Les Cahiers du GERAD ISSN: 0711–2440

Vector generalizations of the Robbins-Siegmund theorem and stochastic gradient descent with delays

A. Mahajan, S.I. Niculescu, M. Vidyasagar

G–2026–27

June 2026

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

Vector generalizations of the Robbins-Siegmund theorem and stochastic gradient descent with delays

Aditya Mahajan [a, b]

Silviu-Iulian Niculescu [c]

Mathukumalli Vidyasagar [d]

[a] *Department of Electrical and Computer Engineering, McGill University, Montréal (QC), Canada, H3A 0E9*

[b] *GERAD, Montréal (QC), Canada, H3T 1J4*

[c] *University of Paris-Saclay, CNRS, CentraleSupelec, Inria Saclay, Laboratory of Signals and Systems (L2S), 91190 Gif-sur-Yvette, France*

[d] *Department of Artificial Intelligence at the Indian Institute of Technology Hyderabad, Kandi, Telangana 502285, India.*

aditya.mahajan@mcgill.ca
silviu.niculescu@centralesupelec.fr
m.vidyasagar@iith.ac.in

June 2026
Les Cahiers du GERAD
G–2026–27

Abstract : The almost supermartingale convergence theorem of Robbins and Siegmund (1971) is a fundamental tool for establishing the convergence of various stochastic iterative algorithms arising in system identification, adaptive control, machine learning, and reinforcement learning. However, the original theorem is limited to scalar-valued stochastic processes. In this paper, we generalize the Robbins-Siegmund theorem to non-negative vector-valued stochastic processes and provide multiple sufficient conditions for such processes to converge almost surely, relying on the convergence of infinite products of matrices. We demonstrate the utility of this framework by introducing the concept of an autoregressive almost supermartingale and applying it to establish the almost sure convergence of stochastic gradient descent with delayed updates.

Keywords: Stochastic approximation; time-delay systems; stochastic gradient descent

Résumé : Le théorème de convergence des presque surmartingales de Robbins et Siegmund (1971) constitue un résultat fondamental pour l'analyse de convergence de nombreux algorithmes itératifs stochastiques intervenant en identification des systèmes, commande adaptative, apprentissage automatique et apprentissage par renforcement. Néanmoins, sa formulation originale se limite aux processus stochastiques scalaires. Dans cet article, nous proposons une généralisation du théorème de Robbins–Siegmund au cadre des processus stochastiques vectoriels non négatifs et établissons plusieurs conditions suffisantes garantissant leur convergence presque sûre, en nous appuyant notamment sur l'étude de la convergence de produits infinis de matrices. Nous illustrons ensuite l'intérêt de ce cadre en introduisant la notion de presque surmartingale autorégressive, que nous appliquons à l'analyse de la convergence presque sûre de méthodes de descente de gradient stochastique en prenant en compte le retard dans les mises à jour.

Mots clés : Approximation stochastique; systèmes à retard; gradient stochastique

Acknowledgements: The work of Aditya Mahajan was supported by DATAIA Fellowship and Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant RGPIN–2021–0351 and Alliance Catalyst Grant ALLRP 592356. The work of Mathukumalli Vidyasagar was supported by the Science and Engineering Research Board of India.

1 Introduction

In optimization, systems and control, signal processing, and machine learning, data is often observed sequentially and corrupted by noise. Algorithms in these settings typically start with an initial guess of a solution and recursively update it as additional data becomes available. Representative examples of such stochastic iterative algorithms include recursive least square methods in system identification [9, 17, 18, 19], certainty equivalent methods in adaptive control [7, 9, 33], stochastic gradient descent methods in machine learning [21, 30], various learning algorithms such as Q-learning in reinforcement learning [22, 35], and distributed consensus [23, 24].

The convergence of these algorithms is commonly studied using stochastic approximation theory [3, 6, 15, 16, 29]. Two main approaches are used. The first, called the *ODE approach*, establishes that the iterates of stochastic iterative algorithms track, in an appropriate sense, the trajectories of deterministic difference or differential equations. The second approach, called the *martingale approach*, analyzes the discrete-time iterates directly and establishes convergence using generalizations of the supermartingale convergence theorem, typically using a generalization called the almost supermartingale convergence theorem which is due to Robbins and Siegmund [28]. We refer the reader to [6, 16] for an overview and applications of the ODE approach and to [34] for an overview of the martingale approach. While these classical results establish asymptotic convergence, the recent literature increasingly focuses on deriving non-asymptotic finite-time bounds to quantify the convergence rate and sample complexity. See [10] for an overview.

In recent years, there has been a proliferation of distributed machine learning where processing is offloaded to a cloud or edge servers, which often leads to network and processing delays. This has led to considerable interest in understanding the behavior of standard machine learning algorithms such as stochastic gradient descent (SGD) under delayed updates (also called stale updates).

Since the seminal Hogwild! paper [27], numerous asynchronous SGD variants have been studied [1, 5, 8, 11, 12, 20]. However, with the exception of [5, 8], this literature primarily establishes finite-time convergence rates for the *average* of the iterates. Crucially, convergence of the average does not guarantee that the iterates themselves remain bounded or converge. To the best of the authors' knowledge, [5, 8] are the only work in this domain that establishes almost sure convergence (and also asymptotic normality in case of [8]) of the iterates themselves.

In this paper, we present a conceptually simple approach to analyze the convergence of SGD with delayed updates. Our intuition rests on a standard technique: discrete-time processes with memory (i.e., delay) can be converted to processes without memory by using an appropriate state augmentation. However, this transformation does not work for SGD with delayed updates because the standard martingale analysis relies on the scalar-valued almost supermartingale convergence theorem of Robbins and Siegmund [28]. State augmentation introduces vector-valued states, which would require a vector generalization of the almost supermartingale convergence theorem which, to the best of our knowledge, does not exist in the open literature.

The main contributions of this paper can be summarized as follows.

- In Sec. 2, we provide vector-valued generalizations of the almost supermartingale convergence theorem under different sets of sufficient conditions, which rely on the convergence of an appropriately defined infinite product of matrices.
- As an application of these vector-valued almost supermartingales, in Sec. 3 we introduce the concept of an *autoregressive almost supermartingale* and provide easy to verify sufficient conditions for its boundedness and convergence.
- In Sec. 4, we show that SGD with delayed updates can be viewed as an autoregressive almost supermartingale and use the results of the previous section to show that it converges under stated assumptions on the objective function and the noise.

The notations are standard and explained when first used.

2 Vector generalizations of Robbins-Siegmund theorem

A sequence $\{X_n\}_{n\geq 0}$ of real-valued random variables adapted to a filtration $\{\mathcal{F}_n\}_{n\geq 0}$ is called a **supermartingale** if

$$\mathbb{E}_n[X_{n+1}] \leq X_n, \quad \forall n \geq 0$$

where the notation $\mathbb{E}_n[\cdot]$ is a shorthand for $\mathbb{E}[\cdot \,|\, \mathcal{F}_n]$. Roughly speaking, supermartingales generalize the notion of monotonically decreasing sequences to stochastic processes, see, e.g., [26]. A fundamental result in martingale theory is the **supermartingale convergence theorem** [25, 31], which states that non-negative supermartingales converge almost surely.

