TAIL BOUNDS FOR STOCHASTIC APPROXIMATION

MICHAEL P. FRIEDLANDER[*] AND GABRIEL GOH[*]

Abstract. Stochastic-approximation gradient methods are attractive for large-scale convex optimization because they offer inexpensive iterations. They are especially popular in data-fitting and machine-learning applications where the data arrives in a continuous stream, or it is necessary to minimize large sums of functions. It is known that by appropriately decreasing the variance of the error at each iteration, the expected rate of convergence matches that of the underlying deterministic gradient method. Conditions are given under which this happens with overwhelming probability.

Key words. stochastic approximation, sample-average approximation, incremental gradient, steepest descent, convex optimization

AMS subject classifications. 90C15, 90C25

1. Introduction.

Stochastic-approximation methods for convex optimization are prized for their inexpensive iterations and applicability to large-scale problems. The convergence analyses for these methods typically rely on expectation-based metrics for gauging progress towards a solution. But because the solution path is itself stochastic, practitioners—especially those relying on ad-hoc applications of such algorithms for a limited number of iterations—may pause and question how far an observed solution path is from the optimal value. The aim of this paper is to develop bounds on the probability of deviating too far from the solution. This result complements expectation-based analysis, and can furnish useful guidance for practitioners.

Consider the differentiable convex optimization problem

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad f(x),$$

and the approximate gradient descent iteration

$$x_{k+1} = x_k - \alpha_k(\nabla f(x_k) + e_k), \tag{1.1}$$

where e_k is a random variable. The gradient residual e_k may, for example, account for the error incurred in the computation of the gradient $\nabla f(x_k)$. Bertsekas and Tsitsiklis [2] give mild conditions on e_k and f under which $f(x_k) \to \inf f(x)$ in probability. Friedlander and Schmidt [5] link the convergence rate to $\mathbf{E}\,\|e_k\|^2$, which measures the variance in the error. Our goal here is to complement these results by providing conditions under which $f(x_k) \to \inf f(x)$ linearly with overwhelming probability; see Theorem 3.5 of section 3.2.

Two applications of this framework are to provide tail bounds for stochastic-approximation and for incremental-gradient methods for minimizing the function

$$f(x) := \mathbf{E}\, F(x, Z),$$

where Z is a random variable. In the context of stochastic approximation, at each iteration a random sample $\{Z_1, \dots, Z_{m_k}\}$ of size m_k is generated to compute the search direction

$$\nabla f(x_k) + e_k = \frac{1}{m_k}\sum_{i=1}^{m_k} F(x_k, Z_i). \tag{1.2}$$

[*] Department of Computer Science, University of British Columbia, Vancouver, B.C. V6T 1Z4, Canada ({mpf,ggoh}@cs.ubc.ca). The work of these authors was supported by NSERC Discovery Grant 312104. April 19, 2013

In the case where Z takes on a finite number of values with uniform probability, then f reduces to the familiar case of sums of functions:

$$f(x) = \frac{1}{M}\sum_{i=1}^{M} f_i(x). \tag{1.3}$$

In this context, the incremental-gradient method chooses search directions

$$\nabla f(x_k) + e_k = \frac{1}{m_k}\sum_{i\in\mathcal{S}_k} \nabla f_i(x_k), \tag{1.4}$$

where the random sample $\mathcal{S}_k \subseteq \{1,\dots,M\}$ of size m_k is chosen without replacement.

At one extreme is a fixed sample size m_k (equal to 1, say), which yields an inexpensive iteration but generally does not converge to a minimizer unless $\alpha_k \to 0$; at best it converges sublinearly to the solution. At the other extreme is steepest descent, which surely converges linearly. As do Friedlander and Schmidt [5] and Byrd, Chin, Nocedal, and Wu [3], we consider a method for interpolating between these two extremes by increasing the sample size at a linear rate. We show that this and related algorithms converge linearly with overwhelming probability; see section 6.

1.1. Assumptions and notation.

We make the blanket assumption throughout that the function f is strongly convex and that its gradient ∇f is uniformly Lipschitz continuous, i.e., there exist positive constants μ and L such that for all $x, y \in \mathbb{R}^n$,

$$f(y) \ge f(x) + \langle y-x, \nabla f(x)\rangle + \mu/2\|y-x\|^2, \tag{1.5a}$$

$$\|\nabla f(y) - \nabla f(x)\| \le L\|y-x\|. \tag{1.5b}$$

Throughout, we use the notation $\rho := 1-\mu/L$, which can be interpreted as a normalized condition number of f. Let $\pi_k := f(x_k) - \inf f(x)$ be the distance to the optimal value, and $R_k := \sum_{i=0}^{k-1}\rho^i$. Let $\mathcal{F}_k = \sigma(e_1, e_2, \dots, e_k)$ be the σ-algebra generated by the error history. When the context is clear, $[z]_i$ denotes the ith component of a vector z. Except for the discussion in section 1.2, we make the assumption that $\alpha_k \equiv 1/L$.

1.2. Existing convergence analysis.

It is well known that deterministic steepest descent with a constant stepsize $\alpha_k = 1/L$ converges linearly with a rate constant that depends on the condition number ρ, i.e.,

$$\pi_k \le \rho^k \pi_0; \tag{1.6}$$

see [9, section 8.6]. Because it is convenient to phrase our convergence results for the stochastic method in terms of its deviation from the deterministic case, we derive most results in terms of a tail bound on

$$\Pr\,(\pi_k - \rho^k\pi_0 \ge \epsilon). \tag{1.7}$$

It is straightforward to recast these results to obtain tail bounds on $\Pr(\pi_k > \epsilon)$.

In general, if $\liminf_k \|e_k\| \ne 0$, then we necessarily require $\alpha_k \to 0$ in (1.1) in order to ensure stationarity of limit points. Solodov [15] describes how bounding the steplengths away from zero yields limit points $\bar{x}$ that satisfy the approximate stationarity condition

$$\|\nabla f(\bar{x})\| = \mathcal{O}(\lim_k \alpha_k).$$

Bertsekas and Tsitsiklis [2] describe conditions for convergence of the iteration (1.1) when the steplengths α_k satisfy the infinite travel and summable conditions

$$\sum_{k=0}^{\infty} \alpha_k = \infty \quad \text{and} \quad \sum_{k=0}^{\infty} \alpha_k^2 < \infty.$$

For a non-vanishing stepsize, i.e., $\liminf_k \alpha_k > 0$, Luo and Tseng [10] show that for a decreasing error sequence that satisfies $\|e_k\| = \mathcal{O}(\|x_{k+1} - x_k\|)$, the sequence $\pi_k \to 0$ at an asymptotic linear rate. For a constant stepsize, Friedlander and Schmidt [5] give non-asymptotic rates that directly depend on the rate at which the error goes to zero.

The convergence in probability of the stochastic-approximation method was first discussed by the classic Robbins [13] paper. More recently, Nemirovski, Juditsky, Lan, and Shapiro [12] show that for decreasing steplengths $\alpha_k = \mathcal{O}(1/k)$, these methods achieve a sublinear rate according to $\mathbf{E}[\pi_k] = \mathcal{O}(1/k)$; the iteration average has similar convergence properties, and it converges sublinearly with overwhelming probability.

Bertsekas and Nedić [11] show that incremental-gradient methods for (1.3), with constant steplength $\alpha_k \equiv \alpha$, converge as

$$\mathbf{E}\, \pi_k \le \mathcal{O}(\rho^k) + \mathcal{O}(\alpha).$$

This expression is telling because the first term on the right-hand side decreases at a linear rate, and depends on the normalized condition number ρ; this term is present for any deterministic first-order method with constant stepsize. Thus, we can see that with the strong-convexity assumption and a constant steplength, the incremental-gradient algorithm has the same convergence characteristics as steepest descent, but with an additional constant-error term.

2. Gradient descent with error. Our point of departure is the following result, which relates the progress in the objective value to the norm of the gradient residual.

Lemma 2.1. *After k iterations of algorithm* (1.1),

$$\pi_k - \rho^k \pi_0 \le \frac{1}{2L} \sum_{i=0}^{k-1} \rho^{k-1-i} \|e_i\|^2.$$

Proof. From [5, Lemma 2.1], $\pi_{k+1} \le \rho \pi_k + \frac{1}{2L}\|e_k\|^2$. The required result follows from applying this inequality recursively from $i = k-1$ down to $i = 0$. □

When the errors $\|e_k\|^2$ are identically zero, the search directions in iteration (1.1) are simply gradient vectors, and this result reduces to the well-known fact that steepest descent decreases the objective value linearly with an error constant proportional to the conditioning of the problem, cf. (1.6). When the gradient residuals are nonzero, however, the inequality in Lemma 2.1 states that the deviation in progress that would have been achieved via steepest descent is bounded by the discounted sum of the errors made at each iteration.

