1}^{\tau}\langle e(t), x(t) - x^* \rangle$$
$$= \sum_{t=\tau+1}^{T}\langle e(t), x(t-\tau) - x^* \rangle + \sum_{t=\tau+1}^{T}\langle e(t), x(t) - x(t-\tau) \rangle + \sum_{t=1}^{\tau}\langle e(t), x(t) - x^* \rangle. \quad (25)$$

The second-to-last term in the sum (25) was bounded in the proofs of Proposition 2 and Theorem 1, where we applied Lemma 9 to see

$$\sum_{t=\tau+1}^{T} \langle e(t), x(t) - x(t-\tau)\rangle \le \sum_{t=\tau+1}^{T} 2\tau\alpha(t-\tau)G^2 \le 2\tau G^2 \sum_{t=1}^{T} \alpha(t).$$

The last term in the sum (25) has each of its terms bounded by $\langle e(t), x(t) - x^*\rangle \le 2GR$.

What remains is to control the first term in the summation (25). We construct a family of τ different martingales from the error sequence above, each of which we control with high probability, and we apply a union bound to get deviation bounds on the entire series. We begin by defining

$$Z_t := \langle e(t), x(t-\tau) - x^*\rangle$$

and the filtration of σ-fields $\mathcal{A}_i^j = \mathcal{F}_{\tau i+j}$ for $j = 1, \dots, \tau$. We can now construct a set of τ Doob martingales by defining

$$\begin{aligned} X_i^j &:= Z_{\tau i+j} - \mathbb{E}[Z_{\tau i+j} \mid \mathcal{A}_{i-1}^j] = Z_{\tau i+j} - \mathbb{E}[Z_{\tau i+j} \mid \mathcal{F}_{\tau(i-1)+j}] \\ &= \langle e(\tau i + j), x(\tau(i-1)+j) - x^*\rangle - \mathbb{E}[\langle e(\tau i + j), x(\tau(i-1)+j) - x^*\rangle \mid \mathcal{F}_{\tau(i-1)+j}]. \end{aligned}$$

By inspection, X_i^j is measurable with respect to the σ-field $\mathcal{A}_i^j$, and $\mathbb{E}[X_i^j \mid \mathcal{A}_{i-1}^j] = 0$. Thus for each j, the sequence X_i^j is a martingale difference sequence adapted to the filtration $\mathcal{A}_i^j$, $i \in \{2, 3, \dots\}$. Define the index set $\mathcal{I}(j)$ to be the indices $\{2, \dots, \lfloor T/\tau \rfloor + 1\}$ for $j \le T - \tau \lfloor T/\tau \rfloor$ and $\{2, \dots, \lfloor T/\tau \rfloor\}$ otherwise. With the definition of X_i^j and the indices $\mathcal{I}(j)$, we see that

$$\sum_{t=\tau+1}^{T} Z_t = \sum_{j=1}^{\tau} \sum_{i \in \mathcal{I}(j)} X_i^j + \sum_{t=\tau+1}^{T} \mathbb{E}[Z_t \mid \mathcal{F}_{t-\tau}] = \sum_{j=1}^{\tau} \sum_{i=2}^{|\mathcal{I}(j)|} X_i^j + \sum_{t=\tau+1}^{T} \mathbb{E}[Z_t \mid \mathcal{F}_{t-\tau}]. \tag{26}$$

Now we note the following important fact: by the compactness assumption (20) and the Lipschitz assumption A,

$$|X_i^j| \le |Z_{\tau i+j}| + \left|\mathbb{E}[Z_{\tau i+j} \mid \mathcal{F}_{\tau(i-1)+j}]\right| \le 4GR.$$

This bound, coupled with the representation (26), shows that $\sum_{t=\tau+1}^{T} Z_t$ can be written as a sum of τ different bounded-difference martingales plus a sum of conditional expectations we will bound later. To control the martingale portion of the sum (26), we apply the triangle inequality, a union bound, and Azuma's inequality to find

$$\mathbb{P}\Big(\sum_{j=1}^{\tau} \sum_{i \in \mathcal{I}(j)} X_i^j > \gamma\Big) \le \sum_{j=1}^{\tau} \mathbb{P}\Big(\sum_{i \in \mathcal{I}(j)} X_i^j > \frac{\gamma}{\tau}\Big) \le \sum_{j=1}^{\tau} \exp\left(-\frac{\gamma^2}{32G^2R^2\tau T}\right),$$

since there are fewer than T/τ terms in each of the sums X_i^j. Noting that $6 > \sqrt{32}$ and substituting $\gamma = 6GR\sqrt{T\tau \log\tau \log(1/\delta)}$, we find

$$\mathbb{P}\Big(\sum_{j=1}^{\tau} \sum_{i \in \mathcal{I}(j)} X_i^j > 6GR\sqrt{T\tau \log\tau \log\frac{1}{\delta}}\Big) \le \delta.$$