When analyzing convergence of stochastic iterative algorithms, it is not always convenient to directly apply the supermartingale convergence theorem and a slight generalization, initially proposed by Robbins and Siegmund [28], is much more useful:

Theorem 1 (Almost supermartingale convergence)**.** Suppose $\{X_n\}_{n\geq 0}$, $\{\beta_n\}_{n\geq 0}$, $\{Y_n\}_{n\geq 0}$, $\{Z_n\}_{n\geq 0}$ are $\mathbb{R}_+$-valued stochastic processes adapted to some filtration $\{\mathcal{F}_n\}_{n\geq 0}$ that satisfy

$$\mathbb{E}_n[X_{n+1}] \leq (1+\beta_n)X_n + Y_n - Z_n, \quad n \geq 0 \tag{1}$$

Define the set Ω_0 by

$$\Omega_0 = \left\{\omega \in \Omega : \sum_{n\geq 0} \beta_n(\omega) < \infty\right\} \cap \left\{\omega : \sum_{n\geq 0} Y_n(\omega) < \infty\right\}.$$

Then, for almost all $\omega \in \Omega_0$, we have that

1. The sequence $\{X_n(\omega)\}_{n\geq 0}$ converges, and hence, is bounded.[1]
2. $\sum_{n\geq 0} Z_n(\omega) < \infty$.

To analyze the convergence of SGD with delayed updates, we need a *delayed* version of Theorem 1. Such delayed almost-supermartingales can be viewed as vector-valued almost supermartingales via state augmentation. Therefore, we start with a generalization of Theorem 1 from $\mathbb{R}_+$ to $\mathbb{R}_+^p$. Such a generalization is trivial for $\mathbb{R}_+^p$-valued stochastic processes if they are *component-wise* supermartingales, that is, $\mathbb{E}_n[X_{n+1}] \leq X_n$, where the inequality holds for each component of X_{n+1}. However, such simple arguments do not work in the almost-supermartingale case.

We present a set of sufficient conditions under which the vector-valued almost supermartingales converge. Our conditions rely on the existence of the limit of infinite product of matrices. So, we review some basic results of infinite product of matrices in Sec. 2.1 and then present the main convergence theorem in Sec. 2.2.

2.1 Review of infinite product of matrices

We start with a review of infinite products of matrices. If $\{B_n\}_{n\geq 0}$ are $p \times p$ real matrices, we define

$$\prod_{k=n}^{m} B_k = \begin{cases} B_m B_{m-1} \cdots B_n & \text{if } n \leq m, \\ I & \text{if } n > m; \end{cases}$$

to be the product where successive terms multiply on the left.

Let $\|\cdot\|$ denote any sub-multiplicative matrix norm on $\mathbb{R}^{p\times p}$. The following are sufficient conditions for the convergence of infinite product of matrices.

[1] Note that the bound will in general depend on ω.

(C1) If $\sum_{n=1}^{\infty}\|B_n\| < \infty$ then $\prod_{n=1}^{\infty}(I + B_n)$ converges.

(C2) Let $\{R_n\}_{n\geq 0}$ be a sequence of $p \times p$ matrices such that $\lim_{n\to\infty} R_n = I$ and $\sum_{n=1}^{\infty}\|(I+B_n)R_n - R_{n+1}\| < \infty$ then $\prod_{n=1}^{\infty}(I + B_n)$ converges.

(C3) Let $\{U_n\}_{n\geq 0}$ be a sequence of $p \times p$ matrices such that $\|U_n\| = 1$ for all n. Suppose that for all $N \geq 0$, the product $\prod_{n=N}^{m} U_n$ converges for $m \to \infty$ and $\sum_{n=1}^{\infty}\|B_n\| < \infty$, then $\prod_{n=1}^{\infty}(U_n + B_n)$ converges.

Condition (C1) is a standard result and stated as Theorem 1 in [32]. Condition (C2) is Theorem 5 of [32]. Condition (C3) is Theorem 2.1 in [2]. It is also argued in [2] that in (C3) if the U_n are invertible, then convergence of $\prod_{n=1}^{m} U_n$ as $m \to \infty$ implies convergence of $\prod_{n=N}^{m} U_n$ for all N.

2.2 First vector generalization of Theorem 1

Suppose $\{X_n\}_{n\geq 0}$, $\{Y_n\}_{n\geq 0}$, and $\{Z_n\}_{n\geq 0}$ are $\mathbb{R}_+^p$-valued stochastic processes and $\{A_n\}_{n\geq 0}$ is a $\mathbb{R}_+^{p\times p}$-valued stochastic process, all adapted to some filtration $\{\mathcal{F}_n\}_{n\geq 0}$, that satisfy

$$\mathbb{E}_n[X_{n+1}] \leq A_n X_n + Y_n - Z_n, \quad n \geq 0. \tag{2}$$

We assume that there exists a deterministic sequence $\{\bar{A}_n\}_{n\geq 0}$, $\bar{A}_n \in \mathbb{R}_+^{p\times p}$ that satisfies the following properties:

(A1) For each n, $A_n \leq \bar{A}_n$ almost surely.

(A2) For each n, $\bar{A}_n$ is invertible.

(A3) The infinite product

$$\bar{M} \coloneqq \prod_{n=0}^{\infty} \bar{A}_n$$

converges. Clearly, $\bar{M} \in \mathbb{R}_+^{p\times p}$.

(A4) $\bar{M}$ is invertible.

Remark 1. The sufficient conditions (C1)–(C3) mentioned above for existence of infinite product of matrices can be used to verify (A3). In particular, if we assume that $A_n = I + B_n$ where $\sum_{n\geq 0}\|B_n\| < \infty$, we have a natural generalization of the conditions in Theorem 1 to the vector case. However, condition (A3) is more general. For example, as shown in [32], $\prod_{n\geq 0}(I + B_n)$ converges even if $\sum_{n\geq 0}\|B_n\| = \infty$ if condition (C2) holds.

Define

$$\bar{\Psi}_n \coloneqq \prod_{m\geq n} \bar{A}_m. \tag{3}$$

It follows from (A2) and (A3) that $\bar{\Psi}_n$ is well defined for all n because

$$\bar{\Psi}_n = \bar{M}\left[\prod_{m=0}^{n-1} \bar{A}_m\right]^{-1}. \tag{4}$$

It is also clear that for all $n \geq 0$, $\bar{\Psi}_n \in \mathbb{R}_+^{p\times p}$, and

$$\bar{\Psi}_{n+1}\bar{A}_n = \bar{\Psi}_n. \tag{5}$$

It is shown in [32] that (A4) implies

$$\lim_{n\to\infty} \bar{\Psi}_n = I. \tag{6}$$

Theorem 2. Suppose (A1)–(A3) hold. Define the set

$$\Omega_1 = \left\{\omega \in \Omega : \sum_{n\geq 0} Y_n(\omega) < \infty\right\}. \tag{7}$$

Then, for almost all $\omega \in \Omega_1$, we have the following:

1. The sequence $\{\bar{\Psi}_n X_n(\omega)\}_{n\geq 0}$ converges, and hence, is bounded.
2. Additionally, if (A4) holds then $\{X_n(\omega)\}_{n\geq 0}$ converges, and hence, is bounded.
3. $\sum_{n\geq 0} \bar{\Psi}_{n+1} Z_n(\omega) < \infty$.

Proof. Observe that since all processes are non-negative, Assumption (A1) implies that we can replace (2) by the following:

$$\mathbb{E}_n[X_{n+1}] \leq \bar{A}_n X_n + Y_n - Z_n, \quad n \geq 0. \tag{8}$$

Define $X'_n = \bar{\Psi}_n X_n$, $Y'_n = \bar{\Psi}_{n+1} Y_n$, and $Z'_n = \bar{\Psi}_{n+1} Z_n$. Now consider

$$\begin{aligned}
\mathbb{E}_n[X'_{n+1}] &= \mathbb{E}_n[\bar{\Psi}_{n+1} X_{n+1}] \\
&\overset{(a)}{=} \bar{\Psi}_{n+1} \mathbb{E}_n[X_{n+1}] \\
&\overset{(b)}{\leq} \bar{\Psi}_{n+1}\big(\bar{A}_n X_n + Y_n - Z_n\big) \\
&\overset{(c)}{=} X'_n + Y'_n - Z'_n
\end{aligned} \tag{9}$$

where (a) follows because $\bar{\Psi}_{n+1}$ is deterministic, (b) follows from (8), and (c) follows from (5) and the definition of X'_n, Y'_n, and Z'_n. Note that (9) holds for every component.

