SUMMARY OF AVERAGING TRAJECTORIES OF STOCHASTIC APPROXIMATION: PROOF OF ASYMPTOTIC NORMALITY OF THE ESTIMATES AND ALMOST SURE CONVERGENCE

ROBERT DAM, YIXING LU, AND SANSKRUTI MORE

Abstract. This report summarises the classical result of the asymptotic normality of Ruppert-Polyak averaging in stochastic approximation. We first discuss the benefits of stochastic approximation and the gains to be had from averaging. Then we discuss the consequences of the normality result like its importance for using stochastic approximation in inference. Finally, we provide a summary of the key ideas in proving asymptotic normality before showing the full proof.

1. Introduction

In this project, we summarise the main results, consequences, and contributions of *Acceleration of Stochastic Approximation by Averaging* by Polyak and Judistsky [PJ92]. This paper proves the asymptotic normality of the average of the iterates of stochastic gradient descent (SGD) in an algorithm called the Polyak-Rupert averaged SGD algorithm. This paper is a classical result and one of the first to address SGD from an inference viewpoint rather than an optimal convergence rate viewpoint. Having inferential results for SGD is important for using SGD in statistical settings as a replacement for the maximum likelihood estimator (MLE). Additionally, this average of iterates as the estimated optimum has benefits over the simple last iterate. It provides stability to SGD, has a faster convergence rate, is asymptotically normal, and is asymptotically efficient. These benefits makes it important to better understand the distribution of such an estimate.

Stochastic gradient descent and its variants is one of the most popular optimization algorithms for many real world applications due to its simplicity, efficiency, and online properties. However, since SGD is a random algorithm, its results are subject to randomness as well. This can be a concern where we cannot fully trust the results of SGD without understanding its random nature. Having an asymptotic distribution for this algorithm is important since it allows us to have inference and uncertainty quantifications for its results.

We first describe the algorithm itself before presenting the main theorems. Then we mainly showed the proof of the almost sure convergence and asymptotic normality of the process under the more general **nonlinear problem** setting, first providing a high level overview of the key points, followed by detailed proofs. Some technical details will be left to the appendix. The proof demonstrates techniques for working with stochastic iterative processes. It requires techniques using the Martingale Central Limit theorem, techniques to prove almost sure convergence, and techniques to decompose iterative algorithms.

2. Algorithm Description

This stochastic approximation algorithm is iterative algorithm that seeks to find the point x^* such that for some function $R(x) : \mathbb{R}^n \to \mathbb{R}^n$, $R(x^*) = 0$. In this setup, our problem

is a root finding one, but in the most common use-case of stochastic approximation is to optimize functions by finding points where the gradient is 0.

In this scenario, we have access to the function values of R, but finding the point x^* directly is typically difficult. To approximate this point, the you have access to a sequence of observations $(y_t)_{t\geq 1}$, where $y_t = R(x_{t-1}) + \xi_t$ is defined as the prediction residual from the previous iteration plus a random error term. The following recursive algorithm is used to estimate x^*:

$$\begin{aligned} x_t &= x_{t-1} - \gamma_t y_t \\ \bar{x}_t &= \frac{1}{t}\sum_{i=0}^{t-1} x_i,\ x_0 \in \mathbb{R}^n \end{aligned} \tag{1}$$

where γ_t is a step size term and x_0 is an initial point.

Recursively, at the t^{th} iteration, x^* is estimated by an average of the previous $t-1$ estimates. Intuitively, averaging will lessen the variability at each iteration to provide stability to the estimate. What this algorithm does is that it starts at an initial point x_0 and descends randomly in small steps towards the direction of the function evaluated at the last step. The idea is that at points where the function is at a higher absolute value you change values quicker to reach the point where your function is 0, but at points where your function is close to 0 you move more slowly.

Note that this is an online algorithm since each iteration only uses the value of one observation point which means we do not need to keep all observations simultaneously in memory. The average itself can be updated by keeping a running sum, so this algorithm is quite efficient at each step. These efficiency factors are one reason why stochastic approximation is so popular in practice.

3. Main Results

Theorem 2: Nonlinear Case

Assumptions

2.1) There exists a function $V(x) : \mathbb{R}^n \to \mathbb{R}$ such that for some $\lambda > 0,\ \alpha > 0,\ \varepsilon > 0,\ L > 0$ and all $x, y \in \mathbb{R}^n$ that the conditions $V(x) \geq \alpha\|x\|^2$, $|\nabla V(x) - \nabla V(y)| \leq L\|x-y\|$, $V(x^*) = 0$, $\nabla V(x-x^*)^T R(x) > 0$ for $x \neq x^*$ hold true. Also, $\nabla V(x-x^*)^T R(x) \geq \lambda V(x)$ for all $\|x - x^*\| \leq \varepsilon$.

2.2) There exists a matrix $G \in \mathbb{R}^n \times \mathbb{R}^n$ and $K_1 < \infty, \varepsilon > 0, 0 < \lambda \leq 1$ such that

$$\|R(x) - G(x-x^*)\| \leq K_1\|x-x^*\|^{1+\lambda}$$

for all $\|x - x^*\| \leq \varepsilon$ and $Re\ \lambda_i(G) > 0$, $i = 1, ..., N$.

2.3) Let $(\xi_t)_{t\geq 1}$ be a martingale difference process defined on the probability space $(\Omega, \mathscr{F}, \mathscr{F}_t, P)$. Then for some K, we have $E(|\xi_t|^2|\mathscr{F}_{t-1}) + |R(x_{t-1})|^2 \leq K(1+|x_{t-1}|^2)\ a.s$ for all $t \geq 1$. Also, in the following decomposition:

$$\xi_t = \xi_t(0) + \zeta_t(x_{t-1}), \tag{2}$$

$\xi_t(0)$ converges to zero a.s. and it's covariance converges to $S > 0$ in probability, and $\sup_t E(|\xi_t(0)|^2 I(|\xi_t(0)| > C)|\mathscr{F}_{t-1}) \xrightarrow{p} 0$ as $C \to \infty$. And for large enough t, $E(|\zeta_t(x_{t-1})|^2|\mathscr{F}_{t-1}) \leq \delta(x_{t-1})\ a.s.$, with $\delta(x) \to 0$ as $x \to 0$.

2.4) It holds that $(\gamma_t - \gamma_{t+1})/\gamma_t = o(\gamma_t)$, $\gamma_t > 0$ for all t; and $\sum_{t=1}^{\infty}(1+\lambda)/\gamma_t^2 t^{-1/2} < \infty$

Proposition If Assumptions 2.1-2.4 are satisfied, then $\bar{x}_t \xrightarrow{a.s.} x^*$ and

$$\sqrt{t}(\bar{x}_t - x^*) \xrightarrow{D} N(0, V), \text{ where } V = G^{-1}S(G^{-1})^T$$

Discussion: Theorem 2 shows that under certain assumptions on our target function, the averaged iterates from the stochastic approximation process is asymptotically normal centered on the true root value. Namely, we must assume the existence of a two functions related to $R(x)$. The scalar valued Lyapunov function $V(x)$ provides a way to prove the convergence of nonlinear processes due to its properties and by assuming there exists a function with information of the level curves of our nonlinear process. The linear function G serves as a local linear approximation to $R(x)$ in ways stated in Assumption 2.2.

Additionally, this result assumes that the step sizes γ_t decrease at a suitable rate that is both not too slow or too rapid. For γ_t of the form $\gamma t^{-\alpha}$, $\frac{1}{2} < \alpha < 1$ would satisfy the requirements of Assumption 2.4 for a suitable γ. In practice, the tuning of the step size is a difficult and crucial problem. Finally, the result also assumes a certain structure on the random error terms ξ_t. The martingale difference process assumption is weaker than the i.i.d. assumption where in this case we only need $E(\xi_t|\mathscr{F}_{t-1}) = 0$ and finite conditional variances. These assumptions are the weakest found for this problem and offer not much issue when it comes to practical use of this result.

