Stochastic optimization with non-i.i.d. noise

Alekh Agarwal **John C. Duchi**
Department of Electrical Engineering and Computer Science
University of California, Berkeley
{alekh,jduchi}@eecs.berkeley.edu

Abstract

We study the convergence of a class of stable online algorithms for stochastic convex optimization in settings where we do not receive independent samples from the distribution over which we optimize, but instead receive samples that are coupled over time. We show the optimization error of the averaged predictor output by any stable online learning algorithm is upper bounded—with high probability—by the average regret of the algorithm, so long as the underlying stochastic process is β- or ϕ-mixing. We additionally show sharper convergence rates when the expected loss is strongly convex, which includes as special cases linear prediction problems including linear and logistic regression, least-squares SVM, and boosting.

1 Introduction

In this paper, we study the performance of online algorithms for solving stochastic optimization problems where data comes from a non-i.i.d. process. Formally, let $\{F(\cdot;\xi), \xi \in \Xi\}$ be a collection of convex functions whose domains contain the closed convex set $\mathcal{X} \subseteq \mathbb{R}^d$ and Π be a probability distribution over the sample space Ξ. Define the convex function $f : \mathcal{X} \to \mathbb{R}$ by

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

we study online algorithms for solving the following convex optimization problem:

$$\operatorname*{minimize}_x \; f(x) \quad \text{subject to} \quad x \in \mathcal{X}. \tag{2}$$

Duchi et al. [5] study mirror descent algorithms for solving the problem (2), noting that while an extensive literature on stochastic gradient descent strategies exists (e.g. [14, 13, 12]), many approaches assume that it is possible to obtain independent and identically distributed samples from the distribution Π. When this assumption is violated—such as when one solves statistical machine learning problems with non-i.i.d. data [9, 11]—the performance of such approaches is not as clear. On the other hand, online learning algorithms have the attractive property that regret guarantees hold for arbitrary sequences of loss functions. That is, the regret of the sequence of points $x(1), \ldots, x(T)$ the algorithm plays measured against a fixed predictor x^*, defined on the sequence $\xi_1, \xi_2, \ldots \subseteq \Xi$ as

$$\mathfrak{R}_T := \sum_{t=1}^T F(x(t);\xi_t) - F(x^*;\xi_t), \tag{3}$$

is provably small for arbitrary sequences of training examples ξ_t without assuming any statistical regularity of the sequence. In stochastic optimization and statistical learning settings, however, it is the optimization error on the expected (or population) function f defined by the expectation (1) that is of central interest.

When data is drawn i.i.d. from an underlying probability distribution, Cesa-Bianchi et al. [3] have shown that online learning algorithms can indeed output predictors that approximately minimize the expected function f. Specifically, for convex loss functions, the average of the T predictors played by the online algorithm has optimization error (nearly) bounded by the average regret $\frac{\mathfrak{R}_T}{T}$ with high probability. We ask the same question when the data ξ_t is drawn according to a dependent process. To that end, we show that online learning algorithms enjoy high-probability convergence guarantees when the samples $\xi_1, \xi_2, \ldots$ form a β- or ϕ-mixing sequence. In particular, we prove that stable online learning algorithms—those that do not change the predictor $x(t)$ too aggressively between iterations—converge to the minimum of the population objective f. In favorable regimes of geometric mixing, we demonstrate convergence rates of $\mathcal{O}(\log T/\sqrt{T})$ after T iterations when the loss functions F are convex and Lipschitz. We also demonstrate faster $\mathcal{O}(\log T/T)$ convergence when the loss function is strongly convex in the hypothesis x, which is often the case for regularized problems such as SVMs, regularized logistic regression, , kernel ridge regression and maximum entropy estimation. In addition, we consider linear prediction settings, and show $\mathcal{O}(\log T/T)$ convergence when a strongly convex loss is applied to a linear predictor $\langle x, \cdot \rangle$, which gives fast rates for problems such as least squares SVMs, linear regression, logistic regression, and boosting over bounded sets, when samples are not independent and identically distributed.