Friedlander and Schmidt [5, §2] note that the iteration (1.1) yields a monotonic decrease in the objective value if the error is small enough. In particular,

$$\pi_{k+1} \le \pi_k \quad \text{if} \quad \|e_k\|^2 \le \|\nabla f(x_k)\|^2.$$

The following is a probabilistic generalization of this result. The dependence on the σ-algebra $\mathcal{F}_{k-1}$ is effectively a conditioning on the history of the algorithm.

THEOREM 2.2 (Supermartingale property).

$$\mathbf{E}\,[\pi_{k+1}\,|\,\mathcal{F}_{k-1}] \le \pi_k \quad \textit{if} \quad \mathbf{E}\,[\|e_k\|^2\,|\,\mathcal{F}_{k-1}] \le \|\nabla f(x_k)\|^2.$$

Proof. It follows from assumption (1.5b) that

$$f(y) \le f(x) + \langle y-x, \nabla f(x)\rangle + \frac{L}{2}\|y-x\|^2.$$

Use $x = x_k$ and $y = x_{k+1}$, as defined in (1.1), and simplify to obtain

$$\begin{aligned} f(x_{k+1}) &\le f(x_k) - \frac{1}{L}\langle \nabla f(x_k) + e_k, \nabla f(x_k)\rangle + \frac{1}{2L}\|\nabla f(x_k) + e_k\|^2 \\ &= f(x_k) - \frac{1}{L}\|\nabla f(x_k)\|^2 - \frac{1}{L}\langle \nabla f(x_k), e_k\rangle \\ &\qquad + \frac{1}{2L}\|\nabla f(x_k)\|^2 + \frac{1}{L}\langle \nabla f(x_k), e_k\rangle + \frac{1}{2L}\|e_k\|^2 \\ &= f(x_k) + \frac{1}{2L}\big(\|e_k\|^2 - \|\nabla f(x_k)\|^2\big). \end{aligned} \tag{2.1}$$

Then

$$\begin{aligned} \mathbf{E}\,[\pi_{k+1}\,|\,\mathcal{F}_{k-1}] &= \mathbf{E}\,\big[f(x_{k+1})\,|\,\mathcal{F}_{k-1}\big] - f(x_*) \\ &\overset{(i)}{\le} \mathbf{E}\left[f(x_k) + \frac{1}{2L}(\|e_k\|^2 - \|\nabla f(x_k)\|^2) \,\middle|\, \mathcal{F}_{k-1}\right] - f(x_*) \\ &= f(x_k) + \frac{1}{2L}\big(\,\mathbf{E}[\|e_k\|^2\,|\,\mathcal{F}_{k-1}] - \|\nabla f(x_k)\|^2\big) - f(x_*) \le \pi_k, \end{aligned}$$

where (i) follows from (2.1). □

EXAMPLE 2.3 (Gradient descent with independent Gaussian noise, part I). *Let $e_k \sim N(0,\sigma^2 I)$. Because $\|e_k\|^2$ is a sum of n independent Gaussians, it follows a chi-squared distribution with mean $\mathbf{E}\,\|e_k\|^2 = n\sigma^2$. Therefore,*

$$\mathbf{E}\,\pi_k - \rho^k\pi_0 \le \frac{1}{2L}\sum_{i=0}^{k-1}\rho^{k-1-i}\,\mathbf{E}\,\|e_i\|^2 = \frac{n\sigma^2}{2L}\sum_{i=0}^{k-1}\rho^{k-1-i}. \tag{2.2}$$

Take the limit inferior of both sides of (2.2), and note that $\lim_{k\to\infty}\sum_{i=0}^{k-1}\rho^{k-1-i} = 1/(1-\rho) = L/\mu$, and thus that

$$\mathbf{E}\liminf_{k\to\infty}\pi_k \le \liminf_{k\to\infty}\mathbf{E}\,\pi_k \le \frac{n\sigma^2}{2\mu},$$

where the first inequality follows from the application of Fatou's Lemma. Hence, even though $\lim_{k\to\infty}\pi_k$ may not exist, we can still bound the lower envelope on the suboptimality that is proportional to the variance of the error term.

An immediate consequence of Lemma 2.1 is a tail bound via Markov's inequality:

$$\Pr(\pi_k - \rho^k\pi_0 \ge \epsilon) \le \Pr\left(\frac{1}{2L}\sum_{i=0}^{k-1}\rho^{k-1-i}\|e_i\|^2 \ge \epsilon\right) \le \frac{1}{2L\epsilon}\sum_{i=0}^{k-1}\rho^{k-1-i}\,\mathbf{E}\,\|e_i\|^2.$$

This inequality is too weak, however, to say anything meaningful about the confidence in our solution after a finite number of iterations. We are instead interested in Chernoff-type bounds that are exponentially decreasing in ϵ, and in the parameters that control the size of the error.

3. Bounds for gradient descent with random error.

Our first bound makes no assumption on the relation of the gradient errors between iterations. Thus, it may or may not allow for history-dependent errors, and we call this a generic error sequence. The second bound makes the stronger assumption about the relationship of the errors between iterations.

3.1. Generic error sequence.

Our first exponential tail bounds are defined in terms of the moment-generating function of the error norms $\|e_k\|^2$, denoted by

$$\gamma_k(\theta) := \mathbf{E}\exp(\theta\|e_k\|^2).$$

We make the convention that $\gamma_k(\theta) = +\infty$ for $\theta \notin \operatorname{dom}\gamma_k$.

Theorem 3.1 (Tail bound for generic errors). *For algorithm* (1.1),

$$\Pr(\pi_k - \rho^k\pi_0 \geq \epsilon) \leq \inf_{\theta>0}\left\{\frac{\exp(-2\theta L\epsilon/R_k)}{R_k}\sum_{i=0}^{k-1}\rho^{k-1-i}\gamma_i(\theta)\right\}. \tag{3.1a}$$

If $\gamma_k \equiv \gamma$ for all k (i.e., the error norms $\|e_k\|^2$ are identically distributed), then the bound simplifies to

$$\Pr(\pi_k - \rho^k\pi_0 \geq \epsilon) \leq \inf_{\theta>0}\left\{\exp(-2\theta L\epsilon/R_k)\,\gamma(\theta)\right\}. \tag{3.1b}$$

Proof. Note that $\left(\sum_{i=0}^{k-1}\rho^{k-1-i}\right)/R_k = 1$. For $\theta > 0$,

$$\begin{aligned}\mathbf{E}\exp\left(\theta\sum_{i=0}^{k-1}\rho^{k-1-i}\|e_i\|^2\right) &\overset{(i)}{=} \mathbf{E}\exp\left(\sum_{i=0}^{k-1}\frac{\rho^{k-1-i}}{R_k}\theta R_k\|e_i\|^2\right)\\ &\overset{(ii)}{\leq} \mathbf{E}\sum_{i=0}^{k-1}\frac{\rho^{k-1-i}}{R_k}\exp(\theta R_k\|e_i\|^2)\\ &= \frac{1}{R_k}\sum_{i=0}^{k-1}\rho^{k-1-i}\gamma_i(\theta R_k),\end{aligned}$$

where (i) follows from the convexity of $\exp(\cdot)$, and (ii) follows from the linearity of the expectation operator and the definition of γ_i. Together with Markov's inequality, the above implies that for all $\theta > 0$,

$$\begin{aligned}\Pr\left(\sum_{i=0}^{k-1}\rho^{k-1-i}\|e_i\|^2 \geq \epsilon\right) &= \Pr\left(\exp\left[\theta\sum_{i=0}^{k-1}\rho^{k-1-i}\|e_i\|^2\right] \geq \exp(\theta\epsilon)\right)\\ &\leq \exp(-\theta\epsilon)\,\mathbf{E}\exp\left(\theta\sum_{i=0}^{k-1}\rho^{k-1-i}\|e_i\|^2\right)\\ &\leq \frac{\exp(-\theta\epsilon)}{R_k}\sum_{i=0}^{k-1}\rho^{k-1-i}\gamma_i(\theta R_k).\end{aligned} \tag{3.2}$$

This inequality, together with Lemma 2.1, implies that for all $\theta > 0$,

$$\begin{aligned}\Pr\left(\pi_k - \rho^k\pi_0 \ge \epsilon\right) &\le \Pr\left(\frac{1}{2L}\sum_{i=0}^{k-1}\rho^{k-1-i}\|e_i\|^2 \ge 2L\epsilon\right)\\ &\le \frac{\exp(-2\theta L\epsilon)}{R_k}\sum_{i=0}^{k-1}\rho^{k-1-i}\gamma_i(\theta R_k),\end{aligned}$$

where we use the elementary fact that $\Pr(X \ge \epsilon) \le \Pr(Y \ge \epsilon)$ if $X \le Y$ almost surely. Redefine θ as θR_k, and take the infimum of the right-hand side over $\theta > 0$, which gives the required inequality (3.1a). The simplified bound (3.1b) follows directly from the definition of R_k. □

In the case where the errors are identically distributed, there is an intriguing connection between the tail bounds described in Theorem 3.1 and the convex conjugate $[\cdot]^*$ of the cumulant-generating function of that distribution, i.e., $\log\circ\gamma$.