There are two major consequences of this result. First, we have almost sure convergence of $\bar{x}_t$ to x^* which is the strongest convergence for random variables. This means that under the right conditions, our estimate will asymptotically be the true value we want. Next, we have an asymptotic distribution for our random estimator, which provides some level of inference for us to perform our estimate which is important for random estimators. More recent research has looked into non-asymptotic normality results [BM11] [ABE19].

Note that the variance of the asymptotic distribution must be estimated for the construction of approximate confidence regions for x^*. One quick way to do this is through the bootstrap. Bootstrapping is possible in smaller data situations, but it does remove the benefits of the online nature of SGD and it becomes computationally expensive with high dimensional data. Other research has found ways to construct confidence regions while maintaining the computational benefits of SGD [SZ18] [LLKC17] [FXY17].

These inference focused results are important for application in statistical settings. Oftentimes, when one cannot find a close-form solution to the maximum likelihood estimator, one would need to resort to some iterative optimization algorithm. However, properties of the MLE are often not established for their iterative estimates. The asymptotic normality of the averaged stochastic approximation estimates circumvents this issue because we can directly use its statistical properties for statistical inferences. Furthermore, SGD is online in nature and efficient which makes it well suited to high dimensional and big data situations. SGD has much practical use on many optimization problems due to its flexibility, and these additional statistical properties make it a stronger practical option when close-form MLE is hard to find.

4. Key Proof Ideas

To prove the asymptotic normality of the normalized error $\sqrt{t}\bar{\Delta}_t = \sqrt{t}(\bar{x}_t - x^*)$, it is first shown that asymptotic normality holds for the linear case. Then this result is used to prove the nonlinear case.

To prove the linear case, the error of the iterative process can be decomposed to three terms (Lemma 2):

$$\begin{aligned}\sqrt{t}\bar{\Delta}_t &= \frac{1}{\sqrt{t}\gamma_0}\alpha_t\Delta_0 + \frac{1}{\sqrt{t}}\sum_{j=1}^{t-1} A^{-1}\xi_j + \frac{1}{\sqrt{t}}\sum_{j=1}^{t-1} w_j^{(t)}\xi_j \\ &= I^{(1)} + I^{(2)} + I^{(3)}\end{aligned}$$

$I^{(1)}$ and $I^{(3)}$ are shown to converge to zero as $t \to \infty$ in mean square using results from Lemma 1 and Lemma 2. Then, we demonstrate that $I^{(2)}$ fulfills the Lindeberg condition and that the asymptotic conditional variance of $I^{(2)}$ converges to V by using the linearity of expectation and Assumption 1.5(a). With these two things verified, the central limit theorem for martingales can be employed for $I^{(2)}$ which verifies asymptotic normality for the linear case.

For the nonlinear case of Theorem 2, there are two parts: the first part proves that the non-linear iterative process converges almost surely; and the second part shows that the normalized error is also asymptotically normal like the linear process in Theorem 1.

The Lyapunov's second method is used to prove convergence of the non-linear iterative process. A non-negative scalar function V(x) (Lyapunov function) is defined and examined on a series of iterative values $\Delta_t = x_t - x^*$. If $V(\Delta_t)$ converges to some bounded value, then the assumptions on V lead to the convergence of the process if we can verify some additional conditions. Specifically, the convergence of $V(\Delta_t)$ leads us to the fact that for every $\epsilon > 0$, there exists some $R < \infty$ such that

$$P(\sup_t |\Delta_t| \leq R \; for \; all \; t \geq \tau_R) \geq 1 - \epsilon \tag{3}$$

where the stopping time τ_t is defined as $\tau_t = \inf\{t \geq 1 : |\Delta_t| > R\}$. Then we further demonstrate that $\lim_{t\to\infty} P(\tau_R > t) \to 0$, which shows that t will be asymptotically greater than the stopping time τ_R so that the probability of deviations of Δ_t from 0 is 0. This completes the proof of the almost sure convergence of $\bar{x}_t - x^*$.

To prove that $\sqrt{t}(\bar{x}_t - x^*)$ is asymptotically normal for the non-linear process, we first define a linear process by the following conditions:

$$\begin{aligned}\Delta_t^1 &= \Delta_{t-1}^1 - \gamma_t G\Delta_{t-1}^1 + \gamma_t\xi_t, \\ \bar{\Delta}_t^1 &= \frac{1}{t}\sum_{i=0}^{t-1}\Delta_i^1\end{aligned} \tag{4}$$

Note here G is a linear transformation, unlike $R(x)$ which is non-linear.

We first show that this linear process fulfills all the conditions needed to prove the asymptotic normality as shown in Theorem 1 proposition (a). Then, we demonstrate that the non-linear process is asymptotically identical to the linear process, thus itself is also asymptotically normal.

5. Full Proof

We now present the proof of the main result: Theorem 2. We first present the proof of Theorem 1 proposition (a) which shows asymptotic normality for the linear stochastic approximation process. This result is needed for the proof of the nonlinear case in Theorem 2. Additional propositions (b) and (c) of Theorem 1 and their proofs are left to the appendix.

Throughout this paper, K will represent asymptotically unimportant constants.

Assumptions of Theorem 1

1.1) The matrix -A is Hurwitz, i.e., $Re\lambda_i(A) > 0$.
1.2) The step size coefficients $\gamma_t > 0$ is either a constant ($0 < \gamma < 2(min_i Re\lambda_i(A))^{-1}$), or it goes to zero as $t \to \infty$ at a sufficiently slow rate ($\frac{\gamma_t - \gamma_{t+1}}{\gamma_t}) = o(\gamma_t)$).
1.3) $(\xi_t)_{t\geq 1}$ is a martingale-difference process, defined on the probability space $(\Omega, \mathscr{F}, \mathscr{F}_t, P)$.
1.4) $lim_{C\to\infty} \limsup_{t\to\infty} E(|\xi_t|^2 I(|\xi_t| > C|\mathscr{F}_{t-1})) \overset{p}{=} 0$.
1.5) The following hold:
  a. $lim_{t\to\infty} E(\xi_t \xi_t^T|\mathscr{F}_{t-1}) \overset{p}{=} S > 0$;
  b. $lim_{t\to\infty} E(\xi_t \xi_t^T) = S > 0$

First we introduce two lemmas that will be useful in proving the linear case. The proofs of Lemma 1 and 2 will be left to the appendix.

Lemma 1 Let Assumptions 1.1 and 1.2 hold: Then there is constant $K < \infty$ such that for all j and $t \geq j$:

$$\|\phi_j^{(t)}\| \leq K, \ and \ \lim_{t\to\infty} \frac{1}{t} \sum_{j=0}^{t-1} \phi_j^{(t)} = 0.$$

where $X_j^{(t+1)} = X_j^{(t)} - \gamma_t A X_j^{(t)}$, $X_j^{(j)} = I$, $\bar{X}_j^{(t)} = \gamma_j \sum_{i=j}^{t-1} X_j^{(i)}$, and $\phi_j^{(t)} = A^{-1} - \bar{X}_j^{(t)}$.

Lemma 2 If the statements in Lemma 1 hold, then

$$\sqrt{t}\bar{\Delta}_t = \frac{1}{\sqrt{t}\gamma_0}\alpha_t \Delta_0 + \frac{1}{\sqrt{t}} \sum_{j=1}^{t-1} A^{-1}\xi_j + \frac{1}{\sqrt{t}} \sum_{j=1}^{t-1} w_j^{(t)} \xi_j, \tag{5}$$

where α_t, $w_j^{(t)} \in \mathbb{R}^{N\times N}$ are such that $\|\alpha_t\| \leq K$, $\|w_j^{(t)}\| \leq K \ for \ some \ K < \infty$, and $\frac{1}{t}\sum_{j=1}^{t-1} \|w_j^t\| \to 0 \ as \ t \to \infty$.

Theorem 1 proposition (a): Let assumptions 1.1-1.4 and 1.5(a) be satisfied, then $\sqrt{t}(\bar{x}_t - x^*) \xrightarrow{D} N(0, V)$, where $V = A^{-1}S(A^{-1})^T$.

