Bias in Stochastic Approximation Cannot Be Eliminated With Averaging

Caio Kalil Lauand
Department of Electrical and Computer Engineering
University of Florida
Gainesville, USA
caio.kalillauand@ufl.edu

Sean Meyn
Department of Electrical and Computer Engineering
University of Florida
Gainesville, USA
meyn@ece.ufl.edu

2022 58th Annual Allerton Conference on Communication, Control, and Computing (Allerton) | 979-8-3503-9998-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ALLERTON49937.2022.9929369

This paper concerns bias and asymptotic statistics for stochastic approximation (SA) driven by Markovian noise. This extended abstract is organized into three parts: 1. Background, 2. Asymptotic statistics with Markovian noise, 3. Quasi stochastic approximation.

1. Background: The goal of SA is to solve root finding problems of the form $\bar{f}(\theta^*) = 0$, where the function $\bar{f}\colon \mathbb{R}^d \to \mathbb{R}^d$ is defined as an expectation $\bar{f}(\theta) := \mathsf{E}[f(\theta, \Phi)]$. Robbins and Monro introduced in [13] the celebrated SA algorithm, expressed as the d-dimensional recursion,

$$\theta_{n+1} = \theta_n + \alpha_{n+1} f(\theta_n, \Phi_{n+1}), \quad n \geq 0 \tag{1}$$

with initial condition $\theta_0 \in \mathbb{R}^d$, $\{\alpha_n\}$ a non-negative step-size sequence and $\mathbf{\Phi} := \{\Phi_n\}$ a sequence of vector-valued random variables such that $\Phi_n \xrightarrow{\text{d}} \Phi$ as $n \to \infty$.

The SA algorithm is motivated by ordinary differential equation theory, and this theory plays a large part in the analysis and design of stochastic algorithms. In short, the SA recursion is viewed as a noisy Euler approximation of the ODE $\frac{d}{dt} x_t = \bar{f}(x_t)$, designed to be globally asymptotically stable with unique equilibrium θ^*. This and minor additional assumptions imply that the estimates $\{\theta_n\}$ converge to θ^* with probability one, from each initial condition. Interest in machine learning has spurred recent growth in theory and application of SA [1], [10], [15].

A typical choice of step-size in theory is

$$\alpha_n = g(1 + n/n_e)^{-\rho}, \tag{2}$$

in which $g, n_e > 0$ and $\frac{1}{2} < \rho \leq 1$; the constraint on ρ is imposed so that the step-size is square summable, but $\sum_n \alpha_n = \infty$. Here there is often a break between "practitioners" and "theoreticians". For reasons that are not clear to the authors, many users of SA advocate a fixed step-size, in which $\alpha_n \equiv \alpha > 0$. There is no hope for convergence in this case, but bounds on bias and variance can be obtained once boundedness of the recursion is established. One approach is to adopt a Markovian framework: if $\mathbf{\Phi}$ is a Markov chain, then the pair process $\mathbf{\Psi} := \{\Psi_n = (\theta_n, \Phi_n)\}$ is also Markovian. Conditions for geometric ergodicity of the joint process are described in [4], where it is shown that the steady-state variance of $\{\theta_n\}$ is typically of order α.

Financial support from ARO award W911NF2010055 and National Science Foundation award EPCN 1935389 is gratefully acknowledged.

Recent work provides a bridge between theory and practice, via the averaging technique of Polyak and Ruppert. For $N \geq 1$ and a scalar $\delta_0 \in (0, 1)$, the *PR-estimate* is defined by

$$\theta_N^{\mathsf{PR}} := \frac{1}{N - N_0} \sum_{n=N_0+1}^{N} \theta_n, \quad N_0 = \lfloor \delta_0 N \rfloor \tag{3}$$

It is well known that the resulting estimates are consistent and asymptotically efficient under special conditions on the step-size sequence and $\mathbf{\Phi}$:

$$\lim_{N\to\infty} \theta_N^{\mathsf{PR}} = \theta^* \qquad \text{A. Unbiased} \tag{4a}$$

$$\lim_{N\to\infty} (N - N_0)\text{Cov}\,(\theta_N^{\mathsf{PR}}) = \Sigma^{\mathsf{PR}} \qquad \text{A. Efficient} \tag{4b}$$

