

SGD with Vanishing Correlated Gradient Errors

(technique notes)

Abstract. Online estimation can improve the accuracy of individual gradient estimates while increasing the asymptotic variance of averaged stochastic gradient descent (SGD). We study this phenomenon through a weighted historical model for vanishing correlated gradient errors. Under strong convexity and smoothness, we establish mean-square and almost-sure convergence and derive a Polyak–Ruppert central limit theorem with an explicit covariance determined by cumulative historical weights. Running averages and recursive weighted averages illustrate how estimator gains affect single-step accuracy and downstream efficiency differently. We also establish a sharp variance lower bound within the class of normalized deterministic historical linear estimators and construct a recursion that attains it. Counterintuitively, this recursion has a slower single-step estimation MSE rate than the running mean, yet achieves the minimum downstream PR asymptotic variance in that class. Experiments on a nonlinear stochastic optimization problem confirm the predicted covariance formulas and demonstrate that minimizing single-step estimation error need not minimize the asymptotic variance of the averaged SGD iterate.

1 Introduction

Consider a nonlinear stochastic optimization problem,

$$\min_{\theta \in \mathbb{R}} F(\theta) = \mathbb{E} \left[\frac{1}{2} (\theta - X)^2 + \log \cosh \theta \right], \quad F'(\theta) = \theta + \tanh \theta - m,$$

where $m = \mathbb{E}X$ is unknown. A stochastic gradient can use either a fresh observation X_t or the running estimate $\hat{m}_t = t^{-1} \sum_{s=1}^t X_s$ in place of m . The latter uses more observations and becomes progressively more accurate, suggesting that it should also improve the final averaged estimate.

For $X \sim N(0, 1)$, however, replacing X_t by $\hat{m}_t$ reduces the gradient estimation MSE from 1 to $1/t$ while increasing the Polyak–Ruppert (PR) asymptotic variance from $1/4$ to $1/2$. How can a progressively more accurate gradient estimate produce a less precise averaged iterate? Successive running estimates reuse observations, so their errors are temporally correlated and accumulate coherently. Their single-step MSE alone does not determine the variance of the downstream average.

Related historical dependence arises from online moment estimation, streaming feature normalization, and online subspace estimation, as well as recursive inner-expectation estimation in compositional optimization [9] and nuisance-dependent gradients [10].

This note studies these gradient errors through a unified weighted historical representation. We characterize their contribution to PR fluctuations through cumulative weights and derive explicit covariance formulas. The examples and a sharp variance bound within the normalized historical linear estimator class clarify why optimizing a single-step estimator and optimizing its downstream PR average can favor different choices.

The same distinction yields a second counterintuitive conclusion. A recursive auxiliary estimator with single-step MSE of order $t^{-\rho}$, $1/2 < \rho < 1$, can attain the smallest downstream PR asymptotic covariance within the normalized historical linear estimator class. A running mean has the faster t^{-1} single-step rate but a larger downstream covariance. Thus improving the auxiliary estimator’s single-step convergence rate can conflict with optimizing the final averaged SGD estimate. Proposition 4 and Section 5.3 establish and illustrate this distinction.

2 Related work

Ruppert [1] and Polyak and Juditsky [2] established the classical averaging principle and its asymptotic efficiency under stochastic approximation conditions. Kushner and Yang [3] extend averaging results to more general stochastic approximation processes. Mou et al. [4] give a refined covariance analysis for averaged linear stochastic approximation with a fixed step size. These results provide the background for understanding how averaging transforms stochastic updates into a limiting covariance.

Coupled stochastic approximation is particularly relevant when one recursion supplies estimates to another. Konda and Tsitsiklis [5] establish asymptotic covariance and normality results for linear two-time-scale recursions. Mokkadem and Pelletier [6] extend the analysis to nonlinear recursions; their Theorem 2 establishes a joint CLT for the two averaged components. Their later work [7] uses two-time-scale algorithms to construct asymptotically efficient estimators. This theory provides a closely related treatment of recursive auxiliary estimation and its propagation into an averaged downstream process.

Biased-gradient analyses, such as Ajalloeian and Stich [8], quantify the effects of gradient errors on convergence and attainable accuracy. Online auxiliary estimation also appears in stochastic compositional gradient descent, where Wang, Fang, and Liu [9] track an inner expectation recursively, and in nuisance-dependent optimization, where Yu et al. [10] study convergence and gradient orthogonalization. These settings motivate the study of estimation errors that persist across successive gradient evaluations.

The focus of this note is the temporal correlation structure generated by deterministic historical weights. Cumulative weights give a direct expression for the PR covariance, including cross-covariance with the SGD noise, and make the effect of auxiliary-estimator gains explicit.

2.1 Our Contributions

1. **Historical correlation and preservation of SGD guarantees.** We study vanishing gradient errors generated by deterministic weighted histories of random innovations. We give moment and cumulative-weight conditions that preserve mean-square and almost-sure convergence and $\sqrt{T}$ -asymptotic normality of the PR average, and characterize the resulting covariance. The classical PR covariance is recovered when the normalized cumulative error vanishes, in particular when $\mathbb{E}\|b_t\|^2 = O(t^{-\beta})$ with $\beta > 1$ under the stated regularity conditions.
2. **A sharp downstream variance bound and an attaining estimator.** Within the normalized deterministic historical linear estimator class, we establish a cumulative variance lower bound and construct a recursive estimator attaining it. Under zero cross-covariance with the SGD noise and the PR conditions, this construction minimizes the downstream asymptotic covariance. Counterintuitively, its single-step estimation MSE decays at order $t^{-\rho}$, $1/2 < \rho < 1$, while a running mean achieves t^{-1} but has a larger downstream covariance. Thus the estimator with the faster single-step MSE rate can yield the less efficient PR average.

3 Main results

Assumption 1 (Objective regularity).

- (a) $F : \mathbb{R}^p \rightarrow \mathbb{R}$ is differentiable, μ -strongly convex, and L -smooth, with $0 < \mu \leq L$. Its unique minimizer is denoted by θ^* .
- (b) F is twice continuously differentiable. Write $H = \nabla^2 F(\theta^*) \succ 0$. For some finite $L_H \geq 0$ and all $\theta, \vartheta \in \mathbb{R}^p$, its Hessian satisfies

$$\|\nabla^2 F(\theta) - \nabla^2 F(\vartheta)\|_{\text{op}} \leq L_H \|\theta - \vartheta\|. \quad (3.1)$$

Part (a) is used for convergence; part (b) additionally controls the nonlinear remainder in the PR limit.

Let $(\mathcal{F}_t)_{t \geq 1}$ be a filtration. We consider

$$\theta_{t+1} = \theta_t - \eta_t [\nabla F(\theta_t) + \xi_{t+1} + b_t], \quad \eta_t = Dt^{-\alpha}, \quad (3.2)$$

where $\theta_t, b_t \in \mathbb{R}^p$ are $\mathcal{F}_t$ -measurable and ξ_{t+1} is $\mathcal{F}_{t+1}$ -measurable. The gradient error has the weighted historical representation

$$b_t = \sum_{s=1}^t W_{t,s} \varepsilon_s + \delta_t, \quad W_{t,s} \in \mathbb{R}^{p \times q}. \quad (3.3)$$

The weights are deterministic; ε_s is $\mathcal{F}_s$ -measurable, and δ_t is $\mathcal{F}_t$ -measurable. All vector norms are Euclidean.

Theorem 1 (Mean-square and almost-sure convergence). *Suppose Assumption 1(a) holds, $a_1 := \mathbb{E}\|\theta_1 - \theta^*\|^2 < \infty$, and $\eta_t = Dt^{-\alpha}$ with $D > 0$ and $1/2 < \alpha < 1$. Assume the following conditions:*

- (i) *The gradient noise satisfies, almost surely,*

$$\mathbb{E}[\xi_{t+1} \mid \mathcal{F}_t] = 0, \quad \mathbb{E}[\|\xi_{t+1}\|^2 \mid \mathcal{F}_t] \leq \sigma^2.$$

- (ii) *The innovations ε_s are independent, centered, and have common finite covariance Q . For finite constants $B_W, B_\delta \geq 0$,*

$$\sum_{s=1}^t \text{tr}(W_{t,s} Q W_{t,s}^\top) \leq \frac{B_W}{t}, \quad \mathbb{E}\|\delta_t\|^2 \leq \frac{B_\delta}{t}, \quad t \geq 1. \quad (3.4)$$

Then, with $a_t = \mathbb{E}\|\theta_t - \theta^*\|^2$, for every integer $t \geq n$,

$$a_t \leq C_{\text{init}} \exp\left[-c_0(t^{1-\alpha} - n^{1-\alpha})\right] + C_\xi t^{-\alpha} + C_b t^{-1}. \quad (3.5)$$

In particular, $\theta_t \rightarrow \theta^*$ in mean square and almost surely.