Proof. From Lemma 2 (5) we can decompose $\sqrt{t}\bar{\Delta}_t$ into a sum of three parts $\sqrt{t}\bar{\Delta}_t = I^{(1)} + I^{(2)} + I^{(3)}$

Now, since $\|\alpha_t\| \leq K$ by Lemma 2, when we take $t \to \infty$ all other parts in $I^{(1)}$ are constant or bounded by constants which means $I^{(1)} \to 0$ in mean square.

By Lemma 2 and the Cauchy-Schwarz inequality we have that for $I^{(3)}$

$$E\left|\frac{1}{\sqrt{t}} \sum_{j=1}^{t-1} w_j^t \xi_j\right|^2 \leq \frac{K}{t} \sum_{j=1}^{t-1} \|w_j^t\|^2 E\|\xi_j\|^2 \leq \frac{K}{t} \sum_{j=1}^{t-1} \|w_j^t\|^2 \leq \frac{K}{t} \sum_{j=1}^{t-1} \|w_j^t\| \to 0 \text{ as } t \to \infty$$

The last inequality comes from the fact that $\frac{1}{\sqrt{t}} \sum_{j=1}^{t-1} \|w_j^t\| \to 0$ as $t \to \infty$ so each $\frac{\|w_j^t\|}{\sqrt{t}}$ must be less than 1 so the square is less than the original.

All this together means that $I^{(3)} \to 0$ as $t \to \infty$.

All that is left to show is that the central limit theorem for martingales holds for $I^{(2)}$. To do this, we must demonstrate the Lindeberg condition and the convergence of the conditional

second moment of $I^{(2)}$ to V. To check the Lindeberg condition we note that for a large enough constant C

$$\limsup_{t\to\infty} \frac{1}{t}\sum_{j=1}^{t-1} E(|A^{-1}\xi_j|^2 I(|A^{-1}\xi_j| > C)|\mathscr{F}_{j-1})$$

$$\leq K^2 \limsup_{t\to\infty} \sum_{j=1}^{t-1} E(|\xi_j|^2 I(|\xi_j| > CK^{-1})|\mathscr{F}_{j-1}) = \mathscr{G}(C)$$

This is true for some K that upper bounds the transformation done by A^{-1} which must be finite since A^{-1} is a linear map. Now, by Assumption 2.4, $\mathscr{G}(C) \xrightarrow{p} 0$ as $C \to \infty$. This means that the Lindeberg condition is fulfilled.

Now, to check the convergence of the variance we evaluate the asymptotic conditional second moment of $I^{(2)}$. By using the linearity of expectation and Assumption 1.5(a) we get

$$\begin{aligned} E(I^{(2)}I^{(2)T}|\mathscr{F}_{j-1}) &= E\left(\frac{1}{t}\sum_{j=1}^{t-1} A^{-1}\xi_j\xi_j^T(A^{-1})^T|\mathscr{F}_{j-1}\right) \\ &= \frac{1}{t}\sum_{j=1}^{t-1} A^{-1}E(\xi_j\xi_j^T|\mathscr{F}_{j-1})(A^{-1})^T \overset{p}{=} A^{-1}S(A^{-1})^T = V \end{aligned} \tag{6}$$

With both conditions verified, we have that $I^{(2)} \xrightarrow{D} N(0, V)$ □

Theorem 2

We now prove Theorem 2. First let us introduce some notation:

$$\begin{aligned} \Delta_t &= x_t - x* \\ y(\Delta_t) &= \bar{R}(\Delta_{t-1}) + \xi_t(\Delta_{t-1}) \\ V_t &= V(\Delta_t) \end{aligned} \tag{7}$$

Part 1 It holds that $V(\Delta_t) \to V(\omega)$, where $V(\omega)$ is bounded.

We introduced a Lyapunov function V(X) which is differentiable on [x,x+y], and $\nabla V(x)$ satisfies a Lipschitz condition on [x,x+y], then we have:

$$\|V(x+y) - V(x) - \nabla V(x)y\| \leq L\|y\|^2/2$$

Using this formula and Assumption 2.1, we have in our case $x = \Delta_{t-1}$, $y = -\gamma_t y(\Delta_t)$, and $V(\Delta_{t-1} - \gamma_t y(\Delta_t)) = V_t$, then it follows that:

$$V_t \leq V_{t-1} - \gamma_t \nabla V_{t-1}^T(\bar{R}(\Delta_{t-1}) + \xi(\Delta_{t-1})) + \frac{L}{2}\gamma_t^2\|\bar{R}(\Delta_{t-1}) + \xi_t(\Delta_{t-1})\|^2 \tag{8}$$

Taking the conditional expectation of both sides of (17), we have:

$$\begin{aligned} \mathbb{E}(V_t|\mathscr{F}_{t-1}) &= V_{t-1} - \gamma_t \nabla V_{t-1}^T \bar{R}(\Delta_{t-1}) - \gamma_t \nabla V_{t-1}^T \mathbb{E}(\xi_t(\Delta_{t-1})|\mathscr{F}_{t-1}) + \frac{L}{2}\gamma_t^2\|\bar{R}(\Delta_{t-1}) + \mathbb{E}(\xi_t(\Delta_{t-1})|\mathscr{F}_{t-1})\|^2 \\ &\leq V_{t-1} - \gamma_t \nabla V_{t-1}^T \bar{R}(\Delta_{t-1}) + K\gamma_t^2(|\Delta_{t-1}|^2 + 1) \\ &\leq V_{t-1}(1 + K\gamma_t^2) + K\gamma_t^2 - \gamma_t \nabla V_{t-1}^T \bar{R}(\Delta_{t-1}) \end{aligned}$$

The first inequality follows from Assumption 2.3 about the martingale-difference process, and the second inequality follows from Assumption 2.1 that implies $V_{t-1} \geq \alpha\|\Delta_{t-1}\|^2$ for some $\alpha > 0$.

When γ_t satisfies the conditions in Assumption 2.4, $\sum_{t=1}^{\infty}\gamma_t^2 < \infty$, so we could obtain from the Robbins-Siegmund theorem that $V_t \to V$ a.s. Also, given $V_{t-1} \geq \alpha\|\Delta_{t-1}\|^2$, $P(sup_t|\Delta_t| < \infty) = 1$. Then, for every $\epsilon > 0$, there exists some $R < \infty$ such that

$$P(sup_t|\Delta_t| \leq R \; for \; all \; t \geq \tau_R) \geq 1 - \epsilon \tag{9}$$

Part 2 Now to show the almost sure convergence of $x \to x*$, it suffices to show that $lim_{t\to\infty}P(\tau_R > t) \to 0$. The stopping time τ_R is defined as $\tau_R = inf\{t \geq 1 : |\Delta_t| > R\}$. From part 1, we have

$$\begin{aligned} E(V_t I(\tau_R > t)|\mathscr{F}_{t-1}) &\leq E(V_t I(\tau_R > t-1)|\mathscr{F}_{t-1}) \\ &\leq [V_{t-1}(1+\gamma_t^2 K) - \alpha\gamma_t V_{t-1}]I(\tau_R > t-1) + \gamma_t^2 K \end{aligned}$$

for some $\alpha,\ K > 0$. The second inequality follows from Assumption 2.1: $\nabla V_{t-1}^T \bar{R}(\Delta_{t-1}) \geq \alpha V_{t-1}$. Taking the expectation of both sides of the inequality, we obtain:

$$E|\Delta_t|^2 I(\tau_R > t) \leq EV_t I(\tau_R > t) \leq EV_{t-1}(1 - \alpha\gamma_t + K\gamma_t^2) + K\gamma_t^2.$$

Finally, by Theorem 24 in [BMP90], we have

$$E|\Delta_t|^2 I(\tau_R > t) \leq E(|\Delta_t|^2)E(I(\tau_R > t)) \leq K\gamma_t P(\tau_R > t) \leq K'\gamma_t \tag{10}$$

Since $\gamma_t \to 0$, $P(\tau_R > t) \to 0$, which completes the proof of almost sure convergence.