Define

$$\Omega'_1 = \left\{\omega \in \Omega : \sum_{n\geq 0} \|\bar{\Psi}_{n+1} Y_n(\omega)\| < \infty\right\}$$

(where $\|\cdot\|$ is any norm on $\mathbb{R}^p$). Since all norms are equivalent in a finite-dimensional space, we have that for any $\omega \in \Omega'_1$, $\sum_{n\geq 0} [Y'_n]_i < \infty$, for all $i \in \{1, \ldots, p\}$, where $[Y'_n]_i$ denotes the i-th component of Y'_n. Therefore, by applying Theorem 1 to every component of (9), we get that for almost all $\omega \in \Omega'_1$, $\lim_{n\to\infty} X'_n(\omega)$ exists and $\sum_{n\geq 0} Z'_n(\omega) < \infty$. This establishes the first and last part of the result.

We now prove part 2. Recall that due to (A2) and (A4), $\bar{\Psi}_n$ is invertible and $\lim_{n\to\infty} \bar{\Psi}_n = I$. Therefore, $\lim_{n\to\infty} \bar{\Psi}_n^{-1} = I$. Hence, for almost all $\omega \in \Omega'_1$,

$$X_n(\omega) = \bar{\Psi}_n^{-1} X'_n(\omega) \to X'_\infty(\omega).$$

So, we have shown that the result holds for almost all $\omega \in \Omega'_1$. We will now show that $\Omega_1 \subseteq \Omega'_1$. For any $\omega \in \Omega_1$, we have that $\sum_{n\geq 0} \|Y_n(\omega)\|_\infty < \infty$. Due to the equivalence of norms in a finite dimensional space, this implies that

$$\sum_{n\geq 0} \|Y_n(\omega)\| < \infty. \tag{10}$$

From (A3), it follows that $\bar{\Psi}^* := \sup_{n\geq 0} \|\bar{\Psi}_n\| < \infty$. Then, for any $\omega \in \Omega_1$,

$$\sum_{n\geq 0} \|\bar{\Psi}_{n+1} Y_n(\omega)\| \leq \bar{\Psi}^* \sum_{n\geq 0} \|Y_n(\omega)\| < \infty$$

where the first inequality follows from the definition of $\bar{\Psi}^*$ and the second inequality follows from (10). Thus, $\Omega_1 \subseteq \Omega'_1$. Hence, the result holds for almost all $\omega \in \Omega_1$. □

2.3 Second vector generalization of Theorem 1

Assumptions (A1)–(A4) require an upper bound on $\{A_n\}_{n\geq 0}$ to satisfy certain properties. We now present a variation of these assumptions that does not require an upper bound, but places a stronger assumption on the infinite product of matrices. In particular, we assume that the following assumptions hold.

(A2') For each n, A_n is invertible.

(A3') The infinite product

$$M \coloneqq \prod_{n=0}^{\infty} A_n$$

converges to a deterministic limit, almost surely. Clearly $M \in \mathbb{R}_+^{p\times p}$.

(A4') M is invertible.

Similar to version 1, define

$$\Psi_n \coloneqq \prod_{m\geq n} A_m. \tag{11}$$

It follows from (A2') and (A3') that Ψ_n is well defined for all n because

$$\Psi_n = M\left[\prod_{m=0}^{n-1} A_m\right]^{-1}. \tag{12}$$

Since M is deterministic, Ψ_n is $\mathcal{F}_n$ measurable. As in version 1, it is clear that for all $n \geq 0$, $\Psi_n \in \mathbb{R}_+^{p\times p}$, and

$$\Psi_{n+1}A_n = \Psi_n. \tag{13}$$

Furthermore, as before, (A4') implies that

$$\lim_{n\to\infty} \Psi_n = I. \tag{14}$$

Theorem 3. Suppose (A2')–(A4') hold. Define the set Ω_1 to be the same as (7). Then, for almost all $\omega \in \Omega_1$, we have that

1. The sequence $\{X_n(\omega)\}_{n\geq 0}$ converges, and hence, is bounded.
2. $\sum_{n\geq 0} \Psi_{n+1}Z_n(\omega) < \infty$.

Proof. The proof argument is similar to that of Theorem 2, with the following changes. Define $X'_n = \Psi_n X_n$, $Y'_n = \Psi_{n+1}Y_n$, and $Z'_n = \Psi_{n+1}Z_n$. Now consider

$$\begin{aligned}
\mathbb{E}_n[X'_{n+1}] &= \mathbb{E}_n[\Psi_{n+1}X_{n+1}] \\
&\overset{(a)}{=} \Psi_{n+1}\mathbb{E}_n[X_{n+1}] \\
&\overset{(b)}{\leq} \Psi_{n+1}\big(A_nX_n + Y_n - Z_n\big) \\
&\overset{(c)}{=} X'_n + Y'_n - Z'_n
\end{aligned} \tag{15}$$

where (a) follows because Ψ_{n+1} is $\mathcal{F}_n$ measurable due to (12), (b) follows from (2), and (c) follows from (13) and the definition of X'_n, Y'_n, and Z'_n. Note that (15) holds for every component.

The rest of the proof argument is the same as that of Theorem 2. □

2.4 Third vector generalization of Theorem 1

Assume that the sequence $\{A_n\}_{n\geq 0}$ satisfies the following assumptions:

(A5) For each n, A_n is invertible.

(A6) For each n, $A_n^{-1} \in \mathbb{R}_+^{p\times p}$.

Remark 2. A necessary and sufficient condition for (A6) is that each A_n is *monomial*, i.e., for every n there exists a permutation matrix P_n and a diagonal matrix D_n with strictly positive diagonal elements such that $A_n = P_n D_n$.

Define $\Phi_n \coloneqq \prod_{m<n} A_m$. It follows from (A5) that Φ_n is invertible and from (A6) that $\Phi_n^{-1} \in \mathbb{R}_+^{p\times p}$. It is also clear that for all $n \geq 0$,

$$\Phi_{n+1}^{-1} A_n = \Phi_n^{-1}. \tag{16}$$

Theorem 4. Suppose assumptions (A5) and (A6) hold. Define

$$\Omega_2 = \left\{\omega \in \Omega : \lim_{n\to\infty} \Phi_n(\omega) \text{exists and is finite}\right\} \cap \left\{\omega \in \Omega : \sum_{n\geq 0} \Phi_{n+1}^{-1} Y_n(\omega) < \infty\right\}. \tag{17}$$

Then, for almost all $\omega \in \Omega_2$, we have that

1. The sequence $\{X_n(\omega)\}_{n\geq 0}$ converges, and hence, is bounded.
2. $\sum_{n\geq 0} \Phi_{n+1}^{-1}(\omega) Z_n(\omega) < \infty$.

