Convergence Rate in Nonlinear Two-Time-Scale Stochastic Approximation with State (Time)-Dependence

Zixi Chen[1], **Yumin Xu**[1], **Ruixun Zhang**[1,2,3,4*]
[1]School of Mathematical Sciences, Peking University
[2]Center for Statistical Science, Peking University
[3]National Engineering Laboratory for Big Data Analysis and Applications, Peking University
[4]Laboratory for Mathematical Economics and Quantitative Finance, Peking University
chenzixi22@stu.pku.edu.cn, xuyumin@pku.edu.cn, zhangruixun@pku.edu.cn

Abstract

The nonlinear two-time-scale stochastic approximation is widely studied under conditions of bounded variances in noise. Motivated by recent advances that allow for variability linked to the current state or time, we consider state- and time-dependent noises. We show that the Lyapunov function exhibits polynomial convergence rates in both cases, with the rate of polynomial delay depending on the parameters of state- or time-dependent noises. Notably, if the state noise parameters fully approach their limiting value, the Lyapunov function achieves an exponential convergence rate. We provide two numerical examples to illustrate our theoretical findings in the context of stochastic gradient descent with Polyak-Ruppert averaging and stochastic bilevel optimization.

1 Introduction

Stochastic approximation (SA) was initially introduced by Robbins and Monro (1951) to find the root of an unobserved function operator. Specifically, it aims to approximate the solution of $f(x) = 0$ under noisy observations $f(x_k) + \xi_k$. This method has been widely adopted in different areas including stochastic optimization, machine learning, and reinforcement learning (RL); see, for example, Sutton, Szepesvári, and Maei (2008); Ermoliev (2009); Borkar (2009); Sutton et al. (2009); Pham (2010); Bottou, Curtis, and Nocedal (2018); Xu, Zou, and Liang (2019); Lan (2020); Wu et al. (2020); Khodadadian et al. (2022).

Traditional single-time-scale SA faces challenges in complex applications, and the two-time-scale SA was proposed by Ermoliev (1983) and developed by Borkar (1997) to manage complex processes with both linear and nonlinear update functions. We define the fast-scale as $x_k \in \mathbb{R}^{d_1}$ and the slow-scale as $y_k \in \mathbb{R}^{d_2}$, with unknown update functions $f : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}^{d_1}$ and $g : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}^{d_2}$. $f(x_k, y_k) + \xi_k$ and $g(x_k, y_k) + \psi_k$ represent their noisy observations, respectively. We consider the following two-time-scale process:

$$\begin{aligned} x_{k+1} - x_k &= -\alpha_k[f(x_k, y_k) + \xi_k], \\ y_{k+1} - y_k &= -\beta_k[g(x_k, y_k) + \psi_k], \end{aligned} \tag{1}$$

where $\{\xi_k\}$ and $\{\psi_k\}$ are martingale differences with decreasing sequences of α_k and β_k. It fits a dynamical system:

$$\begin{aligned} \dot{x} &= -f(x, y), \\ \dot{y} &= -g(x, y). \end{aligned} \tag{2}$$

The goal of a stochastic two-time-scale approximation is to control the sequence so that it converges to a goal (x^*, y^*). Because the two variables x_k and y_k interact with each other, their convergence behavior differs from that of a single variable. Typically, to ensure the convergence across both fast and slow scales, the learning rates α_k and β_k are chosen to satisfy $\alpha_k \gg \beta_k$ (Konda and Tsitsiklis 2004). Under certain assumptions, they are set to a special polynomial order of magnitude, as discussed in Doan (2022). With stricter conditions on functions f and g, the ratio of α_k and β_k can be fixed, leading to convergence degenerating into a single-time-scale scenario (Shen and Chen 2022).

It is subtle to determine the relationship between two timescales and their learning rates α_k and β_k. Our work seeks an explicit relationship between the choice of order of magnitude under certain conditions. Moreover, in many applications, the stochastic term is not bounded. We achieve a faster convergence rate ranging from $\mathcal{O}\left(k^{-2/3}\right)$ to $\mathcal{O}\left(k^{-\infty}\right)$—and even an exponential rate of $\mathcal{O}\left(e^{-\epsilon k}\right)$ in certain cases—with different state noise parameters δ_{ij} and time noise parameters γ_i. Here $\mathcal{O}(\cdot)$ indicates that the term is dominated by $c \times \cdot$ with constant c as $k \to \infty$.

Motivating Applications

Two-time-scale SA is widely adopted in many algorithms because it can precisely describe the relation between variables and achieve fast convergence. We are motivated by two algorithms, stochastic gradient descent (SGD) with Polyak-Ruppert averaging and stochastic bilevel optimization.

SGD with Polyak-Ruppert averaging. In order to minimize a function f with only noisy observations of its gradient, Ruppert (1988) and Polyak and Juditsky (1992) introduced SGD with Polyak-Ruppert averaging where

$$\begin{aligned} x_{k+1} - x_k &= -\alpha_k(\nabla F(x_k) + \xi_k), \\ y_{k+1} - y_k &= -\beta_k(y_k - x_k). \end{aligned} \tag{3}$$

This algorithm has the form of two-time-scale SA (1) with $f(x, y) = \nabla F(x)$ and $g(x, y) = y - x$.

*Corresponding author

Stochastic bilevel optimization. Colson, Marcotte, and Savard (2007) provides a review of bilevel optimization, which focus on solving

$$\begin{aligned} &\min_{y\in\mathbb{R}^{d_2}} l(y) := G(x^*(y), y) \\ \text{s.t.} \quad &x^*(y) := \operatorname*{argmin}_{x\in\mathbb{R}^{d_1}} F(x, y), \end{aligned} \tag{4}$$

where $F(x, y)$ and $l(x)$ are called inner and outer objective functions, respectively. With only noisy observations of gradients and Hessian matrices, Shen and Chen (2022) apply the two-time-scale (1) where

$$\begin{aligned} f(x, y) &= \nabla_x F(x, y), \\ g(x, y) &= \nabla_y G(x, y) \\ &\quad - \nabla^2_{yx} F(x, y)[\nabla^2_{xx} F(x, y)]^{-1} \nabla_x G(x, y). \end{aligned} \tag{5}$$

To accelerate their convergence rate, we consider the noisy term under various conditions. For more details of the experiment, see Section 5.