The constants in (3.5) may be chosen as

$$\begin{aligned} B &= 2(B_W + B_\delta), \quad c_0 = \frac{\mu D}{2(1-\alpha)}, \quad C_\xi = \frac{4D\sigma^2}{\mu}, \quad C_b = \frac{12B}{\mu^2}, \\ n &= \max\left\{1, \left\lceil (DL)^{1/\alpha} \right\rceil, \left\lceil \left(\frac{4}{\mu D}\right)^{1/(1-\alpha)} \right\rceil\right\}, \quad C_{\text{init}} = a_n. \end{aligned} \quad (3.6)$$

An entirely explicit bound for C_{init} in terms of a_1 is provided in Appendix A.1. The result allows biased remainders and history-dependent errors. The same conclusion and constants hold for any adapted b_t satisfying $\mathbb{E}\|b_t\|^2 \leq B/t$. If $\delta_t \equiv 0$, one may use the sharper choice $B = B_W$ throughout.

Theorem 2 (Polyak–Ruppert averaging under weighted historical errors). Consider (3.2)–(3.3), with the adaptedness and deterministic-weight conventions above. Suppose Assumption 1(a) holds, $\mathbb{E}\|\theta_1 - \theta^*\|^2 < \infty$, and $\eta_t = Dt^{-\alpha}$, where $D > 0$ and $1/2 < \alpha < 1$. Assume the following conditions:

(i) The joint innovations $Z_s = (\tilde{\xi}_s^\top, \varepsilon_s^\top)^\top$ are i.i.d., centered, and have finite covariance

$$Q = \text{Cov}(Z_s) = \begin{pmatrix} S & C_0 \\ C_0^\top & Q \end{pmatrix}. \quad (3.7)$$

For every $t \geq 1$, Z_{t+1} is independent of $\mathcal{F}_t$.

(ii) For some $B < \infty$ and $\beta > 0$, the error and remainder satisfy

$$\mathbb{E}\|b_t\|^2 \leq Bt^{-\beta} \quad (t \geq 1), \quad \frac{1}{\sqrt{T}} \sum_{t=1}^T \delta_t \xrightarrow{p} 0. \quad (3.8)$$

Define the cumulative weights and their covariance matrix by

$$K_{T,s} = \sum_{t=s}^T W_{t,s}, \quad V_T = \frac{1}{T} \sum_{s=1}^T K_{T,s} Q K_{T,s}^\top. \quad (3.9)$$

Write $B_T = T^{-1/2} \sum_{t=1}^T b_t$ and $\Gamma_T^{(b)} = \text{Cov}(B_T)$. For the PR central limit theorem, suppose additionally that Assumption 1(b) holds, $\beta > \max\{1/2, 2\alpha - 1\}$, and

$$V_T \rightarrow \Gamma, \quad \frac{1}{T} \sum_{s=1}^T K_{T,s} \rightarrow K, \quad \max_{1 \leq s \leq T} \frac{\|K_{T,s}\|_{\text{op}}}{\sqrt{T}} \rightarrow 0. \quad (3.10)$$

Then, for $\bar{\theta}_T = T^{-1} \sum_{t=1}^T \theta_t$,

$$\sqrt{T}(\bar{\theta}_T - \theta^*) \Rightarrow N(0, \Sigma_{\text{PR}}), \quad \Sigma_{\text{PR}} = H^{-1} \Sigma_{\text{eff}} H^{-\top}, \quad (3.11)$$

where

$$\Sigma_{\text{eff}} = S + \Gamma + C_0 K^\top + K C_0^\top. \quad (3.12)$$

If $\beta > 1$, the cumulative error is negligible and the PR limit holds with $\Sigma_{\text{PR}} = H^{-1} S H^{-\top}$ even without (3.10).

The cumulative error covariance has the following classification under the second-moment bound in (3.8):

Decay exponent	Behavior of $\Gamma_T^{(b)}$
$\beta > 1$	Necessarily converges to zero
$\beta = 1$	Bounded; convergence is not guaranteed
$0 < \beta < 1$	May vanish, have a finite nonzero limit, or diverge

For $0 < \beta < 1$, nonconvergence without divergence is also possible.

The table follows from the second-moment bound on b_t . For the PR limit, $\beta > \max\{1/2, 2\alpha - 1\}$ controls the nonlinear and summation-by-parts remainders and holds in the principal regime $\beta = 1$. The matrix Γ determines the distributional limit of B_T ; it also equals $\lim_T \Gamma_T^{(b)}$ if the cumulative remainder in (3.8) converges to zero in L^2 .

3.1 Examples

Both examples use $\varepsilon_t \in \mathbb{R}^p$ and the joint innovation conditions of Theorem 2, with $C_0 = \text{Cov}(\xi_t, \varepsilon_t) = 0$. Here ξ_t is the SGD sampling noise and ε_t is the fresh innovation of the auxiliary estimator. Independent data streams are one way to ensure this condition. The objective, initial iterate, and SGD step size satisfy the PR conditions of Theorem 2. The calculations below verify its remaining error and weight conditions. Detailed derivations are given in Appendix A.3.

Example 1: Running average. Let

$$b_t = \frac{1}{t} \sum_{s=1}^t \varepsilon_s. \quad (3.13)$$

Then $W_{t,s} = t^{-1}I$, $\delta_t = 0$, and

$$\mathbb{E}[b_t b_t^\top] = \frac{Q}{t}, \quad \mathbb{E}\|b_t\|^2 = \frac{\text{tr}(Q)}{t}.$$

Thus the error lies in the critical regime $\beta = 1$. Writing $h_t = \sum_{j=1}^t j^{-1}$ and $h_0 = 0$, the cumulative weights are $K_{T,s} = (h_T - h_{s-1})I$. They satisfy the exact identities

$$\frac{1}{T} \sum_{s=1}^T K_{T,s} = I, \quad \frac{1}{T} \sum_{s=1}^T K_{T,s} Q K_{T,s}^\top = \left(2 - \frac{h_T}{T}\right) Q. \quad (3.14)$$

Moreover, $\max_s \|K_{T,s}\|_{\text{op}} / \sqrt{T} = h_T / \sqrt{T} \rightarrow 0$. Consequently, $K = I$, $\Gamma = 2Q$, and

$$\boxed{\sqrt{T}(\bar{\theta}_T - \theta^*) \Rightarrow N\left(0, H^{-1}(S + 2Q)H^{-\top}\right)}. \quad (3.15)$$

Although the one-step error covariance is Q/t , temporal reuse of the innovations produces a cumulative covariance of $2Q$.

Example 2: Recursive weighted average. For a fixed gain $c > 1/2$ and a deterministic initial value b_0 , let

$$b_t = \left(1 - \frac{c}{t}\right) b_{t-1} + \frac{c}{t} \varepsilon_t. \quad (3.16)$$

The gain c controls the weight c/t of the new innovation. For $t > c$, this is a convex combination of the previous estimate and the new innovation; the finitely many earlier steps are included in the initial-transient analysis. Expanding the recursion gives

$$W_{t,s} = \frac{c}{s} \prod_{j=s+1}^t \left(1 - \frac{c}{j}\right) I, \quad \delta_t = \prod_{j=1}^t \left(1 - \frac{c}{j}\right) b_0, \quad (3.17)$$

with empty products equal to one. The initial term is $O(t^{-c})$, so $T^{-1/2} \sum_{t=1}^T \delta_t \rightarrow 0$. Furthermore,

$$t \mathbb{E}[b_t b_t^\top] \rightarrow \frac{c^2}{2c-1} Q, \quad \mathbb{E}\|b_t\|^2 = O(t^{-1}). \quad (3.18)$$

For $s/T \rightarrow x \in (0, 1]$, the cumulative weights converge to $k_c(x)I$, where

$$k_c(x) = cx^{c-1} \int_x^1 v^{-c} dv = \begin{cases} \frac{c}{c-1} (1 - x^{c-1}), & c \neq 1, \\ \log(1/x), & c = 1. \end{cases} \quad (3.19)$$

The condition $c > 1/2$ ensures square integrability at $x = 0$ and yields

$$K = \left(\int_0^1 k_c(x) dx \right) I = I, \quad \Gamma = \left(\int_0^1 k_c(x)^2 dx \right) Q = \frac{2c}{2c-1} Q. \quad (3.20)$$

The maximal weight condition also holds. Therefore

$$\sqrt{T}(\bar{\theta}_T - \theta^*) \Rightarrow N\left(0, H^{-1} \left[S + \frac{2c}{2c-1} Q \right] H^{-\top} \right). \quad (3.21)$$

At $c = 1$, (3.16) is exactly the running average, regardless of b_0 , and both covariance formulas agree.