Proof. Define $X'_n = \Phi_n^{-1} X_n$, $Y'_n = \Phi_{n+1}^{-1} Y_n$, $Z'_n = \Phi_{n+1}^{-1} Z_n$. Then,

$$\begin{aligned}\mathbb{E}_n[X'_{n+1}] &\overset{(a)}{=} \Phi_{n+1}^{-1}\mathbb{E}_n[X_{n+1}] \overset{(b)}{\leq} \Phi_{n+1}^{-1}[A_n X_n + Y_n - Z_n] \\ &\overset{(c)}{=} \Phi_n^{-1} X_n + \Phi_{n+1}^{-1} Y_n - \Phi_{n+1}^{-1} Z_n = X'_n + Y'_n - Z'_n,\end{aligned} \tag{18}$$

where (a) follows because Φ_{n+1} is $\mathcal{F}_n$-measurable, (b) follows from (2) and (A6) and (c) follows from (16). Inequality (18) holds for each component of the vectors. Therefore, by applying Theorem 1 on each component, we get that $\lim_{n\to\infty} X'_n(\omega)$ and $\sum_{n\geq 0} Z'_n(\omega)$ exist and are finite for almost all elements in the set $\left\{\omega : \sum_{n\geq 0} Y'_n(\omega) < \infty\right\} \supset \Omega_2$.

Furthermore, on Ω_2, $\lim_{n\to\infty} \Phi_n(\omega)$ converges to some limit Φ_∞. So, $X_n = \Phi_n X'_n$ also converges. □

3 Autoregressive almost supermartingales

Fix an integer $D \geq 0$ and let $\mathcal{D} = \{0, 1, \ldots, D\}$. Consider $\mathbb{R}_+$-valued stochastic processes $\{X_n\}_{n\geq 0}$, $\{Y_n\}_{n\geq 0}$, $\{Z_n\}_{n\geq 0}$, $\{a_{n,d}\}_{n\geq 1}$, $d \in \mathcal{D}$, all adapted to a filtration $\{\mathcal{F}_n\}_{n\geq 0}$, that satisfy,

$$\mathbb{E}_n[X_{n+1}] \leq X_n + \sum_{d\in\mathcal{D}} a_{n,d} X_{n-d} + Y_n - Z_n, \quad \forall n \geq D. \tag{19}$$

Drawing a parallel with (1), we call such a process **autoregressive almost supermartingale**.

Define $\bar{X}_n, \bar{Y}_n, \bar{Z}_n \in \mathbb{R}^{D+1}$ as follows: $\bar{X}_n = \text{vec}(X_{n-D}, \ldots, X_n)$, $\bar{Y}_n = \text{vec}(0, \ldots, 0, Y_n)$, and $\bar{Z}_n = \text{vec}(0, \ldots, 0, Z_n)$. Then, the dynamics (19) may be written in vector form as

$$\mathbb{E}_n[\bar{X}_{n+1}] \leq A_n \bar{X}_n + \bar{Y}_n - \bar{Z}_n, \quad n \geq D \tag{20}$$

where $A_n \in \mathbb{R}_+^{(D+1)\times(D+1)}$ is given by

$$A_n = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \ddots & 0 & 1 \\ a_{n,D} & a_{n,D-1} & \cdots & \cdots & a_{n,0}+1 \end{bmatrix}. \tag{21}$$

Theorem 5. Consider the autoregressive almost supermartingale defined in (19). Suppose there exists a strictly positive deterministic sequence $\{\bar{a}_{n,d}\}_{d\in\mathcal{D},n\geq 0}$ that satisfies the following:

(AS1) For all $d \in \mathcal{D}$, $n \geq 0$, we have $a_{n,d} \leq \bar{a}_{n,d}$.

(AS2) For all $d \in \mathcal{D}$, we have $\sum_{n\geq 0} \bar{a}_{n,d} < \infty$.

Define $\bar{A}_n$ to have a similar structure to A_n with $a_{n,d}$ replaced by $\bar{a}_{n,d}$ and define $\bar{\Psi}_n$ as in (3). Define the set Ω_1 as in (7), that is,

$$\Omega_1 = \left\{\omega \in \Omega : \sum_{n\geq 0} Y_n(\omega) < \infty\right\}.$$

Then, for almost all $\omega \in \Omega_1$, we have the following conclusions:

(Con1) The sequence $\{\bar{\Psi}_n \bar{X}_n(\omega)\}_{n\geq D}$ converges and is therefore bounded.

(Con2) The sequence $\{X_n(\omega)\}_{n\geq 0}$ converges and is therefore bounded.

(Con3) $\sum_{n\geq 0} Z_n(\omega) < \infty$.

Proof. By construction $\bar{A}_n$ is invertible and satisfies (A1) and (A2). We first prove that it satisfies (A3).

Observe that we can write $\bar{A}_n$ as

$$\bar{A}_n = S + B_n$$

where

$$S = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \ddots & 0 & 1 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}, \quad B_n = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \ddots & 0 & 0 \\ \bar{a}_{n,D} & \bar{a}_{n,D-1} & \cdots & \cdots & \bar{a}_{n,0} \end{bmatrix}.$$

Observe that for $k \geq D$, $S^k = L$, where

$$L = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 0 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}.$$

Thus, for every $N \geq 0$, the product $\prod_{n=N}^{m} S = S^{m-N+1}$ converges to L as $m \to \infty$. Moreover, from (AS2),

$$\sum_{n\geq D} \|B_n\|_\infty = \sum_{n\geq D} \sum_{d=0}^{D} \bar{a}_{n,d} < \infty. \tag{22}$$

Hence the perturbations B_n are summable. Thus, condition (C3) of Sec. 2.1 implies that

$$\prod_{n=D}^{\infty} \bar{A}_n = \prod_{n=D}^{\infty} (S + B_n)$$

converges. Hence (A3) holds.

Now, Eq. (20) may be viewed as a vector almost-supermartingale. The matrices $\bar{A}_t$ satisfy (A1)–(A3). Therefore, Theorem 2 implies that $\lim_{n\to\infty} \bar{\Psi}_n \bar{X}_n(\omega)$ exists and is bounded and $\sum_{n\geq 0} \bar{\Psi}_{n+1}\bar{Z}_n(\omega) < \infty$. This establishes the first part.

We now prove the second part. We first show that $\{X_n\}_{n\geq 0}$ is bounded. Since all entries of $\bar{A}_n$ are non-negative,

$$[\bar{\Psi}_n]_{D,D} \geq \prod_{m=n}^{\infty} (1 + \bar{a}_{m,0}) \geq 1, \tag{23}$$

where we use the notation $[M]_{i,j}$ to denote the (i,j)-th entry of a matrix M (with indices ranging from 0 to D) and use a similar notation for vectors.

Now observe that

$$[\bar{\Psi}_n \bar{X}_n]_D = \sum_{d=0}^{D} [\bar{\Psi}_n]_{D,d}[\bar{X}_n]_d \geq [\bar{\Psi}_n]_{D,D} X_n \geq X_n$$

where we have used the structure of $\bar{X}_n$ and (23). By (Con1), we know that $\{\bar{\Psi}_n \bar{X}_n\}_{n\geq 0}$ converges and, therefore, all its components are bounded.

We now prove convergence of $\{X_n\}_{n\geq 0}$. For any $M \geq n$, define

$$P_{n,M} \coloneqq \prod_{m=n}^{M} S \quad \text{and} \quad Q_{n,M} \coloneqq \prod_{m=n}^{M} (S + B_m).$$

By the telescoping identity for products, we have

$$Q_{n,M} - P_{n,M} = \sum_{r=n}^{M} \left(\prod_{m=r+1}^{M} (S + B_m) \right) B_r \left(\prod_{m=n}^{r-1} S \right)$$

Since $\|S\|_\infty = 1$, we get

$$\begin{aligned} \|Q_{n,M} - P_{n,M}\|_\infty &\leq \sum_{r=n}^{M} \prod_{m=r+1}^{M} (1 + \|B_m\|_\infty)\|B_r\|_\infty \\ &\leq \sum_{r=n}^{M} \prod_{m=n}^{M} (1 + \|B_m\|_\infty)\|B_r\|_\infty \\ &\leq \exp\left(\sum_{m=n}^{M} \|B_m\|_\infty \right) \sum_{m=n}^{M} \|B_m\|_\infty. \end{aligned} \tag{24}$$

Taking the limit as $M \to \infty$, we get that

$$\|\bar{\Psi}_n - L\|_\infty \leq \exp\left(\sum_{m=n}^{\infty} \|B_m\|_\infty \right) \sum_{m=n}^{\infty} \|B_m\|_\infty.$$

Combining this with (22), we get that

$$\lim_{n\to\infty} \|\bar{\Psi}_n - L\|_\infty = 0$$

Hence $\lim_{n\to\infty} \bar{\Psi}_n = L$.