Part 3 The following equations described the error $\bar{\Delta}_t$ of the non-linear version of the recursive algorithm: $\Delta_t = \Delta_{t-1} - \gamma_t\bar{R}(\Delta_{t-1}) + \gamma_t\xi_t,\ \bar{\Delta}_t = \frac{1}{t}\sum_{i=0}^{t-1}\Delta_i$. To prove asymptotic normality for the non-linear process, we first define a linear process Δ_t^1 (4), and show this linear process satisfies the conditions of Theorem 1 proposition (a).

We will prove that Δ_t^1 satisfies the Lindeberg condition. The proof that its variance converges is exactly analogous to the proof in Theorem 1 proposition (a). By decomposition (2) in assumption 2.3, we have that

$$I(|\xi_t| > C) \leq I(|\zeta_t(\Delta_{t-1})| > \frac{C}{2}) + I(|\xi_t(0)| > \frac{C}{2});$$

$$\begin{aligned} E(|\xi_t|^2 I(|\xi_t| > C)|\mathscr{F}_{n-1}) &\leq 2E(|\zeta_t(\Delta_{t-1})|^2 I(|\zeta_t(\Delta_{t-1})| > \frac{C}{2})|\mathscr{F}_{t-1}) + 2E(|\xi_t(0)|^2 I(|\xi_t(0)| > \frac{C}{2})|\mathscr{F}_{t-1}) \\ &\leq 2\delta(\Delta_{t-1}) + 2E(|\xi_t(0)|^2 I(|\xi_t(0)| > \frac{C}{2})|\mathscr{F}_{t-1}) = I_1 + I_2 \end{aligned}$$

The last inequality follows from assumption 2.3. $I_1 \to 0$ since $\Delta_t \to 0$; $I_2 \to 0$ as $t \to \infty$ and $C \to \infty$ as stated in assumption 2.3.

Part 4 From Part 3, we have shown that the linear approximation process $\bar{\Delta}_t^1$ satisfies all of the conditions of proposition (a) of Theorem 1. To show that the non-linear process $\bar{\Delta}_t$ also satisfies those conditions we can show that $\bar{\Delta}_t^1$ and $\bar{\Delta}_t$ are asymptotically equivalent. Let $\delta_t = \bar{\Delta}_t^1 - \bar{\Delta}_t$. Similarly to (5) in Lemma 2, we can obtain the decomposition of δ_t by replacing the random disturbance ξ_t with the difference between the non-linear and linear process:

$$\sqrt{t}\delta_t = \frac{1}{\sqrt{t}\gamma_0}\alpha_t\Delta_0 + \frac{1}{\sqrt{t}}\sum_{j=1}^{t-1}(G^{-1} + \omega_j^{(t)})(\bar{R}(\Delta_j) - G\Delta_j) = I_t^{(1)} + I_t^{(2)} \tag{11}$$

It suffices to show that $\sqrt{t}\delta_t \to 0$ as $t \to \infty$.

We can see that $I_t^{(1)} \to 0$ as $t \to \infty$ because by Lemma 2, $\|\alpha_t\| \leq K$ and the rest of the terms are constant with respect to t.

Now, we need to show that $I_t^{(2)} \to 0$. By Lemma 2 and Assumption 2.2

$$I_t^{(2)} \leq \sum_{i=1}^{\infty} \frac{1}{i^{1/2}} |(G^{-1} + w_j^t)(\bar{R}(\Delta_j) - G\Delta_j)| \quad \leq K \sum_{i=1}^{\infty} \frac{1}{i^{1/2}} |\bar{R}(\Delta_j) - G\Delta_j)| \leq K \sum_{i=1}^{\infty} \frac{|\Delta_i|^{1+\lambda}}{i^{1/2}}$$

The first inequality comes from the fact that each $i^{1/2}$ is less than $t^{1/2}$. The second inequality comes from Lemma 2 where the sum of $\|w_j^t\|$ is bounded and the fact that the linear map G^{-1} is bounded. The last inequality comes directly from Assumption 2.2.

Now, from (10) in part 2 and Assumption 2.4, we have

$$\sum_{i=1}^{\infty} \frac{E(|\Delta_t|^{1+\lambda} I(\tau_R > t))}{i^{1/2}} \leq \sum_{i=1}^{\infty} \frac{K\gamma_i^{\frac{1+\lambda}{2}}}{i^{1/2}} < \infty$$

. Since the sum of the expectation is finite, we must also have that the sum itself is finite which gives us $\sum_{i=1}^{\infty} \frac{|\Delta_t|^{1+\lambda} I(\tau_R > t)}{i^{1/2}} < \infty$.

We now want to use (9) from Part 2 to show that $\sum_{i=1}^{\infty} \frac{|\Delta_t|^{1+\lambda}}{i^{1/2}} < \infty$. To do that, note that

$$\left\{ \sum_{i=1}^{\infty} \frac{|\Delta_t|^{1+\lambda}}{i^{1/2}} < \infty \right\} \supseteq \left\{ \sup_i |\Delta_i| \leq R \right\} \cap \left\{ \sum_{i=1}^{\infty} \frac{|\Delta_t|^{1+\lambda} I(\tau_R > t)}{i^{1/2}} < \infty \right\}$$

This is true since the RHS is just restricting the set on the LHS. Now, we know that from (9) the probability of the RHS is greater than $1 - \varepsilon$.

Since the RHS is a subset of the LHS, we have that the probability of LHS is greater than the RHS which gives $P\left(\left\{ \sum_{i=1}^{\infty} \frac{|\Delta_t|^{1+\lambda}}{i^{1/2}} < \infty \right\} \right) \geq 1 - \varepsilon$.

Since this holds $\forall\ \varepsilon > 0$ we have that this probability is arbitrarily close to 1 which means that

$$\sum_{i=1}^{\infty} \frac{|\Delta_t|^{1+\lambda}}{i^{1/2}} < \infty \tag{12}$$

Finally, we can use Kronecker's lemma on (12). To do this, note that each $i^{1/2}$ is increasing in i, but less than $t^{1/2} \to \infty$ as $t \to \infty$. Furthermore, (12) shows that the sum converges to a finite value. Putting this together means that as $t \to \infty$

$$I_t^{(2)} \leq \frac{1}{\sqrt{t}} \sum_{j=1}^{t-1} \frac{|\Delta_t|^{1+\lambda}}{j^{1/2}} \to 0$$

Where the convergence to 0 comes from Kronecker's lemma. This establishes the asymptotic equivalence of the linear approximation process $\bar{\Delta}_t^1$ with the non-linear process $\bar{\Delta}_t$ which proves Theorem 2.

□

References

[ABE19] Andreas Anastasiou, Krishnakumar Balasubramanian, and Murat A. Erdogdu, *Normal approximation for stochastic gradient descent via non-asymptotic rates of martingale clt*, 2019.

[BM11] Francis Bach and Eric Moulines, *Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning*, Neural Information Processing Systems (NIPS) (Spain), 2011, pp. –.

[BMP90] Albert Benveniste, Michel Métivier, and Pierre Priouret, *Adaptive algorithms and stochastic approximations*, Springer Berlin Heidelberg, Berlin, Heidelberg, 1990.

[FXY17] Yixin Fang, Jinfeng Xu, and Lei Yang, *On scalable inference with stochastic gradient descent*, 2017.

[LLKC17] Tianyang Li, Liu Liu, Anastasios Kyrillidis, and Constantine Caramanis, *Statistical inference using sgd*, 2017.

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

[SZ18] Weijie J. Su and Yuancheng Zhu, *Uncertainty quantification for online learning and stochastic approximation via hierarchical incremental gradient descent*, 2018.

Appendix A. Additional Details of the Proofs and Small Simulation

Proof of Lemmas

Let $(X_j^{(t)})_{t\geq j}$, $(\bar{X}_j^{(t)})_{t\geq j}$ be the sequences of $\mathbb{R}^{N\times N}$ matrices determined by the following recursive relations:

$$\begin{aligned} X_j^{(t+1)} &= X_j^{(t)} - \gamma_t A X_j^{(t)} \qquad & X_j^{(j)} &= I \\ \bar{X}_j^{(t)} &= \gamma_j \sum_{i=j}^{t-1} X_j^{(i)}, \qquad & \phi_j^{(t)} &= A^{-1} - \bar{X}_j^{(t)} \end{aligned} \tag{13}$$

Proof of Lemma 1: According to Assumption 1.2, the coefficient γ_t needs to satisfy one of the two conditions, first of which assumes γ_t is a constant, and second of which is more relaxed and requires $\gamma_t \to 0$ and to decrease slowly. Part 1 of the proof of Lemma 1 deals with the first condition, and part 2-4 demonstrate the proof under the second condition.