We build off of a recent paper by Duchi et al. [5], who show high probability bounds on the convergence of the mirror descent algorithm even in dependent noise settings. In particular, while their results only apply to mirror descent algorithm, here we show a broad family of algorithms to optimize with non-i.i.d. data. We also extend their martingale techniques by exploiting recent ideas of Kakade and Tewari [8], and use a weakened versions of β and ϕ-mixing for our high probability results. Our proofs use only relatively elementary martingale convergence arguments, and we do not require that the input data is stationary but only that it is suitably convergent. More details and proofs can be found in the full-length version [1].

2 Setup, Assumptions and Notation

The online algorithm receives data $\xi_1, \ldots, \xi_T$ from the sample space Ξ, where the data is generated according to a stochastic process P. The algorithm plays points (hypotheses) $x \in \mathcal{X}$, and at iteration t suffers the loss $F(x(t); \xi_t)$. The total variation distance between distributions P and Q, with densities p and q w.r.t. a measure μ,[1], defined on a space S is

$$d_{\mathrm{TV}}(P, Q) := \sup_{A \subset S} |P(A) - Q(A)| = \frac{1}{2} \int_S |p(s) - q(s)| d\mu(s). \tag{4}$$

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 $\xi_1, \ldots, \xi_s$. Our main assumption is that there is a stationary distribution Π to which the distribution of ξ_t converges as t grows. We also assume that the distributions $P^t_{[s]}$ and Π are absolutely continuous with respect to a fixed underlying measure μ throughout. We use the following criterion to measure the convergence of P:

Definition 2.1 (Weak β and ϕ-mixing)**.** *The β and ϕ-mixing coefficients of the sampling distribution P are defined, respectively, as*

$$\beta(k) := \sup_{t \in \mathbb{N}} \left\{ 2\mathbb{E}[d_{\mathrm{TV}}(P^{t+k}(\cdot \mid \mathcal{F}_t), \Pi)] \right\} \quad \textit{and} \quad \phi(k) := \sup_{t \in \mathbb{N}} \left\{ 2 d_{\mathrm{TV}}(P^{t+k}(\cdot \mid B), \Pi) \,:\, B \in \mathcal{F}_t \right\}.$$

The process is ϕ-mixing (respectively, β-mixing) if $\phi(k) \to 0$ as $k \to \infty$, and we assume without loss that β and ϕ are non-increasing. The above definition is weaker than the standard definitions of mixing [11, 17], which require mixing over the entire tail σ-field of the process; we require mixing over only the single-slice marginal of ξ_{t+k}. We also note that β-mixing is weaker than ϕ-mixing since $\beta \leq \phi$. Two regimes of mixing are of special interest. A process is called geometrically β-mixing if $\beta(k) \leq \beta_0 \exp(-\beta_1 k^\theta)$ for some $\beta_i, \theta > 0$, and similarly for ϕ. Stochastic processes satisfying geometric mixing include finite-state ergodic Markov chains; we refer the reader to the reference [10] for more examples. A process is

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

algebraically β-mixing if $\beta(k) \le \beta_0 k^{-\theta}$ for some $\beta_0, \theta > 0$ ($\phi(k) \le \phi_0 k^{-\theta}$). Algebraic mixing holds for certain Metropolis-Hastings samplers when the proposal distribution does not have a lower bounded density [7], some queuing systems, and other unbounded processes.

We make the following standard boundedness assumptions on F and the domain $\mathcal{X}$:

Assumption A (Boundedness)**.** *For μ-a.e. ξ, the functions $F(\cdot;\xi)$ are convex and G-Lipschitz with respect to a norm $\|\cdot\|$ over $\mathcal{X}$, and $\mathcal{X}$ is compact with finite radius:*

$$|F(x;\xi) - F(y;\xi)| \le G\|x-y\| \quad \text{and} \quad \|x-y\| \le R \tag{5}$$

for all $x, y \in \mathcal{X}$. Further, $F(x;\xi) \in [0, GR]$.

Assumption A implies that the function f defined by the expectation (1) is G-Lipschitz. We give somewhat stronger results in the presence of the following additional assumption.