Related Works

Two-time-scale SA convergence was initially introduced by Borkar (1997). The convergence rate in the linear case has been established by Konda and Tsitsiklis (2004), where $\beta_k^{-1/2}(y_k - y^*)$ converges to a normal distribution. Kaledin et al. (2020) established a convergence rate with both martingale and Markovian noises. In these studies, the linear two-time-scale SA achieves a convergence rate of $\mathcal{O}\left(k^{-1}\right)$.

For the nonlinear case that we focus on, Doan (2022) sets a baseline convergence rate of $\mathcal{O}\left(k^{-2/3}\right)$ when f has a strongly monotonic property. A stronger smoothness condition leads to a single-time-scale convergence rate of $\mathcal{O}\left(k^{-1}\right)$, as shown in Shen and Chen (2022). Recently, Hu, Doshi, and Eun (2024) established a central limit theorem for convergence with Markov noises, while Han, Li, and Zhang (2024) investigated the case with local linearity, achieving convergence rates of $\mathcal{O}\left(\alpha_k\right)$ and $\mathcal{O}\left(\beta_k\right)$. Doan (2024) studied a variant in which f and g have a special form, also achieving a convergence rate of $\mathcal{O}\left(k^{-1}\right)$.

In cases without strong monotonicity, Shen and Chen (2022) and Zeng, Doan, and Romberg (2024) show different convergence rates under certain smoothness conditions.

In the field of SA, Ilandarideva et al. (2023) and Karandikar and Vidyasagar (2023) recently introduced a general noise assumption termed 'state-dependent noise'. We focus on specific cases of their assumptions, which are connected with the 'overparametrized model' discussed in Sebbouh, Gower, and Defazio (2021). In this particular case, $\mathbb{E}[\|\xi_k\|^2] \to 0$ as $x_k \to x^*$, indicating that the model is overparametrized and has no uncertainty.

Contribution

By requiring that martingale differences $\{\xi_k\}$ and $\{\psi_k\}$ have decreasing variances with respect to state x_k and y_k with parameters δ_{ij}, or time k with parameters γ_i, we are able to improve the convergence rate of $\mathcal{O}\left(k^{-2/3}\right)$ in Doan (2022) and $\mathcal{O}\left(k^{-1}\right)$ in Shen and Chen (2022).

More importantly, our work provides novel insights into understanding the gap between stochastic and deterministic two-time-scale approximations. We show explicitly how the convergence rate depends on the parameters δ_{ij} or γ_i by introducing a linear programming form $m(x)$. In the proof, we propose a novel technique based on Bernoulli's inequality for balancing coefficients under different parameters. Surprisingly, we find that there could be a convergence rate of $\mathcal{O}\left(k^{-\infty}\right)$ for both parameters, and even an exponential rate of $\mathcal{O}\left(e^{-\epsilon k}\right)$ when the parameter δ_{ij} is equal to 1.

Outline

Section 2 formulates the problem and defines notations. Sections 3 and 4 demonstrate theoretical results on convergence rates under state-dependent and time-dependent noises, respectively. Section 5 conducts numerical experiments. Section 6 concludes. The supplementary material provides proofs of all theoretical results (Appendix A) and figures of numerical experiments (Appendix B).

2 Problem Setup

In this section, we introduce several assumptions and establish several basic lemmas for convergence analysis. Generally, these assumptions are set to ensure appropriate propositions of update functions and steady convergence of the two-time-scale SA. Assumptions 1–3 below are commonly used in previous research (Doan 2022; Shen and Chen 2022). However, Assumptions 4–6 introduced here shed light on a new aspect of the noise term, which will be discussed later.

Assumptions

We assume that there is a unique solution (x^*, y^*) s.t.

$$\begin{aligned} f(x^*, y^*) &= 0, \\ g(x^*, y^*) &= 0. \end{aligned} \tag{6}$$

Assumption 1. *Given y there exists an operator λ such that $x = \lambda(y)$ is the unique solution of*

$$f(\lambda(y), y) = 0. \tag{7}$$

Suppose λ is Lipschitz continuous with respect to constant L_λ,

$$\|\lambda(y_1) - \lambda(y_2)\| \le L_\lambda \|y_1 - y_2\|. \tag{8}$$

Given a fixed y, this assumption guarantees a unique equilibrium solution $\lambda(y)$ of x, which is the unique update target of x. Moreover, the Lipschitz smoothness of λ ensures that the update target of x does not change significantly when y changes. These two properties of $\lambda(y)$ will be key when dealing with the differences among x, $\lambda(y)$, and x^*.

Assumption 2. *f is Lipschitz continuous with positive constant L_f, i.e. $\forall x_1, x_2, y_1$, and y_2,*

$$\|f(x_1, y_1) - f(x_2, y_2)\| \le L_f(\|x_1 - x_2\| + \|y_1 - y_2\|), \tag{9}$$

and f is strongly monotone w.r.t x when y is fixed, i.e. there exists a constant $\mu_f > 0$,

$$\langle x_1 - x_2, f(x_1, y) - f(x_2, y)\rangle \ge \mu_f \|x_1 - x_2\|^2. \tag{10}$$

With Lipschitz continuity and strong monotonicity of f, the update of x is controlled to be close to its target. This can be replaced by weaker requirements when only considering linear two-time-scale SA convergence, as shown in Konda and Tsitsiklis (2004) and Kaledin et al. (2020).

Assumption 3. *The operator $g(\cdot,\cdot)$ is Lipschitz continuous with constant L_g, i.e. $\forall x_1, x_2, y_1$, and y_2,*

$$\|g(x_1,y_1)-g(x_2,y_2)\| \leq L_g(\|x_1-x_2\|+\|y_1-y_2\|). \tag{11}$$

Moreover, g is 1-point strongly monotone w.r.t y^*, i.e. there exists a constant $\mu_g > 0$ such that

$$\langle y-y^*, g(\lambda(y),y)\rangle \geq \mu_g\|y-y^*\|^2. \tag{12}$$

For g, similar requirements are necessary to guarantee the convergence of y, where the strong monotonicity of g is directly related to the final solution y^*. This condition is widely used in the literature on nonlinear two-time-scale studies due to the non-linear nature of the problems involved. Note that strong monotonicity is a weaker condition than strong convexity and covers more application scenarios.

To better illustrate the update process, we introduce two residual variables:

$$\hat{x}_k = x_k - \lambda(y_k), \tag{13}$$

$$\hat{y}_k = y_k - y^*. \tag{14}$$

These variables represent the differences between x_k, y_k, and their respective update targets $\lambda(y_k)$ and y^*. We will show that the residual variables decrease to 0 as $k \to +\infty$ when investigating the Lyapunov function.