Since $\{X_n\}_{n\geq 0}$ is bounded, we have that there exists some constant C such that $\|\bar{X}_n\|_\infty \leq C$.

$$\lim_{n\to\infty} \|(\bar{\Psi}_n - L)\bar{X}_n\|_\infty \leq \lim_{n\to\infty} \|(\bar{\Psi}_n - L)\|_\infty C = 0.$$

Now in (Con1), we had shown that $\{\bar{\Psi}_n \bar{X}_n\}_{n\geq 0}$ converges. Combining with the above, we get that $\{L\bar{X}_n\}_{n\geq 0}$ converges. By definition of L, we have that

$$L\bar{X}_n = (X_n, \ldots, X_n)^\top.$$

Hence, $\{X_n\}_{n\geq 0}$ converges. This completes the proof of (Con2).

Now we prove the third part. Observe that

$$[\bar{\Psi}_{n+1}\bar{Z}_n]_D = \sum_{d=0}^{D} [\bar{\Psi}_{n+1}]_{D,d}[\bar{Z}_n]_d \geq [\bar{\Psi}_{n+1}]_{D,D} Z_n \geq Z_n$$

where we have used the structure of $\bar{Z}_n$ and (23). Therefore, Theorem 2 implies that

$$\sum_{n\geq 0} Z_n \leq \sum_{n\geq 0} [\bar{\Psi}_{n+1}\bar{Z}_n]_D < \infty.$$

This proves (Con3). □

4 Stochastic Gradient Descent with Delayed Updates

Let $\{\mathcal{F}_n\}_{n\geq 0}$ be a filtration and let $\{\xi_n\}_{n\geq 0}$, $\xi_n \in \mathbb{R}^p$ be a noise process and $\{\alpha_n\}_{n\geq 0}$, $\alpha_n \in \mathbb{R}_{\geq 0}$, be a sequence of learning rates, both adapted to $\{\mathcal{F}_n\}_{n\geq 0}$.

We consider a version of SGD with delayed updates, where the current update term is a convex combination of past gradients. The coefficients of the convex combination can vary from one step to another, and can also be random. However, it is assumed that there is a known upper bound D on the delayed information.

We are interested in finding the minimum of a $\mathcal{C}^1$ function $J\colon \mathbb{R}^p \to \mathbb{R}$. We start with initial estimates $\theta_0, \cdots, \theta_D \in \mathbb{R}^p$. From step $D+1$ onward, the estimate is updated according to the iteration:

$$\theta_{n+1} = \theta_n - \alpha_n \left[\sum_{k=0}^{D} \lambda_{n,k} \nabla J(\theta_{n-k}) + \xi_{n+1}\right], \quad n \geq D, \tag{25}$$

In the above, for each fixed n, the "weights" $\{\lambda_{n,k}\}_{k=0}^{D}$ are nonnegative, and add up to one. They can also be $\mathcal{F}_n$-measurable random variables. When the delay $D = 0$, the above iteration corresponds to standard SGD. The paper [13] gives the best available results for that problem. Our objective is to analyze the asymptotic behavior of (25) when $D > 0$, in such a way that the results reduce to those of [13] when $D = 0$.

For this purpose, we impose suitable assumptions on the objective function $J(\cdot)$, the noise sequence ξ_{t+1}, and on the learning rate α_t. We begin with the objective function.

(J1) The function J is ℓ-smooth for some $\ell > 0$, i.e., for any $\theta_1, \theta_2 \in \mathbb{R}^p$, we have

$$\|\nabla J(\theta_1) - \nabla J(\theta_2)\| \leq \ell \|\theta_1 - \theta_2\|$$

where $\|\cdot\|$ denotes the Euclidean norm.

(J2) The quantity

$$J^* := \inf_{\theta\in\mathbb{R}^p} J(\theta) > -\infty.$$

Moreover, it is assumed that the infimum is attained. By replacing $J(\cdot)$ by $J(\cdot) - J^*$, it can be assumed without loss of generality that $J^* = 0$.

(J3) $J(\cdot)$ satisfies the relaxed "Kurdyka-Lojasiewicz" condition, denoted by $J \in \mathrm{KL}'$, defined next. A function $\eta\colon \mathbb{R} \to \mathbb{R}$ is said to *belong to Class* $\mathcal{B}$ if (i) $\eta(0) = 0$, and (ii) whenever $\epsilon > 0$ and $M < \infty$, we have

$$\inf_{\epsilon \le r \le M} \eta(r) > 0.$$

The assumption is that there exists a function $\eta(\cdot)$ of Class $\mathcal{B}$ such that

$$\|\nabla J(\theta)\|^2 \ge \eta(J(\theta)).$$

Remark 3. Now we discuss some implications of the above assumptions.

- As shown in [13, Lemma 1], Assumption (J1) implies that

$$\|\nabla J(\theta)\|^2 \le 2\ell J(\theta). \tag{26}$$

- The class of functions known as Kurdyka-Lojasiewicz (KL) is standard in optimization theory. Note that every function that belongs to the class (KL') is **invex**, in the sense that every stationary point θ^* such that $\nabla J(\theta^*) = 0$ is a global minimizer of $J(\cdot)$. The reader is directed to [14] for a detailed discussion of such ideas. Our definition imposes slightly weaker conditions. See [13, Section 4.2] for further details.

Next, we impose the following assumptions on the noise.

(N1) The noise sequence $\{\xi_{n+1}\}_{n\ge 0}$ is a martingale difference sequence with respect to the filtration $\{\mathcal{F}_n\}_{n\ge 0}$, i.e.,

$$\mathbb{E}_n[\xi_{n+1}] = 0, \quad \forall n \ge 0.$$

(N2) There exists a sequence of nonnegative constants σ_n^2 such that

$$\mathbb{E}_n[\|\xi_{n+1}\|^2] \le \sigma_n^2, \quad \forall n \ge 0.$$

Finally, here are the assumptions on the learning rates.

(R1) The learning rate sequence is upper bounded by a deterministic non-negative sequence $\{\bar{\alpha}_n\}_{n\ge 0}$ such that

$$\sum_{n\ge 0} \bar{\alpha}_n^2 < \infty \quad \text{and} \quad \sum_{n\ge 0} \bar{\alpha}_n^2 \sigma_n^2 < \infty.$$

(R2) The learning rate sequence satisfies

$$\sum_{n\ge 0} \alpha_n = \infty, \quad \text{a.s.}$$

Theorem 6. Suppose assumptions (J1)–(J3), (N1) and (N2) hold.

1. If (R1) holds, then the sequences $J(\theta_n)$ and $\nabla J(\theta_n)$ are bounded almost surely.
2. If, in addition, (R2) holds, then the sequences $J(\theta_n)$ and $\|\nabla J(\theta_n)\|$ converge to zero almost surely.

Remark 4. Note that the conclusions of the theorem do not say anything about the behavior of the iterates $\{\theta_n\}$. In order to draw such conclusions, additional assumptions need to be imposed on the objective function $J(\cdot)$. This is done later in the paper in Remark 5.

