Causal Reinforcement Learning: An Instrumental Variable Approach

Jin Li, Ye Luo, Xiaowei Zhang*
Faculty of Business and Economics, The University of Hong Kong, Pokfulam Road, Hong Kong, jli1@hku.hk, kurtluo@hku.hk, xiaoweiz@hku.hk

In the standard data analysis framework, data is first collected (once for all), and then data analysis is carried out. With the advancement of digital technology, decisionmakers constantly analyze past data and generate new data through the decisions they make. In this paper, we model this as a Markov decision process and show that the dynamic interaction between data generation and data analysis leads to a new type of bias—reinforcement bias—that exacerbates the endogeneity problem in standard data analysis.

We propose a class of instrument variable (IV)-based reinforcement learning (RL) algorithms to correct for the bias and establish their asymptotic properties by incorporating them into a two-timescale stochastic approximation framework. A key contribution of the paper is the development of new techniques that allow for the analysis of the algorithms in general settings where noises feature time-dependency.

We use the techniques to derive sharper results on finite-time trajectory stability bounds: with a polynomial rate, the entire future trajectory of the iterates from the algorithm fall within a ball that is centered at the true parameter and is shrinking at a (different) polynomial rate. We also use the technique to provide formulas for inferences that are rarely done for RL algorithms. These formulas highlight how the strength of the IV and the degree of the noise's time dependency affect the inference.

Key words: Endogeneity, Markov Decision Process, Instrumental Variable, Reinforcement Bias, Reinforcement Learning, Q-Learning, Stochastic Approximation

* We thank Chunrong Ai, Victor Chernozhukov, Jim Dai, Ivan Fernandez-Val, Jean-Jacques Forneron, Wei Jiang, Hiroaki Kaido, Tom Luo, as well as seminar participants at Boston University, HKUST, Peking University, and Shenzhen Research Institute of Big Data for helpful discussions. All remaining errors are ours.

1. Introduction

With the advancement of digital technology, business analytics has become increasingly popular. Data-based management has permeated into all policy aspects. Human resource managers rely on data to hire, evaluate, promote, and fire workers. Marketing departments channel their resources to digital advertising with objective metrics (such as the number of viewers and clicks). Operation managers use past data to forecast demand and determine inventory and pricing policy. Firms such as Amazon and Uber update their prices constantly in response to changes in market conditions.

The typical data-based management paradigm assumes, implicitly, that data generation precedes data analysis. The data is first collected, once and for all, and then data analysis is carried out. In practice, however, data generation can also follow data analysis. Managers analyze limited data to make decisions, and these decisions generate new data. Results from the analysis of past data, by affecting the decisions taken—and therefore the new data generated, will then also affect the results of the data analysis in the future. When data generation and data analysis intertwine and interact, how does this back-and-forth affect the long-term decision policies from the data analysis? In particular, if the data is imperfect, and various types of biases—omitted variable, measurement error, model mis-specification, selection, and simultaneous equations—are present, how do these biases manifest in the interaction between data generation and data analysis? And finally, and most importantly, what can be done to deal with these biases?

Consider, for example, a manager who decides each month how much to spend on digital marketing based on the number of monthly active users (MAUs) on a website. The manager cares about the additional sales generated, which depend on MAUs, the amount of spending on digital marketing, and an unknown parameter that measures its effectiveness. The manager, however, cannot observe the additional sales directly. Instead, she observes the total sales, which also depends on unobserved consumer sentiment that can increase the total sales and the number of MAUs. Consumer sentiment, therefore, is an omitted variable that is correlated with the MAUs.

To unpack the biases that arise when data generation and data analysis intertwine, suppose first that the data were generated under the optimal spending policy, which we may assume to be an increasing function of the MAUs. Because the manager doesn't know the effectiveness parameter, she may estimate it by running a regression of the total sales on the number of MAUs and spending (under the optimal policy). Her estimate, however, will be upward biased: When the spending is bigger (because of a bigger number of MAUs), it is likely that consumer sentiment is better, causing the total sales to also be higher. This is a standard omitted variable bias that makes the spending appear more effective than it actually is.

This upward bias leads the manager to believe that the optimal spending policy is wrong and adjusts her policy by increasing the spending (for each MAU). Under this new spending policy, new data will be generated and analyzed. However, the new data is affected by the new spending policy. Thus, the analysis on the new data again leads the manager to believe that this new spending policy is also suboptimal. This causes further adjustments in spending policies.

In this example, the manager plays the dual role of a data analyst and a data generator. As a data analyst, she analyzes existing data to estimate the effectiveness of digital marketing spending. As a data generator, she uses the estimate to decide on the digital marketing spending and generate new data on spending and total sales. This process continues *ad infinitum*, and in the limit, the estimated effectiveness parameter must be consistent with the effectiveness that guides the spending policies. In other words, the estimated parameter must be a fixed point.

This estimated parameter as a fixed point is different from the true parameter. The estimated parameter is also different from the estimate with omitted variable bias (under the optimal spending policy). The additional bias arises because the biases from the previous analysis affect the new data generated and can amplify the original bias in a fashion similar to the multiplier effect in macroeconomics. Therefore, we call such bias the *reinforcement bias* (R-bias).

To study the biases that arise out of the intertwining process of data generation and data analysis, and more importantly, the ways to correct for them, we consider a Markov decision process (MDP) framework. An agent makes a sequence of decisions over time to maximize his discounted rewards. In each period, the *true* reward is an unknown function of an observable current state variable and the agent's decision. However, the *observed* reward is a noisy measure of his true reward. Importantly, the noise can be correlated with the state variable or the action taken, creating bias in the data. Based on the biased data, the agent makes decisions and learns over time the reward function with bias, which conceivably makes his develop a biased, suboptimal policy that links the state variables to his actions.

Our first contribution is to formally establish the R-bias in a special case of the MDP framework. We analyze an example in which the state variables are independent and identically distributed (i.i.d.) over time so that the action has no effect on the evolution of the states, but the noise is correlated with the state variable or action. We calculate explicitly how the bias in policy is amplified because of the iterative process between learning and decision-making.

Our second contribution is to show how to correct for this bias by using an instrument variable (IV) approach. For the i.i.d. case, we propose an algorithm that incorporates the IV into a stochastic gradient descent (IV-SGD) routine and show that it asymptotically learns the optimal policy. For the general Markovian case, we introduce a class of IV-based reinforcement learning (RL) algorithms to again correct for the R-bias, using IV-Q-Learning as a primary example.[1] By using IVs, our

[1] Another approach to causal RL is based on "directed acyclic graph" (DAG) models. However, the DAG approach requires detailed domain knowledge of the causal relationships between states and actions, making it less applicable to large-scale economic problems with complex causal channels. See Imbens (2020) for more comments on use of DAG in social science.

approach can be viewed as performing a *causal exploration.* The notion of "exploration" in the conventional sense in RL algorithms help the agent better learn about the rewards by trying out seemingly inferior actions. But when observed rewards are biased, knowledge from these explorations is also biased, leading to suboptimal policies. In contrast, our IV-based approach, by perturbing the data-generating process, helps the agent identify true rewards and learn the optimal policies.

Our third contribution—the most significant one theoretically—is to establish a variety of theoretical guarantees for the performance of IV-RL algorithms by embedding them in the nonlinear two-timescale stochastic approximation (SA) framework and developing a new set of techniques to analyze them. In particular, we prove finite-time trajectory stability bounds for nonlinear two-timescale SA. They characterize the probability that, after a given finite time, the entire trajectory of the SA iterates will stay in a "shrinking ball" centered at the true parameter value and having a radius that decays at a polynomial rate. The bounds indicate that such probability approaches one with another polynomial rate once the shrinking ball is defined. Our result is substantially sharper than previous studies on trajectory stability, which focus on the probability that the trajectory stays in a "fixed ball" that has a constant radius (Dalal et al. 2018, Thoppe and Borkar 2019). In addition, the assumptions that underlie our results are more general, permitting unbounded state space and Markovian noise structure, whereas existing literature predominately relies on bounded state space and i.i.d. or martingale-difference noise. We are able to derive stronger results under weaker assumptions by developing a new set of techniques, particularly "polynomial partitioning with look-back", which can be used to deal with the technical challenges stemming from the Markovian noise structure.

These techniques also enable us to establish central limit theorems (CLTs) for the nonlinear two-timescale SA iterates, again, under assumptions more general than previous work (Kushner and Yin 2003, Konda and Tsitsiklis 2004, Mokkadem and Pelletier 2006). We further apply the CLTs to IV-Q-Learning to develop inference methods for the optimal policy. Relative to the existing RL literature that focus on risk bounds, our exact formula highlights how the asymptotic distribution is affected by the strength of the IV and the time dependency of the Markov noise.

Our paper contributes to the rapidly growing literature of RL algorithms, which have been widely used to develop artificial intelligence systems that substantially outperform human experts for solving complex, sequential decision-making problems (e.g., MDPs), including Atari video games (Mnih et al. 2015) and board games such as Go and chess (Silver et al. 2016, 2018); see Schrittwieser et al. (2020) for the most recent advance. Inspired by the remarkable success of RL in games, firms—particularly those exposed to a data-rich environment—are increasingly interested in adopting RL to automate their business decision processes, such as hiring, pricing, and advertising (Blake, Nosko, and Tadelis 2015, Calvano et al. 2020, Cowgill and Tucker 2020, Li, Raymond, and

Bergman 2020, Johnson, Rhodes, and Wildenbeest 2020). A critical difference between making decisions in real life and in a simulated environment like games is that the data generated in the former may well be subject to measurement errors or other issues that lead to biased signals. The present paper provides the foundation for applying RL to develop data-driven decision-making systems where the observed rewards are biased.

Our paper is broadly related to the vast literature of IV methods. Theoretical analysis of IVs in linear models includes Griliches (1977), Hausman (1983), Imbens and Angrist (1994), and Chen, Hong, and Tamer (2005). For IVs in nonlinear models, we refer to Hansen (1982), Ai and Chen (2003), Newey and Powell (2003), and Chernozhukov et al. (2007). Chernozhukov and Hansen (2005) apply IVs to quantile regression, and Belloni et al. (2012) discusses the theory of high-dimensional IVs. For reviews of identification problems in economics and the use of IVs in general, we refer to Manski (1995, Chapter 6), Angrist and Pischke (2009), and Imbens (2014). Most of the existing literature focuses on using IVs to estimate the causal effect of a single decision. However, we use IVs to learn the causal effect of a policy, which prescribes state-dependent decisions. One recent exception is Nambiar, Simchi-Levi, and Wang (2019), who use random experiments to create IVs to identify optimal policies in a setting where the state variables are i.i.d. Unlike our paper, they do not consider general IVs or general Markovian environments.

The rest of the paper is organized as follows. We provide an illustrative example of online learning and R-bias in Section 2. We cast IV regression into an online learning framework and propose the IV-SGD algorithm for parameter estimation in Section 3. We set up a general MDP framework with biased rewards and introduce IV-Q-Learning to correct for R-bias in Section 4. We formulate IV-RL algorithms, including IV-Q-Learning, as two-timescale SA in Section 5 and derive various theoretical properties there. We apply the general results for two-timescale SA to IV-Q-Learning in Section 6 to derive its theoretical properties, and develop tools for inference. We present a simulation to illustrate numerically the performance of IV-Q-Learning in Section 7 and make concluding remarks in Section 8. The Appendix collects all the technical proofs and discussions on other IV-RL algorithms.

Notation. Throughout this paper, we use $\|\cdot\|$ to denote the L^2 norm of a vector and maximum eigenvalue norm for a matrix. We use $\|\cdot\|_{\mathsf{F}}$ to denote the Frobenius norm for a matrix. For a square matrix M, we use $\mathrm{tr}(M)$ to denote the trace of M, and $\mathrm{vec}(M)$ to denote its vectorization by stacking its columns on top of one another. We use $\mathbb{E}[\cdot|\mathcal{F}_t]$ to denote the expectation conditional on information filtration at time period t. For probability measure ν, we use $\mathbb{E}_\nu[\cdot]$ to represent the expectation under measure ν. Moreover, $a = o_p(b)$ and $a = O_p(b)$ imply that $\frac{a}{b} \to 0$ and $|\frac{a}{b}| \leq C$ for some constant C, with probability going to one, respectively; $\rightsquigarrow$ represents convergence in distribution.

2. Reinforcement Bias: An Illustrative Example

In this section, we introduce and discuss R-bias via a digital advertising example. Consider a firm that chooses the digital advertising investment each period to maximize the firm's expected discounted return:

$$\max_{\{A_t\}_{t=0}^{\infty}} \mathbb{E} \sum_{t=0}^{\infty} \gamma^t \Big(\theta^* S_t A_t - \frac{1}{2} A_t^2\Big), \tag{1}$$

where S_t is the number of daily active users on the firm's website, and A_t is a measure of advertising intensity, e.g., the level of product exposure. To obtain the level of A_t, the associated cost is $\frac{1}{2}A_t^2$. Moreover, θ^* is an unknown parameter that measures the effectiveness of advertisement. We normalize the price of the product to be one, so that $(\theta^* S_t A_t - \frac{1}{2}A_t^2)$ represents the reward of the firm in period t.

This problem can be viewed as a standard MDP. The state variable in each period is S_t, and the action taken is A_t. Together, S_t and A_t affect the distribution of S_{t+1}. This MDP is typically analyzed by characterizing optimal value function V^* and optimal policy π^*, defined as follows:

$$V^*(s) \coloneqq \max_{\pi} \mathbb{E}_{S_0=s, A_t\sim\pi(S_t)} \sum_{t=0}^{\infty} \gamma^t \Big(\theta^* S_t A_t - \frac{1}{2} A_t^2\Big),$$

$$\pi^*(s) \coloneqq \arg\max_{\pi} \mathbb{E}_{S_0=s, A_t\sim\pi(S_t)} \sum_{t=0}^{\infty} \gamma^t \Big(\theta^* S_t A_t - \frac{1}{2} A_t^2\Big).$$

In general, there are two approaches to solving this class of problems. If the state transition and the reward distribution are known, one can use the standard dynamic programming method. This is often referred to as the "model-based" approach. If either of them is unknown, one can solve the problem via RL. This is often referred to as the "model-free" approach. In both approaches, however, it is typically assumed that the decisionmaker observes either the true reward or an *unbiased* noisy version of it.

2.1. Biases in Observed Rewards

In our example, the true reward of the firm is given by $r_t \coloneqq \theta^* S_t A_t - \frac{1}{2}A_t^2$, while the observed reward is given by $R_t \coloneqq r_t + \epsilon_t$, where ϵ_t is the noise. For instance, the true reward reflects the purchase due to digital advertisement, and the firm observes only the total purchase which is a noisy measure of what the firm truly cares about. Importantly, this noise may be correlated with the state variable or the action taken; i.e., $\mathbb{E}[R_t|S_t, A_t] \neq r_t$, creating a number of issues as will be discussed below.

There are a number of reasons for why such dependence arises. To see why the noise may be correlated with the action taken, notice that the firm typically is unable to measure perfectly the actual cost associated with A_t. For example, digital advertisement costs may include the opportunity cost of investment in other valuable activities, cannibalization of sales of the firm's other products, and additional non-pecuniary costs, such as manpower and coordination. These types of costs are difficult to measure and may contribute to systematic under- or overestimate of the true cost; see, e.g., Alchian and Demsetz (1972) and Williamson (1985) for classic discussions of the difficulties with measurements within organizations. Another reason for the observed reward to be biased has to do with incentive issues. For example, if the digital advertisement investment is championed by a senior manager, then analysts who evaluate the investment return may sugarcoat the data. Incentive concerns, therefore, can cause the managers to intentionally create biased measures; see, e.g., Holmström and Milgrom (1991), Prendergast (1993), and Morris (2001).

To see why the noise may be correlated with the state variable, recall from the introduction that the number of visitors to the website may be correlated with unobserved consumer sentiments. Such sentiment may make consumers more likely to visit the website as well as to purchase the product. This type of omitted variable problem is a common cause of a biased signal and has been studied extensively in the economics literature; see, e.g., Angrist and Krueger (1999). In addition to the omitted variable problem, other reasons that generate a similar form of endogeneity include the selection problem (Heckman 1979, 1990) and the simultaneous equation problem (Hausman 1983, Manski 1995, Chapter 6).

When the observed rewards are biased measures of the true rewards, policies from maximizing the observed rewards will not maximize a firm's true rewards. In other words, while the firm's number may look good, its true profit will suffer. Social scientists have long discussed the harm brought to all kinds of organizations when managers and workers maximize biased measures; see Muller (2018) for a comprehensive discussion. In the digital age, policies developed through algorithms that maximize biased rewards again mislead firms and cause them to take sub-optimal actions. In the next subsection, we provide an example to illustrate how sub-optimal policies arise when the firm maximizes biased rewards, and the rest of the paper discusses how to obtain optimal policies.

2.2. Reinforcement Bias

When the noise is dependent on the state variable or the action, a new form of bias will arise if the decisionmaker is also a data generator. To make things simple, consider a simple transition rule that S_{t+1} depends on S_t only. Namely, the action A_t has no strategic effect on S_{t+1} and beyond. As the true reward is given by $r_t = \theta^* S_t A_t - \frac{1}{2}A_t^2$, the optimal policy is $\pi^*(s) = \theta^* s$ for any state s. We focus on whether the parameter θ^* can be learned in the long run.

For illustrative purposes, we first consider the case in which the noise depends on the action only. In particular, suppose that

$$\epsilon_t = \beta A_t^2 + o_t,$$

where o_t's are i.i.d. random variables with mean zero. When $\beta > 0$, this corresponds with our previous discussion that the true cost is undermeasured in proportion.

Before going into the interactive environment, let us first assume that the firm uses a *fixed* linear policy $A_t = \tilde{\theta} S_t$, where $\tilde{\theta}$ is exogenously given and not necessarily the same as the true parameter value θ^*. To estimate θ^*, one common strategy is to regress $R_t + \frac{1}{2}A_t^2 = \theta^* S_t A_t + \epsilon_t$ on $S_t A_t$, assuming that the cost $\frac{1}{2}A_t^2$ is known. The standard regression formula shows that, in the long run, the estimate of θ^* is given by

$$\operatorname{plim} \hat{\theta}_{\mathsf{static}}(\tilde{\theta}) = \theta^* + \frac{\operatorname{Cov}(S_t A_t, \epsilon_t)}{\operatorname{Var}(S_t A_t)} = \theta^* + \frac{\operatorname{Cov}(\tilde{\theta} S_t^2, \beta \tilde{\theta}^2 S_t^2)}{\operatorname{Var}(\tilde{\theta} S_t^2)} = \theta^* + \beta \tilde{\theta}. \tag{2}$$

Therefore, the long-run bias of the estimate is $\beta\tilde{\theta}$. This bias reveals the endogeneity in the setting because the noise ϵ_t is correlated with the action taken. Notice that even if the firm chooses the optimal policy $A_t = \theta^* S_t$, such bias remains, and it equals $\theta^* + \beta\theta^*$.

Now, consider the interactive environment, where the firm can adjust its policy based on existing data. This adjustment, of course, affects the data generated in the future and will lead to further adjustment of the policy. In particular, suppose the firm starts with a policy $A_t = \hat{\theta}^0 S_t$, for some $\hat{\theta}^0$. If this policy is used for an extended period of time, then following (2) with $\tilde{\theta} = \hat{\theta}^0$, the estimate of θ^* from the data generated under this policy is given by $\hat{\theta}^0_{\mathsf{est}} = \theta^* + \beta\hat{\theta}^0$.

Suppose the firm uses this estimate as a guide and adjusts its policy to $A_t = \hat{\theta}^1 S_t$, where $\hat{\theta}^1 = \hat{\theta}^0_{\mathsf{est}}$. After this adjusted policy is used for an extended period of time, the new estimate of θ^* is then given by $\hat{\theta}^1_{\mathsf{est}} = \theta^* + \beta\hat{\theta}^1$. Continuing this process, the estimate of θ^* after k iterations is given by

$$\hat{\theta}^k_{\mathsf{est}} = \theta^* + \beta\hat{\theta}^{k-1} = \cdots = \theta^* + \beta\theta^* + \cdots + \beta^{k-2}\theta^* + \beta^{k-1}\hat{\theta}^0 \to \frac{\theta^*}{1-\beta},$$

as $k \to \infty$, provided that $\beta \in (-1, 1)$. We call the following the *R-bias*

$$\frac{\theta^*}{1-\beta} - \theta^* = \frac{\beta\theta^*}{1-\beta}.$$

Notice that the R-bias can also be calculated via a fixed-point argument. Suppose the long-run policy converges to $A_t = \hat{\theta} S_t$. Then, consistency between the estimate and the decision requires $\operatorname{plim} \hat{\theta}_{\mathsf{static}}(\hat{\theta}) = \hat{\theta}$, i.e., $\theta^* + \beta\hat{\theta} = \hat{\theta}$. This implies that $\hat{\theta} = \frac{\theta^*}{1-\beta}$, which, again, gives the same R-bias.

Thus far, we have examined the case that the noise is correlated with the action. Next, we consider the case that the noise is correlated with the state variable. Suppose, for example, that

$$\epsilon_t = \beta S_t^2 + o_t,$$

where, again, o_t's are i.i.d. random variables with mean zero. If the firm uses a fixed policy $A_t = \tilde{\theta} S_t$, the standard regression formula implies that the long-run estimate of θ is given by

$$\operatorname{plim} \hat{\theta}_{\mathsf{static}}(\tilde{\theta}) = \theta^* + \frac{\operatorname{Cov}(S_t A_t, \epsilon_t)}{\operatorname{Var}(S_t A_t)} = \theta^* + \frac{\operatorname{Cov}(\tilde{\theta} S_t^2, \beta S_t^2)}{\operatorname{Var}(\tilde{\theta} S_t^2)} = \theta^* + \frac{\beta}{\tilde{\theta}}.$$

Consequently, to calculate the estimate of θ in the interactive environment, the fixed-point argument used earlier implies that the long-run estimate should satisfy

$$\theta^* + \frac{\beta}{\hat{\theta}} = \hat{\theta}.$$

Thus, the R-bias is given by $\hat{\theta} - \theta^* = \frac{1}{2}(\sqrt{(\theta^*)^2 + 4\beta} - \theta^*)$.

2.3. Discussion

Comment 1. We have illustrated R-bias in two special scenarios; i.e., the noise depends only on the action, and the noise depends only on the state variable. More generally, the noise can depend on both at the same time. Moreover, the noise structure can be more general. In the digital advertisement example, the noise does not depend on what happens in the past, and the action does not affect the state transition. Relaxing these assumptions is important for real-world applications. We will discuss it in more detail in later sections. □

Comment 2. Our example studies digital advertising. There are many scenarios in which R-bias can arise, particularly because algorithms have become more prevalent in the workplace and the social arena. For example, algorithms dictate the key performance indicators (KPIs) of delivery riders for online shopping platforms such as Meituan-Dianping and Uber Eats. These KPIs—delivery time and customer ratings—are adjusted according to the actual performances of the riders. If these performances are not well measured (riders may violate traffic rules to speed up delivery), our mechanism implies that the KPIs will be distorted. This could possibly result in excessively demanding requirements for the riders, and in the long run, loss of profit for the firm and loss of welfare for the society (McMorrow and Liu 2020). Similarly, online news outlets often use algorithms today to send recommendations to readers based on their reading patterns. These patterns, of course, are affected by what are recommended, and therefore, the recommendation affects the future reading patterns and future recommendations. If the reading preferences are not perfectly measured, our mechanism then implies that, in the long run, the readers do not receive the best recommendations they can; many readers often complain about receiving repetitive articles or videos. Moreover, such recommendations may favor mainstream tastes and eventually, crowd out specialized content providers (Smith 2020). □

3. IV-SGD

Our approach to correcting for R-bias is based on IVs that have been widely used for causal estimation and inference. Unlike standard causal analysis which is usually done in a static environment, the problem that we try to address has two new features: First, the environment requires online learning rather than batch learning; second, the environment is interactive, leading to potential R-bias.

3.1. Instrument Variables

IV is a common and popular approach to deal with endogeneity in data. In general, IVs are random variables Z_t that are correlated with the covariates of the model—S_tA_t in the example in Section 2.2. In addition, we require that Z_t and ϵ_t are conditionally independent given (S_t, A_t). A weaker condition is that $Z_t\epsilon_t$ has mean 0 conditional on (S_t, A_t).

The existing applications of IV have mostly been causal inferences of the treatment effect in a static environment; see, for example, Angrist (1990), Angrist and Krueger (1991), Levitt (1997), Angrist and Evans (1998), and Imbens (2014), and Angrist and Pischke (2009) and Imbens (2014) for reviews. These studies typically rely on finding clever choices of instrumental variables for their analysis. But IVs can also be constructed directly. Using the digital advertisement example, below we show how to construct IV for the two cases that we analyzed before: The noise is dependent on S_t only, and the noise is dependent on A_t only.

Case 1: Noise correlated with S_t only. Let S_t be the number of visitors to the website in time period t. Both ϵ_t and S_t are positively affected by unobserved consumer sentiment, which causes the endogeneity issue. Therefore, valid IVs may be chosen to be variables that affect the website traffic without affecting the consumer sentiment. For example, firms can use Google Ads ranking of the website to cause exogenous variation in the number of visitors. In addition, firms can perturb the website traffic via randomized control trials (RCTs) such as A/B tests on the web design. Both policies can create valid IVs.

Case 2: Noise correlated with A_t only. When action A_t affects ϵ_t, it is generally more difficult to find valid IVs if A_t is generated by a *deterministic* policy as a function of S_t. One possible solution is to consider a *randomized* policy that executes random actions with certain probability in certain periods. When A_t is randomized, one can then find or generate exogenous shocks to S_t to serve as IVs, as these shocks affect S_t and are independent of ϵ_t and A_t. A similar idea is used by Nambiar, Simchi-Levi, and Wang (2019), where a randomized pricing policy is employed to correct for possible mis-specification of the demand model.

Notice that our approach points to a new role business experiments and RCTs can play. The existing emphasis of RCTs has been on helping compare alternative choices; see Athey and Imbens (2017) and Thomke (2020) for reviews. In our framework, RCTs serve as part of the learning process. It helps provide an exogenous shock that allows the algorithm (to be specify below) to consistently learn about the unknown parameter. Notice that such parameter informs not just a single choice but also a policy that maps different states to final choices. In other words, RCTs help us identify not just (state-independent) choices but also policies (that are state-dependent choices).

3.2. IV-SGD for Data from a Fixed Policy

We now cast IV regression into an online learning framework for a fixed policy, which can be applied to the digital advertisement example, and present IV-SGD to algorithmically solve for the estimate. This algorithm offers insight and motivates general IV-RL algorithms presented later sections.

Consider the linear regression model:

$$Y_t = X_t^\intercal \theta^* + \epsilon_t, \quad \forall t = 1, \dots, T,$$

where $\theta^* \in \mathbb{R}^p$ is a vector of unknown parameters, $X_t \in \mathbb{R}^p$ is a vector of covariates, $Y_t \in \mathbb{R}$ is the dependent variable, and $\epsilon_t \in \mathbb{R}$ is the noise. The goal is to estimate the value of θ^*. In this subsection, we assume that (X_t, Y_t) is generated by a fixed policy, and ϵ_t is potentially endogenous; i.e., $\mathbb{E}[\epsilon_t | X_t] \neq 0$. The interactive policy will be considered in the next subsection. In the digital advertisement example, $X_t = S_t A_t$ and $Y_t = R_t + \frac{1}{2} A_t^2 = \theta^* S_t A_t + \epsilon_t$. Here, ϵ_t can depend on S_t or A_t or both.

When all the data $\{(X_t, Y_t) : t = 1, \dots, T\}$ is collected at once, the estimation of θ can be formulated as an optimization problem:

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^p} \sum_{t=1}^{T} (Y_t - X_t^\intercal \theta_t)^2. \tag{3}$$

This approach is often known as "statistical learning", where an agent observes the data and learns key parameters of a model given the observations.

When the data arrives in a streaming fashion, estimation of θ^* is often achieved with online learning methods such as stochastic gradient descent (SGD); see, e.g., Bottou (2010). Specifically, the optimization problem (3) can be solved in the following sequential manner

$$\theta_{t+1} = \theta_t + \alpha_t X_t (Y_t - X_t^\intercal \theta_t), \quad \forall t = 1, \dots, T, \tag{4}$$

where $\alpha_t > 0$ is called the *learning rate* at time t.

Comment 3. Notice that one can also perform standard linear regression to update the estimate of θ^* each time a new data point arrives (Escobar and Moser 1993), which is also known as recursive least squares (RLS; Young 2011, Chapter 3). In a dynamic environment with big data, SGD is the preferred method due its computational advantages relative to RLS. This is because the former relies only on first-order information (i.e., the gradient) of the objective function, whereas the latter is essentially equivalent to using Newton's method to solve the optimization problem (3), which additionally requires second-order information (e.g., the Hessian). Such advantage is particularly important if predictions or decisions need to be made continuously and in real time. □

In the presence of endogeneity, it is well-known that standard linear regression in (3) leads to biased estimates of θ^*. If a set of IVs, denoted as $Z_t \in \mathbb{R}^q$, are available, one common approach is to use two-stage least squares (2SLS) to consistently estimate θ^*:

$$\text{First-Stage Regression:} \quad \hat{\Gamma} := \arg\min_{\Gamma \in \mathbb{R}^{p\times q}} \sum_{t=1}^{T} \|X_t - \Gamma Z_t\|^2,$$

$$\text{Second-Stage Regression:} \quad \hat{\theta} := \arg\min_{\theta \in \mathbb{R}^p} \sum_{t=1}^{T} (Y_t - Z_t^\intercal \hat{\Gamma}^\intercal \theta)^2,$$

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

To extend 2SLS to the data-streaming setting, we propose Algorithm 1, hereafter referred to as IV-SGD. In general, the learning rates α_t and β_t may converge to zero at different rates, which effectively implies that the estimates θ_t and Γ_t are updated at different timescales. The dynamics of (θ_t, Γ_t) in (5) and (6) is often called a two-timescale system; see, e.g., Borkar and Konda (1997).

Algorithm 1: IV-SGD for a Fixed Policy

Input : Learning rates (α_t, β_t)
Output : θ_T and Γ_T

```
Initialize θ_0 and Γ_0
for all t = 0, 1, ..., T−1 do
    Observe (X_t, Y_t, Z_t)
    Update the estimates of θ* and Γ* via
end
```

$$\theta_{t+1} = \theta_t + \alpha_t (Y_t - X_t^\intercal \theta_t) \Gamma_t Z_t \tag{5}$$

$$\Gamma_{t+1} = \Gamma_t + \beta_t (X_t - \Gamma_t Z_t) Z_t^\intercal \tag{6}$$

Comment 4. In Algorithm 1, θ_t converges to θ^* as t goes to infinity under mild conditions. (We relegate the formal statement and analysis of the result to Appendix E.) The updating equations (5) and (6) correspond to the second- and first-stage regressions in 2SLS, respectively. Thus, IV-SGD can be viewed as an algorithm that performs causal analysis at each time t in an online learning fashion. □

3.3. IV-SGD for Data from an Interactive Policy

Previously, we described IV-SGD for the fixed policy, we now study how to modify the IV-SGD when the firm chooses an interactive policy. We describe the algorithm using the digital advertising example. We will discuss algorithms for a more general setting in Section 4.

Suppose the firm can adjust its policy in the digital advertising example. We now provide an algorithm that allows the firm to learn the truth parameter in the long run. In particular, let the policy at time t be $A_t = \theta_t S_t$. For arbitrary θ_0 and Γ_0, we update the parameters as follows.

$$\theta_{t+1} = \theta_t + \alpha_t (Y_t - X_t \theta_t^2) \Gamma_t Z_t, \tag{7}$$

$$\Gamma_{t+1} = \Gamma_t + \beta_t (X_t - Z_t \Gamma_t) Z_t^\intercal, \tag{8}$$

where $X_t = S_t^2$, $Y_t = R_t + \frac{1}{2} A_t^2 = \theta^* \theta_t X_t + \epsilon_t$, and $Z_t \in \mathbb{R}^q$ is a vector of IVs.

One can show that θ_t converges to θ^* as time goes to infinity. This convergence result, however, is established using different techniques. The critical difference between the algorithm above and Algorithm 1 is that the updating scheme (7) is nonlinear in θ_t, whereas it is linear for (5). Therefore, a theory that can tackle nonlinear stochastic systems is needed. We develop this theory in Section 5, which is then applied to this algorithm. This theory is based on a general MDP framework, and in particular, it allows for Markovian noises.

4. Endogeneity in a General MDP Framework and RL Algorithms

We now consider a general sequential decision-making problem. An agent is engaged in an infinite-horizon MDP with state space $\mathcal{S}$, action space $\mathcal{A}$, and a discount factor $\gamma \in (0,1)$. At each time $t = 0, 1, \ldots$, the agent observes the current state $S_t = s \in \mathcal{S}$ of the environment and chooses policy π to take an action $A_t = a \in \mathcal{A}$. The policy can be either deterministic ($a = \pi(s)$) or random ($a \sim \pi(\cdot|s)$; i.e., $\pi(\cdot|s)$ is a probability distribution on $\mathcal{A}$ from which a is sampled). Following the action, reward $r(s,a) \in \mathbb{R}$ is realized, and state $S_{t+1} = s'$ occurs with probability $\mathsf{P}(s'|s,a) := \mathbb{P}(S_{t+1} = s' | S_t = s, A_t = a)$.

The agent chooses a policy π to maximize his expected cumulative discounted future reward:

$$V^*(s) := \max_{\pi} \mathbb{E}_{a \sim \pi(\cdot|s)} \left[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t) \Big| S_0 = s \right], \tag{9}$$

where the expectation is taken with respect to the probability distribution induced by the policy π. The function V^* defined in (9) represents the value of the state s under optimal policy π^*. Likewise, we define

$$Q^*(s,a) := r(s,a) + \max_{\pi} \mathbb{E}_{a \sim \pi(\cdot|s)} \left[\sum_{t=1}^{\infty} \gamma^t r(S_t, A_t) \Big| S_0 = s, A_0 = a \right], \tag{10}$$

where $Q^*(s,a)$ is the value of playing action a at state s while playing optimal policy afterward. By construction, the Q^* function satisfies the Bellman equation:

$$Q^*(s,a) = r(s,a) + \gamma \max_{a' \in \mathcal{A}} \mathbb{E}_{s' \sim \mathsf{P}(\cdot|s,a)} Q^*(s', a'). \tag{11}$$

Notice that

$$V^*(s) = \max_{a \in \mathcal{A}} Q^*(s,a),$$

and the optimal policy π^* satisfies

$$\pi^*(s) \in \arg\max_{a \in \mathcal{A}} Q^*(s,a).$$

In general, it is difficult to solve for $\pi^*(s)$ and $V^*(s)$ analytically and numerically, especially for complex MDPs with large state and action spaces. One can, however, approximate them using RL algorithms, which we discuss next.

4.1. Q-Learning

Q-Learning is a leading RL algorithm for solving complex MDP problems, achieving superb performances in Atari video games (Mnih et al. 2015) and board games such as Go (Silver et al. 2016) and chess (Silver et al. 2018). The basic idea of Q-Learning is to approximate the Q-function by iterating on the right-hand-side (RHS) of the Bellman equation (11). To carry out such iterations efficiently, one usually approximates the Q-function by linearization[2]; see, e.g., Bertsekas and Tsitsiklis (1996). Let $Q^*(s,a;\theta)$ denote the Q-function approximation parameterized by $\theta \in \mathbb{R}^p$, that is,

$$Q^*(s,a;\theta) := \phi^\intercal(s,a)\theta = \sum_{\ell=1}^{p} \theta_\ell \phi_\ell(s,a),$$

where $\phi(s,a) := (\phi_1(s,a), \ldots, \phi_p(s,a))^\intercal$ is a column vector of p basis functions.[3] In particular, $Q^*(s,a;\theta^*) = \phi^\intercal(s,a)\theta^*$ can be viewed as the "best" linear projection of $Q^*(s,a)$ on the linear space spanned by $\phi(s,a)$; see, e.g., Melo, Meyn, and Ribeiro (2008). Under such approximation, Q-Learning updates the estimate of θ^* as follows:

$$\theta_{t+1} = \theta_t + \alpha_t \Big[r_t + \gamma \max_{a' \in \mathcal{A}} (\phi^\intercal(S_{t+1}, a')\theta_t) - \phi^\intercal(S_t, A_t)\theta_t \Big] \phi(S_t, A_t), \tag{12}$$

where $r_t = r(S_t, A_t)$ is the true reward, $\alpha_t > 0$ is learning rate at time t, and A_t is generated by some behavior policy π.

Notice that the recursive equation (12) bears a resemblance to SGD-type algorithms like (4). It solves for the fixed point θ^* the equation

$$\mathbb{E}_{S_{t+1} \sim \mathrm{P}(\cdot | S_t = s, A_t = a)} \Big[r_t + \gamma \max_{a' \in \mathcal{A}} (\phi^\intercal(S_{t+1}, a')\theta^*) - \phi^\intercal(S_t, A_t)\theta^* | S_t = s, A_t = a \Big] = 0,$$

[2] The function Q and policy π can also be approximated with nonlinear architectures such as deep neural networks; see François-Lavet et al. (2018) for a recent introduction.

[3] The features may be chosen based on domain knowledge of the specific problem or extracted using sophisticated feature engineering techniques such as autoencoders (Rumelhart and McClelland 1986, Vincent et al. 2008). In general, there is an approximation error given a fixed set of p basis functions and such error should decay as p grows. In this paper, we focus on analyzing the IV-RL algorithms in fixed dimensions for simplicity and leave the non-parametric problems to future research.

which is derived by combining the Bellman equation (11) with the linear approximation. Standard convergence analysis of Q-Learning can be found in Tsitsiklis (1994) and Melo, Meyn, and Ribeiro (2008).

4.2. IV-Q-Learning

The convergence properties of Q-Learning are mostly obtained under the premise that the observed reward R_t is either the true reward r_t or an unbiased measure of r_t. As discussed before, there are many cases where the reward $R_t = r(S_t, A_t) + \epsilon_t$ is biased; i.e., $\mathbb{E}[\epsilon_t | S_t, A_t] \neq 0$. In these cases, the Q-Learning algorithm would conceivably lead to suboptimal policies in the long run.

To correct for such biases, we again take an IV approach that is similar to IV-SGD. Suppose, in particular, that there exists $Z_t \in \mathbb{R}^q$, a vector of IVs, available at time t. Formally, we require that

$$\mathbb{E}[Z_t \epsilon_t | S_t, A_t] = 0,$$

and there exists a full rank matrix $\Gamma^* \in \mathbb{R}^{p \times q}$ such that

$$\phi(S_t, A_t) = \Gamma^* Z_t + \eta_t, \tag{13}$$

where $\mathbb{E}[\eta_t | Z_t, X_t] = 0$ and $\mathbb{E}[Z_t \epsilon_t | X_t] = 0$. These assumptions are standard for the use of IV in treatment effect estimation.[4] The IV-Q-Learning algorithm can be described as follows.

Algorithm 2: IV-Q-Learning with Linear Function Approximation

Input : Behavior policy π, learning rates (α_t, β_t), features ϕ, and time horizon T
Output : θ_T and Γ_T

Initialize θ_0 and Γ_0
for *all* $t = 0, 1, \ldots, T-1$ **do**
  Observe the state S_t and take action $A_t \sim \pi(\cdot | S_t)$
  Observe IV Z_t, reward R_t, and the new state S_{t+1}
  Update the estimates of θ^* and Γ^* via

$$\theta_{t+1} = \theta_t + \alpha_t \Big[R_t + \gamma \max_{a' \in \mathcal{A}} \phi^\intercal(S_{t+1}, a')\theta_t - \phi^\intercal(S_t, A_t)\theta_t \Big] \Gamma_t Z_t, \tag{14}$$

$$\Gamma_{t+1} = \Gamma_t + \beta_t [\phi(S_t, A_t) - \Gamma_t Z_t] Z_t^\intercal. \tag{15}$$

end

Like IV-SGD, the IV-Q-Learning algorithm is a two-timescale system, and θ_t is updated in the direction of $\Gamma_t Z_t$, which is a proxy of the projection of $\phi(S_t, A_t)$ on Z_t. This is in contrast to Q-Learning, in which θ_t is updated in the direction of $\phi(S_t, A_t)$. If Z_t is a vector of valid IVs, iterations along these directions resolve the endogeneity issue and result in convergence to the true parameter value.

[4] A more general set of assumptions will be adopted later in Section 5.

Comment 5. The use of IV in our algorithms appears to be similar to "exploration" in RL algorithms. But they are fundamentally different. The notation of exploration in the conventional sense means that the agent ought to be allowed to take seemingly suboptimal actions to better learn about the rewards. When the observed rewards are biased, however, knowledge from these explorations is also biased and cannot correct for R-bias. The use of IV, in contrast, can be viewed as a form of "causal exploration". It perturbs the data-generating process, permitting the agent to de-bias the signals to identify the true rewards and learn optimal policies. □

There are a great of variety of RL algorithms, including temporal difference (TD) learning and actor-critic (AC) algorithms, which may suffer from similar endogeneity issues in the presence of biased reward observations. We may correct for the biases in these algorithms with similar IV-based approaches. Due to space limits, however, we focus on the analysis of IV-Q-Learning in the sequel and discuss IV-TD and IV-AC in Appendix F. Theoretical properties of these IV-RL algorithms can be analyzed under a general framework based on nonlinear two-timescale SA, which we discuss in the next section.

5. Theoretical Properties of Nonlinear Two-Timescale SA

To provide a unified analysis, we express IV-SGD and IV-Q-Learning in the general SA framework. Let $\{W_t : t \geq 0\}$ denote a sequence of random variables with a common state space $\mathcal{W}$. The updating schemes for IV-SGD and IV-Q-Learning can be expressed in the following form:

$$\begin{aligned} \lambda_{t+1} &= \lambda_t + \alpha_t G(W_t, \lambda_t, \xi_t), \\ \xi_{t+1} &= \xi_t + \beta_t H(W_t, \xi_t), \end{aligned} \tag{16}$$

where $\lambda_t \in \mathbb{R}^{\mathsf{d}_\lambda}$ and $\xi_t \in \mathbb{R}^{\mathsf{d}_\xi}$ for all $t \geq 0$ and some positive integers d_λ and d_ξ; moreover, $G : \mathcal{W} \times \mathbb{R}^{\mathsf{d}_\lambda} \times \mathbb{R}^{\mathsf{d}_\xi} \mapsto \mathbb{R}^{\mathsf{d}_\lambda}$ and $H : \mathcal{W} \times \mathbb{R}^{\mathsf{d}_\xi} \mapsto \mathbb{R}^{\mathsf{d}_\xi}$ are two measurable mappings.

In Algorithm 1, for example, by setting $\lambda_t = \theta_t$, $\xi_t = \text{vec}(\Gamma_t)$, $W_t = (X_t, Y_t, Z_t)$, $w = (x, y, z)$, and

$$\begin{aligned} G(w, \lambda, \xi) &= (y - x^\intercal \lambda)\Gamma z, \\ H(w, \xi) &= (x - \Gamma z) z^\intercal, \end{aligned}$$

we may convert IV-SGD to the form (16). Here, $\text{vec}(\cdot)$ denotes the vectorization of a matrix.

Likewise, in Algorithm 2 we may set $W_t = (S_t, A_t, R_t, Z_t, S_{t+1})$, $w = (s, a, r, z, s')$, and

$$\begin{aligned} G(w, \lambda, \xi) &= (r + \gamma \max_{a' \in \mathcal{A}} \phi^\intercal(s', a')\lambda - \phi^\intercal(s, a)\lambda)\Gamma z, \\ H(w, \xi) &= (\phi(s, a) - \Gamma z) z^\intercal, \end{aligned} \tag{17}$$

to convert IV-Q-Learning to the form (16) as well.

Comment 6. The system (16) is general. Other IV-RL algorithms such as IV-TD and IV-AC in Appendix F can be fit into it and systems alike. □

A major challenge for theoretical analysis of the system (16) stems from the nonlinear dependence of $(\lambda_{t+1},\xi_{t+1})$ on (λ_t,ξ_t). To address the issue, we introduce an additional projection step between iterations, a standard treatment in SA literature (Kushner and Yin 2003), which projects $(\lambda_{t+1},\xi_{t+1})$ back to a bounded region if it falls outside after the updating. Specifically, we revise (16) to

$$\begin{aligned}\lambda_{t+1} &= \Pi_B\big(\lambda_t+\alpha_t G(W_t,\lambda_t,\xi_t)\big),\\ \xi_{t+1} &= \Pi_B\big(\xi_t+\beta_t H(W_t,\xi_t)\big),\end{aligned} \tag{18}$$

for some constant $B>0$, where Π_B is the operator that projects a point in $\mathbb{R}^{\mathsf{d}}$ to the d-dimensional ball $\mathsf{Ball}(B) := \{\widetilde{X}\in\mathbb{R}^{\mathsf{d}} : \|\widetilde{X}\|\leq B\}$.

5.1. Key Assumptions

To establish convergence properties of the two-timescale SA (18), we impose Assumption 1 on the random variables $\{W_t : t\geq 0\}$ and Assumption 2 on non-linear functions G and H.

Assumption 1. *The sequence $\{W_t : t\geq 0\}$ forms an irreducible and aperiodic Markov chain that admits a unique stationary distribution ν. It is geometrically ergodic with respect to $U(\cdot) := 1+\|\cdot\|^{d_U}$ for some $d_U>0$ in the following sense.*

(i) There exists a positive integer N, a measurable set $\mathcal{W}_0\subseteq\mathcal{W}$, a constant $b_1>0$, and a probability distribution ψ on $\mathcal{W}$, such that

$$\mathbb{P}(W_N\in\mathcal{B}|W_0=w)\geq b_1\psi(\mathcal{B}),$$

for all $w\in\mathcal{W}_0$ and all measurable set $\mathcal{B}\subseteq\mathcal{W}$.

(ii) There exist constants $\rho\in[0,1)$ and $b_2>0$ such that

$$\mathbb{E}[U(W_t)|W_0=w]\leq\rho U(w)+b_2\mathbb{I}_{\mathcal{W}_0}(w), \tag{19}$$

for all $w\in\mathcal{W}$, where $\mathbb{I}_{\mathcal{W}_0}(\cdot)$ is the indicator function of the set $\mathcal{W}_0$.