Single-step and downstream parameter choices. For fixed Q, S, H , the gain-dependent coefficients in the single-step and downstream asymptotic covariances are, respectively,

$$f_{\text{single}}(c) = \frac{c^2}{2c-1}, \quad f_{\text{PR}}(c) = \frac{2c}{2c-1}, \quad c > 1/2. \quad (3.22)$$

Their derivatives are

$$f'_{\text{single}}(c) = \frac{2c(c-1)}{(2c-1)^2}, \quad f'_{\text{PR}}(c) = -\frac{2}{(2c-1)^2}.$$

When $Q \neq 0$, the single-step asymptotic MSE is minimized at $c = 1$, whereas the downstream PR covariance decreases in the positive semidefinite order as c increases. Its infimum is $H^{-1}(S + Q)H^{-\top}$ as $c \rightarrow \infty$ and is not attained at a finite gain. The two asymptotic criteria therefore favor different gain choices. These limits hold for fixed c as $T \rightarrow \infty$.

4 Variance lower bound and attainability

The examples suggest asking how small the cumulative error covariance can be. We compare deterministic scalar historical weights applied to i.i.d. centered innovations $\varepsilon_s \in \mathbb{R}^p$ with covariance Q , subject to the normalization

$$b_t = \sum_{s=1}^t w_{t,s} \varepsilon_s, \quad \sum_{s=1}^t w_{t,s} = 1. \quad (4.1)$$

This normalization preserves constant inputs, as required for a linear estimator of a common mean. The lower bound below concerns this specified estimator class.

Proposition 3 (Variance lower bound for normalized historical weights). *Under (4.1), let $k_{T,s} = \sum_{t=s}^T w_{t,s}$ and $\Gamma_T = \text{Cov}(T^{-1/2} \sum_{t=1}^T b_t)$. Then, for every $T \geq 1$,*

$$\Gamma_T - Q = \frac{Q}{T} \sum_{s=1}^T (k_{T,s} - 1)^2 \succeq 0. \quad (4.2)$$

In particular,

$$\frac{1}{T} \sum_{s=1}^T (k_{T,s} - 1)^2 \rightarrow 0 \implies \Gamma_T \rightarrow Q. \quad (4.3)$$

If $Q \neq 0$, this condition is also necessary for $\Gamma_T \rightarrow Q$. Whenever the downstream PR CLT holds with $C_0 = 0$ and $\Gamma_T \rightarrow \Gamma$, its covariance satisfies

$$\Sigma_{\text{PR}} \succeq H^{-1}(S + Q)H^{-\top}. \quad (4.4)$$

The running average has cumulative covariance $2Q$, and the fixed-gain c/t recursion approaches Q only as $c \rightarrow \infty$. The next construction attains Q using a fixed gain schedule, while its single-step MSE decays more slowly than t^{-1} .

Proposition 4 (Attainment by a polynomial-gain recursion). *Let the innovations be as above, fix $0 < a \leq 1$ and $1/2 < \rho < 1$, and set*

$$b_1 = \varepsilon_1, \quad b_t = (1 - \gamma_t)b_{t-1} + \gamma_t \varepsilon_t, \quad \gamma_t = at^{-\rho}, \quad t \geq 2. \quad (4.5)$$

Equivalently, take $\gamma_1 = 1$. The resulting deterministic weights are nonnegative and exactly normalized. They satisfy

$$\mathbb{E}\|b_t\|^2 = O(t^{-\rho}), \quad \frac{1}{T} \sum_{s=1}^T (k_{T,s} - 1)^2 = O(T^{\rho-1}), \quad \Gamma_T \rightarrow Q. \quad (4.6)$$

Thus the lower bound of Proposition 3 is attained. If the objective, SGD step size, and joint innovation conditions of Theorem 2 hold, $C_0 = 0$, and $\rho > 2\alpha - 1$, then

$$\boxed{\sqrt{T}(\bar{\theta}_T - \theta^*) \Rightarrow N\left(0, H^{-1}(S + Q)H^{-\top}\right)}. \quad (4.7)$$

The initialization $b_1 = \varepsilon_1$ ensures exact normalization. The additional condition $\rho > 2\alpha - 1$ controls the downstream PR remainders. The three gain schedules compare as follows:

Gain schedule	Single-step MSE	Cumulative covariance limit
$1/t$	$O(t^{-1})$	$2Q$
$c/t, c > 1/2$	$O(t^{-1})$	$\frac{2c}{2c-1}Q$
$at^{-\rho}, 1/2 < \rho < 1$	$O(t^{-\rho})$	Q

The third row uses (4.5); the c/t row uses Example 2 with its cumulatively negligible initialization term.

The bound is sharp within the normalized deterministic historical linear estimator class.

5 Experiments

5.1 More accurate gradients, larger PR variance

We compare two gradient estimators for the same stochastic problem,

$$\min_{\theta \in \mathbb{R}} F(\theta) = \mathbb{E} \left[\frac{1}{2}(\theta - X)^2 + \log \cosh \theta \right], \quad X \sim N(m, 1), \quad (5.1)$$

where the algorithms observe samples but do not use the population mean m . Since $F'(\theta) = \theta + \tanh \theta - m$ and $F''(\theta) = 1 + \text{sech}^2 \theta \in [1, 2]$, the objective is strongly convex and smooth. Moreover, $|F'''(\theta)| \leq 2$, so its Hessian is Lipschitz. We set $m = 0$, giving $\theta^* = 0$ and $H = 2$.

At iteration t , the fresh-sample method uses

$$g_t(\theta) = \theta + \tanh \theta - X_t,$$

whereas the historical-mean method updates $\hat{m}_t = \hat{m}_{t-1} + (X_t - \hat{m}_{t-1})/t$ from $\hat{m}_0 = 0$ and uses

$$\tilde{g}_t(\theta) = \theta + \tanh \theta - \hat{m}_t.$$

Each method takes one SGD step using its own current iterate. At any fixed θ , their gradient estimation MSEs are

$$\mathbb{E}|g_t(\theta) - F'(\theta)|^2 = 1, \quad \mathbb{E}|\tilde{g}_t(\theta) - F'(\theta)|^2 = \frac{1}{t}. \quad (5.2)$$

The fresh-sample method has independent additive noise of variance one. The historical method corresponds to $\xi_{t+1} = 0$ and $b_t = -(\hat{m}_t - m)$, whose cumulative covariance is two by Example 1. Consequently, the predicted limits are

$$\sqrt{T}(\bar{\theta}_T - \theta^*) \Rightarrow \begin{cases} N(0, 1/4), & \text{fresh-sample gradient,} \\ N(0, 1/2), & \text{historical-mean gradient.} \end{cases} \quad (5.3)$$

These are total variances for two gradient strategies solving the same objective.

We run 10,000 independent replications with $\theta_1 = 2$ and $\eta_t = 0.5t^{-0.65}$. Within each replication, both methods receive the same sample sequence and one new observation per iteration. We average $\theta_1, \dots, \theta_T$ without discarding an initial interval, and record $\sqrt{T}(\bar{\theta}_T - \theta^*)$ at $T = 200,000$. The random seed is 20261016.