The efficiency is in a strong sense: if $\Sigma := \lim_{N\to\infty} N\text{Cov}\,(\theta'_N)$ is the asymptotic covariance obtained with another consistent SA recursion, then $\Sigma \geq \Sigma^{\mathsf{PR}}$, in the sense that $\Sigma - \Sigma^{\mathsf{PR}}$ is positive semi-definite [12], [14]. Until recently, theory was developed only for algorithms with vanishing step-size. For the special case (2), asymptotic efficiency requires $\frac{1}{2} < \rho < 1$ (the value $\rho = 1$ is excluded).

This theory was generalized to SA recursions with fixed step-size in [5], [11], along with finer finite-n bounds on estimation error. These results come with a large price: it is assumed that $\mathbf{\Phi}$ is i.i.d., and also that f is *linear* in the parameter, of the form $f(\theta_n, \Phi_{n+1}) = A_{n+1}\theta_n - b_{n+1}$ (so that $\Phi_n := (A_n; b_n)$). Subject to second order moment bounds on $\mathbf{\Phi}$, and a density assumption to obtain a form of irreducibility, it follows from [4] that there is $\alpha_0 > 0$ such that the joint process $\mathbf{\Psi}$ is geometrically ergodic for any $\alpha \in (0, \alpha_0)$ for the nonlinear SA recursion.

For linear SA the irreducibility assumption can be relaxed, since it is not difficult to instead consider a topological form of coupling: let $\{\theta_n^i : i = 1, 2; n \geq 0\}$ denote two parameter estimate sequences with different initial conditions. The difference evolves as a linear system without additive disturbance:

$$\mathcal{E}_n := \theta_n^2 - \theta_n^1 = \prod_{k=1}^{n} [I - \alpha A_k]\mathcal{E}_0$$

If $A^* = \mathsf{E}[A_k]$ is Hurwitz, then the right hand side converges to 0 almost surely and in L_2, for sufficiently small $\alpha > 0$. This implies geometric ergodicity in a topological sense, and asymptotic consistency: $\theta_n \to \theta^* = [A^*]^{-1}b$ a.s., from each initial condition, with $b = \mathsf{E}[b_n]$.

The proof of efficiency is obtained through a representation of the SA recursion with additive white noise,

$$\theta_{n+1} = \theta_n + \alpha_{n+1}\big(A^*[\theta_n - \theta^*] + \mathcal{W}_{n+1}\big) \tag{5}$$

in which $\{\mathcal{W}_n = \tilde{A}_n\theta_{n-1} - \tilde{b}_n\}$ is a martingale-difference sequence, with $\tilde{A}_n = A_n - A^*$ and $\tilde{b}_n = b_n - b$.

The martingale difference sequence has bounded second moment, and its covariance is convergent to the matrix $\Sigma_{\mathcal{W}} = \text{Cov}\,(\tilde{A}_k\theta^* - \tilde{b}_k)$ under the i.i.d. assumptions, and whenever $\theta_n \to \theta^*$ in L_2. These facts lead to a simple proof of efficiency of PR-averaging, where for simplicity we take $N_0 = 1$: Summing both sides of (5) from $n = 1$ to N gives for the fixed step-size recursion,

$$\theta_{N+1} = \theta_1 + \alpha\sum_{n=1}^{N}\big(A^*[\theta_n - \theta^*] + \mathcal{W}_{n+1}\big) \tag{6}$$

and after rearranging terms and dividing by N,

$$\theta_N^{\mathsf{PR}} = \theta^* + G\frac{1}{N}\Big[\frac{1}{\alpha}(\theta_{N+1} - \theta_1) + \sum_{n=1}^{N}\mathcal{W}_{n+1}\Big] \tag{7}$$

where $G = -[A^*]^{-1}$ is the stochastic Newton-Raphson matrix gain [14]. It is clear from (7) that the PR-estimates are asymptotically unbiased, and also asymptotically efficient:

$$\lim_{N\to\infty} N\text{Cov}\,(\theta_N) = G\Sigma_{\mathcal{W}}G^{\intercal} = \Sigma^{\mathsf{PR}}$$