Assumption B (Strong Convexity)**.** *The expected function f is σ-strongly convex with respect to the norm $\|\cdot\|$, that is, for all $g \in \partial f(x)$,*

$$f(y) \ge f(x) + \langle g, y-x\rangle + \frac{\sigma}{2}\|x-y\|^2 \quad \text{for } x, y \in \mathcal{X}. \tag{6}$$

To prove generalization error bounds for online learning algorithms, we require them to be appropriately stable, as described in the next assumption.

Assumption C. *There is a non-increasing sequence $\kappa(t)$ such that if $x(t)$ and $x(t+1)$ are successive iterates of the online algorithm, then $\|x(t) - x(t+1)\| \le \kappa(t)$.*

Here $\|\cdot\|$ is the same norm as that used in Assumption A. This assumption is different from the stability condition of Mohri and Rostamizadeh [11], and neither implies the other. It is common (or at least often straightforward) to bound $\kappa(t)$ as a part of the regret analysis of online algorithms (e.g. [16, Lemma 10]), which motivates our assumption here.

3 Main Results

Our goal is to use the sequence $x(1), \ldots, x(T)$ to construct an estimator $\widehat{x}_T$ that has a small optimization error $f(\widehat{x}_T) - f(x^*)$. In the setting where the samples ξ_t are independent and identically distributed [3, 8], the average $\widehat{x}_T$ of the T predictors $x(1), x(2), \ldots, x(T)$ the online algorithm plays satisfies this condition. We state all our convergence results for the same averaged predictor $\widehat{x}_T = [x(1) + \cdots + x(T)]/T$, and provide proofs of all our results in the full length version of this paper [1].

3.1 Convergence rate for convex functions

We begin with a bound on the expected generalization error of $\widehat{x}_T$ for convex Lipschitz losses; since we measure expectation, β-mixing is a sufficient condition for the result.

Theorem 1. *Under Assumptions A and C, for any $\tau \in \mathbb{N}$ the predictor $\widehat{x}_T$ satisfies*

$$\mathbb{E}[f(\widehat{x}_T)] - f(x^*) \le \frac{1}{T}\mathbb{E}[\mathfrak{R}_T] + \beta(\tau)GR + \frac{\tau G}{T}\left(R + \sum_{t=1}^{T}\kappa(t)\right).$$

The above result holds only in expectation and provides no control over deviations of the optimization error. We can apply new martingale concentration techniques to achieve stronger high-probability bounds as in the following theorem.

Theorem 2. *Under Assumptions A and C,*

(a) with probability at least $1-\delta$, for any $\tau \in \mathbb{N}$ the predictor $\widehat{x}_T$ satisfies

$$f(\widehat{x}_T) - f(x^*) \le \frac{1}{T}\mathfrak{R}_T + \frac{\tau G}{T}\sum_{t=1}^{T}\kappa(t) + 2GR\sqrt{\frac{2\tau}{T}\log\frac{\tau}{\delta}} + \phi(\tau)GR + \frac{\tau GR}{T}.$$

(b) with probability at least $1 - 2\delta$*, for any* $\tau \in \mathbb{N}$ *the predictor* $\widehat{x}_T$ *satisfies*

$$f(\widehat{x}_T) - f(x^*) \leq \frac{1}{T}\mathfrak{R}_T + \frac{\tau G}{T}\sum_{t=1}^{T}\kappa(t) + 2GR\sqrt{\frac{2\tau}{T}\log\frac{2\tau}{\delta}} + \frac{2\beta(\tau)GR}{\delta} + \frac{\tau GR}{T}.$$

To better illustrate our results, we now specialize them under concrete mixing assumptions. We begin with a corollary giving error bounds for geometrically ϕ-mixing processes (as defined in Section 2).