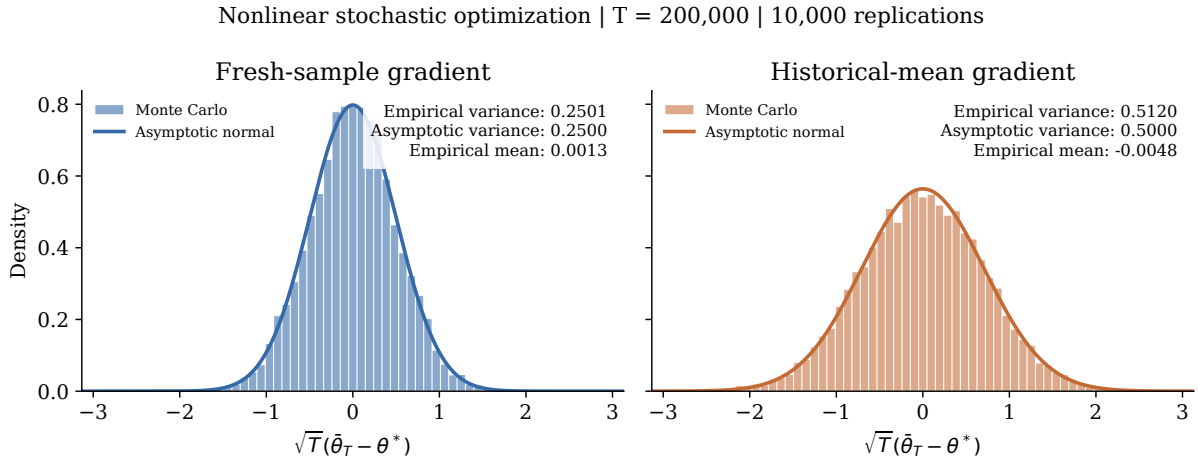


Figure 1. PR fluctuations for the same nonlinear stochastic objective. Each histogram contains 10,000 independent replications at $T = 200,000$. The panels share bins and axes; solid curves are the asymptotic normal densities in (5.3). Historical averaging reduces single-step gradient MSE but increases the PR asymptotic variance.

At $T = 200,000$, the empirical PR variances are 0.2501 for fresh samples and 0.5120 for the historical mean, with estimated Monte Carlo standard errors 0.0035 and 0.0072, respectively. Figure 1 displays the wider historical-mean distribution. The empirical single-step gradient MSEs at the largest horizon are 0.9906 and 4.9532×10^{-6} , compared with the exact values 1 and 5×10^{-6} . Thus a substantial improvement in single-step gradient accuracy accompanies a larger downstream PR variance.

5.2 Estimator gain and downstream efficiency

We retain the objective (5.1) with $m = 0$ and replace the running mean by the recursive estimator

$$\hat{m}_t = \left(1 - \frac{a}{t}\right) \hat{m}_{t-1} + \frac{a}{t} X_t, \quad \hat{m}_0 = 0, \quad a > 1/2. \quad (5.4)$$

The SGD gradient is $\theta_t + \tanh \theta_t - \hat{m}_t$. The two theoretical criteria, from Example 2, are

$$T \mathbb{E}(\hat{m}_T - m)^2 \longrightarrow \frac{a^2}{2a - 1}, \quad \text{Var}_{\text{asyp}} \left[\sqrt{T}(\bar{\theta}_T - \theta^*) \right] = \frac{1}{4} \frac{2a}{2a - 1}. \quad (5.5)$$

The first criterion is minimized at $a = 1$, while the second decreases with a toward $1/4$.

We scan $a \in \{0.6, 0.75, 1, 1.25, 1.5, 2, 3, 5, 8\}$ using 5,000 independent replications per gain. The initial SGD iterate, step size, and averaging convention are the same as in Section 5.1. Within each replication, all gains receive the same sample sequence. We record both criteria at $T = 50,000$ and $200,000$, using seed 20261017. Equation (5.4) is applied throughout the finite initial interval as well.

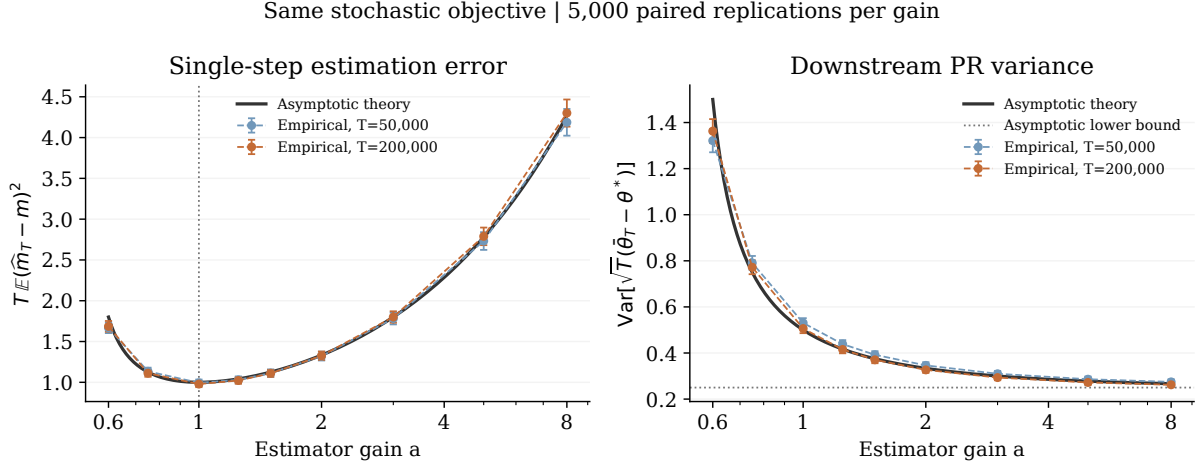


Figure 2. Estimator gain and the two efficiency criteria. The left panel shows the scaled single-step estimation MSE; the right panel shows the PR empirical variance. Solid black curves are the limits in (5.5); error bars are 1.96 Monte Carlo standard errors. The vertical reference marks $a = 1$, and the horizontal reference marks the downstream asymptotic lower bound $1/4$.

Figure 2 exhibits the predicted difference in parameter choices. At $T = 200,000$, the scaled single-step MSE is 0.9785 at $a = 1$ and increases to 4.3006 at $a = 8$, while the PR empirical variance decreases from 0.5062 to 0.2621. The corresponding theoretical values are 1 and 4.2667 for the scaled MSE, and 0.5 and 0.2667 for the PR variance. Across the sampled gains, the single-step criterion is smallest at $a = 1$, whereas the downstream variance is smallest at the largest sampled gain. At $a = 0.6$, the empirical PR variance increases from 1.3213 to 1.3628 between the two horizons, approaching its asymptotic value 1.5 more slowly near the threshold $a = 1/2$.

5.3 Attaining the downstream variance bound

We first state the theoretical predictions for the objective (5.1) with $m = 0$ and $H = 2$. The three auxiliary gain schedules yield

$$\Sigma_{\text{PR}} = \begin{cases} 1/2, & \gamma_t = 1/t, \\ 4/15, & \gamma_t = 8/t, \\ 1/4, & \gamma_t = t^{-0.6}. \end{cases} \quad (5.6)$$

The last value is the sharp downstream bound of Proposition 3, attained by the polynomial-gain construction of Proposition 4. Its single-step estimation MSE has the slower rate $t^{-0.6}$, compared with t^{-1} for the other two schedules. This predicts the counterintuitive combination of slower auxiliary estimation and optimal downstream PR variance.

We simulate all three estimators using

$$\hat{m}_t = (1 - \gamma_t)\hat{m}_{t-1} + \gamma_t X_t, \quad \hat{m}_0 = 0,$$

and the same downstream SGD update as before. The polynomial schedule has $\gamma_1 = 1$, so its historical weights are exactly normalized. Its exponent $\rho = 0.6$ satisfies $\rho > 2\alpha - 1$ with $\alpha =$

0.65. We use 10,000 independent replications, paired across the three methods, $T = 1,000,000$, and seed 20261018. All remaining SGD settings match Section 5.1.