Proof. The proof consists of finding a bound for the conditional expectation $\mathbb{E}_n[J(\theta_{n+1})]$ in such a way that Theorem 5 can be applied. For this purpose, the following bound from [4, Eq. (2.4)] is very useful. It states that, when Assumption (J1) holds, we have

$$J(\theta_2) \le J(\theta_1) + \langle \nabla J(\theta_1), \theta_2 - \theta_1 \rangle + \frac{\ell}{2}\|\theta_2 - \theta_1\|^2. \tag{27}$$

We will also make use of the following inequalities:

1. Given vectors $x_1, \cdots, x_N \in \mathbb{R}^p$, we have that

$$\left\| \sum_{i=1}^{N} x_i \right\|^2 \leq N \sum_{i=1}^{N} \|x_i\|^2. \tag{28}$$

   This follows directly from Cauchy-Schwartz inequality.

2. Given vectors $x_0, \cdots, x_D \in \mathbb{R}^p$ and non-negative weights $\{w_k\}_{k=0}^D$ which add to 1, we have

$$\left\| \sum_{k=0}^{D} w_k x_k \right\|^2 \leq \sum_{k=0}^{D} w_k \|x_k\|^2. \tag{29}$$

   This follows directly from Jensen's inequality

With these preliminaries out of the way, we begin our analysis of (25). Assumption (N1) means that, whenever $x_n \in \mathbb{R}^p$ is $\mathcal{F}_n$-measurable, we have

$$\mathbb{E}_n[\langle x_n, \xi_{n+1} \rangle] = 0, \quad \mathbb{E}_n[\|x_n + \xi_{n+1}\|^2] = \|x_n\|^2 + \mathbb{E}_n[\|\xi_{n+1}\|^2].$$

Substituting this into (25) and using (27) gives

$$\begin{aligned} \mathbb{E}_n[J(\theta_{n+1})] \leq J(\theta_n) &- \alpha_n \sum_{k=0}^{D} \lambda_{n,k} \langle \nabla J(\theta_n), \nabla J(\theta_{n-k}) \rangle \\ &+ \alpha_n^2 \frac{\ell}{2} \left[\left\| \sum_{k=0}^{D} \lambda_{n,k} \nabla J(\theta_{n-k}) \right\|^2 \right] + \alpha_n^2 \frac{\ell}{2} \mathbb{E}_n[\|\xi_{n+1}\|^2]. \end{aligned} \tag{30}$$

Next, define

$$F_{1,n} := -\alpha_n \sum_{k=0}^{D} \lambda_{n,k} \langle \nabla J(\theta_n), \nabla J(\theta_{n-k}) \rangle.$$

Since $\sum_{k=0}^{D} \lambda_{n,k} = 1$, it follows that

$$F_{1,n} = -\alpha_n \|\nabla J(\theta_n)\|^2 + \alpha_n \sum_{k=1}^{D} \lambda_{n,k} \langle \nabla J(\theta_n), \nabla J(\theta_n) - \nabla J(\theta_{n-k}) \rangle.$$

To proceed further, define $F_{2,n}$ to be the second summation, i.e.,

$$F_{2,n} := \alpha_n \sum_{k=1}^{D} \lambda_{n,k} \langle \nabla J(\theta_n), \nabla J(\theta_n) - \nabla J(\theta_{n-k}) \rangle.$$

Then

$$\begin{aligned} F_{2,n} &\leq \alpha_n \sum_{k=1}^{D} \lambda_{n,k} \|\nabla J(\theta_n)\| \cdot \|\nabla J(\theta_n) - \nabla J(\theta_{n-k})\| \\ &\leq \frac{\alpha_n}{2} \sum_{k=1}^{D} \lambda_{n,k} [\|\nabla J(\theta_n)\|^2 + \|\nabla J(\theta_n) - \nabla J(\theta_{n-k})\|^2] \\ &\leq \frac{\alpha_n}{2} \|\nabla J(\theta_n)\|^2 + \alpha_n \frac{\ell^2}{2} \sum_{k=1}^{D} \lambda_{n,k} \|\theta_n - \theta_{n-k}\|^2 \end{aligned} \tag{31}$$

where the first inequality follows from $ab \leq (a^2 + b^2)/2$ and the second follows from (J1). Substituting into (30) gives

$$\mathbb{E}_n[J(\theta_{n+1})] \leq J(\theta_n) - \frac{\alpha_n}{2} \|\nabla J(\theta_n)\|^2 + V_n + S_n, \tag{32}$$

where

$$V_n := \alpha_n \frac{\ell^2}{2} \sum_{k=1}^{D} \lambda_{n,k} \|\theta_n - \theta_{n-k}\|^2, \tag{33}$$

$$S_n := \alpha_n^2 \frac{\ell}{2} \left[\left\| \sum_{k=0}^{D} \lambda_{n,k} \nabla J(\theta_{n-k}) \right\|^2 \right] + \alpha_n^2 \frac{\ell}{2} \mathbb{E}_n[\|\xi_{n+1}\|^2]. \tag{34}$$

We will now find bounds for V_n and S_n, starting with V_n.Note that if $D = 0$, then $V_n = 0$. Hence, in the following, we will assume $D \geq 1$. We write $\theta_n - \theta_{n-k}$ as a telescoping sum, as in

$$\theta_n - \theta_{n-k} = \sum_{j=1}^{k} [\theta_{n-j+1} - \theta_{n-j}].$$

Thus, from (28), we have

$$\|\theta_n - \theta_{n-k}\|^2 = \left\| \sum_{j=1}^{k} [\theta_{n-j+1} - \theta_{n-j}] \right\|^2 \leq k \sum_{j=1}^{k} \|\theta_{n-j+1} - \theta_{n-j}\|^2. \tag{35}$$

Now we invoke the updating formula (25), giving

$$\theta_{n-j+1} - \theta_{n-j} = \alpha_{n-j} \left[\sum_{m=0}^{D} \lambda_{n-j,m} \nabla J(\theta_{n-j-m}) + \xi_{n-j+1} \right].$$

Therefore

$$\begin{aligned} \|\theta_{n-j+1} - \theta_{n-j}\|^2 &\leq 2\alpha_{n-j}^2 \left[\sum_{m=0}^{D} \lambda_{n-j,m} \|\nabla J(\theta_{n-j-m})\|^2 + \|\xi_{n-j+1}\|^2 \right] \\ &\leq 2\alpha_{n-j}^2 \left[\sum_{m=0}^{D} 2\ell \lambda_{n-j,m} J(\theta_{n-j-m}) + \|\xi_{n-j+1}\|^2 \right] \end{aligned} \tag{36}$$

where the first inequality follows from (28) and (29) and the second from (26). We can substitute this bound into (35), which gives

$$\|\theta_n - \theta_{n-k}\|^2 \leq 2k \sum_{j=1}^{k} \alpha_{n-j}^2 \left[\sum_{m=0}^{D} 2\ell \lambda_{n-j,m} J(\theta_{n-j-m}) + \|\xi_{n-j+1}\|^2 \right].$$

Now this can be substituted into (33), leading to

$$V_n \leq \alpha_n \ell^2 \sum_{k=1}^{D} \lambda_{n,k} k \sum_{j=1}^{k} \alpha_{n-j}^2 \left[\sum_{m=0}^{D} 2\ell \lambda_{n-j,m} J(\theta_{n-j-m}) + \|\xi_{n-j+1}\|^2 \right].$$

This expression can be simplified considerably by (i) replacing k by its upper bound D, and (ii) replacing each λ by its upper bound 1. This gives

$$V_n \leq \alpha_n \ell^2 D \sum_{k=1}^{D} \sum_{j=1}^{k} \alpha_{n-j}^2 \left[\sum_{m=0}^{D} 2\ell \lambda_{n-j,m} J(\theta_{n-j-m}) + \|\xi_{n-j+1}\|^2 \right].$$