2. Asymptotic statistics with Markovian noise: It might be hoped that the optimistic conclusions extend to the Markovian setting. One source of optimism is from the work of Metivier and Priouret [2], [9], who demonstrate that the noise can be "whitened" when the sequence $\mathbf{\Phi}$ is Markovian. This technique is based on the change of notation,

$$f(\theta_n, \Phi_{n+1}) = \bar{f}(\theta_n) + \Delta_{n+1} \tag{8}$$

Under mild conditions, the sequence $\mathbf{\Delta} := \{\Delta_n : n \geq 1\}$ can be expressed as a martingale difference sequence, plus terms that are negligible when considering second order statistics. However, all prior work is based on SA with vanishing step-size. For reasons that will soon be clear, it is not possible in general to extend these results to recursions with fixed step-size. Linearity of f does not provide much benefit, except to simplify analysis.

Suppose that $\mathbf{\Phi}$ is a geometrically ergodic Markov chain on an abstract set X (assume this is finite if you are not familiar with the general theory), with invariant measure π. The *disturbance decomposition* of $\mathbf{\Delta}$ is based on a solution to *Poisson's equation*: for each $\theta \in \mathbb{R}^d$,

$$\mathsf{E}[\hat{f}(\theta, \Phi_{n+1}) - \hat{f}(\theta, \Phi_n) \mid \Phi_n = x] = -\tilde{f}(\theta, x)\,, \quad x \in \mathsf{X}\,, \tag{9}$$

in which $\tilde{f}(\theta, x) := f(\theta, x) - \bar{f}(\theta)$.

The following goes back to the 1984 paper [9]:

$$\Delta_{n+1} = \mathcal{W}_{n+2} - \mathcal{T}_{n+2} + \mathcal{T}_{n+1} - \alpha\Upsilon_{n+2}, \tag{10a}$$

with $\mathcal{T}_{n+1} := \hat{f}(\theta_n, \Phi_{n+1})\,,$ *and*

$$\mathcal{W}_{n+2} := \hat{f}(\theta_n, \Phi_{n+2}) - \mathsf{E}[\hat{f}(\theta_n, \Phi_{n+2}) \mid \mathcal{F}_{n+1}] \tag{10b}$$

$$\Upsilon_{n+2} := -\frac{1}{\alpha}\big[\hat{f}(\theta_{n+1}, \Phi_{n+2}) - \hat{f}(\theta_n, \Phi_{n+2})\big] \tag{10c}$$

The identity (10) expresses the disturbance as the sum of the *martingale difference* $\mathcal{W}_{n+2}$, a telescoping sequence, and the final term that is small if α is small.

Substituting (8) into (1) and applying the same steps to obtain (7) gives

$$\frac{1}{N}[\theta_{N+1} - \theta_1] = \frac{\alpha}{N}\sum_{n=1}^{N}[\bar{f}(\theta_n) + \Delta_{n+1}] \tag{11a}$$

$$= \frac{\alpha}{N}\sum_{n=1}^{N}[A^*(\theta_n - \theta^*) + \epsilon_n + \Delta_{n+1}] \tag{11b}$$

in which the second equation is obtained using a first order Taylor series around θ^*, which implies the upper bound $\|\varepsilon_n\| = O(\|\theta_n - \theta^*\|^2)$.

Assumptions are required to ensure the existence of $\hat{f}$, along with L_2 bounds on the martingale difference sequence and other terms. Subject to (A2)–(A5) of [3] we can be assured that the bivariate process $\{\Psi_n = (\theta_n, \Phi_n)\}$ is Markovian, and $\{\theta_n\}$ admits a bounded fourth moment. Subject to an irreducibility condition, the joint process is geometrically ergodic with unique invariant measure ϖ. Strong Lyapunov bounds in [3] imply bounds on the solutions to Poisson's equation of interest.