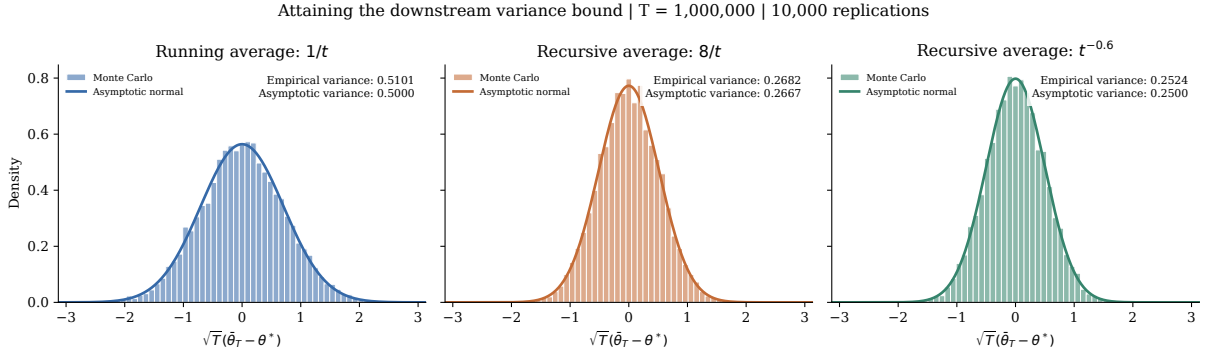


Figure 3. Downstream PR distributions for three auxiliary gain schedules. Each panel contains 10,000 independent replications at $T = 1,000,000$, with shared bins and axes. Solid curves are the theoretical normal densities with variances $1/2$, $4/15$, and $1/4$, respectively. The polynomial-gain estimator has slower single-step MSE decay but attains the smallest downstream asymptotic variance.

The empirical PR variances are 0.5101, 0.2682, and 0.2524, respectively, with Monte Carlo standard errors 0.0073, 0.0038, and 0.0036. Figure 3 compares the empirical distributions with the predictions in (5.6). At the same horizon, the auxiliary-estimation MSEs are 1.0031×10^{-6} , 4.3386×10^{-6} , and 1.2750×10^{-4} . Thus the polynomial estimator has the largest single-step error among the three methods, yet the smallest empirical PR variance, illustrating its theoretically optimal downstream limit.

References

- [1] D. Ruppert. Efficient estimations from a slowly convergent Robbins–Monro process. Technical Report 781, School of Operations Research and Industrial Engineering, Cornell University, 1988. Cornell University repository.
- [2] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992. doi:10.1137/0330046.
- [3] H. J. Kushner and J. Yang. Stochastic approximation with averaging of the iterates: Optimal asymptotic rate of convergence for general processes. *SIAM Journal on Control and Optimization*, 31(4):1045–1062, 1993. doi:10.1137/0331047.
- [4] W. Mou, C. J. Li, M. J. Wainwright, P. L. Bartlett, and M. I. Jordan. On linear stochastic approximation: Fine-grained Polyak–Ruppert and non-asymptotic concentration. In *Proceedings of the Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2947–2997, 2020. PMLR:125/mou20a.
- [5] V. R. Konda and J. N. Tsitsiklis. Convergence rate of linear two-time-scale stochastic approximation. *The Annals of Applied Probability*, 14(2):796–819, 2004. doi:10.1214/105051604000000116.
- [6] A. Mokkadem and M. Pelletier. Convergence rate and averaging of nonlinear two-time-scale stochastic approximation algorithms. *The Annals of Applied Probability*, 16(3):1671–1702, 2006. doi:10.1214/105051606000000448.
- [7] A. Mokkadem and M. Pelletier. A generalization of the averaging procedure: The use of two-time-scale algorithms. *SIAM Journal on Control and Optimization*, 49(4):1523–1543, 2011. doi:10.1137/080718395.
- [8] A. Ajalloeian and S. U. Stich. On the convergence of SGD with biased gradients. *arXiv preprint*, 2020. arXiv:2008.00051.

- [9] M. Wang, E. X. Fang, and H. Liu. Stochastic compositional gradient descent: Algorithms for minimizing compositions of expected-value functions. *Mathematical Programming*, 161(1–2):419–449, 2017. doi:10.1007/s10107-016-1017-3.
- [10] F. Yu, R. Mehta, A. Luedtke, and Z. Harchaoui. Stochastic gradients under nuisances. *Advances in Neural Information Processing Systems*, 2025. arXiv:2508.20326.
- [11] H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing Methods in Statistics*, pages 233–257. Academic Press, 1971. doi:10.1016/B978-0-12-604550-5.50015-8.

A Proofs

A.1 Proof of Theorem 1

Proof. Write $e_t = \theta_t - \theta^*$, $g_t = \nabla F(\theta_t)$, and $V_t = \|e_t\|^2$. We first bound the historical error, then establish a conditional contraction and use it separately for the quantitative mean-square bound and almost-sure convergence.

Historical error. Independence and centering of the innovations imply

$$\mathbb{E} \left\| \sum_{s=1}^t W_{t,s} \varepsilon_s \right\|^2 = \sum_{s=1}^t \text{tr}(W_{t,s} Q W_{t,s}^\top).$$

Using $\|u + v\|^2 \leq 2\|u\|^2 + 2\|v\|^2$ and (3.4) gives

$$\mathbb{E} \|b_t\|^2 \leq \frac{B}{t}, \quad B = 2(B_W + B_\delta). \quad (\text{A.1})$$

If $\delta_t \equiv 0$, the identity above gives $B = B_W$ directly.

Conditional contraction. For a convex L -smooth function, the gradient is $1/L$ -cocoercive. Since $\nabla F(\theta^*) = 0$, this and strong convexity give

$$\|g_t\|^2 \leq L \langle e_t, g_t \rangle, \quad \langle e_t, g_t \rangle \geq \mu V_t.$$

Consequently, whenever $\eta_t \leq 1/L$,

$$\begin{aligned} \|e_t - \eta_t g_t\|^2 &\leq V_t - (2\eta_t - L\eta_t^2) \langle e_t, g_t \rangle \\ &\leq (1 - \mu\eta_t) V_t. \end{aligned} \quad (\text{A.2})$$

In particular, $\|e_t - \eta_t g_t\| \leq \|e_t\|$. Young's inequality yields

$$2\eta_t |\langle e_t - \eta_t g_t, b_t \rangle| \leq \frac{\mu\eta_t}{2} V_t + \frac{2\eta_t}{\mu} \|b_t\|^2.$$

Conditioning the squared update on $\mathcal{F}_t$ eliminates the cross term with ξ_{t+1} . Since $\eta_t \leq 1/L \leq 1/\mu$, we obtain

$$\begin{aligned} \mathbb{E}[V_{t+1} \mid \mathcal{F}_t] &\leq \left(1 - \frac{\mu\eta_t}{2}\right) V_t + \left(\frac{2\eta_t}{\mu} + \eta_t^2\right) \|b_t\|^2 + \sigma^2 \eta_t^2 \\ &\leq \left(1 - \frac{\mu\eta_t}{2}\right) V_t + \frac{3\eta_t}{\mu} \|b_t\|^2 + \sigma^2 \eta_t^2. \end{aligned} \quad (\text{A.3})$$

All second moments are finite, including before the small-step regime. Indeed, Lipschitz continuity of the gradient implies $\|e_t - \eta_t g_t\| \leq (1 + \eta_t L)\|e_t\|$. Since $\eta_t \leq D$, the inequality $\|u + v + w\|^2 \leq 3(\|u\|^2 + \|v\|^2 + \|w\|^2)$ gives, for every $t \geq 1$,

$$a_{t+1} \leq Aa_t + 3D^2(\sigma^2 + B), \quad A = 3(1 + DL)^2 > 1.$$

Thus the initial interval is controlled explicitly by

$$C_{\text{init}} = a_n \leq A^{n-1}a_1 + 3D^2(\sigma^2 + B)\frac{A^{n-1} - 1}{A - 1}. \quad (\text{A.4})$$

The right-hand side may replace C_{init} in (3.5).

Explicit mean-square bound. Taking expectations in (A.3) and using (A.1) gives, for $t \geq n$,

$$a_{t+1} \leq (1 - ct^{-\alpha})a_t + qt^{-2\alpha} + ht^{-\alpha-1}, \quad c = \frac{\mu D}{2}, \quad q = D^2\sigma^2, \quad h = \frac{3DB}{\mu}. \quad (\text{A.5})$$

By the definition of n , $\eta_t \leq 1/L$ and $t^{1-\alpha} \geq 2/c$ for all $t \geq n$. For $p \in \{\alpha, 1\}$, convexity of $x \mapsto x^{-p}$ implies

$$(t+1)^{-p} \geq t^{-p}(1 - p/t) \geq t^{-p} \left(1 - \frac{c}{2}t^{-\alpha}\right).$$

Set

$$f_t = \frac{2q}{c}t^{-\alpha} + \frac{2h}{c}t^{-1} = C_\xi t^{-\alpha} + C_b t^{-1}.$$

The preceding inequality shows that

$$f_{t+1} \geq (1 - ct^{-\alpha})f_t + qt^{-2\alpha} + ht^{-\alpha-1}.$$

Let $P_{n,n} = 1$ and $P_{n,t} = \prod_{j=n}^{t-1} (1 - cj^{-\alpha})$ for $t > n$. The factors lie in $[1/2, 1]$. Comparing with (A.5) and inducting from $t = n$ yields

$$a_t \leq a_n P_{n,t} + f_t.$$

Finally, $1 - x \leq e^{-x}$ and an integral comparison give

$$P_{n,t} \leq \exp\left(-c \sum_{j=n}^{t-1} j^{-\alpha}\right) \leq \exp\left[-\frac{c}{1-\alpha}(t^{1-\alpha} - n^{1-\alpha})\right].$$

This proves (3.5), with the constants in (3.6), and hence mean-square convergence.