Corollary 3.2 (Tail bound for identically-distributed errors). *Suppose that the error norms $\|e_k\|^2$ are identically distributed. Then for algorithm (1.1),*

$$\log\Pr(\pi_k - \rho^k\pi_0 \ge \epsilon) \le -\left[\log\gamma(\cdot)\right]^*(2L\epsilon/R_k).$$

Proof. Take the log of both sides of (3.1b) to get

$$\begin{aligned}\log\Pr(\pi_k - \rho^k\pi_0 \ge \epsilon) &\le \log\inf_{\theta>0}\left\{\exp(-2\theta L\epsilon/R_k)\,\gamma(\theta)\right\}\\ &= -\sup_{\theta>0}\left\{(2L\epsilon/R_k)\theta - \log\gamma(\theta)\right\},\end{aligned}$$

which we recognize as the negative of the conjugate of $\log\gamma(\cdot)$ evaluated at $2L\epsilon/R_k$. □

Note that these bounds are invariant with regard to scaling, in the sense that if the objective function f is scaled by some $\alpha > 0$, then the bounds hold for $\alpha\epsilon$.

The following example illustrates an application of this tail bound to the case where the errors follow a simple distribution with a known moment-generating function.

Example 3.3 (Gradient descent with independent Gaussian noise, part II). *As in Example 2.3, let $e_k \sim N(0, \sigma^2 I)$. Then $\|e_k\|^2$ is a scaled chi-squared distribution with moment-generating function*

$$\gamma_k(\theta) = (1 - 2\sigma^2\theta)^{-n/2}, \quad \theta \in \left[0, 1/2\sigma^2\right).$$

Note that

$$\left[\log\gamma(\cdot)\right]^*(\mu) = \frac{\mu - n\sigma^2}{2\sigma^2} + \frac{n}{2}\log(n\sigma^2/\mu) \quad \textit{for} \quad \mu > n\sigma^2.$$

We can then apply Corollary 3.2 to this case to deduce the bound

$$\Pr(\pi_k - \rho^k\pi_0 \ge \epsilon) \le \left(\frac{2\exp(1)}{n}\cdot\frac{L\epsilon}{\sigma^2 R_k}\right)^{n/2}\exp\left(-\frac{L\epsilon}{\sigma^2 R_k}\right) \quad \textit{for} \quad \epsilon > \frac{n\sigma^2 R_k}{2L}.$$

The bound can be further simplified by introducing an additional perturbation $\delta > 0$ that increases the base of the exponent:

$$\Pr(\pi_k - \rho^k\pi_0 \ge \epsilon) = \mathcal{O}\left[\exp\left(-\delta\frac{L\epsilon}{\sigma^2 R_k}\right)\right] \quad \textit{for all} \quad \delta \in [0,1), \tag{3.3}$$

which highlights the exponential decrease of the bound in terms of ϵ.

3.2. Unconditionally bounded error sequence.

In contrast to the previous section, we now assume that there exists a deterministic bound on the conditional expectation $\mathbf{E}\left[\exp(\theta\|e_k\|^2)\,|\,\mathcal{F}_{k-1}\right]$. We say that this bound holds unconditionally because it holds irrespective of the history of the error sequence.

Assumption 3.4. *Assume that* $\mathbf{E}\left[\exp(\theta\|e_k\|^2)\,|\,\mathcal{F}_{k-1}\right]$ *is finite over* $[0,\sigma)$, *for some* $\sigma > 0$. *Therefore there exists, for each* k, *a deterministic function* $\bar{\gamma}_k : \mathbb{R}_+ \to \mathbb{R}_+ \cup \{\infty\}$ *such that*

$$\bar{\gamma}_k(0) = 1 \quad \textit{and} \quad \mathbf{E}\left[\exp(\theta\|e_k\|^2)\,|\,\mathcal{F}_{k-1}\right] \le \bar{\gamma}_k(\theta).$$

(Thus, the bound is tight at $\theta = 0$.*)*

The existence of such a function in fact implies a bound on the moment-generating function of $\|e_k\|^2$. In particular,

$$\gamma_k(\theta) := \mathbf{E}\exp(\theta\|e_k\|^2) = \mathbf{E}\left[\mathbf{E}\left[\exp(\theta\|e_k\|^2)\,|\,\mathcal{F}_{k-1}\right]\right] \le \mathbf{E}\,\bar{\gamma}_k(\theta) = \bar{\gamma}_k(\theta). \tag{3.4}$$

The converse, however, is not necessarily true. To see this, consider the case where the errors $e_1, \ldots, e_{k-1}$ are independent Bernoulli-distributed random variables, and e_k is a deterministic function of all the previous errors, e.g., $\Pr(e_i = 0) = \Pr(e_i = 1) = 1/2$ for $i = 1, \ldots, k-1$, and the error on the last iteration is completely determined by the previous errors:

$$e_k = \begin{cases} 1 & \text{if } e_1 = e_2 = \cdots = e_{k-1}, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, $\Pr(e_k = 1) = (1/2)^{k-1}$ and $\Pr(e_k = 0) = 1 - (1/2)^{k-1}$, and the moment-generating function of e_k is $\gamma_k(\theta) = 1 - 2^{1-k}(1 + \exp(\theta))$. Then,

$$\mathbf{E}[\exp(\theta e_k^2)\,|\,e_1, \ldots, e_{k-1}] = \begin{cases} \exp(\theta) & \text{if } e_1 = e_2 = \cdots = e_{k-1}, \\ 1 & \text{otherwise,} \end{cases}$$

whose tightest deterministic upper bound is $\bar{\gamma}_k(\theta) = \exp(\theta)$. However, $\bar{\gamma}_k(\theta) \ge \gamma_k(\theta)$ for all $\theta \ge 0$.

The following result is analogous to Theorem 3.1.

Theorem 3.5 (Tail bounds for unconditionally bounded errors). *Suppose that Assumption 3.4 holds. Then for algorithm* (1.1),

$$\Pr(\pi_k - \rho^k\pi_0 \ge \epsilon) \le \inf_{\theta>0}\left\{\exp(-2\theta L\epsilon)\prod_{i=0}^{k-1}\bar{\gamma}_i(\theta\rho^{k-i-1})\right\}.$$

Proof. The proof follows the same outline as many martingale-type inequalities [1, 4]. We obtain the following relationships:

$$\begin{aligned}
\mathbf{E}\exp\left[\theta\sum_{i=0}^{k-1}\rho^{k-1-i}\|e_i\|^2\right] &\overset{(i)}{=} \mathbf{E}\left[\mathbf{E}\left[\exp\left[\theta\sum_{i=0}^{k-1}\rho^{k-1-i}\|e_i\|^2\right]\middle|\,\mathcal{F}_{k-2}\right]\right]\\
&= \mathbf{E}\left[\mathbf{E}\left[\exp\left[\theta\rho^0\|e_{i-1}\|^2+\theta\sum_{i=0}^{k-2}\rho^{k-1-i}\|e_i\|^2\right]\middle|\,\mathcal{F}_{k-2}\right]\right]\\
&\overset{(ii)}{=} \mathbf{E}\left[\exp\left[\theta\sum_{i=0}^{k-2}\rho^{k-1-i}\|e_i\|^2\right]\mathbf{E}\left[\exp\left(\theta\|e_{k-1}\|^2\right)\middle|\,\mathcal{F}_{k-2}\right]\right]\\
&\overset{(iii)}{\le} \mathbf{E}\left[\exp\left[\theta\sum_{i=0}^{k-2}\rho^{k-i-1}\|e_i\|^2\right]\right]\bar{\gamma}_{k-1}(\theta)\\
&\overset{(iv)}{\le} \prod_{i=0}^{k-1}\bar{\gamma}_i(\theta\rho^{k-i-1}),
\end{aligned}$$

where (i) follows from the law of total expectations, i.e., $\mathbf{E}_Y[\mathbf{E}[X|Y]] = \mathbf{E}[X]$; (ii) follows from the observation that the sum $\exp(\theta\sum_{i=0}^{k-2}\rho^{k-1-i}\|e_i\|^2)$ is a deterministic function of $e_0,\dots,e_{k-2}$, and hence is measurable with respect to $\mathcal{F}_{k-1}$ and can be factored out of the expectation; (iii) uses Assumption 3.4; and to obtain (iv) we simply repeat the process recursively.