In the context of state- or time-dependent SA, the variance of the stochastic term can be a function of state or time (Karandikar and Vidyasagar 2023), rather than being simply bounded. We formulate the following assumptions and introduce parameters δ_{ij} and γ_i.

Assumption 4. *We denote by $\mathcal{Q}_k$ the filtration containing all the history generated by up to the iteration k, i.e.,*

$$\mathcal{Q}_k = \{x_0, y_0, \xi_0, \psi_0, \xi_1, \psi_1, \ldots, \xi_k, \psi_k\}, \tag{15}$$

where $\{\xi_k\}_{k\geq 0}$ are independent random variables with zero mean and bounded variances a.s., and so $\{\psi_k\}_{k\geq 0}$ are. Specifically, there exist Γ_{11} and Γ_{22} such that variances can be controlled as follows almost surely,

$$\mathbb{E}[\|\xi_k\|^2|\mathcal{Q}_{k-1}] \leq \Gamma_{11}\|\hat{x}_k\|^{2\delta_{11}} + \Gamma_{12}\|\hat{y}_k\|^{2\delta_{12}}, \tag{16}$$

$$\mathbb{E}[\|\psi_k\|^2|\mathcal{Q}_{k-1}] \leq \Gamma_{21}\|\hat{x}_k\|^{2\delta_{21}} + \Gamma_{22}\|\hat{y}_k\|^{2\delta_{22}}. \tag{17}$$

where $\delta_{11}, \delta_{12}, \delta_{21}$, and δ_{22} are in $[0,1)$. Meanwhile, denote $\Delta_{ij} = 1-\delta_{ij}$ for all $i,j \in \{1,2\}$.

Assumption 5. *$\{\xi_k\}_{k\geq 0}$ and $\{\psi_k\}_{k\geq 0}$ are martingale sequences similarly defined in Assumption 4 with $\delta_{ij} \equiv 1$, in other words a.s.,*

$$\mathbb{E}[\|\xi_k\|^2|\mathcal{Q}_{k-1}] \leq \Gamma_{11}\|\hat{x}_k\|^2 + \Gamma_{12}\|\hat{y}_k\|^2, \tag{18}$$

$$\mathbb{E}[\|\psi_k\|^2|\mathcal{Q}_{k-1}] \leq \Gamma_{21}\|\hat{x}_k\|^2 + \Gamma_{22}\|\hat{y}_k\|^2. \tag{19}$$

Here, we introduce four variables $\{\delta_{ij}\}$ to demonstrate the relationship between the variances of stochastic terms and two errors. If $\delta_{ij} \equiv 0$, it reduces to the form already studied in Doan (2022). However, when the stochastic term is generated from a more general form of stochastic observation, we will have $\delta_{ij} > 0$, which are the so-called state-dependent noise assumptions introduced by Ilandarideva et al. (2023) and Karandikar and Vidyasagar (2023). Assumptions 4 and 5 introduce more fundamental assumptions of an overparametrized model studied in Sebbouh, Gower, and Defazio (2021) with respect to two-time-scale convergence. Examples are widely used in Robbins-Monro setting SA, least squares, logistic regression (Moulines and Bach 2011), and articles on stochastic gradient descent (Chen et al. 2020).

With $\delta_{ij} > 0$, when x_k and y_k are closer to their targets, the variance of the stochastic term will be lower. Intuitively, this can accelerate the convergence rate compared to the case where $\delta_{ij} \equiv 0$ because the variance decreases. We will theoretically demonstrate that these assumptions accelerate the convergence rate. Additionally, there is a qualitative acceleration when $\delta_{ij} \equiv 1$, where Assumption 5 is sufficient to achieve an exponential convergence rate of $\mathcal{O}\left(e^{-\epsilon k}\right)$.

It should be noted that we only make separate assumptions about the variance for both $\{\xi_k\}$ and $\{\psi_k\}$, respectively. However, we do not impose any limits on the correlation between the two martingale differences.

For simplicity, we only consider the condition under a certain form of coefficients α_k and β_k, where

$$\alpha_k = \frac{\alpha}{(k+1+k_0)^a}, \beta_k = \frac{\beta}{(k+1+k_0)^b}. \tag{20}$$

Assumption 6. *$\{\xi_k\}_{k\geq 0}$ and $\{\psi_k\}_{k\geq 0}$ are similarly defined in (15), and for a fixed k_0 there exist constants $\Gamma'_{11}, \Gamma'_{22}, \gamma_1, \gamma_2 \geq 0$ such that $\gamma_1 - \gamma_2 \in [-1, \frac{1}{2})$ and variances can be controlled as follows almost surely,*

$$\mathbb{E}[\|\xi_k\|^2|\mathcal{Q}_{k-1}] \leq \Gamma'_{11}(k+1+k_0)^{-\gamma_1}, \tag{21}$$

$$\mathbb{E}[\|\psi_k\|^2|\mathcal{Q}_{k-1}] \leq \Gamma'_{22}(k+1+k_0)^{-\gamma_2}. \tag{22}$$

In contrast to the former definition, this could be regarded as time-dependent noise assumptions. It represents a special case of the conditions studied in Karandikar and Vidyasagar (2023). We will also study how the convergence rate changes with respect to γ_i. Similarly, as iteration k increases, the variance of the update will decrease, potentially accelerating the convergence rate. However, as $\gamma_i \to \infty$, it will not achieve an exponential convergence rate. This reflects an inherent difference between state-independent and time-dependent two-time-scale SA. Note that the condition $\gamma_1 - \gamma_2 \in [-1, \frac{1}{2})$ is necessary in the proof of Theorem 3 to balance the two different convergence rates.

Connection Between Assumptions and Application

Assumptions 5–6, along with the exponential convergence rate, are intended to demonstrate a specific scenario where fast convergence is achievable and valuable in practical applications of stochastic approximation. Although these assumptions may appear restrictive, they are motivated by and

have strong connections with control systems and reinforcement learning, where exponential convergence is essential for efficient operations. For example, recent studies such as Chen et al. (2020) incorporate a second-order moment assumption for SGD analysis, and Kaledin et al. (2020) highlights how contraction-based assumptions can accelerate convergence, even in multi-scale frameworks.