Almost-sure convergence. Define the nonnegative $\mathcal{F}_t$ -measurable remainder

$$u_t = \frac{3\eta_t}{\mu} \|b_t\|^2 + \sigma^2 \eta_t^2.$$

Because $\alpha > 1/2$ and (A.1) holds,

$$\mathbb{E} \sum_{t=n}^{\infty} u_t \leq \frac{3DB}{\mu} \sum_{t=n}^{\infty} t^{-\alpha-1} + D^2\sigma^2 \sum_{t=n}^{\infty} t^{-2\alpha} < \infty.$$

Tonelli's theorem therefore gives $\sum_{t=n}^{\infty} u_t < \infty$ almost surely. Applying the Robbins–Siegmund almost-supermartingale theorem [11] to (A.3), V_t converges to a finite nonnegative limit and

$$\sum_{t=n}^{\infty} \eta_t V_t < \infty \quad \text{almost surely.}$$

Since $\sum_t \eta_t = \infty$, a positive limit of V_t would contradict this summability. Thus $V_t \rightarrow 0$ almost surely, completing the proof. $\square$

A.2 Proof of Theorem 2

Proof. Throughout the proof, write $\|X\|_{L^2} = (\mathbb{E}\|X\|^2)^{1/2}$. We first prove the classification, which only uses (3.8), and then establish the PR central limit theorem.

Classification of cumulative errors. Minkowski's inequality yields

$$\|B_T\|_{L^2} \leq \frac{\sqrt{B}}{\sqrt{T}} \sum_{t=1}^T t^{-\beta/2}. \quad (\text{A.6})$$

Since $\|\text{Cov}(B_T)\|_{\text{op}} \leq \mathbb{E}\|B_T\|^2$, integral comparisons give

$$\|\Gamma_T^{(b)}\|_{\text{op}} \leq \begin{cases} O(T^{1-\beta}), & 0 < \beta < 2, \\ O(T^{-1} \log^2 T), & \beta = 2, \\ O(T^{-1}), & \beta > 2. \end{cases} \quad (\text{A.7})$$

Thus $B_T \rightarrow 0$ in L^2 for $\beta > 1$, whereas the covariance is bounded for $\beta = 1$.

The remaining possibilities are realized by scalar historical errors. Let U be centered with variance one, measurable at time one, and set $\delta_t = 0$. Each example below has the form $b_t = w_t U$, obtained by using only the first innovation in the historical representation. For $0 < \beta < 1$, the choices $w_t = 0$ and $w_t = t^{-1/2}$ give limiting cumulative variances zero and four, respectively, and both satisfy (3.8). Taking $w_t = t^{-\beta/2}$ gives

$$\text{Var}(B_T) \sim \frac{T^{1-\beta}}{(1-\beta/2)^2} \rightarrow \infty.$$

For $\beta = 1$, take $w_t = t^{-1/2}[1 + \sin(\log t)]$. Integral comparison with a bounded sum-integral error gives

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T w_t = 2 + \text{Im} \left(\frac{T^i}{1/2 + i} \right) + o(1).$$

Its square is bounded and does not converge. This example also satisfies (3.8) for every $0 < \beta < 1$. These examples establish the possibilities in the table; they are not assertions that all of the additional CLT conditions hold.

Mean-square control for a general exponent. Now assume all the PR conditions. Independence of Z_{t+1} from $\mathcal{F}_t$ implies the gradient-noise conditions of Theorem 1, with $\sigma^2 = \text{tr}(S)$. The conditional contraction (A.3) only requires adaptedness of b_t . Inserting (3.8) gives, for all sufficiently large t ,

$$a_{t+1} \leq (1 - ct^{-\alpha})a_t + qt^{-2\alpha} + ht^{-\alpha-\beta}, \quad c = \frac{\mu D}{2}, \quad q = D^2 \text{tr}(S), \quad h = \frac{3DB}{\mu}.$$

For $t^{1-\alpha} \geq 2 \max\{\alpha, \beta\}/c$ and $p \in \{\alpha, \beta\}$,

$$(t+1)^{-p} \geq t^{-p}(1 - p/t) \geq t^{-p} \left(1 - \frac{c}{2} t^{-\alpha}\right).$$

The comparison argument from Appendix A.1, now with $f_t = (2q/c)t^{-\alpha} + (2h/c)t^{-\beta}$, therefore proves

$$a_t = O(t^{-\alpha} + t^{-\beta}). \quad (\text{A.8})$$

All finite initial moments are controlled as before, since $\mathbb{E}\|b_t\|^2 \leq B$. The same almost-supermartingale argument also gives almost-sure convergence, since $\alpha + \beta > 1$ under $\beta > \max\{1/2, 2\alpha - 1\}$.

PR linearization. Assumption 1(b) and the integral form of Taylor's formula give, with $e = \theta - \theta^*$,

$$\begin{aligned} \|\nabla F(\theta) - He\| &= \left\| \int_0^1 [\nabla^2 F(\theta^* + ue) - H]e \, du \right\| \\ &\leq \int_0^1 L_H u \|e\|^2 \, du = \frac{L_H}{2} \|e\|^2. \end{aligned} \quad (\text{A.9})$$

Thus the quadratic remainder constant is $M = L_H/2$.

Write $e_t = \theta_t - \theta^*$ and $\rho_t = \nabla F(\theta_t) - He_t$. The update gives the exact identity

$$H \frac{1}{\sqrt{T}} \sum_{t=1}^T e_t = A_T - \frac{1}{\sqrt{T}} \sum_{t=1}^T (\xi_{t+1} + b_t) - \frac{1}{\sqrt{T}} \sum_{t=1}^T \rho_t, \quad (\text{A.10})$$

where summation by parts yields

$$A_T = \frac{1}{\sqrt{T}} \sum_{t=1}^T \frac{e_t - e_{t+1}}{\eta_t} = \frac{1}{\sqrt{T}} \left[\frac{e_1}{\eta_1} - \frac{e_{T+1}}{\eta_T} + \sum_{t=2}^T (\eta_t^{-1} - \eta_{t-1}^{-1}) e_t \right].$$

By (A.8), $\|e_t\|_{L^2} = O(t^{-\alpha/2} + t^{-\beta/2})$. Also $\eta_t^{-1} - \eta_{t-1}^{-1} \leq Ct^{\alpha-1}$ for $t \geq 2$, where C depends only on D, α . Consequently,

$$\|A_T\|_{L^2} \leq O(T^{-1/2}) + O\left(T^{(\alpha-1)/2} + T^{\alpha-(1+\beta)/2}\right) + \frac{C}{\sqrt{T}} \sum_{t=2}^T \left(t^{\alpha/2-1} + t^{\alpha-\beta/2-1}\right) \longrightarrow 0. \quad (\text{A.11})$$

Indeed, the first power sum is $O(T^{\alpha/2})$, and the second is $O(T^{\alpha-\beta/2})$ if $\beta < 2\alpha$, $O(\log T)$ if $\beta = 2\alpha$, and $O(1)$ if $\beta > 2\alpha$. The resulting terms vanish because $\alpha < 1$ and $\beta > 2\alpha - 1$. Moreover, (A.9) and (A.8) imply

$$\mathbb{E} \left\| \frac{1}{\sqrt{T}} \sum_{t=1}^T \rho_t \right\| \leq \frac{M}{\sqrt{T}} \sum_{t=1}^T a_t = \frac{1}{\sqrt{T}} \sum_{t=1}^T O(t^{-\alpha} + t^{-\beta}) \longrightarrow 0,$$

since $\alpha, \beta > 1/2$. Hence (A.10) becomes

$$\sqrt{T}(\bar{\theta}_T - \theta^*) = -H^{-1} \frac{1}{\sqrt{T}} \sum_{t=1}^T (\xi_{t+1} + b_t) + o_p(1). \quad (\text{A.12})$$

The index shift in the noise is harmless: its difference from the sum of ξ_t is $(\xi_{T+1} - \xi_1)/\sqrt{T}$, which converges to zero in L^2 by the finite second moment assumption.