Thus, we now have a bound on the moment-generating function of the discounted sum of errors $\theta\sum_{i=0}^{k-1}\rho^{k-1-i}\|e_i\|^2$, and we can continue by using the same approach used to derive (3.2). The remainder of the proof then follows that of Theorem 3.1, except that the sums over $i = 0,\dots,k$ are replaced by products over that same range.

□

In an application where both γ_k and $\bar{\gamma}_k$ are available, it is not true in general that either of the bounds obtained in Theorems 3.1 and 3.5 are tighter than the other. When only a bound $\bar{\gamma}_k$ that satisfies Assumption 3.4 is available, however, (which is the case in the sampling application of section 6) we could leverage (3.4) and apply Theorem 3.1 to obtain a valid bound in terms of $\bar{\gamma}_k$ by simply substituting it for γ_k. However, as shown below, in this case it is better to apply Theorem 3.5 because it yields a uniformly better bound:

$$\Pr\left(\pi_k-\rho^k\pi_k\ge\epsilon\right)\le\inf_{\theta>0}\left\{\exp\left(-2\theta L\epsilon+\sum_{i=0}^{k-1}\log\bar{\gamma}_i\left(\theta\rho^{k-1-i}\right)\right)\right\},\tag{3.5}$$

while Theorem 3.1 (with γ_k replaced by $\bar{\gamma}_k$) gives us

$$\Pr\left(\pi_k-\rho^k\pi_0\ge\epsilon\right)\le\inf_{\theta>0}\left\{\exp\left(-2\theta L\epsilon+\log\left[\frac{1}{R_k}\sum_{i=0}^{k-1}\rho^{k-1-i}\bar{\gamma}_i(\theta R_k)\right]\right)\right\},\tag{3.6}$$

where we rescale θ by R_k. A direct comparison of the two bounds show that they only differ by one term:

$$\log\left[\frac{1}{R_k}\sum_{i=0}^{k-1}\rho^{k-i-1}\bar{\gamma}_i(\theta R_k)\right]\qquad\text{vs.}\qquad\sum_{i=0}^{k-1}\log\bar{\gamma}_i(\theta\rho^{k-1-i}).$$

Because $R_k = \sum_{i=0}^{k} \rho^{k-i-1}$, the term in the log on the left is a convex combination of the functions $\bar\gamma_i$. Therefore,

$$\begin{aligned}\log\left[\frac{1}{R_k}\sum_{i=0}^{k-1}\rho^{k-i-1}\bar\gamma_i(\theta R_k)\right] &\overset{(i)}{\geq} \sum_{i=0}^{k-1}\frac{\rho^{k-1-i}}{R_k}\log\bar\gamma_i(\theta R_k)\\ &\overset{(ii)}{\geq} \sum_{i=0}^{k-1}\log\bar\gamma_i(\theta R_k\,\rho^{k-1-i}/R_k)\\ &= \sum_{i=0}^{k-1}\log\bar\gamma_i(\theta\rho^{k-1-i}),\end{aligned}$$

where (i) is an application of Jensen's inequality and the concavity of log, and (ii) follows from the convexity of the cumulant generating function. It is then evident that (3.5) implies (3.6).

As with Corollary 3.2, a connection can be made between our bound and the infimal convolution when $\bar\gamma$ is log-concave:

$$\log\Pr(\pi_k - \rho^k\pi_0 \geq \epsilon) \leq \left[\bigotimes_{i=0}^{k-1}[\log\bar\gamma_i(\,\cdot\,\rho^{k-i-1})]^*\right](2L\epsilon/R_k),$$

where $\otimes$ denotes the infimal convolution operator.

Example 3.6 (Gradient descent with independent Gaussian noise, part III). *As in Example 3.3, let $e_k \sim N(0,\sigma^2 I)$. Because the errors e_k are independent, $\mathbf{E}\left[\exp(\theta\|e_k\|^2)\,|\,\mathcal{F}_{k-1}\right] = \mathbf{E}\exp(\theta\|e_k\|^2) = \gamma_k(\theta)$, which satisfies Assumption 3.4 with $\bar\gamma_k(\theta) := \gamma_k(\theta)$. Apply Theorem 3.5 to obtain the bound*

$$\Pr(\pi_k - \rho^k\pi_0 \geq \epsilon) \leq \inf_{\theta>0}\left\{\exp(-2\theta L\epsilon)\cdot\prod_{i=0}^{k-1}(1-2\sigma^2\theta\rho^{k-1-i})^{-n/2}\right\}. \tag{3.7}$$

Apply Lemma A.2 to obtain

$$\Pr(\pi_k - \rho^k\pi_0 \geq \epsilon) \leq \left(\frac{2\exp(1)}{n\alpha}\cdot\frac{L\epsilon}{\sigma^2}\right)^{\frac{n\alpha}{2}}\exp\left(-\frac{L\epsilon}{\sigma^2}\right) \quad \textit{for} \quad \epsilon > \frac{n\alpha\sigma^2}{2L},$$

where $\alpha = 1-(\log\rho)^{-1}$. We simplify the bound to obtain

$$\Pr(\pi_k - \rho^k\pi_0 \geq \epsilon) = \mathcal{O}\left[\exp\left(-\delta\cdot\frac{L\epsilon}{\sigma^2}\right)\right] \quad \textit{for all} \quad \delta\in(0,1); \tag{3.8}$$

cf. (3.3).

As an aside, we note that we can easily accommodate correlated noise, i.e., $e_k \sim N(0,\Sigma^2)$ where Σ is an $n\times n$ positive definite matrix. The error $\|e_k\|^2$ then has the distribution of a sum of chi-squared random variables that are weighted according to the eigenvalues σ_j of Σ [7]

$$\|e_k\|^2 \sim \sum_{j=1}^{n}\sigma_j^2\chi_1^2,$$

and so the above tail bounds hold with $\sigma = \sigma_{\max}$.

The bounds obtained in Examples 3.3 and 3.6 illustrate the relative strengths of Theorems 3.1 and 3.5. Comparing (3.3) and (3.8), we see that the asymptotic bounds only differ by a factor of $1/R_k$. Hence, for large ϵ, the bound in Example 3.3 is uniformly weaker than the bound in Example 3.6. Note that this holds despite the relaxation (i.e., Lemma A.2) used to simply (3.7).

4. From tail bounds to moment-generating bounds.

Let $\mathcal{G}$ be a σ-algebra. In this section we show that an exponential bound

$$\Pr(X_i \geq \epsilon \,|\, \mathcal{G}) := \mathbf{E}[\mathbb{1}_{X_i \geq \epsilon} \,|\, \mathcal{G}\,] \leq \exp(-\epsilon^2/\nu) \tag{4.1}$$

on the conditional probability [8, Definition 8.11] on a sequence of univariate random variables X_i translates into a deterministic bound on the conditional moment-generating function

$$\mathbf{E}[\exp(\theta\|X\|^2) \,|\, \mathcal{G}],$$

where $X = (X_1, X_2, \dots, X_n)$ is an n-vector.

Lemma 4.1 (Bounds on moments). *If* (4.1) *holds for some* $\nu > 0$, *then*

$$\mathbf{E}[X_i^{2v} \,|\, \mathcal{G}] \leq v!\nu^v \qquad \textit{for all} \qquad v = 0, 1, 2, \dots.$$

Proof. Use the substitution $\epsilon^{2v} = \tau$ to obtain

$$\Pr\left(Y^{2v} \geq \tau \,|\, \mathcal{G}\right) \leq \exp\left(-\tau^{1/v}/\nu\right).$$

Integrate to get

$$\mathbf{E}[Y^{2v} \,|\, \mathcal{G}] = \int_0^\infty \mathbf{E}[\mathbb{1}_{Y^{2v} \geq \tau} \,|\, \mathcal{G}]\, d\tau \;\leq\; \int_0^\infty \exp(-\tau^{1/v}/\nu)\, d\tau \;=\; \Gamma(1+v)\nu^v \;=\; v!\nu^v,$$

where the first equality comes from the conditional layer-cake representation of positive random variables [16]. □

With this result, we can translate the bound (4.1) into a bound on the moment-generating function of Y^2.