Moreover, Assumption 5 supports various applications in non-linear optimization and two-time-scale stochastic control, where stringent convergence rates facilitate computational efficiency. Practical examples include high-frequency trading algorithms Abergel et al. (2016); Gatheral (2011) and adaptive filtering Haykin (2002), where stability and rapid convergence are critical to performance. Furthermore, recent work on stochastic gradient methods Bottou, Curtis, and Nocedal (2018) underscores the importance of exponential rates in improving the training efficiency of complex models. By rigorously analyzing this case, we provide a theoretical benchmark that could inform future research in settings where fast convergence is feasible and advantageous.

Lemmas

In general, we have

$$\hat{x}_{k+1} = \hat{x}_k - \alpha_k f(x_k, y_k) + \lambda(y_k) - \lambda(y_{k+1}) - \alpha_k \xi_k, \tag{23}$$

$$\begin{aligned}\hat{y}_{k+1} = \hat{y}_k &- \beta_k g(\lambda(y_k), y_k) \\ &+ \beta_k (g(\lambda(y_k), y_k) - g(x_k, y_k)) - \beta_k \psi_k.\end{aligned} \tag{24}$$

Here, we show the upper bounds for $\hat{x}_k$ and $\hat{y}_k$. Prior to this, we propose the following lemmas to provide the recursive relational formulas for errors. Here we denote $a \lesssim b$ when there exists an irrelevant constant c such that $a \leq cb$.

Lemma 1. *Suppose that Assumptions 1–4 hold. Let x_k and y_k be updated by (1). Then, for all $k \geq 0$, we have*

$$\begin{aligned}&\mathbb{E}[\|\hat{x}_{k+1}\|^2 | \mathcal{Q}_{k-1}] - (1 - \mu_f \alpha_k)\|\hat{x}_k\|^2 \\ \lesssim &(\beta_k + \alpha_k^2)\|\hat{x}_k\|^2 + (\beta_k + \beta_k^2)\|\hat{y}_k\|^2 \\ &+ \frac{\beta_k^2}{\alpha_k}\mathbb{E}[\|\psi_k\|^2 | \mathcal{Q}_{k-1}] + \alpha_k^2 \mathbb{E}[\|\xi_k\|^2 | \mathcal{Q}_{k-1}].\end{aligned} \tag{25}$$

Proof. We provide a detailed statement and proof in Appendix A.1 of the supplementary material. □

Lemma 2. *Suppose that Assumptions 1–4 hold. Let x_k and y_k be generated by (1). Then, for all $k \geq 0$, we have*

$$\begin{aligned}&\mathbb{E}[\|\hat{y}_{k+1}\|^2 | \mathcal{Q}_{k-1}] - (1 - \mu_g \beta_k)\|\hat{y}_k\|^2 \\ \lesssim &(\beta_k + \beta_k^2)\|\hat{x}_k\|^2 + \beta_k^2 \|\hat{y}_k\|^2 + \beta_k^2 \mathbb{E}[\|\psi_k\|^2 | \mathcal{Q}_{k-1}].\end{aligned} \tag{26}$$

Proof. We provide a detailed statement and proof in Appendix A.2 of the supplementary material. □

These two lemmas provide a well-controlled upper bound on the expected error between two adjacent iterations, establishing our basic foundation. We also observe that the two main terms that affect the convergence rate are $\alpha_k \|\hat{x}_k\|^2$ and $\beta_k \|\hat{y}_k\|^2$. These terms play an important role when considering the Lyapunov function stated below.

We define $c = 4\frac{L_g^2}{\mu_f \mu_g}$ and the Lyapunov function here, denote it as V_k, is:

$$V_k = V(\hat{x}_k, \hat{y}_k) = c\frac{\beta_k}{\alpha_k}\|\hat{x}_k\|^2 + \|\hat{y}_k\|^2. \tag{27}$$

An insightful observation is that we need to balance two different variables, x_k and y_k. This will lead to the aforementioned Lyapunov function V_k.

Combining the above two lemmas allows us to analyze the update of the Lyapunov function in the next lemma. It is important to note that the difference between two iterations depends primarily on the magnitude of β_k and the distinct forms of $\mathbb{E}[\|\xi_k\|^2]$ and $\mathbb{E}[\|\psi_k\|^2]$.

Lemma 3. *Suppose that Assumptions 1–4 hold. Let x_k and y_k be generated by (1). Suppose α_k and β_k decrease (not necessarily strictly), and $\frac{\alpha_k}{\beta_k}$ is sufficiently large. With a constant $C(\alpha_0, \beta_0)$ generated by α_0 and β_0, we have*

$$\begin{aligned}&\mathbb{E}[V_{k+1}] - (1 - \frac{1}{2}\mu_g \beta_k)\mathbb{E}[V_k] \lesssim C(\alpha_0, \beta_0)\alpha_k^2 \mathbb{E}[V_k] \\ &+ \alpha_k \beta_k \mathbb{E}[\|\xi_k\|^2] + \left(\frac{\beta_k^3}{\alpha_k^2} + \beta_k^2\right)\mathbb{E}[\|\psi_k\|^2].\end{aligned} \tag{28}$$

Proof. We provide a detailed statement and proof in Appendix A.3 of the supplementary material. □

3 Convergence Under State-Dependent Noise Assumptions

Now we consider the convergence results under certain forms of coefficients α_k and β_k as in (20). These step sizes form an algebraic fraction with respect to k with a translation to ensure a moderate update rate. Here, we denote $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. In Assumption 4, we define a linear programming function

$$\begin{aligned}m(x) = &\frac{(1 + \delta_{11})x - \delta_{11}}{1 - \delta_{11}} \wedge \frac{x}{1 - \delta_{12}} \\ &\wedge \frac{(2 - \delta_{21})(1 - x)}{1 - \delta_{21}} \wedge \frac{2(1 - x)}{1 - \delta_{22}},\end{aligned} \tag{29}$$

which is set to be the minimum of four linear functions. $m(x)$ depending on all four parameters δ_{ij} is a delicate function set to satisfy

$$\begin{aligned}&-1 + 2a \\ \geq &-1 + a + t + (1 - a - t)\delta_{11} \vee -1 + a + t - t\delta_{12} \\ &\vee -3 + 4a + t + (1 - a - t)\delta_{21} \vee -3 + 4a + t - t\delta_{22},\end{aligned} \tag{30}$$

where $a = \operatorname{argmax}_{x \in (\frac{1}{2}, 1]} m(x)$ and $t = \max_{x \in (\frac{1}{2}, 1]} m(x)$. This balances all terms with different orders in the proof and achieves the maximum convergence rate of $\mathcal{O}(k^{-t})$. Here, we draw the following conclusion.