Joint triangular-array central limit theorem. Exchange the finite historical sums and use (3.8) to get

$$B_T = \frac{1}{\sqrt{T}} \sum_{s=1}^T K_{T,s} \varepsilon_s + o_p(1).$$

Define the independent triangular-array summands

$$Y_{T,s} = \frac{1}{\sqrt{T}} \begin{pmatrix} \xi_s \\ K_{T,s} \varepsilon_s \end{pmatrix}.$$

Their total covariance converges, by (3.10), to

$$\sum_{s=1}^T \mathbb{E}[Y_{T,s} Y_{T,s}^\top] \longrightarrow \begin{pmatrix} S & C_0 K^\top \\ K C_0^\top & \Gamma \end{pmatrix}. \quad (\text{A.13})$$

For completeness, the finite-second-moment assumption and the maximal weight condition imply Lindeberg even if Q is singular. If Q has positive rank, choose a full-column-rank matrix P with $Q = PP^\top$. There are i.i.d. centered vectors U_s with identity covariance and finite second moment such that $Z_s = PU_s$ almost surely: take $U_s = P^\dagger Z_s$, where $P^\dagger$ is the left pseudoinverse. The equality follows because the component of Z_s orthogonal to the range of P has zero second moment. Write

$$Y_{T,s} = A_{T,s}U_s, \quad A_{T,s} = \frac{1}{\sqrt{T}} \begin{pmatrix} I_p & 0 \\ 0 & K_{T,s} \end{pmatrix} P.$$

Then $m_T := \max_s \|A_{T,s}\|_{\text{op}} \rightarrow 0$ and $\sum_s \|A_{T,s}\|_F^2 = O(1)$ by (A.13). For any $\epsilon > 0$,

$$\sum_{s=1}^T \mathbb{E} \left[\|Y_{T,s}\|^2 \mathbf{1}_{\{\|Y_{T,s}\| > \epsilon\}} \right] \leq \left(\sum_{s=1}^T \|A_{T,s}\|_F^2 \right) \mathbb{E} [\|U_1\|^2 \mathbf{1}_{\{\|U_1\| > \epsilon/m_T\}}] \rightarrow 0.$$

Rows with $m_T = 0$ are identically zero. If $Q = 0$, the whole array is identically zero. The multivariate Lindeberg–Feller theorem therefore gives the joint Gaussian limit with covariance (A.13). Adding its two blocks yields

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T (\xi_{t+1} + b_t) \Rightarrow N(0, S + \Gamma + C_0 K^\top + K C_0^\top).$$

Combining this with (A.12) and Slutsky’s theorem proves (3.11)–(3.12).

Finally, if $\beta > 1$, (A.6) gives $B_T \rightarrow 0$ in L^2 . The ordinary i.i.d. CLT for the gradient noise and (A.12) then yield the standard covariance $H^{-1}SH^{-\top}$ without requiring (3.10). $\square$

A.3 Derivations for the examples

Running average. Independence and centering of ε_s give $\mathbb{E}[b_t b_t^\top] = Q/t$. Set $k_{T,s} = \sum_{t=s}^T t^{-1}$. Exchanging sums yields

$$\sum_{s=1}^T k_{T,s} = \sum_{t=1}^T \frac{1}{t} \sum_{s=1}^t 1 = T.$$

For the squared weights, counting the innovations common to times t and u gives

$$\begin{aligned} \sum_{s=1}^T k_{T,s}^2 &= \sum_{t=1}^T \sum_{u=1}^T \frac{\min(t, u)}{tu} \\ &= h_T + 2 \sum_{u=2}^T \frac{u-1}{u} = 2T - h_T. \end{aligned}$$

This proves (3.14). Since $h_T = O(\log T)$, all weight conditions of Theorem 2 hold and (3.15) follows.

Recursive representation and product estimates. Fix $c > 1/2$. Iteration of (3.16) proves (3.17). Write $W_{t,s} = w_{t,s}I$ and $P_{t,s} = \prod_{j=s+1}^t (1 - c/j)$. Choose a fixed integer $m > 2c + 1$. Uniformly for $t \geq s \geq m$, Taylor expansion of the logarithm and a harmonic-sum comparison give

$$\log P_{t,s} = -c \sum_{j=s+1}^t \frac{1}{j} + O\left(\sum_{j=s+1}^t \frac{1}{j^2}\right) = -c \log(t/s) + O(s^{-1}).$$

The constants in the O terms depend only on c . Hence

$$P_{t,s} = (s/t)^c (1 + O(s^{-1})), \quad |w_{t,s}| \leq C s^{c-1} t^{-c}. \quad (\text{A.14})$$

The finitely many indices $s < m$ satisfy the same upper bound after increasing C : their products factor into a fixed finite prefix and $P_{t,m}$. Possible zero factors cause no difficulty. The same argument gives $\|\delta_t\| = O(t^{-c})$. Thus

$$\left\| \frac{1}{\sqrt{T}} \sum_{t=1}^T \delta_t \right\| \leq \begin{cases} O(T^{1/2-c}), & 1/2 < c < 1, \\ O(T^{-1/2} \log T), & c = 1, \\ O(T^{-1/2}), & c > 1, \end{cases}$$

which tends to zero.

Single-step covariance. The initial term is deterministic, so

$$\mathbb{E}[b_t b_t^\top] = \left(\sum_{s=1}^t w_{t,s}^2 \right) Q + \delta_t \delta_t^\top.$$

We show that $t \sum_{s=1}^t w_{t,s}^2 \rightarrow c^2/(2c-1)$. For every fixed $\epsilon \in (0, 1)$, (A.14) holds with relative error $O((\epsilon t)^{-1})$ uniformly for $\epsilon t \leq s \leq t$. Consequently, a Riemann sum gives

$$t \sum_{s=\lceil \epsilon t \rceil}^t w_{t,s}^2 \longrightarrow c^2 \int_{\epsilon}^1 x^{2c-2} dx.$$

For the omitted indices, the product bound gives

$$t \sum_{s < \epsilon t} w_{t,s}^2 \leq C t^{1-2c} \sum_{s < \epsilon t} s^{2c-2} \leq C' \epsilon^{2c-1}.$$

The last bound follows from $2c-2 > -1$; an empty sum is zero. Letting $\epsilon \downarrow 0$ yields

$$t \sum_{s=1}^t w_{t,s}^2 \longrightarrow c^2 \int_0^1 x^{2c-2} dx = \frac{c^2}{2c-1}.$$

The initial term satisfies $t \|\delta_t\|^2 \rightarrow 0$, proving (3.18). The bounds also give

$$\sup_{t \geq 1} t \mathbb{E} \|b_t\|^2 < \infty.$$

Thus the direct error condition holds with $\beta = 1$. The weight and remainder bounds in Theorem 1 likewise hold for finite constants.

Cumulative kernel and boundary control. Write $K_{T,s} = k_{T,s} I$, where $k_{T,s} = \sum_{t=s}^T w_{t,s}$. For $s/T \rightarrow x > 0$, (A.14) and a Riemann-sum approximation imply

$$k_{T,s} = \sum_{t=s}^T \frac{c}{s} (s/t)^c (1 + O(s^{-1})) \longrightarrow c x^{c-1} \int_x^1 v^{-c} dv = k_c(x).$$

The convergence is uniform when $s/T \in [\epsilon, 1]$ for fixed $\epsilon > 0$. To justify passage to the first and second moment integrals, including near $x = 0$, sum the product bound to obtain

$$|k_{T,s}| \leq C \begin{cases} (T/s)^{1-c}, & 1/2 < c < 1, \\ 1 + \log(T/s), & c = 1, \\ 1, & c > 1. \end{cases} \quad (\text{A.15})$$

Thus the small-index contribution to $T^{-1} \sum_s k_{T,s}^2$ has limsup bounded by $C\epsilon^{2c-1}$ when $c < 1$, by $C \int_0^\epsilon [1 + \log(1/x)]^2 dx$ when $c = 1$, and by $C\epsilon$ when $c > 1$. Each bound tends to zero with ϵ . Cauchy–Schwarz gives the corresponding vanishing bound for the first moments. Combining these estimates with the uniform interior convergence proves

$$\frac{1}{T} \sum_{s=1}^T k_{T,s} \rightarrow \int_0^1 k_c(x) dx, \quad \frac{1}{T} \sum_{s=1}^T k_{T,s}^2 \rightarrow \int_0^1 k_c(x)^2 dx.$$

The envelope also gives

$$\max_{s \leq T} \frac{|k_{T,s}|}{\sqrt{T}} \leq \begin{cases} O(T^{1/2-c}), & 1/2 < c < 1, \\ O(T^{-1/2} \log T), & c = 1, \\ O(T^{-1/2}), & c > 1, \end{cases} \rightarrow 0.$$

This verifies the maximal weight condition, including the fixed early innovations.