Comment 7. In Assumption 1, irreducibility, aperiodicity, and geometric ergodicity are all standard conditions for analyzing general state space Markov chains; see Meyn and Tweedie (2009) for a thorough exposition on the subject. Under this assumption, the d_U-th moment of the transition distribution converges to that of the stationary distribution ν at an exponential rate. This assumption holds for a variety of common models, including linear autoregressive processes whose eigenvalues of the characteristic function are upper bounded in absolute value away from one. Essentially the same assumption is imposed by Konda and Tsitsiklis (2003) to establish the convergence of AC algorithms which are formulated as two-timescale SA. □

Assumption 2. *Let $\bar{G}(\lambda,\xi) := \mathbb{E}_\nu[G(W,\lambda,\xi)]$ and $\bar{H}(\xi) := \mathbb{E}_\nu[H(W,\xi)]$, where ν is the distribution defined in Assumption 1.*

(i) The equation $\bar{H}(\xi) = 0$ has a unique solution ξ^*. The equation $\bar{G}(\lambda, \xi^*) = 0$ has a unique solution λ^*.

(ii) Let $B > \max(\|\lambda^*\|, \|\xi^*\|)$. There exists a positive function $h : \mathcal{W} \mapsto \mathbb{R}_+$ with $h(w) \leq L(1 + \|w\|)$ for some constant $L > 0$, such that for all $w \in \mathcal{W}$, $\lambda, \tilde{\lambda} \in \mathbb{R}^{\mathsf{d}_\lambda}$, and $\xi, \tilde{\xi} \in \mathbb{R}^{\mathsf{d}_\xi}$ with $\max(\|\lambda\|, \|\tilde{\lambda}\|, \|\xi\|, \|\tilde{\xi}\|) \leq B$,

$$\|G(w, \lambda, \xi) - G(w, \tilde{\lambda}, \widetilde{\xi})\| \leq h(w)(\|\lambda - \tilde{\lambda}\| + \|\xi - \tilde{\xi}\|), \tag{20}$$

$$\|H(w, \lambda) - H(w, \tilde{\lambda})\| \leq h(w)\|\lambda - \tilde{\lambda}\|. \tag{21}$$

Moreover, for all $w \in \mathcal{W}$,

$$\sup_{\|\lambda\| \leq B, \|\xi\| \leq B} \|G(w, \lambda, \xi)\| \leq L(1 + \|w\|), \tag{22}$$

$$\sup_{\|\xi\| \leq B} \|H(w, \xi)\| \leq L(1 + \|w\|). \tag{23}$$

(iii) There exist constants $\zeta > 0$ and $\psi > 0$ such that for all $\lambda \in \mathbb{R}^{\mathsf{d}_\lambda}$ and $\xi \in \mathbb{R}^{\mathsf{d}_\xi}$,

$$(\lambda - \lambda^*)^\intercal(\bar{G}(\lambda, \xi^*) - \bar{G}(\lambda^*, \xi^*)) \leq -\zeta\|\lambda - \lambda^*\|^2, \tag{24}$$

$$(\xi - \xi^*)^\intercal(\bar{H}(\xi) - \bar{H}(\xi^*)) \leq -\psi\|\xi - \xi^*\|^2. \tag{25}$$

Comment 8. Assumption 2(i) defines contractors λ^*, ξ^* of the two-timescale system (18). It is a common assumption in SA literature and is satisfied for IV-SGD, IV-Q-Learning, and other IV-RL algorithms under mild conditions; see Appendices E and F. Assumption 2(ii) is more general than the standard Lipschitz condition employed in the literature; see, e.g., Melo, Meyn, and Ribeiro (2008). We allow the Lipschitz coefficient h to grow with $\|w\|$, whereas it is a constant in the standard Lipschitz condition. Assumption 2(iii) is standard in RL literature; see, e.g., Konda and Tsitsiklis (2003). It guarantees the nonlinear mapping (G, H) is a contraction, which is necessary for the convergence of (λ_t, ξ_t). □

Previous work on two-timescale SA usually requires stronger assumptions in one or more of the following aspects—updating rule, state space, and noise structure. Although extensive theoretical analysis has done for linear SA (Konda and Tsitsiklis 2004, Bhandari, Russo, and Singal 2018, Kaledin et al. 2020), there are fewer results for nonlinear SA (Mokkadem and Pelletier 2006, Zou, Xu, and Liang 2019). In the recent RL and SA literature including these papers, the state space is mostly assumed to be finite or bounded (Dalal et al. 2018, Gupta, Srikant, and Ying 2019) and the noise is typically assumed to be i.i.d. or form martingale-differences (Tsitsiklis and Van Roy 1997, Konda and Tsitsiklis 2004). The present paper relaxes these assumptions; see Table 1 for a summary. Most importantly, we relax the typical assumptions on noise structure—i.i.d. or martingale-differences—to general Markovian noise. It substantially expands the scope of application of RL algorithms, because noise in MDP problems is often serially correlated due to the inter-temporal effects of the actions taken. The analysis of Markovian noise, however, requires a new set of techniques, which we discuss in detail in Section 5.3.

Table 1 **Summary of Assumptions in SA Literature.**

	Timescale	Updating Rule	State Space	Noise
	single-timescale	linear	finite bounded	i.i.d. martingale-difference
Our setting	*two-timescale*	*nonlinear*	*unbounded*	*Markovian*

Our assumptions are general compared to the literature. When applying to specific IV-RL algorithms such as IV-Q-Learning, all of these assumptions can be verified with more primitive conditions discussed in Section 4.

5.2. Finite-Time Risk Bounds

We now present finite-time risk bounds for mean-squared errors. The proof is deferred to Appendix A.

Proposition 1. *Let* $\{(\lambda_t, \xi_t): t \geq 0\}$ *be the SA iterates following* (18)*. For all* $t \geq 1$*, let* $\alpha_t = \alpha_0 t^{-\kappa}$ *and* $\beta_t = \beta_0 t^{-\delta}$ *for some* $\alpha_0 > 0$*,* $\beta_0 > 0$*, and* $0 < \kappa \leq \delta \leq 1$*. Assume* $\zeta\alpha_0 > 1$ *in the case of* $\kappa = 1$*, and* $\psi\beta_0 > 1$ *in the case of* $\delta = 1$*, where* ζ *and* ψ *are the constants defined in* (24) *and* (25)*, respectively. Suppose Assumption 1 holds with* $d_U = 2$ *and Assumption 2 holds. Then,*

(i) $\mathbb{E}\left[\|\lambda_t - \lambda^*\|^2 | W_0 = w\right] \leq t^{-\kappa}(k_1 \ln t + k_2)$*, and*

(ii) $\mathbb{E}\left[\|\xi_t - \xi^*\|^2 | W_0 = w\right] \leq t^{-\delta}(k_1 \ln t + k_2)$*,*

for all t *large enough, where* $k_1 \geq 0$ *and* $k_2 > 0$ *are some constants that depend on* w*. Moreover,* $k_1 = 0$ *if* $\{W_t : t \geq 0\}$ *are i.i.d.*

Similar results have been obtained in recent RL literature, in more restrictive settings such as finite state space (Chen et al. 2019) or linear SA (Kaledin et al. 2020). Proposition 1 is unique in that it allows for a Markovian noise structure, unbounded state space, as well as linear growth in the Lipschitz coefficient of the nonlinear mapping (G, H). The setting of Konda and Tsitsiklis (2003) is the most similar to ours in terms of generality, but the iterates in the two-timescale SA corresponding to AC algorithms are only shown to be convergent, and no finite-time bounds are obtained there. Proposition 1 serves as an important intermediate step for the main results presented in the next two subsections.

5.3. Finite-Time Trajectory Stability Bounds

Following Proposition 1, and under the same set of assumptions, we establish the finite-time high-probability bounds regarding the behavior of the trajectories $\{\lambda_t : t \geq T\}$ and $\{\xi_t : t \geq T\}$ for all T large enough. We outline its proof below and defer the detailed proof to Appendix B.

THEOREM 1. *Let* $\{(\lambda_t, \xi_t) : t \geq 0\}$ *be the SA iterates following* (18). *Suppose the assumptions in Proposition 1 hold, and in addition,* $1/2 < \kappa \leq \delta < 1$. *Fix arbitrary* $C > 0$, $\kappa_c \in [0, \kappa - 1/2)$, *and* $\delta_c \in [0, \delta - 1/2)$. *Then, for all* T *large enough,*

(i) $\mathbb{P}\big(\|\lambda_t - \lambda^*\| \leq C t^{-\kappa_c},\, \forall t \geq T | W_0 = w\big) \geq 1 - C_\lambda T^{-b_\kappa} \ln T$, *and*

(ii) $\mathbb{P}\big(\|\xi_t - \xi^*\| \leq C t^{-\delta_c},\, \forall t \geq T | W_0 = w\big) \geq 1 - C_\xi T^{-b_\delta} \ln T$,

where $b_\kappa := 2\kappa - 1 - 2\kappa_c$, $b_\delta := 2\delta - 1 - 2\delta_c$, C_λ *and* C_ξ *are some positive constants that depends on* $(\delta, \delta_c, \kappa, \kappa_c, C, w)$ *and* (δ, δ_c, C, w), *respectively. Moreover, if* $\{W_t : t \geq 0\}$ *are i.i.d., then the* $\ln T$ *term can be removed from the bounds.*

Theorem 1 is substantially stronger in several aspects than existing high-probability uniform bounds for SA iterates in the literature. Most notably, previous results, which are often referred to as concentration bounds, focus on the event that the trajectory $\{\lambda_t : t \geq T\}$ stays in a *fixed* ball $\{\mathsf{Ball}(C) : t \geq T\}$; see, e.g., Dalal et al. (2018) and Thoppe and Borkar (2019). In contrast, we characterize the probability of the event that the trajectory stays in a *shrinking* ball of $\{\mathsf{Ball}(Ct^{-\kappa_c}) : t \geq T\}$. Moreover, the clearly sharper result is established in the presence of Markovian noise, a more general setting than the common assumption of martingale-difference noise. Typical techniques for establishing concentration bounds rely on showing that $\|\lambda_t - \lambda\|^2$ is a sub-martingale. This approach, however, is applicable only if the noise forms a martingale-difference sequence. It breaks down in the presence of Markovian noise, because $\|\lambda_t - \lambda\|^2$ is no longer necessarily a sub-martingale.

To tackle the challenge stemming from the Markovian noise, we develop a set of new techniques. To illustrate our main idea, let us consider a simplified case and focus on the bound on the probability of $\{\xi_t : t \geq T\}$ leaving the shrinking ball. A critical technique of ours is "look-back"—approximating ξ_t, the SA iterate at time t, with an earlier iterate in a judiciously chosen time period. A second critical technique that accompanies the look-back is "polynomial partitioning"—producing a sequence of reference points in such a way that the length of each interval grows at a polynomial rate. Specifically, we select a strictly increasing sequence of positive integers $I_j \in \mathbb{N}$ to partition $[1, \infty)$ to a sequence of finite intervals $\{[I_j, I_{j+1}) : j = 0, 1, ...\}$. We choose $I_0 = 1$ and $I_j = \lfloor C_q j^q \rfloor$, where C_q and q are positive constants to be determined later. Notice that for a fixed C_q, the larger q is, the more dispersed the points $\{I_j : j = 0, 1, \ldots\}$ are in $\mathbb{N}$. The polynomial partitioning allows us to carefully control the errors from two types of approximations—(i) approximating ξ^* with ξ_{I_j}, and (ii) approximating ξ_t with ξ_{I_j} for all $t \in [I_j, I_{j+1})$. A trade-off needs to be made between the two types of errors, as they vary in opposite directions as one alters the level of dispersion of the reference points.

By the triangular inequality

$$\|\xi_t - \xi^*\| \le \|\xi_{I_j} - \xi^*\| + \|\xi_t - \xi_{I_j}\|,$$

to bound $\mathbb{P}\Big(\|\xi_t - \xi^*\| \ge Ct^{-\delta_c},\ \exists t \ge T \,\Big|\, W_0 = w\Big)$, it suffices to separately bound

$$\mathbb{P}\left(\|\xi_{I_j} - \xi^*\| \ge \frac{1}{3} C I_{j+1}^{-\delta_c},\ \exists I_j \ge T \,\middle|\, W_0 = w\right), \tag{26}$$

and

$$\mathbb{P}\left(\|\xi_t - \xi_{I_j}\| \ge \frac{2}{3} C I_{j+1}^{-\delta_c},\ \exists t \in [I_j, I_{j+1}), \exists I_j \ge T \,\middle|\, W_0 = w\right). \tag{27}$$

First, we bound the probability (26). Given a value of C, it follows immediately from Markov's inequality and Proposition 1 that there exists some constant $K_1 > 0$ such that for all t large enough,

$$\mathbb{P}\left(\|\xi_{I_j} - \xi^*\| \ge \frac{1}{3} C I_{j+1}^{-\delta_c},\ \exists I_j \ge T \,\middle|\, W_0 = w\right) \le \frac{K_1}{C^2} T^{-[(\delta - 2\delta_c) - 1/q]} \ln T. \tag{28}$$

Next, we bound the probability (27). The SA updating scheme (18) implies that if $\|\xi_t\| \le B$ for all $t \in [I_j, I_{j+1})$, then $\xi_t - \xi_{I_j} = \sum_{l=I_j}^{t-1} \beta_l H(W_l, \xi_l)$. However, this expression involves ξ_l's, each of which depends on the past iterates, thus difficult to analyze. Thus, we approximate $\|\xi_t - \xi_{I_j}\|$ by a partial sum process $S^*_{t,j} := \sum_{l=I_j}^{t-1} \beta_l H(W_l, \xi^*)$. By the triangular inequality,

$$\|\xi_t - \xi_{I_j}\| \le \|S^*_{t,j}\| + \|(\xi_t - \xi_{I_j}) - S^*_{t,j}\|.$$

For the approximation error $\|(\xi_t - \xi_{I_j}) - S^*_{t,j}\|$, we are able to show that there exists some constant $K_2 > 0$ such that for all t large enough,

$$\mathbb{P}\left(\max_{t \in [I_j, I_{j+1})} \|(\xi_t - \xi_{I_j}) - S^*_{t,j}\| \ge \frac{1}{3} C I_{j+1}^{-\delta_c},\ \exists I_j \ge T \,\middle|\, W_0 = w\right) \le \frac{K_2}{C^2} T^{-[(3\delta - 2\delta_c - 2) + 1/q]} \ln T. \tag{29}$$

Finally, for the partial sum process $\|S^*_{t,j}\|$, by virtue of the extended Lévy inequality for dependent data (Loève 1978, Page 51), we are able to prove that there exists some constant $K_3 > 0$ such that for all t large enough,

$$\mathbb{P}\left(\max_{t \in [I_j, I_{j+1})} \|S^*_{t,j}\| \ge \frac{1}{3} C I_{j+1}^{-\delta_c},\ \exists I_j \ge T \,\middle|\, W_0 = w\right) \le \frac{K_3}{C^2} T^{-(2\delta - 2\delta_c - 1)}. \tag{30}$$

As stated, there is a tension between the bounds in (28) and (29), which demands a careful choice of q. Specifically, as q decreases, the exponent of T on the RHS of (28) (i.e., $-[(\delta - 2\delta_c - 1/q])$ increases, yielding a looser bound. Intuitively, this is because for small q, the points $I_j, I_{j+1}, \ldots$ grows to infinity relatively slowly, so the approximation error $\|\xi_{I_j} - \xi^*\|$ does not decay fast enough.

However, as q increases, the exponent of T on the RHS of (29) (i.e., $-[(3\delta-2\delta_c-2)+1/q]$) increases, which also yields a looser bound. This, however, is because each interval $[I_j, I_{j+1})$ is too long for large q, so there are many values of $t\in[I_j, I_{j+1})$ for which the approximation error $\|(\xi_t-\xi_{I_j})-S^*_{t,j}\|$ is large. It turns out that there exists a unique value of $q=1/(1-\delta)$ such that the exponent of T on the RHS of (28), (29), and (30) are identical, which represents a delicate balance among the three bounds and produces the result in Theorem 1.

Comment 9. Theorem 1 assumes $\delta<1$. The polynomial partitioning technique does not apply to the case that $\delta=1$, because the optimal value of q would be $1/(1-\delta)=\infty$. Nevertheless, we may use an "exponential partitioning" technique and follow a similar approach to derive similar finite-time trajectory stability bounds for this case.

5.4. Asymptotic Normality

In this subsection, we show that the SA iterates λ_t and ξ_t are asymptotically normal. For ease of presentation, we introduce the following notations:

$$A_{11}:=\frac{\partial\bar{G}(\lambda^*,\xi^*)}{\partial\lambda},\quad A_{12}:=\frac{\partial\bar{G}(\lambda^*,\xi^*)}{\partial\xi},\quad \text{and}\quad A_{22}:=\frac{\partial\bar{H}(\xi^*)}{\partial\xi},$$

where $\bar{G}(\lambda,\xi)=\mathbb{E}_\nu[G(W,\lambda,\xi)]$ and $\bar{H}(\xi)=\mathbb{E}_\nu[H(W,\xi)]$. We can then expand $\bar{G}$ and $\bar{H}$ in some open neighborhoods of (λ^*,ξ^*) and ξ^*, respectively, as

$$\begin{aligned}\bar{G}(\lambda,\xi)&=A_{11}(\lambda-\lambda^*)+A_{12}(\xi-\xi^*)+r_G(\lambda,\xi),\\ \bar{H}(\xi)&=A_{22}(\xi-\xi^*)+r_H(\xi),\end{aligned}$$

where $r_G(\lambda,\xi)$ and $r_H(\xi)$ are high-order residual terms. Note that by Assumption 2(iii), A_{11} (resp., A_{22}) is a negative definite matrix with the largest eigenvalue no greater than $-\zeta$ (resp., $-\psi$). We make the following assumption on r_G and r_H.

Assumption 3. *There exist positive constants C_G and C_H such that*

$$\|r_G(\lambda,\xi)\|\le C_G(\|\lambda-\lambda^*\|^2+\|\xi-\xi^*\|^2), \tag{31}$$

$$\|r_H(\xi)\|\le C_H\|\xi-\xi^*\|^2, \tag{32}$$

for all $\lambda\in\mathbb{R}^{\mathsf{d}_\lambda}$ and $\xi\in\mathbb{R}^{\mathsf{d}_\xi}$ with $\|\lambda\|\le B$ and $\|\xi\|\le B$.

Comment 10. Assumption 3 allows us to control the nonlinear error term in the SA updating scheme (18). For many applications that include Q-Learning, the nonlinear mapping $G(W,\lambda,\xi)$ may be non-differentiable with respect to (λ,ξ). The theory developed in the present paper has the advantage of making assumptions on the smoother objects $\bar{G}(W,\cdot,\cdot)$ and $\bar{H}(W,\cdot)$. Provided that they are twice differentiable with bounded derivatives, conditions (31) and (32) will hold. □

To express the asymptotic covariance matrices of λ_t and ξ_t, we use

$$\widetilde{G}(W,\lambda,\xi) := G(W,\lambda,\xi) - \bar{G}(W,\lambda,\xi),$$
$$\widetilde{H}(W,\xi) := H(W,\xi) - \bar{H}(W,\xi),$$

to denote the *residuals* of $G(W,\lambda,\xi)$ and $H(W,\xi)$, respectively. Clearly, $\widetilde{G}(W,\lambda^*,\xi^*) = G(W,\lambda^*,\xi^*)$ and $\widetilde{H}(W,\xi^*) = H(W,\xi^*)$, by the definition of (λ^*,ξ^*).

For each $l = 0,1,\ldots$, we denote the steady-state lag-l covariance of $\{G(W_t,\lambda^*,\xi^*) : t \geq 0\}$ by

$$L_G(l) := \mathrm{Cov}_\nu[G(W_0,\lambda^*,\xi^*), G(W_l,\lambda^*,\xi^*)] = \mathbb{E}_\nu[\widetilde{G}(W_0,\lambda^*,\xi^*)\widetilde{G}(W_l,\lambda^*,\xi^*)].$$

Likewise, we denote the steady-state lag-l covariance of $\{H(W_t,\xi^*) : t \geq 0\}$ by

$$L_H(l) := \mathrm{Cov}_\nu[H(W_0,\lambda^*), H(W_l,\lambda^*)] = \mathbb{E}_\nu[\widetilde{H}(W_0,\xi^*)\widetilde{H}(W_l,\xi^*)].$$

Note that under the stationary distribution ν,

$$L_G(l) = \mathrm{Cov}_\nu[G(W_t,\lambda^*,\xi^*), G(W_{t+l},\lambda^*,\xi^*)],$$
$$L_H(l) = \mathrm{Cov}_\nu[H(W_t,\xi^*), H(W_{t+l},\xi^*)],$$

for all $t \geq 0$. In addition, if W_t's are i.i.d., then $L_G(l) = L_H(l) = 0$ for all $l \geq 1$.

Further, we denote the *time average variance constant* (TAVC) of $\{G(W_t,\lambda^*,\xi^*) : t \geq 0\}$ by

$$\bar{L}_G := L_G(0) + \sum_{l=1}^{\infty}(L_G(l) + L_G^{\intercal}(l)).$$

It can be shown that $T^{-1}\,\mathrm{Var}[\sum_{t=1}^T G(W_t,\lambda^*,\xi^*)] \to \bar{L}_G$ as $T \to \infty$ if the limit exists. Likewise, we denote the TAVC of $\{H(W_t,\xi^*) : t \geq 0\}$ by

$$\bar{L}_H := L_H(0) + \sum_{l=1}^{\infty}(L_H(l) + L_H^{\intercal}(l)).$$

We are now ready to state the asymptotic distribution of ξ_t and λ_t. The proof is provided in Appendix C.

THEOREM 2. *Let* $\{(\lambda_t,\xi_t) : t \geq 0\}$ *be the SA iterates following* (18). *Suppose the assumptions in Theorem 1 hold, and in addition,* $1/2 < \kappa < \delta < 1$, *Assumption 1 holds with* $d_U = 4$, *and Assumption 3 holds. Then, as* $t \to \infty$,

(i) $\alpha_t^{-1/2}(\lambda_t - \lambda^*) \rightsquigarrow \mathcal{N}(0,\Sigma_G)$, *where* $\Sigma_G := \int_0^\infty e^{mA_{11}}\bar{L}_G e^{mA_{11}^{\intercal}}\,\mathrm{d}m$, *and*

(ii) $\beta_t^{-1/2}(\xi_t - \xi^*) \rightsquigarrow \mathcal{N}(0,\Sigma_H)$, *where* $\Sigma_H := \int_0^\infty e^{mA_{22}}\bar{L}_H e^{mA_{22}^{\intercal}}\,\mathrm{d}m$.

Here, $\rightsquigarrow$ *means convergence in distribution, and* $\mathcal{N}(0,\Sigma)$ *denotes the multivariate normal distribution with mean 0 and covariance matrix* Σ.

Existing CLTs for SA are mostly established under the assumption that the residuals—e.g., $\{\widetilde{G}(W_t,\lambda,\xi): t\geq 0\}$—are i.i.d. or form a martingale-difference sequence (Kushner and Yin 2003, Konda and Tsitsiklis 2004, Mokkadem and Pelletier 2006). Theorem 2 generalizes these results in that the residuals are assumed to be Markovian, a much more realistic setting which, meanwhile, presents a substantial technical challenge. Such generalization allows us to extend our SA results to dynamic problem such as MDP.

A common approach to proving the asymptotic normality of SA iterates is to reformulate the standardized sequence as the sum of a martingale and some extra terms that are asymptotically negligible, and then invoke a martingale CLT. We, too, follow this approach. However, the Markovian noise structure dramatically complicates the task of verifying the conditions of the martingale CLT and the task of showing the extra terms are negligible. To address the challenge, we adopt a strategy similar to that used for proving Theorem 1. Specifically, we partition the interval $[1,t]$ on the time line into q sub-intervals with equal length for some positive integer q. A martingale will be constructed for the SA iterates at the time epochs corresponding to the end points of these sub-intervals. In contrast, had the residuals formed a martingale-difference sequence, such a martingale would have been constructed for the iterates at all time epochs.

With a small q, each sub-interval is long, so the SA iterates at the end points of the sub-intervals are merely weakly dependent. Then, the differences between them can be well approximated with a martingale-difference sequence upon proper centering, thus suitable for invoking a martingale CLT. We approximate each SA iterate by the one corresponding to a nearby endpoint of a sub-interval. Approximation errors of this kind grow in the length of each sub-interval, so q cannot be too small. Again, we need to strike a balance between the two competing forces when choosing the value of q, a situation similar to the proof of Theorem 1. In particular, we are able to identify q as an increasing function of t in such way that the dependence across the resulting sub-intervals is sufficiently weak, vanishing to 0 as t increases, while at the same time the approximation errors within each sub-interval retains negligible.

The asymptotic covariance matrices in Theorem 2 highlight the impact of the Markovian noise structure. For example, Σ_G will be large, if $L_G(l)$'s are large; that is, the time dependence of the Markov chain $\{G(W_t,\lambda^*,\xi^*): t\geq 0\}$ is strong. Such strong serial dependence is particularly salient in modern large-scale applications. For instance, if each action taken has a relatively small impact on the evolution of the states, which often occurs when the state space is large, then the states tend to be "persistent", that is, the auto-correlation decays slowly as the lag increases, yielding a strong serial dependence structure. Theorem 2 sheds light on why "long-term stategies" are particularly challenging to learn (Mnih et al. 2015). The theorem also motivates the need and benefit of more exploration that would reduce the serial dependence of the noise. In addition, the matrices A_{11}

and A_{22} will partially depend on the choice of IV when the SA theory is applied to the IV-RL algorithms, which can potentially affect their asymptotic property.

COMMENT 11. The asymptotic covariance matrices Σ_G and Σ_H can be estimated by plugging in consistent estimators of A_{11}, A_{22}, $\bar{L}_G$, and $\bar{L}_H$. For example, A_{11} can be consistently estimated via local-linear regression with the SA iterates $\{\lambda_l : 1 \le l \le t\}$, because λ_t is an consistent estimator of λ^* as $t \to \infty$. Moreover, $\bar{L}_G$ can be consistently estimated by

$$\hat{L}_G \coloneqq \hat{L}_G(0) + \sum_{l=1}^{\infty} \omega_l (\hat{L}_G(l) + \hat{L}_G^{\intercal}(l)), \tag{33}$$

where

$$\hat{L}_G(l) \coloneqq \sum_{k=1}^{t-l} \hat{G}(W_k, \lambda_t, \xi_t) \hat{G}(W_{k+l}, \lambda_t, \xi_t)^{\intercal},$$
$$\hat{G}(W_k, \lambda_t, \xi_t) \coloneqq G(W_k, \lambda_t, \xi_t) - \frac{1}{t} \sum_{l=1}^{t} G(W_l, \lambda_t, \xi_t),$$

and $\{\omega_l : l \ge 1\}$ is a scheme of weights indexed by l, e.g., $\omega_l = 1$ for all $l \ge 1$, or as proposed in Newey and West (1987), $\omega_l = (1 - \frac{l}{h+1}) \mathbb{I}(l \le h)$, where h is chosen as $cT^{\frac{1}{3}}$ for some constant $c > 0$. □

COMMENT 12. We have assumed $\kappa < \delta$ in Theorem 2. If $\kappa = \delta$, then the SA updating scheme (18) is reduced to a a single-timescale system. We can apply part (ii) of Theorem 2 to the joint parameter (λ_t, ξ_t) to obtain its asymptotic normality. In addition, for the case of $\delta = 1$, the same results can be obtained by using slightly modified arguments in our proof. □

6. Applying SA Theory to IV-Q-Learning

To apply the general results for nonlinear two-timescale SA, we consider the projected version of IV-Q-Learning in Algorithm 2. That is,

$$\begin{aligned} \theta_{t+1} &= \Pi_B\big(\theta_t + \alpha_t G(W_t, \theta_t, \Gamma_t)\big), \\ \Gamma_{t+1} &= \Pi_B\big(\Gamma_t + \beta_t H(W_t, \Gamma_t)\big), \end{aligned} \tag{34}$$

for some constant $B > 0$, where

$$G(w, \theta, \Gamma) = (r + \gamma \max_{a' \in \mathcal{A}} \phi^{\intercal}(s', a')\theta - \phi^{\intercal}(s, a)\theta)\Gamma z,$$
$$H(w, \Gamma) = (\phi(s, a) - \Gamma z) z^{\intercal},$$

and W_t and w are properly defined in Assumption 4 below. Assuming $\{W_t : t \ge 0\}$ admits a unique stationary distribution ν, we denote the steady-state lag-l covariance of $\{G(W_t, \theta^*, \Gamma^*) : t \ge 0\}$ and its TAVC, respectively, by

$$L_{G,\mathrm{Q}}(l) \coloneqq \mathrm{Cov}_\nu[G(W_0, \theta^*, \Gamma^*), G(W_l, \theta^*, \Gamma^*)], \quad l \ge 0,$$

$$\bar{L}_{G,\mathrm{Q}} := L_{G,\mathrm{Q}}(0) + \sum_{l=1}^{\infty}(L_{G,\mathrm{Q}}(l) + L_{G,\mathrm{Q}}^{\intercal}(l)).$$

We discuss the theoretical properties of IV-Q-Learning in Section 6.1 as a corollary of the SA theory developed, and discuss the related inference issues in Section 6.2. The proofs are provided in Appendix D.

6.1. Theoretical Properties of IV-Q-Learning

Assumption 4. *There exists a vector of IVs $Z_t \in \mathbb{R}^q$ with $q \geq p$ at time t that can be observed prior to period $t+1$. Let $W_t = (S_t, A_t, R_t, Z_t, S_{t+1}, R_t Z_t, \|S_t\|\|Z_t\|, \|S_{t+1}\|\|Z_t\|, Z_t Z_t^{\intercal})$. Then, $\{W_t : t \geq 0\}$ forms a Markov chain that is geometrically ergodic with respect to $U(\cdot) = 1 + \|\cdot\|^{d_U}$ for some $d_U > 0$, having a unique stationary distribution ν.*

Comment 13. In IV-Q-Learning, there are product terms of the form $\phi(S_t, A_t)Z_t$, $\phi(S_{t+1}, A_t)Z_t$, etc. To apply the SA results, we essentially need to assume the U-ergodicity on all the product terms. Therefore, in Assumption 4, the definition of W_t includes auxiliary terms $R_t Z_t$, $\|S_t\|\|Z_t\|$, $\|S_{t+1}\|\|Z_t\|$, and $Z_t Z_t^{\intercal}$. □

Assumption 5. *(i) $\phi(S_t, A_t)$ is a vector of linearly independent random variables under the distribution ν.*

(ii) There exist positive constants B and K such that $B > \max(\|\theta^*\|, \|\Gamma^*\|_{\mathsf{F}})$ and $\max_{a \in \mathcal{A}} \|\phi(s,a)\| \leq K(\|s\| + 1)$ for all $s \in \mathcal{S}$, where $\|\cdot\|_{\mathsf{F}}$ denotes the Frobenius norm.

(iii) There exists $\zeta > 0$ such that for all $\theta \in \mathbb{R}^p$,

$$\gamma^2\, \mathbb{E}_\nu\left[\max_{a' \in \mathcal{A}} (\phi^{\intercal}(S', a')\theta)^2\right] - \mathbb{E}_\nu\left[((\Gamma^* Z)^{\intercal}\theta)^2\right] \leq -2\zeta\|\theta\|^2. \tag{35}$$

(iv) $\phi(S_t, A_t) = \Gamma^* Z_t + \eta_t$, where $\Gamma^* \in \mathbb{R}^{p \times q}$ is a deterministic full-rank matrix, and $\eta_t \in \mathbb{R}^p$ is a random vector; moreover, $\mathbb{E}_\nu[\epsilon_t Z_t] = 0$, $\mathbb{E}_\nu[\eta_t Z_t^{\intercal}] = 0$, and $\mathbb{E}_\nu[Z_t Z_t^{\intercal}]$ has full rank.

Comment 14. Conditions (i) and (ii) in Assumption 5 will be used to verify conditions (i) and (ii) in Assumption 2, respectively.

Condition (iii) in Assumption 5, which will be used to verify the condition (24) in Assumption 2(iii), is weaker than those usually used for the convergence analysis of Q-Learning in Melo, Meyn, and Ribeiro (2008) and Zou, Xu, and Liang (2019). This condition requires that, for IV-Q-Learning to converge well, either (a) discount factor γ is small enough, or (b) the behavior policy π is close enough to π^*, and the instruments Z_t are strong enough. The possible divergence of Q-Learning is well-known in RL literature. Thus, more sophisticated algorithms such as AC are often used instead in practice to avoid the divergence issue; see more discussion on AC in Appendix F.

Condition (iv) in Assumption 5 is used to verify the condition (25) in Assumption 2(iii). Similar full-rank conditions are commonly imposed in IV regression literature. □

At last, we impose a sufficient condition for Assumption 3 to hold.

ASSUMPTION 6. *$\mathbb{E}_\nu[\max_{a'\in\mathcal{A}}(\phi^\intercal(S,a')\theta)]$ is twice differentiable in an open neighborhood of θ^*.*

COROLLARY 1. *Let $\{\theta_t : t\geq 0\}$ be the iterates from the projected IV-Q-Learning (34). For all $t\geq 1$, let $\alpha_t=\alpha_0 t^{-\kappa}$ and $\beta_t=\beta_0 t^{-\delta}$ for some $\alpha_0>0$, $\beta_0>0$, and $0<\kappa\leq\delta\leq 1$. Suppose Assumptions 5 holds. Assume $\zeta\alpha_0>1$ in the case of $\kappa=1$, and $\psi\beta_0>1$ in the case of $\delta=1$, where ζ is the constant defined in (35), and ψ is the smallest eigenvalue of $\mathbb{E}_\nu[Z_tZ_t^\intercal]$.*

(i) If Assumption 4 holds with $d_U=2$, then

$$\mathbb{E}[\|\theta_t-\theta^*\|^2|W_0=w]\leq t^{-\kappa}(k_1\ln t+k_2),\quad \forall t\geq T,$$

for all t large enough, where $k_1\geq 0$ and $k_2>0$ are some constants that depend on w. Moreover, $k_1=0$ if $\{W_t : t\geq 0\}$ are i.i.d.

(ii) Fix arbitrary $C>0$, $\kappa_c\in[0,\kappa-1/2)$, and $\delta_c\in[0,\delta-1/2)$. If Assumption 4 holds with $d_U=2$, and $1/2<\kappa\leq\delta<1$, then

$$\mathbb{P}\big(\|\theta_t-\theta^*\|\leq Ct^{-\kappa_c},\,\forall t\geq T|W_0=w\big)\geq 1-C_\theta T^{-b_\kappa}\ln T,$$

for all t large enough, where $b_\kappa:=2\kappa-1-2\kappa_c$, and C_θ is some positive constant that depends on $(\delta,\delta_c,\kappa,\kappa_c,C,w)$. Moreover, if $\{W_t : t\geq 0\}$ are i.i.d., then the term $\ln T$ can be removed from the bound.

(iii) If Assumption 4 holds with $d_U=4$, Assumption 6 holds, and $1/2<\kappa<\delta<1$, then

$$\alpha_t^{-1/2}(\theta_t-\theta^*)\rightsquigarrow\mathcal{N}(0,\Sigma_{G,\mathrm{Q}}),$$

as $t\to\infty$, where $\Sigma_{G,\mathrm{Q}}:=\int_0^\infty \exp\left(mA_{11,\mathrm{Q}}\right)\bar{L}_{G,\mathrm{Q}}\exp\left(mA_{11,\mathrm{Q}}^\intercal\right)\mathrm{d}t$ and

$$A_{11,\mathrm{Q}}:=\gamma\frac{\partial}{\partial\theta}\mathbb{E}_\nu\Big[\max_{a'\in\mathcal{A}}\big(\phi^\intercal(S_{t+1},a')\theta^*\big)\Gamma^*Z_t\Big]-\mathbb{E}_\nu[\phi(S_t,A_t)\Gamma^*Z_t].$$

COMMENT 15. When the data (S_t,A_t,R_t) are fully exogenous under stationary distribution; i.e., $\mathbb{E}_\nu[(R_t-r_t)\phi(S_t,A_t)]=0$, we can simply allow $Z_t=\phi(S_t,A_t)$. In this case, the IV-Q-Learning reduces to the original Q-Learning in (12). Therefore, all the theoretical results apply to the standard Q-Learning when the reward function is unbiased. □

6.2. Inference for IV-Q-Learning

We now discuss how to carry out inference on the optimal policy function $\pi^*(\cdot)$ by using Corollary 1. Specifically, we can first consider the null hypothesis

$$H_0(s): a_0=\pi^*(s),$$

with the alternative hypothesis that $a_0 \neq \pi^*(s)$.

Define $\hat{a} := \arg\max_{a\in\mathcal{A}} \phi^\intercal(s,a)\theta_t$. Existence of such a maximizer is guaranteed by Assumption 5(ii), and asymptotically $\hat{a}$ is unique if the maximizer of the true $Q^*(s,\cdot)$ function is unique. Denote $\Sigma(s) := \frac{\partial^2 \phi^\intercal(s,a)\theta^*}{\partial a \partial a^\intercal}\big|_{a=\pi^*(s)}$. The following lemma provides inference for $a_0 = \pi^*(s)$ for a fixed state s.

Proposition 2. *Suppose that the assumptions in Corollary 1(iii) hold. Assume that, for a given state $s\in\mathcal{S}$, $\phi(s,a)$ is twice differentiable around $a_0 = \pi^*(s)$ and $\Sigma(s)$ has full rank. Suppose that $a_0 := \pi^*(s)$ is a unique optimal action at s. Suppose $\hat{a} := \arg\max_{a\in\mathcal{A}} \phi^\intercal(s,a)\theta_t$. Then,*

$$\alpha_t^{-\frac{1}{2}}(\hat{a}-a_0) \rightsquigarrow \mathcal{N}(0,\Omega_s),$$

as $t\to\infty$, where

$$\Omega_s := \Sigma(s)^{-1}\frac{\partial \phi^\intercal(s,a_0)}{\partial a}\Sigma_{G,\mathrm{Q}}\frac{\partial \phi(s,a_0)}{\partial a}\Sigma(s)^{-1}.$$

Moreover, the test statistic

$$T(s) := \alpha_t^{-1}(\hat{a}-a_0)^\intercal \Omega_s^{-1}(\hat{a}-a_0) \rightsquigarrow \chi^2(d_A),$$

where d_A represents the dimensionality of the action space.

Proposition 2 allows us to test whether action a_0 is optimal conditional on s. The test statistics $T(s)$ can be approximated with a consistent estimator of Ω_s. For instance, the term $\Sigma(s)^{-1}\frac{\partial \phi^\intercal(s,a)}{\partial a}$ for each (s,a) pair can be consistently estimated by replacing θ^* with $\hat{\theta}_t$; matrix $\bar{L}_{G,\mathrm{Q}}$ can be estimated according to (33).

Below, we present a result for inference on optimal policy $\pi^*(\cdot)$.

Proposition 3. *Suppose that the assumptions in Corollary 1(iii) hold. Suppose that for almost every $s\in\mathcal{S}$, there exists a unique optimal action at s, denoted as $\pi^*(s)$. Suppose that $\phi(s,a)$ is twice differentiable with respect to a with the first- and second-order derivatives being uniformly bounded over $s\in\mathcal{S}$, and for every $s\in\mathcal{S}$, $\Sigma(s)$ has full rank. Assume that for almost every all $s\in\mathcal{S}$, $\hat{\pi}(s) := \arg\max_{a\in\mathcal{A}} \phi^\intercal(s,a)\theta_t$ is uniquely defined. Then,*

$$\alpha_t^{-\frac{1}{2}}(\hat{\pi}(s)-\pi^*(s)) \rightsquigarrow \mathbb{G}(s),$$

as $t\to\infty$, where $\mathbb{G}(s)$ is a tight Gaussian process with covariance function

$$\mathrm{Cov}(\mathbb{G}(s),\mathbb{G}(s')) = \Sigma(s)^{-1}\left(\frac{\partial \phi^\intercal(s,\pi^*(s))}{\partial a}\right)\bar{L}_{G,\mathrm{Q}}\left(\frac{\partial \phi(s',\pi^*(s'))}{\partial a}\right)\Sigma(s')^{-1}.$$

Moreover, as $t\to\infty$,

$$T_2 := \alpha_t^{-1}\int_{\mathcal{S}} \|\hat{\pi}(s)-\pi^*(s)\|^2 \nu(s)\,\mathrm{d}s \rightsquigarrow \int_{\mathcal{S}} \mathbb{G}(s)^2\nu(s)\,\mathrm{d}s\,.$$

Comment 16. We can use $\hat{T}_2 := \alpha_t^{-1}\frac{1}{t}\sum_{l=1}^{t}\|\hat{\pi}(s_l)-\pi^*(s_l)\|^2$ to approximate T_2. To obtain the critical values of $\int_{\mathcal{S}}\mathbb{G}(s)^2\nu(s)\,\mathrm{d}s$, given a consistent estimator $\hat{\Omega}_s$ of Ω_s, we can perform the following:

(i) Draw $\hat{\theta}_t^b \sim \mathcal{N}(\theta_t, \alpha_t\hat{\Omega}_s)$, $b=1,2,\ldots,B$ for some large enough value B.

(ii) Compute $\hat{\pi}^b(s_l) := \arg\max_{a\in\mathcal{A}}\phi^\intercal(s_l,a)\hat{\theta}_t^b$, for $l=1,2,\ldots,t$.

(iii) Compute $G^b := \frac{1}{t}\sum_{l=1}^{t}\|\hat{\pi}^b(s_l)-\hat{\pi}(s_l)\|^2$, $b=1,2,\ldots,B$.

(iv) Compute the $(1-\alpha)$-quantile of $G^b, b=1,2,\ldots,B$ as an estimator of the $(1-\alpha)$-quantile of $\int_{\mathcal{S}}\mathbb{G}(s)^2\nu(s)\,\mathrm{d}s$, for a fixed value $\alpha\in(0,1)$, e.g., $\alpha=0.05$. □

7. Numerical Experiments

In this section, we use simulation experiments to illustrate how SGD and Q-Learning lead to biased estimates in the presence of reward endogeneity. We then demonstrate that IV-SGD (in Section 7.1) and IV-Q-Learning (in Section 7.2) can correct them.

7.1. IV-SGD

Let us revisit the digital advertising example in Section 2.2 and consider the model

$$Y_t = \theta_0^* + \theta_1^* S_t A_t + \epsilon_t, \quad t=1,\ldots,T,$$

where S_t's are i.i.d. with the log-normal distribution $\log(S_t)\sim\mathcal{N}(0,\frac{1}{4})$. Suppose that $\epsilon_t = bA_t^2 + o_t$, where b is some constant, and o_t's are i.i.d. with distribution $\mathcal{N}(0,\sigma_\epsilon^2)$. In addition, assume that the agent plays the following randomized policy at each time t:

(i) Draw a random sample $q_t \sim \mathsf{Bernoulli}(1-p)$.

(ii) Take the action $A_t = \tilde{\theta}S_t$, where $\tilde{\theta}$ is some pre-specified value if $q_t=1$; take a random action $A_t\sim\mathsf{Uniform}(0,1)$ if $q_t=0$.

Namely, the agent performs random exploration $A_t\sim\mathsf{Uniform}(0,1)$ with probability p, and plays a fixed policy $\tilde{\theta}S_t$ with probability $(1-p)$.

Let $X_t := (1, S_tA_t)^\intercal$. The standard SGD is biased. Assume that the agent knows that ϵ_t may depend on A_t but not S_t. Then, $q_t(S_t-\mathbb{E}[S_t])$ is a valid IV.[5] The validity and the calculation of the limit bias of the standard SGD are deferred to Appendix G.1.

The IV-SGD algorithm is then reduced to the following recursive equations:

$$\theta_{t+1} = \Pi_B\left(\theta_t + \alpha_t(Y_t - \theta_t^\intercal X_t)\begin{pmatrix}1\\ \Gamma_t Z_t\end{pmatrix}\right),$$
$$\Gamma_{t+1} = \Pi_B\left(\Gamma_t + \beta_t(S_tA_t - \Gamma_t Z_t)Z_t^\intercal\right),$$

where $Z_t := (1, q_t(S_t-\mathbb{E}[S_t]))^\intercal$ is the vector of IVs, and $\Gamma_t\in\mathbb{R}^{1\times 2}$.

[5] We assume that $\mathbb{E}[S_t]$ is known. Otherwise, we can replace it with the mean of S_t in the sample path. The algorithms work in a similar fashion to the one presented in the paper.

To compare the convergence of the SGD and IV-SGD algorithms, we consider eight different experimental designs, depending on specifications of three parameters: $p = 0.3, 0.7$ (probability of exploration), $b = 0.3, 0.7$ (magnitude of the bias in the residual term of the observed reward), and $T = 10000, 50000$ (number of iterations). Other parameters are specified as follows: $\sigma_\epsilon = 1.0$, $\widetilde{\theta} = 0.5$, $\kappa = 0.7$, $\delta = 0.9$, $\alpha_0 = 10$, $\beta_0 = 5$, $B = 3$, and $\theta_0 = (0.5, 0.5)^\intercal$, $\Gamma_0 = (1, 1)$. Each experiment is replicated independently for 1000 times. We focus on the estimation of θ_1^* and report the results in Tables 2 and 3.

Table 2 **Bias and RMSE for Estimating θ_1^* with IV-SGD and SGD.**

			IV-SGD		SGD	
T	p	b	Bias	RMSE	Bias	RMSE
10,000	0.3	0.3	-0.0065	0.1012	0.148	0.196
		0.7	-0.0228	0.1080	0.345	0.365
	0.7	0.3	-0.0074	0.0933	0.138	0.182
		0.7	-0.0026	0.0929	0.321	0.339
50,000	0.3	0.3	-0.0009	0.0532	0.148	0.159
		0.7	-0.0035	0.0555	0.325	0.329
	0.7	0.3	0.0004	0.0484	0.138	0.151
		0.7	0.0014	0.0542	0.326	0.330

The results are computed over 1000 independent replications.