Lemma 4.2 (Bound on conditional MGF). *If* (4.1) *holds for some* $\nu > 0$, *then*

$$\mathbf{E}[\exp\left(\theta Y^2\right) \,|\, \mathcal{G}] \leq \frac{1}{1-\theta\nu} \qquad \textit{for} \qquad \theta \in [0, 1/\nu).$$

Proof. Using the Taylor expansion of $\mathbf{E}[\exp\left(\theta Y^2\right) \,|\, \mathcal{G}]$,

$$\begin{aligned}
\mathbf{E}[\exp\left(\theta Y^2\right) \,|\, \mathcal{G}] &= \mathbf{E}\left[\sum_{i=0}^\infty \theta^i \frac{Y^{2i}}{i!} \,\middle|\, \mathcal{G}\right] \\
&\overset{(i)}{=} \sum_{i=0}^\infty \theta^i \frac{\mathbf{E}[Y^{2i} \,|\, \mathcal{G}]}{i!} \\
&\overset{(ii)}{\leq} \sum_{i=0}^\infty \theta^i \frac{i!\nu^i}{i!} = \sum_{i=0}^\infty (\theta\nu)^i = \frac{1}{1-\theta\nu}.
\end{aligned}$$

Equality (i) is obtained via the conditional monotone convergence theorem [17, Theorem 9.7e], which allows us to exchange limits and conditional expectations; inequality (ii) is obtained using Lemma 4.1. □

We now generalize this last result to the case where X is a random n-vector.

THEOREM 4.3 (From tail bounds to moment-generating bounds). *Let X be a random n-vector for which the tail bound* (4.1) *holds for each i for some $\nu > 0$. Then*

$$\mathbf{E}[\exp(\theta\|X\|^2)\,|\,\mathcal{G}] \leq \frac{1}{1-\theta\nu n} \qquad \textit{for} \qquad \theta \in [0, 1/\nu n).$$

Proof. From Lemma 4.2,

$$\mathbf{E}\left[\exp\left(\theta n[X]_i^2\right)\,|\,\mathcal{G}\right] \leq \frac{1}{1-\theta n\nu}. \tag{4.2}$$

The following inequalities hold:

$$\begin{aligned}
\mathbf{E}\left[\exp\left(\theta\|X\|^2\right)\middle|\,\mathcal{G}\right] &= \mathbf{E}\left[\exp\left(\theta\sum_{i=1}^n [X]_i^2\right)\middle|\,\mathcal{G}\right] \\
&= \mathbf{E}\left[\exp\left(\theta n\sum_{i=1}^n \frac{1}{n}[X]_i^2\right)\middle|\,\mathcal{G}\right] \\
&\overset{(i)}{\leq} \mathbf{E}\left[\sum_{i=1}^n \frac{1}{n}\exp\left(\theta n[X]_i^2\right)\middle|\,\mathcal{G}\right] \\
&= \sum_{i=1}^n \frac{1}{n}\mathbf{E}\left[\exp\left(\theta n[X]_i^2\right)\middle|\,\mathcal{G}\right] \overset{(ii)}{\leq} \frac{1}{1-\theta n\nu},
\end{aligned}$$

where (i) follows from Jensen's inequality and (ii) follows from (4.2). □

5. Convergence rates for linearly decreasing errors.

Section 3 describes tail bounds for (1.7) in terms of any available bound on the moment-generating function of the error e_k. A goal of this section is to show that an exponential tail bound on the error translates to an exponential tail bound on (1.7). Thus we consider the case where the tails on each component of e_k are exponentially decreasing (cf. Hypothesis 5.1.B below). We also consider two additional conditions on the error sequence, which illustrate the exponential tail bound's relative strength in the following hierarchy of assumptions. In section 6 we show how various sampling strategies satisfy these conditions.

HYPOTHESIS 5.1 (Uniform bounds). *Suppose that*

$$U_k \leq \mathcal{O}(1)\beta^k \tag{5.1}$$

for all k and for some $\beta < 1$. Assume that the following hold.

A. [Variance] $\mathbf{E}\,\|e_k\|^2 \le U_k$;

B. [Exponential Tail] $\Pr\left([e_k]_i \ge \epsilon \,|\, \mathcal{F}_{k-1}\right) \le \exp\left(-\epsilon^2/U_k\right)$;

C. [Norm] $\|e_k\|^2 \le U_k$.

These conditions are ordered in increasing strength: if (C) holds, then (B) holds by Hoeffding's inequality (Theorem A.5), and if (B) holds, then (A) holds because the exponential bound implies a bound on the second moment, i.e.,

$$\mathbf{E}\left[[e_k]_i^2 \,|\, \mathcal{F}_{k-1}\right] = \int_0^\infty \Pr([e_k]_i^2 \ge \epsilon \,|\, \mathcal{F}_{k-1})\, d\epsilon \le \int_0^\infty \exp\left(-\epsilon^2/U_k\right) d\epsilon < \infty.$$

5.1. Expectation-based and deterministic bounds. Although our main goal is to derive tail bounds, it is useful to compare these tail bounds against the expectation-based and deterministic bounds derived in Friedlander and Schmidt [5, Theorem 3.3]. We give here a reformulation of these results, which rely on parts A and C of Hypothesis 5.1.

THEOREM 5.2 (Bound in expectation). *Suppose that Hypothesis 5.1.A holds. Then*

$$\mathbf{E}\,\pi_k - \rho^k \pi_0 = \mathcal{O}([\max\{\beta,\rho\} + \zeta]^k) \quad \textit{for all} \quad \zeta > 0.$$

If $\rho \ne \beta$, then the bound holds with $\zeta = 0$.

Proof. The assumptions give a bound on $\mathbf{E}\,\|e_k\|^2$ in terms of β:

$$\mathbf{E}\,\|e_k\|^2 \le U_k \le \mathcal{O}(1)\beta^k \le 2\tau\beta^k$$

for some τ. For $\beta \le \rho$, it follows from Lemma 2.1 that

$$\mathbf{E}\,\pi_k - \rho^k \pi_0 \le \frac{1}{2L}\sum_{i=0}^{k-1} \rho^{k-i-1}\,\mathbf{E}\,\|e_i\|^2 \le \frac{\tau\rho^{k-1}}{L}\sum_{i=0}^{k-1}(\beta/\rho)^i \le \frac{\tau}{L}\rho^{k-1}k. \tag{5.2}$$

Similarly, for $\beta > \rho$,

$$\mathbf{E}\,\pi_k - \rho^k \pi_0 \le \frac{\tau\beta^{k-1}}{L}\sum_{i=0}^{k-1}(\rho/\beta)^i \le \frac{\tau}{L}\beta^{k-1}k. \tag{5.3}$$

We summarize these last two bounds in the single expression

$$\mathbf{E}\,\pi_k - \rho^k \pi_0 \le \frac{\tau}{L}\max\{\beta,\rho\}^{k-1} k = \mathcal{O}([\max\{\beta,\rho\} + \zeta]^k)$$

for all $\zeta > 0$.

If $\beta \ne \rho$, then it follows from the second inequality in (5.2) and the first inequality in (5.3), and the summation formula for geometric series, that

$$\mathbf{E}\,\pi_k - \rho^k \pi_0 \le \frac{\tau}{L}\max\{\beta,\rho\}^{k-1}\frac{1}{|\beta-\rho|} = \mathcal{O}(\max\{\beta,\rho\}^k). \tag{5.4}$$

□

The following result is identical to Theorem 5.2, except that π_k is deterministic.

THEOREM 5.3 (Deterministic bound). *Suppose that Hypothesis 5.1.C holds. Then*

$$\pi_k - \rho^k \pi_0 = \mathcal{O}([\max\{\beta,\rho\} + \zeta]^k) \quad \textit{for all} \quad \zeta > 0.$$

If $\rho \neq \beta$, then the bound holds with $\zeta = 0$.

5.2. Tail bounds. The next result gives exponential tail bounds in terms the iteration k, and the deviation ϵ from the linear rate of deterministic steepest descent.

THEOREM 5.4 (Tail bounds). *Suppose that Hypothesis 5.1.B holds. Then*

$$\Pr(\pi_k - \rho^k \pi_0 \geq \epsilon) = \mathcal{O}\left(\exp\left[-\frac{\epsilon}{\max\{\beta,\rho\}^k} \cdot \zeta\right]\right) \quad \textit{for some} \quad \zeta > 0. \tag{5.5}$$

Proof. From Theorem 4.3 the conditioned moment-generating function of $\|e_k\|^2$ is bounded:

$$\mathbf{E}[\exp(\theta\|e_k\|^2) \,|\, \mathcal{F}_{k-1}] \leq \frac{1}{1-\theta U_k n} \quad \text{for} \quad \theta \in [0, 1/U_k n)\,.$$

We can now use Theorem 3.5, where we identify $\bar\gamma$ with this bound (and define $\bar\gamma(\theta) = \infty$ outside of the required interval), to obtain the tail bound

$$\Pr(\pi_k - \rho^k \pi_0 \geq \epsilon) \leq \inf_{\theta \in [0,1/\sigma)} \left\{ \frac{\exp(-2\theta L \epsilon)}{\prod_{i=0}^{k-1} \left(1 - \theta U_k n\right) \rho^{k-i-1})} \right\},$$

where $\sigma := \max_k \rho^{k-i-1} U_k n$. By (5.1), there exists some constant τ independent of β, ρ, and L such that

$$nU_k \leq 2\tau\beta^k. \tag{5.6}$$

Define $\alpha = \max\{\beta, \rho\}$. Now

$$\begin{aligned} \Pr\left(\pi_k - \rho^k \pi_0 \geq \epsilon\right) &\overset{(i)}{\leq} \inf_{\theta \in [0,1/\sigma)} \left\{ \frac{\exp(-2\theta\epsilon L)}{\prod_{i=0}^{k-1} (1 - 2\theta\tau\beta^i \rho^{k-i-1})} \right\} \\ &\overset{(ii)}{=} \inf_{\theta \in [0,1/\sigma)} \left\{ \frac{\exp(-2\theta\epsilon L)}{\prod_{i=0}^{k-1} (1 - 2\theta\tau\alpha^{k-1} \min\{\beta/\rho, \rho/\beta\}^i)} \right\}. \end{aligned} \tag{5.7}$$

where inequality (i) is obtained by substituting in (5.6), and equality (ii) follows from the definition of α.