Bias and inefficiency with fixed gain algorithms: The first conclusion is an extension of [4, Thm. 2.3]: there is $\alpha_0 > 0$ and $b_0 < \infty$ such that

$$\mathsf{E}_{\varpi}[\|\theta_1 - \theta^*\|^2] = \lim_{n\to\infty}\mathsf{E}[\|\theta_n - \theta^*\|^2] \leq b_0\alpha \tag{12}$$

where the limit holds for each $\theta_0 \in \mathbb{R}^d$. Hence the standard deviation is of order $\sqrt{\alpha}$ as in the i.i.d. setting.

The bias is far smaller. First, on combining (10) and (11a) we obtain a formula for the asymptotic *target bias*

$$b_{\bar{f}} := \lim_{n\to\infty}\mathsf{E}[\bar{f}(\theta_n)] = \lim_{n\to\infty}\mathsf{E}[\Upsilon_n] = \alpha\bar{\Upsilon} \tag{13}$$

A Taylor series approximation then implies that the bias itself is of the same order: Combining (13) and (11b),

$$\mathsf{E}_{\varpi}[\theta_1] - \theta^* = G\lim_{n\to\infty}\mathsf{E}[\epsilon_n + \Delta_{n+1}] = O(\alpha) \tag{14}$$

Obviously averaging cannot remove bias, but it may reduce variance. From (11b) we obtain the approximation,

$$\lim_{n\to\infty} n\text{Cov}\,(\theta_n^{\mathsf{PR}}) = \Sigma^{\mathsf{PR}} + \alpha Z + O(\alpha^{3/2}) \tag{15}$$

in which $Z_{i,j} = \langle\mathcal{W}_i, \widehat{\Upsilon}_j\rangle + \langle\mathcal{W}_j, \widehat{\Upsilon}_i\rangle - \langle\mathcal{W}_i, \Upsilon_j\rangle$, with $\widehat{\Upsilon}$ defined in analogy with $\hat{f}$. It is not yet known if the matrix Z is positive semi-definite.

Numerical example: While bias is only of order α, the magnitude of $\bar{\Upsilon}$ may be large when the Markov chain has long memory. A scalar linear recursion is used to illustrate this point:

$$\begin{aligned} f(\theta_n, \Phi_{n+1}) &= A_{n+1}\theta_n - b + \mathcal{W}_{n+1} \\ \textit{where } A_{n+1} &= -1 + \mathcal{W}_{n+1}\,, \\ \mathcal{W}_{n+1} &= \beta\mathcal{W}_n + \sqrt{1-\beta^2}N_{n+1} \end{aligned} \tag{16}$$

With $\{N_n\}$ i.i.d. and Gaussian N(0,1), it follows that $A^* = -1$ and $\theta^* = -b$.

The sequence $\{\mathcal{W}_n\}$ resembles the eligibility vector appearing in the TD-algorithms of reinforcement learning [10], [15].

The steady-state variance of $\{\mathcal{W}_n\}$ is unity, but its asymptotic variance Σ^W_{CLT} is large when $\beta \sim 1$:

$$\Sigma^W_{\text{CLT}} = \sum_{n=-\infty}^{\infty} \mathsf{E}[\mathcal{W}_n\mathcal{W}_0] = \frac{1+\beta}{1-\beta}$$

with expectations in steady-state, giving $\mathsf{E}[\mathcal{W}_n\mathcal{W}_0] = \beta^{|n|}$.

Bias can be approximated for small α by

$$\begin{aligned} \lim_{n\to\infty} \mathsf{E}[\theta_n] - \theta^* &= -\alpha \lim_{n\to\infty} \mathsf{E}\big[\Upsilon_n\big] \\ &\approx \frac{\beta}{1-\beta}[1+\theta^*]\alpha = \theta^* + 99\alpha \end{aligned} \tag{17}$$

The Polyak-Ruppert covariance is the scalar,

$$\Sigma^{\text{PR}} = G\Sigma^W_{\text{CLT}}G^{\intercal} = [1+\theta^*]^2\Sigma^W_{\text{CLT}} \tag{18}$$

The numerical results that follow are based on the Gaussian model using $\beta = 0.9$, $\theta^* = 10$ and a short time-horizon of $N = 10^4$. Five values of α were tested for the fixed step-size algorithm, and five values of ρ for the vanishing step-size case:

$\alpha =$ 5e-4	2.8e-3	1.58e-2	8.89e-2	0.5
$\rho = 0.4000$	0.5375	0.6750	0.8125	0.9

The values of α are spaced equally on a logarithmic scale, and the values of ρ are spaced linearly. The remaining two terms in (2) where chosen to be $g = \alpha^{\max} = 0.5$ and $n_e = 1$.