Table 3 **Coverage of Confidence Intervals of θ_1^* Using IV-SGD.**

T	p	b	SD (est.)	SD (theo.)	Coverage
10,000	0.3	0.3	0.1010	0.0890	91.5%
		0.7	0.1050	0.0890	89.7%
	0.7	0.3	0.0929	0.0890	94.1%
		0.7	0.0929	0.0890	94.0%
50,000	0.3	0.3	0.0532	0.0507	93.2%
		0.7	0.0554	0.0507	91.7%
	0.7	0.3	0.0484	0.0507	94.9%
		0.7	0.0542	0.0507	94.4%

The results are computed over 1000 independent replications.

Table 2 shows the bias and root-mean-square error (RMSE) of IV-SGD and SGD to characterize their asymptotic behavior. First, in all eight experimental designs, the RMSE for IV-SGD asymptotically vanishes as T increases, whereas it is not the case for SGD, whose the non-diminishing RMSE is mostly attributed to its bias. Second, with everything else equal, the bias for SGD is larger for a larger value of b, and it is persistent as T increases. The bias for IV-SGD is substantially smaller, and it decays to zero as T increases, demonstrating the efficacy of the IV in de-biasing the estimation. Third, with T and b fixed, a larger p generally leads to a smaller bias and thus, a smaller RMSE, suggesting exploration can help to reduce bias.

Table 3 demonstrates distributional results for IV-SGD based on Corollary SM.1. Using the terminal iterate θ_T and the limiting distribution computed from Corollary SM.1, we construct 95% asymptotic confidence intervals for θ_1^* and report their coverage probabilities. The table shows that the coverage probabilities from our experiments are close to the theoretical predictions for T at $10,000$ and at $50,000$. A closer examination of the data suggests that the coverage is more accurate—closer to 95%—when the bias (b) is small, and the exploration probability (p) is large. These differences, however, are largely eliminated when T is increased from $10,000$ to $50,000$. Even for the worst case ($b = 0.7$ and $p = 0.3$) among all eight experimental designs, the coverage is 91.7%. For the best scenario, the coverage is at $= 94.9\%$, merely 0.1 percentage point from the theoretical prediction 95%.

7.2. IV-Q-Learning

We now study the performances of the IV-Q-Learning algorithm in an MDP setting. To demonstrate its performances and compare them to Q-Learning, we consider the following styled MDP problem whose optimal policy has an analytical solution. In the experimental design, the action is comprised by two parts where randomizing one part leads to causal exploration while randomizing the other part, albeit exploring, does not help us to discuss the true reward due to correlation with the error term. Specifically, the setup is as follows:

(i) The state transition follows $S_{t+1} = c_0 + c_1 S_t + c_2 A_t + \eta_t$, where c_0, c_1, and c_2 are some constants, and η_t is i.i.d. noise with mean 0 and variance σ_η^2.

(ii) The behavior policy that generates action A_t is a randomized policy as follows: $A_t = A_{t,1} + A_{t,2}$, where $A_{t,1}$ and $A_{t,2}$ are independent random variables.

(iii) The true reward function for a state-action pair is $r(s,a) = r_0 + r_1 a + r_2 sa + r_3 a^2$, where r_0, r_1, and r_2 are some constants.

(iv) The observed reward is $R(S_t, A_t) = r(S_t, A_t) + \epsilon_t$, where ϵ_t is endogenous; i.e., $\mathbb{E}_\nu[\epsilon_t | S_t, A_t] \neq 0$; specifically, $\epsilon_t = b A_{t,2}^2 + o_t$, where b is a constant, and o_t is i.i.d. noise with mean 0.

(v) $A_{t,1}$ is observable and taken as an IV at period t. Randomizing $A_{t,1}$ leads to causal exploration.

In this setup, if (i) the parameters involved in the state transition and the true reward function are all known, and (ii) there is no reward endogeneity, then it can be shown that the MDP problem that maximizes the expected cumulative discounted reward is essentially the same as the linear quadratic control problem. This allows us to derive analytically the optimal policy, which serves as a benchmark for our comparison; see Bertsekas (2017, Chapter 4.1). Similar calculations can be done in the presence of endogeneity as well, and the details are provided in Appendix G.2.

Note that even if the state transition and the functional form of the true reward functions are unknown, Q-Learning and IV-Q-Learning lead to respective policy functions from observing past rewards. Because the rewards are biased (due to the noise being correlated with part of the actions: $\epsilon_t = bA_{t,2}^2 + o_t$), Q-Learning leads to sub-optimal policy in the long run. This is because explorations in actions in Q-Learning also perturb the noises, making it impossible to discover the true rewards. In IV-Q-Learning, in contrast, exploration through $A_{t,1}$, which serves as an instrument as it affects the policy without affecting the noises, allows the true rewards to be discovered. This type of exploration, which is causal in nature, enables IV-Q-Learning to discover the optimal policy as we show below.

Assume that $Q^*(s,a) = \phi^\intercal(s,a)\theta^*$,[6] where $\phi(s,a) = (1, s, a, sa, s^2, a^2)^\intercal$ is a vector of basis functions, and $\theta^* \in \mathbb{R}^6$ is a vector of unknown parameters to be learned. Because the agent knows only that ϵ_t does not depend on $A_{t,1}$, a transformation of $A_{t,1}$ and S_t can serve as an IV, provided that A_t is executed by a random policy that does not depend on S_t. Note that three variables—A_t, S_tA_t, and A_t^2—are potentially endogenous. Thus, we construct three IVs—$A_{t,1}$, $S_tA_{t,1}$, and $A_{t,1}^2$, populate the vector of IVs in IV-Q-Learning with $Z_t = (1, S_t, A_{t,1}, S_tA_{t,1}, S_t^2, A_{t,1}^2)^\intercal$, where 1, S_t, and S_t^2 are the IVs for themselves.

The parameters involved in the MDP are chosen as follows: the discount factor $\gamma = 0.8$; the state transition is specified by $c_0 = 0.5$, $c_1 = 0.4$, $c_2 = 0.2$, and $\eta \sim \mathsf{Uniform}[-\sqrt{3}, \sqrt{3}]$ (so that it has mean 0 and variance $\sigma_\eta = 1$); the behavior policy is specified by $A_{t,1} \sim \mathsf{Beta}(1.0, 1.5)$, $A_{t,2} \sim \mathsf{Beta}(1.0, 1.5)$; the true reward function is specified by $r_0 = r_1 = 0$, $r_2 = 1$, and $r_3 = -1$; and the error term of reward R_t is specified by $b = 0.8$ and $o_t \sim \mathsf{Uniform}(-0.25, 0.25)$. In addition, the parameters involved in IV-Q-Learning and Q-Learning are specified as follows: $\alpha_0 = 15$, $\beta_0 = 10$, $\kappa = 0.7$, $\delta = 1.0$, $\theta_0 = (2.5, 0.5, 0.2, 0.5, 0.5, -1.5)^\intercal$, and Γ_0 is initialized randomly. Both algorithms are run for $T = 10^6$ iterations. The results are presented in Figures 1–3.

Figure 1 presents sample paths of the estimates of $\theta^* = (\theta_0^*, \theta_1^*, \ldots, \theta_5^*)$ produced by IV-Q-Learning and Q-Learning, respectively. Clearly, the estimates from IV-Q-Learning converge to the true value of θ^*. In contrast, Q-Learning fails to do so, notably for θ_2^* and θ_5^*, both of which directly affect the estimation of the optimal policy function; see Appendix G.2 for the expression of the optimal policy.

To assess the impact of the bias in parameter estimation on the induced policy, we compute the long-term opportunity cost (LTOC), defined as the difference between $V^*(s)$, the optimal value function, and $V_{\hat{\pi}_t}(s)$, the value function of the policy induced by the estimate of θ^* at t, for IV-Q-Learning and Q-Learning. See Appendix G.2 for the detailed calculation of the value functions. Such a cost represents the counterfactual loss in cumulative value if one executes a policy based on

[6] This assumption is valid for the linear quadratic problem considered here, which can be shown by solving Bellman's equation; see Appendix G.2.

Figure 1 **Sample Paths of Estimates of θ^*.**

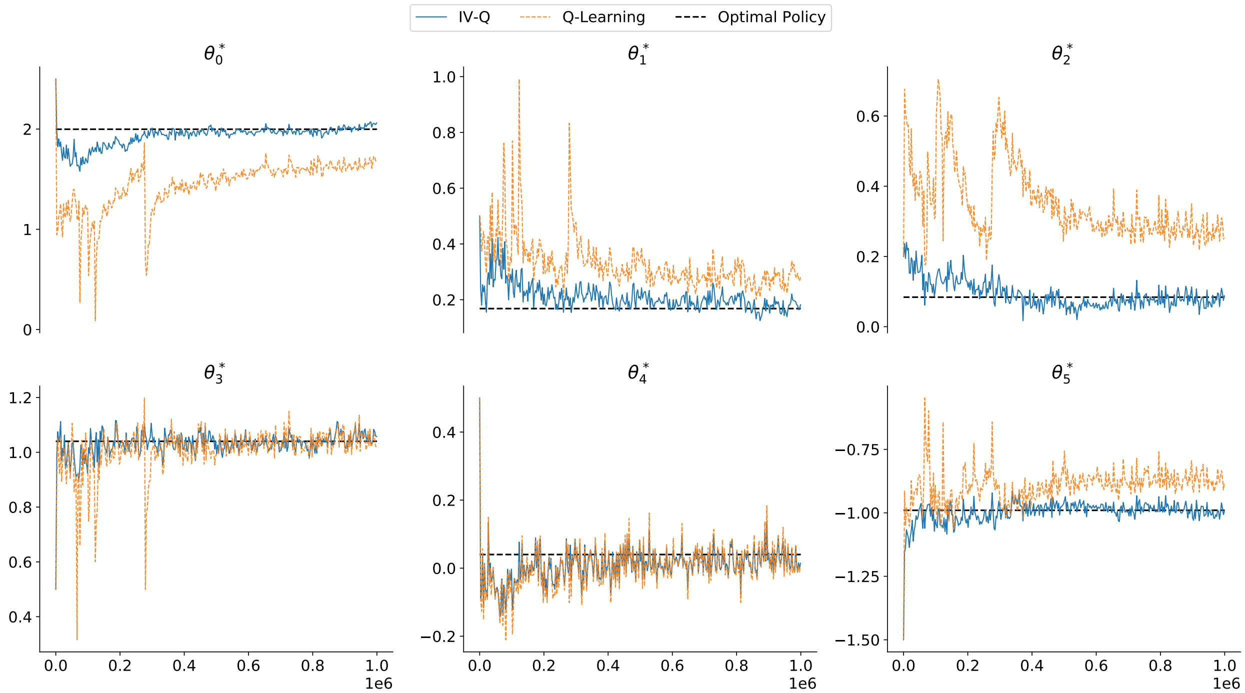
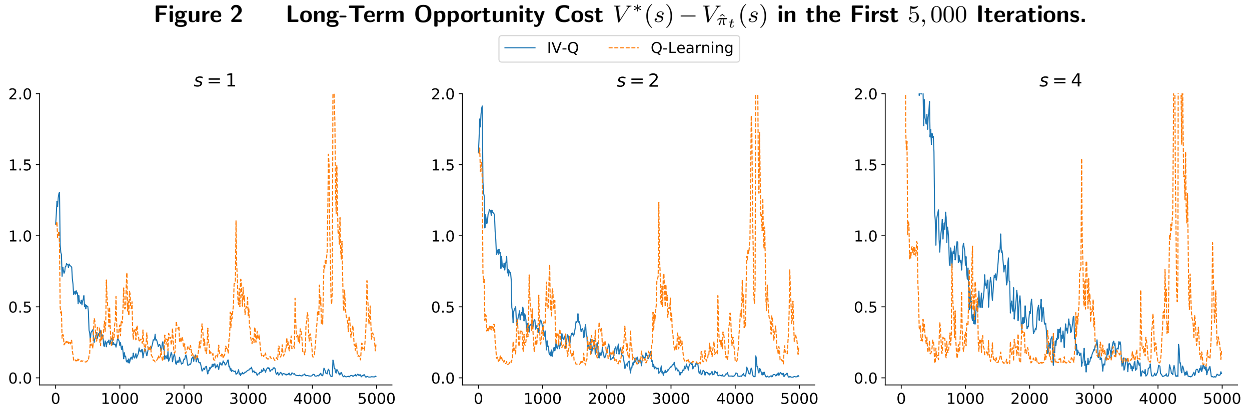


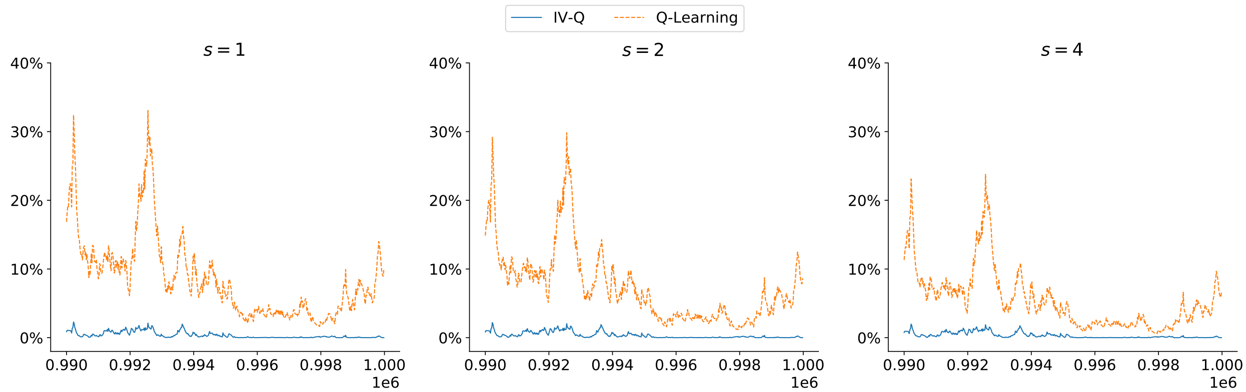
Figure 2 **Long-Term Opportunity Cost $V^*(s) - V_{\hat{\pi}_t}(s)$ in the First $5,000$ Iterations.**



the estimate of θ^* at time t. We plot in Figure 2 the LTOC for early iterations $t = 1, \ldots, 5000$ for three representative states $s = 1, 2, 4$. Clearly, the policy induced by IV-Q-Learning quickly reduces the gap to the optimal policy in several thousand iterations. In contrast, during the same time periods, the policy induced by Q-Learning suffers from a sizable and highly varying LTOC.

Further, we plot in Figure 3 the relative LTOC, that is, $1 - \frac{V_{\hat{\pi}_t}(s)}{V^*(s)}$, for the last $10,000$ iterations to examine the asymptotic behavior of the two algorithms. The policy induced by Q-Learning results in 5% to 40% loss in the value overall, whereas, in stark contrast, the relative LTOC for IV-Q-Learning in these late iterations is nearly zero.

Figure 3 **Relative Long-Term Opportunity Cost** $1 - \frac{V_{\hat{\pi}_t}(s)}{V^*(s)}$ **in the Last** $10,000$ **Iterations.**



8. Conclusions

In the present paper, we examine causality problems in general MDPs when the observed rewards are biased. In this class of problems, the data analyst is also part of the data generator, and this interaction generates R-bias. To deal with R-bias, we propose a class of IV-RL algorithms. We establish various theoretical properties of our algorithms by formulating them as two-timescale SA under Markovian noise. New techniques are developed to establish risk bounds, finite-time bounds for trajectory stability, and asymptotic normality for nonlinear two-timescale SA. We apply these results to IV-Q-Learning and beyond.

In the present paper, we focus on theoretical advances of IV-RL algorithms and provide the first step toward a rich research program. On the applied side, we have not considered the economic cost of creating instrument variables. If firms can create different instruments, it will be useful to examine the "optimal" instruments (in terms of economic costs rather than in a statistical sense). On the theoretical side, we did not investigate multi-agent MDPs with biased rewards. We also did not examine the use of IVs in hierarchical RL algorithms that are useful when decisions can have long-term consequences. In addition, this study points toward a framework that allows the comparison between short-term and long-term treatment effects. Explorations of these topics form a rich avenue for future research.

References

Ai C, Chen X (2003) Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica* 71(6):1795–1843.

Alchian AA, Demsetz H (1972) Production, information costs, and economic organization. *American Economic Review* 62(5):777–795.

Angrist JD (1990) Lifetime earnings and the Vietnam era draft lottery: Evidence from social security administrative records. *American Economic Review* 80(3):313–336.

Angrist JD, Evans WN (1998) Children and their parents' labor supply: Evidence from exogenous variation in family size. *American Economic Review* 88(3):450–477.

Angrist JD, Krueger AB (1991) Does compulsory school attendance affect schooling and earnings. *Quarterly Journal of Economics* 106(4):979–1014.

Angrist JD, Krueger AB (1999) Empirical strategies in labor economics. Ashenfelter OC, Card D, eds., *Handbook of Labor Economics: Volume 3*, 1277–1366 (Elsevier).

Angrist JD, Pischke S (2009) *Mostly Harmless Econometrics* (Princeton University Press), 1st edition.

Athey S, Imbens G (2017) The econometrics of randomized experimentsa. Banerjee AV, Duflo E, eds., *Handbook of Field Experiments*, volume 1, 73–140 (North-Holland).

Belloni A, Chen D, Chernozhukov V, Hansen C (2012) Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* 80(6):2369–2429.

Bertsekas DP (2017) *Dynamic Programming and Optimal Control: Volume 1* (Athena Scientific), 4th edition.

Bertsekas DP, Tsitsiklis JN (1996) *Neuro-Dynamic Programming* (Athena Scientific).

Bhandari J, Russo D, Singal R (2018) A finite time analysis of temporal difference learning with linear function approximation. *Proceedings of the 31st Conference on Learning Theory*, 1691–1692.

Blake T, Nosko C, Tadelis S (2015) Consumer heterogeneity and paid search effectiveness: A large-scale field experiment. *Econometrica* 83(1):155–174.

Borkar VS, Konda VR (1997) The actor-critic algorithm as multi-time-scale stochastic approximation. *Sādhanā* 22(4):525–543.

Bottou L (2010) Large-scale machine learning with stochastic gradient descent. *Proceedings of the 19th International Symposium on Computational Statistics*, 177–186.

Calvano E, Calzolari G, Denicolò V, Pastorello S (2020) Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review* 110(10):3267–3297.

Chen X, Hong H, Tamer E (2005) Measurement error models with auxiliary data. *Review of Economic Studies* 72(2):343–366.

Chen Z, Zhang S, Doan TT, Maguluri ST, Clarke JP (2019) Performance of Q-learning with linear function approximation: Stability and finite-time analysis. *arXiv:1905.11425*.

Chernozhukov V, Hansen C (2005) An IV model of quantile treatment effects. *Econometrica* 73(1):245–261.

Chernozhukov V, Imbens GW, Newey WK (2007) Instrumental variable estimation of nonseparable models. *Journal of Econometrics* 139(1):4–14.

Cowgill B, Tucker CE (2020) Algorithmic fairness and economics. Preprint available at *SSRN:3361280*.

Dalal G, Thoppe G, Szörényi B, Mannor S (2018) Finite sample analysis of two-timescale stochastic approximation with applications to reinforcement learning. *Proceedings of the 31st Conference on Learning Theory*, 1199–1233.

Escobar LA, Moser EB (1993) A note on the updating of regression estimates. *The American Statistician* 47(3):192–194.

Ethier SN, Kurtz TG (1986) *Markov Processes: Characterization and Convergence* (Wiley).

François-Lavet V, Henderson P, Islam R, Bellemare MG, Pineau J (2018) An introduction to deep reinforcement learning. *Foundations and Trends® in Machine Learning* 11(3-4):219–354.

Griliches Z (1977) Estimating the returns to schooling: Some econometric problems. *Econometrica* 45(1):1–22.

Gupta H, Srikant R, Ying L (2019) Finite-time performance bounds and adaptive learning rate selection for two time-scale reinforcement learning. *Advances in Neural Information Processing Systems 32*, 4704–4713.

Hansen LP (1982) Large sample properties of generalized method of moments estimators. *Econometrica* 50(4):1029–1054.

Hausman JA (1983) Specification and estimation of simultaneous equation models. *Handbook of Econometrics*, volume 1, chapter 7, 391–448 (Elsevier).

Heckman JJ (1979) Sample selection bias as a specification error. *Econometrica* 47(1):153–161.

Heckman JJ (1990) Varieties of selection bias. *American Economic Review: Papers and Proceedings* 80(2):313–318.

Holmström B, Milgrom P (1991) Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, & Organization* 7:24–52.

Imbens GW (2014) Instrumental variables: An econometrician's perspective. *Statistical Science* 29(3):323–358.

Imbens GW (2020) Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature* 58(4):1129–1179.

Imbens GW, Angrist JD (1994) Identification and estimation of local average treatment effects. *Econometrica* 62(2):467–475.

Johnson J, Rhodes A, Wildenbeest MR (2020) Platform design when sellers use pricing algorithms. Preprint available at *SSRN:3691621*.

Kaledin M, Moulines E, Naumov A, Tadic V, Wai HT (2020) Finite time analysis of linear two-timescale stochastic approximation with Markovian noise. *Proceedings of the 33th Conference on Learning Theory*, 2144–2203.

Konda VR, Tsitsiklis JN (2003) On actor-critic algorithms. *SIAM Journal on Control and Optimization* 42(4):1143–1166.

Konda VR, Tsitsiklis JN (2004) Convergence rate of linear two-time-scale stochastic approximation. *Annals of Applied Probability* 14(2):796–819.

Kushner HJ, Yin GG (2003) *Stochastic Approximation and Recursive Algorithms and Applications* (Springer-Verlag), 2nd edition.

Levitt S (1997) Using electoral cycles in police hiring to estimate the effect of police on crime. *American Economic Review* 87(3):270–290.

Li D, Raymond L, Bergman P (2020) Hiring as exploration. Preprint available at *SSRN:3630630*.

Loève M (1978) *Probability Theory II* (Springer-Verlag), 4th edition.

Manski CF (1995) *Identification Problems in the Social Sciences* (Harvard University Press).

McMorrow R, Liu N (2020) China's Meituan and Ele.me tackle backlash against demands on couriers. *Financial Times* (September 10), https://www.ft.com/content/bcc453b2-0da5-41a5-af76-eb43b044e822.

Melo F, Meyn S, Ribeiro I (2008) An analysis of reinforcement learning with function approximation. *Proceedings of the 25th Annual International Conference on Machine Learning*, 664–671.

Meyn SP, Tweedie RL (2009) *Markov Chains and Stochastic Stability* (Cambridge University Press), 2nd edition.

Mnih V, Kavukcuoglu K, Silver D, et al (2015) Human-level control through deep reinforcement learning. *Nature* 518:529–533.

Mokkadem A, Pelletier M (2006) Convergence rate and averaging of nonlinear two-time-scale stochastic approximation algorithms. *Annals of Applied Probability* 16(3):1671–1702.

Morris S (2001) Political correctness. *Journal of Political Economy* 109(2):231–265.

Muller JZ (2018) *The Tyranny of Metrics* (Princeton University Press).

Nambiar M, Simchi-Levi D, Wang H (2019) Dynamic learning and pricing with model misspecification. *Management Science* 65(11):4980–5000.

Newey WK, Powell JL (2003) Instrumental variable estimation of nonparametric models. *Econometrica* 71(5):1565–1578.

Newey WK, West KD (1987) A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55(3):703–708.

Prendergast C (1993) A theory of "yes men.". *American Economic Review* 83(4):757–70.

Rumelhart DE, McClelland JL (1986) Learning internal representations by error propagation. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1, 318–362 (MIT Press).

Schrittwieser J, Antonoglou I, Hubert T, et al (2020) Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature* 588(7839):604–609.

Silver D, Huang A, Maddison CJ, et al (2016) Mastering the game of Go with deep neural networks and tree search. *Nature* 529(7587):484–489.

Silver D, Hubert T, Schrittwieser J, et al (2018) A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* 362(6419):1140–1144.

Smith R (2020) How algorithms are changing what we read online. *The Walrus* (September 15), https://thewalrus.ca/how-algorithms-are-changing-what-we-read-online.

Sutton R (1988) Learning to predict by the methods of temporal differences. *Machine Learning* 3:9–44.

Sutton RS, Barto AG (2018) *Reinforcement Learning: An Introduction* (MIT Press), 2nd edition.

Thomke SH (2020) *Experimentation Works: The Surprising Power of Business Experiments* (Harvard Business Review Press).

Thoppe G, Borkar V (2019) A concentration bound for stochastic approximation via Alekseev's formula. *Stochastic Systems* 9(1):1–26.

Tsitsiklis J, Van Roy B (1997) An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control* 42(5):674–690.

Tsitsiklis JN (1994) Asynchronous stochastic approximation and Q-learning. *Machine Learning* 16:185–202.

Vincent P, Larochelle H, Bengio Y, Manzagol PA (2008) Extracting and composing robust features with denoising autoencoders. *Proceedings of the 25th Annual International Conference on Machine Learning*, 1096–1103.

Williamson OE (1985) *The Economic Institutions of Capitalism: Firms, Markets, Relational Contracting* (The Free Press).

Young PC (2011) *Recursive Estimation and Time-Series Analysis: An Introduction for the Student and Practitioner* (Springer), 2nd edition.

Zou S, Xu T, Liang Y (2019) Finite-sample analysis for SARSA with linear function approximation. *Advances in Neural Information Processing Systems 32*, 8668–8678.

Supplemental Material

A. Proof of Proposition 1

We provide several technical lemmas used for the proof of Proposition 1 in Section A.1, prove the error bound on ξ_t in Section A.2, and prove the error bound on λ_t in Section A.3. Throughout this section, we let $\alpha_t = \alpha_0 t^{-\kappa}$ for some $\alpha_0 > 0$ and $\kappa \in (0,1]$ for all $t \geq 1$ and let U and $\rho \in [0,1)$ be the function and constant, respectively, defined in Assumption 1; we also suppose that Assumption 1 holds with $d_U = 1, 2$ and Assumption 2 holds.

A.1. Technical Lemmas

Lemma SM.1. *There exists some constant $c_1 > 0$ such that*

$$\int_{\mathcal{W}} |f(w_t)||\mathbb{P}(w_t|W_0 = w) - \nu(w_t)|\,\mathrm{d}w_t \leq c_1 \rho^t U(w), \tag{A.1}$$

for all $t \geq 0$, $w \in \mathcal{W}$, and measurable function $f : \mathcal{W} \mapsto \mathbb{R}$ for which $|f(\cdot)| \leq U(\cdot)$. Moreover,

$$K_{d_U} := \mathbb{E}_\nu[U(W)] = \mathbb{E}_\nu[1 + \|W_t\|^{d_U}] < \infty, \tag{A.2}$$

and

$$\mathbb{E}[(1 + \|W_t\|^{d_U})|W_0 = w] \leq C_{w,d_U}, \tag{A.3}$$

for all $t \geq 0$ and $w \in \mathcal{W}$, where $C_{w,d_U} := c_1(1 + \|w\|^{d_U}) + K_{d_U}$.

Proof of Lemma SM.1. The statements (A.1) and (A.2) follow from Theorem 16.0.1 and Theorem 14.0.1 of Meyn and Tweedie (2009), respectively. To show (A.3), note that

$$\begin{aligned} &\mathbb{E}[(1 + \|W_t\|^{d_U})|W_0 = w] \\ \leq &\int_{\mathcal{W}} |\mathbb{P}(w_t|W_0 = w) - \nu(w_t)|(1 + \|w_t\|^{d_U})\,\mathrm{d}w_t + \mathbb{E}_\nu[(1 + \|W_t\|^{d_U})] \\ \leq &\, c_1\rho(1 + \|w\|^{d_U}) + K_{d_U}. \quad \square \end{aligned}$$

Lemma SM.2. *For all $t \geq 0$, $w \in \mathcal{W}$, $\lambda \in \mathbb{R}^{\mathsf{d}_\lambda}$, and $\xi \in \mathbb{R}^{\mathsf{d}_\xi}$ with $\|\lambda\| \leq B$ and $\|\xi\| \leq B$, we have:*

$$\left\|\mathbb{E}[G(W_t, \lambda, \xi)|W_0 = w] - \bar{G}(\lambda, \xi)\right\| \leq c_1 \mathsf{d}_\lambda L \rho^t (1 + \|w\|), \tag{A.4}$$

$$\left\|\mathbb{E}[H(W_t, \xi)|W_0 = w] - \bar{H}(\xi)\right\| \leq c_1 \mathsf{d}_\xi L \rho^t (1 + \|w\|). \tag{A.5}$$

Proof of Lemma SM.2. We only prove (A.4), and (A.5) can be proved similarly. Applying Lemma SM.1 with $U(\cdot) = 1 + \|\cdot\|$ leads to

$$\int_{\mathcal{W}} |f(w_t)||\mathbb{P}(w_t|W_0 = w) - \nu(w_t)|\,\mathrm{d}w_t \leq c_1 \rho^t (1 + \|w\|), \tag{A.6}$$

for all $t \ge 0$, $w \in \mathcal{W}$, and $f : \mathcal{W} \mapsto \mathbb{R}$ such that $|f(\cdot)| \le 1 + \|\cdot\|$.

Fix arbitrary $\lambda \in \mathbb{R}^{\mathsf{d}_\lambda}$ and $\xi \in \mathbb{R}^{\mathsf{d}_\xi}$ with $\|\lambda\| \le B$ and $\|\xi\| \le B$. Let G_i and $\bar{G}_i$ denote the i-th component of G and $\bar{G}$, respectively. Then, $|G_i(\cdot, \lambda, \xi)| \le L(1 + \|\cdot\|)$ by the condition (22) in Assumption 2. Thus, for all $t \ge 0$, $w \in \mathcal{W}$,

$$\begin{aligned} \left\|\mathbb{E}[G(W_t, \lambda, \xi)|W_0 = w] - \bar{G}(\lambda, \xi)\right\| &\le \sum_{i=1}^{\mathsf{d}_\lambda} |\mathbb{E}[G_i(W, \lambda, \xi)|W_0 = w] - \mathbb{E}_\nu[G_i(W, \lambda, \xi)]| \\ &= \sum_{i=1}^{\mathsf{d}_\lambda} \int_{\mathcal{W}} |G_i(w_t, \lambda, \xi)| \, |\mathbb{P}(w_t|W_0 = w) - \nu(w_t)| \, \mathrm{d}w_t \\ &\le \mathsf{d}_\lambda \int_{\mathcal{W}} L(1 + \|w_t\|) |\mathbb{P}(w_t|W_0 = w) - \nu(w_t)| \, \mathrm{d}w_t \\ &\le c_1 \mathsf{d}_\lambda L \rho^t (1 + \|w\|). \qquad \square \end{aligned}$$

Lemma SM.3. *For all $t \ge 1$, define $\sigma_t := \lfloor c_2 \ln t + c_3 \rfloor$, where*

$$c_2 := \frac{\kappa}{\ln(1/\rho)} \quad \text{and} \quad c_3 := \frac{\ln(2c_1 \mathsf{d}_\lambda L/(\rho \alpha_0^2))}{\ln(1/\rho)} + 1. \tag{A.7}$$

Then, for all t large enough such that $t > \sigma_t$, $u \in [\sigma_t, t]$, $\lambda \in \mathbb{R}^{\mathsf{d}_\lambda}$, and $\xi \in \mathbb{R}^{\mathsf{d}_\xi}$ with $\|\lambda\| \le B$ and $\|\xi\| \le B$, we have:

$$\|\mathbb{E}[G(W_t, \lambda, \xi)|W_{t-u}] - \bar{G}(\lambda, \xi)\| \le \alpha_t^2 (1 + \|W_{t-u}\|), \tag{A.8}$$

$$\|\mathbb{E}[H(W_t, \xi)|W_{t-u}] - \bar{H}(\xi)\| \le \beta_t^2 (1 + \|W_{t-u}\|). \tag{A.9}$$

Proof of Lemma SM.3. Fix arbitrary $u \le t$, $\lambda \in \mathbb{R}^{\mathsf{d}_\lambda}$, and $\xi \in \mathbb{R}^{\mathsf{d}_\xi}$ with $\|\lambda\| \le B$ and $\|\xi\| \le B$. The inequality (A.4) in Lemma SM.2 implies that

$$\left\|\mathbb{E}[G(W_t, \lambda, \xi)|W_{t-u}] - \bar{G}(\lambda, \xi)\right\| \le c_1 \mathsf{d}_\lambda L \rho^u (1 + \|W_{t-u}\|).$$

It is easy to see that if $u \ge \sigma_t$, then

$$u \ge \frac{2\ln(1/\alpha_t) + \ln(2c_1 \mathsf{d}_\lambda L)/\rho)}{\ln(1/\rho)},$$

which is equivalent to $c_1 \mathsf{d}_\lambda L \rho^u \le \alpha_t^2$. Hence, (A.8) holds if $u \in [\sigma_t, t]$. $\square$

Lemma SM.4. *For all $t \in [0, T]$, define $\alpha_{t,T} := \sum_{\ell=t}^T \alpha_\ell$ and $\beta_{t,T} := \sum_{\ell=t}^T \beta_\ell$. Then,*

$$\|\lambda_{T+1} - \lambda_t\| \le L \sum_{\ell=t}^T \alpha_\ell (1 + \|W_\ell\|), \tag{A.10}$$

$$\mathbb{E}[\|\lambda_{T+1} - \lambda_t\| |W_0 = w] \le L C_{w,1} \alpha_{t,T}, \tag{A.11}$$

$$\|\xi_{T+1} - \xi_t\| \le L \sum_{\ell=t}^T \beta_\ell (1 + \|W_\ell\|), \tag{A.12}$$

$$\mathbb{E}[\|\xi_{T+1} - \xi_t\| |W_0 = w] \le L C_{w,1} \beta_{t,T}, \tag{A.13}$$

for all $w \in \mathcal{W}$, where $C_{w,1} = c_1(1 + \|w\|) + K_1$ is the constant C_{w,d_U} defined in Lemma SM.1 with $d_U = 1$.

Proof of Lemma SM.4. We only prove (A.10) and (A.11); the inequalities (A.12) and (A.13) can be proved similarly. Note that

$$\begin{aligned}\|\lambda_{T+1}-\lambda_t\| &\le \sum_{\ell=t}^{T}\|\lambda_{\ell+1}-\lambda_\ell\| \\ &= \sum_{\ell=t}^{T}\|\Pi_B(\lambda_\ell+\alpha_\ell G(W_\ell,\lambda_\ell,\xi_\ell))-\Pi_B(\lambda_\ell)\| \\ &\le \sum_{\ell=t}^{T}\|(\lambda_\ell+\alpha_\ell G(W_\ell,\lambda_\ell,\xi_\ell))-\lambda_\ell\| \\ &= \sum_{\ell=t}^{T}\alpha_\ell\|G(W_\ell,\lambda_\ell,\xi_\ell)\| \\ &\le \sum_{\ell=t}^{T}\alpha_\ell L(1+\|W_\ell\|),\end{aligned}$$

where the second inequality follows from the fact that Π_B is an orthogonal projection operator onto a convex set, thereby being non-expansive, while the last step follows from the condition (22) in Assumption 2. Therefore,

$$\mathbb{E}[\|\lambda_{T+1}-\lambda_t\||W_0=w]\le\sum_{\ell=t}^{T}\alpha_\ell L\mathbb{E}[1+\|W_\ell\||W_0=w]\le\sum_{\ell=t}^{T}\alpha_\ell LC_{w,1}=LC_{w,1}\alpha_{t,T},$$

where the second inequality follows from (A.3). □

LEMMA SM.5. *Fix $w\in\mathcal{W}$. For all $t\ge1$, define*

$$\begin{aligned}f_t &:= \max\left\{\mathbb{E}[\|\lambda_t-\lambda^*\|^2|W_0=w],(L/\zeta)^2 g_t\right\}, \\ g_t &:= \mathbb{E}[\|\xi_t-\xi^*\|^2|W_0=w],\end{aligned} \tag{A.14}$$

where ζ is the constant defined in (24)*. Then,*

$$\mathbb{E}[\|\lambda_{t+1}-\lambda^*\|^2|W_0=w]\le\left(1-\frac{k_0}{t^\kappa}\right)f_t+\frac{\widetilde{k}_1\ln t+\widetilde{k}_2}{t^{2\kappa}}, \tag{A.15}$$

for all t large enough, where $k_0:=\zeta\alpha_0$, and $\widetilde{k}_1\ge0$ and $\widetilde{k}_2>0$ are some constants that depend on w. Moreover, $\widetilde{k}_1=0$ if $\{W_t:t\ge0\}$ are i.i.d.

Proof of Lemma SM.5. The key of the proof is to construct an upper bound on

$$\mathbb{E}[\|\lambda_{t+1}-\lambda^*\|^2|W_0=w]-\mathbb{E}[\|\lambda_t-\lambda^*\|^2|W_0=w],$$

for all t large enough. To that end, let us begin with decomposing $\mathbb{E}\big[\|\lambda_{t+1}-\lambda^*\|^2\big]$ as follows,

$$\|\lambda_{t+1}-\lambda^*\|^2=\big\|\Pi_B\big(\lambda_t+\alpha_t G(W_t,\lambda_t,\xi_t)\big)-\lambda^*\big\|^2$$

$$
\begin{aligned}
&= \left\|\Pi_B\big(\lambda_t + \alpha_t G(W_t,\lambda_t,\xi_t)\big) - \Pi_B(\lambda^*)\right\|^2 \\
&\le \|\lambda_t + \alpha_t G(W_t,\lambda_t,\xi_t) - \lambda^*\|^2 \\
&\le \|\lambda_t - \lambda^*\|^2 + 2\alpha_t(\lambda_t-\lambda^*)^\intercal G(W_t,\lambda_t,\xi_t) + \alpha_t^2\|G(W_t,\lambda_t,\xi_t)\|^2, \qquad \text{(A.16)}
\end{aligned}
$$

where the second equality holds because $\|\lambda^*\| \le B$.

Noting that $\bar{G}(\lambda^*,\xi^*)=0$,

$$
\begin{aligned}
&G(W_t,\lambda_t,\xi_t) \\
&= [G(W_t,\lambda_t,\xi_t) - G(W_t,\lambda_t,\xi^*)] + [G(W_t,\lambda_t,\xi^*) - \bar{G}(W_t,\lambda_t,\xi^*)] + [\bar{G}(\lambda_t,\xi^*) - \bar{G}(\lambda^*,\xi^*)]. \qquad \text{(A.17)}
\end{aligned}
$$

It then follows from (A.16) and (A.17) that

$$
\|\lambda_{t+1}-\lambda^*\|^2 \le \|\lambda_t-\lambda^*\|^2 + 2\alpha_t(\Psi_1+\Psi_2+\Psi_3) + \alpha_t^2\Psi_4, \qquad \text{(A.18)}
$$

where

$$
\begin{aligned}
\Psi_1 &:= (\lambda_t-\lambda^*)^\intercal[G(W_t,\lambda_t,\xi_t) - G(W_t,\lambda_t,\xi^*)], \\
\Psi_2 &:= (\lambda_t-\lambda^*)^\intercal[G(W_t,\lambda_t,\xi^*) - \bar{G}(\lambda_t,\xi^*)], \\
\Psi_3 &:= (\lambda_t-\lambda^*)^\intercal[\bar{G}(\lambda_t,\xi^*) - \bar{G}(\lambda^*,\xi^*)], \\
\Psi_4 &:= \|G(W_t,\lambda_t,\xi_t)\|^2.
\end{aligned}
$$

In the sequel, we will upper bound Ψ_i in terms of $\|\lambda_t-\lambda^*\|^2$ and α_t , for each $i=1,\dots,4$.

Step 1: Bounding Ψ_1. Our goal in this step is to show

$$
\mathbb{E}[|\Psi_1||W_0=w] \le \frac{\zeta}{2} f_t + C_{\Psi_1}\alpha_t\sigma_t + 4BC_{w,1}\alpha_t^2, \qquad \text{(A.19)}
$$

for all t large enough, where $C_{\Psi_1} := 8\sqrt{2}L^2C_{w,2}(1+B+B\beta_0\alpha_0^{-1}) + 8BLC_{w,1}$ and σ_t is defined in Lemma SM.3. Define

$$
\Psi_{1,\sigma_t} := (\lambda_{t-\sigma_t}-\lambda^*)^\intercal[G(W_t,\lambda_{t-\sigma_t},\xi_{t-\sigma_t}) - G(W_t,\lambda_{t-\sigma_t},\xi^*)].
$$

By the triangular inequality,

$$
|\Psi_1| \le |\Psi_1 - \Psi_{1,\sigma_t}| + |\Psi_{1,\sigma_t}|. \qquad \text{(A.20)}
$$

Our strategy is to establish bounds for $\mathbb{E}[|\Psi_1-\Psi_{1,\sigma_t}||W_0=w]$ and $\mathbb{E}[|\Psi_{1,\sigma_t}||W_0=w]$ separately. It is easy to see that

$$
\begin{aligned}
\Psi_1 - \Psi_{1,\sigma_t} &= (\lambda_{t-\sigma_t}-\lambda_t)^\intercal[G(W_t,\lambda_{t-\sigma_t},\xi_{t-\sigma_t}) - G(W_t,\lambda_{t-\sigma_t},\xi^*] \\
&\quad + (\lambda_t-\lambda^*)^\intercal[G(W_t,\lambda_t,\xi_t) - G(W_t,\lambda_{t-\sigma_t},\xi_{t-\sigma_t}) - G(W_t,\lambda_t,\xi^*) + G(W_t,\lambda_{t-\sigma_t},\xi^*)].
\end{aligned}
$$

Hence,

$$\begin{aligned}
&\mathbb{E}[|\Psi_1 - \Psi_{1,\sigma_t}||W_0 = w] \\
&\leq \mathbb{E}[\|\lambda_{t-\sigma_t} - \lambda_t\|\|G(W_t, \lambda_{t-\sigma_t}, \xi_{t-\sigma_t}) - G(W_t, \lambda_{t-\sigma_t}, \xi^*)\||W_0 = w] \\
&\quad + \mathbb{E}[\|\lambda_t - \lambda^*\|\|G(W_t, \lambda_t, \xi_t) - G(W_t, \lambda_{t-\sigma_t}, \xi_{t-\sigma_t}) - G(W_t, \lambda_t, \xi^*) + G(W_t, \lambda_{t-\sigma_t}, \xi^*)\||W_0 = w] \\
&\leq 2L\,\mathbb{E}[\|\lambda_{t-\sigma_t} - \lambda_t\|(1 + \|W_t\|)|W_0 = w] \\
&\quad + 4BL\,\mathbb{E}[(\|\lambda_t - \lambda_{t-\sigma_t}\| + \|\xi_t - \xi_{t-\sigma_t}\|)(1 + \|W_t\|)|W_0 = w],
\end{aligned} \tag{A.21}$$

where the second inequality follows from the conditions (20) an (22) in Assumption 2 and the fact that $\max\{\|\lambda_t\|, \|\lambda^*\|\} \leq B$.

Next, note that

$$\mathbb{E}[(1 + \|W_t\|)^2|W_0 = w] \leq 2\,\mathbb{E}[1 + \|W_t\|^2|W_0 = w] \leq 2C_{w,2}, \tag{A.22}$$

where $C_{w,2} = c_1(1 + \|w\|^2) + K_2$ is the constant C_{w,d_U} defined in Lemma SM.1 with $d_U = 2$.

Since $\alpha_t = \alpha_0 t^{-\kappa}$, it is easy to see that for any t large enough, $\alpha_t \leq \alpha_{t-u} \leq \sqrt{2}\alpha_t$, for all $u = 1, \ldots, \sigma_t$, by the definition of σ_t. We then apply the inequality (A.10) in Lemma SM.4 to deduce that

$$\begin{aligned}
\mathbb{E}[\|\lambda_{t-\sigma_t} - \lambda_t\|^2|W_0 = w] &\leq \mathbb{E}\left(\sum_{\ell=t-\sigma_t}^{t-1} L\alpha_\ell(1 + \|W_\ell\|)\big|W_0 = w\right)^2 \\
&\leq 2L^2\alpha_t^2\,\mathbb{E}\left(\sum_{\ell=t-\sigma_t}^{t-1} (1 + \|W_\ell\|)\big|W_0 = w\right)^2 \\
&\leq 2L^2\sigma_t\alpha_t^2 \sum_{\ell=t-\sigma_t}^{t-1} \mathbb{E}[(1 + \|W_\ell\|)^2|W_0 = w] \\
&\leq 4L^2 C_{w,2}\sigma_t^2\alpha_t^2,
\end{aligned} \tag{A.23}$$

for all t large enough, where the last step follows from (A.22).

Likewise, it can be shown that for all t large enough,

$$\mathbb{E}[\|\xi_{t-\sigma_t} - \xi_t\|^2] \leq 4L^2 C_{w,2}\sigma_t^2\beta_t^2. \tag{A.24}$$

Using the Cauchy–Schwarz inequality and plugging (A.23) and (A.24) in (A.21),

$$\begin{aligned}
&\mathbb{E}[|\Psi_1 - \Psi_{1,\sigma_t}||W_0 = w] \\
&\leq 2L\,\mathbb{E}[\|\lambda_{t-\sigma_t} - \lambda_t\|^2|W_0 = w]^{\frac{1}{2}}\,\mathbb{E}[(1 + \|W_t\|)^2|W_0 = w]^{\frac{1}{2}} \\
&\quad + 4BL(\mathbb{E}[\|\lambda_t - \lambda_{t-\sigma_t}\|^2|W_0 = w]^{\frac{1}{2}} + \mathbb{E}[\|\xi_t - \xi_{t-\sigma_t}\|^2|W_0 = w]^{\frac{1}{2}})\,\mathbb{E}[(1 + \|W_t\|)^2|W_0 = w]^{\frac{1}{2}} \\
&\leq 2L \cdot (2LC_{w,2}^{\frac{1}{2}}\sigma_t\alpha_t) \cdot (2C_{w,2})^{\frac{1}{2}} + 4BL \cdot (2LC_{w,2}^{\frac{1}{2}}\sigma_t\alpha_t + 2LC_{w,2}^{\frac{1}{2}}\sigma_t\beta_t) \cdot (2C_{w,2})^{\frac{1}{2}} \\
&= 4\sqrt{2}L^2 C_{w,2}\sigma_t(\alpha_t + 2B\alpha_t + 2B\beta_t)
\end{aligned}$$

$$\leq 8\sqrt{2}L^2 C_{w,2}\sigma_t\alpha_t(1+B+B\beta_0\alpha_0^{-1}), \tag{A.25}$$

for all t large enough, where the last step holds because $\beta_t\alpha_t^{-1} = \beta_0/\alpha_0 t^{-(\delta-\kappa)} \leq \beta_0\alpha_0^{-1}$ since $\delta \geq \kappa$.