Theorem 1. *Suppose that Assumptions 1–4 hold. Let x_k and y_k be generated by (1). We assume*

$$b = 1, a = \operatorname*{argmax}_{x \in (\frac{1}{2}, 1]} m(x), t = \max_{x \in (\frac{1}{2}, 1]} m(x). \tag{31}$$

In addition, β and $\frac{\alpha}{\beta}$ are sufficiently large. Let $C_{\alpha\beta} = C(\alpha,\beta) \geq C(\alpha_0,\beta_0)$ defined in Lemma 3 and set k_0 as large enough. There exist two functions generated by M with $C_1(M) = \mathcal{O}(M)$ and $C_2(M) = \mathcal{O}\left(M^{\delta_{11}\vee\delta_{12}\vee\delta_{21}\vee\delta_{22}}\right)$ as $M \to +\infty$. Note that $\delta_{ij} < 1$, there exists a M s.t.

$$M \geq 3k_0^t V_0 \vee \frac{3}{a} C_2(M). \tag{32}$$

Together, we have $\forall k \geq 0$,

$$\mathbb{E}[V_k] \leq \frac{M}{(k+k_0)^t}. \tag{33}$$

Proof. We provide a detailed statement and proof in Appendix A.4 of the supplementary material. □

Theorem 1 provides a general view of how the convergence rate relates to the parameters δ_{ij} through a linear programming function $m(x)$. To better understand the implications of Theorem 1, particularly regarding the behavior of $m(x)$, we provide two enlightening special cases.

Corollary 1. *Suppose that Assumptions 1–4 hold. Let x_k and y_k be generated by (1). Moreover, we assume $\delta_{11} = \delta_{12}$ and $\delta_{21} = \delta_{22}$. Recall that $\Delta_{ij} = 1 - \delta_{ij}$, then under the condition of Theorem 1,*

$$a = \frac{\Delta_{11}+\Delta_{22}}{\Delta_{11}+2\Delta_{22}}, t = \frac{1+\Delta_{22}}{\Delta_{11}+2\Delta_{22}}. \tag{34}$$

There exist α, β, k_0, and M, s.t. $\forall k \geq 0$,

$$\mathbb{E}[V_k] \leq \frac{M}{(k+k_0)^t}. \tag{35}$$

Proof. We provide a detailed statement and proof in Appendix A.5 of the supplementary material. □

It is evident that under the assumption that $\delta_{11} = \delta_{12}$ and $\delta_{21} = \delta_{22}$, if $\delta_{11} \neq 0$, we always have $a < t$. This implies that the condition $\delta_{ij} > 0$ indeed accelerates the convergence rate to be faster than $\mathcal{O}(k^{-a})$, which is not demonstrated in earlier studies. Furthermore, when $\delta_{ij} = 0 (\Delta_{ij} = 1)$, we find that $a = t = \frac{2}{3}$. This is consistent with the results in Doan (2022).

Moreover, this bridges the gap between stochastic and deterministic two-time-scale SA. The latter is well-known for achieving at least an exponential convergence rate of $\mathcal{O}\left(e^{-\epsilon k}\right)$. As $\delta_{ij} \to 1^- (\Delta_{ij} \to 0^+)$, we observe that $t \to +\infty$, which naturally leads to the following result.

Now we investigate the case in Assumption 5 with constant learning rates $\alpha_k = \alpha$ and $\beta_k = \beta$. We define $c = 4\frac{L_g^2}{\mu_f \mu_g}$. The Lyapunov function is defined as follows and denoted as V_k,

$$V_k = V(\hat{x}_k, \hat{y}_k) = c\frac{\beta}{\alpha}\|\hat{x}_k\|^2 + \|\hat{y}_k\|^2. \tag{36}$$

Theorem 2. *Suppose that Assumptions 1–3,5 hold. Let x_k and y_k be generated by (1). Let $C_\beta = B_1 + B_2\beta$ with two constants B_1, B_2, and we set $\omega = \frac{\alpha}{\beta}$. There exist three functions generated by ω with $D_1(\omega) = -\frac{1}{2}\mu_g + \mathcal{O}\left(\frac{1}{\omega}\right)$ as $\omega \to +\infty$, $D_2(\omega)$, and $D_3(\omega)$. We assume ω to be sufficiently large and*

$$D_1(\omega) \leq -\frac{1}{4}\mu_g. \tag{37}$$

Then there exists a $\beta > 0$ s.t.

$$e^{-\epsilon} \triangleq 1 + D_1(\omega)\beta + D_2(\omega)\beta^2 + D_3(\omega)\beta^3 < 1. \tag{38}$$

Together, we have $\forall k \geq 0$,

$$\mathbb{E}[V_k] \leq e^{-\epsilon k} V_0. \tag{39}$$

Proof. We provide a detailed statement and proof in Appendix A.6 of the supplementary material. □

Combining Theorem 1 and 2, it is now clear that the boundary between the polynomial convergence rate and the exponential convergence rate occurs at $\delta_{ij} \equiv 1$. Specifically, if $\delta_{ij} < 1$, the convergence rate remains polynomial, but as $\delta_{ij} \to 1^-$, the convergence rate increases significantly. When $\delta_{ij} \equiv 1$, an exponential convergence rate is achieved.

4 Convergence Under Time-Dependent Noise Assumptions

We continue with the specified form of the coefficients α_k and β_k as in (20), and proceed to derive our conclusion under Assumption 6, where the noises satisfy the decay property at a rate of $\mathcal{O}(k^{-\gamma_i})$.