The estimates for the fixed step-size algorithm remain bounded in n for this range of α. Boundedness and asymptotic consistency holds for the vanishing step-size algorithm, as predicted by theory for $\rho \in (\frac{1}{2}, 1]$ [3].

In application of PR-averaging (3), the value $N_0 = 0.2N$ was chosen in all ten cases. With the given numerical values, applying (18) gives the approximation for the vanishing gain algorithm,

$$(N - N_0)\mathsf{E}[(\theta^{\text{PR}}_N - \theta^*)^2] \approx \Sigma^{\text{PR}} \approx 2.3 \times 10^3$$

To obtain estimates of mean and bias, for each experiment $M = 500$ independent runs were conducted, initialized independently $\theta^i_0 \sim N(0, 25)$ and $\mathcal{W}_0 \sim N(0, 1)$.

Fig. 1 shows the estimates of mean and variance obtained in each case. The plot does not reveal much information for the fixed step-size algorithms because most values of α gave very poor results. The singular winner over all fixed step-size gains was $\alpha^\star = 2.8 \times 10^{-3}$, resulting in $\lim_{n\to\infty} \mathsf{E}[\theta_n] \approx 10.29$ and $\lim_{N\to\infty}(N - N_0)\text{Cov}\,[\theta_N] \approx 0.7 * \Sigma^{\text{PR}}$. The other four performed far worse.

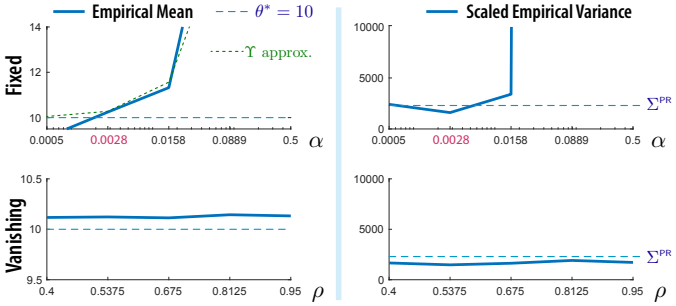

Fig. 1. Comparison of mean and bias obtained from PR-averaging in a scalar random linear system.

Each of the experiments using a vanishing gain resulted in variance of approximately equal to what was obtained using $\alpha^\star$ and with smaller bias.

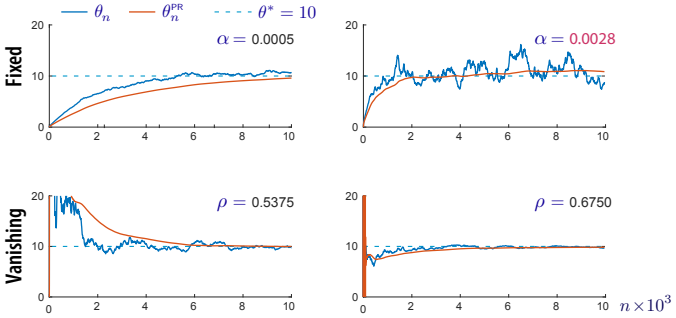

Fig. 2. Sample paths from four experiments.

Large bias can be anticipated for the fixed step-size algorithms by the approximation in (17). For $\alpha^\star$ we have $\theta^* + 99\alpha^\star \approx 10.28$, so this approximation nearly matches the approximation $\lim_{n\to\infty} \mathsf{E}[\theta_n] \approx 10.29$ obtained through simulation for $\alpha^\star = 2.8 \times 10^{-3}$.