To bound $\mathbb{E}[|\Psi_{1,\sigma_t}||W_0 = w]$, note that

$$\begin{aligned}
&\mathbb{E}[\Psi_{1,\sigma_t}|W_0 = w] \\
&= \mathbb{E}[(\lambda_{t-\sigma_t} - \lambda^*)^\intercal \mathbb{E}[(G(W_t, \lambda_{t-\sigma_t}, \xi_{t-\sigma_t}) - G(W_t, \lambda_{t-\sigma_t}, \xi^*))|W_{t-\sigma_t}]|W_0 = w].
\end{aligned} \tag{A.26}$$

Moreover, note that

$$\begin{aligned}
&\mathbb{E}[(G(W_t, \lambda_{t-\sigma_t}, \xi_{t-\sigma_t}) - G(W_t, \lambda_{t-\sigma_t}, \xi^*))|W_{t-\sigma_t}] \\
&= \mathbb{E}_\nu[(G(W_t, \lambda_{t-\sigma_t}, \xi_{t-\sigma_t}) - G(W_t, \lambda_{t-\sigma_t}, \xi^*))] \\
&\quad + (\mathbb{E}[G(W_t, \lambda_{t-\sigma_t}, \xi_{t-\sigma_t})|W_{t-\sigma_t}] - \mathbb{E}_\nu[G(W_t, \lambda_{t-\sigma_t}, \xi_{t-\sigma_t})]) \\
&\quad - (\mathbb{E}[G(W_t, \lambda_{t-\sigma_t}, \xi^*)|W_{t-\sigma_t}] - \mathbb{E}_\nu[G(W_t, \lambda_{t-\sigma_t}, \xi^*)]).
\end{aligned} \tag{A.27}$$

For the first term of the summation (A.27), the condition (20) in Assumption 2 implies that

$$\|\mathbb{E}_\nu[(G(W_t, \lambda_{t-\sigma_t}, \xi_{t-\sigma_t}) - G(W_t, \lambda_{t-\sigma_t}, \xi^*))]\| \leq L\|\xi_{t-\sigma} - \xi^*\|. \tag{A.28}$$

For the second and third terms, Lemma SM.3 implies that

$$\|\mathbb{E}[G(W_t, \lambda_{t-\sigma_t}, \xi_{t-\sigma_t})|W_{t-\sigma_t}] - \mathbb{E}_\nu[G(W_t, \lambda_{t-\sigma_t}, \xi_{t-\sigma_t})]\| \leq \alpha_t^2(1 + \|W_{t-\sigma_t}\|), \tag{A.29}$$

and

$$|\mathbb{E}[G(W_t, \lambda_{t-\sigma_t}, \xi^*)|W_{t-\sigma_t}] - \mathbb{E}_\nu[G(W_t, \lambda_{t-\sigma_t}, \xi^*)]\| \leq \alpha_t^2(1 + \|W_{t-\sigma_t}\|). \tag{A.30}$$

Combining (A.26), (A.27), (A.28), (A.29), and (A.30) yields

$$\begin{aligned}
\mathbb{E}[|\Psi_{1,\sigma_t}|] &\leq L\mathbb{E}[\|\lambda_{t-\sigma_t} - \lambda^*\|\|\xi_{t-\sigma} - \xi^*\||W_0 = w] + 2\alpha_t^2\mathbb{E}[\|\lambda_{t-\sigma_t} - \lambda^*\|(1 + \|W_{t-\sigma_t}\|)|W_0 = w] \\
&\leq L\mathbb{E}[\|\lambda_{t-\sigma_t} - \lambda^*\|\|\xi_{t-\sigma} - \xi^*\||W_0 = w] + 4BC_{w,1}\alpha_t^2.
\end{aligned} \tag{A.31}$$

By the triangular inequality,

$$\begin{aligned}
&\mathbb{E}[\|\lambda_{t-\sigma_t} - \lambda^*\|\|\xi_{t-\sigma_t} - \xi^*\|] \\
&\leq \mathbb{E}[\|\lambda_t - \lambda^*\|\|\xi_t - \xi^*\|] + \mathbb{E}[\|\lambda_t - \lambda_{t-\sigma_t}\|\|\xi_t - \xi^*\|] + \mathbb{E}[\|\lambda_{t-\sigma_t} - \lambda^*\|\|\xi_t - \xi_{t-\sigma_t}\|] \\
&\leq \mathbb{E}[\|\lambda_t - \lambda^*\|^2]^{\frac{1}{2}}\,\mathbb{E}[\|\xi_t - \xi^*\|^2]^{\frac{1}{2}} + 2B\,\mathbb{E}[\|\lambda_t - \lambda_{t-\sigma_t}\|] + 2B\,\mathbb{E}[\|\xi_t - \xi_{t-\sigma_t}\|] \\
&\leq \frac{\zeta}{2L}f_t + 2BC_{w,1}(\alpha_{t-\sigma_t,t} + \beta_{t-\sigma_t,t}) \\
&\leq \frac{\zeta}{2L}f_t + 8BC_{w,1}\sigma_t\alpha_t,
\end{aligned} \tag{A.32}$$

for all t large enough, where f_t is defined in (A.14). Here, the second inequality follows from the Cauchy-Schwarz inequality and the fact that $\max\{\|\xi_t\|, \|\xi^*\|, \|\lambda_{t-\sigma_t}\|, \|\lambda^*\|\} \le B$; the third inequality holds because of (A.11) and (A.13) in Lemma SM.4; the last step holds because $\beta_{t-\sigma_t,t} \le \alpha_{t-\sigma_t,t} \le 2\sigma_t\alpha_t$ for all t large enough by the definitions of α_t, β_t, and σ_t.

Plugging (A.32) into (A.31) yields that for all t large enough,

$$\mathbb{E}[|\Psi_{1,\sigma_t}|] \le \frac{\zeta}{2} f_t + 8BLC_{w,1}\sigma_t\alpha_t + 4BC_{w,1}\alpha_t^2. \tag{A.33}$$

The inequality (A.19) is then established by combining (A.20), (A.25), and (A.33)

Step 2: Bounding Ψ_2. First, we decompose Ψ_2 into two parts:

$$\begin{aligned}\Psi_2 &= (\lambda_t - \lambda^*)^\intercal[G(W_t, \lambda_t, \xi^*) - \bar{G}(\lambda_t, \xi^*)] \\ &= \underbrace{(\lambda_t - \lambda_{t-\sigma_t})^\intercal[G(W_t, \lambda_t, \xi^*) - \bar{G}(\lambda_t, \xi^*)]}_{\Psi_{2,1}} + \underbrace{(\lambda_{t-\sigma_t} - \lambda^*)^\intercal[G(W_t, \lambda_t, \xi^*) - \bar{G}(\lambda_t, \xi^*)]}_{\Psi_{2,2}}.\end{aligned} \tag{A.34}$$

To bound $\Psi_{2,1}$, we use the Cauchy-Schwarz inequality,

$$|\Psi_{2,1}| \le \|\lambda_t - \lambda_{t-\sigma_t}\| \|G(W_t, \lambda_t, \xi^*) - \bar{G}(\lambda_t, \xi^*)\| \le 2L(1 + \|W_t\|)\|\lambda_t - \lambda_{t-\sigma_t}\|, \tag{A.35}$$

where the second inequality follows from the condition (22) in Assumption 2.

To bound $\Psi_{2,2}$, consider the decomposition

$$\begin{aligned}\Psi_{2,2} = &\underbrace{\Psi_{2,2} - (\lambda_{t-\sigma_t} - \lambda^*)^\intercal[G(W_t, \lambda_{t-\sigma_t}, \xi^*) - \bar{G}(\lambda_{t-\sigma_t}, \xi^*)]}_{\Psi_{2,2,1}} \\ &+ \underbrace{(\lambda_{t-\sigma_t} - \lambda^*)^\intercal[G(W_t, \lambda_{t-\sigma_t}, \xi^*) - \bar{G}(\lambda_{t-\sigma_t}, \xi^*)]}_{\Psi_{2,2,2}}.\end{aligned}$$

For $\Psi_{2,2,1}$, we have that,

$$\begin{aligned}|\Psi_{2,2,1}| &= |(\lambda_{t-\sigma_t} - \lambda^*)^\intercal[G(W_t, \lambda_t, \xi^*) - G(W_t, \lambda_{t-\sigma_t}, \xi^*) + \bar{G}(\lambda_{t-\sigma_t}, \xi^*) - \bar{G}(\lambda_t, \xi^*)]| \\ &\le \|\lambda_{t-\sigma_t} - \lambda^*\| \big(\|G(W_t, \lambda_t, \xi^*) - G(W_t, \lambda_{t-\sigma_t}, \xi^*)\| + \|\bar{G}(\lambda_t, \xi^*) - \bar{G}(\lambda_{t-\sigma_t}, \xi^*)\|\big) \\ &\le 2L\|\lambda_{t-\sigma_t} - \lambda^*\| \|\lambda_t - \lambda_{t-\sigma_t}\|(1 + \|W_t\|) \\ &\le 4BL\|\lambda_t - \lambda_{t-\sigma_t}\|(1 + \|W_t\|),\end{aligned} \tag{A.36}$$

where the second inequality follows from the condition (22) in Assumption 2.

For $\Psi_{2,2,2}$, we will derive an upper bound on $\mathbb{E}[\Psi_{2,2,2}|W_{t-\sigma_t}]$. Specifically,

$$\begin{aligned}\mathbb{E}[|\Psi_{2,2,2}||W_{t-\sigma_t}] &= \big|(\lambda_{t-\sigma_t} - \lambda^*)^\intercal\big(\mathbb{E}[G(W_t, \lambda_{t-\sigma_t}, \xi^*)|W_{t-\sigma_t}] - \bar{G}(\lambda_{t-\sigma_t}, \xi^*)\big)\big| \\ &\le \|\lambda_{t-\sigma_t} - \lambda^*\| \|\mathbb{E}[G(W_t, \lambda_{t-\sigma_t}, \xi^*)|W_{t-\sigma_t}] - \bar{G}(\lambda_{t-\sigma_t}, \xi^*)\| \\ &\le \alpha_t^2\|\lambda_{t-\sigma_t} - \lambda^*\|(1 + \|W_{t-\sigma_t}\|)\end{aligned}$$

$$\leq 2B\alpha_t^2(1+\|W_{t-\sigma_t}\|),$$

for all t large enough, where the second inequality follows from (A.8) in Lemma SM.3. Therefore,

$$\mathbb{E}[|\Psi_{2,2,2}||W_0=w] \leq 2B\alpha_t^2\mathbb{E}[(1+\|W_{t-\sigma_t}\|)|W_0=w] \leq 2BC_{w,1}\alpha_t^2, \tag{A.37}$$

for all t large enough. It follows from (A.34), (A.35), (A.36), and (A.37) that

$$\begin{aligned}
&\mathbb{E}[|\Psi_2||W_0=w] \\
&\leq \mathbb{E}[|\Psi_{2,1}||W_0=w] + \mathbb{E}[|\Psi_{2,2,1}||W_0=w] + \mathbb{E}[|\Psi_{2,2,2}||W_0=w] \\
&\leq 2L\mathbb{E}[(1+\|W_t\|)\|\lambda_t-\lambda_{t-\sigma_t}\||W_0=w] + 4BL\mathbb{E}[(1+\|W_t\|)\|\lambda_t-\lambda_{t-\sigma_t}\||W_0=w] + 2BC_{w,1}\alpha_t^2 \\
&\leq (4BL+2L)\mathbb{E}[(1+\|W_t\|)\|\lambda_t-\lambda_{t-\sigma_t}\||W_0=w] + 2BC_{w,1}\alpha_t^2.
\end{aligned}$$

for all t large enough. Further, note that by the Cauchy-Schwarz inequality,

$$\begin{aligned}
\mathbb{E}[(1+\|W_t\|)\|\lambda_t-\lambda_{t-\sigma_t}\||W_0=w] &\leq \mathbb{E}[(1+\|W_t\|)^2|W_0=w]^{\frac{1}{2}}\mathbb{E}[\|\lambda_t-\lambda_{t-\sigma_t}\|^2|W_0=w]^{\frac{1}{2}} \\
&\leq 2LC_{w,2}\sigma_t\alpha_t,
\end{aligned}$$

where the last inequality follows from (A.23). Therefore, for all t large enough,

$$\mathbb{E}[|\Psi_2||W_0=w] \leq (8BL^2+4L^2)C_{w,2}\sigma_t\alpha_t + 2BC_{w,1}\alpha_t^2. \tag{A.38}$$

Step 3: Bounding Ψ_3. By the definition of Ψ_3 and the condition (24) in Assumption 2,

$$\Psi_3 \leq -\zeta\|\lambda_t-\lambda^*\|^2. \tag{A.39}$$

Step 4: Bounding Ψ_4. By the condition (22) in Assumption 2,

$$\Psi_4 = \|G(W_t,\lambda_t,\xi^*)\|^2 \leq 2L^2(1+\|W_t\|^2).$$

Therefore,

$$\mathbb{E}[\Psi_4|W_0=w] \leq 2L^2\mathbb{E}[1+\|W_t\|^2|W_0=w] \leq 2L^2C_{w,2}. \tag{A.40}$$

Step 5: Putting All Together. Combining (A.18), (A.19), (A.38), (A.39), and (A.40),

$$\begin{aligned}
&\mathbb{E}[\|\lambda_{t+1}-\lambda^*\|^2|W_0=w] \\
&\leq \mathbb{E}[\|\lambda_t-\lambda^*\|^2|W_0=w] + \zeta\alpha_t f_t + 2\alpha_t(C_{\Psi_1}\sigma_t\alpha_t + 4BC_{w,1}\alpha_t^2) \\
&\quad + 2\alpha_t[(8BL^2+4L^2)C_{w,2}\sigma_t\alpha_t + 2BC_{w,1}\alpha_t^2] - 2\zeta\alpha_t\,\mathbb{E}[\|\lambda_t-\lambda^*\|^2|W_0=w] + 2L^2C_{w,2}\alpha_t^2 \\
&= (1-2\zeta\alpha_t)\,\mathbb{E}[\|\lambda_t-\lambda^*\|^2|W_0=w] + \zeta\alpha_t f_t
\end{aligned}$$

$$
\begin{aligned}
&\quad + 2\alpha_t^2[C_{\Psi_1}\sigma_t + (8BL^2+4L^2)C_{w,2}\sigma_t + 6BC_{w,1}\alpha_t + L^2C_{w,2}] \\
&\le (1-\zeta\alpha_t)f_t + 2\widetilde{C}\alpha_t^2\sigma_t \\
&\le (1-\zeta\alpha_t)f_t + 2\widetilde{C}\alpha_t^2(c_2\ln t + c_3),
\end{aligned}
\tag{A.41}
$$

for all t large enough, where $\widetilde{C} := C_{\Psi_1} + (8BL^2+4L^2)C_{w,2} + 6BC_{w,1} + L^2C_{w,2}$, c_2 and c_3 are the constants defined in (A.7). Here, the second inequality follows from the definition of f_t in (A.14), and the last from the definition of σ_t in Lemma SM.3. Then, by setting $k_0 := \zeta\alpha_0$, $\widetilde{k}_1 := 2\widetilde{C}c_2$, and $\widetilde{k}_2 := 2\widetilde{C}c_3$, the inequality (A.41) is reduced to (A.15).

At last, if $\{W_t : t \ge 0\}$ are i.i.d. with distribution ν, then $\mathbb{P}(W_t \in \cdot | W_0 = w) \equiv \nu(\cdot)$, so the parameter ρ in the condition (19) can be taken to be 0. Hence, the parameter $c_2 = \kappa/\ln(1/\rho) = 0$ by its definition (A.7), so $\widetilde{k}_1 = 2\widetilde{C}c_2 = 0$. □

LEMMA SM.6. *Fix $w \in \mathcal{W}$. Following the assumptions and notation of Lemma SM.5, if $\kappa \in (0,1)$ and*

$$
(L/\zeta)^2 g_t \le \frac{\widetilde{k}_1 \ln t + \widetilde{k}_2}{t^\kappa},
\tag{A.42}
$$

for all t large enough, then there exist positive constants k_3 and k_4 such that

$$
\mathbb{E}\big[\|\lambda_t - \lambda^*\|^2 | W_0 = w\big] \le k_3 \exp\left(-\frac{k_0 t^{1-\kappa}}{1-\kappa}\right) + k_4 \cdot \frac{\widetilde{k}_1 \ln t + \widetilde{k}_2}{t^\kappa},
\tag{A.43}
$$

for all t large enough.

Proof of Lemma SM.6. Fix an arbitrary integer T such that (A.42) holds for all $t \ge T$ and $T^{1-\kappa} \ge 2\kappa/k_0$. Define $k_3 := f_T \cdot \exp(k_0 T^{1-\kappa}/(1-\kappa))$ and $k_4 := \max(2/k_0, 1)$. We now prove by induction that the inequality (A.43) holds for all $t \ge T$.

For $t = T$, the right-hand-side (RHS) of (A.43) becomes

$$
k_3 \exp\left(-\frac{k_0 T^{1-\kappa}}{1-\kappa}\right) + k_4 \cdot \frac{\widetilde{k}_1 \ln T + \widetilde{k}_2}{T^\kappa} = f_T + k_4 \cdot \frac{\widetilde{k}_1 \ln T + \widetilde{k}_2}{T^\kappa} \ge f_T \ge \mathbb{E}\big[\|\lambda_N - \lambda^*\|^2 | W_0 = w\big],
$$

where the last step follows from the definition of f_T in (A.14). Thus, (A.43) holds for $t = T$.

Assume that (A.43) holds for $t = \ell$ for some $\ell \ge T$, that is,

$$
\mathbb{E}\big[\|\lambda_\ell - \lambda^*\|^2 | W_0 = w\big] \le k_3 \exp\left(-\frac{k_0 \ell^{1-\kappa}}{1-\kappa}\right) + k_4 \cdot \frac{\widetilde{k}_1 \ln \ell + \widetilde{k}_2}{\ell^\kappa}.
$$

Hence,

$$
\begin{aligned}
f_\ell &= \max\Big\{\mathbb{E}\big[\|\lambda_\ell - \lambda^*\|^2 | W_0 = w\big],\, (L/\zeta)^2 g_\ell\Big\} \\
&\le \max\Big\{\mathbb{E}\big[\|\lambda_\ell - \lambda^*\|^2 | W_0 = w\big],\, \frac{\widetilde{k}_1 \ln \ell + \widetilde{k}_2}{\ell^\kappa}\Big\}
\end{aligned}
$$

$$\leq k_3 \exp\left(-\frac{k_0 \ell^{1-\kappa}}{1-\kappa}\right) + k_4 \cdot \frac{\widetilde{k}_1 \ln \ell + \widetilde{k}_2}{\ell^\kappa}, \tag{A.44}$$

where the first inequality follows from (A.42), and the second from the fact that $k_4 \geq 1$.

We now consider the case $t = \ell + 1$. By (A.15) in Lemma SM.5 and (A.44),

$$\begin{aligned}
&\mathbb{E}\big[\|\lambda_{\ell+1} - \lambda^*\|^2 | W_0 = w\big] \\
&\leq \left(1 - \frac{k_0}{\ell^\kappa}\right) f_\ell + \frac{\widetilde{k}_1 \ln \ell + \widetilde{k}_2}{\ell^{2\kappa}} \\
&\leq \underbrace{\left(1 - \frac{k_0}{\ell^\kappa}\right) k_3 \exp\left(-\frac{k_0 \ell^{1-\kappa}}{1-\kappa}\right)}_{I_1} + \underbrace{\left(1 - \frac{k_0}{\ell^\kappa}\right) k_4 \cdot \frac{\widetilde{k}_1 \ln \ell + \widetilde{k}_2}{\ell^\kappa} + \frac{\widetilde{k}_1 \ln \ell + \widetilde{k}_2}{\ell^{2\kappa}}}_{I_2}.
\end{aligned} \tag{A.45}$$

Next, we analyze I_1 and I_2 separately. For I_1, note that

$$1 - \frac{k_0}{\ell^\kappa} \leq \exp\left(-\frac{k_0}{\ell^\kappa}\right) \leq \exp\left(-\int_\ell^{\ell+1} \frac{k_0}{u^\kappa}\, \mathrm{d}u\right) = \exp\left[-\frac{k_0}{1-\kappa}\left((\ell+1)^{1-\kappa} - \ell^{1-\kappa}\right)\right],$$

where the first inequality follows from the basic relation that $1 + u \leq e^u$ for all $u \in \mathbb{R}$. Thus,

$$\begin{aligned}
I_1 &\leq \exp\left[-\frac{k_0}{1-\kappa}\left((\ell+1)^{1-\kappa} - \ell^{1-\kappa}\right)\right] \cdot k_3 \exp\left(-\frac{k_0 \ell^{1-\kappa}}{1-\kappa}\right) \\
&= k_3 \exp\left(-\frac{k_0 (\ell+1)^{1-\kappa}}{1-\kappa}\right).
\end{aligned} \tag{A.46}$$

For I_2, note that

$$\begin{aligned}
I_2 - k_4 \cdot \frac{\widetilde{k}_1 \ln(\ell+1) + \widetilde{k}_2}{(\ell+1)^\kappa} &\leq I_2 - k_4 \cdot \frac{\widetilde{k}_1 \ln \ell + \widetilde{k}_2}{(\ell+1)^\kappa} \\
&= (\widetilde{k}_1 \ln \ell + \widetilde{k}_2)\left[\left(1 - \frac{k_0}{\ell^\kappa}\right)\frac{k_4}{\ell^\kappa} + \frac{1}{\ell^{2\kappa}} - \frac{k_4}{(\ell+1)^\kappa}\right] \\
&\leq \frac{k_4(\widetilde{k}_1 \ln \ell + \widetilde{k}_2)}{\ell^{2\kappa}}\left[\left(\frac{1}{k_4} - k_0\right) + \ell^\kappa\left(1 - \frac{\ell^\kappa}{(\ell+1)^\kappa}\right)\right] \\
&\leq \frac{k_4(\widetilde{k}_1 \ln \ell + \widetilde{k}_2)}{\ell^{2\kappa}}\left[-\frac{k_0}{2} + \ell^\kappa\left(1 - \frac{\ell^\kappa}{(\ell+1)^\kappa}\right)\right],
\end{aligned} \tag{A.47}$$

where the last inequality holds because $k_4 = \max(2/k_0, 1) \geq 2/k_0$. Furthermore, using the basic relations that $(1 + 1/u)^u \leq e$ and $1 + u \leq e^u$ for all $u \in \mathbb{R}$, we have

$$\ell^\kappa\left(1 - \frac{\ell^\kappa}{(\ell+1)^\kappa}\right) = \ell^\kappa\left[1 - \left(\frac{1}{(1+1/\ell)^\ell}\right)^{\kappa/\ell}\right] \leq \ell^\kappa\left(1 - e^{-\kappa/\ell}\right) \leq \frac{\kappa}{\ell^{1-\kappa}} \leq \frac{\kappa}{T^{1-\kappa}} \leq \frac{k_0}{2},$$

where the last inequality follows from the definition of T. Thus, the RHS of (A.47) is no greater than 0. Hence,

$$I_2 \leq k_4 \cdot \frac{\widetilde{k}_1 \ln(\ell+1) + \widetilde{k}_2}{(\ell+1)^\kappa}. \tag{A.48}$$

At last, combining the inequalities (A.45), (A.46), and (A.48) yields

$$\mathbb{E}\big[\|\lambda_{\ell+1}-\lambda^*\|^2|W_0=w\big]\le k_3\exp\left(-\frac{k_0(\ell+1)^{1-\kappa}}{1-\kappa}\right)+k_4\cdot\frac{\widetilde{k}_1\ln(\ell+1)+\widetilde{k}_2}{(\ell+1)^{\kappa}},$$

that is, (A.43) holds for $t=\ell+1$, completing the induction argument. Therefore, (A.43) holds for all $t\ge T$. $\square$

A.2. Proof of Part (ii) of Proposition 1

Define a mapping $F(W,\xi,\varsigma):=(H(w,\xi),\varsigma)^\intercal$ for all $w\in\mathcal{W}$, $\xi\in\mathbb{R}^{\mathsf{d}}$, and $\varsigma\in\mathbb{R}$. Let $\bar{F}(\xi,\varsigma):=\mathbb{E}_\nu[F(W,\xi,\varsigma)]$. Then, (ξ^*,ς^*) with $\varsigma^*=0$ is a solution to the equation $\bar{F}(\xi,\varsigma)=0$. We set $\varsigma_t\equiv 0$ for all $t\ge 0$. Then, it is straightforward to verify that F satisfies the assumptions of Lemma SM.5. Hence,

$$\mathbb{E}\big[\|\xi_{t+1}-\xi^*\|^2|W_0=w\big]\le\left(1-\frac{k_0}{t^\delta}\right)\mathbb{E}\big[\|\xi_t-\xi^*\|^2|W_0=w\big]+\frac{\widetilde{k}_1\ln t+\widetilde{k}_2}{t^{2\delta}}, \tag{A.49}$$

for all t large enough and $w\in\mathcal{W}$, where $k_0:=\psi\beta_0$, and $\widetilde{k}_1\ge 0$ and $\widetilde{k}_2>0$ are some constants. Moreover, $\widetilde{k}_1=0$ if $\{W_t:t\ge 0\}$ are i.i.d.

Case 1: $\delta\in(0,1)$**.** We may directly invoke Lemma SM.6 by setting $g_t\equiv 0$ there, and conclude that there exists some positive constants k_3 and k_4 such that

$$\mathbb{E}\big[\|\xi_t-\xi^*\|^2|W_0=w\big]\le k_3\exp\left(-\frac{k_0t^{1-\delta}}{1-\delta}\right)+k_4\cdot\frac{\widetilde{k}_1\ln t+\widetilde{k}_2}{t^\delta},$$

for all t large enough and $w\in\mathcal{W}$.

Note that $k_3\exp\left(-k_0t^{1-\delta}/(1-\delta)\right)\le t^{-\delta}$ for all t large enough. Hence, by setting $k_1=\widetilde{k}_1k_4$ and $k_2=1+\widetilde{k}_2k_4$, we have

$$\mathbb{E}\big[\|\xi_t-\xi^*\|^2|W_0=w\big]\le\frac{k_1\ln t+k_2}{t^\delta},$$

for all t large enough and $w\in\mathcal{W}$.

Case 2: $\delta=1$**.** Let $h_t=(\widetilde{k}_1\ln t+\widetilde{k}_2)/t^2$. Applying (A.49) iteratively yields that, for all $t\ge T$,

$$\begin{aligned}
&\mathbb{E}\big[\|\xi_t-\xi^*\|^2|W_0=w\big]\\
\le&\left(1-\frac{k_0}{t-1}\right)\mathbb{E}\big[\|\xi_{t-1}-\xi^*\|^2|W_0=w\big]+h_{t-1}\\
\le&\left(1-\frac{k_0}{t-1}\right)\left[\left(1-\frac{k_0}{t-2}\right)\mathbb{E}\big[\|\xi_{t-2}-\xi^*\|^2|W_0=w\big]+h_{t-2}\right]+h_{t-1}\\
=&\left(1-\frac{k_0}{t-1}\right)\left(1-\frac{k_0}{t-2}\right)\mathbb{E}\big[\|\xi_{t-2}-\xi^*\|^2|W_0=w\big]+\left(1-\frac{k_0}{t-1}\right)h_{t-2}+h_{t-1}\\
\le&\cdots\\
\le&\prod_{\ell=T}^{t-1}\left(1-\frac{k_0}{\ell}\right)\mathbb{E}\big[\|\xi_T-\xi^*\|^2|W_0=w\big]+\sum_{n=T}^{t-1}h_n\prod_{\ell=n+1}^{t-1}\left(1-\frac{k_0}{\ell}\right)
\end{aligned}$$

$$\leq \prod_{\ell=T}^{t-1}\left(1-\frac{k_0}{\ell}\right)\mathbb{E}\left[\|\xi_T-\xi^*\|^2|W_0=w\right]+(\widetilde{k}_1\ln t+\widetilde{k}_2)\sum_{n=T}^{t-1}\frac{1}{n^2}\prod_{\ell=n+1}^{t-1}\left(1-\frac{k_0}{\ell}\right). \tag{A.50}$$

Note that

$$\prod_{\ell=T}^{t-1}\left(1-\frac{k_0}{\ell}\right)\leq \exp\left(-k_0\sum_{\ell=T}^{t-1}\frac{1}{\ell}\right)\leq \exp\left(-k_0\int_T^t\frac{\mathrm{d}u}{u}\right)=\frac{T^{k_0}}{t^{k_0}}. \tag{A.51}$$

Hence,

$$\sum_{n=T}^{t-1}\frac{1}{n^2}\prod_{\ell=n+1}^{t-1}\left(1-\frac{k_0}{\ell}\right)\leq \sum_{n=T}^{t-1}\frac{1}{n^2}\frac{(n+1)^{k_0}}{t^{k_0}}\leq\left(\frac{2}{t}\right)^{k_0}\sum_{n=T}^{t-1}\frac{1}{n^{2-k_0}}\leq\left(\frac{2}{t}\right)^{k_0}\int_{T-1}^{t-1}\frac{\mathrm{d}u}{u^{2-k_0}}, \tag{A.52}$$

where the first inequality follows from (A.51). Obviously, the integral on the RHS of (A.52) depends on the value of k_0, which we discuss as follows.

- If $k_0\in(0,1)$, then for all $t\geq T$,

$$\int_{T-1}^{t-1}\frac{\mathrm{d}u}{u^{2-k_0}}=\frac{1}{1-k_0}\left(\frac{1}{T^{1-k_0}}-\frac{1}{t^{1-k_0}}\right)\leq\frac{1}{1-k_0};$$

- If $k_0=1$, then for all $t\geq T$

$$\int_{T-1}^{t-1}\frac{\mathrm{d}u}{u}=\ln(t-1)-\ln(T-1)\leq\ln t;$$

- If $k_0>1$, then for all $t\geq T$

$$\int_{T-1}^{t-1}\frac{\mathrm{d}u}{u^{2-k_0}}=\frac{1}{k_0-1}\left(t^{k_0-1}-T^{k_0-1}\right)\leq\frac{t^{k_0-1}}{k_0-1}.$$

It then follows from (A.50), (A.51), and (A.52) that if $\delta=1$, then for all $t\geq T$:

$$\mathbb{E}\left[\|\xi_t-\xi^*\|^2|W_0=w\right]\leq\begin{cases}\dfrac{T^{k_0}}{t^{k_0}}+\dfrac{2^{k_0}}{1-k_0}\cdot\dfrac{\widetilde{k}_1\ln t+\widetilde{k}_2}{t^{k_0}}, & \text{if } k_0\in(0,1),\\ \dfrac{T}{t}+\dfrac{2(\widetilde{k}_1\ln t+\widetilde{k}_2)\ln t}{t}, & \text{if } k_0=1,\\ \dfrac{T^{k_0}}{t^{k_0}}+\dfrac{2^{k_0}}{k_0-1}\cdot\dfrac{\widetilde{k}_1\ln t+\widetilde{k}_2}{t}, & \text{if } k_0>1.\end{cases} \tag{A.53}$$

Since it is assumed in Proposition 1 that $k_0=\psi\beta_0>1$ in the case of $\delta=1$, note that $(T/t)^{k_0}\leq t^{-1}$ for all t large enough. Hence, by setting $k_1=2^{k_0}\widetilde{k}_1/(k_0-1)$ and $k_2=1+2^{k_0}\widetilde{k}_2/(k_0-1)$, we have

$$\mathbb{E}\left[\|\xi_t-\xi^*\|^2|W_0=w\right]\leq\frac{k_1\ln t+k_2}{t^{\delta}},$$

for all t large enough and $w\in\mathcal{W}$. $\square$

A.3. Proof of Part (i) of Proposition 1

Let f_t and g_t be defined as (A.14). Applying Lemma SM.5,

$$\mathbb{E}\big[\|\lambda_{t+1}-\lambda^*\|^2|W_0=w\big] \le \left(1-\frac{k_0}{t^\kappa}\right) f_t + \frac{\widetilde{k}_1 \ln t + \widetilde{k}_2}{t^{2\kappa}},$$

for all t large enough and $w \in \mathcal{W}$, where $k_0 := \zeta\alpha_0$, and $\widetilde{k}_1 \ge 0$ and $\widetilde{k}_2$ are some constants. Moreover, $\widetilde{k}_1 = 0$ if $\{W_t : t \ge 0\}$ are i.i.d.

Since $0<\kappa\le\delta\le 1$ and $\psi\beta_0 > 1$ in the case of $\delta = 1$, it follows from part (ii) of Proposition 1 that

$$g_t = \mathbb{E}\big[\|\xi_t-\xi^*\|^2|W_0=w\big] \le \frac{k_1 \ln t + k_2}{t^\delta} \le \frac{k_1 \ln t + k_2}{t^\kappa}$$

for all t large enough, for some constants $k_1 \ge 0$ and $k_2 > 0$. The proof is then concluded by applying Lemma SM.6. □

B. Proof of Theorem 1

Throughout this section, we suppose that Assumption 1 holds with $d_U = 2$ and Assumption 2 holds; we also assume that $1/2 < \kappa \le \delta < 1$. Moreover, we abbreviate $\mathbb{P}(\cdot|W_0 = w)$ as $\mathbb{P}(\cdot)$ for notational simplicity, unless explicitly stated otherwise.

The following lemma is needed for proving Theorem 1.

Lemma SM.7. *For any $l_1 \ge l_2 \ge m$, we have:*

$$|\,\mathbb{E}[H(W_{l_1},\xi^*)^\intercal H(W_{l_2},\xi^*)|W_m]| \le 2L^2 c_1 \rho^{l_1-l_2}(c_1(1+||W_m||^2)+K_{d_U}).$$

Proof. It is easy to see that

$$\begin{aligned}
&|\,\mathbb{E}[H(W_{l_1},\xi^*)^\intercal H(W_{l_2},\xi^*)|W_m]| \\
&= |\,\mathbb{E}[\mathbb{E}[H(W_{l_1},\xi^*)^\intercal|W_{l_2}]H(W_{l_2},\xi^*)|W_m]| \\
&\le |\,\mathbb{E}[\mathbb{E}_\nu[H(W_{l_1},\xi^*)^\intercal]H(W_{l_2},\xi^*)|W_m]| \\
&\quad + \mathbb{E}\left[\int_{\mathcal{W}} |\,\mathbb{P}(w_{l_1}|W_{l_2}) - \nu(w_{l_1})|\|H(w_{l_1},\xi^*)\|\|H(W_{l_2},\xi^*)\|\Big|W_m\right] \mathrm{d}w_{l_1} \\
&\le 0 + 2L^2c_1\rho^{l_1-l_2}\,\mathbb{E}[(1+||W_{l_2}||^2)|W_m] \\
&\le 2L^2c_1\rho^{l_1-l_2}(c_1(1+||W_m||^2)+K_{d_U}),
\end{aligned}$$

where the second inequality follows from Lemma SM.1 and Assumption 2, and the last inequality follows from Lemma SM.1. □

We focus on proving statement (ii) of Theorem 1. The proof of statement (i) is similar.

Cut-off Strategy. For any $\delta_c \in [0, \frac{1}{2}(2\delta-1))$, consider an increasing sequence $\{l_t = Ct^{-\delta_c}: t = 1,2,\dots\}$, and an increasing sequence of positive integers $\{I_j = \lfloor j^q \rfloor : j = 1,2,\dots\}$ for some constant $q > 1$.

The proof is divided into two parts. In the first part, we establish the uniform convergence of the subsequence $\{\xi_m : m = I_j, j = 1, 2, \ldots\}$ to ξ^*. In the second part, we use a partitioning argument to show that the entire sequence $\{\xi_m : m = 1, 2, \ldots\}$ converges to ξ^* uniformly.

Uniform Convergence in Probability on $\{\xi_{I_j} : j = 1, 2, \ldots\}$. By Proposition 1 and Markov's inequality, we know that:

$$\mathbb{P}(\|\xi_{I_j} - \xi^*\| > l_{I_j}) \leq \frac{(k_1 \ln t + k_2)}{l_{I_j}^2 I_j^\delta},$$

for any $I_j \geq T$, for some constant $T > 0$. Therefore, plugging in $I_j = \lfloor j^q \rfloor$, we have:

$$\mathbb{P}(\|\xi_{I_j} - \xi^*\| \leq l_{I_j}, \forall j = s, s+1, \ldots) \geq 1 - \sum_{j=s}^{\infty} \frac{(k_1 \ln I_j + k_2)}{l_{I_j}^2 I_j^\delta}.$$

For $s = \lfloor T^{\frac{1}{q}} \rfloor + 1$ being large enough and $q > \frac{1}{\delta - 2\delta_c}$,

$$\begin{aligned}\sum_{j=s}^{\infty} \frac{(k_1 \ln I_j + k_2)}{l_{I_j}^2 I_j^{-\delta}} &\leq \sum_{j=s}^{\infty} \frac{(k_1 q \ln j + k_2)}{C^2 j^{-q(\delta - 2\delta_c)}} \\ &\leq \sum_{j \geq s} j^{-q(\delta - 2\delta_c)} \frac{k_1 q \ln j + k_2}{C^2} \\ &\leq s^{1 - q(\delta - 2\delta_c)} \frac{k_1 q \ln s + k_2}{|1 - q(\delta - 2\delta_c)| C^2}.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{P}(\|\xi_{I_j} - \xi^*\| \leq l_{I_j}, \forall I_j \geq T) &\geq 1 - T^{-\frac{(q(\delta - 2\delta_c) - 1)}{q}} \frac{k_1 \ln T + k_2}{|1 - q(\delta - 2\delta_c)| C^2} \\ &\geq 1 - C_{S,1} T^{-\frac{(q(\delta - 2\delta_c) - 1)}{q}} \ln T, \qquad \text{(B.1)}\end{aligned}$$

where

$$C_{S,1} := \frac{k_1 + k_2}{|1 - q(\delta - 2\delta_c)| C^2}.$$

Since δ_c is a positive constant, for T being large enough, with probability no less than $1 - C_{S,1} T^{-\frac{(q(\delta - 2\delta_c) - 1)}{q}} \ln T$, each element of the sequence $\{\xi_{I_j} : I_j \geq T\}$ is in the neighborhood of ξ^*. In particular, we have $\|\xi_{I_j} - \xi^*\| \leq B$ for all $I_j \geq T$, and thus the projection Π_B is the identity operator in all iterations $I_j \geq T$.

We claim that we can assume that the projection operator Π_B equals the identity operator for iteration number $m \geq T$. The argument is as follows. For every interval $[I_j, I_{j+1})$ such that $I_j \geq T$, consider to construct a new sequence such that $\widetilde{\xi}_{m+1} = \widetilde{\xi}_m + \beta_m H(W_m, \widetilde{\xi}_m)$, for $m = I_j, ..., I_{j+1} - 2$ while setting $\widetilde{\xi}_{I_j} = \xi_{I_j}$. Then, the rest of the proof in this section holds for $\widetilde{\xi}_{I_j}$. That is, there exists some constant C'_ξ such that

$$\sum_{j \geq s = T^{1/q}} \mathbb{P}(\max_{m \in [I_j, I_{j+1})} \|\widetilde{\xi}_m - \xi_{I_j}\| \leq C I_{j+1}^{-\delta_c}) \geq 1 - C'_\xi T^{-(2\delta - 2\delta_c - 1)} \ln T.$$

Given that the event

$$\Big\{\|\xi_{I_j}-\xi^*\|\leq CI_j^{-\delta_c}, I_j\geq T\Big\}\bigcap\Big\{\max_{m\in[I_j,I_{j+1})}\|\widetilde{\xi}_m-\xi_{I_j}\|\leq CI_{j+1}^{-\delta_c}, I_j\geq T\Big\}$$

holds, for T large enough, $\widetilde{\xi}_m$ must be in the interior of $\mathsf{Ball}(B)$ for all $m\geq T$. It implies that the path $\{\widetilde{\xi}_m, m\geq T\}$ must be the same as the path $\{\xi_m, m\geq T\}$; i.e., the projection operator Π_B equals the identity operator.

Therefore, it suffices to prove the rest of results by using the sequence $\{\widetilde{\xi}_m, m\geq T\}$ instead of $\{\xi_m, m\geq T\}$; i.e., the projection operator Π_B is replaced by the identity operator in the iteration process. For notational simplicity, we still use ξ_m instead of $\widetilde{\xi}_m$. This assumption allows $\xi_{m_2}=\xi_{m_1}+\sum_{m=m_2}^{m_1-1}\beta_m H(W_m,\xi_m)$, which facilitates the rest of the analysis in this section.

Chaining Method. We establish bounds on $\max_{I_j\leq m\leq I_{j+1}-1}\|\xi_m-\xi_{I_j}\|$.

It is easy to see that

$$I_j<I_{j+1}\leq 2I_j, \tag{B.2}$$

for all j large enough. Indeed, $\frac{I_{j+1}}{I_j}\to 1$ as $j\to\infty$. Moreover, note that

$$I_{j+1}-I_j\leq 2qj^{q-1} \tag{B.3}$$

for all j large enough. Notice that we can choose t large enough to guarantee that for all $j\geq T^{\frac{1}{q}}$, j is large enough to allow (B.2) and (B.3) to hold.

Define the partial sums

$$S_{j,m}:=\sum_{l=I_j}^{m-1}\beta_l H(W_l,\xi_l),$$
$$S^*_{j,m}:=\sum_{l=I_j}^{m-1}\beta_l H(W_l,\xi^*),$$
$$\widetilde{S}_{j,m}:=\sum_{l=I_j}^{m-1}\beta_l H(W_l,\xi_{l-\sigma}),$$

for $m=I_j,\dots,I_{j+1}-1$, and $\sigma=d\ln I_j$ and $d=\frac{3}{|\ln\rho|}$ is a constant.

By the triangular inequality,

$$\|S_{j,m}-S^*_{j,m}\|\leq\|S_{j,m}-\widetilde{S}_{j,m}\|+\|\widetilde{S}_{j,m}-S^*_{j,m}\|.$$

Therefore,

$$\begin{aligned}\max_{I_j\leq m<I_j}\|S_{j,m}-S^*_{j,m}\|&\leq\max_{I_j\leq m<I_{j+1}}\|S_{j,m}-\widetilde{S}_{j,m}\|+\max_{I_j\leq m<I_{j+1}}\|\widetilde{S}_{j,m}-S^*_{j,m}\|\\&\leq\sum_{I_j\leq m<I_{j+1}}\beta_m\|H(W_m,\xi_m)-H(W_m,\xi_{m-\sigma})\|\end{aligned}$$

$$
\begin{aligned}
&+ \sum_{I_j \le m < I_{j+1}} \beta_m \|H(W_m,\xi^*) - H(W_m,\xi_{m-\sigma})\| \\
&\le \beta_{I_j} L \underbrace{\sum_{I_j \le m < I_{j+1}} (1+\|W_m\|)\|\xi_m - \xi_{m-\sigma}\|}_{\Psi_{S,1,j}} \\
&+ \beta_{I_j} L \underbrace{\sum_{I_j \le m < I_{j+1}} (1+\|W_m\|)\|\xi^* - \xi_{m-\sigma}\|}_{\Psi_{S,2,j}},
\end{aligned}
$$

where the last inequality follows from Assumption 2.

First, we bound $\Psi_{S,1,j}$. By Lemma SM.7,

$$
\begin{aligned}
\mathbb{E}[\Psi_{S,1,j}] &\le \sum_{I_j \le m < I_{j+1}} \mathbb{E}[(1+\|W_m\|) \sum_{l=m-\sigma}^{m-1} \beta_l \|H(W_l,\xi_l)\|] \\
&\le 2\sigma L (I_{j+1} - I_j)\beta_{I_j} C_{w,2}.
\end{aligned}
$$

If follows that, by Markov's inequality,

$$
\begin{aligned}
\mathbb{P}\left(\beta_{I_j} L \Psi_{S,j,1} \ge \frac{C}{3} I_{j+1}^{-\delta_c}\right) &\le \frac{6\sigma C_{w,2} L^2}{C} I_j^{-2}(I_{j+1}-I_j) I_{j+1}^{\delta_c} \\
&\le \frac{6(d\ln I_j) q C_{w,2} L^2}{C} j^{-q(2\delta-\delta_c-1)-1} \\
&\le \frac{6dq^2 C_{w,2} L^2}{C} j^{-q(2\delta-2\delta_c-1)-q\delta_c-1} \ln j.
\end{aligned} \tag{B.4}
$$

Second, we bound $\Psi_{S,2,j}$. Note that

$$
\Psi_{S,2,j} = \bar{\Psi}_{S,2,j} + (\Psi_{S,2,j} - \bar{\Psi}_{S,2,j}),
$$

where $\bar{\Psi}_{S,2,j} := \sum_{I_j \le m < I_{j+1}} \mathbb{E}_{W_m \sim \nu}[(1+\|W_m\|)]\|\xi^* - \xi_{m-\sigma}\|$.