Define $\hat\gamma = 1 + 1/\log(1/\min\{\beta/\rho, \rho/\beta\}) = 1 + 1/|\log\beta - \log\rho|$, and apply Lemma A.4 to (5.7) to obtain

$$\Pr\left(\pi_k - \rho^k \pi_0 \geq \epsilon\right) \leq \left(\frac{\exp(1)}{\hat\gamma} \cdot \frac{L\epsilon}{\tau\alpha^{k-1}}\right)^{\hat\gamma} \exp\left(-\frac{L\epsilon}{\tau\alpha^{k-1}}\right), \quad \epsilon \geq \hat\gamma\alpha^{k-1}\tau/L. \tag{5.8}$$

Next, note that $\min\{\beta/\rho, \rho/\beta\} \leq 1$, and so from (5.7),

$$\begin{aligned}\Pr\left(\pi_k - \rho^k \pi_0 \geq \epsilon\right) &\leq \inf_{\theta\in[0,1/\sigma)} \left\{ \frac{\exp(-2\theta\epsilon L)}{(1-2\theta\tau\alpha^{k-1}L)^k} \right\} \\ &\overset{(i)}{\leq} \left(\frac{\exp(1)}{k}\cdot\frac{L\epsilon}{\tau\alpha^{k-1}}\right)^k \exp\left(-\frac{L\epsilon}{\tau\alpha^{k-1}}\right), \quad \epsilon \geq k\alpha^{k-1}\tau/L, \quad (5.9)\end{aligned}$$

where inequality (i) follows from Lemma A.4. Inequalities (5.8) and (5.9) can be expressed together as

$$\Pr(\pi_k - \rho^k\pi_0 \geq \epsilon) \leq \left(\frac{\exp(1)}{\gamma_k}\cdot\frac{L\epsilon}{\tau\alpha^{k-1}}\right)^{\gamma_k} \exp\left(-\frac{L\epsilon}{\tau\alpha^{k-1}}\right), \quad \epsilon \geq \gamma_k\alpha^{k-1}\tau/L, \tag{5.10}$$

where $\gamma_k = \min\{1 + 1/|\log\beta - \log\rho|, k\}$.

As $\epsilon \to \infty$,

$$\Pr(\pi_k - \rho^k\pi_0 \geq \epsilon) \leq \mathcal{O}\left[\exp\left(-\delta\cdot\frac{L\epsilon}{\alpha^{k-1}}\right)\right],$$

for some positive δ independent of L and α. Also, as $k \to \infty$,

In order to further simplify this bound, take the logarithm of both sides:

$$\log\Pr(\pi_k - \rho^k\pi_0 \geq \epsilon) \leq \gamma\log\left(\frac{\epsilon}{\gamma\tau\alpha^{k-1}}\right) + \gamma - \frac{\epsilon}{\tau\alpha^{k-1}} = \mathcal{O}\left(-\frac{\epsilon}{\alpha^{k-1}}\right),$$

which implies (5.10), as required. □

Corollary 5.5. *[Overwhelming tail bounds] Suppose that* (5.5)*. For k fixed, for all $A > 0$ there exists C_A such that*

$$\Pr\left(\pi_k - \rho^k\pi_0 \geq \epsilon\right) \leq C_A\epsilon^{-A}.$$

Furthermore, for ϵ fixed, there exists C_A such that for all $A > 0$,

$$\Pr\left(\pi_k - \rho^k\pi_0 \geq \epsilon\right) \leq C_A A^{-k}$$

Proof. The right-hand side of (5.10) can be equivalently expressed in two ways as

$$\left(\frac{\exp(1)}{\gamma_k}\cdot\frac{L\epsilon}{\tau\alpha^{k-1}}\right)^{\gamma_k}\exp\left(-\frac{L\epsilon}{\tau\alpha^{k-1}}\right) = \begin{cases}\mathcal{O}(1)\cdot\epsilon^{\gamma}\exp(-\epsilon\cdot\mathcal{O}(1)) \\ \exp\left(f(k)\right)\exp\left(-\exp\left(g(k)\right)\right),\end{cases}$$

where $f(k) = [\gamma\log\left(L\epsilon\alpha/\gamma\tau\right) + \gamma] - k[\gamma\log\alpha]$ and $g(k) = \log L\epsilon\alpha - k\log\alpha$. The result then follows from Lemma A.1. □

6. Stochastic and sample average approximations. The results of section 5 are agnostic to the source of the gradient errors that are made at each iteration. We translate these generic results into a sampling policies that yields a linear convergence rate, both in expectation and with overwhelming probability.

Theorem 6.1 (Stochastic-approximation convergence rates). *Consider the stochastic-approximation algorithm described by* (1.1) *and* (1.2) *where*

$$\frac{1}{m_k} \leq \mathcal{O}(1)\beta^k$$

for all k for some $\beta < 1$. Then the following hold.

1. *[Expectation bound] If the variance of the error is bounded, i.e.,*

$$\sup_x \mathbf{E} \left\| \nabla f(x) - \nabla F(x,Z) \right\|^2 < \infty,$$

then

$$\mathbf{E}\, \pi_k - \rho^k \pi_0 = \mathcal{O}([\max\{\beta,\rho\} + \zeta]^k) \quad \textit{for all} \quad \zeta > 0.$$

If $\rho \neq \beta$, then the bound holds with $\zeta = 0$.

2. *[Tail bound] If the diameter of the error is bounded, i.e.,*

$$\sup_x \left\{ \sup_{z\in\Omega} [\nabla F(x,z)]_i - \inf_{z\in\Omega} [\nabla F(x,z)]_i \right\} < \infty,$$

for all $i = 1, \dots, n$, then

$$\Pr(\pi_k - \rho^k \pi_0 \geq \epsilon) = \mathcal{O}\left(\exp\left[-\frac{\epsilon}{\max\{\beta,\rho\}^k} \cdot \zeta \right] \right) \quad \textit{for some} \quad \zeta > 0.$$

Proof.

Part 1 (Expectation Bound). Because the random variables $Z_1, \dots, Z_{m_k}$ are independent, the expected sample error is equal to the sample average. Thus,

$$\begin{aligned} \mathbf{E} \left\| e_k \right\|^2 &= \mathbf{E} \left\| \nabla f(x_k) - \nabla F(x_k, Z_i) \right\|^2 / m_k \\ &\leq \sup_x \mathbf{E} \left\| \nabla f(x) - \nabla F(x,Z) \right\|^2 / m_k \leq \mathcal{O}(1)\beta^k, \end{aligned}$$

therefore satisfying Hypothesis 5.1.A and thus the hypothesis of Theorem 5.2.

Part 2 (Tail Bound). This follows from Hoeffding's Inequality; see Theorem A.5. Thus we satisfy Hypothesis 5.1.B and therefore the hypothesis of Theorem 5.4. □

Theorem 6.2 (Sample average gradient convergence rates). *Consider the stochastic-approximation algorithm described by* (1.1) *and* (1.4) *where*

$$\frac{1}{m_k}\left(1 - \frac{m_k - 1}{M}\right) \leq \mathcal{O}(1)\beta^k \tag{6.1}$$

for all k for some $\beta < 1$. Then the following hold.

1. [Expectation bound] If the population variance is bounded, i.e.,

$$\sup_x \frac{1}{M-1}\sum_{i=1}^{M}\|f(x)-f_i(x)\|^2 < \infty,$$

then

$$\mathbf{E}\,\pi_k - \rho^k\pi_0 = \mathcal{O}([\max\{\beta,\rho\}+\zeta]^k) \quad \textit{for all} \quad \zeta > 0.$$

If $\rho \neq \beta$, then the bound holds with $\zeta = 0$.