Part 1. Proposition of the lemma is true with the first condition in Assumption 1.2 holds:

$$\gamma_t \equiv \gamma, \qquad 0 < \gamma < 2\big(\min_i \ Re\ \gamma_i(A)\big)^{-1} \tag{14}$$

Proof. Given $\gamma_t \equiv \gamma$ and (3) we can get:

$$\begin{aligned} X_j^{(t)} &= X_j^{(t-1)}(I - \gamma A) \\ &= X_j^{(t-2)}(I - \gamma A)^2 \\ &\cdots\cdots \\ &= X_j^{(j)}(I - \gamma A)^{t-j} \\ &= (I - \gamma A)^{t-j} \qquad given\ X_j^{(j)} = I \end{aligned} \tag{15}$$

By (4), we have:

$$\begin{aligned} \bar{X}_j^{(t)} &= \gamma(X_j^{(j)} + X_j^{(j+1)} + \cdots + X_j^{(t-1)}) \\ &= \gamma(I + (I - \gamma A) + \cdots + (I - \gamma A)^{t-j-1}) \end{aligned} \tag{16}$$

Since $I - \gamma A$ is a square matrix, then the last expression in (10) by definition is a geometric series generated by the matrix $I - \gamma A$. By Assumption 1.2, $Re\ \lambda_i(A) > 0$, then

$$\lambda_i(I - \gamma A) = I - \gamma\lambda_i(A),\ and\ |\lambda_i(I - \gamma A)| < 1 \tag{17}$$

Given condition (11),

$$\begin{aligned} \bar{X}_j^{(t)} &= \gamma(I + (I - \gamma A) + \cdots + (I - \gamma A)^{t-j-1}) \\ &= \gamma[I - (I - \gamma A)]^{-1}[I - (I - \gamma A)^{t-j}] \\ &= \gamma(\gamma A)^{-1}[I - (I - \gamma A)^{t-j}] \\ &= A^{-1} - (I - \gamma A)^{t-j}A^{-1} \end{aligned} \tag{18}$$

Note that condition (11) is equivalent to

$$\lim_{t\to\infty}(I - \gamma A)^t = 0 \tag{19}$$

From (5), we have:

$$\begin{aligned}\phi_j^{(t)} &= A^{-1} - \bar{X}_j^{(t)} \\ &= (I - \gamma A)^{t-j} A^{-1}\end{aligned} \tag{20}$$

It is easy to see that (6) hold given (13) and (14); and

$$\frac{1}{t}\sum_{j=0}^{t-1}\phi_j^{(t)} = \frac{1}{t}\sum_{j=0}^{t-1}(I-\gamma A)^{t-j}A^{-1} = \frac{1}{t}\sum_{k=2}^{t}(I-\gamma A)^k A^{-1} \ \rightarrow \ 0 \ as \ t \ \rightarrow \ \infty. \tag{21}$$

Now we have completed the proof of Lemma 1 under the condition of $\gamma_t \equiv \gamma$.

Part 2. It holds that $t\gamma_t \to \infty$ with the second condition in Assumption 1.2:

$$\gamma_t \to 0, \qquad \frac{\gamma_t - \gamma_{t+1}}{\gamma_t} = o(\gamma_t) \tag{22}$$

Proof. We define $\alpha_t = \frac{1}{t\gamma_t}$. Then,

$$\begin{aligned}\alpha_{t+1} &= \frac{1}{(t+1)\gamma_{t+1}} = \frac{1}{t+1}(\gamma_t^- 1 + o(1)) \qquad \text{as } t \to \infty \\ &= \alpha_t \frac{t}{t+1} + \frac{o(1)}{t+1} \\ &= \alpha_t\left(1 - \frac{1}{t+1}\right) + \frac{o(1)}{t+1},\end{aligned} \tag{23}$$

where $o(1) \to 0$ as $t \to \infty$. Since $\sum_{t=1}^{\infty}\frac{1}{(t+1)}$ is a harmonic series without it's first term it diverges. So, $\sum_{t=1}^{\infty}\frac{1}{t+1} = \infty$ and we obtain, $\alpha_t \to 0$ and consequently, $t\gamma_t \to \infty$.

Part 3. There are $\alpha \geq 0$ and $K \leq \inf$ such that for all j and $t \geq j$ we have

$$\|X\| \leq Kexp(-\alpha\sum_{i=j}^{t-1}\gamma_i) \tag{24}$$

Proof. From the second condition of the Lemma, given

$$Re\lambda_i(A) \geq 0, i = \overline{1, N}$$

Then there is a constant $K \leq \infty$ such that for all j and $t \geq j$

$$\|X\| \leq K \tag{25}$$

$$\lim_{t\to\infty}\frac{1}{t}\sum_{j=0}^{t-1}\phi_j^t = 0 \tag{26}$$

and from the Lyapunov theorem, we know that there exists the solution $V = V^T \geq 0$ for the Lyapunov equation $A^TV + VA = I$.

Define $L = max\lambda_i(V), l = min\lambda_i(V)$ and $U_t = (X_j^t)^T V X_j^t$. Substituting the defined terms in U_{t+1} and simplifying, we get

$$\begin{aligned}U_{t+1} &= (X_j^t)^T(I-\gamma_t A)^T V(I-\gamma_t A)X_j^t \\ &= U_t - \gamma_t(X_j^t)^T(A^TV + VA)X_j^t + \gamma_t^2(X_j^t)^T A^T V A X_j^t\end{aligned} \tag{27}$$

From the norm inequality, we get that

$$(X_j^t)^T X_j^t \geq 1/L(X_j^t)^T V X_j^t$$

and, where $c = \frac{\|A\|^2 L}{l}$. Then from

$$(X_j^t)^T A^T V A X \leq c (X_j^t)^T V X_j^t$$

Then from U_{t+1} equation (27), for a sufficiently large t and some $\lambda \geq 0$, we get that:

$$U_{t+1} \geq_t \left(1 - \frac{1}{L}\gamma_t + c\gamma_t^2\right) \quad \leq (1 - \lambda\gamma_t) U_t \leq e^{-\lambda\gamma_t} U_t$$

As a result, $U_t \leq U_j exp(-\lambda \sum_{i=j}^{t-1} \gamma_t)$. However,

$$\|U_t\| \geq l \|X_j^t\|^2 \text{ and } \|U_j\| \leq L \|X_j^j\|^2 = L$$

Thus we obtain that

$$\|X_j^t\| \leq \sqrt{\frac{L}{l}} exp\left(-\frac{\lambda}{2} \sum_{i=j}^{t-1} \gamma_i\right)$$

Part 4. Equations (25) and (26) hold.