It is easy to see that

$$
\begin{aligned}
&|\,\mathbb{E}[\Psi_{S,2,j} - \bar{\Psi}_{S,2,j}]| \\
&\le \sum_{I_j \le m < I_{j+1}} \mathbb{E}[(1+\|W_m\|)\|\xi_{m-\sigma} - \xi^*\|] - \mathbb{E}_{W_m\sim\nu}[(1+\|W_m\|)\|\xi_{m-\sigma} - \xi^*\|] \\
&= \sum_{I_j \le m < I_{j+1}} \mathbb{E}\Big[\big(\mathbb{E}[1+\|W_m\||W_{m-\sigma}] - \mathbb{E}_{W_m\sim\nu}[1+\|W_m\|]\big)\|\xi_{m-\sigma} - \xi^*\|\Big] \\
&\le c_1 \rho^\sigma\, \mathbb{E}[(1+\|W_{m-\sigma}\|)\|\xi_{m-\sigma} - \xi^*\|] \\
&\le 2c_1 B C_{w,1} I_j^{-3},
\end{aligned}
$$

where the second inequality follows from Lemma SM.1. Then, by Markov's inequality,

$$
\mathbb{P}\left(\beta_{I_j} L |\,\mathbb{E}[\Psi_{S,2,j} - \bar{\Psi}_{S,2,j}]| \ge \frac{C}{3} I_{j+1}^{-\delta_c}\right) \le \frac{6c_1 B C_{w,1} L}{C} I_j^{-3-\delta} I_{j+1}^{\delta_c}
$$

$$\leq \frac{6c_1 B C_{w,1} L}{C} j^{-q(2\delta-2\delta_c-1)-q(4-\delta+\delta_c)}. \tag{B.5}$$

Moreover, note that

$$\begin{aligned}\mathbb{E}[|\bar{\Psi}_{S,2,j}|^2] &\leq 2 \sum_{I_j \leq l_1 \leq l_2 < I_{j+1}} \mathbb{E}\Big[\big(\mathbb{E}_{W_{l_1}\sim\nu}[(1+\|W_{l_1}\|)]\|\xi^* - \xi_{l_1-\sigma}\|\big)\big(\mathbb{E}_{W_{l_2}\sim\nu}[(1+\|W_{l_2}\|)]\|\xi^* - \xi_{l_2-\sigma}\|\big)\Big] \\ &\leq 2 \sum_{I_j \leq l_1 \leq l_2 < I_{j+1}} K_{d_U}^2\, \mathbb{E}[\|\xi^* - \xi_{l_1-\sigma}\|\|\xi^* - \xi_{l_2-\sigma}\|].\end{aligned}$$

By Proposition 1 and the Cauchy-Schwarz inequality,

$$\begin{aligned}\mathbb{E}[\|\xi^* - \xi_{l_1-\sigma}\|)\|\xi^* - \xi_{l_2-\sigma}\|)] &\leq \mathbb{E}[\|\xi^* - \xi_{l_1-\sigma}\|^2]^{\frac{1}{2}}\, \mathbb{E}[\|\xi^* - \xi_{l_2-\sigma}\|^2]^{\frac{1}{2}} \\ &\leq I_{j-1}^{-\delta}(k_1 \ln I_{j-1} + k_2) \\ &\leq 2 I_j^{-\delta}(k_1 \ln I_j + k_2),\end{aligned}$$

for all j large enough such that $\min(l_1, l_2) - \sigma \geq I_j - d\ln I_j \geq I_{j-1}$. Therefore,

$$\mathbb{E}[|\bar{\Psi}_{S,2,j}|^2] \leq 4K_{d_U}^2 (I_{j+1} - I_j)^2 I_j^{-\delta}(k_1 \ln I_j + k_2).$$

By Markov's inequality,

$$\mathbb{P}\left(\beta_{I_j} L|\bar{\Psi}_{S,2,j}| \geq \frac{C}{3} I_{j+1}^{-\delta_c}\right) \leq \frac{36 q^2 L^2 K_{d_U}^2}{C^2} j^{-q(3\delta-2\delta_c-2)-2}(k_1 \ln I_j + k_2). \tag{B.6}$$

Let $s := T^{\frac{1}{q}}$, for all q such that $q(3\delta - 2\delta_c - 2) + 1 > 0$ and s large enough. Combining (B.4), (B.5), and (B.6), we have:

$$\begin{aligned}&\sum_{j\geq s} \mathbb{P}\left(\max_{I_j \leq m < I_{j+1}} \|S_{j,m} - S_{j,m}^*\| \leq C I_{j+1}^{-\delta_c}, j \geq s\right) \\ &\geq \sum_{j\geq s} \mathbb{P}\left(\beta_{I_j} L \Psi_{S,1,j} \leq \frac{C}{3} I_{j+1}^{-\delta_c}\right) + \mathbb{P}\left(\beta_{I_j} L|\bar{\Psi}_{S,2,j}| \leq \frac{C}{3} I_{j+1}^{-\delta_c}\right) + \mathbb{P}\left(\beta_{I_j} L|\Psi_{S,2,j} - \bar{\Psi}_{S,2,j}| \leq \frac{C}{3} I_{j+1}^{-\delta_c}\right) \\ &\geq 1 - \sum_{j\geq s} \frac{6dq^2 C_{w,2} L^2}{C} j^{-q(2\delta-2\delta_c-1)-q\delta_c-1} \ln j \\ &\quad - \sum_{j\geq s} \frac{6c_1 B C_{w,1} L}{C} j^{-q(2\delta-2\delta_c-1)-q(4-\delta+\delta_c)} \\ &\quad - \sum_{j\geq s} \frac{36 q^2 L^2 K_{d_U}^2}{C^2} j^{-q(3\delta-2\delta_c-2)-2}(k_1 \ln I_j + k_2).\end{aligned} \tag{B.7}$$

If we take $q \geq \frac{1}{1-\delta}$, then

$$-q(3\delta - 2\delta_c - 2) - 2 \geq -q(2\delta - 2\delta_c - 1) - q\delta_c - 1 > -q(2\delta - 2\delta_c - 1) - q(4 - \delta + \delta_c).$$

Consequently, it follows from (B.7) that

$$\sum_{j\geq s} \mathbb{P}\left(\max_{I_j \leq m < I_{j+1}} \|S_{j,m} - S_{j,m}^*\| \leq C I_{j+1}^{-\delta_c}, j \geq s\right)$$

$$\geq 1 - 2\sum_{j\geq s}\left(\frac{6dq^2C_{w,2}L^2}{C} + \frac{36q^2L^2K_{d_U}^2}{C^2}\right)j^{-q(3\delta-2\delta_c-2)-2}(k_1 q\ln j + k_2)$$
$$\geq 1 - \frac{2}{q|q(3\delta-2\delta_c-2)+1|}\left(\frac{6dq^2C_{w,2}L^2}{C} + \frac{36q^2L^2K_{d_U}^2}{C^2}\right)\max(k_1,k_2)T^{-(3\delta-2\delta_c-2)-1/q}\ln T$$
$$= 1 - C_{S,2}T^{-(3\delta-2\delta_c-2)-1/q}\ln T, \tag{B.8}$$

where

$$C_{S,2} := \frac{2}{|q(3\delta-2\delta_c-2)+1|}\left(\frac{6dqC_{w,2}L^2}{C} + \frac{36qL^2K_{d_U}^2}{C^2}\max(k_1,k_2)\right).$$

Now we establish bound on $\max_{I_j\leq m\leq I_{j+1}-1}\|S^*_{j,m}\|$. We first establish in Lemma SM.8 bounds on the median of $S^*_{j,m} - S^*_{j,I_{j+1}-1}$ conditional on W_m for $m\in[I_j, I_{j+1})$.

LEMMA SM.8. *(i) For any $I_{j+1}-1\geq m_1\geq m_2\geq I_j$,*

$$\mathbb{E}[\|S^*_{j,m_1} - S^*_{j,m_2}\|^2|W_{m_2}] \leq \frac{4L^2c_1\beta_{I_j}^2}{1-\rho}\left((m_1-m_2)K_{d_U} + \frac{2c_1}{1-\rho}(1+\|W_{m_2}\|^2)\right).$$

(ii) For any $I_j\leq m < I_{j+1}$, define $\mathsf{med}(S^*_{j,m} - S_{j,I^*_{j+1}-1}|W_m)$ as the median of $S^*_{j,m} - S^*_{j,I_{j+1}-1}$ conditional on W_m. Then,

$$\sum_{j\geq s}\mathbb{P}\left(\max_{I_j\leq m<I_{j+1}}\|\mathsf{med}(S^*_{j,m} - S^*_{j,I_{j+1}-1}|W_m)\| \geq \frac{C}{2}I_{j+1}^{-\delta_c}\right)$$
$$\leq \frac{64c_1^2C_{w,2}(1+4\mathsf{d}_\xi L)^2}{C^2(1-\rho)^2(2\delta-2\delta_c-1)}T^{-(2\delta-2\delta_c-1)}.$$

Proof of Lemma SM.8 (i). It is easy to see that for any $I_{j+1}-1\geq m_1 > m_2\geq I_j$,

$$\mathbb{E}[\|S^*_{j,m_1} - S^*_{j,m_2}\|^2|W_{m_2}] = \sum_{m_2<l_1,l_2\leq m_1}\beta_{l_1}\beta_{l_2}\,\mathbb{E}[H(W_{l_1},\xi^*)^\intercal H(W_{l_2},\xi^*)|W_{m_2}]$$
$$\leq 2\sum_{m_2<l_1\leq l_2\leq m_1}\beta_{l_1}\beta_{l_2}|\,\mathbb{E}[H(W_{l_1},\xi^*)^\intercal\,\mathbb{E}[H(W_{l_2},\xi^*)|W_{l_1}]|W_{m_2}]|$$
$$\leq 2\sum_{m_2<l_1\leq l_2\leq m_1}\beta_{l_1}\beta_{l_2}\,\mathbb{E}[\|H(W_{l_1},\xi^*)\|c_1\rho^{l_2-l_1}L(1+\|W_{l_1}\|)|W_{m_2}]$$
$$\leq 2L^2\sum_{m_2<l_1\leq l_2\leq m_1}\beta_{l_1}\beta_{l_2}\,\mathbb{E}[c_1\rho^{l_2-l_1}(1+\|W_{l_1}\|)^2|W_{m_2}]$$
$$\leq \frac{4L^2c_1\beta_{I_j}^2}{1-\rho}\sum_{m_2<l_1\leq m_1}\mathbb{E}[(1+\|W_{l_1}\|)^2|W_{m_2}], \tag{B.9}$$

where the second inequality follows from Lemma SM.1.

Then, we rewrite the term $\sum_{m_2<l_1\leq m_1}\mathbb{E}[(1+\|W_{l_1}\|)^2|W_{m_2}]$ as:

$$\sum_{m_2<l_1\leq m_1}\mathbb{E}[(1+\|W_{l_1}\|)^2|W_{m_2}]$$
$$= \sum_{m_2<l_1\leq m_1}\mathbb{E}_\nu[(1+\|W_{l_1}\|)^2] + \sum_{m_2<l_1\leq m_1}\big(\mathbb{E}[(1+\|W_{l_1}\|)^2|W_{m_2}] - \mathbb{E}_\nu[(1+\|W_{l_1}\|)^2]\big).$$

$$
\begin{aligned}
&\le (m_1-m_2)K_{d_U} + \sum_{m_2<l_1\le m_1} \int_{\mathcal{W}} |\mathbb{P}(w_{l_1}|W_{m_2}) - \nu(w_{l_1})|(1+\|w_{l_1}\|)^2\,\mathrm{d}w_{l_1} \\
&\le (m_1-m_2)K_{d_U} + \sum_{m_2<l_1\le m_1} c_1\rho^{l_1-m_2}\cdot 2(1+\|W_{m_2}\|^2) \\
&\le (m_1-m_2)K_{d_U} + \frac{2c_1}{1-\rho}(1+\|W_{m_2}\|^2).
\end{aligned}
$$

Statement (i) is obtained by combining the last inequality and (B.9). □

Proof of Lemma SM.8 (ii). It is well-known that $|\mathsf{med}(x)-\mathbb{E}[x]|\le \mathsf{std}(x)$ for any one dimensional random variable x, where $\mathsf{std}(x)$ is the standard deviation of x.

Let $x = S_{j,m} - S_{j,I_{j+1}-1}|W_m$. Then, x is a d_ξ-dimensional vector. We have:

$$
\begin{aligned}
&\|\mathsf{med}(S^*_{j,m} - S^*_{j,I_{j+1}-1}|W_m)\| \\
&\le \|\,\mathbb{E}[(S^*_{j,m} - S^*_{j,I_{j+1}-1})|W_m]\| + \mathsf{d}_\xi\,\mathbb{E}[\|S^*_{j,m} - S^*_{j,I_{j+1}-1}\|^2|W_m]^{\frac{1}{2}} \\
&= \left\|\sum_{l=m}^{I_{j+1}-1}\beta_l\,\mathbb{E}[H(W_l,\xi^*)|W_m]\right\| + \mathsf{d}_\xi\,\mathbb{E}[\|S^*_{j,m} - S^*_{j,I_{j+1}-1}\|^2|W_m]^{\frac{1}{2}} \\
&\le \frac{c_1\beta_{I_j}}{1-\rho}(1+\|W_m\|) + \left(\frac{4\mathsf{d}_\xi^2 L^2 c_1}{1-\rho}\right)^{\frac{1}{2}}\beta_{I_j}\left[(I_{j+1}-I_j)^{\frac{1}{2}}K_{d_U}^{\frac{1}{2}} + \left(\frac{2c_1}{1-\rho}\right)^{\frac{1}{2}}(1+\|W_m\|)\right] \\
&\le \left(\frac{4q\mathsf{d}_\xi^2 L^2 c_1 K_{d_U}}{1-\rho}\right)^{\frac{1}{2}}\beta_{I_j} j^{\frac{q-1}{2}} + \frac{c_1(1+4\mathsf{d}_\xi L)\beta_{I_j}}{1-\rho}(1+\|W_m\|),
\end{aligned}
$$

where the second inequality follows from Lemma SM.1 and statement (i) of Lemma SM.8.

First, for j large enough,

$$
\left(\frac{4q\mathsf{d}_\xi^2 L^2 c_1 K_{d_U}}{1-\rho}\right)^{\frac{1}{2}}\beta_{I_j} j^{\frac{q-1}{2}} = \left(\frac{4q\mathsf{d}_\xi^2 L^2 c_1 K_{d_U}}{1-\rho}\right)^{\frac{1}{2}} j^{-q\delta_c} j^{-q(\delta-\delta_c-\frac{1}{2})-\frac{1}{2}} < \frac{C}{8} j^{-q\delta_c} < \frac{C}{4} I_{j+1}^{-\delta_c}. \tag{B.10}
$$

Second, by Markov's inequality,

$$
\begin{aligned}
&\mathbb{P}\left(\max_{I_j\le m<I_{j+1}} \frac{c_1(1+4\mathsf{d}_\xi L)\beta_{I_j}}{1-\rho}(1+\|W_m\|) \ge \frac{C}{4}I_{j+1}^{-\delta_c}\right) \\
&\le \sum_{I_j\le m<I_{j+1}} \mathbb{P}\left(\frac{c_1(1+4\mathsf{d}_\xi L)\beta_{I_j}}{1-\rho}(1+\|W_m\|) \ge \frac{C}{4}I_{j+1}^{-\delta_c}\right) \\
&\le (I_{j+1}-I_j) I_{j+1}^{2\delta_c}\beta_{I_j}^2 \frac{16c_1^2(1+4\mathsf{d}_\xi L)^2}{C^2(1-\rho)^2}\mathbb{E}[(1+\|W_m\|)^2] \\
&\le \frac{64c_1^2 q C_{w,2}(1+4\mathsf{d}_\xi L)^2}{C^2(1-\rho)^2} j^{-q(2\delta-2\delta_c-1)-1}.
\end{aligned} \tag{B.11}
$$

Therefore, combining (B.10) and (B.11) yields

$$
\begin{aligned}
&\sum_{j\ge s}\mathbb{P}\left(\max_{I_j\le m<I_{j+1}} \|\mathsf{med}(S^*_{j,m} - S^*_{j,I_{j+1}-1}|W_m)\| \ge \frac{C}{2}I_{j+1}^{-\delta_c}\right) \\
&\le \sum_{j\ge s}\frac{64c_1^2 q C_{w,2}(1+4\mathsf{d}_\xi L)^2}{C^2(1-\rho)^2} j^{-q(2\delta-2\delta_c-1)-1}
\end{aligned}
$$

$$\le \frac{64c_1^2 C_{w,2}(1+4\mathsf{d}_\xi L)^2}{C^2(1-\rho)^2(2\delta-2\delta_c-1)} s^{-q(2\delta-2\delta_c-1)}$$
$$\le \frac{64c_1^2 C_{w,2}(1+4\mathsf{d}_\xi L)^2}{C^2(1-\rho)^2(2\delta-2\delta_c-1)} T^{-(2\delta-2\delta_c-1)}. \qquad \square$$

Now let us return to the proof of Theorem 1. By the extended Lévy inequality for dependent partial sum process, e.g., see Page 51 of Loève (1978), for any $z \in \mathbb{R}$, we have:

$$\mathbb{P}\left(\max_{I_j \le m < I_{j+1}} \|S_{j,m}^* - \mathsf{med}(S_{j,m}^* - S_{j,I_{j+1}-1}^* | W_m)\| \ge z\right) \le 2\,\mathbb{P}(\|S_{j,I_{j+1}-1}^*\| \ge z). \tag{B.12}$$

Let $z = \frac{3}{2} C I_j^{-\delta_c}$. By (B.12) and statement (ii) of Lemma SM.8, we have that for j large enough,

$$\begin{aligned}
&\mathbb{P}\left(\max_{I_j \le m < I_{j+1}} \|S_{j,m}^*\| \ge 2C I_{j+1}^{-\delta_c}\right)\\
&\le \mathbb{P}\left(\max_{I_j \le m < I_{j+1}} \|S_{j,m}^* - \mathsf{med}(S_{j,m}^* - S_{j,I_{j+1}-1}^* | W_m)\| \ge \frac{3}{2} C I_{j+1}^{-\delta_c}\right)\\
&\quad + \mathbb{P}\left(\max_{I_j \le m < I_{j+1}} \|\mathsf{med}(S_{j,m}^* - S_{j,I_{j+1}-1}^* | W_m)\| \ge \frac{C}{2} I_{j+1}^{-\delta_c}\right)\\
&\le 2\,\mathbb{P}\left(\|S_{j,I_{j+1}-1}^*\| \ge \frac{3}{2} C I_{j+1}^{-\delta_c}\right) + \frac{64c_1^2 q C_{w,2}(1+4\mathsf{d}_\xi L)^2}{C^2(1-\rho)^2} j^{-q(2\delta-2\delta_c-1)-1}\\
&\le \frac{8}{9C^2} I_{j+1}^{2\delta_c} \mathbb{E}[\|S_{j,I_{j+1}-1}^*\|^2] + \frac{64c_1^2 q C_{w,2}(1+4\mathsf{d}_\xi L)^2}{C^2(1-\rho)^2} j^{-q(2\delta-2\delta_c-1)-1}\\
&\le \frac{8}{9C^2} I_{j+1}^{2\delta_c} \beta_{I_j}^2 \frac{4L^2 c_1}{1-\rho}\left((I_{j+1}-I_j) K_{d_U} + \frac{2c_1}{1-\rho} C_{w,2}\right) + \frac{64c_1^2 q C_{w,2}(1+4\mathsf{d}_\xi L)^2}{C^2(1-\rho)^2} j^{-q(2\delta-2\delta_c-1)-1}\\
&\le \left(\frac{64 q K_{d_U} L^2 c_1}{9C^2(1-\rho)} + \frac{64c_1^2 q C_{w,2}(1+4\mathsf{d}_\xi L)^2}{C^2(1-\rho)^2}\right) j^{-q(2\delta-2\delta_c-1)-1},
\end{aligned}$$

where the third inequality follows from Markov's inequality, and the fourth follows from statement (i) of Lemma SM.8. It follows that

$$\begin{aligned}
&\sum_{j \ge s = T^{\frac{1}{q}}} \mathbb{P}\left(\max_{I_j \le m \le I_{j+1}-1} \|S_{j,m}^*\| \le 2C I_{j+1}^{-\delta_c}\right)\\
&\ge 1 - \left(\frac{64 q K_{d_U} L^2 c_1}{9C^2(1-\rho)} + \frac{64c_1^2 q C_{w,2}(1+4\mathsf{d}_\xi L)^2}{C^2(1-\rho)^2}\right) \frac{1}{q(2\delta-2\delta_c-1)} s^{-q(2\delta-2\delta_c-1)}\\
&= 1 - \left(\frac{64 K_{d_U} L^2 c_1}{9C^2(1-\rho)(2\delta-2\delta_c-1)} + \frac{64c_1^2 C_{w,2}(1+4\mathsf{d}_\xi L)^2}{C^2(1-\rho)^2(2\delta-2\delta_c-1)}\right) T^{-(2\delta-2\delta_c-1)}\\
&= 1 - C_{S,3} T^{-(2\delta-2\delta_c-1)},
\end{aligned} \tag{B.13}$$

where

$$C_{S,3} := \left(\frac{64 K_{d_U} L^2 c_1}{9C^2(1-\rho)(2\delta-2\delta_c-1)} + \frac{64c_1^2 C_{w,2}(1+4\mathsf{d}_\xi L)^2}{C^2(1-\rho)^2(2\delta-2\delta_c-1)}\right).$$

Combining (B.1), (B.8), (B.13), and statement (ii) of Lemma SM.8, for any $\delta_c \in [0, \delta - \frac{1}{2})$, we require the following conditions for q and δ_c:

(i) $q > \frac{1}{\delta - 2\delta_c}$;

(ii) $3q\delta - 2q - 1 + 2q\delta_c > 0$;

(iii) $2\delta - 1 - 2\delta_c > 0$;

(iv) $q \geq \frac{1}{1-\delta}$.

Condition (iii) holds automatically when $\delta_c < \delta - \frac{1}{2}$. Condition (ii) is equivalent to $q(-3\delta + 2 + 2\delta_c) < 1$. If $-3\delta + 2 + 2\delta_c \leq 0$, it holds. If $-3\delta + 2 + 2\delta_c > 0$, we require that $q < \frac{1}{2-3\delta+2\delta_c}$. Combining with (i), q exists if and only if

$$\frac{1}{\delta - 2\delta_c} < \frac{1}{2 - 3\delta + 2\delta_c}.$$

The inequality above is equivalent to

$$\delta_c < \delta - \frac{1}{2}.$$

Therefore, when $\delta_c \in [0, \delta - \frac{1}{2})$, q always exists such that conditions (i)–(iv) hold at the same time. Therefore, by (B.1), (B.8), and (B.13), the joint event

$$\begin{aligned}\mathcal{A} := &\{\|\xi_{I_j} - \xi^*\| \leq C I_{j+1}^{-\delta_c}, \forall I_j \geq T\} \\ &\cap \{\|S_{j,m}^*\| \leq 2C I_{j+1}^{-\delta_c}, \forall I_j \geq T \text{ and } I_j \leq m < I_{j+1}\} \\ &\cap \{\|S_{j,m}^* - S_{j,m}\| \leq C I_{j+1}^{-\delta_c}, \forall I_j \geq T \text{ and } I_j \leq m < I_{j+1}\}\end{aligned}$$

holds with probability no less than

$$1 - C_{S,1} T^{-\frac{(q(\delta - 2\delta_c) - 1)}{q}} \ln T - C_{S,2} T^{-(3\delta - 2\delta_c - 2) - 1/q} \ln T - C_{S,3} T^{-(2\delta - 2\delta_c - 1)}. \tag{B.14}$$

Conditional on the event $\mathcal{A}$, for all $I_j \geq T$ and $I_j \leq m < I_{j+1}$, we have

$$\begin{aligned}\|\xi_t - \xi^*\| &\leq \|\xi_{I_j} - \xi^*\| + \|S_{j,m}\| \\ &\leq \|\xi_{I_j} - \xi^*\| + \|S_{j,m}^*\| + \|S_{j,m} - S_{j,m}^*\| \\ &\leq C I_{j+1}^{-\delta_c} + 2C I_{j+1}^{-\delta_c} + C I_{j+1}^{-\delta_c} \\ &= 4C I_{j+1}^{-\delta_c}.\end{aligned}$$

Let $q = \frac{1}{1-\delta}$, we have:

$$2\delta - 1 - 2\delta_c = \frac{3q\delta - 2q + 1 - 2q\delta_c}{q} = \frac{q(\delta - 2\delta_c) - 1}{q} > 0.$$

Therefore, the probability bound in (B.14) is maximized when $q = \frac{1}{1-\delta}$, which yields

$$\mathbb{P}(\|\xi_t - \xi^*\| \leq 4Ct^{-\delta_c}, t \geq T) \geq 1 - C_\xi T^{-(2\delta - 1 - 2\delta_c)} \ln T,$$

where $C_\xi := (C_{S,1} + C_{S,2} + C_{S,3})$ is a fixed constant that depends on (δ, δ_c, w, C). Replacing C with $\frac{C}{4}$, we have the conclusion for statement (ii) of Theorem 1.

For statement (i) of Theorem 1, similar to statement (ii), we can select cut-off points $\{I_j = j^{q'} : j = 1, 2, \ldots\}$, where $q' = \frac{1}{1-\kappa}$. Then the proof of statement (i) follows the same argument as before.

C. Proof of Theorem 2

Throughout this section, we suppose that Assumption 1 holds with $d_U = 4$ and Assumptions 2 and 3 hold; we also assume that $1/2 < \kappa < \delta < 1$. Moreover, we abbreviate $\mathbb{P}(\cdot|W_0 = w)$ as $\mathbb{P}(\cdot)$ for notational simplicity, unless explicitly stated otherwise.

We focus on proving statement (ii); i.e., CLT for ξ_t. The proof of statement (i) is similar. WLOG., we assume that $\beta_0 = 1$ for notational simplicity in the analysis.

By Theorem 1, with probability going to 1 as $T \to \infty$, the sample path $\{\xi_t : t \geq T\}$ is close enough to ξ^* for T large enough, so that the Taylor expansion in Assumption 3 can be applied for all $t \geq T$. Hence, we have

$$\xi_t - \xi^* = Q_{t,T}(\xi_T - \xi^*) + \sum_{m=T}^{t-1} Q_{t,m}\beta_m r_H(\xi_m) + \sum_{m=T} Q_{t,m}\beta_m \widetilde{H}(W_t, \xi_t), \tag{C.1}$$

where, with I being the identity matrix,

$$Q_{t,m} := \prod_{l=m}^{t-1} (I + \beta_l A_{22}) \leq \exp\left(A_{22} \sum_{l=m}^{t-1} \beta_l\right).$$

Recall that A_{22} is a negative definite matrix with the largest eigenvalue $\leq -\psi$. So,

$$\|Q_{t,T}(\xi_T - \xi^*)\| \leq 2B \exp\left(-\psi(t-T)t^{-\delta}\right) < 2B \exp\left(-\frac{\psi}{2} t^{1-\delta}\right),$$

when $t > 2T$ is large enough. Therefore, $\|Q_{t,T}(\xi_T - \xi^*)\|$ decays exponentially in fractional polynomial of t, thus it is dominated by other terms in (C.1), which will be explained later.

The overall strategy of the proof can be divided into three different steps.

1. First, we show that $\sum_{m=T}^{t-1} Q_{t,m}\beta_k r_H(\Gamma_k)$ converges to 0 faster than $t^{-\frac{\delta}{2}}$ with probability no less than $1 - C_1 t^{-b_1}$ for some constants $b_1 > 0$ and $C_1 > 0$.
2. Second, we decompose $\sum_{m=T} Q_{t,m}\beta_m \widetilde{H}(W_t, \Gamma_t)$ into two parts. We show that one part converges to 0 faster than $t^{-\frac{\delta}{2}}$ with probability no less than $1 - C_2 t^{-b_2}$, for some constants $b_2 > 0$ and $C_2 > 0$.
3. Third, for the other part, we establish a martingale CLT.

The following lemma will be used through out the rest of the paper.

Lemma SM.9. *Let $\{\widetilde{Q}_{t,m} : t \geq m\}$ be a sequence of real numbers indexed by (t, m) such that $0 < \widetilde{Q}_{t,m} \leq \exp\left(-C(t-m)t^{-\tilde{\delta}}\right)$, where $C > 0$ is a fixed constant. Then, for any $a \in [0, 4]$, $\tilde{\delta} \in (0, 1)$, and all t large enough, we have:*

$$\sum_{m=1}^{t} \widetilde{Q}_{t,m} m^{-a} \leq \frac{2}{C} t^{\tilde{\delta}-a}.$$

Proof of Lemma SM.9. Let $d=(1+\widetilde{\delta}+a)/C$ be a positive constant. If $m\le t-dt^{\widetilde{\delta}}\ln t$, then

$$\widetilde{Q}_{t,m}\le \exp\left(-Cd\ln t\right)\le t^{-Cd}\le t^{-(1+\widetilde{\delta}+a)}.$$

On the other hand, for all t large enough, we have: $(t-dt^{\widetilde{\delta}}\ln t)^{-a}\le \frac{3}{2}t^{-a}$, since $a\in[0,4]$. Therefore,

$$\begin{aligned}
\sum_{m=1}^{t}\widetilde{Q}_{t,m}m^{-a} &= \sum_{1\le m\le t-dt^{\widetilde{\delta}}\ln t}\widetilde{Q}_{t,m}m^{-a}+\sum_{t\ge m>t-dt^{\widetilde{\delta}}\ln t}\widetilde{Q}_{t,m}m^{-a}\\
&\le t^{-(1+\widetilde{\delta}+a)}\int_1^t m^{-a}\,\mathrm{d}m+\sum_{t\ge m>t-dt^{\widetilde{\delta}}\ln t}\widetilde{Q}_{t,m}\frac{3}{2}t^{-a}\\
&\le t^{-(1+\widetilde{\delta}+a)}\int_1^t 1\,\mathrm{d}m+\frac{3}{2}t^{-a}\int_{t-dt^{\widetilde{\delta}}\ln t}^{t}\exp\left(-Ct^{-\widetilde{\delta}}(t-m)\right)\mathrm{d}m\\
&\le t^{-\widetilde{\delta}-a}+\frac{3}{2}\frac{1}{C}t^{\widetilde{\delta}-a}\\
&\le \frac{2}{C}t^{\widetilde{\delta}-a},
\end{aligned}$$

for all t large enough. □

C.1. Step 1: Fast Convergence of the Nonlinear Term $\sum_{m=T}^{t-1}Q_{t,m}\beta_m r_H(\Gamma_m)$

LEMMA SM.10. *For t large enough,*

$$\mathbb{P}\left(\left\|\sum_{m=T}^{t-1}Q_{t,m}\beta_m r_H(\xi_m)\right\|\le t^{-\frac{2}{3}\delta}\right)\ge 1-\frac{4C_H}{\psi}\max(k_1,k_2)t^{-\frac{1}{3}\delta}\ln t,$$

where k_1 and k_2 are the constants defined in Proposition 1.

Proof of Lemma SM.10. By Assumption 3, $r_H(\xi_m)\le C_H\|\xi_m-\xi^*\|^2$. By Proposition 1, for t large enough, we have:

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{m=T}^{t-1}Q_{t,m}\beta_m r_H(\xi_m)\right\|\right] &\le C_H\sum_{m=T}^{t-1}Q_{t,m}\beta_m\,\mathbb{E}[\|\xi_m-\xi^*\|^2]\\
&\le C_H\sum_{m=T}^{t-1}Q_{t,m}\beta_m m^{-\delta}(k_1\ln t+k_2)\\
&\le \frac{4C_H\max(k_1,k_2)}{\psi}t^{-\delta}\ln t,
\end{aligned}$$

where the last inequality follows from Lemma SM.9 by setting $C=\frac{\psi}{1-\delta}$ and $a=2\delta$ there.

By Markov's inequality, we have:

$$\begin{aligned}
\mathbb{P}\left(\left\|\sum_{m=T}^{t-1}Q_{t,m}\beta_m r_H(\xi_m)\right\|\le t^{-\frac{2}{3}\delta}\right) &\ge 1-t^{\frac{2\delta}{3}}\mathbb{E}\left[\left\|\sum_{m=T}^{t-1}Q_{t,m}\beta_m r_H(\xi_m)\right\|\right]\\
&\ge 1-\frac{4C_H}{\psi}\max(k_1,k_2)t^{-\frac{1}{3}\delta}\ln t. \qquad \square
\end{aligned}$$

Lookback Strategy and Cut-off Points. Define $q = t^{\delta_q}$, where $\frac{2}{3}\delta > \delta_q > \frac{\delta}{2}$ is a generic constant. Let T be chosen such that $T = q$. Define $J_t := \lfloor \frac{t}{q} \rfloor$, and define $I_j := jq$, $j = 1, 2, \ldots, J_t$.

For simplicity, we assume that q is an integer. Otherwise we can replace q as $\lfloor t^{\delta_q} \rfloor$ and the proof remains valid as $t \to \infty$. Similarly, without loss of generality, we assume that $t = J_t q$ and the proof remains valid for general values of t since by definition, $t \in [qJ_t, q(J_t+1))$.

Now, we decompose the principle term $\sum_{m=T}^{t} Q_{t,m}\beta_m \widetilde{H}(W_m, \xi_m)$ in (C.1) as:

$$
\begin{aligned}
&\sum_{m=T}^{t} Q_{t,m}\beta_m \widetilde{H}(W_m, \xi_m) \\
&= \sum_{m=T}^{t} Q_{t,m}(H(W_m, \xi_m) - \bar{H}(\xi_m)) \\
&= \underbrace{\sum_{j=1}^{J_t-1} \sum_{m=I_j}^{I_{j+1}-1} Q_{t,m}\beta_m \widetilde{H}(W_m, \xi_{I_j})}_{\Psi_{H,1}} + \underbrace{\sum_{j=1}^{J_t-1} \sum_{m=I_j}^{I_{j+1}-1} Q_{t,m}\beta_m (\widetilde{H}(W_m, \xi_m) - \widetilde{H}(W_m, \xi_{I_j}))}_{\Psi_{H,2}}.
\end{aligned}
$$

Our next step is to show that $\Psi_{H,2}$ converges to 0 at a rate faster than $t^{-\frac{\delta}{2}}$, and then, establish a "martingale-difference array CLT" on $\Psi_{H,1}$.

From now on, we claim that we can assume that the projection operator Π_B is identity for $t \geq T = t^{\delta_q}$ as t is large enough. Consider to construct a new sequence that obeys $\widetilde{\xi}_{m+1} = \widetilde{\xi}_m + \beta_m H(W_m, \widetilde{\xi}_m)$ for all $m \geq T$, and let $\widetilde{\xi}_T = \xi_T$. By Theorem 1, we know that for any fixed C, with probability no less than $1 - C_\xi T^{-(2\delta-1)}$, we have $\|\xi_m - \xi^*\| \leq C$. By choosing C such that $\{\xi \| \|\xi - \xi^*\| \leq C\} \subset \mathsf{Ball}(B)$, we know that the projection operator is identity given that the event $\mathcal{A} := \{\|\xi_m - \xi^*\| \leq C, m = T, T+1, \ldots\}$ hold. Conditional on event $\mathcal{A}$, the path of $\xi_m, m \geq T$ will be the same as the path of $\widetilde{\xi}_m, m \geq T$.

Therefore, it suffices to establish faster than $t^{-\frac{\delta}{2}}$ of $\Psi_{H,2}$ rate and CLT on $\Psi_{H,1}$ based on the sequence $\widetilde{\xi}_m, m \geq T$ rather than using the original sequence $\xi_m, m \geq T$, as the limit distribution of $\frac{1}{\sqrt{\beta_t}}(\xi_t - \xi^*)$ and $\frac{1}{\sqrt{\beta_t}}(\widetilde{\xi}_t - \xi^*)$ will be the same asymptotically as $t \to \infty$. For notational simplicity, we still use ξ_m for $\widetilde{\xi}_m$; i.e., we assume that the projection operator Π_B is identity for $m \geq T$. In this case, we have $\xi_{m_1} = \xi_{m_2} + \sum_{m=m_2}^{m_1-1} \beta_m H(W_m, \xi_m)$ for any $m_1 \geq m_2 \geq T$, which is helpful for our analysis in the rest of this section.

C.2. Step 2: Fast Convergence of $\Psi_{H,2}$

In the lemma below we show that $\Psi_{H,2} = O_p(t^{-\delta_q})$ for a fixed $\delta_q \in (\frac{\delta}{2}, \frac{2\delta}{3})$.

Lemma SM.11. *For t large enough,*

$$\mathbb{P}\left(\|\Psi_{H,2}\| \leq t^{-\delta_q}\right) \geq 1 - C_{\Psi,2} t^{-\min\left(\frac{5}{6}\delta - \delta_q, \delta - \frac{3}{2}\delta_q\right)\ln t},$$

where $C_{\Psi,2}$ is a constant defined in (C.4).

Proof of Lemma SM.11. For any $m \in [I_j, I_{j+1})$, by the Lipshitz condition in Assumption 2, we have that

$$\|\widetilde{H}(W_m, \xi_m) - \widetilde{H}(W_m, \xi_{I_j})\| \le 2L(1 + \|W_m\|)\|\xi_m - \xi_{I_j}\|.$$

Therefore, we have

$$\begin{aligned}
\mathbb{E}[\|\psi_{H,2}\|] &\le \sum_{j=1}^{J_t} \sum_{m=I_j}^{I_{j+1}-1} \|Q_{t,m}\|\beta_m \, \mathbb{E}[2L(1+\|W_m\|)\|\xi_m - \xi_{I_j}\|] \\
&\le \sum_{j=1}^{J_t} \sum_{m=I_j}^{I_{j+1}-1} \|Q_{t,m}\|\beta_m \, \mathbb{E}[\|\sum_{l=I_j}^{m-1} \beta_l H(W_l, \xi_l)\| 2L(1+\|W_m\|)] \\
&\le \underbrace{2L \sum_{j=1}^{J_t} \sum_{m=I_j}^{I_{j+1}-1} \|Q_{t,m}\|\beta_m \, \mathbb{E}[(1+\|W_m\|)\|\sum_{l=I_j}^{m-1} \beta_l (H(W_l,\xi_l) - H(W_l,\xi^*))\|]}_{\Psi_{H,2,a}} \\
&\quad \underbrace{+ 2L \sum_{j=1}^{J_t} \sum_{m=I_j}^{I_{j+1}-1} \|Q_{t,m}\|\beta_m \, \mathbb{E}[(1+\|W_m\|)\|\sum_{l=I_j}^{m-1} \beta_l H(W_l,\xi^*)\|]}_{\Psi_{H,2,b}}.
\end{aligned}$$

By Assumption 2, for t large enough, we have

$$\begin{aligned}
\Psi_{H,2,a} &\le 2L \sum_{j=1}^{J_t} \sum_{m=I_j}^{I_{j+1}-1} \|Q_{t,m}\|\beta_k \, \mathbb{E}[(1+\|W_m\|)\|\sum_{l=I_j}^{m-1} \beta_l (H(W_l,\xi_l) - H(W_l,\xi^*))\|] \\
&\le 2L^2 \sum_{j=1}^{J_t} \sum_{m=I_j}^{I_{j+1}-1} \|Q_{t,m}\|\beta_m \, \mathbb{E}[(1+\|W_m\|)^2 \|\sum_{l=I_j}^{m-1} \beta_l (1+\|W_l\|)\|^2]^{\frac{1}{2}} \, \mathbb{E}[\|\xi_l - \xi^*\|^2]^{\frac{1}{2}} \\
&\le 8L^2 C_{w,4}^{\frac{1}{2}} \sum_{j=1}^{J_t} \sum_{m=I_j}^{I_{j+1}-1} \|Q_{t,m}\|\beta_m \sum_{l=I_j}^{m-1} \beta_l m^{-\frac{\delta}{2}} (k_1 \ln m + k_2)^{\frac{1}{2}} \\
&\le 16L^2 C_{w,4}^{\frac{1}{2}} \max(k_1,k_2,1) \ln t \sum_{j=1}^{J_t} \sum_{m=I_j}^{I_{j+1}-1} q\|Q_{t,I_m}\| m^{-\frac{5}{2}\delta} \\
&\le \frac{32L^2 C_{w,4}^{\frac{1}{2}} \max(k_1,k_2,1)}{\psi} t^{-\frac{3}{2}\delta + \delta_q} \ln t \\
&\le \frac{32L^2 C_{w,4}^{\frac{1}{2}} \max(k_1,k_2,1)}{\psi} t^{-\frac{5}{6}\delta} \ln t, \qquad \text{(C.2)}
\end{aligned}$$

where the second inequality follows from Assumption 3 and the Cauchy-Schwarz inequality, the third follows from Lemma SM.1 and Proposition 1, and the second to last follows from Lemma SM.9.

For $\Psi_{H,2,b}$, define $d = \frac{4}{|\ln \rho|}$, consider

$$\mathbb{E}\left[\left\|\sum_{l=I_j}^{m-1} \beta_l H(W_l, \xi^*)\right\|^2\right] \le \sum_{I_j \le l_1, l_2 \le I_{j+1}-1} \beta_{l_1}\beta_{l_2} \, \mathbb{E}[H(W_{l_1}, \xi^*)^\intercal H(W_{l_2}, \xi^*)]$$

$$= \sum_{I_j \le l_1, l_2 \le I_{j+1}-1, |l_1-l_2| \le d \ln t} \beta_{l_1}\beta_{l_2}\, \mathbb{E}[H(W_{l_1},\xi^*)^\intercal H(W_{l_2},\xi^*)] + \sum_{I_j \le l_1, l_2 \le I_{j+1}-1, |l_1-l_2| > d \ln t} \beta_{l_1}\beta_{l_2}\, \mathbb{E}[H(W_{l_1},\xi^*)^\intercal H(W_{l_2},\xi^*)].$$

On the one hand, by Lemma SM.1, we have:

$$\begin{aligned}
&\sum_{I_j \le l_1, l_2 \le I_{j+1}-1, |l_1-l_2| > d \ln t} \beta_{l_1}\beta_{l_2}\, \mathbb{E}[H(W_{l_1},\xi^*)H(W_{l_2},\xi^*)] \\
&\le 2 \sum_{I_j \le l_1, l_2 \le I_{j+1}-1, l_2-l_1 > d \ln t} \beta_{l_1}\beta_{l_2}\, \mathbb{E}[H(W_{l_1},\xi^*)\, \mathbb{E}[H(W_{l_2},\xi^*)|W_{l_1}]] \\
&\le 2 \sum_{I_j \le l_1, l_2 \le I_{j+1}-1, l_2-l_1 > d \ln t} \beta_{l_1}\beta_{l_2}\, \mathbb{E}[H(W_{l_1},\xi^*)L(1+\|W_{l_1}\|)c_1\rho^{l_2-l_1}] \\
&\le 4c_1L^2C_{w,2}q^2\beta_{I_j}^2 t^{-4}.
\end{aligned}$$

On the other hand, the number of pairs (l_1,l_2) such that $I_j \le l_1, l_2 \le I_{j+1}-1, |l_1-l_2| \le d \ln t$ is less than or equal to $2dq\ln t$. Therefore,

$$\begin{aligned}
&\sum_{I_j \le l_1, l_2 \le I_{j+1}-1, |l_1-l_2| \le d \ln t} \beta_{l_1}\beta_{l_2}\, \mathbb{E}[H(W_{l_1},\xi^*)H(W_{l_2},\xi^*)] \\
&\le \sum_{I_j \le l_1, l_2 \le I_{j+1}-1, |l_1-l_2| \le d \ln t} \beta_{l_1}\beta_{l_2}\mathbb{E}[L^2(1+\|W_{l_1}\|)(1+\|W_{l_2}\|)] \\
&\le L^2 \sum_{I_j \le l_1, l_2 \le I_{j+1}-1, |l_1-l_2| \le d \ln t} \beta_{l_1}\beta_{l_2}\mathbb{E}[(1+\|W_{l_1}\|^2)+(1+\|W_{l_2}\|^2)] \\
&\le 4L^2C_{w,2}dq\beta_{I_j}^2 \ln t,
\end{aligned}$$

where the last inequality is followed by Lemma SM.1

Therefore, for $I_j \ge T$ and t, T large enough,

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{l=I_j}^{m-1}\beta_l H(W_l,\xi^*)\right\|^2\right] &\le 4c_1L^2C_{w,2}q^2\beta_{I_j}^2 t^{-4} + 4L^2C_{w,2}dq\beta_{I_j}^2 \ln t \\
&\le 8L^2C_{w,2}dq\beta_{I_j}^2 \ln t.
\end{aligned}$$

Therefore, we have:

$$\begin{aligned}
\Psi_{H,2,b} &\le L\sum_{j=1}^{J_t}\sum_{m=I_j}^{I_{j+1}-1}\|Q_{t,m}\|\beta_m\,\mathbb{E}\left[(1+\|W_m\|)\left\|\sum_{l=I_j}^{m-1}\beta_l H(W_l,\xi^*)\right\|\right] \\
&\le L\sum_{j=1}^{J_t}\sum_{m=I_j}^{I_{j+1}-1}\|Q_{t,m}\|\beta_m\,\mathbb{E}[(1+\|W_m\|)^2]^{\frac{1}{2}}\,\mathbb{E}\left[\left\|\sum_{l=I_j}^{m-1}\beta_l H(W_l,\xi^*)\right\|^2\right]^{\frac{1}{2}} \\
&\le 2L\sqrt{2dqC_{w,2}\ln t}\sqrt{2C_{w,2}}\sum_{j=1}^{J_t}\sum_{m=I_j}^{I_{j+1}-1}\|Q_{t,m}\|\beta_m\beta_{I_j}
\end{aligned}$$

$$
\begin{aligned}
&\leq 4L\sqrt{dq\ln t}C_{w,2}\frac{2}{\psi}t^{-\delta} \\
&= 4L\sqrt{d}KC_{w,2}\frac{2}{\psi}t^{-\delta+\frac{\delta_q}{2}}\ln t \\
&\leq 4L\sqrt{d}C_{w,2}\frac{2}{\psi}t^{-\delta_q+(\frac{3\delta_q}{2}-\delta)}\ln t, \qquad \text{(C.3)}
\end{aligned}
$$

where the forth inequality follows from Lemma SM.9.