2. [Tail bound] If the population diameter is bounded, i.e.,

$$\sup_x \left\{\max_j[\nabla f_j(x)]_i - \min_j[\nabla f_j(x)]_i\right\} < \infty,$$

for all $i = 1,\dots,n$, then

$$\Pr(\pi_k - \rho^k\pi_0 \geq \epsilon) = \mathcal{O}\left(\exp\left[-\frac{\epsilon}{\max\{\beta,\rho\}^k}\cdot\zeta\right]\right) \quad \textit{for some} \quad \zeta > 0.$$

3. [Deterministic bound] If the diameter of the error is bounded, i.e.,

$$\sup_x \|f_i(x)\|^2 < \infty$$

for all $i = 1,\dots,n$, then

$$\pi_k - \rho^k\pi_0 = \mathcal{O}([\max\{\beta,\rho\}+\zeta]^k) \quad \textit{for all} \quad \zeta > 0.$$

If $\rho \neq \beta$, then the bound holds with $\zeta = 0$.

Proof.
Part 1 (Expectation Bound). Let

$$S(x) := \frac{1}{M-1}\sum_{i=1}^{M}\|f(x)-f_i(x)\|^2.$$

Then from Friedlander and Schmidt [5, §3.2],

$$\mathbf{E}\,\|e_k\|^2 = \left(1-\frac{m_k}{M}\right)\frac{S(x_k)}{m_k} \leq \left(1-\frac{m_k-1}{M}\right)\frac{\sup_x S(x)}{m_k} \leq \mathcal{O}(1)\beta^k,$$

therefore satisfying Hypothesis 5.1.A and thus the hypothesis of Theorem 5.2.

Part 2 (Tail Bound). This follows from Serfling's Inequality; see Theorem A.6. Thus we satisfy Hypothesis 5.1.B and therefore the hypothesis of Theorem 5.4.

Part 3 (Deterministic Bound). Refer to Friedlander and Schmidt [5, §3.1]. □

The asymptotic notation in the theorem statements helps us simplify the results, however non asymptotic bounds are available explicitly within the proofs. Figure 6.1 illustrates the non asymptotic bounds (5.4) and (5.10) that correspond to parts 1 and 2 of Theorem 6.2; the deterministic bounds follow from Friedlander and Schmidt [5, Theorem 3.1].

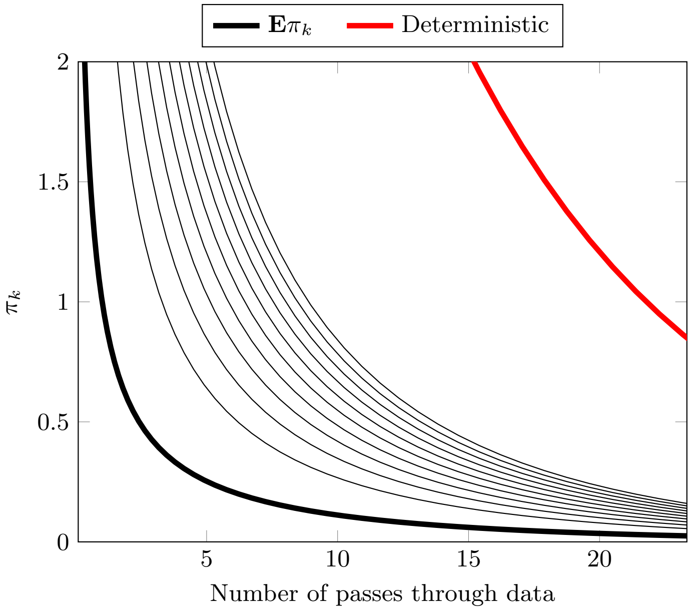

Fig. 6.1. *An illustration of the bounds derived in Theorem 6.2; this figure plots the nonasymptotic bound shown in* (5.10)*. The thick black line (bottom left) shows the bound in expectation (see Part 1 of Theorem 6.2). For comparison, the thick red line (top right) shows the deterministic bound on the distance to the solution (see [5, Theorem 3.1]). The thin lines in between give the bounds on* $\pi_k - \rho^k \pi_0$ *that correspond to probabilities* 10^{-i} *for* $i = 10, 20, \ldots, 100$*. Assume* $M = 300$, $\beta = 0.9$, *and* $\rho = 0.9$.

7. Numerical experiments.

Figure 7.1 shows the results of a Monte Carlo simulation on a logistic regression problem, where

$$f_i(x) = \log(1 + \exp[-b_i \langle a_i, x\rangle]),$$

$a_i \in \mathbb{R}^n$ is a vector of input features, and $b_i \in \{-1, 1\}$ is the corresponding observation. For this problem, we generate a dataset with $M = 100$ pairs (a_i, b_i) of random points. Algorithm (1.1) and (1.4), where the sample size satisfies (6.1) with $\beta \approx .91$, is run 10K times on this fixed dataset. The starting point between runs is the same, and the only difference is the randomness of the sampling. Figure 7.1 summarizes the results of this experiment. As expected, the sample paths are concentrated tightly around the mean. Furthermore, the probability of deviating from the mean decays doubly-exponentially (cf. 6.2), as evidenced by the linear tail shown in the bottom panel.

A. Auxiliary results.

Lemma A.1. *Suppose* $f(x) = \mathcal{O}(x^{\mathcal{O}(1)}) \exp(-\mathcal{O}(x^{\mathcal{O}(1)}))$*. Then for all* $A > 0$ *there exists a constant* C_A *that depends only on* A *such that*

$$f(x) \le C_A \epsilon^{-A}. \tag{A.1}$$

Suppose $f(x) = \exp(\mathcal{O}(x^{\mathcal{O}(1)})) \exp(-\exp(\mathcal{O}(x^{\mathcal{O}(1)})))$*. Then for all* $A > 0$ *there exists* C_A *such that*

$$f(x) \le C_A A^{-x} \tag{A.2}$$

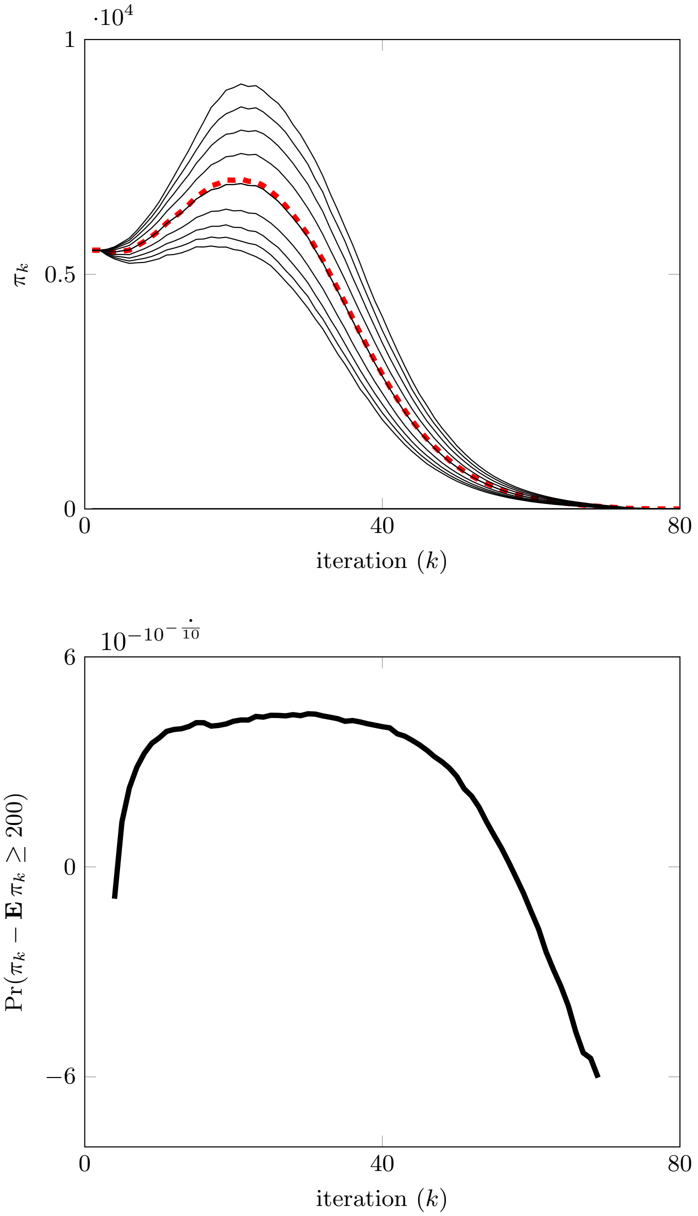

FIG. 7.1. *Top panel: distance to solution for quantiles* $1 - 0.5^j$ *and* 0.5^j, $j = -5 : 5$. *Bottom panel: probability of the deviation from expected value against a log-log y-axis, which exhibits the tail that converges with a doubly-exponential tail.*

Proof. The statement follows by taking the logarithms on both sides of (A.1) and (A.2). □