Theorem 3. *Suppose that Assumptions 1–3,6 hold. Let x_k and y_k be generated by (1). We assume*

$$b = 1, a = \frac{2-\gamma_1+\gamma_2}{3} \in (\frac{1}{2}, 1], t = \frac{2+2\gamma_1+\gamma_2}{3}. \tag{40}$$

In addition, β and $\frac{\alpha}{\beta}$ are sufficiently large. Let $C_{\alpha\beta} = C(\alpha,\beta) \geq C(\alpha_0,\beta_0)$ defined in Lemma 3 and set k_0 as large enough. There exists a function generated by M with $C_1(M) = \mathcal{O}(M)$ as $M \to +\infty$ and a constant C_2'. There exists a M s.t.

$$M \geq 3k_0^t V_0 \vee \frac{3}{a} C_2'. \tag{41}$$

Together, we have $\forall k \geq 0$,

$$\mathbb{E}[V_k] \leq \frac{M}{(k+k_0)^t}. \tag{42}$$

Proof. We provide a detailed statement and proof in Appendix A.7 of the supplementary material. □

Similarly, when $\gamma_i \equiv 0$, this reduces to the special case studied in Doan (2022). Furthermore, we observe that $t \to +\infty$ as $\gamma_i \to +\infty$, which aligns with the intuition that a higher order of the time-dependent noise parameters γ_i also accelerates convergence. However, note that the magnitude t increases linearly with respect to γ_i, which contrasts with the findings of Theorem 1. This difference explains why we cannot reach the same conclusion about the exponential convergence rate of $\mathcal{O}\left(e^{-\epsilon k}\right)$ as stated in Theorem 2.

5 Numerical Experiment

We conduct numerical experiments using two classical examples to illustrate the convergence rates in Sections 3 and 4. We initialize $(x_0, y_0) = (1, 1)$, and choose (α, β) to satisfy the same conditions as in Theorems 1-3. The step sizes α_k and β_k are set as $\alpha_k = \frac{\alpha}{(k+1+k_0)^a}$ and $\beta_k = \frac{\beta}{(k+1+k_0)^b}$, where k_0 is chosen optimally. It is important to note that k_0 increases rapidly with increasing (α, β) and affects the step sizes exponentially. Therefore, within constraints, we set the initial values of (α, β) as small as possible to prevent slow convergence caused by the excessively small initial step sizes (Moulines and Bach 2011). For simplicity of analysis, we set $\delta_{ij} \equiv \delta, \Gamma_{ij} \equiv \Gamma$, and $\gamma_k \equiv \gamma$. $\Gamma^{'}_{kk} \equiv \Gamma^{'}$ as noise parameters. All experiments are repeated 1000 times.

For $\delta \in [0, 1)$, we present log-log plots of the averaged V_k with respect to the number of iterations. For $\delta = 1$, we use the logarithmic plot. The slope of the line indicates the convergence rate. Specifically, a relationship of the form $y = rx^{-s}$ corresponds to $\log y = -s \log x + \log r$. To ensure stability, we calculate the slope of the error lines using iterations after 10^6.

SGD with Polyak-Ruppert averaging. We employ SGD with Polyak-Ruppert averaging (3) under 5-dimensional setting to minimize the function $F(x) = (x_1^2 + \sin x_1, \cdots, x_5^2 + \sin x_5)$ under normal white noise $\{\xi_k\}$, which satisfies either Assumption 4 or 5. The function $f(x, y) = \nabla F(x)$ meets Assumptions 1–3 in Section 2, with parameters set to $L_f = 3, \mu_f = 1, L_g = 2, \mu_g = 1$, and $L_\lambda = 0$.

Figure 1 illustrates the convergence results under various parameters. For $\delta_{ij} \equiv \delta \in [0, 1)$, we examine five values: $\delta = 0.0, 0.2, 0.4, 0.6, 0.8$ with $\Gamma = 0.02$. As shown in Figure 1, at the beginning of iterations, the convergence rate is no slower than our theoretical results. Moreover, with increasing iterations and the decay of errors (or Lyapunov functions V_k), the convergence rate is also close to our theoretical bound. Specifically, the slope closely approximates t as defined in Theorem 1. Moreover, as δ increases, the absolute value of the slope also increases, indicating faster convergence of the algorithm. This result is consistent with the intuition that a smaller noise volatility allows faster convergence without significant fluctuations. Figure 1 demonstrates that the algorithm actually achieves an exponential convergence rate of $\mathcal{O}\left(e^{-\epsilon k}\right)$ when $\delta = 1$ and $\Gamma^{'} = 0.1$.

Stochastic bilevel optimization. We consider an example of stochastic bilevel optimization (5) specified as follows:

$$F(x, y) = 10(x + \widetilde{h}_2(y))^2 + 10\sin(x + \widetilde{h}_2(y)), \quad (43)$$

$$G(x, y) = (x + \widetilde{h}_2(y))^2 + \sin(y) + y^2, \quad (44)$$

where $\widetilde{h}_2(z) = \text{sign}(z)z^2/2$ if $|z| \leq 1$ and $\text{sign}(z)(|z| - 1/2)$ otherwise. We directly take the partial derivative of equation (43) to obtain $f(x, y)$ and $g(x, y)$. It is straightforward to verify that Assumptions 1–3 hold with parameters set to $L_f = 60, \mu_f = 10, L_g = 3, \mu_g = 1$, and $L_\lambda = 3$. The noise terms $\{\xi_k\}$ and $\{\psi_k\}$ follow independent normal distributions satisfying either Assumption 4 or 5.

Figure 2 illustrates the convergence results of the stochastic bilevel optimization algorithm under two cases $\delta \in [0, 1)$ and $\delta = 1$. The convergence rate increases as δ approaches 1 within the interval $[0, 1)$. Notably, all experiments in this problem converge no slower than the theoretical predictions. Furthermore, the algorithm achieves an exponential convergence rate when $\delta = 1$, which matches the theoretical results of Theorem 2.

Moreover, we replicate the above experiments under time-dependent noise assumptions. Figure 3 displays the convergence results for the two examples mentioned above. We calculate the slope of the lines in Figure 3 using data from iterations after 10^5. The observed trends are analogous to those in the state-dependent case. Notably, when $\gamma = 4$, the error curve appears flat due to limitations in computational accuracy. For this reason, we compute the slope using data from iterations between 10^5 and 10^6 for $\gamma = 4$. As depicted in Figure 3, the convergence rate increases with higher values of γ, and all experiments in this problem converge even faster than the theoretical predictions.

6 Conclusion

This paper studies the convergence rate of two-time-scale SA under state- and time-dependent noises. The noises decay with respect to the state residual variables $\|\hat{x}_k\|^2$ and $\|\hat{y}_k\|^2$, or time (iteration number) k, respectively. We introduce a linear programming function $m(x)$ to establish a general relationship between the parameters δ_{ij} and the convergence rate of V_k. Specifically, in the state-dependent case where $\delta_{ij} \in [0, 1)$, the Lyapunov function V_k decays at a polynomial rate of $\mathcal{O}(k^{-t})$, and a similar result holds in the time-dependent case. Moreover, we derive explicit solutions for the optimal step size when all noise decay parameters are identical.