Combing (C.2) and (C.3), we have:

$$
\begin{aligned}
\Psi_{H,2} &\leq \Psi_{H,2,a} + \Psi_{H,2,b} \\
&\leq \frac{32L^2C_{w,4}^{\frac{1}{2}}\max(k_1,k_2,1)}{\psi}t^{-\delta_q-(\frac{5}{6}\delta-\delta_q)}\ln t + 4L\sqrt{d}C_{w,2}\frac{2}{\psi}t^{-\delta_q+(\frac{3\delta_q}{2}-\delta)}\ln t \\
&\leq C_{\Psi,2}t^{-\delta_q-\min(\frac{5}{6}\delta-\delta_q,\delta-\frac{3}{2}\delta_q)}\ln t,
\end{aligned}
$$

where

$$
C_{\Psi,2} := \frac{32L^2C_{w,4}^{\frac{1}{2}}\max(k_1,k_2,1)}{\psi} + 4L\sqrt{d}C_{w,2}\frac{2}{\psi}. \qquad \text{(C.4)}
$$

By Markov's inequality, we have

$$
\mathbb{P}\left(\Psi_{H_2} \leq t^{-\delta_q}\right) \geq 1 - C_{\Psi,2}t^{-\min(\frac{5}{6}\delta-\delta_q,\delta-\frac{3}{2}\delta_q)}\ln t. \qquad \square
$$

C.3. Step 3: Martingale Central Limit Theorem

Now we a focus on the analysis of $\Psi_{H,1}$. Define

$$
\begin{aligned}
X_{t,j} &:= \beta_t^{-\frac{1}{2}}\sum_{m=I_j}^{I_{j+1}-1}Q_{t,m}\beta_m\widetilde{H}(W_m,\xi_{I_j}), \\
X_{t,j}^* &:= (X_{t,j} - \mathbb{E}[X_{t,j}|\mathcal{F}_{I_j}]).
\end{aligned}
$$

Therefore, we can rewrite $\Psi_{H,1}$ as:

$$
\Psi_{H,1} = \underbrace{\beta_t^{\frac{1}{2}}\sum_{j=1}^{J_t-1}\mathbb{E}[X_{t,j}|\mathcal{F}_{I_j}]}_{\Psi_{H,3,1}} + \underbrace{\beta_t^{\frac{1}{2}}\sum_{j=1}^{J_t-1}X_{t,j}^*}_{\Psi_{H,3,2}}.
$$

By definition, $\Psi_{H,3,2}$ is a sum of martingale difference arrays. We strategy is to prove that $\Psi_{H,3,1} = o_p(t^{-\frac{\delta}{2}})$ and establish martingale CLT on $\Psi_{H,3,2}$. We divide the rest of the proof as Section C.4 and Section C.5.

C.4. Step 3a: Fast Convergence of $\Psi_{H,3,1}$

In this step, we bound the bias term $\Psi_{H,3,1}$ and show that $\Psi_{H,3,1}$ has faster than $t^{-\frac{\delta}{2}}$ convergence in Lemma SM.12.

Lemma SM.12. *For each $j \geq T$, t large enough, we have:*

$$\beta_t^{\frac{1}{2}} \| \mathbb{E}[X_{t,j}|\mathcal{F}_{I_j}]\| \leq \left(c_1 t^{-(1+\delta-\delta_q)} + \frac{1+\delta-\delta_q}{|\ln \rho|} \|Q_{t,I_{j+1}}\| \beta_{I_j} \ln t \right) L(1+||W_{I_j}||),$$

and

$$\mathbb{E}[\|\Psi_{H,3,1}\|] \leq \left(\frac{c_1}{1-\rho} t^{-\delta} + \frac{4(1+\delta-\delta_q)}{\psi |\ln \rho|} t^{-\delta_q} \ln t \right) LC_{w,1}.$$

Proof of Lemma SM.12. By Lemma SM.1, for $m > I_j$,

$$\begin{aligned} \| \mathbb{E}[\widetilde{H}(W_m, \xi_{I_j})|\mathcal{F}_{I_j}]\| &= \left\| \int_{\mathcal{W}} (\mathbb{P}(w_m|\mathcal{F}_{I_j}) - \nu(w_m)) H(w_m, \xi_{I_j}) \, \mathrm{d}w_m \right\| \\ &\leq c_1 \rho^{m-I_j} L(1+||W_{I_j}||). \end{aligned} \tag{C.5}$$

Given a constant $d > 0$, for each j, by (C.5), it is easy to see that

$$\begin{aligned} &\beta_t^{\frac{1}{2}} \| \mathbb{E}[X_{t,j}|\mathcal{F}_{I_j}]\| \\ &= \sum_{m=I_j}^{I_{j+1}-1} \beta_m \|Q_{t,m}\| \| \mathbb{E}[\widetilde{H}(W_m, \xi_{I_j})|\mathcal{F}_{I_j}]\| \\ &= \sum_{I_{j+1} > m \geq d \ln t + I_j} \beta_m \|Q_{t,m}\| \| \mathbb{E}[\widetilde{H}(W_m, \xi_{I_j})|\mathcal{F}_{I_j}]\| + \sum_{I_j \leq m < d \ln t + I_j} \beta_m \|Q_{t,m}\| \| \mathbb{E}[\widetilde{H}(W_m, \xi_{I_j})|\mathcal{F}_{I_j}]\| \\ &\leq \frac{c_1}{1-\rho} \rho^{d \ln t} L(1+||W_{I_j}||) + \sum_{I_j \leq m < d \ln t + I_j} \|Q_{t,m}\| \beta_m L(1+||W_{I_j}||) \\ &\leq \frac{c_1}{1-\rho} t^{-d|\ln \rho|} L(1+||W_{I_j}||) + d \ln t \|Q_{t,I_{j+1}}\| \beta_{I_j} L(1+||W_{I_j}||), \end{aligned}$$

where the first and second inequalities follow from Lemma SM.1 and Lemma SM.9. It follows that

$$\begin{aligned} &\beta_t^{\frac{1}{2}} \sum_{j=1}^{J_t-1} \mathbb{E}[\| \mathbb{E}[X_{t,j}|\mathcal{F}_{I_j}]\|] \\ &\leq \sum_{j=1}^{J_t-1} \frac{c_1}{1-\rho} t^{-d|\ln \rho|} LC_{w,1} + d \ln t \sum_{j=1}^{J_t-1} \|Q_{t,I_{j+1}}\| \beta_{I_j} LC_{w,1} \\ &\leq \left(\frac{c_1}{1-\rho} t^{-d|\ln \rho|+1-\delta_q} + d \ln t \sum_{j=1}^{J_t-1} \exp^{-\psi t^{-\delta}(t-(j+1)q)} (jq)^{-\delta} \right) LC_{w,1}. \end{aligned} \tag{C.6}$$

Similar to Lemma SM.9, it is easy to see that for t large enough, we have:

$$\sum_{j=1}^{J_t-1} \exp^{-\psi t^{-\delta}(t-(j+1)q)} (jq)^{-\delta} \leq \frac{2}{\psi} \frac{t^{\delta}}{q} (J_t q)^{-\delta} \leq \frac{4}{\psi} t^{-\delta_q},$$

for t large enough. Therefore, by choosing $d = \frac{1+\delta-\delta_q}{|\ln \rho|}$, (C.6) implies that

$$\begin{aligned} \mathbb{E}[\|\Psi_{H,3,1}\|] &\leq \beta_t^{\frac{1}{2}} \sum_{j=1}^{J_t-1} \mathbb{E}[\| \mathbb{E}[X_{t,j}|\mathcal{F}_{I_j}]\|] \\ &\leq \left(\frac{c_1}{1-\rho} t^{-\delta} + d \frac{4}{\psi} t^{-\delta_q} \ln t \right) LC_{w,1}, \end{aligned}$$

which implies that $\|\Psi_{H,3,1}\| = O_p(t^{-\delta_q} \ln t) = o_p(t^{-\frac{\delta}{2}})$. □

C.5. Step 3b: Martingale Central Limit Theorem on $\Psi_{H,3,2}$

At last, we establish a martingale CLT on $\Psi_{H,3,2} \coloneqq \sum_{j=1}^{J_t-1} X^*_{t,j}$. It is easy to see that $X^*_{t,j}$, $j = 1,2,\ldots,J_t-1$. is a martingale-difference array indexed by j. According to Chapter 7.1 of Ethier and Kurtz (1986), to establish a martingale CLT, it suffices to verify that:

(a) $\sum_{j=1}^{J_t-1} X^*_{t,j}(X^*_{t,j})^\intercal \to_p \Sigma$ for some positive definite matrix Σ.

(b) $\mathbb{E}[\max_{1\le j\le J_t-1} \|X^*_{t,j}\|] \to 0$ as $t\to\infty$.

C.5.1. Verifying Condition (a). We first verify condition (a). For each j, we consider approximating $\mathbb{E}[X^*_{t,j}(X^*_{t,j})^\intercal|\mathcal{F}_{I_j}]$ by $\mathbb{E}[X_{t,j}X^\intercal_{t,j}|\mathcal{F}_{I_j}]$. By definition, we have:

$$\mathbb{E}[X^*_{t,j}(X^*_{t,j})^\intercal|\mathcal{F}_{I_j}] - \mathbb{E}[X_{t,j}X^\intercal_{t,j}|\mathcal{F}_{I_j}] = \mathbb{E}[X_{t,j}|\mathcal{F}_{I_j}]\,\mathbb{E}[X_{t,j}|\mathcal{F}_{I_j}]^\intercal.$$

By Lemma SM.12, we have:

$$\begin{aligned}
&\|\,\mathbb{E}[X^*_{t,j}(X^*_{t,j})^\intercal|\mathcal{F}_{I_j}] - \mathbb{E}[X_{t,j}X^\intercal_{t,j}|\mathcal{F}_{I_j}]\|\\
&\le \|\,\mathbb{E}[X_{t,j}|\mathcal{F}_{I_j}]\|^2\\
&\le t^\delta L^2(1+||W_{I_j}||)^2(c_1 t^{-(1+\delta-\delta_q)} + \frac{1+\delta-\delta_q}{|\ln\rho|}\frac{4}{\psi}\|Q_{t,I_{j+1}}\|\beta_{I_j}\ln t)^2\\
&\le 2L^2(1+||W_{I_j}||)^2(c_1^2 t^{-2-\delta+2\delta_q} + t^\delta(\frac{1+\delta-\delta_q}{|\ln\rho|}\frac{4}{\psi})^2\|Q_{t,I_{j+1}}\|^2\beta^2_{I_j}\ln^2 t).
\end{aligned}$$

Consequently, as $t\to\infty$, we have:

$$\begin{aligned}
&\sum_{j=1}^{J_t-1}\mathbb{E}[\|\,\mathbb{E}[X^*_{t,j}(X^*_{t,j})^\intercal|\mathcal{F}_{I_j}] - \mathbb{E}[X_{t,j}X^\intercal_{t,j}|\mathcal{F}_{I_j}]\|]\\
&\le 2L^2C_{w,2}\left(c_1^2 J_t t^{-2-2\delta+2\delta_q} + t^\delta(\frac{1+\delta-\delta_q}{|\ln\rho|})^2\sum_{j=1}^{J_t-1}\|Q_{t,I_{j+1}}\|^2\beta^2_{I_j}\ln^2 t\right)\\
&\le 2L^2C_{w,2}\left(c_1^2 t^{-1-2\delta+\delta_q} + (\frac{1+\delta-\delta_q}{|\ln\rho|})^2\frac{2}{2\psi}t^{-\delta_q}\ln^2 t\right)\to 0,
\end{aligned}$$

as $t\to\infty$, where the second inequality follows from Lemma SM.9. Therefore, it suffices to show that

$$\sum_{j=1}^{J_t}[X_{t,j}(X_{t,j})^\intercal] \to_p \Sigma_H,$$

where Σ_H is defined in statement (ii) of Theorem 2.

By definition of $X_{t,j}$, we have:

$$\begin{aligned}
&\mathbb{E}[X_{t,j}(X_{t,j})^\intercal|\mathcal{F}_{I_j}]\\
&= \frac{t^\delta}{\beta_0}\mathbb{E}\Bigg[\Bigg(\sum_{m=I_j}^{I_{j+1}-1} Q_{t,m}\beta_m\widetilde{H}(W_m,\xi_{I_j})\Bigg)\Bigg(\sum_{m=I_j}^{I_{j+1}-1} Q_{t,m}\beta_m\widetilde{H}(W_m,\xi_{I_j})\Bigg)^\intercal\Bigg|\mathcal{F}_{I_j}\Bigg]
\end{aligned}$$

$$= \frac{t^\delta}{\beta_0} \sum_{m=I_j}^{I_{j+1}-1} \mathbb{E}[Q_{t,m}\beta_m^2 \widetilde{H}(W_m, \xi_{I_j}) \widetilde{H}(W_m, \xi_{I_j})^\intercal Q_{t,m}^\intercal | \mathcal{F}_{I_j}] \tag{C.7}$$

$$+ \frac{t^\delta}{\beta_0} \mathbb{E}\Bigg[\sum_{I_j \le m_1 < m_2 \le I_{j+1}-1} Q_{t,m_1}\beta_{m_1}\beta_{m_2} \widetilde{H}(W_{m_1}, \xi_{I_j}) \widetilde{H}(W_{m_2}, \xi_{I_j})^\intercal Q_{t,m_2}^\intercal \Bigg| \mathcal{F}_{I_j} \Bigg] \tag{C.8}$$

$$+ \frac{t^\delta}{\beta_0} \Bigg(\mathbb{E}\Bigg[\sum_{I_j \le m_1 < m_2 \le I_{j+1}-1} Q_{t,m_1}\beta_{m_1}\beta_{m_2} \widetilde{H}(W_{m_1}, \xi_{I_j}) \widetilde{H}(W_{m_2}, \xi_{I_j})^\intercal Q_{t,m_2}^\intercal \Bigg| \mathcal{F}_{I_j} \Bigg] \Bigg)^\intercal. \tag{C.9}$$

The term in (C.7) can be written as:

$$\frac{t^\delta}{\beta_0} \sum_{m=I_j}^{I_{j+1}-1} \mathbb{E}[Q_{t,m}\beta_m^2 \widetilde{H}(W_m, \xi_{I_j}) \widetilde{H}(W_m, \xi_{I_j})^\intercal Q_{t,m}^\intercal | \mathcal{F}_{I_j}]$$
$$= \frac{t^\delta}{\beta_0} \sum_{m=I_j}^{I_{j+1}-1} \mathbb{E}[Q_{t,m}^2\beta_m^2 \widetilde{H}(W_m, \xi^*) \widetilde{H}(W_m, \xi^*)^\intercal Q_{t,m}^\intercal | \mathcal{F}_{I_j}] + r_{j,1},$$

where

$$r_{j,1} \coloneqq \frac{t^\delta}{\beta_0} \sum_{m=I_j}^{I_{j+1}-1} \mathbb{E}[Q_{t,m}\beta_m^2 \widetilde{H}(W_m, \xi_{I_j}) \widetilde{H}(W_m, \xi_{I_j})^\intercal Q_{t,m}^\intercal | \mathcal{F}_{I_j}]$$
$$- \frac{t^\delta}{\beta_0} \sum_{m=I_j}^{I_{j+1}-1} \mathbb{E}[Q_{t,m}\beta_m^2 \widetilde{H}(W_m, \xi^*) \widetilde{H}(W_m, \xi^*)^\intercal Q_{t,m}^\intercal | \mathcal{F}_{I_j}].$$

By Assumption 1 and Proposition 1,

$$\mathbb{E}[\|r_{j,1}\|] \coloneqq \mathbb{E}\Bigg[\Bigg\| \frac{t^\delta}{\beta_0} \sum_{m=I_j}^{I_{j+1}-1} \mathbb{E}[Q_{t,m}\beta_m^2 \widetilde{H}(W_m, \xi_{I_j}) \widetilde{H}(W_m, \xi_{I_j})^\intercal Q_{t,m}^\intercal | \mathcal{F}_{I_j}]$$
$$- \frac{t^\delta}{\beta_0} \sum_{m=I_j}^{I_{j+1}-1} \mathbb{E}[Q_{t,m}\beta_m^2 \widetilde{H}(W_m, \xi^*) \widetilde{H}(W_m, \xi^*)^\intercal Q_{t,m}^\intercal | \mathcal{F}_{I_j}] \Bigg\|\Bigg]$$
$$\le 2L^2 \frac{t^\delta}{\beta_0} \sum_{m=I_j}^{I_{j+1}-1} \|Q_{t,m}^2\| \beta_m^2 \mathbb{E}[(1 + \|W_m\|)\|^2 \xi_{I_j} - \xi^*\|]$$
$$\le 2L^2 \frac{t^\delta}{\beta_0} \sum_{m=I_j}^{I_{j+1}-1} \|Q_{t,m}^2\| \beta_m^2 \mathbb{E}[\|(1 + \|W_m\|)\|^4]^{\frac{1}{2}} \, \mathbb{E}[\|\xi_{I_j} - \xi^*\|^2]^{\frac{1}{2}}.$$
$$\le 4L^4 C_{w,4}^{\frac{1}{2}} \frac{t^\delta}{\beta_0} \sum_{m=I_j}^{I_{j+1}-1} \|Q_{t,m}^2\| \beta_m^2 I_j^{-\frac{\delta}{2}} (k_1 \ln I_j + k_2)^{\frac{1}{2}}.$$

Consequently, for t large enough, we have:

$$\sum_{j=1}^{J_t-1} \mathbb{E}[\|r_{j,1}\|] \le \sum_{j=1}^{J_t} 4L^2 C_{w,4}^{\frac{1}{2}} \frac{t^\delta}{\beta_0} \sum_{m=I_j}^{I_{j+1}-1} Q_{t,m}^2 \beta_m^2 I_j^{-\frac{\delta}{2}} (k_1 \ln I_j + k_2)^{\frac{1}{2}}$$
$$\le 8L^2 C_{w,4}^{\frac{1}{2}} \frac{2}{2\psi\beta_0} t^{-\frac{\delta}{2}} \max(k_1, k_2)^{\frac{1}{2}} \ln^{\frac{1}{2}} t \to 0,$$

where the second inequality follows from Lemma SM.9 by setting $a=\frac{5\delta}{2}$.

Next, we will analyze the term in (C.8). The term in (C.9) will follow the same analysis since it is simply the transpose of the term in (C.8). For any $m_2-m_1=l\geq 0$, recall that

$$L_H(l):=\int_{\mathcal{W}\times\mathcal{W}}\mathbb{P}(w_{m_2}|w_{m_1})\nu(w_{m_1})\widetilde{H}(w_{m_1},\xi^*)\widetilde{H}(w_{m_2},\xi^*)^\intercal\,\mathrm{d}w_{m_2}\,\mathrm{d}w_{m_1}\,.$$

Since $W_m, m\geq 1$ is a Markov chain and ν is the stationary distribution of it, the term $L_H(l)$ does not depend on m_1,m_2 but only depend on the difference $l:=m_2-m_1$. By Assumption 1,

$$\begin{aligned}
\|L_H(l)\| &\leq \left\|\int_{\mathcal{W}\times\mathcal{W}}\mathbb{P}(w_{m_2}|w_{m_1})\nu(w_{m_1})\widetilde{H}(w_{m_1},\xi^*)\widetilde{H}(w_{m_2},\xi^*)^\intercal\,\mathrm{d}w_{m_2}\,\mathrm{d}w_{m_1}\right\|\\
&=\left\|\int_{\mathcal{W}\times\mathcal{W}}\nu(w_{m_1})\widetilde{H}(w_{m_1},\xi^*)\,\mathbb{E}[\widetilde{H}(w_{m_2},\xi^*)^\intercal|w_{m_1}]\,\mathrm{d}w_{m_2}\,\mathrm{d}w_{m_1}\right\|\\
&\leq c_1\rho^l\,\mathbb{E}_\nu[\|\widetilde{H}(W_{m_1},\xi^*)L(1+||W_{m_1}||)\|]\\
&\leq 2L^2K_{d_U}c_1\rho^l,
\end{aligned}$$

where the second inequality follows from Lemma SM.1.

It follows that $\bar{L}_H=L_H(0)+\sum_{l=1}^\infty(L_H(l)+L_H(l)^\intercal)$ exists since $\sum_{l=0}^\infty 2L^2K_{d_U}c_1\rho^l<\infty$.

For any $I_{j+1}>m_2>m_1\geq I_j$,

$$\begin{aligned}
&\mathbb{E}[\widetilde{H}(W_{m_1},\xi_{I_j})\widetilde{H}(W_{m_2},\xi_{I_j})^\intercal|\mathcal{F}_{I_j}]\\
&=\int_{\mathcal{W}\times\mathcal{W}}\mathbb{P}(w_{m_2}|\mathcal{F}_{m_1+1})\,\mathbb{P}(w_{m_1}|\mathcal{F}_{I_j})\widetilde{H}(w_{m_1},\xi_{I_j})\widetilde{H}(w_{m_2},\xi_{I_j})^\intercal\,\mathrm{d}w_{m_2}\,\mathrm{d}w_{m_1}\\
&=L_H(m_2-m_1)\\
&\quad+\underbrace{\int_{\mathcal{W}\times\mathcal{W}}\mathbb{P}(w_{m_2}|\mathcal{F}_{m_1+1})\nu(w_{m_1})\left(\widetilde{H}(w_{m_1},\xi_{I_j})\widetilde{H}(w_{m_2},\xi_{I_j})^\intercal-\widetilde{H}(w_{m_1},\xi^*)\widetilde{H}(w_{m_2},\xi^*)^\intercal\right)\mathrm{d}w_{m_2}\,\mathrm{d}w_{m_1}}_{\Psi_{H,4,1}}\\
&\quad+\underbrace{\int_{\mathcal{W}\times\mathcal{W}}\mathbb{P}(w_{m_2}|\mathcal{F}_{m_1+1})\big(\mathbb{P}(w_{m_1}|\mathcal{F}_{I_j})-\nu(w_{m_1})\big)\widetilde{H}(w_{m_1},\xi_{I_j})\widetilde{H}(w_{m_2},\xi_{I_j})^\intercal\,\mathrm{d}w_{m_2}\,\mathrm{d}w_{m_1}}_{\Psi_{H,4,2}}\,.
\end{aligned}$$

By the triangular inequality, consider to bound $\Psi_{H,4,1}$:

$$\begin{aligned}
&\|\Psi_{H,4,1}\|\\
\leq&\left\|\int_{\mathcal{W}\times\mathcal{W}}\mathbb{P}(w_{m_2}|\mathcal{F}_{m_1+1})\nu(w_{m_1})\left(\widetilde{H}(w_{m_1},\xi_{I_j})-\widetilde{H}(w_{m_1},\xi^*)\right)\widetilde{H}(w_{m_2},\xi_{I_j})^\intercal\,\mathrm{d}w_{m_2}\,\mathrm{d}w_{m_1}\right\|\\
&+\left\|\int_{\mathcal{W}\times\mathcal{W}}\nu(w_{m_2})\nu(w_{m_1})\widetilde{H}(w_{m_1},\xi^*)\left(\widetilde{H}(w_{m_2},\xi_{I_j})-\widetilde{H}(w_{m_2},\xi^*)\right)^\intercal\mathrm{d}w_{m_2}\,\mathrm{d}w_{m_1}\right\|\\
&+\left\|\int_{\mathcal{W}\times\mathcal{W}}(\mathbb{P}(w_{m_2}|\mathcal{F}_{m_1+1})-\nu(w_{m_2}))\nu(w_{m_1})\widetilde{H}(w_{m_1},\xi^*)\right.\\
&\qquad\left.\left(\widetilde{H}(w_{m_2},\xi_{I_j})-\widetilde{H}(w_{m_2},\xi^*)\right)^\intercal\mathrm{d}w_{m_2}\,\mathrm{d}w_{m_1}\right\|
\end{aligned}$$

$$
\begin{aligned}
&\leq \int_{\mathcal{W}} \left\| \int_{\mathcal{W}} \mathbb{P}(w_{m_2}|\mathcal{F}_{m_1+1}) \widetilde{H}(w_{m_2}, \xi_{I_j}) \, \mathrm{d}w_{m_2} \right\| \cdot \nu(w_{m_1}) \left\| \widetilde{H}(w_{m_1}, \xi_{I_j}) - \widetilde{H}(w_{m_1}, \xi^*) \right\| \mathrm{d}w_{m_1} \\
&\quad + \left\| \int_{\mathcal{W}} \nu(w_{m_2}) (\widetilde{H}(w_{m_2}, \xi_{I_j}) - \widetilde{H}(w_{m_2}, \xi^*)) \, \mathrm{d}w_{m_2} \right\| \left\| \int_{\mathcal{W}} \nu(w_{m_1}) \widetilde{H}(w_{m_1}, \xi^*) \, \mathrm{d}w_{m_1} \right\| \\
&\quad + \int_{\mathcal{W}} \left\| \int_{\mathcal{W}} (\mathbb{P}(w_{m_2}|\mathcal{F}_{m_1+1}) - \nu(w_{m_2})) \left(\widetilde{H}(w_{m_2}, \xi_{I_j}) - \widetilde{H}(w_{m_2}, \xi^*) \right) \mathrm{d}w_{m_2} \right\| \\
&\qquad\qquad \cdot \left\| \nu(w_{m_1}) \widetilde{H}(w_{m_1}, \xi^*) \right\| \mathrm{d}w_{m_1} \\
&\leq 2 c_1 \rho^{m_2 - m_1} L K_{d_U} \|\xi_{I_j} - \xi^*\| + 2 L K_{d_U} c_1 \rho^{m_1 - I_j} \|\xi_{I_j} - \xi^*\| + 2 L K_{d_U} c_1 \rho^{m_2 - m_1} \|\xi_{I_j} - \xi^*\| \\
&\leq 6 L K_{d_U} c_1 \rho^{\min(m_2 - m_1, m_1 - I_j)} \|\xi_{I_j} - \xi^*\|,
\end{aligned}
$$

where the second to last inequality follows from Lemma SM.1 and Assumption 2. Therefore, by Proposition 1,

$$
\begin{aligned}
\mathbb{E}[||\Psi_{H,4,1}||] &\leq 6 L K_{d_U} c_1 \rho^{\min(m_2 - m_1, m_1 - I_j)} \mathbb{E}[\|\xi_{I_j} - \xi^*\|] \\
&\leq 6 L K_{d_U} c_1 \rho^{\min(m_2 - m_1, m_1 - I_j)} I_j^{-\frac{\delta}{2}} (k_1 \ln I_j + k_2).
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
\|\Psi_{H,4,2}\| &\leq \int_{\mathcal{W}} \left\| \int_{\mathcal{W}} \mathbb{P}(w_{m_2}|\mathcal{F}_{m_1+1}) \widetilde{H}(w_{m_2}, \xi_{I_j}) \, \mathrm{d}w_{m_2} \right\| |\mathbb{P}(w_{m_1}|\mathcal{I}_j) - \nu(w_{m_1})| \|\widetilde{H}(w_{m_1}, \xi_{I_j})\| \, \mathrm{d}w_{m_1} \\
&\leq \int_{\mathcal{W}} L c_1 (\|w_{m_1}\| + 1) |\mathbb{P}(w_{m_1}|\mathcal{I}_j) - \nu(w_{m_1})| \|\widetilde{H}(w_{m_1}, \xi_{I_j})\| \, \mathrm{d}w_{m_1} \\
&\leq L^2 c_1 \rho^{m_2 - m_1} \rho^{m_1 - I_j} (1 + \|W_{I_j}\|^2) c_1 \rho^{m_2 - I_j} \\
&= L^2 c_1^2 \rho^{m_2 - I_j} (1 + \|W_{I_j}\|^2).
\end{aligned}
$$

We rewrite (C.8) as:

$$
\begin{aligned}
&\frac{t^\delta}{\beta_0} \mathbb{E}\left[\sum_{I_j \leq m_1 < m_2 \leq I_{j+1}-1} Q_{t,m_1} \beta_{m_1} \beta_{m_2} \widetilde{H}(W_{m_1}, \xi_{I_j}) \widetilde{H}(W_{m_2}, \xi_{I_j})^\intercal Q_{t,m_2}^\intercal \middle| \mathcal{F}_{I_j} \right] \\
&= \frac{t^\delta}{\beta_0} \sum_{I_j \leq m_1 < m_2 \leq I_{j+1}-1} Q_{t,m_1} \beta_{m_1} \beta_{m_2} L_H (m_2 - m_1) Q_{t,m_2}^\intercal \qquad \text{(C.10)} \\
&\quad + \frac{t^\delta}{\beta_0} \sum_{I_j \leq m_1 < m_2 \leq I_{j+1}-1} Q_{t,m_1} \beta_{m_1} \beta_{m_2} (\Psi_{H,4,1} + \Psi_{H,4,2}) Q_{t,m_2}^\intercal. \qquad \text{(C.11)}
\end{aligned}
$$

Define $d = \frac{4}{|\ln \rho|}$. The number of pairs of $I_j \leq m_1 < m_2 < I_{j+1} - 1$ such that $\min(m_2 - m_1, m_1 - I_j) \leq d \ln t$, denoted as $N_{j,1}$, satisfies that

$$
N_{j,1} \leq 2 q d \ln t.
$$

Similarly, number of pairs of $I_j \leq m_1 < m_2 < I_{j+1} - 1$ such that $m_2 - I_j \leq d \ln t$, denoted as $N_{j,2}$, satisfies:

$$
N_{j,2} \leq d^2 \ln^2 t.
$$

For t large enough, we have:

$$\begin{aligned}
&\frac{t^\delta}{\beta_0}\mathbb{E}\left[\left\|\sum_{j=1}^{J_t-1}\sum_{I_j\le m_1<m_2\le I_{j+1}} Q_{t,m_1}\beta_{m_1}\beta_{m_2}(\Psi_{H,4,1}+\Psi_{H,4,2})Q_{t,m_2}^\intercal\right\|\right]\\
&\le \frac{t^\delta}{\beta_0}\sum_{j=1}^{J_t-1}\sum_{I_j\le m_1<m_2\le I_{j+1}}\|Q_{t,m_1}\|\|Q_{t,m_2}\|\beta_{m_1}\beta_{m_2}\\
&\qquad\left[6LK_{d_U}c_1\rho^{\min(m_2-m_1,m_1-I_j)}I_j^{-\frac{\delta}{2}}(k_1\ln I_j+k_2)+L^2c_1^2C_{w,2}\rho^{m_2-I_j}\right]\\
&\le \frac{t^\delta}{\beta_0}\sum_{j=1}^{J_t-1}\sum_{I_j\le m_1<m_2\le I_{j+1}}\|Q_{t,m_1}\|\|Q_{t,m_2}\|\beta_{m_1}\beta_{m_2}\left[12LK_{d_U}c_1t^{-3}I_j^{-\frac{\delta}{2}}\max(k_1,k_2)\ln I_j+L^2c_1^2C_{w,2}t^{-3}\right]\\
&\quad+\frac{t^\delta}{\beta_0}\sum_{j=1}^{J_t}\|Q_{t,I_{j+1}}\|^2\beta_{I_j}^2(12LK_{d_U}c_1dqI_j^{-\frac{\delta}{2}}\ln t+L^2c_1^2C_{w,2}d^2\ln^2 t)\\
&\le \frac{2}{\psi\beta_0}(12LK_{d_U}c_1t^{-\frac{\delta}{2}}\ln t+L^2c_1^2t^{-\delta_q}\ln^2 t)\to 0,
\end{aligned}\tag{C.12}$$

where the last inequality follows from Lemma SM.9. Therefore, plug in (C.12) into (C.10) and (C.11), as $t\to\infty$, we have:

$$\begin{aligned}
&\mathbb{E}\left[\frac{t^\delta}{\beta_0}\left\|\sum_{1\le j\le J_t-1}\mathbb{E}[\sum_{I_j\le m_1<m_2\le I_{j+1}-1} Q_{t,m_1}\beta_{m_1}\beta_{m_2}\widetilde{H}(W_{m_1},\xi_{I_j})\widetilde{H}(W_{m_2},\xi_{I_j})^\intercal Q_{t,m_2}^\intercal|\mathcal{F}_{I_j}]\right\|\right]\\
&\to \frac{t^\delta}{\beta_0}\left\|\sum_{1\le j\le J_t-1}\sum_{I_j\le m_1<m_2\le I_{j+1}-1}Q_{t,m_1}\beta_{m_1}\beta_{m_2}L_H(m_2-m_1)Q_{t,m_2}^\intercal\right\|\\
&\le \sum_{1\le j\le J_t-1}\frac{t^\delta}{\beta_0}\sum_{I_j\le m_1<m_2\le I_{j+1}-1}\|Q_{t,m_1}\|\|Q_{t,m_2}\|\beta_{m_1}\beta_{m_2}c_12L^2K_{d_U}\rho^{m_2-m_1}\\
&\le \sum_{1\le j\le J_t-1}\frac{t^\delta}{\beta_0}\sum_{I_j\le m_1\le I_{j+1}-1}\|Q_{t,I_{j+1}}\|^2\beta_{I_j}^2c_12L^2K_{d_U}\frac{1}{1-\rho}\\
&\le t^\delta c_12L^2K_{d_U}\frac{2}{2\psi\beta_0}\frac{1}{1-\rho}t^{\delta-2\delta}\\
&=2c_1L^2K_{d_U}\frac{1}{\psi\beta_0}\frac{1}{1-\rho},
\end{aligned}$$

is bounded from above by a constant that does not depend on t. Here the last inequality follows from Lemma SM.9. Therefore, the sum

$$\frac{t^\delta}{\beta_0}\sum_{1\le j\le J_t-1}\sum_{I_j\le m_1<m_2\le I_{j+1}-1}Q_{t,m_1}\beta_{m_1}\beta_{m_2}L_H(m_2-m_1)Q_{t,m_2}^\intercal$$

exists with its max-eigenvalue norm bounded by $2c_1L^2K_{d_U}\frac{1}{\psi}\frac{1}{1-\rho}$.

Define

$$\Sigma_{H,t}:=\frac{t^\delta}{\beta_0}\sum_{1\le j\le J_t-1}\sum_{I_j\le m<I_{j+1}}Q_{t,m}\beta_m^2L_H(0)Q_{t,m}^\intercal$$

$$+\frac{t^\delta}{\beta_0}\sum_{1\le j\le J_t-1}\sum_{I_j\le m_1<m_2\le I_{j+1}-1} Q_{t,m_1}\beta_{m_1}\beta_{m_2}L_H(m_2-m_1)Q_{t,m_2}^\intercal$$
$$+\frac{t^\delta}{\beta_0}\sum_{1\le j\le J_t-1}\sum_{I_j\le m_1<m_2\le I_{j+1}-1} Q_{t,m_2}\beta_{m_1}\beta_{m_2}L_H(m_2-m_1)^\intercal Q_{t,m_1}^\intercal.$$

By (C.12), we have that:

$$\left\|\sum_{j=1}^{J_t}\mathbb{E}[X_{t,j}(X_{t,j})^\intercal]-\Sigma_{H,t}\right\|=o(1). \tag{C.13}$$

Next, we show in Lemma SM.13 that as $t\to\infty$, we have:

$$\|\Sigma_{H,t}-\Sigma_H\|\to 0,$$

where

$$\Sigma_H\coloneqq \beta_0\int_0^\infty \exp(mA_{22})\bar{L}_H\exp(mA_{22}^\intercal)\,\mathrm{d}m$$

is defined in the statement (ii) of Theorem 2.

Denote

$$\Sigma_{H,t}^0\coloneqq\frac{t^\delta}{\beta_0}\sum_{1\le j\le J_t-1}\sum_{I_j\le m\le I_{j+1}-1} Q_{t,m}\beta_m^2L_H(0)Q_{t,m}^\intercal,$$
$$\Sigma_{H,t}^1\coloneqq\frac{t^\delta}{\beta_0}\sum_{1\le j\le J_t-1}\sum_{I_j\le m_1<m_2\le I_{j+1}-1} Q_{t,m_1}\beta_{m_1}\beta_{m_2}L_H(m_2-m_1)Q_{t,m_2}^\intercal.$$

and

$$\Sigma_H^0\coloneqq\int_0^\infty \exp^{mA_{22}}L(0)\exp(mA_{22}^\intercal)\,\mathrm{d}m\,,$$
$$\Sigma_H^1\coloneqq\int_0^\infty \exp^{mA_{22}}\sum_{s\ge1}L(s)\exp(mA_{22}^\intercal)\,\mathrm{d}m\,.$$

By definition, we know that $\Sigma_H=\Sigma_H^0+\Sigma_H^1+(\Sigma_H^1)^\intercal$. Similarly, $\Sigma_{H,t}=\Sigma_{H,t}^0+\Sigma_{H,t}^1+(\Sigma_{H,t}^1)^\intercal$. Lemma SM.13 below validates that $\|\Sigma_H-\Sigma_{H,t}\|\to0$ by showing that $\|\Sigma_H^0-\Sigma_{H,t}^0\|\to0$ and $\|\Sigma_H^1-\Sigma_{H,t}^1\|\to0$ separately, as $t\to\infty$.

Lemma SM.13. *As $t\to\infty$, we have:*

$$\|\Sigma_{H,t}^0-\Sigma_H^0\|\to0\quad\textit{and}\quad\|\Sigma_{H,t}^1-\Sigma_H^1\|\to0.$$

Proof of Lemma SM.13. First, we prove that $\|\Sigma_{H,t}^0-\Sigma_H^0\|\to0$. Let $d=\frac{2}{\psi\beta_0}$.

By definition, we know that

$$\Sigma_{H,t}^0\coloneqq\frac{t^\delta}{\beta_0}\sum_{1\le j\le J_t-1}\sum_{I_j\le m\le I_{j+1}-1} Q_{t,m}\beta_m^2L_H(0)Q_{t,m}^\intercal$$

$$
\begin{aligned}
&= \beta_0 t^{-\delta} \sum_{I_1 \le m \le t-1} Q_{t,m} L_H(0) Q_{t,m}^\intercal \\
&= \underbrace{\beta_0 t^{-\delta} \sum_{I_t \le m \le t-dt^\delta \ln t} Q_{t,m} L_H(0) Q_{t,m}^\intercal}_{\Psi_{\Sigma,0}} + \underbrace{\beta_0 t^{-\delta} \sum_{t-dt^\delta \ln t < m \le t-1} Q_{t,m} L_H(0) Q_{t,m}^\intercal}_{\Psi_{\Sigma,1}}
\end{aligned}
$$

By definition,

$$
Q_{t,m} = \prod_{l=m}^{t-1} (I + \beta_l A_{22}) \le \exp\left(\sum_{l=m}^{t-1} \beta_l A_{22} \right) \le \exp\left(\beta_0 (m-t) t^{-\delta} A_{22} \right). \tag{C.14}
$$

Then, for any $m < t - dt^\delta \ln t$, we have that $\|Q_{t,m}\| \le \exp(-\beta_0 \psi d \ln t) \le t^{-2}$. Consequently,

$$
\begin{aligned}
\|\Psi_{\Sigma,0}\| &\le \beta_0 t^{-\delta} \sum_{I_t \le m \le t-dt^\delta \ln t} t^{-4} \|L_H(0)\| \\
&\le \beta_0 \|L_H(0)\| t^{-3-\delta} \to 0.
\end{aligned} \tag{C.15}
$$

Notice that $\delta < 1$. For any arbitrary $c > 0$, for t large enough, and for any $l \in (t - dt^\delta \ln t, t)$, we have that:

$$
I + \beta_l A_{22} \ge \exp(-(1+c)\beta_l A_{22}).
$$

It follows that for t large enough, and for any $m \in (t - dt^\delta \ln t, t)$

$$
\begin{aligned}
Q_{t,m} &= \prod_{l=m}^{t-1} (I + \beta_l A_{22}) \\
&\ge \prod_{l=m}^{t-1} \exp\big(-(1+c)\beta_l A_{22}\big) \\
&= \exp\left(-(1+c) \sum_{l=m}^{t-1} \beta_l A_{22} \right) \\
&\ge \exp\left((1+c)^2 \beta_0 (t-m) t^{-\delta} A_{22} \right),
\end{aligned} \tag{C.16}
$$

where the last inequality follow from the fact that $\beta_l \le (1+c)\beta_t = (1+c)\beta_0 t^{-\delta}$ for t large enough.

By (C.14), we know that

$$
\begin{aligned}
\Psi_{\Sigma,1} &\le \beta_0 t^{-\delta} \sum_{t-dt^\delta \ln t \le m \le t-1} \exp\left(-\beta_0 (m-t) t^{-\delta} A_{22}\right) L_H(0) \exp\left(\beta_0 (m-t) t^{-\delta} A_{22}^\intercal\right) \\
&\le \beta_0 t^{-\delta} \int_1^{dt^\delta \ln t} \exp\left(\beta_0 x t^{-\delta} A_{22}\right) L_H(0) \exp\left(\beta_0 x t^{-\delta} A_{22}^\intercal\right) \mathrm{d}x \\
&= \int_0^{d \ln t} \exp\left(m A_{22}\right) L_H(0) \exp\left(m A_{22}^\intercal\right) \mathrm{d}m \\
&\le \int_0^{\infty} \exp\left(m A_{22}\right) L_H(0) \exp\left(m A_{22}^\intercal\right) \mathrm{d}m \\
&= \Sigma_H^0,
\end{aligned}
$$

where the first equality in the above follows from change of variables by setting $m = xt^{-\delta}$.

Similarly, by (C.16), for t large enough, we have:

$$
\begin{aligned}
\Psi_{\Sigma,1} &\geq \beta_0 t^{-\delta} \sum_{t-dt^{\delta}\ln t \leq m \leq t-1} \exp\left(-\beta_0(m-t)t^{-\delta}A_{22}\right) L_H(0) \exp\left(\beta_0(m-t)t^{-\delta}A_{22}^{\intercal}\right) \\
&\geq \beta_0 t^{-\delta} \int_0^{dt^{\delta}\ln t - 1} \exp\left((1+c)^2\beta_0 x t^{-\delta}A_{22}\right) L_H(0) \exp\left(\beta_0 x t^{-\delta} A_{22}^{\intercal}\right) \mathrm{d}x \\
&= \frac{1}{(1+c)^2} \int_0^{\frac{1}{(1+c)^2}(d\ln t - \frac{1}{t^{\delta}})} \exp\left(mA_{22}\right) L_H(0) \exp\left(mA_{22}^{\intercal}\right) \mathrm{d}m \\
&\geq \frac{1}{(1+c)^3} \int_0^{\infty} \exp\left(mA_{22}\right) L_H(0) \exp\left(mA_{22}^{\intercal}\right) \mathrm{d}m \\
&= \frac{1}{(1+c)^3} \Sigma_H^0,
\end{aligned}
$$

where the first equality in the above follows from change of variables by setting $m = (1+c)^2 x t^{-\delta}$. Since c can be arbitrarily close to 0, as $t \to \infty$, we have that:

$$
\|\Psi_{\Sigma,1} - \Sigma_H^0\| \to 0. \tag{C.17}
$$

Combining (C.15) and (C.17), as $t \to 0$, we have:

$$
\|\Sigma_{H,t}^0 - \Sigma_H^0\| \to 0.
$$

For the analysis of $\Sigma_{H,t}^1$, by re-arranging the terms, we have:

$$
\begin{aligned}
\Sigma_{H,t}^1 &= \frac{t^{\delta}}{\beta_0} \sum_{1\leq j \leq J_t - 1} \sum_{I_j \leq m_1 < m_2 \leq I_{j+1}-1} Q_{t,m_1} \beta_{m_1} \beta_{m_2} L_H(m_2 - m_1) Q_{t,m_2}^{\intercal} \\
&= \sum_{s=1}^{q-1} \sum_{1\leq j \leq J_t - 1} \sum_{I_j \leq m_1 \leq I_{j+1}-s-1} Q_{t,m_1} L_H(s) Q_{t,m_1+s}^{\intercal} \\
&= \underbrace{\sum_{1\leq s \leq d\ln t} \sum_{1\leq j \leq J_t - 1} \sum_{I_j \leq m_1 \leq I_{j+1}-s-1} Q_{t,m_1} L_H(s) Q_{t,m_1+s}^{\intercal}}_{\Psi_{\Sigma,2}} \\
&\quad + \underbrace{\sum_{d\ln t < s \leq q-1} \sum_{1\leq j \leq J_t - 1} \sum_{I_j \leq m_1 \leq I_{j+1}-s-1} Q_{t,m_1} L_H(s) Q_{t,m_1+s}^{\intercal}}_{\Psi_{\Sigma,3}}.
\end{aligned}
$$

Similar to the proof for $\Sigma_{H,t}^0$, as $t \to \infty$, we have that:

$$
\|\Psi_{\Sigma,2}\| \to 0 \quad \text{and} \quad \|\Psi_{\Sigma,3} - \Sigma_H^1\| \to 0.
$$

Therefore, as $t \to \infty$, we have that

$$
\|\Sigma_{H,t}^1 - \Sigma_H^1\| \leq \|\Psi_{\Sigma,2}\| + \|\Psi_{\Sigma,3} - \Sigma_H^1\| \to 0. \qquad \square
$$

By Lemma SM.14 in next Subsection and the martingale law of large numbers, combining results in Lemma SM.13 and (C.13), as $t \to \infty$, we have:

$$\sum_{j=1}^{J_t-1} X^*_{t,j}(X^*_{t,j})^\intercal \to_p \mathbb{E}[\sum_{j=1}^{J_t-1} \mathbb{E}[X^*_{t,j}(X^*_{t,j})^\intercal|\mathcal{F}_{I_j}]] \to \mathbb{E}[\sum_{j=1}^{J_t-1} \mathbb{E}[X_{t,j}(X_{t,j})^\intercal|\mathcal{F}_{I_j}]] \to \Sigma_H.$$

Therefore, Condition (a) of the martingale CLT is verified.

C.5.2. Verifying Condition (b). Denote $\widetilde{H}(W_m,\xi_{I_j}) := (H(W_m,\xi_{I_j},1),\dots,H(W_m,\xi_{I_j},d_2))$, where d_2 is the dimension of $\widetilde{H}$. Denote $X_{t,j,l}$ as the l^{th} component of $X_{t,j}$. We require the following Lemma to verify Condition (b) of the martingale CLT. Without loss of generality, we can assume that $\beta_0 = 1$ and $d_U = 4$ in the rest of the proof.

LEMMA SM.14. *For all t,j large enough and $1 \le l \le d_2$, we have:*

$$\mathbb{E}[|X_{t,j,l}|^4] \le \frac{12}{\psi}\left[(C_{H,1}+C_{H,2})t^{-1}\ln t + C_{H,3}t^{-\delta+\delta_q}\ln^2 t\right],$$

where $C_{H,1}, C_{H,2}, C_{H,3}$ are defined in (C.20), (C.22) *and* (C.24)*, respectively.*

Proof of Lemma SM.14. It is easy to see that

$$\begin{aligned}\sum_{j=1}^{J_t-1}\mathbb{E}[|X_{t,j,l}|^4] &\le \sum_{j=1}^{J_t-1}\mathbb{E}[\mathbb{E}[|X_{t,j,l}|^4|\mathcal{F}_{I_j}]]\\ &\le t^{2\delta}\,\mathbb{E}\Big[\sum_{j=1}^{J_t-1}\sum_{I_j\le m_1,m_2,m_3,m_4\le I_{j+1}-1}\mathbb{E}\Big[\Big(\prod_{l=1}^{t}\|Q_{t,k_l}\|\beta_{m_l}\Big)\\ &\qquad \widetilde{H}(W_{m_1},\xi_{I_j},l)\widetilde{H}(W_{m_2},\xi_{I_j},l)\widetilde{H}(W_{m_3},\xi_{I_j},l)\widetilde{H}(W_{m_4},\xi_{I_j},l)\Big|\mathcal{F}_{I_j}\Big]\Big]. \qquad (C.18)\end{aligned}$$

Now let's assume that $I_j \le m_1 \le m_2 \le m_3 \le m_4 \le I_{j+1}-1$.