LEMMA A.2. *For $y \in (0,1)$ and $x \in [0,1]$,*

$$(1-x)^{1-1/\log y} \leq \prod_{i=0}^{\infty}(1-xy^i) \tag{A.3}$$

Proof. To prove the lower bound, we use the following fact:

$$\ln(1-x) \geq -\frac{x}{1-x} \quad \text{for all} \quad x \in [0,1).$$

Therefore,

$$\begin{aligned}\prod_{i=1}^{\infty}(1-xy^i) &= \exp\left(\sum_{i=1}^{\infty}\log\left(1-xy^i\right)\right)\\ &\geq \exp\left(\sum_{i=1}^{\infty}-\frac{y^i}{1/x-y^i}\right)\\ &\geq \exp\left(-\int_0^{\infty}\frac{y^i}{1/x-y^i}di\right)\\ &= \exp\left(-\frac{\log(1-x)}{\log(y)}\right)\\ &\geq (1-x)^{-1/\log y}.\end{aligned}$$

Thus,

$$\prod_{i=0}^{\infty}(1-xy^i) = (1-x)\prod_{i=1}^{\infty}(1-xy^i) \geq (1-x)^{1-1/\log y},$$

as required. □

LEMMA A.3. *For $y \in (0,1)$ and $x \in [0,1]$,*

$$\exp\left(-\frac{\log(1-x/y)-\log(1-xy^{N+1})}{\log(y)}\right) \leq \prod_{i=0}^{N}(1-xy^i).$$

Proof. Similar to the proof of the previous inequality

$$\begin{aligned}\prod_{i=1}^{N}(1-xy^i) &= \exp\left(\sum_{i=1}^{N}\log\left(1-xy^i\right)\right)\\ &\geq \exp\left(\sum_{i=1}^{N}-\frac{xy^i}{1-xy^i}\right)\\ &\geq \exp\left(-\int_0^{N}\frac{xy^i}{1-xy^i}di\right)\\ &\geq \exp\left(-\frac{\log(1-x)-\log\left(1-xy^N\right)}{\log(y)}\right).\end{aligned}$$

Thus,

$$\prod_{i=0}^{N}(1-xy^i) = \prod_{i=1}^{N+1}(1-(x/y)y^i)$$
$$\geq \exp\left(-\frac{\log(1-x/y)-\log(1-xy^{N+1})}{\log(y)}\right),$$

as required. □

Lemma A.4. *Let $k>0$, $\mu>0$, and $\epsilon>0$. Then for $y\in(0,1)$ and $x\in(0,1]$,*

$$\inf_{\theta>0}\left\{\exp(-\theta\epsilon\nu)\prod_{i=0}^{N-1}\left(1-\theta xy^i\right)^{-k}\right\} \leq \left(\frac{\exp(1)}{\alpha}\cdot\frac{\epsilon\nu}{x}\right)^{\alpha}\exp\left(-\frac{\epsilon\nu}{x}\right),$$

where $\alpha = \frac{1}{k}\left(\frac{1}{\log(1/y)}+1\right)$.

Proof. By inverting both sides of (A.3) we obtain the following inequality

$$\prod_{i=0}^{\infty}(1-xy^i)^{-k} \leq \exp\left(-\log(1-x)\left[\frac{1}{\log(1/y)}+1\right]\right). \tag{A.4}$$

Therefore, for $\epsilon\geq\alpha x/v$,

$$\inf_{\theta>0}\left\{\exp(-\theta\epsilon\nu)\prod_{i=0}^{N-1}(1-\theta xy^i)^{-k}\right\}$$
$$\leq \inf_{\theta>0}\left\{\exp(-\theta\epsilon\nu)\prod_{i=0}^{\infty}(1-\theta xy^i)^{-k}\right\}$$
$$\overset{(i)}{\leq} \inf_{\theta>0}\left\{\exp\left(-\frac{1}{k}\left[\frac{1}{\log(1/y)}+1\right]\log(1-\theta x)-\theta v\epsilon\right)\right\}$$
$$= \inf_{\theta>0}\left\{\exp\left(-\alpha\log(1-\theta x)-\theta\epsilon\nu\right)\right\}$$
$$\overset{(ii)}{=} \exp\left(-\alpha\log\left(1-\left(\frac{1}{x}-\frac{\alpha}{v\epsilon}\right)x\right)-\left(\frac{1}{x}-\frac{\alpha}{v\epsilon}\right)v\epsilon\right)$$
$$= \left(\frac{\exp(1)}{\alpha}\cdot\frac{\epsilon\nu}{x}\right)^{\alpha}\exp\left(-\frac{\epsilon\nu}{x}\right),$$

where (i) comes from (A.4); (ii) uses the substitution $\theta = 1/x-\alpha/v\epsilon$, which can be shown to be the optimal choice of θ. Because $\theta>0$, $\epsilon>\alpha x/v$. □

For the following theorems, we define the sample average

$$S_m := \frac{1}{m}\sum_{i}^{m}X_i$$

for a sequence of random variables $\{X_1,\ldots,X_m\}$.

Theorem A.5 (Hoeffding [6, Theorem 2]). *Consider independent random variables $\{X_1,\ldots,X_m\}$, $X_i:\Omega\to\Re$. If the random variables are bounded, i.e.,*

$$d := \sup_{\omega\in\Omega}X_i(\omega) - \inf_{\omega\in\Omega}X_i(\omega)$$

is finite, then

$$\Pr\left(S_m - \mathbf{E}\, S_m \geq \epsilon\right) \leq \exp\left(-\epsilon^2/\eta_m\right) \qquad \textit{where} \qquad \eta_m = \frac{d^2}{2m}.$$

THEOREM A.6 (Serfling [14, Corollary 1.1]). *Let $x_1, \dots, x_M$ be a population, $\{X_1, \dots, X_m\}$ be samples drawn without replacement from the population, and let*

$$d := \max_i x_i - \min_i x_i.$$

Then

$$\Pr\left(S_m - \mathbf{E}\, S_m \geq \epsilon\right) \leq \exp\left(-\epsilon^2/\eta_m\right) \qquad \textit{where} \qquad \eta_m = \frac{d^2}{2m}\left(1 - \frac{m-1}{M}\right).$$

Because η_m is strictly decreasing in m, the Serfling bound is uniformly better than the Hoeffding bound. Note that the Serfling bound is not tight: in particular, when $M = m$ (i.e., $S_m = \mathbf{E}\, S_m$), the bound is not zero (except for degenerate population).

Acknowledgements. We are grateful to Bill Aiello for valuable comments, and for pointing us to Hoeffding [6], which helped to catalyze this paper. We are also indebted to Jason Swanson for generously answering our questions regarding conditional expectations, and for his excellent lecture notes [16] on the topic.

REFERENCES

[1] K. Azuma, *Weighted sums of certain dependent random variables*, Tohoku Mathematical Journal, 19 (1967), pp. 357–367.

[2] D. P. Bertsekas and J. N. Tsitsiklis, *Gradient convergence in gradient methods with errors*, SIAM J. Optim., 10 (2000), pp. 627–642.

[3] R. H. Byrd, G. Chin, J. Nocedal, and Y. Wu, *Sample size selection in optimization methods for machine learning*, Math. Program., 134 (2012), pp. 127–155.

[4] F. Chung and L. Lu, *Concentration inequalities and martingale inequalities: a survey*, Internet Mathematics, 3 (2006), pp. 79–127.

[5] M. P. Friedlander and M. Schmidt, *Hybrid deterministic-stochastic methods for data fitting*, SIAM J. Sci. Comp., 34 (2012).

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

[7] J. P. Imhof, *Computing the distribution of quadratic forms in normal variables*, Biometrika, (1961), pp. 419–426.

[8] A. Klenke, *Probability theory: a comprehensive course*, Springer Verlag, London, 2008.

[9] D. Luenberger and Y. Ye, *Linear and nonlinear programming*, Springer Verlag, 2008.

[10] Z. Luo and P. Tseng, *Error bounds and convergence analysis of feasible descent methods: A general approach*, Ann. Oper. Res., 46 (1993), pp. 157–178.

[11] A. Nedic and D. Bertsekas, *Convergence rate of incremental subgradient algorithms*, Stochastic Optimization: Algorithms and Applications, (2000), pp. 263–304.

[12] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM J. Optim., 19 (2009), pp. 1574–1609.

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

[14] R. Serfling, *Probability inequalities for the sum in sampling without replacement*, Ann. Statist., 2 (1974), pp. 39–48.

[15] M. Solodov, *Incremental gradient algorithms with stepsizes bounded away from zero*, Computational Optimization and Applications, 11 (1998), pp. 23–35.

[16] J. Swanson, *Conditional expectation.* Unpublished lecture notes, http://www.swansonsite.com/W/instructional/condexp.pdf, April 2009.

[17] D. Williams, *Probability with Martingales*, Cambridge University Press, England, 8th ed., 1991.