See the plot on the upper left hand side of Fig. 1 for a comparison of this approximation with the empirical mean. For the smallest value of α tested, the parameter estimates are far from steady-state by the end of the run. In this case we typically observe *negative* bias. The cause of the negative bias for $\alpha = 5\times10^{-4}$ is explained by the fact that θ^i_0 is drawn from $N(0, 25)$ (so zero mean, while $\theta^* = 10$). Fig. 2 shows sample paths with and without averaging for two selected values of fixed step-size, and two values of ρ for vanishing step-size, with initialization $\theta_0 = 0$ in each case. It is clear from these plots why $\alpha = 5 \times 10^{-4}$ fails, and $\alpha = 2.8 \times 10^{-3}$ performs much better.

The ten subplots in Fig. 3 show histograms of $\{\theta^i_N\,, \theta^{\text{PR}\,i}_N : 1 \le i \le M\}$ for each of the ten settings. The results using a vanishing step-size are not sensitive to ρ, even though theory is violated for the smallest value $\rho = 0.4$ (recall that standard theory requires $\frac{1}{2} < \rho < 1$ in (2)).

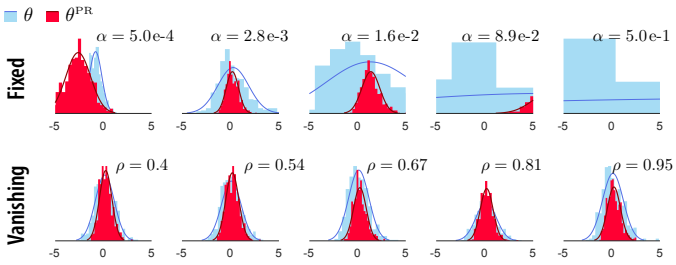

Fig. 3. Histograms of the error $\tilde{\theta}_N$ and $\tilde{\theta}_N^{\text{PR}}$ obtained from $M = 500$ independent runs in each of 10 experiments. The top row shows results using the fixed step-size algorithm for various values of α, and the bottom row shows results from the vanishing step-size (2).

3. Quasi stochastic approximation: Theory does not offer much encouragement to use a fixed step-size in SA (outside of linear SA with additive white noise). There is bias that cannot be removed with averaging, and we see in experiments that variance may also be large.

Performance can be improved when the sequence $\mathbf{\Phi}$ can be chosen by the user, such as in gradient free optimization or reinforcement learning, provided all randomness is a product of *exploration*. In recent work [6], [8] it is shown that bias can be reduced to $O(\alpha^2)$ in *quasi stochastic approximation* for which $\mathbf{\Phi}$ remains Markovian, but *deterministic*. An example is the $K-$dimensional clock process $\mathbf{\Phi}$ with entries $\Phi_n^i = \exp(2\pi j[\omega_i n + \phi_i])$, so that X is a bounded subset of $\mathbb{C}^K$.

The pair process $\{\Psi_n = (\theta_n, \Phi_n)\}$ is a Feller Markov chain; if the sample paths are bounded, then it admits at least one invariant measure ϖ. Provided ϖ is unique, the identity (13) holds in modified form:

$$b_{\bar{f}} := \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \bar{f}(\theta_n) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \Upsilon_n = \alpha\bar{\Upsilon} \quad (19)$$

in which the definitions are unchanged: Υ in (10c) and its mean $\bar{\Upsilon} = \int \Upsilon(\theta, x)\, \varpi(d\theta, dx)$. The existence of $\hat{f}$ is assured when f is smooth and $\{\omega_i\}$ chosen with care [7], [8].

This theory is developed in continuous time, where the definition of Υ is modified slightly and is shown that $\bar{\Upsilon}$ is *zero* provided the frequencies $\{\omega_i\}$ are irrationally related. We can expect the same bias bounds provided we are careful in choice of approximation of the ODE—this is a topic for future research.