Proof. Summing the first equation of (13) from j to t,

$$X_j^t = X_j^j - A \sum_{i=j}^{t-1} \gamma_i X_j^i = I - A \sum_{i=j}^{t-1} \gamma_i X_j^i \tag{28}$$

We consider sum in the RHS of (28), summing by parts, we get that

$$\sum_{i=j}^{t-1} \gamma_i X_j^i = \gamma_j \sum_{i=j}^{t-1} X_j^i + \sum_{i=j}^{t-1} (\gamma_i - \gamma_j) X_j^i$$

From the second relation in (13) we get,

$$\sum_{i=j}^{t-1} \gamma_i X_j^i = \bar{X_j^t} + S_j^t$$

We now estimate S_j^t by using the results of Part 3, and assumption 1.2 to obtain:

$$\begin{aligned} \|S_j^t\| &\leq \left\| \sum_{i=1}^{t} \left[\sum_{k=j}^{i=1} (\gamma_{k+1} - \gamma_k) \right] X_j^i \right\| \leq \sum_{i=j}^{t} \sum_{k=j}^{i-1} \gamma_k o(\gamma_k) \|X_j^j\| \\ &\leq o(\gamma_j) \sum_{i=j}^{t} m_j^i e^{\lambda m_j^i} = o(\gamma_j) \sum_{i=j}^{t} \frac{m_j^i e^{-\lambda m_j^i} (m_j^i - m_j^i - 1)}{\gamma_i} \end{aligned} \tag{29}$$

, where $m_j^i = \sum_{k=j}^{i} \gamma_k$. From part 2, it follows that $j\gamma_j \leq Ki\gamma_i$ for a sufficiently large i. Since $m_j^i = \sum_{k=j}^{i} \gamma_k \geq \mu(ln(i/j))$, where μ is an arbitrarily large constant, we can estimate $1/\gamma_i$ as

$$\frac{1}{\gamma_i} \leq K \frac{i}{j\gamma_j} \leq \frac{K}{\gamma_j} exp\left(\frac{m_i^j}{\mu}\right)$$

From (29) we get

$$\|S_j^t\| \leq \frac{Ko(\gamma_j)}{\gamma_j} \sum_{i=j}^{t} m_j^i e^{-\lambda m_j^i}(m_j^i - m_j^{i-1}) \equiv \frac{Ko(\gamma_j}{\gamma_j} \int_0^\infty m e^{-\lambda m} dm \epsilon_j$$

, where the Riemann sum goes to zero as $j \to \infty$, such that, for all $t \geq j$,

$$\lim_{j\to\infty} \epsilon_j = 0 \tag{30}$$

From (25), rearranging the equation yields $\bar{X_j^t} + S_j^t = A^{-1} - A^{-1}X_j^t$, we have, by definition of ϕ_j^t, that

$$\phi_j^t = S_j^t + A^{-1}X_j^t$$

From part 3, we have that $\|X_j^t\| \leq K$; thus from (30) we obtain $\phi_j^t \leq K$ i.e (25). Since $\|X_j^t\| \leq Kexp(-\mu(ln(t/j))) = K(j/t)^\mu$ for μ arbitrarily large, we get that

$$\frac{1}{t}\sum_{j=j_o}^{t-1} \|X_j^t\| \leq K(\mu+1)^{-1}$$

, for j_o large enough. For some K, recall $\|X_j^t\| \leq K$,

$$\frac{1}{t}\sum_{j=0}^{t-1} \|X_j^t\| = \frac{1}{t}\sum_{j=0}^{j_o} \|X_j^t\| + \frac{1}{t}\sum_{j=j_0+1}^{t-1} \|X_j^t\| \quad \leq \frac{1}{t}\sum_{j=0}^{j_o} K + \frac{1}{t}\sum_{j=j_0+1}^{t-1} \|X_j^t\|$$

For arbitrary $\epsilon \geq 0$, we can choose μ and a j_o dependent on mu, $j_o(\mu)$ such that,

$$\frac{1}{t}\sum_{j=0}^{t-1} \|X_j^t\| \leq K(\mu+1)^{-1} \quad \leq \epsilon/2$$

Then, choosing a sufficiently large t, we get that $1/t\sum_{j=0}^{j_0} K \leq \epsilon/2$. Hence $1/t\sum_{j=0}^{t-1}\|X_j^t\| \leq \epsilon$. From (30), we know,

$$\lim_{t\to\infty} \frac{1}{t}\sum_{j=0}^{t-1} \|S_j^t\| = 0.$$

Therefore, we obtain, (26) $\lim_{t\to\infty} 1/t\sum_{j=0}^{t-1}\phi_j^{it} = 0$.

□

Proof of Lemma 2: First, from algorithm 1, we have:

$$\begin{aligned}
\Delta_t &= \Delta_{t-1} - \gamma_t(Ax_{t-1} - b) + \gamma_t\xi_t \\
&= \Delta_{t-1} - \gamma_t(Ax_{t-1} - Ax^*) + \gamma_t\xi_t \\
&= \Delta_{t-1} - \gamma_t(A\Delta_{t-1}) + \gamma_t\xi_t, \\
\bar{\Delta_t} &= \frac{1}{t}\sum_{i=0}^{t-1} \Delta_i.
\end{aligned} \tag{31}$$

$$\begin{aligned}\Delta_t &= \Delta_{t-1} - \gamma_t A\Delta_{t-1} + \gamma_t\xi_t \\ &= \Delta_{t-1}(I-\gamma_t A) + \gamma_t\xi_t \\ &= \Delta_{t-2}(I-\gamma_{t-1}A)(I-\gamma_t A) + [(I-\gamma_t A)\gamma_{t-1}\xi_{t-1} + \gamma_t\xi_t] \\ &\cdots\cdots \\ &= \Delta_0\prod_{j=1}^{t}(I-\gamma_j A) + \sum_{j=1}^{t}\gamma_j\xi_j\prod_{i=j+1}^{t}(I-\gamma_i A)\end{aligned} \tag{32}$$

Then by definition, we have:

$$\begin{aligned}\bar{\Delta}_t &= \frac{1}{t}\sum_{j=0}^{t-1}\Delta_0\prod_{i=1}^{j}(I-\gamma_i A) + \frac{1}{t}\sum_{k=1}^{t-1}\sum_{j=1}^{k}\gamma_j\xi_j[\prod_{i=j+1}^{k}(I-\gamma_i A)] \\ &= \frac{1}{t}\sum_{j=0}^{t-1}\Delta_0\prod_{i=1}^{j}(I-\gamma_i A) + \frac{1}{t}\sum_{j=1}^{t-1}\gamma_j\xi_j[\sum_{k=j}^{t-1}\prod_{i=j+1}^{k}(I-\gamma_i A)]\end{aligned} \tag{33}$$

The second equality is obtained by expanding the term after the minus sign and rearranging. Set

$$\begin{aligned}\alpha_j^{(t)} &= \gamma_j\sum_{k=j}^{t-1}\prod_{i=j+1}^{k}(I-\gamma_i A) \\ \omega_j^{(t)} &= \alpha_j^{(t)} - A^{-1}.\end{aligned} \tag{34}$$

Then the last equation can be expressed as:

$$\bar{\Delta}_t = \frac{1}{t\gamma_0}\alpha_0^{(t)}\Delta_0 + \frac{1}{t}\sum_{j=1}^{t-1}\omega_j^{(t)}\xi_j + \frac{1}{t}\sum_{j=1}^{t-1}A^{-1}\xi_j \tag{35}$$

By definition of $X_j^{(t)}$ and $\bar{X}_j^{(t)}$, we have

$$\begin{aligned}\alpha_j^{(t)} &= \bar{X}_j^{(t)}, \\ \|\omega_j^{(t)}\| &= \|\alpha_j^{(t)} - A^{-1}\| = \|A^{-1} - \alpha_j^{(t)}\| \\ &= \|A^{-1} - \bar{X}_j^{(t)}\| \\ &= \|\phi_j^{(t)}\|\end{aligned} \tag{36}$$

From lemma 1, we can get $lim_{t\to\infty}\frac{1}{t}\|\omega_j^{(t)}\| = 0$, and $\|\omega_j^{(t)}\| \le K$.
By triangle inequality, we also have $\|\alpha_j^{(t)}\| = \|\omega_j^{(t)} + A^{-1}\| \le \|\omega_j^{(t)}\| + \|A^{-1}\| \le K$.

□

Additional Propositions of Theorem 1

(b) If assumptions 1.1-1.3 and 1.5(b) are satisfied, then $lim_{t\to\infty}E[t(\bar{x}_t - x^*)(\bar{x}_t - x^*)^T] = V$.

(c) If assumptions 1.1-1.3 are satisfied and all the $(\xi_t)_{t\ge1}$ are mutually i.i.d, then $\bar{x}_t - x^* \to 0,\ a.s.$

Proof. [Proof of Theorem 1 Proposition (b)] We must prove that

$$lim_{t\to\infty}tE(\bar{\Delta}_t\bar{\Delta}_t^T) = V$$

under Assumptions 1.1-1.3, 1.5(b).