The square summability of $\bar{\alpha}_n$ implies that $\bar{\alpha}_n \to 0$ as $n \to \infty$. Since $\alpha_n \leq \bar{\alpha}_n$, we can choose a deterministic integer N sufficiently large that $\alpha_n \leq 1/\ell^2 D$ whenever $n \geq N$. Thus, for all $n \geq N$, we can write

$$V_n \leq \sum_{k=1}^{D} \sum_{j=1}^{k} \alpha_{n-j}^2 \left[\sum_{m=0}^{D} 2\ell \lambda_{n-j,m} J(\theta_{n-j-m}) + \|\xi_{n-j+1}\|^2 \right].$$

It is henceforth assumed that $n \geq N$, and this is not explicitly mentioned every time. Next, observe that

$$\sum_{k=1}^{D}\sum_{j=1}^{k} = \sum_{j=1}^{D}\sum_{k=j}^{D}.$$

Therefore

$$\begin{aligned} V_n &\leq \sum_{j=1}^{D}\sum_{k=j}^{D} \alpha_{n-j}^2 \left[\sum_{m=0}^{D} 2\ell\lambda_{n-j,m} J(\theta_{n-j-m}) + \|\xi_{n-j+1}\|^2\right] \\ &\leq D\sum_{j=1}^{D} \alpha_{n-j}^2 \left[\sum_{m=0}^{D} 2\ell J(\theta_{n-j-m}) + \|\xi_{n-j+1}\|^2\right]. \end{aligned}$$

Here we use the fact that k does not appear anywhere in the summand. Since k ranges from j to D, we can simply multiply the summand by D.

We now consider S_n, defined in (34).

$$\begin{aligned} S_n &= \alpha_n^2 \frac{\ell}{2}\left[\left\|\sum_{k=0}^{D}\lambda_{n,k}\nabla J(\theta_{n-k})\right\|^2\right] + \alpha_n^2\frac{\ell}{2}\mathbb{E}_n[\|\xi_{n+1}\|^2] \\ &\leq \alpha_n^2 \frac{\ell}{2}\left[\sum_{k=0}^{D}\lambda_{n,k}\|\nabla J(\theta_{n-k})\|^2\right] + \alpha_n^2\frac{\ell}{2}\mathbb{E}_n[\|\xi_{n+1}\|^2] \\ &\leq \alpha_n^2\ell^2\sum_{k=0}^{D}\lambda_{n,k}J(\theta_{n-k}) + \alpha_n^2\frac{\ell}{2}\sigma_n^2 \\ &\leq \alpha_n^2\ell^2\sum_{k=0}^{D}J(\theta_{n-k}) + \alpha_n^2\frac{\ell}{2}\sigma_n^2. \end{aligned}$$

where the first inequality follows from (29), second follows from (26), and the last holds as $\lambda_{n,k} \leq 1$. Substituting these bounds into (32) and rearranging terms gives an inequality to which Theorem 5 can be applied, namely

$$\begin{aligned} \mathbb{E}_n[J(\theta_{n+1})] \leq J(\theta_n) - \frac{\alpha_n}{2}\|\nabla J(\theta_n)\|^2 + 2\ell D\sum_{j=1}^{D}\alpha_{n-j}^2\sum_{m=0}^{D}J(\theta_{n-j-m}) + \alpha_n^2\ell^2\sum_{k=0}^{D}J(\theta_{n-k}) \\ + D\sum_{j=1}^{D}\alpha_{n-j}^2\|\xi_{n-j+1}\|^2 + \alpha_n^2\frac{\ell}{2}\sigma_n^2. \end{aligned} \tag{37}$$

This is of the form (19) where $X_n = J(\theta_n)$, with the maximum delay D replaced by $2D$ (because of the presence of the term $J(\theta_{n-j-m})$ where j ranges from 1 to D and m ranges from 0 to D), and coefficients that are easily identifiable. Moreover

$$Y_n = D\sum_{j=1}^{D}\alpha_{n-j}^2\|\xi_{n-j+1}\|^2 + \alpha_n^2\frac{\ell}{2}\sigma_n^2, \quad Z_n = \frac{\alpha_n}{2}\|\nabla J(\theta_n)\|^2.$$

We now verify the hypotheses of Theorem 5. The autoregressive coefficients in (37) are bounded above by deterministic sequences obtained by replacing each α_n^2 by $\bar{\alpha}_n^2$. Thus, (R1) implies (AS1) and (AS2). It remains to show that Y_n is summable. Observe that Y_n consists of two terms (after ignoring constants): $\alpha_n^2\sigma_n^2$, which is summable as per (R1), and a sum of terms of the form $\alpha_{n-j}^2\|\xi_{n-j+1}\|^2$. Moreover, by the tower property of conditional expectations, we have that

$$\mathbb{E}[\|\xi_{n-j+1}\|^2] = \mathbb{E}[\mathbb{E}_{n-j}[\|\xi_{n-j+1}\|^2]] \leq \sigma_{n-j}^2.$$

Note that in the above equation, we are taking the *expected value*, not the conditional expectation. Now applying (R1) shows that

$$\mathbb{E}\left[\sum_{n\geq 0} Y_n\right] \leq \sum_{n\geq 0} D\sum_{j=1}^{D} \bar{\alpha}_{n-j}^2\sigma_{n-j}^2 + \bar{\alpha}_n^2\frac{\ell}{2}\sigma_n^2 < \infty.$$

This in turn implies that Y_n is summable along almost all sample paths.

Therefore when hypothesis (R1) holds, the hypotheses of Theorem 5 apply, and we conclude from Conclusion (Con2) that $J(\theta_n)$ converges and is, therefore, bounded. Since $\|\nabla J(\theta_n)\|^2 \leq 2\ell J(\theta_n)$, it follows that $\nabla J(\theta_n)$ is also bounded. This establishes the first conclusion of the theorem.

To prove the second conclusion under the additional Assumption (R2). We start with Conclusions (Con2) and (Con3) of Theorem 5. Conclusion (Con2) states that the sequence $\{J(\theta_n)\}_{n\geq 0}$ converges almost surely. Let C denote the limit.

Conclusion (Con3) states that

$$\sum_{n\geq 0} Z_n = \sum_{n\geq 0} \frac{1}{2}\alpha_n\|\nabla J(\theta_n)\|^2 < \infty, \tag{38}$$

almost surely.

We now show by contradiction that $C = 0$. Suppose $C > 0$. Since $J(\theta_n) \to C$, there exists a N_1 such that for all $n \geq N_1$

$$J(\theta_n) \in \left[\tfrac{1}{2}C, 2C\right].$$

Now let $\eta(\cdot)$ be the function of Class $\mathcal{B}$ in Assumption (J3), and define

$$\rho := \inf_{C/2\leq r\leq 2C} \eta(r).$$

Then, from (J3), we get

$$\|\nabla J(\theta_n)\|^2 \geq \rho, \quad \forall n \geq N_1.$$

However, since

$$\sum_{n=N_1}^{\infty} \alpha_n = \infty,$$

this contradicts (38). Thus it is not possible for C to be greater than zero. Hence $J(\theta_n) \to 0$. In turn, this implies that $\nabla J(\theta_n)$ also approaches zero due to (26). □

Remark 5. Note that Theorem 6 does not say anything about the convergence of the iterates θ_n. This is because of the *very mild* assumptions on the function $J(\cdot)$. If $J(\cdot)$ is assumed to be strongly convex, in that, for some θ^* and some $R > 0$, it is true that $J(\theta^*) = 0$, and for every pair θ_1, θ_2,

$$J(\theta_2) \geq J(\theta_1) + \langle \nabla J(\theta_1), \theta_2 - \theta_1\rangle + R\|\theta_2 - \theta_1\|^2,$$

Then it follows readily from the convergence of $J(\theta_n)$ to zero that θ_n converges to θ^*.