To bound the final term $\mathbb{E}[Z_t \mid \mathcal{F}_{t-\tau}]$ in the sum (26) we recall from the proof of Proposition 2—specifically the inequality (24)—that

$$|\mathbb{E}[\langle e(t), x(t - \tau_{\text{mix}}) - x^* \rangle \mid \mathcal{F}_{t-\tau_{\text{mix}}}]| \leq \|\mathbb{E}[e(t) \mid \mathcal{F}_{t-\tau_{\text{mix}}}]\|_* R \leq \epsilon G R.$$

Thus $\sum_{t=\tau+1}^{T} \mathbb{E}[Z_t \mid \mathcal{F}_{t-\tau}] \leq T\epsilon RG$, and substituting $\tau_{\text{mix}}(P, \epsilon)$ for τ completes the proof. □

Proof of Theorem 2 The proof is a straightforward combination of the proofs of previous results. We simply replace the sum of errors $\sum_{t=1}^{T} \langle e(t), x(t) - x^* \rangle$ in the non-probabilistic Proposition 1 with the upper bound in Proposition 3, which holds with probability at least $1 - \delta$. □

6 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 $\xi_1, \xi_2, \ldots$ 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 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. [Yu94]). A natural extension of this work, which we hope to be able to accomplish, is to relax the assumptions on the uniformity of the mixing times in Assumption B, which would allow a wider range of applications of our results.

Acknowledgments

We thank Lester Mackey for several interesting questions he posed that helped lead to this work.

A Proofs of Optimization Results

Proof of Lemma 8 By the first-order convexity inequality and definition of the subgradient $f'(x(t))$, we have

$$f(x(t)) - f(x^*) \leq \left\langle f'(x(t)), x(t) - x^* \right\rangle = \langle g(t), x(t) - x^* \rangle + \langle e(t), x(t) - x^* \rangle .$$

We expand the stochastic subgradient term to see that

$$\langle g(t), x(t) - x^* \rangle = \langle g(t), x(t+1) - x^* \rangle + \langle g(t), x(t+1) - x(t) \rangle . \tag{27}$$

For any $y \in \mathcal{X}$, the first-order optimality conditions for $x(t+1)$ in the update (5) imply

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

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

$$\alpha(t) \langle g(t), x(t+1) - x^* \rangle \leq \langle \nabla\psi(x(t+1)) - \nabla\psi(x(t)), x^* - x(t+1) \rangle . \tag{28}$$

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

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

Combining this result with the expanded gradient term (27) and the consequence (28) of the first-order convexity inequality, we get

$$\begin{aligned}
f(x(t)) - f(x^*) &\leq \frac{1}{\alpha(t)} D_\psi(x^*, x(t)) - \frac{1}{\alpha(t)} D_\psi(x^*, x(t+1)) - \frac{1}{\alpha(t)} D_\psi(x(t+1), x(t)) \\
&\quad + \langle g(t), x(t+1) - x(t) \rangle + \langle e(t), x(t) - x^* \rangle \\
&\overset{(i)}{\leq} \frac{1}{\alpha(t)} D_\psi(x^*, x(t)) - \frac{1}{\alpha(t)} D_\psi(x^*, x(t+1)) - \frac{1}{\alpha(t)} D_\psi(x(t+1), x(t)) \\
&\quad + \frac{\alpha(t)}{2} \|g(t)\|_*^2 + \frac{1}{2\alpha(t)} \|x(t+1) - x(t)\|^2 + \langle e(t), x(t) - x^* \rangle \\
&\overset{(ii)}{\leq} \frac{1}{\alpha(t)} D_\psi(x^*, x(t)) - \frac{1}{\alpha(t)} D_\psi(x^*, x(t+1)) + \frac{\alpha(t)}{2} \|g(t)\|_*^2 + \langle e(t), x(t) - x^* \rangle .
\end{aligned}$$

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., [BV04, Example 3.27]), while the last inequality (ii) follows by the strong convexity of ψ as in (4), which shows that $D_\psi(x(t+1), x(t)) \geq \frac{1}{2}\|x(t+1) - x(t)\|^2$. □

Proof of Proposition 1 By summing the inequality in Lemma 8 from $t = 1$ to T, we have

$$\begin{aligned}
&\sum_{t=1}^{T} f(x(t)) - f(x^*) \\
&\leq \sum_{t=2}^{T} D_\psi(x^*, x(t)) \left[\frac{1}{\alpha(t)} - \frac{1}{\alpha(t-1)} \right] + \frac{1}{\alpha(1)} D_\psi(x^*, x(1)) - \frac{1}{\alpha(T)} D_\psi(x^*, x(T+1)) \\
&\quad + \sum_{t=1}^{T} \frac{\alpha(t)}{2} \|g(t)\|_*^2 + \sum_{t=1}^{T} \langle e(t), x(t) - x^* \rangle .
\end{aligned}$$

Recalling the compactness assumption that $D_\psi(x^*, x(t)) \leq \frac{1}{2}R^2$, the above bound implies

$$\begin{aligned}
\sum_{t=1}^{T} f(x(t)) - f(x^*) &\leq \frac{R^2}{2} \sum_{t=2}^{T} \left[\frac{1}{\alpha(t)} - \frac{1}{\alpha(t-1)} \right] + \frac{R^2}{2\alpha(1)} + \sum_{t=1}^{T} \frac{\alpha(t)}{2} \|g(t)\|_*^2 + \sum_{t=1}^{T} \langle e(t), x(t) - x^* \rangle \\
&= \frac{R^2}{2\alpha(T)} + \frac{1}{2} \sum_{t=1}^{T} \alpha(t) \|g(t)\|_*^2 + \sum_{t=1}^{T} \langle e(t), x(t) - x^* \rangle .
\end{aligned} \tag{29}$$