Case (i). Suppose $m_4 - m_3 > d\ln t$, $d \ge \frac{\delta+3}{|\ln\rho|}$. Then, by Lemma SM.1, we have:

$$\left|\mathbb{E}[\widetilde{H}(W_{m_4},\xi_{I_j},l)|\mathcal{F}_{m_3+1}]\right| \le c_1\rho^{d\ln t}L(1+\|W_{m_3}\|) \le Lc_1t^{-3-\delta}(1+\|W_{m_3}\|).$$

Therefore,

$$\begin{aligned}&\left|\mathbb{E}\left[\widetilde{H}(W_{m_1},\xi_{I_j},l)\widetilde{H}(W_{m_2},\xi_{I_j},l)\widetilde{H}(W_{m_3},\xi_{I_j},l)\widetilde{H}(W_{m_4},\xi_{I_j},l)\big|\mathcal{F}_{I_j}\right]\right|\\ &\le \mathbb{E}\left[\left|\mathbb{E}\Big[\widetilde{H}(W_{m_4},\xi_{I_j},l)|W_{m_3+1}\Big]\right|\cdot\left|\widetilde{H}(W_{m_1},\xi_{I_j},l)\widetilde{H}(W_{m_2},\xi_{I_j},l)\widetilde{H}(W_{m_3},\xi_{I_j},l)\right|\Big|\mathcal{F}_{I_j}\right]\\ &\le Lc_1t^{-3-\delta}\,\mathbb{E}\Big[(1+\|W_{m_3}\|)\Big|\widetilde{H}(W_{m_1},\xi_{I_j},l)\widetilde{H}(W_{m_2},\xi_{I_j},l)\widetilde{H}(W_{m_3},\xi_{I_j},l)\Big|\Big|\mathcal{F}_{I_j}\Big]\\ &\le 16L^4c_1(c_1(1+\|W_{I_j}\|^4)+K_{d_U})t^{-3-\delta},\end{aligned}$$

where the last inequality follows from Lemma SM.1 and Lipschitz conditions of Assumption 2.

Let S_1 be that:

$$
\begin{aligned}
S_1 &:= t^{2\delta}\Big|\mathbb{E}\Big[\sum_{j=1}^{J_t-1}\sum_{\substack{I_j\le m_3+d\ln t<m_4\\ I_j\le m_1\le m_2\le m_3\le m_4\le I_{j+1}-1}}\mathbb{E}\Big[\Big(\prod_{l'=1}^{4}\|Q_{t,m_{l'}}\|\beta_{m_{l'}}\Big)\\
&\qquad \widetilde{H}(W_{m_1},\xi_{I_j},l)\widetilde{H}(W_{m_2},\xi_{I_j},l)\widetilde{H}(W_{m_3},\xi_{I_j},l)\widetilde{H}(W_{m_4},\xi_{I_j},l)\Big|\mathcal{F}_{I_j}\Big]\Big]\Big|\\
&\le t^{2\delta}16L^4c_1(c_1C_{w,4}+K_{d_U})t^{-3-\delta}q^4\sum_{j=1}^{J_t-1}Q_{t,I_{j+1}}^4\beta_{I_j}^4\\
&\le \frac{C_{H,1}}{2\psi}t^{-3+\delta+4\delta_q+\delta-4\delta}\ln t\\
&\le \frac{C_{H,1}}{2\psi}t^{-1}\ln t, \qquad\qquad (C.19)
\end{aligned}
$$

where the second inequality holds because of and Lemma SM.9, and

$$
C_{H,1} := 16L^4c_1(c_1C_{w,4}+K_{d_U}). \qquad (C.20)
$$

Case (ii). Suppose $m_2-m_1>d\ln t$ and $m_1-I_j>d\ln t$. Then,

$$
\begin{aligned}
&\mathbb{E}[\widetilde{H}(w_{m_4},\xi_{I_j},l)\widetilde{H}(w_{m_3},\xi_{I_j},l)\widetilde{H}(w_{m_2},\xi_{I_j},l)|\mathcal{F}_{m_1+1}]\\
&= \int_{\mathcal{W}\times\mathcal{W}\times\mathcal{W}}\widetilde{H}(w_{m_4},\xi_{I_j},l)\widetilde{H}(w_{m_3},\xi_{I_j},l)\widetilde{H}(w_{m_2},\xi_{I_j},l)\\
&\qquad \mathbb{P}(w_{m_4}|w_{m_3})\,\mathbb{P}(w_{m_3}|w_{m_2})\,\mathbb{P}(w_{m_2}|w_{m_1})\,\mathrm{d}w_{m_4}\,\mathrm{d}w_{m_3}\,\mathrm{d}w_{m_2}\\
&= \Psi_{H,5,1}+\Psi_{H,5,2}
\end{aligned}
$$

where

$$
\begin{aligned}
\Psi_{H,5,1} := \int_{\mathcal{W}\times\mathcal{W}\times\mathcal{W}}&\widetilde{H}(w_{m_4},\xi_{I_j},l)\widetilde{H}(w_{m_3},\xi_{I_j},l)\widetilde{H}(w_{m_2},\xi_{I_j},l)\\
&\mathbb{P}(w_{m_4}|w_{m_3})\,\mathbb{P}(w_{m_3}|w_{m_2})\nu(w_{m_2})\,\mathrm{d}w_{m_4}\,\mathrm{d}w_{m_3}\,\mathrm{d}w_{m_2}\,,
\end{aligned}
$$

and

$$
\begin{aligned}
\Psi_{H,5,2} := &\widetilde{H}(w_{m_4},\xi_{I_j},l)\widetilde{H}(w_{m_3},\xi_{I_j},l)\widetilde{H}(w_{m_2},\xi_{I_j},l)\\
&\int_{\mathcal{W}\times\mathcal{W}\times\mathcal{W}}\mathbb{P}(w_{m_4}|\mathcal{F}_{m_2+1})\,\mathbb{P}(w_{m_3}|w_{m_2})(\mathbb{P}(w_{m_2}|w_{m_1})-\nu(w_{m_2}))\,\mathrm{d}w_{m_4}\,\mathrm{d}w_{m_3}\,\mathrm{d}w_{m_2}\,.
\end{aligned}
$$

By Assumption 2 and Lemma SM.1, we have:

$$
\begin{aligned}
|\Psi_{H,5,1}| &\le \int_{\mathcal{W}\times\mathcal{W}\times\mathcal{W}}L^3(1+\|w_{m_4}\|)(1+\|w_{m_3}\|)(1+\|w_{m_2}\|)\\
&\qquad \mathbb{P}(w_{m_4}|w_{m_3})\,\mathbb{P}(w_{m_3}|w_{m_2})\nu(w_{m_2})\,\mathrm{d}w_{m_4}\,\mathrm{d}w_{m_3}\,\mathrm{d}w_{m_2}\\
&\le \frac{1}{3}\sum_{l'=2}^{4}\int_{\mathcal{W}}(1+\|w_{m_{l'}}\|)^3\nu(w_{m_{l'}})\,\mathrm{d}w_{m_{l'}}
\end{aligned}
$$

$$\leq 8L^3 K_{d_U}.$$

Therefore, $\Psi_{H,5,1}$ is a generic constant that does not depend on W_{m_1} and bounded by $8L^3K_{d_U}$. Similarly, by Assumption 2 and Lemma SM.1, we have:

$$\begin{aligned}\mathbb{E}[|\Psi_{H,5,2}||\mathcal{F}_{m_1+1}] &\leq 8L^3(c_1+K_{d_U})^2 c_1 \rho^{m_2-m_1}(1+||W_{m_1}||^3)\\ &\leq 8L^3(c_1+K_{d_U})^2(1+||W_{m_1}||^3)c_1 t^{-3-\delta}.\end{aligned}$$

Therefore,

$$\begin{aligned}&\left|\mathbb{E}\left[\widetilde{H}(W_{m_1},\xi_{I_j},l)\widetilde{H}(W_{m_2},\xi_{I_j},l)\widetilde{H}(W_{m_3},\xi_{I_j},l)\widetilde{H}(W_{m_4},\xi_{I_j},l)\middle|\mathcal{F}_{I_j}\right]\right|\\ &\leq 8L^3(c_1+K_{d_U})^2 c_1 t^{-3-\delta}\,\mathbb{E}\left[(1+\|W_{m_1}\|^3)\left|\widetilde{H}(W_{m_1},\xi_{I_j},l)\right|\middle|\mathcal{F}_{I_j}\right]+|\Psi_{H,5,1}|\left|\mathbb{E}\left[\widetilde{H}(W_{m_1},\xi_{I_j},l)\middle|\mathcal{F}_{I_j}\right]\right|\\ &\leq 16L^4(c_1+K_{d_U})^3c_1(1+\|W_{I_j}\|^4)t^{-3-\delta}+8L^4K_{d_U}c_1\rho^{m_1-I_j}(1+||W_{I_j}||)\\ &\leq \left[16L^4(c_1+K_{d_U})^3c_1(1+||W_{I_j}||^4)+8L^4K_{d_U}c_1(1+||W_{I_j}||)\right]t^{-3-\delta},\end{aligned}$$

where the second last inequality follows from Lemma SM.1.

Similar to Case (i) above, we have

$$\begin{aligned}S_2 &:= t^{2\delta}\Bigg|\mathbb{E}\Bigg[\sum_{j=1}^{J_t-1}\sum_{\substack{m_2-m_1>d\ln t\\ m_1-I_j>d\ln t\\ I_j\leq m_1\leq m_2\leq m_3\leq m_4\leq I_{j+1}-1}}\mathbb{E}\Big[\Big(\prod_{l=1}^{t}\|Q_{t,m_l}\|\beta_{m_l}\Big)\\ &\qquad \widetilde{H}(W_{m_1},\xi_{I_j},l)\widetilde{H}(W_{m_2},\xi_{I_j},l)\widetilde{H}(W_{m_3},\xi_{I_j},l)\widetilde{H}(W_{m_4},\xi_{I_j},l)\Big|\mathcal{F}_{I_j}\Big]\Bigg]\Bigg|\\ &\leq t^{2\delta}C_{H,2}t^{-3-\delta}q^4\sum_{j=1}^{J_t-1}\|Q_{t,I_{j+1}}\|^4\beta_{I_j}^4\\ &\leq \frac{C_{H,2}}{2\psi}t^{-3+\delta+4\delta_q+\delta-4\delta}\ln t\\ &\leq \frac{C_{H,2}}{2\psi}t^{-1}\ln t,\end{aligned}\tag{C.21}$$

where the second to last inequality holds because of Lemma SM.9, and

$$C_{H,2} := 16L^4(c_1+K_{d_U})^3c_1C_{w,4}+8L^4K_{d_U}c_1C_{w,4}.\tag{C.22}$$

Case (iii). Consider the complement of Case (i) and Case (ii); i.e., the following event

$$\begin{aligned}\mathcal{M}_j := &\left\{(m_1,m_2,m_3,m_4)\middle|\, I_j\leq m_1\leq m_2\leq m_3\leq m_4<I_{j+1},\, m_4-m_3\leq d\ln t,\, m_2-m_1\leq d\ln t\right\}\\ &\bigcup\left\{(m_1,m_2,m_3,m_4)\middle|\, I_j\leq m_1\leq m_2\leq m_3\leq m_4<I_{j+1},\, m_4-m_3\leq d\ln t,\, m_1-I_j\leq d\ln t\right\}.\end{aligned}$$

Let N_j be the cardinality of $\mathcal{M}_j$. Then, we have:

$$N_j\leq 2d^2q^2\ln^2 t.$$

By Lemma SM.1,

$$\begin{aligned}
&\left|\mathbb{E}\left[\widetilde{H}(W_{m_1},\xi_{I_j},l)\widetilde{H}(W_{m_2},\xi_{I_j},l)\widetilde{H}(W_{m_3},\xi_{I_j},l)\widetilde{H}(W_{m_4},\xi_{I_j},l)\middle|\mathcal{F}_{I_j}\right]\right|\\
&\le \frac{1}{4}\mathbb{E}\left[\sum_{l'=1}^{4}\left|\widetilde{H}(W_{m_{l'}},\xi_{I_j},l)\right|^4\middle|\mathcal{F}_{I_j}\right]\\
&\le 16L^4(\max(c_1,1)(1+||W_{I_j}||^4)+K_{d_U}).
\end{aligned}$$

Therefore,

$$\begin{aligned}
S_3 &:= t^{2\delta}\left|\mathbb{E}\left[\sum_{j=1}^{J_t-1}\sum_{(m_1,m_2,m_3,m_4)\in\mathcal{M}_j}\mathbb{E}\Big[\Big(\prod_{l=1}^{t}\|Q_{t,m_l}\|\beta_{m_l}\Big)\right.\right.\\
&\qquad\left.\left.\widetilde{H}(W_{m_1},\xi_{I_j},l)\widetilde{H}(W_{m_2},\xi_{I_j},l)\widetilde{H}(W_{m_3},\xi_{I_j},l)\widetilde{H}(W_{m_4},\xi_{I_j},l)\middle|\mathcal{F}_{I_j}\Big]\right]\right|\\
&\le t^{2\delta}16L^4(\max(c_1,1)C_{w,4}+K_{d_U})N_j\sum_{j=1}^{J_t}\|Q_{t,I_{j+1}}\|^4\beta_{I_j}^4\\
&\le \frac{2C_{H,3}}{4\psi}t^{2\delta+\delta-4\delta+2\delta_q-\delta_q}\ln^2 t\\
&= \frac{C_{H,3}}{2\psi}t^{-\delta+\delta_q}\ln^2 t,
\end{aligned}\tag{C.23}$$

where the second to last inequality follows from Lemma SM.9 with

$$C_{H,3} := 32L^4(\max(c_1,1)C_{w,4}+K_{d_U})d^2. \tag{C.24}$$

Combining (C.19),(C.21) and (C.23), for (C.18), we have that:

$$\begin{aligned}
\sum_{j=1}^{J_t-1}\mathbb{E}[|X_{t,j,l}|^4] &\le t^{2\delta}\,\mathbb{E}\left[\sum_{j=1}^{J_t-1}\sum_{I_j\le m_1,m_2,m_3,m_4\le I_{j+1}-1}\mathbb{E}\Big[\Big(\prod_{l=1}^{t}\|Q_{t,m_l}\|\beta_{m_l}\Big)\right.\\
&\qquad\left.\widetilde{H}(W_{m_1},\xi_{I_j},l)\widetilde{H}(W_{m_2},\xi_{I_j},l)\widetilde{H}(W_{m_3},\xi_{I_j},l)\widetilde{H}(W_{m_4},\xi_{I_j},l)\middle|\mathcal{F}_{I_j}\Big]\right]\\
&\le 4!(S_1+S_2+S_3)\\
&\le \frac{24}{2\psi}((C_{H,1}+C_{H,2})t^{-1}\ln t + C_{H,3}t^{-\delta+\delta_q}\ln^2 t).
\end{aligned}\tag{C.25}$$

By (C.25), we have that $\sum_{j=1}^{J_t-1}\mathbb{E}[|X_{t,j,l}|^4]\to 0$ as $t\to\infty$. □

With Lemma SM.14, we now proceed to verify condition (b) for the martingale CLT.

Define $X_{t,\max} := \max_{1\le j\le J_t-1}\|X_{t,j}\|$, and $X^*_{t,\max} := \max_{1\le j\le J_t-1}\|X^*_{t,j}\|$. It is easy to know that

$$\|\mathbb{E}[X^*_{t,\max}]-\mathbb{E}[X_{t,\max}]\| \le \mathbb{E}\big[\max_{1\le j\le J_t-1}\|X_{t,j}-X^*_{t,j}\|\big] \le \mathbb{E}\left[\sum_{j=1}^{J_t-1}\|X_{t,j}-X^*_{t,j}\|\right]$$

By Lemma SM.12, we know that

$$\mathbb{E}\left[\sum_{j=1}^{J_t-1}\left\|X_{t,j}-X^*_{t,j}\right\|\right] = \sum_{j=1}^{J_t-1}\mathbb{E}\big[\|\mathbb{E}[X_{t,j}|\mathcal{F}_{I_j}]\|\big]$$

$$\leq c_1 t^{-\frac{\delta}{2}} + \frac{4(1+\delta-\delta_q)}{\psi|\ln\rho|} t^{-\delta_q+\frac{\delta}{2}} \ln t \to 0.$$

Therefore, it suffices to show that $\mathbb{E}[X_{t,\max}] \to 0$.

It is easy to know that

$$\begin{aligned}
\mathbb{E}[X_{t,\max}] &\leq \mathbb{E}[X_{t,\max}1(X_{t,\max} < M)] + \mathbb{E}[X_{t,\max}1(X_{t,\max} \geq M)] \\
&\leq M + \int_{x\geq M} \mathbb{P}\left(\max_{1\leq j\leq J_t-1} \|X_{t,j}\| \geq x\right) \mathrm{d}x \\
&\leq M + \sum_{1\leq j\leq J_t-1} \int_{x\geq M} \mathbb{P}(\|X_{t,j}\| \geq x) \,\mathrm{d}x \\
&\leq M + \int_{x\geq M} \frac{1}{x^4} \sum_{1\leq j\leq J_t-1} \mathbb{E}[\|X_{t,j}\|^4] \\
&= M + \frac{1}{3M^3} \sum_{1\leq j\leq J_t-1} \mathbb{E}[\|X_{t,j}\|^4],
\end{aligned}$$

where the fourth inequality follows from Markov's inequality. It suffices to show that there exists a constant $c_3 > 0$ such that $\sum_{1\leq j\leq J_t-1} \mathbb{E}[\|X_{t,j}\|^4] < t^{-c_3}$, then M can be chosen as $t^{-\frac{1}{4}c_3}$, and consequently we have $\mathbb{E}[X_{t,\max}] \leq \frac{4}{3} t^{-\frac{1}{4}c_3} \to 0$.

By Lemma SM.14,

$$\begin{aligned}
\sum_{1\leq j\leq J_t-1} \mathbb{E}[\|X_{t,j}\|^4] &\leq d_2 \sum_{1\leq j\leq J_t-1} \sum_{1\leq l\leq d_2} \mathbb{E}[|X_{t,j,l}|^4] \\
&\leq \frac{24 d_2^2}{4\psi} \left[(C_{H,1}+C_{H,2}) t^{-1} \ln t + C_{H,3} t^{-\delta+\delta_q} \ln^2 t\right] \to 0,
\end{aligned}$$

as $t \to \infty$. Therefore, c_3 can be chosen as any positive number such that $c_3 < \delta - \delta_q$. Therefore, the martingale CLT holds for the martingale-difference array $X^*_{t,j}$, $1 \leq j \leq J_t - 1$. It the follows from Lemmas SM.10 and SM.11 that as $t \to \infty$,

$$\frac{1}{\sqrt{\beta_t}} (\xi_t - \xi^*) \rightsquigarrow \mathcal{N}(0, \Sigma_H).$$

C.6. Asymptotics of λ_t

When $\delta = \kappa$, we can treat (λ_t, ξ_t) together as a parameter, and the result of the theorem applies directly.

When $\delta > \kappa$, by Proposition 1, $\mathbb{E}[\|\xi_t - \xi^*\|^2] \leq t^{-\delta}(k_1 \ln T + k_2)$, for any $t \geq T$. Therefore, we can approximate $G(W_t, \lambda_t, \xi_t)$ by $G(W_t, \lambda_t, \xi^*)$ while the approximation error is negligible compared to $t^{-\frac{\kappa}{2}}$ since $\kappa < \delta$. Therefore, such approximation does not affect to apply our proof for λ_t. Hence,

$$\frac{1}{\sqrt{\alpha_t}} (\lambda_t - \lambda^*) \rightsquigarrow \mathcal{N}(0, \Sigma_G),$$

as $t \to \infty$, where

$$\Sigma_G := \int_m^\infty \exp(mA_{11})\, \bar{L}_G \exp(mA_{11}^\intercal)\, \mathrm{d}m\,.$$

D. Analysis of IV-Q-Learning

We now prove the theoretical properties of IV-Q-Learning—Corollary 1, Proposition 2, and Proposition 3. We use the following notation throughout this section:

$$\begin{aligned}
W_t &= (S_t, A_t, R_t, Z_t, S_{t+1}, R_t Z_t, \|S_t\|\|Z_t\|, \|S_{t+1}\|\|Z_t\|, Z_t Z_t^\intercal),\\
w &= (s, a, r, z, s', rz, \|s\|\|z\|, \|s'\|\|z\|, zz^\intercal),\\
G_{\mathrm{Q}}(w, \theta, \Gamma) &= (r + \gamma \max_{a' \in \mathcal{A}} \phi^\intercal(s', a)\theta - \phi^\intercal(s, a)\theta)\Gamma z,\\
H_{\mathrm{Q}}(w, \Gamma) &= (\phi(s, a) - \Gamma z) z^\intercal.
\end{aligned}$$

D.1. Proof of Corollary 1

The three parts of Corollary 1 are direct results of Proposition 1, Theorem 1, and Theorem 2, respectively, so it suffices to verify their conditions, namely, Assumptions 1–3. Note that Assumptions 1 is implied by Assumptions 4, while Assumption 3 is implied by Assumptions 5(iii) and 6. We next use a sequence of lemmas to verify Assumption 2.

Lemma SM.15. *For all $s' \in \mathcal{S}$ and $\theta, \tilde{\theta} \in \mathbb{R}^p$,*

$$\left| \max_{a' \in \mathcal{A}} \phi^\intercal(s', a')\theta - \max_{a' \in \mathcal{A}} \phi^\intercal(s', a')\tilde{\theta} \right| \le \max_{a' \in \mathcal{A}} |\phi^\intercal(s', a')(\theta - \tilde{\theta})|.$$

Proof of Lemma SM.15. Let $a_1 \in \arg\max_{a' \in \mathcal{A}} \phi^\intercal(s', a')\theta$. Then,

$$\begin{aligned}
\max_{a' \in \mathcal{A}} \phi^\intercal(s', a')\theta - \max_{a' \in \mathcal{A}} \phi^\intercal(s', a')\tilde{\theta} &= \phi^\intercal(s', a_1)\theta - \max_{a' \in \mathcal{A}} \phi^\intercal(s', a')\tilde{\theta}\\
&\le \phi^\intercal(s', a_1)(\theta - \tilde{\theta})\\
&\le \max_{a' \in \mathcal{A}} \phi^\intercal(s', a')(\theta - \tilde{\theta}).
\end{aligned}$$

Likewise, it can be shown that

$$\max_{a' \in \mathcal{A}} \phi^\intercal(s', a')\theta - \max_{a' \in \mathcal{A}} \phi^\intercal(s', a')\tilde{\theta} \ge \min_{a' \in \mathcal{A}} \phi^\intercal(s', a')(\theta - \tilde{\theta}).$$

Hence,

$$\begin{aligned}
\left| \max_{a' \in \mathcal{A}} \phi^\intercal(s', a')\theta - \max_{a' \in \mathcal{A}} \phi^\intercal(s', a')\tilde{\theta} \right| &\le \max\left(\left| \min_{a' \in \mathcal{A}} \phi^\intercal(s', a')(\theta - \tilde{\theta}) \right|, \left| \max_{a' \in \mathcal{A}} \phi^\intercal(s', a')(\theta - \tilde{\theta}) \right| \right)\\
&= \max_{a' \in \mathcal{A}} |\phi^\intercal(s', a')(\theta - \tilde{\theta})|. \qquad \square
\end{aligned}$$

Lemma SM.16. *For all $w \in \mathcal{W}$, $\theta, \tilde{\theta} \in \mathbb{R}^p$ with $\|\theta\| \le B$, and $\Gamma, \tilde{\Gamma} \in \mathbb{R}^{p \times q}$ with $\|\Gamma\|_{\mathsf{F}} \le B$, we have*

$$\begin{aligned}
\|G_{\mathrm{Q}}(w, \theta, \Gamma) - G_{\mathrm{Q}}(w, \tilde{\theta}, \Gamma)\| &\le 2BK(\gamma + 1)(1 + \|w\|)\|\theta - \tilde{\theta}\|,\\
\|G_{\mathrm{Q}}(w, \theta, \Gamma) - G_{\mathrm{Q}}(w, \theta, \tilde{\Gamma})\| &\le (1 + 2BK(\gamma + 1))(1 + \|w\|)\|\Gamma - \tilde{\Gamma}\|_{\mathsf{F}},\\
\|G_{\mathrm{Q}}(w, \theta, \Gamma)\| &\le B(1 + 2BK(\gamma + 1))(1 + \|w\|).
\end{aligned}$$

Proof of Lemma SM.16. It follows from the definition of G_{Q} that

$$\begin{aligned}
&\|G_{\mathrm{Q}}(w,\theta,\Gamma)-G_{\mathrm{Q}}(w,\tilde{\theta},\Gamma)\| \\
=&\left\|\left[\gamma\Big(\max_{a'\in\mathcal{A}}\phi^{\intercal}(s',a')\theta-\max_{a'\in\mathcal{A}}\phi^{\intercal}(s',a')\tilde{\theta}\Big)-\Big(\phi^{\intercal}(s,a)\theta-\phi^{\intercal}(s,a)\tilde{\theta}\Big)\right]\Gamma z\right\| \\
\le&\left[\gamma\left|\max_{a'\in\mathcal{A}}\phi^{\intercal}(s',a')\theta-\max_{a'\in\mathcal{A}}\phi^{\intercal}(s',a')\tilde{\theta}\right|+\left|\phi^{\intercal}(s,a)\theta-\phi^{\intercal}(s,a)\tilde{\theta}\right|\right]\|\Gamma z\| \\
\le&\left[\gamma\max_{a'\in\mathcal{A}}|\phi^{\intercal}(s',a')(\theta-\tilde{\theta})|+|\phi^{\intercal}(s,a)(\theta-\tilde{\theta})|\right]\|\Gamma z\| \\
\le&\left[\gamma\max_{a'\in\mathcal{A}}\|\phi(s',a')\|\|\theta-\tilde{\theta}\|+\|\phi(s,a)\|\|\theta-\tilde{\theta}\|\right]\|\Gamma\|\cdot\|z\| \\
\le&\left[\gamma K(1+\|s'\|)\|\theta-\tilde{\theta}\|+K(1+\|s\|)\|\theta-\tilde{\theta}\|\right]\|\Gamma\|_{\mathsf{F}}\cdot\|z\| \\
\le& 2BK(\gamma+1)(1+\|w\|)\|\theta-\tilde{\theta}\|,
\end{aligned}$$

for all $w\in\mathcal{W}$, $\theta,\tilde{\theta}\in\mathbb{R}^p$, and $\Gamma\in\mathbb{R}^{p\times q}$ with $\|\Gamma\|_{\mathsf{F}}\le B$, where the second inequality follows from Lemma SM.15, the fourth from Assumption 5(ii), and the last from the definition of w.

Likewise,

$$\begin{aligned}
\|G_{\mathrm{Q}}(w,\theta,\Gamma)-G_{\mathrm{Q}}(w,\theta,\tilde{\Gamma})\| &= \left\|\Big(r+\gamma\max_{a'\in\mathcal{A}}\phi^{\intercal}(s',a')\theta-\phi^{\intercal}(s,a)\theta\Big)(\Gamma-\tilde{\Gamma})z\right\| \\
&\le\left|r+\Big(\gamma\max_{a'\in\mathcal{A}}\phi(s',a')-\phi(s,a)\Big)^{\intercal}\theta\right|\cdot\|\Gamma-\tilde{\Gamma}\|\cdot\|z\| \\
&\le(1+2BK(\gamma+1))(1+\|w\|)\|\Gamma-\tilde{\Gamma}\|_{\mathsf{F}},
\end{aligned}$$

for all $w\in\mathcal{W}$, $\theta\in\mathbb{R}^p$ with $\|\theta\|\le B$, and $\Gamma,\tilde{\Gamma}\in\mathbb{R}^{p\times q}$.

For $\|G_{\mathrm{Q}}(w,\theta,\Gamma)\|$, it is easy to see that

$$\begin{aligned}
\|G_{\mathrm{Q}}(w,\theta,\Gamma)\| &= \left\|\Big(r+\gamma\max_{a'\in\mathcal{A}}\phi^{\intercal}(s',a')\theta-\phi^{\intercal}(s,a)\theta\Big)\Gamma z\right\| \\
&\le\left|r+\Big(\gamma\max_{a'\in\mathcal{A}}\phi(s',a')-\phi(s,a)\Big)^{\intercal}\theta\right|\cdot B\cdot\|z\| \\
&\le B(1+2BK(\gamma+1))(1+\|w\|). \qquad \square
\end{aligned}$$

Lemma SM.17. *Let*

$$\bar{G}_{\mathrm{Q}}(\theta,\Gamma)\coloneqq\mathbb{E}_{\nu}[G_{\mathrm{Q}}(w_t,\theta,\Gamma)]=\mathbb{E}_{\nu}\left[\Big(R_t+\gamma\max_{a'\in\mathcal{A}}\phi^{\intercal}(S_{t+1},a')\theta-\phi^{\intercal}(S_t,A_t)\theta\Big)\Gamma Z_t\right]$$

Then, for all $\theta\in\mathbb{R}^p$,

$$(\theta-\theta^*)^{\intercal}(\bar{G}_{\mathrm{Q}}(\theta,\Gamma^*)-\bar{G}_{\mathrm{Q}}(\theta^*,\Gamma^*))\le-\zeta\|\theta-\theta^*\|^2.$$

Proof of Lemma SM.17. For simplicity, let $\Psi=(\theta-\theta^*)^{\intercal}(\bar{G}_{\mathrm{Q}}(\theta,\Gamma^*)-\bar{G}_{\mathrm{Q}}(\theta^*,\Gamma^*))$. Note that

$$\begin{aligned}
\Psi = &\underbrace{(\theta-\theta^*)^{\intercal}\left[\gamma\,\mathbb{E}_{\nu}\left[\Big(\max_{a'\in\mathcal{A}}\phi^{\intercal}(S_{t+1},a')\theta-\max_{a'\in\mathcal{A}}\phi^{\intercal}(S_{t+1},a')\theta^*\Big)\Gamma^*Z_t\right]\right]}_{\Psi_1} \\
&-\underbrace{(\theta-\theta^*)^{\intercal}\,\mathbb{E}_{\nu}\left[\phi^{\intercal}(S_t,A_t)(\theta-\theta^*)\Gamma^*Z_t\right]}_{\Psi_2}.
\end{aligned} \tag{D.1}$$

We first apply the Cauchy-Schwarz inequality to Ψ_1,

$$\begin{aligned}
\Psi_1 &\le \sqrt{\mathbb{E}_\nu\big[((\theta-\theta^*)^\intercal\Gamma^* Z_t)^2\big]}\cdot\sqrt{\gamma^2\,\mathbb{E}_\nu\Big[\Big(\max_{a'\in\mathcal{A}}\phi^\intercal(S_{t+1},a')\theta-\max_{a'\in\mathcal{A}}\phi^\intercal(S_{t+1},a')\theta^*\Big)^2\Big]}\\
&\le \sqrt{\mathbb{E}_\nu\big[((\theta-\theta^*)^\intercal\Gamma^* Z_t)^2\big]}\cdot\sqrt{\gamma^2\,\mathbb{E}_\nu\Big[\Big(\max_{a'\in\mathcal{A}}|\phi^\intercal(S_{t+1},a')(\theta-\theta^*)|\Big)^2\Big]}\\
&\le \sqrt{\mathbb{E}_\nu\big[((\theta-\theta^*)^\intercal\Gamma^* Z_t)^2\big]}\cdot\sqrt{\gamma^2\,\mathbb{E}_\nu\Big[\max_{a'\in\mathcal{A}}\big(\phi^\intercal(S_{t+1},a')(\theta-\theta^*)\big)^2\Big]},
\end{aligned}\tag{D.2}$$

where the second inequality follows from Lemma SM.15. We then apply Assumption 5(iv) to Ψ_2,

$$\begin{aligned}
\Psi_2 &= \mathbb{E}_\nu\big[(\theta-\theta^*)^\intercal\Gamma^* Z_t\phi^\intercal(S_t,A_t)(\theta-\theta^*)\big]\\
&= \mathbb{E}_\nu\big[(\theta-\theta^*)^\intercal\Gamma^* Z_t(\Gamma^* Z_t+\eta_t)^\intercal(\theta-\theta^*)\big]\\
&= \mathbb{E}_\nu\big[(\theta-\theta^*)^\intercal\Gamma^* Z_t(\Gamma^* Z_t)^\intercal(\theta-\theta^*)\big]\\
&= \mathbb{E}_\nu\big[((\theta-\theta^*)^\intercal\Gamma^* Z_t)^2\big],
\end{aligned}\tag{D.3}$$

where the third equality holds because

$$\mathbb{E}_\nu[\Gamma^* Z_t\eta_t^\intercal]=\mathbb{E}_\nu\big[\mathbb{E}[\Gamma^* Z_t\eta_t^\intercal|Z_t]\big]=\mathbb{E}_\nu\big[\Gamma^* Z_t\,\mathbb{E}[\eta_t^\intercal|Z_t]\big]=0.$$

Let

$$\Psi_3 := \gamma^2\,\mathbb{E}_\nu\Big[\max_{a'\in\mathcal{A}}\big(\phi^\intercal(S_{t+1},a')(\theta-\theta^*)\big)^2\Big].$$

It follows from (D.1), (D.2) and (D.3) that

$$\begin{aligned}
\Psi &\le \sqrt{\Psi_2}\cdot\sqrt{\Psi_3}-\Psi_2\\
&= \frac{\sqrt{\Psi_2}(\Psi_3-\Psi_2)}{\sqrt{\Psi_3}+\sqrt{\Psi_2}}\\
&= \frac{1}{\sqrt{\Psi_3/\Psi_2}+1}\Big[\gamma^2\,\mathbb{E}_\nu\Big[\max_{a'\in\mathcal{A}}\big(\phi^\intercal(S_{t+1},a')(\theta-\theta^*)\big)^2\Big]-\mathbb{E}_\nu\big[((\theta-\theta^*)^\intercal\Gamma^* Z_t)^2\big]\Big]\\
&\le \frac{-2\zeta\|\theta-\theta^*\|^2}{\sqrt{\Psi_3/\Psi_2}+1}\\
&\le -\zeta\|\theta-\theta^*\|^2,
\end{aligned}$$

where the second inequality follows from Assumption 5(iii). □

LEMMA SM.18. *For all $w\in\mathcal{W}$ and $\Gamma,\tilde{\Gamma}\in\mathbb{R}^{p\times q}$ with $\|\Gamma\|_{\mathsf{F}}\le B$, we have*

$$\begin{aligned}
\|H_{\mathrm{Q}}(w,\Gamma)-H_{\mathrm{Q}}(w,\tilde{\Gamma})\|_{\mathsf{F}} &\le \|w\|\|\Gamma-\tilde{\Gamma}\|_{\mathsf{F}},\\
\|H_{\mathrm{Q}}(w,\Gamma)\|_{\mathsf{F}} &\le \sqrt{p}(2K+B)\|w\|.
\end{aligned}$$

Proof of Lemma SM.18. Direct calculation yields

$$\begin{aligned}\|H_{\mathrm{Q}}(w,\Gamma)-H_{\mathrm{Q}}(w,\tilde{\Gamma})\|_{\mathsf{F}}^2 &= \|(\Gamma-\tilde{\Gamma})zz^{\intercal}\|_{\mathsf{F}}^2\\ &= \mathrm{tr}\Big(zz^{\intercal}(\Gamma-\tilde{\Gamma})^{\intercal}(\Gamma-\tilde{\Gamma})zz^{\intercal}\Big)^2\\ &= z^{\intercal}(\Gamma-\tilde{\Gamma})^{\intercal}(\Gamma-\tilde{\Gamma})z\cdot\mathrm{tr}(zz^{\intercal})\\ &= \|(\Gamma-\tilde{\Gamma})z\|^2\cdot\|z\|^2.\end{aligned}$$

Thus,

$$\|H_{\mathrm{Q}}(w,\Gamma)-H_{\mathrm{Q}}(w,\tilde{\Gamma})\|_{\mathsf{F}}\le\|\Gamma-\tilde{\Gamma}\|\cdot\|z\|^2\le\|w\|\|\Gamma-\tilde{\Gamma}\|_{\mathsf{F}},$$

for all $\Gamma,\tilde{\Gamma}\in\mathbb{R}^{p\times q}$, where the last inequality follows from the definition of w.

For $\|H_{\mathrm{Q}}(w,\Gamma)\|_{\mathsf{F}}$, it is easy to see that by Assumption 5(ii),

$$\|H_{\mathrm{Q}}(w,\Gamma)\|_{\mathsf{F}}=\|(\phi(s,a)-\Gamma z)z^{\intercal}\|_{\mathsf{F}}\le\sqrt{p}(\|z\phi(s,a)\|+B\|zz^{\intercal}\|)\le\sqrt{p}(2K+B)\|w\|. \qquad\square$$

Lemma SM.19. *Let $\psi>0$ denote the smallest eigenvalue of $\mathbb{E}_{\nu}[Z_tZ_t^{\intercal}]$, and*

$$\bar{H}_{\mathrm{Q}}(\Gamma):=\mathbb{E}_{\nu}[H_{\mathrm{Q}}(w_t,\Gamma)]=\mathbb{E}_{\nu}[(\phi(S_t,A_t)-\Gamma Z_t)Z_t^{\intercal}].$$

Then, for all $\Gamma\in\mathbb{R}^{p\times q}$,

$$\mathrm{vec}(\Gamma-\Gamma^*)^{\intercal}\mathrm{vec}(\bar{H}_{\mathrm{Q}}(\Gamma)-\bar{H}_{\mathrm{Q}}(\Gamma^*))\le-\psi\|\Gamma-\Gamma^*\|_{\mathsf{F}}^2.$$

Proof of Lemma SM.19. It follows from Assumption 5(iv) that

$$\bar{H}_{\mathrm{Q}}(\Gamma^*)=\mathbb{E}_{\nu}\big[[\phi(S_t,A_t)-\Gamma^*Z_t]Z_t^{\intercal}\big]=\mathbb{E}_{\nu}[\eta_tZ_t^{\intercal}]=0.$$

Note that $\bar{H}_{\mathrm{Q}}(\Gamma)-\bar{H}_{\mathrm{Q}}(\Gamma^*)=\mathbb{E}_{\nu}[-(\Gamma-\Gamma^*)Z_tZ_t^{\intercal}]$. Hence, for all $\Gamma\in\mathbb{R}^{p\times q}$,

$$\begin{aligned}\mathrm{vec}(\Gamma-\Gamma^*)^{\intercal}\mathrm{vec}(\bar{H}_{\mathrm{Q}}(\Gamma)-\bar{H}_{\mathrm{Q}}(\Gamma^*)) &= \mathrm{tr}\big((\Gamma-\Gamma^*)(\bar{H}_{\mathrm{Q}}(\Gamma)-\bar{H}_{\mathrm{Q}}(\Gamma^*))^{\intercal}\big)\\ &= \mathrm{tr}\big(\mathbb{E}_{\nu}[-(\Gamma-\Gamma^*)Z_tZ_t^{\intercal}(\Gamma-\Gamma^*)^{\intercal}]\big)\\ &= -\,\mathrm{tr}\big((\Gamma-\Gamma^*)\,\mathbb{E}_{\nu}[Z_tZ_t^{\intercal}](\Gamma-\Gamma^*)^{\intercal}\big).\end{aligned}\tag{D.4}$$

Let $PDP^{\intercal}$ be its eigenvalue decomposition, where P is a unitary matrix and D is the diagonal matrix whose main diagonal contains the eigenvalues of $\mathbb{E}_{\nu}[Z_tZ_t^{\intercal}]$. For simplicity, let $M=\Gamma-\Gamma^*$, and let $D^{1/2}$ denote the square root of D. Therefore,

$$\mathrm{tr}(MPDP^{\intercal}M^{\intercal})=\big\|MPD^{1/2}\big\|_{\mathsf{F}}^2=\sum_{i,j}\big(MPD^{1/2}\big)_{ij}^2=\sum_{i,j}(MP)_{ij}^2(D^{1/2})_{jj}^2\ge\psi\sum_{i,j}(MP)_{ij}^2.\tag{D.5}$$

Further,

$$\sum_{i,j}(MP)_{ij}^2=\mathrm{tr}(MPP^{\intercal}M^{\intercal})=\mathrm{tr}(MM^{\intercal})=\|M\|_{\mathsf{F}}^2=\|\Gamma-\Gamma^*\|_{\mathsf{F}}^2.\tag{D.6}$$

The proof is completed by combining (D.4), (D.5), and (D.6) $\square$

Let us now return to the proof of Corollary 1. We have verified Assumption 2(ii) via Lemmas SM.16 and SM.18, and have verified Assumption 2(iii) via Lemmas SM.17 and SM.19. Consequently, to apply the theoretical results for two-timescale SA—Proposition 1, Theorem 1, and Theorem 2—to IV-Q-Learning, it suffices to verify Assumption 2(i).

It is easy to see that

$$\begin{aligned}\mathbb{E}_\nu[G_{\mathrm{Q}}(W_t,\theta^*,\Gamma^*)] &= \mathbb{E}_\nu\left[\Big(R_t+\gamma\max_{a'\in\mathcal{A}}\phi^\intercal(S_{t+1},a')\theta^*-\phi^\intercal(S_t,A_t)\theta^*\Big)\Gamma^*Z_t\right]\\ &= \mathbb{E}_\nu[\epsilon_t\Gamma^*Z_t]+\mathbb{E}_\nu\left[\Big(r(S_t,A_t)+\gamma\max_{a'\in\mathcal{A}}\phi^\intercal(S_{t+1},a')\theta^*-\phi^\intercal(S_t,A_t)\theta^*\Big)\Gamma^*Z_t\right]\\ &= \mathbb{E}_\nu\left[\Big(r(S_t,A_t)+\gamma\max_{a'\in\mathcal{A}}\phi^\intercal(S_{t+1},a')\theta^*-\phi^\intercal(S_t,A_t)\theta^*\Big)\Gamma^*Z_t\right],\end{aligned}$$

where the last inequality follows from Assumption 5(iv).

Recall the definition of the Q^* function in (10):

$$Q^*(s,a)=\mathbb{E}_{s'\sim\mathsf{P}(\cdot|s,a)}\left[r(s,a)+\gamma\max_{a'\in\mathcal{A}}Q^*(s',a')\right],\quad \forall(s,a).$$

Hence, by the assumption that $Q^*(s,a)=\phi^\intercal(s,a)\theta^*$,

$$\phi^\intercal(S_t,A_t)\theta^*=\mathbb{E}_{S_{t+1}\sim\mathsf{P}(\cdot|S_t,A_t)}\left[r(S_t,A_t)+\gamma\max_{a'\in\mathcal{A}}\phi^\intercal(S_{t+1},a')\theta^*\right]. \tag{D.7}$$

Multiplying Γ^*z on both sides of (D.7) and applying $\mathbb{E}_\nu$, we have:

$$\mathbb{E}_\nu[\phi^\intercal(S_t,A_t)\theta^*\Gamma^*z]=\mathbb{E}_\nu\left[\Big(r(S_t,A_t)+\gamma\max_{a'\in\mathcal{A}}\phi^\intercal(S_{t+1},a')\theta^*\Big)\Gamma^*Z_t\right].$$

Therefore, $\mathbb{E}_\nu[G_{\mathrm{Q}}(w,\theta^*,\Gamma^*)]=0$; that is, θ^* solves the equation $\bar{G}_{\mathrm{Q}}(\theta,\Gamma^*)=0$.

It is also easy to show $\mathbb{E}_\nu[H_{\mathrm{Q}}(w,\Gamma^*)]=0$, so Γ^* solves the equation $\bar{H}_{\mathrm{Q}}(\Gamma)=0$.

The uniqueness of θ^* and Γ^* are guaranteed by linear independence of $\phi(s,a)$ and full rankness of Γ^* and $\mathbb{E}_\nu[Z_tZ_t^\intercal]$ by conditions (i) and (iv) in Assumption 5. Hence, Assumption 2(i) holds.

D.2. Proof of Proposition 2

By Corollary 1, since $\alpha_t^{-\frac{1}{2}}(\theta_t-\theta^*)\rightsquigarrow\mathcal{N}(0,\Sigma_{G,\mathrm{Q}})$, $\hat{a}=\arg\max_{a\in\mathcal{A}}\phi^\intercal(s,a)\theta_t$ must be a consistent estimator for $a_0=\arg\max_{a\in\mathcal{A}}$.

By the delta method, note that $\phi(s,a)$ is twice-differentiable with respect to a around a_0. Under the null hypothesis, we have that as t approaches infinity,

$$\frac{\partial^2\phi^\intercal(s,a_0)\theta^*}{\partial a\partial a^\intercal}(\hat{a}-a_0)=-(1+o_p(1))\frac{\partial\phi^\intercal(s,a_0)}{\partial a}(\theta_t-\theta^*).$$

By Corollary 1, it follows that

$$\alpha_t^{-\frac{1}{2}}(\hat{a}-a_0)=-(1+o_p(1))\left(\frac{\partial^2\phi^\intercal(s,a_0)\theta^*}{\partial a\partial a^\intercal}\right)^{-1}\frac{\partial\phi^\intercal(s,a_0)}{\partial a}\alpha_t^{-\frac{1}{2}}(\theta_t-\theta^*)$$

$$\begin{aligned}&\rightsquigarrow \left(\frac{\partial^2\phi^\intercal(s,a_0)\theta^*}{\partial a\partial a^\intercal}\right)^{-1}\frac{\partial\phi^\intercal(s,a_0)}{\partial a}\mathcal{N}(0,\Sigma_{G,\mathrm{Q}})\\&=\mathcal{N}(0,\Omega_s),\end{aligned}$$

where

$$\Omega_s=\left(\frac{\partial^2\phi^\intercal(s,a_0)\theta^*}{\partial a\partial a^\intercal}\right)^{-1}\frac{\partial\phi^\intercal(s,a_0)}{\partial a}\Sigma_{G,\mathrm{Q}}\frac{\partial\phi(s,a_0)}{\partial a}\left(\frac{\partial^2\phi^\intercal(s,a_0)\theta^*}{\partial a\partial a^\intercal}\right)^{-1}.$$

Hence,

$$T(s)=(\hat a-a_0)^\intercal\Omega_s^{-1}(\hat a-a_0)\rightsquigarrow\chi^2(d_A).$$

D.3. Proof of Proposition 3

The proof of this proposition is a natural extension of Proposition 2. Since $\phi(s,a)$ has bounded second-order derivatives, by the delta method we have that

$$\alpha_t^{-\frac{1}{2}}(\hat\pi(s)-\pi^*(s))=(1+r_t(s))\Sigma(s)^{-1}\frac{\partial\phi^\intercal(s,a)}{\partial a}\bigg|_{a=\pi^*(s)}\alpha_t^{-\frac{1}{2}}(\theta_t-\theta^*),$$

where $r_t(s)\to 0$ uniformly as $t\to\infty$.