Furthermore, it is notable that two-time-scale algorithms can achieve an exponential convergence rate under the condition that the variances of noise are bounded by quadratic forms of the state variables, i.e., $\|\hat{x}_k\|^2$ and $\|\hat{y}_k\|^2$.

We present two numerical examples: SGD with Polyak-Ruppert averaging and stochastic bilevel optimization. The convergence results corroborate our theoretical insights. As $\delta_{ij} \to 1^-$ or $\gamma_i \to +\infty$, the exponent t in the convergence rate $\mathcal{O}\left(k^{-t}\right)$ increases, potentially reaching $\mathcal{O}\left(k^{-\infty}\right)$. Furthermore, numerical experiments indicate that when $\delta_{ij} \equiv 1$, the exponential convergence rate of $\mathcal{O}\left(e^{-\epsilon k}\right)$ in Theorem 2 can indeed be achieved.

In conclusion, our findings show that under the assumptions outlined in Ilandarideva et al. (2023) and Karandikar and Vidyasagar (2023), two-time-scale SA can achieve a convergence rate faster than $\mathcal{O}\left(k^{-1}\right)$ and, in certain cases, even attain an exponential convergence rate of $\mathcal{O}\left(e^{-\epsilon k}\right)$.

Acknowledgments

Research support from the National Key R&D Program of China (2022YFA1007900), the National Natural Science Foundation of China (12271013,72342004), Yinghua Education Foundation, and the Fundamental Research Funds for the Central Universities is gratefully acknowledged.

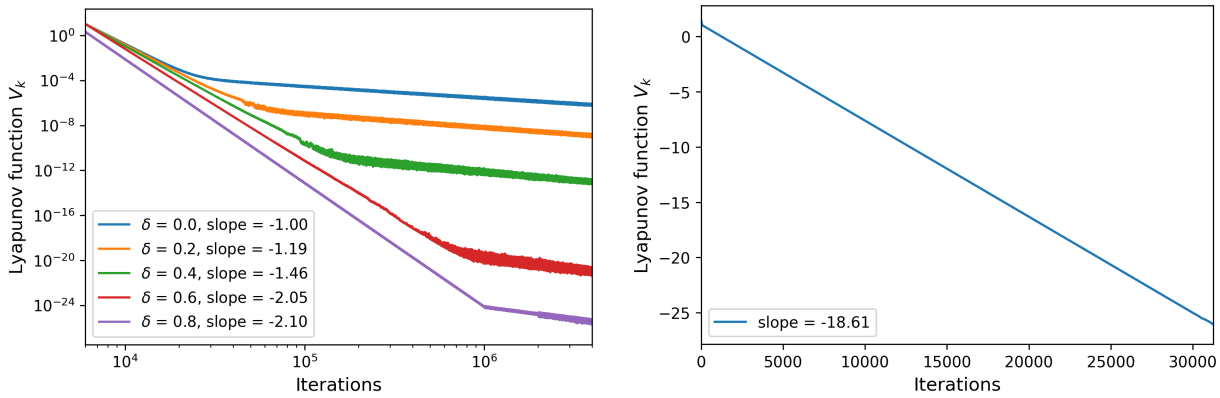

Figure 1: The convergence results of SGD with Polyak-Ruppert averaging. The figure on the left is a log-log plot in case $\delta = 0.0, 0.2, 0.4, 0.6, 0.8$; the figure on the right is a logarithmic plot in case $\delta = 1$. All R-squares of the fitted slopes do not exceed 1e-6.

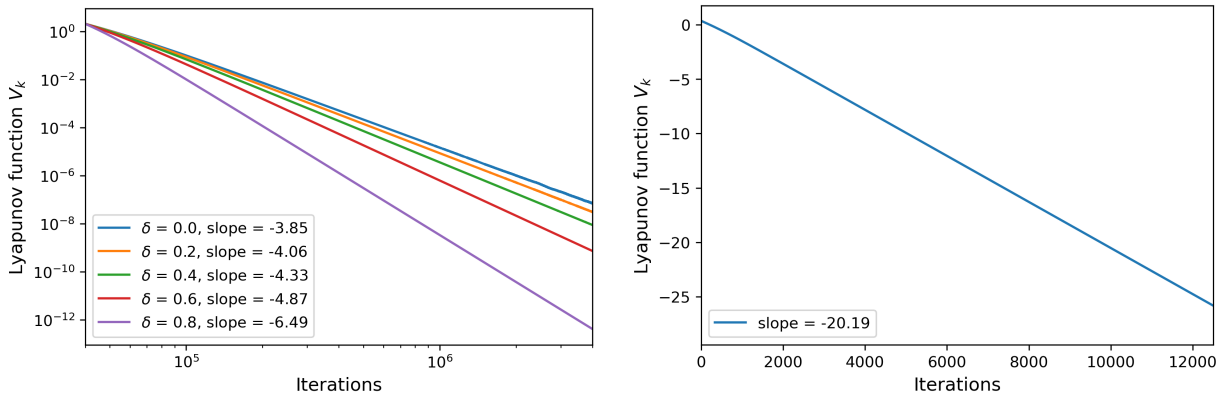

Figure 2: The convergence results of stochastic bilevel optimization. The figure on the left is a log-log plot in case $\delta = 0.0, 0.2, 0.4, 0.6, 0.8$; the figure on the right is a logarithmic plot in case $\delta = 1$. All R-squares of the fitted slopes do not exceed 0.01.