Lastly, we use the Lipschitz assumption to bound $\|g(t)\|_* \leq G$ which completes the proof. □

Proof of Lemma 9 By the first-order condition for the optimality of $x(t+1)$ for the update (5), 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

$$\begin{aligned}\langle \nabla\psi(x(t)) - \nabla\psi(x(t+1)), x(t) - x(t+1)\rangle &\leq \alpha(t)\langle g(t), x(t) - x(t+1)\rangle \\ &\leq \alpha(t)\left\|g(t)\right\|_* \left\|x(t) - x(t+1)\right\|\end{aligned}$$

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 thus see that

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

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

References

[Azu67] K. Azuma, *Weighted sums of certain dependent random variables*, Tohoku Mathematical Journal **68** (1967), 357–367.

[Bil86] Patrick Billingsley, *Probability and measure*, Second ed., Wiley, 1986.

[Bra05] R. C. Bradley, *Basic properties of strong mixing conditions. a survey and some open questions*, Probability Surveys **2** (2005), 107–144.

[BT03] A. Beck and M. Teboulle, *Mirror descent and nonlinear projected subgradient methods for convex optimization*, Operations Research Letters **31** (2003), 167–175.

[BV04] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge University Press, 2004.

[Chu98] F.R.K. Chung, *Spectral graph theory*, AMS, 1998.

[DAW11] J. C. Duchi, A. Agarwal, and M. Wainwright, *Dual averaging for distributed optimization: convergence analysis and network scaling*, URL `http://arxiv.org/abs/1005.2012`, 2011.

[HUL96] J. Hiriart-Urruty and C. Lemaréchal, *Convex analysis and minimization algorithms i*, Springer, 1996.

[IZ89] R. Impagliazzo and D. Zuckerman, *How to recycle random bits*, 30th Annual Symposium on Foundations of Computer Science, 1989, pp. 248–253.

[JR02] Sren F. Jarner and Gareth O. Roberts, *Polynomial convergence rates of Markov chains*, The Annals of Applied Probability **12** (2002), no. 1, pp. 224–247.

[JRJ09] B. Johansson, M. Rabi, and M. Johansson, *A randomized incremental subgradient method for distributed optimization in networked systems*, SIAM Journal on Optimization **20** (2009), no. 3, 1157–1170.

[JS96] M. Jerrum and A. Sinclair, *The Markov chain Monte Carlo method: an approach to approximate counting and integration*, Approximation Algorithms for NP-hard Problems (D. S. Hochbaum, ed.), PWS Publishing, 1996.

[KK91] A. Karzanov and L. Khachiyan, *On the conductance of order Markov chains*, Order **8** (1991), 7–15.

[KY03] H. J. Kushner and G. Yin, *Sotchastic approximation and recursive algorithms and applications*, Second ed., Springer, 2003.

[Mar98] K. Marton, *Measure concentration for a class of random processes*, Probability Theory and Related Fields **110** (1998), 427–439.

[MT09] S. Meyn and R. L. Tweedie, *Markov chains and stochastic stability*, Second ed., Cambridge University Press, 2009.

[NB01] A. Nedić and D. P. Bertsekas, *Incremental subgradient methods for nondifferentiable optimization*, SIAM Journal on Optimization **12** (2001), 109–138.

[NJLS09] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM Journal on Optimization **19** (2009), no. 4, 1574–1609.

[NY83] A. Nemirovski and D. Yudin, *Problem complexity and method efficiency in optimization*, Wiley, New York, 1983.

[PJ92] B. T. Polyak and A. B. Juditsky, *Acceleration of stochastic approximation by averaging*, SIAM Journal on Control and Optimization **30** (1992), no. 4, 838–855.

[RC04] C. Robert and G. Casella, *Monte carlo statistical methods*, Second ed., Springer, 2004.

[RM51] H. Robbins and S. Monro, *A stochastic approximation method*, Annals of Mathematical Statistics **22** (1951), 400–407.

[RW82] 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), 173–182.

[Sam00] Paul-Marie Samson, *Concentration of measure inequalities for Markov chains and ϕ-mixing processes*, Annals of Probability **28** (2000), no. 1, 416–461.

[Wil04] D. B. Wilson, *Mixing times of lozenge tiling and card shuffling Markov chains*, Annals of Applied Probability **14** (2004), no. 1, 274–325.

[WT90] 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), no. 411, 699–704.

[Yu94] Bin Yu, *Rates of convergence for empirical processes of stationary mixing sequences*, Annals of Probability **22** (1994), no. 1, 94–116.

[Zin03] M. Zinkevich, *Online convex programming and generalized infinitesimal gradient ascent*, Proceedings of the Twentieth International Conference on Machine Learning, 2003.