Corollary 1. *In addition to the conditions of Theorem 2(a), assume* $\phi(k) \leq \phi_0 \exp(-\phi_1 k^\theta)$*. There exists a finite universal constant* C *such that with probability at least* $1 - \delta$

$$f(\widehat{x}_T) - f(x^*) \leq \frac{1}{T}\mathfrak{R}_T + C \cdot \left[\frac{(\log T)^{1/\theta} G}{T\phi_1}\sum_{t=1}^{T}\kappa(t) + GR\sqrt{\frac{(\log T)^{1/\theta}}{\phi_1 T}\log\frac{(\log T)^{1/\theta}}{\delta}}\right].$$

Under geometric ϕ-mixing, the probabilistic terms are of the same order as the i.i.d. setting [3] to within poly-logarithmic factors. Algebraic mixing yields somewhat slower rates (see the full version [1] of this paper). More generally, under the same condition on the stability, an argument similar to that for Corollary 7 of Duchi et al. [5] implies $f(\widehat{x}_T) - f(x^*) \to 0$ with probability one when $\phi(k) \to 0$ as $k \to \infty$. Theorem 2(b) can also be extended to similar corollaries but we omit such discussion here due to lack of space.

To obtain a concrete convergence rate from our results, we need to know bounds on the stability sequence $\kappa(t)$ (and the regret $\mathfrak{R}_T$). Indeed, for two common first-order methods, online mirror-descent (see e.g. Theorem 11.1 of [4] or the paper [12]) and regularized dual averaging [16], known convergence results combined with Theorem 2 give that with probability at least $1 - \delta$,

$$f(\widehat{x}_T) - f(x^*) \leq \frac{1}{T}\mathfrak{R}_T + C \cdot \inf_{\tau \in \mathbb{N}}\left[\frac{GR\tau}{\sqrt{T}} + \frac{GR}{\sqrt{T}}\sqrt{\tau\log\frac{\tau}{\delta}} + \phi(\tau)GR\right], \tag{7}$$

for some universal constant C. The bound (7) captures the known convergence rates for i.i.d. sequences [3, 12, 16, 4] by taking $\tau = 1$, since $\phi(1) = 0$. Further specializing to the geometric mixing rate of Corollary 1, one obtains an error bound of $\mathcal{O}(1/(\phi_1\sqrt{T}))$ to poly-logarithmic factors; this is essentially same as the generalization error in i.i.d. settings.

3.2 Convergence rates for strongly convex functions

It is by now well-known that the regret of online learning algorithms scales as $\mathcal{O}(\log T)$ for strongly convex functions, a result which is originally due to Hazan et al. [6]. Many online algorithms, including gradient and mirror descent [2, 6, 15] and dual averaging [16, Lemma 10], satisfy the stability bound $\|x(t) - x(t+1)\| \leq G/(\sigma t)$ when the loss functions are σ-strongly convex. Under these conditions, Theorem 1 gives an expected generalization bound of $\mathcal{O}(\inf_{\tau\in\mathbb{N}}\{\beta(\tau) + \tau\log T/T\})$, which is faster than the standard rate of $\mathcal{O}(\inf_{\tau\in\mathbb{N}}\{\beta(\tau) + \sqrt{\tau/T}\})$, but the improvement in rates does not directly apply to Theorem 2. In the next theorem, we derive sharper guarantees when the expected function f is strongly convex by extending a self-bounding martingale argument due to Kakade and Tewari [8].

Theorem 3. *Let Assumptions A–C hold. For any* $\delta < 1/e$*,* $T \geq 3$*, with probability at least* $1 - 4\delta\log n$*, for any* $\tau \in \mathbb{N}$ *the predictor* $\widehat{x}_T$ *satisfies*

$$f(\widehat{x}_T) - f(x^*) \leq \frac{2}{T}\mathfrak{R}_T + \frac{2\tau G}{T}\sum_{t=1}^{T}\kappa(t) + \frac{32G^2\tau}{\sigma T}\log\frac{\tau}{\delta} + \frac{4\tau RG}{T}\left(3\log\frac{\tau}{\delta} + 1\right) + 2RG\phi(\tau).$$

We can establish the following corollary under the previously mentioned stability bound $\|x(t) - x(t+1)\| \leq G/(\sigma t)$ for strongly convex online learning algorithms.