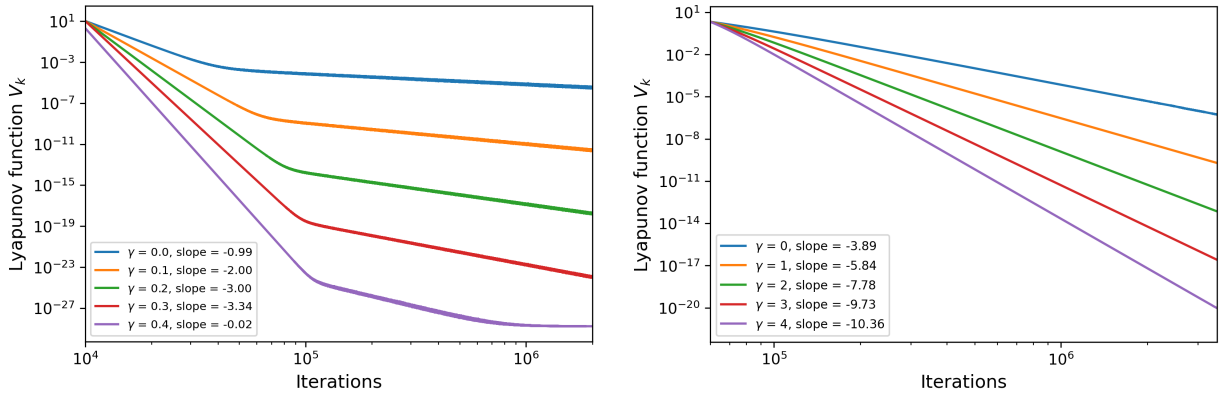

Figure 3: The convergence results of SGD with Polyak-Ruppert averaging (left) and stochastic bilevel optimization (right) under time-dependent noise assumption. The figure is a log-log plot in case $\gamma = 0, 1, 2, 3, 4$. All R-squares of the fitted slopes do not exceed 0.1.

References

Abergel, F.; Anane, M.; Chakraborti, A.; Jedidi, A.; and Toke, I. M. 2016. *Limit order books*. Cambridge University Press.

Borkar, V. S. 1997. Stochastic approximation with two time scales. *Systems & Control Letters*, 29(5): 291–294.

Borkar, V. S. 2009. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer.

Bottou, L.; Curtis, F. E.; and Nocedal, J. 2018. Optimization methods for large-scale machine learning. *SIAM review*, 60(2): 223–311.

Chen, X.; Lee, J. D.; Tong, X. T.; and Zhang, Y. 2020. Statistical inference for model parameters in stochastic gradient descent. *Annals of Statistics*, 48(1): 251–273.

Colson, B.; Marcotte, P.; and Savard, G. 2007. An overview of bilevel optimization. *Annals of operations research*, 153: 235–256.

Doan, T. T. 2022. Nonlinear two-time-scale stochastic approximation convergence and finite-time performance. *IEEE Transactions on Automatic Control*.

Doan, T. T. 2024. Fast Nonlinear Two-Time-Scale Stochastic Approximation: Achieving $\mathcal{O}(1/k)$ Finite-Sample Complexity. *arXiv preprint arXiv:2401.12764*.

Ermoliev, Y. 1983. Stochastic quasigradient methods and their application to system optimization. *Stochastics: An International Journal of Probability and Stochastic Processes*, 9(1-2): 1–36.

Ermoliev, Y. 2009. *Two-stage stochastic programming: quasigradient methodTwo-Stage Stochastic Programming: Quasigradient Method*, 3955–3959. Boston, MA: Springer US. ISBN 978-0-387-74759-0.

Gatheral, J. 2011. *The volatility surface: A Practitioner's Guide*. John Wiley and Sons, Inc.

Han, Y.; Li, X.; and Zhang, Z. 2024. Finite-Time Decoupled Convergence in Nonlinear Two-Time-Scale Stochastic Approximation. *arXiv preprint arXiv:2401.03893*.

Haykin, S. S. 2002. *Adaptive filter theory*. Pearson Education India.

Hu, J.; Doshi, V.; and Eun, D. Y. 2024. Central Limit Theorem for Two-Timescale Stochastic Approximation with Markovian Noise: Theory and Applications. *arXiv preprint arXiv:2401.09339*.

Ilandarideva, S.; Juditsky, A.; Lan, G.; and Li, T. 2023. Accelerated stochastic approximation with state-dependent noise. *arXiv preprint arXiv:2307.01497*.

Kaledin, M.; Moulines, E.; Naumov, A.; Tadic, V.; and Wai, H.-T. 2020. Finite time analysis of linear two-timescale stochastic approximation with Markovian noise. In *Conference on Learning Theory*, 2144–2203. PMLR.

Karandikar, R. L.; and Vidyasagar, M. 2023. Convergence Rates for Stochastic Approximation: Biased Noise with Unbounded Variance, and Applications. *arXiv preprint arXiv:2312.02828*.

Khodadadian, S.; Doan, T. T.; Romberg, J.; and Maguluri, S. T. 2022. Finite sample analysis of two-time-scale natural actor-critic algorithm. *IEEE Transactions on Automatic Control*.

Konda, V. R.; and Tsitsiklis, J. N. 2004. Convergence rate of linear two-time-scale stochastic approximation. *The Annals of Applied Probability*, 14(2): 796–819.

Lan, G. 2020. *First-order and stochastic optimization methods for machine learning*, volume 1. Springer.

Moulines, E.; and Bach, F. 2011. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24.

Pham, K. 2010. New risk-averse control paradigm for stochastic two-time-scale systems and performance robustness. *Journal of optimization theory and applications*, 146: 511–537.

Polyak, B. T.; and Juditsky, A. B. 1992. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4): 838–855.

Robbins, H.; and Monro, S. 1951. A stochastic approximation method. *The annals of mathematical statistics*, 400–407.

Ruppert, D. 1988. Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University Operations Research and Industrial Engineering.

Sebbouh, O.; Gower, R. M.; and Defazio, A. 2021. Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball. In *Conference on Learning Theory*, 3935–3971. PMLR.

Shen, H.; and Chen, T. 2022. A single-timescale analysis for stochastic approximation with multiple coupled sequences. *Advances in Neural Information Processing Systems*, 35: 17415–17429.

Sutton, R. S.; Maei, H. R.; Precup, D.; Bhatnagar, S.; Silver, D.; Szepesvári, C.; and Wiewiora, E. 2009. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th annual international conference on machine learning*, 993–1000.

Sutton, R. S.; Szepesvári, C.; and Maei, H. R. 2008. A convergent O(n) algorithm for off-policy temporal-difference learning with linear function approximation. *Advances in neural information processing systems*, 21(21): 1609–1616.

Wu, Y. F.; Zhang, W.; Xu, P.; and Gu, Q. 2020. A finite-time analysis of two time-scale actor-critic methods. *Advances in Neural Information Processing Systems*, 33: 17617–17628.

Xu, T.; Zou, S.; and Liang, Y. 2019. Two time-scale off-policy TD learning: Non-asymptotic analysis over Markovian samples. *Advances in neural information processing systems*, 32.

Zeng, S.; Doan, T. T.; and Romberg, J. 2024. A two-time-scale stochastic optimization framework with applications in control and reinforcement learning. *SIAM Journal on Optimization*, 34(1): 946–976.