From the decomposition (19) used in the proof of proposition (a), we have that

$$\begin{aligned} tE(\bar{\Delta}_t\bar{\Delta}_t^T) &= tE((I^{(1)}+I^{(2)}+I^{(3)})(I^{(1)}+I^{(2)}+I^{(3)})^T) \\ &= tE(I^{(1)}I^{(1)T}+I^{(1)}I^{(2)T}+I^{(1)}I^{(3)T}+I^{(2)}I^{(1)T}+I^{(2)}I^{(2)T}+I^{(2)}I^{(3)T} \\ &+ I^{(3)}I^{(1)T}+I^{(3)}I^{(2)T}+I^{(3)}I^{(3)T}) \end{aligned}$$

By the same argument as in proposition (a), all the parts involving $I^{(1)}$ and $I^{(3)}$ $\to 0$ as $t \to \infty$ which leaves us with

$$tE(\bar{\Delta}_t\bar{\Delta}_t^T) = tE(I^{(2)}I^{(2)T})$$

So taking the limit and substituting gives us

$$\begin{aligned} lim_{t\to\infty}tE(\bar{\Delta}_t\bar{\Delta}_t^T) = lim_{t\to\infty}tE(I^{(2)}I^{(2)T}) &= lim_{t\to\infty}E\left(\frac{1}{t}\sum_{j=1}^{t-1}A^{-1}\xi_j\xi_j^T(A^{-1})^T\right) \\ &= lim_{t\to\infty}\frac{1}{t}\sum_{j=1}^{t-1}A^{-1}E(\xi_j\xi_j^T)(A^{-1})^T = A^{-1}S(A^{-1})^T = V \end{aligned}$$

The third equality comes from the linearity of expectation and the fourth equality comes from Assumption 1.5(b).

Proof. [Proof of Theorem 1 Proposition (c)] We must prove that

$$\bar{\Delta}_t \overset{a.s.}{\to} 0 \text{ as } t \to \infty$$

under Assumptions 1.1-1.3 and the additional assumption that $(\xi_t)_{t\geq 1}$ are i.i.d. with mean 0. To do this we will use the same decomposition (26) as as part 1. In this section, we will assume that ξ_i are univariate for simplicity of notation, but the argument is exactly the same for the multivariate case

By Lemma 2 and the same argument as part 1, $\frac{I^{(1)}}{\sqrt{t}} \to 0$ as $t \to \infty$.

Now let's consider

$$\frac{I^{(2)}}{\sqrt{t}} = \sum_{j=1}^{t-1}A^{-1}\xi_j = A^{-1}\sum_{j=1}^{t-1}\xi_j$$

Under the assumption of i.i.d., by the SLLN $\sum_{j=1}^{t-1}\xi_j \overset{a.s.}{\to} tE(\xi_1) = 0$ which means that $I^{(2)} \overset{a.s.}{\to} 0$.

All that is left to show is that $\frac{I^{(3)}}{\sqrt{t}} \to 0$. To do this, let's first consider the random sequence $(\bar{\xi_t})_{t\geq 1}$ defined as

$$\bar{\xi_t} = \begin{cases} \xi_t & |\xi_t| \leq t^{3/4} \\ 0 & |\xi_t| > t^{3/4} \end{cases}$$

By Chebyshev's inequality,

$$P(|\xi_t| > t^{3/4}) \leq E(|\xi_t|^2)t^{-3/2} \leq Kt^{-3/2}$$

This means that $\sum_{i=1}^{\infty} P(|\xi_i| > i^{3/4}) = \sum_{i=1}^{\infty} Ki^{-3/2} < \infty$. So we have

$$P(|\xi_t| > t^{3/4} \text{ infinitely often}) = 0$$

Which means that we only need to consider the case when $|\xi_t| \leq t^{3/4}$. Since w_j^t are uniformly bounded by Lemma 2, we now can show $\frac{I^{(3)}}{\sqrt{t}} \to 0$ by showing

$$\frac{1}{t}\sum_{j=1}^{t-1} w_j^t \bar{\xi}_j = t^{-1}S_t \to 0$$

To do this we will use the Chernoff bound on $t^{-1}S_t$ with the function $f(X) = X^4$. This will involve bounding all the terms that come from the fourth power expansion of $t^{-1}S_t$ which has the form

$$S_t^4 = \sum_{j=0}^{t-1}(w_j^t)^4\bar{\xi}_j^4 + K\sum_{i<j}^{t-1}(w_j^t)^2(w_i^t)^2\bar{\xi}_j^2\bar{\xi}_i^2 + \sum_{\substack{i\neq j\\ i\neq k\\ j<k}}^{t-1}(w_i^t)^2 w_j^t w_k^t \bar{\xi}_i^2\bar{\xi}_j\bar{\xi}_k$$
$$+ K\sum_{i<j<k<l}^{t-1} w_i^t w_j^t w_k^t w_l^t \bar{\xi}_i\bar{\xi}_j\bar{\xi}_k\bar{\xi}_l + K\sum_{i\neq j}^{t-1} w_i^t (w_j^t)^3 \bar{\xi}_i\bar{\xi}_j^3 = \sum_{i=1}^{5} I_t^{(i)}$$

We will demonstrate some bounds for each $E(I_t^{(i)})$ for $i = 1, ..., 5$. To do this, let's first note that since $E(\xi_t) = 0$

$$|E(\bar{\xi}_t)| = E[\xi_t I(|\xi_t| > t^{3/4})] \leq [E(\xi_t^2)]^{1/2}[P(|\xi_t| > t^{3/4})]^{1/2} \leq Kt^{-3/4}$$

The first inequality comes from the Cauchy Schwarz inequality for random variables and the second inequality comes from the earlier result using Chebyshev's inequality.

Also, for any higher moment $E(\bar{\xi}_t^k)$ for $k > 1$ we that since $\bar{\xi}_t \leq t^{3/4}$ by definition, $E(\bar{\xi}_t^k) \leq E(\bar{\xi}_t t^{3k/4}) = t^{3k/4}E(\bar{\xi}_t)$. With these two observations, we are now able to provide some bounds for $E(I_t^{(i)})$ for $i = 1, ..., 5$.

$$E(I_t^{(1)}) = E\left(\sum_{j=0}^{t-1}(w_j^t)^4\bar{\xi}_j^4\right) = KE\left(\sum_{j=0}^{t-1}\bar{\xi}_j^4\right) \leq Kt^{3/2}\sum_{j=0}^{t-1}E(\bar{\xi}_j^2)$$
$$\leq Kt^{3/2}\sum_{j=0}^{t-1} j^{-3/4}j^{3/4} \leq Kt^{3/2}\sum_{j=0}^{t-1} 1 = Kt^{5/2}$$

$$E(I_t^{(2)}) = K\sum_{i<j} E(\bar{\xi}_i^2)E(\bar{\xi}_j^2) \leq K\sum_{i<j} i^{-3/4}i^{3/4}j^{-3/4}j^{3/4} = K\sum_{i<j} 1 = Kt^2$$

$$|E(I_t^{(3)})| = K\left|\sum_{\substack{i\neq j\\ i\neq k\\ j<k}}^{t-1} E(\bar{\xi}_i^2)E(\bar{\xi}_j)E(\bar{\xi}_k)\right| \leq K\sum_{\substack{i\neq j\\ i\neq k\\ j<k}}^{t-1} i^{-3/4}i^{3/4}j^{-3/4}k^{-3/4}$$
$$\leq Kt^{-3/2}\sum_{\substack{i\neq j\\ i\neq k\\ j<k}}^{t-1} 1 \leq Kt^{-3/2}t^3 = Kt^{3/2}$$

$$E(I_t^{(4)}) = K\sum_{i<j<k<l} E(\bar{\xi}_i)E(\bar{\xi}_j)E(\bar{\xi}_k)E(\bar{\xi}_l) \leq K\sum_{i<j<k<l} i^{-3/4}j^{-3/4}k^{-3/4}l^{-3/4}$$
$$\leq K\sum_{i<j<k<l} t^{-3} = Kt^{-3}t^4 = Kt$$

$$|E(I_t^{(5)})| = K\left|\sum_{i\neq j}^{t-1} E(\bar{\xi}_i)E(\bar{\xi}_j^3)\right| \leq K\sum_{i\neq j}^{t-1} i^{-3/4}j^{3/4}E(\bar{\xi}_j^2) \leq K\sum_{i\neq j}^{t-1} 1 = Kt^2$$