Corollary 2. *In addition to the conditions of Theorem 3, assume the stability bound* $\kappa(t) \leq G/\sigma t$*. There is a universal constant* C *such that with probability at least* $1 - 4\delta\log T$*,*

$$f(\widehat{x}_T) - f(x^*) \leq \frac{2}{T}\mathfrak{R}_T + C \cdot \inf_{\tau\in\mathbb{N}}\left[\frac{\tau G^2}{\sigma T}\log T + \frac{\tau G^2}{\sigma T}\log\frac{\tau}{\delta} + \frac{G^2}{\sigma}\phi(\tau)\right].$$

The factor of 2 in front of $\mathfrak{R}_T$ is insignificant, since $\mathfrak{R}_T = o(T)$ for any low-regret online learning algorithm, so we have no loss in rates. For a few concrete examples, we note that when the losses $F(\cdot;\xi)$ are σ-strongly convex, online gradient and mirror descent [6, 15], as well as dual averaging [16], all have $\mathfrak{R}_T \le C \cdot G^2 \log T/\sigma T$, so Corollary 2 implies the convergence bound $f(\widehat{x}_T) - f(x^*) = \mathcal{O}((G^2/\sigma)\inf_{\tau\in\mathbb{N}}[\tau \log T/T + \phi(\tau)])$ with high probability. For example, we can now observe that that online algorithms for regularized SVMs satisfy a sharp high-probability generalization guarantee, even for non-i.i.d. data.

3.3 Fast rates for linear prediction

We now turn to the common statistical prediction setting in which samples come in pairs of the form $(\xi, \nu) \in \Xi \times \mathcal{V}$, where ν is the label or target value of the sample $\xi \in \mathbb{R}^d$, where $\|\xi\|_2 \le r$. We measure the performance of the hypothesis x on the example (ξ, ν) by

$$F(x; (\xi, \nu)) = \ell(\nu, \langle \xi, x\rangle), \quad \ell : \mathcal{V} \times \mathbb{R} \to \mathbb{R}_+, \tag{8}$$

where ℓ is a loss measuring the accuracy of the prediction $\langle \xi, x\rangle$. Many learning problems fall into the framework (8): linear regression, where the loss is $\ell(\nu, \langle \xi, x\rangle) = \frac{1}{2}(\nu - \langle \xi, x\rangle)^2$; logistic regression, where $\ell(\nu, \langle \xi, x\rangle) = \log(1 + \exp(-\nu \langle \xi, x\rangle))$; boosting; and SVMs all have the form (8). The natural curvature assumption in this setting is the following.

Assumption D (Linear strong convexity)**.** *For fixed ν, the loss function $\ell(\nu, \cdot)$ is a σ-strongly convex and L-Lipschitz scalar function over $[-Rr, Rr]$:*

$$\ell(\nu, b) \ge \ell(\nu, a) + \ell'(\nu, a)(b - a) + \frac{\sigma}{2}(b - a)^2 \quad \textit{and} \quad |\ell(\nu, b) - \ell(\nu, a)| \le L|a - b|$$

for any $a, b \in \mathbb{R}$ with $\max\{|a|, |b|\} \le Rr$.

Logistic and linear regression, least-squares SVMs, and boosting on a bounded domain satisfy Assumption D. Whenever the covariance $\mathrm{Cov}(\xi)$ of ξ is non-degenerate under the stationary distribution Π, Assumption D implies the expected function f is strongly convex, putting us in the setting of Theorem 3. If we had access to a stable online learning algorithm with small regret (i.e. both $\mathfrak{R}_T = \mathcal{O}(\log T)$ and $\sum_{t=1}^T \kappa(t) = \mathcal{O}(\log T)$) for losses of the form (8) satisfying Assumption D, we could simply apply Theorem 3 to guarantee good generalization properties of the predictor $\widehat{x}_T$. We do not know of an existing algorithm satisfying our desiderata of logarithmic regret and stability. Online gradient descent and dual averaging algorithms guarantee stability, but do not attain $\mathcal{O}(\log T)$ regret since $F(\cdot;\xi)$ is no longer strongly convex, and while Hazan et al. [6] show that online Newton and follow the approximate leader (FTAL) algorithms have logarithmic regret, neither algorithm satisfies even a weakened form of the stability assumption C. Thus we define a new update that combines FTAL with the Vovk-Azoury-Warmuth forecaster [4, Chapter 11.8]:

$$x(t) = \operatorname*{argmin}_{x\in\mathcal{X}} \left\{ \sum_{i=1}^{t-1} \langle \nabla F(x(i); (\xi_i, \nu_i)), x\rangle + \frac{\sigma}{2} \sum_{i=1}^{t-1} \langle x(i) - x, \xi_i\rangle^2 + \frac{\sigma}{2} x^\top (\xi_t \xi_t^\top + \epsilon I) x \right\}. \tag{9}$$

We have the following result for the algorithm (9) under Assumption D.