While these results are encouraging, results obtained using a vanishing step-size algorithm are much better. It is likely that theory from [6] can be extended to the discrete time setting of this paper to obtain the following bounds using step-size (2), and subject to the smoothness and stability assumptions imposed in this prior work: for a fixed constant b_f and any initial condition $(\theta_0; \Phi_0)$,

$$\lim_{n\to\infty} n^{4\rho} \|\theta_n^{\text{PR}} - \theta^*\|^2 \le b_f$$

That is, instead of the $O(1/n)$ rate of convergence that is found in the most efficient SA recursions, we obtain a rate that is arbitrarily close to $O(1/n^4)$ by choosing ρ close to unity.

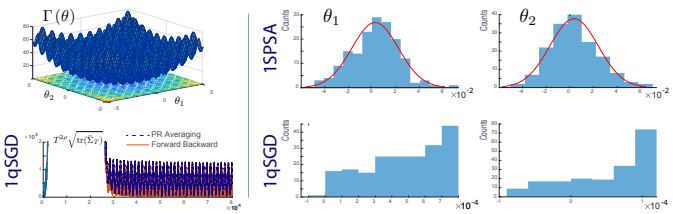

Fig. 4. Rastrigin objective (top left), scaled trace of empirical covariance (bottom left), histograms of estimation error for 1SPSA with PR averaging (top middle and top right), histograms of estimation error for 1qSGD with PR averaging (bottom middle and bottom right).

Fig. 4 from [6] shows what can be expected: this is a comparison of Spall's 1SPSA and its QSA counterpart 1qSGD on a non-convex objective function. The standard deviation of the estimation error is reduced by two orders of magnitude through the use of pseudo randomness.

REFERENCES

[1] F. Bach. *Learning Theory from First Principles*. In Preparation, 2021.

[2] A. Benveniste, M. Métivier, and P. Priouret. *Adaptive algorithms and stochastic approximations*, volume 22. Springer Science & Business Media, Berlin Heidelberg, 2012.

[3] V. Borkar, S. Chen, A. Devraj, I. Kontoyiannis, and S. Meyn. The ODE method for asymptotic statistics in stochastic approximation and reinforcement learning. *arXiv e-prints:2110.14427*, pages 1–50, 2021.

[4] V. S. Borkar and S. P. Meyn. The ODE method for convergence of stochastic approximation and reinforcement learning. *SIAM J. Control Optim.*, 38(2):447–469, 2000.

[5] A. Durmus, E. Moulines, A. Naumov, S. Samsonov, K. Scaman, and H.-T. Wai. Tight high probability bounds for linear stochastic approximation with fixed step-size. *Advances in Neural Information Processing Systems and arXiv:2106.01257*, 34:30063–30074, 2021.

[6] C. K. Lauand and S. Meyn. Approaching quartic convergence rates for quasi-stochastic approximation with application to gradient-free optimization. *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[7] C. K. Lauand and S. Meyn. Extremely fast convergence rates for extremum seeking control with Polyak-Ruppert averaging. *arXiv 2206.00814*, 2022.

[8] C. K. Lauand and S. Meyn. Markovian foundations for quasi stochastic approximation with applications to extremum seeking control. *arXiv 2207.06371*, 2022.

[9] M. Métivier and P. Priouret. Applications of a Kushner and Clark lemma to general classes of stochastic algorithms. *Trans. on Information Theory*, 30(2):140–151, March 1984.

[10] S. Meyn. *Control Systems and Reinforcement Learning*. Cambridge University Press, Cambridge, 2021.

[11] W. Mou, C. Junchi Li, M. J. Wainwright, P. L. Bartlett, and M. I. Jordan. On linear stochastic approximation: Fine-grained Polyak-Ruppert and non-asymptotic concentration. *Conference on Learning Theory and arXiv:2004.04719*, pages 2947–2997, 2020.

[12] B. T. Polyak. A new method of stochastic approximation type. *Avtomatika i telemekhanika (in Russian). translated in Automat. Remote Control, 51 (1991)*, pages 98–107, 1990.

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

[14] D. Ruppert. Efficient estimators from a slowly convergent Robbins-Monro processes. Technical Report Tech. Rept. No. 781, Cornell University, School of Operations Research and Industrial Engineering, Ithaca, NY, 1988.

[15] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 2nd edition, 2018.