5 Conclusion

In this paper, we presented a generalization of the Robbins-Siegmund almost supermartingale convergence theorem to non-negative vector-valued stochastic processes. We established multiple sufficient conditions for the almost sure convergence of these processes. These conditions rely on the stability properties of infinite products of matrices. Building on this theoretical foundation, we introduced the

concept of *autoregressive almost supermartingales*, a structure that naturally arises in systems modeled via state augmentation.

To demonstrate the practical utility of our framework, we analyzed the convergence of SGD with delayed updates. We proved that, under standard assumptions on the objective function and noise, delayed SGD drives the objective values to the optimum almost surely, i.e., $J(\theta_n) \to 0$, and $\|\nabla J(\theta_n)\| \to 0$. This result highlights the robustness of SGD to asynchrony and delays.

We believe that the generalizations presented in this work will be useful in the analysis of other stochastic iterative algorithms subject to delays. Future work includes applying this framework to delay-sensitive domains such as asynchronous reinforcement learning and federated reinforcement learning.

References

[1] Alekh Agarwal and John C Duchi. "Distributed Delayed Stochastic Optimization". In: *Advances in Neural Information Processing Systems*. Vol. 24. 2011.

[2] Marc Artzrouni. "On the convergence of infinite products of matrices". In: *Linear Algebra and its Applications* 74 (Feb. 1986), pp. 11–21. ISSN: 0024-3795. DOI: 10.1016/0024-3795(86)90112-6.

[3] Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive algorithms and stochastic approximations*. Vol. 22. Springer Science & Business Media, 2012.

[4] Dmitri P. Bertsekas and John N. Tsitsiklis. "Global convergence in gradient methods with errors". In: *SIAM Journal on Optimization* 10.3 (2000), pp. 627–642.

[5] Shalabh Bhatnagar. "The Borkar–Meyn theorem for asynchronous stochastic approximations". In: *Systems & control letters* 60.7 (2011), pp. 472–478.

[6] Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*. Hindustan Book Agency, 2008.

[7] Peter E Caines. *Linear stochastic systems*. SIAM, 2018.

[8] Sorathan Chaturapruek, John C Duchi, and Christopher Ré. "Asynchronous stochastic convex optimization: the noise is in the noise and SGD don't care". In: *Advances in Neural Information Processing Systems*. 2015.

[9] Han-Fu Chen and Lei Guo. *Identification and Stochastic Adaptive Control*. Birkhäuser Boston, 1991. DOI: 10.1007/978-1-4612-0429-9.

[10] Aymeric Dieuleveut et al. "Stochastic approximation beyond gradient for signal processing and machine learning". In: *IEEE Transactions on Signal Processing* 71 (2023), pp. 3117–3148.

[11] Sanghamitra Dutta et al. "Slow and Stale Gradients Can Win the Race: Error-Runtime Tradeoffs in Distributed SGD". In: *International Conference on Artificial Intelligence and Statistics*. Vol. 84. Apr. 2018, pp. 803–812.

[12] Hamid Reza Feyzmahdavian, Arda Aytekin, and Mikael Johansson. "An Asynchronous Mini-Batch Algorithm for Regularized Stochastic Optimization". In: *IEEE Transactions on Automatic Control* 61.12 (2016), pp. 3740–3754. DOI: 10.1109/TAC.2016.2525015.

[13] Rajeeva Laxman Karandikar and Mathukumalli Vidyasagar. "Convergence Rates for Stochastic Approximation: Biased Noise with Unbounded Variance, and Applications". en. In: *Journal of Optimization Theory and Applications* 203.3 (Dec. 2024), pp. 2412–2450. ISSN: 1573-2878. DOI: 10.1007/s10957-024-02547-7.

[14] Hamed Karimi, Julie Nutini, and Mark Schmidt. "Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak- Lojasiewicz Condition". In: *Lecture Notes in Computer Science* 9851 (2016), pp. 795–811.

[15] J. Kiefer and J. Wolfowitz. "Stochastic estimation of the maximum of a regression function". In: *Annals of Mathematical Statistics* 23.3 (1952), pp. 462–466.

[16] Harold J. Kushner and G. George Yin. *Stochastic Approximation Algorithms and Applications*. Springer New York, 1997. DOI: 10.1007/978-1-4899-2696-8.

[17] Tze Leung Lai and Zhiliang Ying. "Recursive identification and adaptive prediction in linear stochastic systems". In: *SIAM Journal on Control and Optimization* 29.5 (1991), pp. 1061–1090.

[18] L. Ljung. "Analysis of recursive stochastic algorithms". In: *IEEE Transactions on Automatic Control* 22.4 (1977), pp. 551–575. DOI: 10.1109/TAC.1977.1101561.

[19] Lennart Ljung. "Perspectives on system identification". In: *Annual Reviews in Control* 34.1 (2010), pp. 1–12.

[20] Horia Mania et al. "Perturbed Iterate Analysis for Asynchronous Stochastic Optimization". In: *SIAM Journal on Optimization* 27.4 (2017), pp. 2202–2229. DOI: 10.1137/16M1057000.

[21] Panayotis Mertikopoulos et al. "On the Almost Sure Convergence of Stochastic Gradient Descent in Non-Convex Problems". In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020.

[22] Sean Meyn. *Control systems and reinforcement learning*. Cambridge University Press, 2022.

[23] Angelia Nedić and Ji Liu. "Distributed optimization for control". In: *Annual Review of Control, Robotics, and Autonomous Systems* 1 (2018), pp. 77–103.

[24] Angelia Nedić and Alex Olshevsky. "Distributed optimization over time-varying directed graphs". In: *IEEE Transactions on Automatic Control* 60.3 (2014), pp. 601–615.

[25] J. Neveu. *Discrete Parameter Martingales*. North Holland, 1975.

[26] B.T. Polyak. *Introduction to optimization*. Optimization Software Inc: New York, 1987.

[27] Benjamin Recht et al. "Hogwild!: A Lock-Free Approach to Parallelizing Stochastic Gradient Descent". In: *Advances in Neural Information Processing Systems*. Vol. 24. 2011.

[28] H. Robbins and D. Siegmund. "A convergence theorem for non-negative almost supermartingales and some applications". In: *Optimizing Methods in Statistics*. Elsevier, 1971, pp. 233–257. DOI: 10.1016/b978-0-12-604550-5.50015-8.

[29] Herbert Robbins and Sutton Monro. "A Stochastic Approximation Method". In: *The Annals of Mathematical Statistics* 22.3 (1951), pp. 400–407.

[30] Othmane Sebbouh, Robert M Gower, and Aaron Defazio. "Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball". In: *Conference on Learning Theory*. PMLR. 2021, pp. 3935–3971.

[31] A.N. Shiryayev. *Probability*. Springer-Verlag: New York, 1984.

[32] William F. Trench. "Invertibly convergent infinite products of matrices". In: *Journal of Computational and Applied Mathematics* 101.1–2 (Jan. 1999), pp. 255–263. ISSN: 0377-0427. DOI: 10.1016/s0377-0427(98)00191-5.

[33] Y. Tsypkin. *Adaptation and learning in automatic systems*. Academic Press: New York, 1971.

[34] Mathukumalli Vidyasagar. *Stochastic Algorithms for Nonconvex Optimization and Reinforcement Learning*. https://people.iith.ac.in/m_vidyasagar/RL-Notes.pdf. 2025.

[35] Christopher J. C. H. Watkins and Peter Dayan. "Q-learning". In: *Machine Learning* 8.3-4 (May 1992), pp. 279–292. DOI: 10.1007/bf00992698.