With all of these bounds, we can now see that

$$t^{-4}E(S_t^4) = t^{-4}E\left(\sum_{i=1}^{5} I_t^{(i)}\right) \leq Kt^{-4}(t^{5/2}+t^2+t^{3/2}+t) \leq Kt^{-4}(4t^{5/2}) = Kt^{-3/2}$$

So we can now apply the Chernoff bound to get

$$\sum_{t=1}^{\infty} P(|t^{-1}S_t| > \delta) \leq \sum_{t=1}^{\infty}(t\delta)^{-4}E(S_t^4) \leq K\sum_{t=1}^{\infty} t^{-3/2} < \infty$$

Since δ here is arbitrary, we get that this probability is arbitrarily small which means that $t^{-1}S_t \overset{a.s.}{\to} 0$ as $t \to \infty$. All this together proves the almost sure convergence.

□

Simulation with Various Step Sizes

Finally, we provide a small simulation to demonstrate what can happen when the step size assumptions are not met. As we note, for step sizes γ_t of the form $\gamma t^{-\alpha}$, $\frac{1}{2} < \alpha < 1$ would satisfy Assumption 2.4. In this case we consider simple linear regression on simulated data. The setup of this simulation is as follows: 300 random values $(x_i)_{i=1}^{300}$ were generated uniformly on the interval $[0, 10]$, 300 random error terms $(\varepsilon_i)_{i=1}^{300}$ were generated independently from a $N(0, 1)$ distribution, and our outcome variable y_i was obtained by

$$y_i = 5x_i + \varepsilon_i, \text{ for } i = 1, 2, ..., 300$$

We use the averaged stochastic gradient descent algorithm on the squared error loss function to estimate $\beta_1 = 5$. We run 1000 simulations of this process starting at initial value $\beta_1^{(0)} = 10$ to obtain a distribution for this estimate. We do this for two values of γ and three values of α. The first value of γ is 0.01 which was found to be a suitable value for

convergence, and the second value is 0.03 which is a value a bit higher than the convergence value. For the three α levels which corresponds to the decay rate of the step size, one was chosen to satisfy Assumption 2.4, while another was chosen to be lower, and the last was chosen to be higher. Figures 1 and 2 contain the outcomes of this simulation.

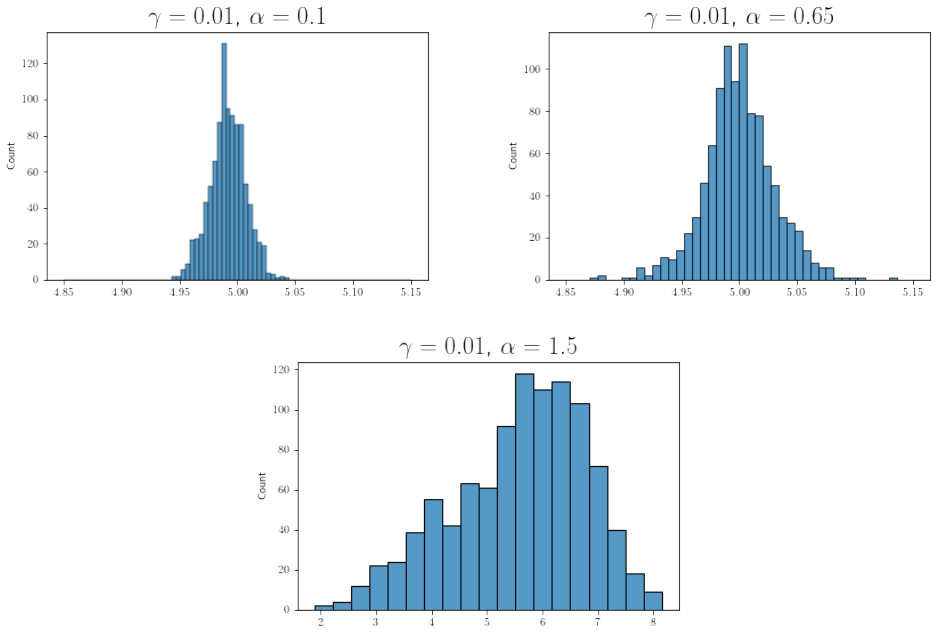

FIGURE 1. Distributions of $\bar{\beta}_1$ at lower initial step size

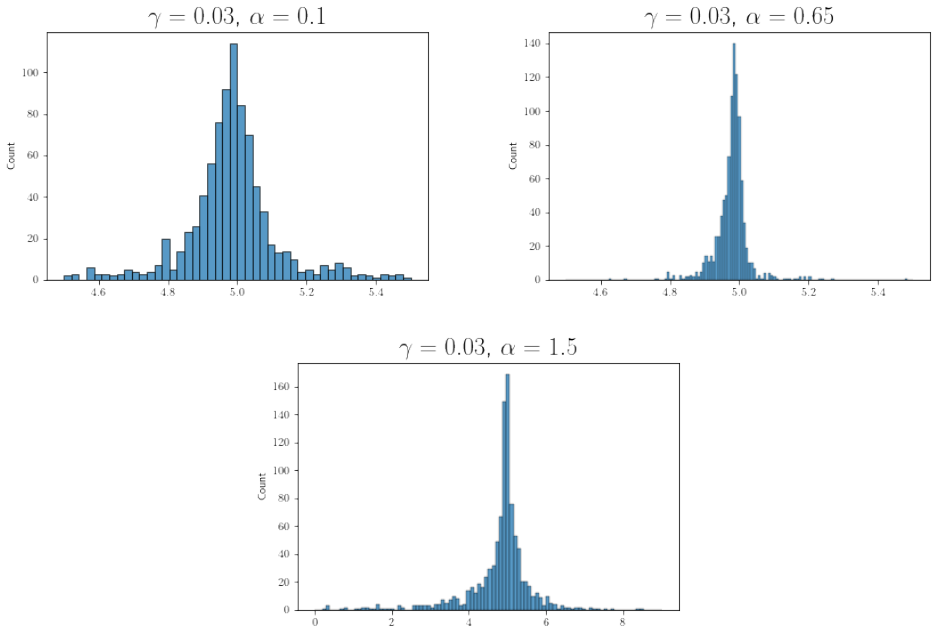

FIGURE 2. Distributions of $\bar{\beta}_1$ at higher initial step size

From these results, we can see that in the lower initial step size $\gamma = 0.01$ scenario, we get reasonable convergence for both the theoretically sound decay value of 0.65 and the low decay value of 0.1. Qualitatively, the $\alpha = 0.65$ case looks more normally distributed since the $\alpha = 0.1$ case has light tails, but in practice that would be a desirable outcome. However, for the higher decay rate of $\alpha = 1.5$ there was no real convergence as the distribution is skewed towards the initial value 10. This is an intuitive result since a fast step size rate of decay would mean that the algorithm would take step too small before reaching the optimal value.

For the higher initial step size $\gamma = 0.03$, we have poorer convergence for the $\alpha = 0.65$ and $\alpha = 0.1$ cases than before. In the theoretically sound case, we are still generally centered at the true value, but there are higher incidents of tail events now. This suggests a higher variance and more numerical instability brought about by the higher initial step size. In the low decay rate case this phenomenon is very pronounced. The tails are much heavier than a normal distribution which suggests high variance in the estimate. The benefit brought about by the theoretically sound decay rate seems to be more robustness against improperly set initial step sizes. Finally, for the higher decay rate we have a left skewed distribution. This suggests that in some cases the initial values decreased too quickly and the vanishing step size left the algorithm stuck in this unoptimal region.

Email address: rvdam@ucdavis.edu

Email address: yixlu@ucdavis.edu

Email address: smore@ucdavis.edu