Evaluation of the covariance integrals. For $c \neq 1$, direct integration of (3.19) gives

$$\begin{aligned} \int_0^1 k_c(x) dx &= \frac{c}{c-1} \left(1 - \frac{1}{c}\right) = 1, \\ \int_0^1 k_c(x)^2 dx &= \frac{c^2}{(c-1)^2} \left(1 - \frac{2}{c} + \frac{1}{2c-1}\right) = \frac{2c}{2c-1}. \end{aligned}$$

For $c = 1$, the corresponding integrals are $\int_0^1 \log(1/x) dx = 1$ and $\int_0^1 \log^2(1/x) dx = 2$. This establishes (3.20) in every case. Applying Theorem 2 with $C_0 = 0$ proves (3.21).

Finally, differentiating (3.22) proves the stated parameter comparison. Since $H^{-1}QH^{-\top}$ is positive semidefinite, the scalar monotonicity of f_{PR} implies monotonicity of the downstream covariance in the positive semidefinite order. If $Q \neq 0$, its trace strictly decreases, whereas the single-step asymptotic MSE has its unique minimum at $c = 1$.

A.4 Proof of Proposition 3

Proof. Independence and centering give

$$\Gamma_T = \frac{Q}{T} \sum_{s=1}^T k_{T,s}^2.$$

Normalization and exchange of the finite sums imply

$$\sum_{s=1}^T k_{T,s} = \sum_{t=1}^T \sum_{s=1}^t w_{t,s} = T.$$

Therefore

$$\frac{1}{T} \sum_{s=1}^T (k_{T,s} - 1)^2 = \frac{1}{T} \sum_{s=1}^T k_{T,s}^2 - 1,$$

which proves (4.2) and sufficiency in (4.3). If $Q \neq 0$, taking traces gives

$$\text{tr}(\Gamma_T - Q) = \text{tr}(Q) \frac{1}{T} \sum_{s=1}^T (k_{T,s} - 1)^2,$$

so the condition is necessary as well. Finally, if the PR covariance is $H^{-1}(S + \Gamma)H^{-\top}$ with $\Gamma_T \rightarrow \Gamma$, closedness of the positive semidefinite cone gives $\Gamma \succeq Q$. Congruence by H^{-1} proves (4.4). $\square$

A.5 Proof of Proposition 4

Proof. Set $\gamma_1 = 1$. Expanding (4.5) gives

$$w_{t,s} = \gamma_s \prod_{j=s+1}^t (1 - \gamma_j), \quad 1 \leq s \leq t.$$

Since $0 < \gamma_t \leq 1$, the weights are nonnegative. Their row sum is one at $t = 1$ and remains one by induction, because the next row has sum $(1 - \gamma_t) \cdot 1 + \gamma_t = 1$.

Single-step error. To control the deterministic weights without imposing $Q \neq 0$, let u_s be independent centered scalar random variables of variance one, and define the auxiliary scalar recursion

$$z_1 = u_1, \quad z_t = (1 - \gamma_t)z_{t-1} + \gamma_t u_t.$$

Then $z_t = \sum_s w_{t,s} u_s$ and $v_t := \mathbb{E} z_t^2 = \sum_s w_{t,s}^2$ satisfies

$$v_t = (1 - \gamma_t)^2 v_{t-1} + \gamma_t^2 \leq (1 - at^{-\rho})v_{t-1} + a^2 t^{-2\rho}. \quad (\text{A.16})$$

For completeness, $v_t \leq Ct^{-\rho}$ follows by a direct comparison. For all sufficiently large t ,

$$(t-1)^{-\rho} - t^{-\rho} \leq C_1 t^{-\rho-1} \leq \frac{a}{2} t^{-2\rho},$$

where the last inequality uses $\rho < 1$. If $v_{t-1} \leq C(t-1)^{-\rho}$, then (A.16) implies

$$v_t \leq Ct^{-\rho} + \left(-\frac{Ca}{2} + a^2\right) t^{-2\rho} \leq Ct^{-\rho}$$

for $C \geq 2a$. Increase C to cover the finite initial interval and apply induction. Consequently,

$$\mathbb{E} \|b_t\|^2 = \text{tr}(Q)v_t = O(t^{-\rho}).$$

The rate is genuinely of order $t^{-\rho}$ when $Q \neq 0$. Indeed, for the last $\lfloor t^\rho \rfloor$ indices $s \leq t$, one has $\gamma_s \geq at^{-\rho}$ and, for sufficiently large t ,

$$\prod_{j=s+1}^t (1 - \gamma_j) \geq \exp\left(-2 \sum_{j=s+1}^t \gamma_j\right) \geq c_* > 0,$$

where c_* is independent of t, s . Here $\gamma_j \leq 1/2$ on that interval, and its length is $O(t^\rho)$ with $j \asymp t$. Thus $v_t \geq c_*^2 a^2 t^{-2\rho} \lfloor t^\rho \rfloor$, proving $v_t = \Theta(t^{-\rho})$.

Cumulative weights approach one. The key estimate follows from summation by parts, without approximating the individual historical products. Define $d_t = \gamma_t^{-1} - 1$, so $d_1 = 0$. Rearranging the scalar recursion gives

$$u_t = z_t + d_t(z_t - z_{t-1}),$$

where z_0 can be set to zero since $d_1 = 0$. Summing yields

$$\sum_{t=1}^T (z_t - u_t) = -d_T z_T + \sum_{t=1}^{T-1} (d_{t+1} - d_t) z_t. \quad (\text{A.17})$$

For $t \geq 2$, $d_t = a^{-1}t^\rho - 1$, so $d_t = O(t^\rho)$ and $d_{t+1} - d_t = O(t^{\rho-1})$; the first increment is finite. Using $\|z_t\|_{L^2} = O(t^{-\rho/2})$ and Minkowski's inequality,

$$\begin{aligned} \left\| \frac{1}{\sqrt{T}} \sum_{t=1}^T (z_t - u_t) \right\|_{L^2} &\leq O(T^{(\rho-1)/2}) + O(T^{-1/2}) + \frac{C}{\sqrt{T}} \sum_{t=2}^{T-1} t^{\rho/2-1} \\ &= O(T^{(\rho-1)/2}). \end{aligned}$$

On the other hand, exchanging sums gives the exact identity

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T (z_t - u_t) = \frac{1}{\sqrt{T}} \sum_{s=1}^T (k_{T,s} - 1) u_s.$$

Independence and unit variance therefore imply

$$\frac{1}{T} \sum_{s=1}^T (k_{T,s} - 1)^2 = \mathbb{E} \left| \frac{1}{\sqrt{T}} \sum_{t=1}^T (z_t - u_t) \right|^2 = O(T^{\rho-1}) \longrightarrow 0. \quad (\text{A.18})$$

This is a deterministic weight estimate, valid for any innovation covariance Q . Proposition 3 now gives

$$\Gamma_T = Q[1 + O(T^{\rho-1})] \longrightarrow Q.$$

Downstream PR limit. Here $W_{t,s} = w_{t,s}I$, $\delta_t = 0$, and $\beta = \rho$ in Theorem 2. Normalization gives $T^{-1} \sum_s K_{T,s} = I$ exactly, and the covariance limit is Q . Moreover, (A.18) implies

$$\max_{s \leq T} \frac{|k_{T,s}|}{\sqrt{T}} \leq T^{-1/2} + \left(\frac{1}{T} \sum_{s=1}^T (k_{T,s} - 1)^2 \right)^{1/2} \longrightarrow 0.$$

Thus every cumulative-weight condition holds. Since $\rho > 1/2$, the remaining exponent condition is exactly $\rho > 2\alpha - 1$. Theorem 2 with $C_0 = 0$ proves (4.7). $\square$