Theorem 4. *Assume $\|x\|_2 \le R$ for $x \in \mathcal{X} \subseteq \mathbb{R}^d$, and let $x(t)$ be generated according to the update (9) with $\epsilon = 1$. Then with probability at least $1 - 4\delta \log n$, for any $\tau \in \mathbb{N}$,*

$$f(\widehat{x}_T) - f(x^*) \le \frac{L^2 d}{\sigma T}(9 + 14\tau) \log\left(r^2 T + 1\right) + \frac{\sigma}{T}\|x^*\|_2^2 + \frac{32 L^2 r^2 \tau}{\sigma T \cdot \lambda_{\min}(\mathrm{Cov}(\xi))} \log \frac{\tau}{\delta} + \frac{8\tau L^2}{\sigma T}\left(3 \log \frac{\tau}{\delta} + 1\right) + \frac{4L^2}{\sigma}\phi(\tau).$$

The proof of the theorem requires showing a regret guarantee for the update (9), which adapts related arguments [6, 16] in the literature, as well as developing and controlling a new weakened form of the stability assumption C. We can further specialize Theorem 4 using different mixing mixing assumptions on the process. As in Corollary 1, we have

Corollary 3. *In addition to the conditions of Theorem 4, assume P is geometrically ϕ-mixing. There exists a universal constant C such that with probability at least $1 - \delta \log T$,*

$$f(\widehat{x}_T) - f(x^*) \leq C \cdot \left[\frac{L^2 d (\log T)^{1+\frac{1}{\theta}}}{\phi_1 \sigma T} + \frac{L^2 (\log T)^{\frac{1}{\theta}}}{\phi_1 \sigma T \cdot \lambda_{\min}(\mathrm{Cov}(\xi))} \log\left(\frac{\log T}{\delta}\right) \right].$$

4 Conclusions and Discussion

In this paper, we have shown that the martingale concentration arguments used to derive online to batch conversions for independent data [3, 8] can be extended to situations where the somewhat brittle assumptions of i.i.d. samples do not hold. As is the case for earlier generalization guarantees for online learning, our arguments require only elementary martingale convergence arguments, and we do not need the more powerful tools of empirical process theory (e.g. [17]). Our results are of particular interest for machine learning problems: just as guarantees for stochastic optimization imply generalization error bounds in the i.i.d. case [3], our results establish that any stable online learning algorithm produces a hypothqesis that can generalize well even on non-i.i.d. data samples. Additionally, our results extend to the settings considered by Duchi et al. [5], yielding a family of algorithms for distributed optimization and optimization over combinatorial spaces.

References

[1] A. Agarwal and J. C. Duchi. The generalization ability of online algorithms for dependent data. URL `http://arxiv.org/abs/1110.2529`, 2011.

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

[3] N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50:2050–2057, 2004.

[4] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

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

[6] E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69, 2007.

[7] S.F. Jarner and G.O. Roberts. Polynomial convergence rates of markov chains. *The Annals of Applied Probability*, 12(1):pp. 224–247, 2002.

[8] S. M. Kakade and A. Tewari. On the generalization ability of online strongly convex programming algorithms. In *Advances in Neural Information Processing Systems 21*, 2009.

[9] R. Meir. Nonparametric time series prediction through adaptive model selection. *Machine Learning*, 39:5–34, 2000.

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

[11] M. Mohri and A. Rostamizadeh. Stability bounds for stationary ϕ-mixing and β-mixing processes. *Journal of Machine Learning Research*, 11:789–814, 2010.

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

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

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

[15] S. Shalev-Shwartz and Y. Singer. Logarithmic regret algorithms for strongly convex repeated games. Technical Report 42, The Hebrew University, 2007.

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

[17] Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *Annals of Probability*, 22(1):94–116, 1994.