By Corollary 1, we know that $\alpha_t^{-\frac{1}{2}}(\theta_t-\theta^*)\rightsquigarrow\mathcal{N}(0,\Sigma_{G,\mathrm{Q}})$. Then, $\alpha_t^{-\frac{1}{2}}(\hat\pi(s)-\pi^*(s))\rightsquigarrow\mathbb{G}(s)$ with covariance function

$$\mathrm{Cov}(\mathbb{G}(s),\mathbb{G}(s'))=\Sigma(s)^{-1}\left(\frac{\partial\phi^\intercal(s,\pi^*(s))}{\partial a}\right)\bar L_{G,\mathrm{Q}}\left(\frac{\partial\phi(s',\pi^*(s'))}{\partial a}\right)\Sigma(s')^{-1}.$$

Hence,

$$T_2=\alpha_t^{-1}\int_{\mathcal{S}}\|\hat\pi(s)-\pi^*(s)\|^2\nu(s)\,\mathrm{d}s=\mathbb{E}_\nu[\|\hat\pi(s)-\pi^*(s)\|^2]\rightsquigarrow\mathbb{E}_\nu[\mathbb{G}(s)^2]=\int_{\mathcal{S}}\mathbb{G}(s)^2\nu(s)\,\mathrm{d}s\,.$$

E. Analysis of IV-SGD

In this section, we present theoretical results for Algorithm 1; i.e., IV-SGD for data from a fixed policy. Similar results for the IV-SGD algorithm for data from an interactive policy outlined in Section 3.3 can also be established, by casting it into our two-timescale SA framework. We omit the details.

Assumption SM.1. *Assume that the data $\{(X_t,Y_t,Z_t):t\geq 0\}$ are i.i.d. Suppose for all $t\geq 0$,*

$$\begin{aligned}Y_t&=X_t^\intercal\theta^*+\epsilon_t,\\X_t&=\Gamma^*Z_t+\eta_t.\end{aligned}$$

(i) There exists a constant $C>0$ such that $\mathbb{E}[\|Z_tX_t^\intercal\|^{d_U}]<C$, $\mathbb{E}[\|Z_t\epsilon_t\|^{d_U}]<C$, $\mathbb{E}[\|Z_t\eta_t^\intercal\|^{d_U}]<C$, $\mathbb{E}[\|Z_tZ_t^\intercal\|^{d_U}]<C$ for some constant $d_U>0$.

(ii) $\mathbb{E}[Z_t\epsilon_t]=0$ and $\mathbb{E}[Z_t\eta_t^\intercal]=0$.

(iii) The matrices Γ^* and $\mathbb{E}[Z_t Z_t^\intercal]$ have full rank.

COROLLARY SM.1. *Let $\{\theta_t : t \geq 0\}$ be the iterates from the projected IV-SGD. For all $t \geq 1$, let $\alpha_t = \alpha_0 t^{-\kappa}$ and $\beta_t = \beta_0 t^{-\delta}$ for some $\alpha_0 > 0$, $\beta_0 > 0$, and $0 < \kappa \leq \delta \leq 1$. Assume $\zeta\alpha_0 > 1$ in the case of $\kappa = 1$, and $\psi\beta_0 > 1$ in the case of $\delta = 1$, where ζ and ψ are the smallest eigenvalues of $\mathbb{E}_\nu[\Gamma^* Z_t Z_t^\intercal (\Gamma^*)^\intercal]$ and $\mathbb{E}_\nu[Z_t Z_t^\intercal]$, respectively.*

(i) If Assumption SM.1 holds with $d_U = 2$, then for all t large enough,

$$\mathbb{E}[\|\theta_t - \theta^*\|^2] \leq k t^{-\kappa},$$

where k is some positive constant.

(ii) Fix arbitrary $C > 0$, $\kappa_c \in [0, \kappa - 1/2)$, and $\delta_c \in [0, \delta - 1/2)$. If Assumption SM.1 holds with $d_U = 2$ and $1/2 < \kappa \leq \delta < 1$, then for all t large enough,

$$\mathbb{P}\big(\|\theta_t - \theta^*\| \leq C t^{-\kappa_c},\, \forall t \geq T\big) \geq 1 - C_\theta T^{-b_\kappa},$$

where $b_\kappa := 2\kappa - 1 - 2\kappa_c$ and C_θ is some positive constant that depends on $(\delta, \delta_c, \kappa, \kappa_c, C)$.

(iii) If Assumption SM.1 holds with $d_U = 4$ and $1/2 < \kappa < \delta < 1$, then as $t \to \infty$,

$$\alpha_t^{-1/2}(\theta_t - \theta^*) \rightsquigarrow \mathcal{N}(0, \Sigma_{G,\mathrm{SGD}}),$$

where

$$\begin{aligned}
\Sigma_{G,\mathrm{SGD}} &:= \int_0^\infty \exp\left(tA_{11,\mathrm{SGD}}\right) L_{G,\mathrm{SGD}} \exp\left(tA_{11,\mathrm{SGD}}^\intercal\right) \mathrm{d}t\,,\\
L_{G,\mathrm{SGD}} &:= \mathbb{E}[\epsilon_t^2 Z_t^\intercal (\Gamma^*)^\intercal \Gamma^* Z_t],\\
A_{11,\mathrm{SGD}} &:= -\mathbb{E}[Z_t^\intercal (\Gamma^*)^\intercal \Gamma^* Z_t].
\end{aligned}$$

Corollary SM.1 requires less conditions than Corollary 1 for IV-Q-Learning in a general MDP environment. For example, the condition (35) in Assumption 5 is specific for IV-Q-Learning and it is not required for IV-SGD. Note that in Corollary SM.1, we restrict ourselves to the setting of i.i.d. data, and thus the term $\ln T$ is removed in the results compared to Corollary 1. However, Corollary SM.1 can be easily generalized to the setting of Markovian data.

Proof of Corollary SM.1. We use the following notation throughout this proof:

$$\begin{aligned}
W_t &= (X_t, Y_t, Z_t, Z_t X_t^\intercal, Z_t Y_t, Z_t Z_t^\intercal),\\
w &= (x, y, z, zx^\intercal, zy, zz^\intercal),\\
G_{\mathrm{SGD}}(w, \theta, \Gamma) &= (y - x^\intercal\theta)\Gamma z,\\
H_{\mathrm{SGD}}(w, \Gamma) &= (x - \Gamma z) z^\intercal.
\end{aligned}$$

Then, $\{W_t : t \geq 0\}$ forms an i.i.d. sequence. Let ν denote the distribution of W_t.

We verify that W_t, G_{SGD}, and H_{SGD} satisfy Assumptions 1, 2, and 3, so that Proposition 1, Theorem 1, and Theorem 2 can be invoked. Note that Assumptions 1 holds because of condition (i) in Assumption SM.1, while Assumption 3 holds due to differentiability of G_{SGD} and H_{SGD}. Hence, it suffices to verify Assumption 2.

Verifying Assumption 2(i). Note that

$$\bar{G}_{\mathrm{SGD}}(\theta^*, \Gamma^*) := \mathbb{E}_\nu[G_{\mathrm{SGD}}(W_t, \theta^*, \Gamma^*) = \mathbb{E}_\nu[(Y_t - X_t^\intercal \theta^*)\Gamma^* Z_t] = \mathbb{E}_\nu[\epsilon_t \Gamma^* Z_t] = \Gamma^* \, \mathbb{E}_\nu[\epsilon_t Z_t] = 0,$$

and

$$\bar{H}_{\mathrm{SGD}}(\Gamma^*) := \mathbb{E}_\nu[H_{\mathrm{SGD}}(W_t, \Gamma^*)] = \mathbb{E}_\nu[(X_t - \Gamma^* Z_t) Z_t^\intercal] = \mathbb{E}_\nu[\eta_t Z_t^\intercal] = 0.$$

Hence,

$$\begin{aligned}
\bar{G}_{\mathrm{SGD}}(\theta, \Gamma^*) &= \mathbb{E}_\nu[G_{\mathrm{SGD}}(W_t, \theta, \Gamma^*)] - \mathbb{E}_\nu[G_{\mathrm{SGD}}(W_t, \theta^*, \Gamma^*)] \\
&= \mathbb{E}_\nu[\Gamma^* Z_t X_t^\intercal (\theta - \theta^*)] \\
&= \mathbb{E}_\nu[\Gamma^* Z_t (\Gamma^* Z_t + \eta_t)^\intercal (\theta - \theta^*)] \\
&= \mathbb{E}_\nu[\Gamma^* Z_t Z_t^\intercal (\Gamma^*)^\intercal](\theta - \theta^*)^\intercal.
\end{aligned}$$

and

$$\bar{H}_{\mathrm{SGD}}(\Gamma) = \mathbb{E}_\nu[H_{\mathrm{SGD}}(W_t, \Gamma)] = \mathbb{E}_\nu[H_{\mathrm{SGD}}(W_t, \Gamma)] - \mathbb{E}_\nu[H_{\mathrm{SGD}}(W_t, \theta^*, \Gamma^*)] = \mathbb{E}_\nu[(\Gamma - \Gamma^*) Z_t Z_t^\intercal].$$

Therefore, Γ^* is the unique solution to $\bar{H}_{\mathrm{SGD}}(\Gamma) = 0$, because $\mathbb{E}_\nu[Z_t Z_t^\intercal]$ has full rank. Moreover, since Γ^* has full rank, we conclude that $\mathbb{E}_\nu[\Gamma^* Z_t Z_t^\intercal (\Gamma^*)^\intercal]$ has full rank, and thus θ^* is the unique solution to $\bar{G}_{\mathrm{SGD}}(\theta, \Gamma^*) = 0$.

Verifying Assumption 2(ii). Let $B > \max(\|\theta^*\|, \|\Gamma^*\|_{\mathsf{F}})$ be a constant. Note that for all $\theta, \tilde{\theta} \in \mathbb{R}^p$ and $\Gamma, \tilde{\Gamma} \in \mathbb{R}^{p \times q}$ with $\|\Gamma\|_{\mathsf{F}} \leq B$,

$$\|G_{\mathrm{SGD}}(w, \theta, \Gamma) - G_{\mathrm{SGD}}(w, \tilde{\theta}, \Gamma)\| = \|\Gamma z x^\intercal (\theta - \tilde{\theta})\| \leq \|\Gamma\| \|z x^\intercal\| \|\theta - \tilde{\theta}\| \leq B \|w\| \|\theta - \tilde{\theta}\|.$$

Moreover, for all $\theta, \tilde{\theta} \in \mathbb{R}^p$ with $\|\theta\| \leq B$ and $\Gamma, \tilde{\Gamma} \in \mathbb{R}^{p \times q}$,

$$\|G_{\mathrm{SGD}}(w, \theta, \Gamma) - G_{\mathrm{SGD}}(w, \theta, \tilde{\Gamma})\| = \|(y - x^\intercal \theta)(\Gamma - \tilde{\Gamma}) z\| \leq \|yz - \theta^\intercal x z\| \|\Gamma - \tilde{\Gamma}\| \leq (1 + B) \|w\| \|\Gamma - \tilde{\Gamma}\|_{\mathsf{F}},$$

where the last step follows from the definition of w. Likewise,

$$\|H_{\mathrm{SGD}}(w, \Gamma) - H_{\mathrm{SGD}}(w, \tilde{\Gamma})\|_{\mathsf{F}} \leq \sqrt{p} \|H_{\mathrm{SGD}}(w, \Gamma) - H_{\mathrm{SGD}}(w, \tilde{\Gamma})\| = \sqrt{p} \|(\Gamma - \tilde{\Gamma}) z z^\intercal\| \leq \sqrt{p} \|\Gamma - \tilde{\Gamma}\|_{\mathsf{F}} \|w\|.$$

In addition, for all $\theta, \tilde{\theta} \in \mathbb{R}^p$ and $\Gamma, \tilde{\Gamma} \in \mathbb{R}^{p \times q}$ with $\max(\|\theta\|, \|\Gamma\|_{\mathsf{F}}) \leq B$,

$$\|G_{\mathrm{SGD}}(w, \theta, \Gamma)\| = \|(y - x^\intercal \theta) \Gamma z\| \leq \|yz - \theta^\intercal x z\| \|\Gamma\| \leq B(1 + B) \|w\|$$

and

$$\|H_{\mathrm{SGD}}(w,\Gamma)\|_{\mathsf{F}} \leq \sqrt{p}\|H_{\mathrm{SGD}}(w,\Gamma)\| = \sqrt{p}\|(x-\Gamma z)z^{\intercal}\| \leq \sqrt{p}(1+B)\|w\|.$$

Therefore, by setting $L = \max(1+B, B(1+B), \sqrt{p}B(1+B))$, we have verified Assumption 2(ii).

Verifying Assumption 2(iii). By full rankness of Γ^* and $\mathbb{E}_{\nu}[Z_t Z_t^{\intercal}]$, we have

$$\begin{aligned}
&(\theta-\theta^*)^{\intercal}(\mathbb{E}[(Y_t - X_t^{\intercal}\theta)\Gamma^* Z_t] - (Y_t - X_t^{\intercal}\theta^*)\Gamma^* Z_t])\\
&= -(\theta-\theta^*)^{\intercal}\mathbb{E}[\Gamma^* Z_t X_t^{\intercal}](\theta-\theta^*)\\
&= -(\theta-\theta^*)^{\intercal}\mathbb{E}[\Gamma^* Z_t(\Gamma^* Z_t + \eta_t)^{\intercal}](\theta-\theta^*)\\
&= -(\theta-\theta^*)^{\intercal}\mathbb{E}[\Gamma^* Z_t Z_t^{\intercal}(\Gamma^*)^{\intercal}](\theta-\theta^*)\\
&\leq -\zeta\|\theta-\theta^*\|^2,
\end{aligned}$$

since $\zeta > 0$ is the smallest eigenvalue of $\mathbb{E}[\Gamma^* Z_t Z_t^{\intercal}(\Gamma^*)^{\intercal}]$.

Moreover, with a proof similar to that of Lemma SM.19, we can show that for all $\Gamma \in \mathbb{R}^{p\times q}$,

$$\mathrm{vec}(\Gamma-\Gamma^*)^{\intercal}\mathrm{vec}(\bar{H}_{\mathrm{SGD}}(\Gamma) - \bar{H}_{\mathrm{SGD}}(\Gamma^*)) \leq -\psi\|\Gamma-\Gamma^*\|_{\mathsf{F}}^2. \qquad \square$$

F. Other IV-RL Algorithms

A great variety of RL algorithm have been developed in recent years. We refer to Sutton and Barto (2018) for a general introduction. In this section, we focus on temporal-difference (TD) learning and actor-critic (AC) methods, due to their prominent roles in RL literature. We discuss how to combine them with the IV approach to address the reward endogeneity issue. The use of IVs in other RL algorithms can be pursued in a similar fashion.

F.1. IV-TD

TD-Learning (Sutton 1988) is a fundamental algorithm for policy evaluation—that is, computation of the value function of a *fixed* policy—via value iteration. It is a building block for many sophisticated RL algorithms.

For any policy π, define $Q_\pi(s,a)$ as the value of playing action a at state s while playing the policy π afterwards. Namely,

$$Q_\pi(s,a) := r(s,a) + \mathbb{E}_{a\sim\pi(\cdot|s)}\left[\sum_{t=1}^{\infty}\gamma^t r(S_t, A_t)\middle| S_0 = s, A_0 = a\right].$$

Then, Q_π satisfies the following equation:

$$Q_\pi(s,a) = r(s,a) + \gamma\mathbb{E}_{s'\sim\mathsf{P}(\cdot|s,a)}Q_\pi(s',a'). \tag{F.1}$$

Assume that Q_π can be well approximated with a linear architecture, that is, $Q_\pi(s,a) = \phi^\intercal(s,a)\theta_\pi$ for some unknown parameter $\theta_\pi \in \mathbb{R}^p$, where $\phi(s,a) \in \mathbb{R}^p$ is a vector of possibly nonlinear transformations of (s,a). Under such approximation, TD-Learning iteratively updates the estimate of θ_π as follows:

$$\theta_{t+1} = \theta_t + \alpha_t[r_t + \gamma\phi^\intercal(S_{t+1}, A_{t+1})\theta_t - \phi^\intercal(S_t, A_t)\theta_t]\phi(S_t, A_t),$$

where $\alpha_t > 0$ is learning rate at time t.

In the presence reward endogeneity, the observed signal $R(s,a)$ is biased from the true reward $r(s,a)$. Following the same assumption with regard to the IVs as that for IV-Q-Learning, we propose the IV-TD algorithm (Algorithm 3) to learn $Q_\pi(s,a)$ for a given policy π.

Algorithm 3: IV-TD-Learning with Linear Function Approximation

Input : Behavior policy π, learning rates (α_t, β_t), features ϕ, and time horizon T
Output : θ_T and Γ_T

Initialize θ_0 and Γ_0
for *all* $t = 0, 1, \ldots, T-1$ **do**
  Observe the state S_t and take action $A_t \sim \pi(\cdot|S_t)$
  Observe IV Z_t, reward R_t, and the new state S_{t+1}
  Update the estimates of θ_π and Γ_π via

$$\theta_{t+1} = \theta_t + \alpha_t\Big[R_t + \gamma\phi^\intercal(S_{t+1}, A_{t+1})\theta_t - \phi^\intercal(S_t, A_t)\theta_t\Big]\Gamma_t Z_t,$$
$$\Gamma_{t+1} = \Gamma_t + \beta_t[\phi(S_t, A_t) - \Gamma_t Z_t]Z_t^\intercal.$$

end

To facilitate the presentation of theoretical properties of the IV-TD algorithm, we use the following notation throughout this section:

$$W_t = (S_t, A_t, R_t, Z_t, S_{t+1}, A_{t+1}, \|Z_t\|\|S_t\|, \|Z_t\|\|S_{t+1}\|, Z_t Z_t^\intercal, R_t Z_t),$$
$$w = (s, a, r, z, s', a', \|z\|\|s\|, \|z\|\|s'\|, zz^\intercal, rz),$$
$$G_{\mathrm{TD}}(w, \theta, \Gamma) = (r + \gamma\phi^\intercal(s', a')\theta - \phi^\intercal(s,a)\theta)\Gamma z,$$
$$H_{\mathrm{TD}}(w, \Gamma) = (\phi(s,a) - \Gamma z)z^\intercal.$$

Moreover, we denote the steady-state lag-l covariance of $\{G_{\mathrm{TD}}(W_t, \theta^*, \Gamma^*) : t \geq 0\}$ and its TAVC, respectively, by

$$L_{G,\mathrm{TD}}(l) := \mathrm{Cov}_\nu[G(W_0, \theta^*, \Gamma^*), G(W_l, \theta^*, \Gamma^*)], \quad l \geq 0,$$
$$\bar{L}_{G,\mathrm{TD}} := L_{G,\mathrm{TD}}(0) + \sum_{l=1}^{\infty}(L_{G,\mathrm{TD}}(l) + L_{G,\mathrm{TD}}^\intercal(l)).$$

Similar to IV-SGD, below we present the convergence rates for IV-TD algorithm. Note that IV-TD does not require statement (iii) of Assumption 5 compared to IV-Q-Learning if additional time non-dependence condition $\mathbb{E}[\eta_{t+1} Z_t^\intercal] = 0$ holds.

COROLLARY SM.2. *Let $\{\theta_t : t \geq 0\}$ be the iterates from the projected IV-TD. For all $t \geq 1$, let $\alpha_t = \alpha_0 t^{-\kappa}$ and $\beta_t = \beta_0 t^{-\delta}$ for some $\alpha_0 > 0$, $\beta_0 > 0$, and $0 < \kappa \leq \delta \leq 1$. Suppose conditions (i), (ii) and (iv) of Assumption 5 hold with θ^* there being replaced by θ_π. Assume $\zeta\alpha_0 > 1$ in the case of $\kappa = 1$, and $\psi\beta_0 > 1$ in the case of $\delta = 1$, where $\zeta = (1-\gamma)\varphi$, φ is the smallest eigenvalue of $\mathbb{E}_\nu[\Gamma^* Z_t Z_t^\intercal (\Gamma^*)^\intercal]$, and ψ is the smallest eigenvalue of $\mathbb{E}_\nu[Z_t Z_t^\intercal]$. In addition, assume that $\mathbb{E}_\nu[Z_t \eta_{t+1}^\intercal] = 0$.*

(i) If Assumption 4 holds with $d_U = 2$, then for all t large enough,

$$\mathbb{E}[\|\theta_t - \theta_\pi\|^2 | W_0 = w] \leq t^{-\kappa}(k_1 \ln t + k_2), \quad \forall t \geq T,$$

where k_1 and k_2 are some nonnegative constants that depend on w. Moreover, $k_1 = 0$ if $\{W_t : t \geq 0\}$ are i.i.d.

(ii) Fix arbitrary $C > 0$, $\kappa_c \in [0, \kappa - 1/2)$, and $\delta_c \in [0, \delta - 1/2)$. If Assumption 4 holds with $d_U = 2$ and $1/2 < \kappa \leq \delta < 1$, then for all t large enough,

$$\mathbb{P}\big(\|\theta_t - \theta_\pi\| \leq C t^{-\kappa_c}, \forall t \geq T \big| W_0 = w\big) \geq 1 - C_\theta T^{-b_\kappa} \ln T,$$

where $b_\kappa := 2\kappa - 1 - 2\kappa_c$, and C_θ is some positive constant that depends on $(\delta, \delta_c, \kappa, \kappa_c, C, w)$, and the term $\ln T$ can be removed if $\{W_t : t \geq 0\}$ are i.i.d.

(iii) If Assumption 4 holds with $d_U = 4$ and $1/2 < \kappa < \delta < 1$, then as $t \to \infty$,

$$\alpha_t^{-1/2}(\theta_t - \theta_\pi) \rightsquigarrow \mathcal{N}(0, \Sigma_{G,\mathrm{TD}}),$$

where

$$\begin{aligned}
\Sigma_{G,\mathrm{TD}} &:= \int_0^\infty \exp\left(t A_{11,\mathrm{TD}}\right) \bar{L}_{G,\mathrm{TD}} \exp\left(t A_{11,\mathrm{TD}}^\intercal\right) \mathrm{d}t\,, \\
A_{11,\mathrm{TD}} &:= -(1-\gamma)\mathbb{E}_\nu[Z_t^\intercal (\Gamma^*)^\intercal \Gamma^* Z_t].
\end{aligned}$$

Similar to the proofs of Corollary 1 and Corollary SM.1, we verify Assumptions 1–3 to apply Proposition 1 and Theorems 1 and 2. The following two lemmas will be used to verify conditions (ii) and (iii) of Assumption 2.

LEMMA SM.20. *For all $w \in \mathcal{W}$, $\theta, \tilde{\theta} \in \mathbb{R}^p$ with $\|\theta\| \leq B$, and $\Gamma, \tilde{\Gamma} \in \mathbb{R}^{p \times q}$ with $\|\Gamma\|_\mathsf{F} \leq B$, we have*

$$\begin{aligned}
\|G_{\mathrm{TD}}(w, \theta, \Gamma) - G_{\mathrm{TD}}(w, \tilde{\theta}, \Gamma)\| &\leq 2BK(\gamma+1)(1 + \|w\|)\|\theta - \tilde{\theta}\|, \\
\|G_{\mathrm{TD}}(w, \theta, \Gamma) - G_{\mathrm{TD}}(w, \theta, \tilde{\Gamma})\| &\leq (1 + 2BK(\gamma+1))(1 + \|w\|)\|\Gamma - \tilde{\Gamma}\|_\mathsf{F}, \\
\|G_{\mathrm{TD}}(w, \theta, \Gamma)\| &\leq B(1 + 2BK(\gamma+1))(1 + \|w\|).
\end{aligned}$$

Proof of Lemma SM.20. It follows from the definition of G_{TD} that

$$\begin{aligned}\|G_{\mathrm{TD}}(w,\theta,\Gamma)-G_{\mathrm{TD}}(w,\tilde{\theta},\Gamma)\| &= \left\|(\gamma\phi(s',a')-\phi(s,a))^{\intercal}(\theta-\tilde{\theta})\Gamma z\right\| \\ &\le \big(\|\gamma\phi(s',a')\|+\|\phi(s,a)\|\big)\cdot\|\theta-\tilde{\theta}\|\cdot\|\Gamma\|\cdot\|z\| \\ &\le \big(\gamma K(1+\|s'\|)+K(1+\|s\|)\big)\|\Gamma\|_{\mathsf{F}}\|z\|\|\theta-\tilde{\theta}\| \\ &\le 2BK(\gamma+1)(1+\|w\|)\|\theta-\tilde{\theta}\|,\end{aligned}$$

for all $w\in\mathcal{W}$, $\theta,\tilde{\theta}\in\mathbb{R}^p$, and $\Gamma\in\mathbb{R}^{p\times q}$ with $\|\Gamma\|_{\mathsf{F}}\le B$, where the second from Assumption 5(ii), and the last from the definition of w. Likewise,

$$\begin{aligned}\|G_{\mathrm{TD}}(w,\theta,\Gamma)-G_{\mathrm{TD}}(w,\theta,\tilde{\Gamma})\| &= \left\|(r+\gamma\phi^{\intercal}(s',a')\theta-\phi^{\intercal}(s,a)\theta)(\Gamma-\tilde{\Gamma})z\right\| \\ &\le |r+(\gamma\phi(s',a')-\phi(s,a))^{\intercal}\theta|\cdot\|\Gamma-\tilde{\Gamma}\|\cdot\|z\| \\ &\le (1+2BK(\gamma+1))(1+\|w\|)\|\Gamma-\tilde{\Gamma}\|_{\mathsf{F}},\end{aligned}$$

for all $w\in\mathcal{W}$, $\theta\in\mathbb{R}^p$ with $\|\theta\|\le B$, and $\Gamma,\tilde{\Gamma}\in\mathbb{R}^{p\times q}$.

For $\|G_{\mathrm{TD}}(w,\theta,\Gamma)\|$, it is easy to see that

$$\begin{aligned}\|G_{\mathrm{TD}}(w,\theta,\Gamma)\| &= \|(r+\gamma\phi^{\intercal}(s',a')\theta-\phi^{\intercal}(s,a)\theta)\Gamma z\| \\ &\le |r+(\gamma\phi(s',a')-\phi(s,a))^{\intercal}\theta|\cdot B\cdot\|z\| \\ &\le B(1+2BK(\gamma+1))(1+\|w\|). \qquad \square\end{aligned}$$

LEMMA SM.21. *Let φ denote the smallest eigenvalue of $\mathbb{E}_\nu[\Gamma^* Z_t Z_t^{\intercal}(\Gamma^*)^{\intercal}]$, $\zeta=(1-\gamma)\varphi$, and*

$$\bar{G}_{\mathrm{TD}}(\theta,\Gamma) := \mathbb{E}_\nu[G_{\mathrm{TD}}(W_t,\theta,\Gamma)] = \mathbb{E}_\nu[(R_t+\gamma\phi^{\intercal}(S_{t+1},A_{t+1})\theta-\phi^{\intercal}(S_t,A_t)\theta)\Gamma Z_t]$$

Then, $\zeta>0$ and for all $\theta\in\mathbb{R}^p$,

$$(\theta-\theta_\pi)^{\intercal}(\bar{G}_{\mathrm{TD}}(\theta,\Gamma^*)-\bar{G}_{\mathrm{TD}}(\theta_\pi,\Gamma^*)) \le -\zeta\|\theta-\theta_\pi\|^2.$$

Proof of Lemma SM.21. By Assumption 5(iv), $\mathbb{E}[\Gamma^* Z_t Z_t^{\intercal}(\Gamma^*)^{\intercal}]$ has full rank, so $\zeta>0$. By the definition of Γ^* and θ_π, it is easy to see that

$$\begin{aligned}\bar{G}_{\mathrm{TD}}(\theta,\Gamma^*)-\bar{G}_{\mathrm{TD}}(\theta_\pi,\Gamma^*) &= \mathbb{E}_\nu[(R_t-r_t+\gamma\phi^{\intercal}(S_{t+1},A_{t+1})(\theta-\theta_\pi)-\phi^{\intercal}(S_t,A_t)(\theta-\theta_\pi))\Gamma^* Z_t] \\ &= \mathbb{E}_\nu[\epsilon_t\Gamma^* Z_t]+\mathbb{E}_\nu[\Gamma^* Z_t(\gamma\phi^T(S_{t+1,A_{t+1}})-\phi^T(S_t,A_t))(\theta-\theta_\pi)]. \qquad (\text{F.2})\end{aligned}$$

The first term of (F.2) is 0 by Assumption 5(iv).

The second term of (F.2) can be re-written as:

$$\mathbb{E}_\nu[\Gamma^* Z_t(\gamma\phi^{\intercal}(S_{t+1,A_{t+1}})-\phi^{\intercal}(S_t,A_t))(\theta-\theta_\pi)]$$

$$= \mathbb{E}_\nu[\Gamma^* Z_t(\gamma Z_{t+1}^\intercal (\Gamma^*)^\intercal + \eta_{t+1} - Z_t^\intercal (\Gamma^*)^\intercal - \eta_t)](\theta - \theta_\pi)$$
$$= (\gamma \mathbb{E}_\nu[\Gamma^* Z_t Z_{t+1}^\intercal (\Gamma^*)^\intercal] - \mathbb{E}_\nu[\Gamma^* Z_t Z_t^\intercal \Gamma^\intercal] + \gamma \mathbb{E}_\nu[\Gamma^* Z_t \eta_{t+1}^\intercal] - \mathbb{E}_\nu[\Gamma^* Z_t \eta_t^\intercal])(\theta - \theta_\pi)$$

By Assumption 5(iv) and the assumption that $\mathbb{E}_\nu[Z_t\eta_{t+1}^\intercal] = 0$, we have

$$\mathbb{E}_\nu[\Gamma^* Z_t \eta_{t+1}^\intercal] = \mathbb{E}_\nu[\Gamma^* Z_t \eta_t^\intercal] = 0.$$

By the Cauchy-Schwartz inequality,

$$|(\theta-\theta_\pi)^\intercal \gamma \mathbb{E}_\nu[\Gamma^* Z_t Z_{t+1}^\intercal (\Gamma^*)^\intercal](\theta-\theta_\pi)| \le \gamma(\mathbb{E}_\nu[((\theta-\theta_\pi)^\intercal \Gamma^* Z_t)^2])^{\frac{1}{2}} (\mathbb{E}_\nu[((\theta-\theta_\pi)^\intercal \Gamma^* Z_{t+1})^2])^{\frac{1}{2}}.$$

By the Markov property of Z_t, $\mathbb{E}_\nu[\Gamma^* Z_{t+1} Z_{t+1}^\intercal (\Gamma^*)^\intercal] = \mathbb{E}_\nu[\Gamma^* Z_t Z_t^\intercal (\Gamma^*)^\intercal]$. Therefore,

$$|(\theta-\theta_\pi)^\intercal \gamma \mathbb{E}_\nu[\Gamma^* Z_t Z_{t+1}^\intercal (\Gamma^*)^\intercal](\theta-\theta_\pi)| \le \gamma(\theta-\theta_\pi)\mathbb{E}_\nu[\Gamma^* Z_t Z_t^\intercal (\Gamma^*)^\intercal](\theta-\theta_\pi). \tag{F.3}$$

Plugging (F.3) into (F.2), we have:

$$(\theta-\theta_\pi)^\intercal(\bar{G}_{\mathrm{TD}}(\theta,\Gamma^*) - \bar{G}_{\mathrm{TD}}(\theta_\pi,\Gamma^*)) \le (\gamma-1)(\theta-\theta_\pi)^\intercal \mathbb{E}_\nu[\Gamma^* Z_t Z_t^\intercal (\Gamma^*)^\intercal](\theta-\theta_\pi) \le -\zeta\|\theta-\theta_\pi\|^2,$$

for all $\theta \in \mathbb{R}^p$. □

Proof of Corollary SM.2. Assumption 1 holds automatically by its definition. The local Taylor expansion in Assumption 3 holds due to the linearity of G_{TD} and H_{TD} in (θ, Γ). The negative definiteness of A_{11} and A_{22} in Assumption 3 holds due to the full rankness of Γ^* and $\mathbb{E}_\nu[Z_t Z_t^\intercal]$.

Moreover, note that H_{Q} and H_{TD} are identical. Hence, Assumption 2(ii) is verified by Lemmas SM.18 and SM.20, while Assumption 2(iii) is verified by Lemmas SM.19 and SM.21. Hence, to apply Proposition 1, Theorem 1, and Theorem 2 to prove Corollary SM.2, it suffices to verify Assumption 2(i); that is, (θ_π, Γ^*) is the unique solution to the system of equations

$$\mathbb{E}_\nu[G_{\mathrm{TD}}(W_t, \theta, \Gamma)] = 0,$$
$$\mathbb{E}_\nu[H_{\mathrm{TD}}(W_t, \Gamma)] = 0.$$

It is easy to see that

$$\begin{aligned}
\mathbb{E}_\nu[G_{\mathrm{TD}}(W_t,\theta_\pi,\Gamma^*)] &= \mathbb{E}_\nu[(R_t + \gamma\phi^\intercal(S_{t+1},A_{t+1})\theta^* - \phi^\intercal(S_t,A_t)\theta\pi)\Gamma^* Z_t] \\
&= \mathbb{E}_\nu[\epsilon_t \Gamma^* Z_t] + \mathbb{E}_\nu[(r(S_t,A_t) + \gamma\phi^\intercal(S_{t+1},A_{t+1})\theta_\pi - \phi^\intercal(S_t,A_t)\theta_\pi)\Gamma^* Z_t] \\
&= \mathbb{E}_\nu[(r(S_t,A_t) + \gamma\phi^\intercal(S_{t+1},A_{t+1})\theta_\pi - \phi^\intercal(S_t,A_t)\theta_\pi)\Gamma^* Z_t],
\end{aligned}$$

where the last inequality follows from Assumption 5(iv).

By the definition of Q_π, particularly the equation (F.1) it satisfies, and the assumption that $Q_\pi(s,a)=\phi^\intercal(s,a)\theta_\pi$,

$$\phi^\intercal(S_t,A_t)\theta_\pi=\mathbb{E}_{S_{t+1}\sim \mathsf{P}(\cdot|S_t,A_t)}[r(S_t,A_t)+\gamma\phi^\intercal(S_{t+1},A_{t+1})\theta_\pi]. \tag{F.4}$$

Multiplying $\Gamma^* z$ on both sides of (F.4) and applying $\mathbb{E}_\nu$, we have:

$$\mathbb{E}_\nu[\phi^\intercal(S_t,A_t)\theta_\pi\Gamma^* z]=\mathbb{E}_\nu[(r(S_t,A_t)+\gamma\phi^\intercal(S_{t+1},A_{t+1})\theta_\pi)\Gamma^* Z_t].$$

Therefore, $\mathbb{E}_\nu[G_{\text{TD}}(w,\theta_\pi,\Gamma^*)]=0$; that is, θ^* solves the equation $\bar{G}_{\text{TD}}(\theta,\Gamma^*)=0$.

It is also easy to show $\mathbb{E}_\nu[H_{\text{TD}}(w,\Gamma^*)]=0$, so Γ^* solves the equation $\bar{H}_{\text{TD}}(\Gamma)=0$.

The uniqueness of θ_π and Γ^* are guaranteed by linear independence of $\phi(s,a)$ and full rankness of Γ^* and $\mathbb{E}_\nu[Z_tZ_t^\intercal]$ by conditions (i) and (iv) in Assumption 5. Hence, Assumption 2(i) holds. □

F.2. IV-AC

In contrast to Q-Learning and TD-Learning, both of which rely on a fixed behavior policy to generate data, actor-critic (AC) algorithms gradually improve the behavior policy towards the optimal. There are essentially two learning processes that are executed by a "critic" and an "actor" simultaneously. The critic performs value iteration and aims to learn the value function Q_π of the behavior policy π that is currently undertaken by the actor. Meanwhile, the actor performs policy iteration and aims to improve the behavior policy based on the evaluation provided by the critic.

Intuitively, the critic may need multiple rounds of value iteration between two policy updates from the actor to provide accurate estimation of the value function. Nevertheless, to avoid *ad hoc* schemes to specify the length of the time interval between two policy updates, a common treatment is to operate value iteration and policy iteration with different learning rates, so that the critic lives on a faster effective timescale than the actor. This effectively ensures that the current behavior policy of the actor appears approximately static in the view of the critic; see, e.g., Borkar and Konda (1997) and Konda and Tsitsiklis (2003) for details.

Formally, we consider a randomized policy π and parametrize it as $\pi_\mu(\cdot|s)$. That is, for a given state s, $\pi_\mu(\cdot|s)$ is a probability distribution with parameter μ on the action space $\mathcal{A}$. Assume that the optimal action-value function is linearized; i.e., $Q^*(s,a)=\phi^\intercal(s,a)\theta^*$. AC algorithms usually alternate between the critic step and actor step as follows

$$\begin{aligned}
&\text{(Critic)} \qquad \theta_{t+1}=\theta_t+\alpha_t[R_t+\gamma\phi^\intercal(S_{t+1},A_{t+1})\theta_t-\phi^\intercal(S_t,A_t)\theta_t]\Gamma_t Z_t,\\
&\text{(Actor)} \qquad \mu_{t+1}=\mu_t+\frac{1}{1-\gamma}\tau_t\phi^\intercal(S_t,A_t)\theta_t\nabla_\mu\log\pi_{\mu_t}(A_t|S_t),
\end{aligned}$$

where α_t and τ_t are the learning rates.

Similar to Q-Learning and TD-Learning, AC algorithms can also be modified naturally to take advantage of IVs to address potential reward endogeneity. We hereby present the IV-AC algorithm (Algorithm 4). In light of the varying learning rates $(\alpha_t, \beta_t, \tau_t)$, IV-AC can be viewed as three-timescale SA. It is conceivable that the techniques developed in the present paper for analyzing two-timescale SA can serve as the foundation for theoretical analysis of IV-AC. We leave it for future research.

Algorithm 4: IV-AC with Linear Function Approximation

Input : Learning rates $(\alpha_t, \beta_t, \tau_t)$, features ϕ, and time horizon T
Output : θ_T, μ_T, Γ_T

Initialize θ_0, μ_0, and Γ_0
for *all* $t = 0, 1, \dots$ **do**
  Observe the state S_t and take action $A_t \sim \pi_{\mu_t}(\cdot|S_t)$
  Observe IV Z_t, reward R_t, and the new state S_{t+1}
  Update the estimates via

$$\begin{aligned}
\theta_{t+1} &= \theta_t + \alpha_t[R_t + \gamma\phi^\intercal(S_{t+1}, A_{t+1})\theta_t - \phi^\intercal(S_t, A_t)\theta_t]\Gamma_t Z_t,\\
\mu_{t+1} &= \mu_t + \frac{1}{1-\gamma}\tau_t\phi^\intercal(S_t, A_t)\theta_t\nabla_\mu \log\pi_\mu(A_t|S_t),\\
\Gamma_{t+1} &= \Gamma_t + \beta_t[\phi(S_t, A_t) - \Gamma_t Z_t]Z_t^\intercal.
\end{aligned}$$

end

G. Calculations for Numerical Experiments

G.1. Experiment for IV-SGD

To establish the validity of IV $q_t(S_t - \mathbb{E}[S_t])$ in Section 7.1, we note that

$$\mathrm{Cov}(q_t(S_t - \mathbb{E}[S_t]), S_t A_t) = \frac{1}{2}(1-p)\left(e^{1/2} - 1\right) \neq 0,$$

while

$$\mathrm{Cov}(q_t(S_t - \mathbb{E}[S_t]), f(A_t)) = 0,$$

for any measurable function $f(\cdot)$.

For SGD, the bias in estimating θ_1^* is

$$\begin{aligned}
\mathrm{plim}_{T\to\infty}(\hat\theta_{1,T}^{\mathrm{SGD}} - \theta^*) &= \frac{\mathrm{Cov}(bA_t^2, S_t A_t)}{\mathrm{Var}(S_t A_t)}\\
&= b\frac{\left(p\widetilde{\theta}^3 e^2 + \frac{1}{4}e^{1/8}(1-p)\right) - \left(p\widetilde{\theta}^2 e^{1/2} + \frac{1}{3}(1-p)\right)\left(p\widetilde{\theta} e^{1/2} + \frac{1}{2}e^{1/8}(1-p)\right)}{\left(p\widetilde{\theta}^2 e^2 + \frac{1}{5}(1-p)\right) - \left(p\widetilde{\theta} e^{1/2} + \frac{1}{3}(1-p)\right)^2}.
\end{aligned}$$

G.2. Experiment for IV-Q-Learning

We calculate the optimal policy as well as the value function of linear policies for the linear quadratic problem in Section 7.2.

G.2.1. Optimal Policy Let $\theta^* = (\theta_0^*, \theta_1^*, \dots, \theta_5^*)^\intercal$ with $\theta_5^* < 0$. Assume that

$$Q^*(s,a) = \phi^\intercal(s,a)\theta^* = \theta_0^* + \theta_1^* s + \theta_2^* a + \theta_3^* sa + \theta_4^* s^2 + \theta_5^* a^2.$$

Then, given state s, the optimal action is $\arg\max_a Q^*(s,a) = \omega_0^* + \omega_1^* s$, where

$$\omega_0^* = -\frac{\theta_2^*}{2\theta_5^*} \quad \text{and} \quad \omega_1^* = -\frac{\theta_3^*}{2\theta_5^*};$$

moreover,

$$\max_a Q^*(s,a) = \theta_0^* + \theta_1^* s - \frac{(\theta_2^* + \theta_3^* s)^2}{4\theta_5^*} + \theta_4^* s^2.$$

It then follows from Bellman's equation that

$$\begin{aligned}
Q^*(S_t, A_t) &= \mathbb{E}\big[R(S_t,A_t) + \gamma \max_{a'} Q^*(S_{t+1}, a') | S_t, A_t\big] \\
&= r(S_t,A_t) + \mathbb{E}[\epsilon_t | S_t, A_t] + \gamma \mathbb{E}\left[\theta_0^* + \theta_1^* S_{t+1} - \frac{(\theta_2^* + \theta_3^* S_{t+1})^2}{4\theta_5^*} + \theta_4^* S_{t+1}^2 \middle| S_t, A_t\right] \\
&= r_0 + r_1 A_t + r_2 S_t A_t + r_3 A_t^2 + b\,\mathbb{E}[A_{t,2}^2] \\
&\quad + \gamma \mathbb{E}\left[\left(\theta_0^* - \frac{(\theta_2^*)^2}{4\theta_5^*}\right) + \left(\theta_1^* - \frac{\theta_2^*\theta_3^*}{2\theta_5^*}\right) S_{t+1} + \left(\theta_4^* - \frac{(\theta_3^*)^2}{4\theta_5^*}\right) S_{t+1}^2 \middle| S_t, A_t\right] \\
&= r_0 + r_1 A_t + r_2 S_t A_t + r_3 A_t^2 + b\,\mathbb{E}[A_{t,2}^2] + \gamma\left[\chi_0 + \chi_1 \bar{S}_{t+1} + \chi_2\big(\bar{S}_{t+1}^2 + \sigma_\eta^2\big)\right],
\end{aligned}$$

where $\bar{S}_{t+1} := \mathbb{E}[S_{t+1} | S_t, A_t] = c_0 + c_1 S_t + c_2 A_t$, and

$$\chi_0 := \theta_0^* - \frac{(\theta_2^*)^2}{4\theta_5^*}, \quad \chi_1 := \theta_1^* - \frac{\theta_2^*\theta_3^*}{2\theta_5^*}, \quad \text{and} \quad \chi_2 := \theta_4^* - \frac{(\theta_3^*)^2}{4\theta_5^*}.$$

Here, the last equality holds because the state transition follows $S_{t+1} = c_0 + c_1 S_t + c_2 A_t + \eta_t$ with η_t having mean 0 and variance σ_η^2. We may further expand the RHS of the above equation to express it as a quadratic function in (S_t, A_t). That is,

$$\begin{aligned}
Q^*(S_t, A_t) = r_0 &+ b\,\mathbb{E}[A_{t,2}^2] + \gamma\chi_0 + \gamma c_0 \chi_1 + \gamma(c_0^2 + \sigma_\eta^2)\chi_2 + \gamma c_1(\chi_1 + 2c_0\chi_2) S_t \\
&+ [r_1 + \gamma c_2(\chi_1 + 2\beta_0\chi_2)] A_t + (r_2 + 2\gamma\beta_1 c_2 \chi_2) S_t A_t + \gamma c_1^2 \chi_2 S_t^2 + (r_3 + \gamma c_2^2 \chi_2) A_t^2
\end{aligned}$$

Note that $Q^*(S_t, A_t)$ is also quadratic in (S_t, A_t). Thus, by matching the coefficients, we have

$$\begin{aligned}
\theta_0^* &= r_0 + b\,\mathbb{E}[A_{t,2}^2] + \gamma\left(\theta_0^* - \frac{(\theta_2^*)^2}{4\theta_5^*}\right) + \gamma c_0\left(\theta_1^* - \frac{\theta_2^*\theta_3^*}{2\theta_5^*}\right) + \gamma(c_0^2 + \sigma_\eta^2)\left(\theta_4^* - \frac{(\theta_3^*)^2}{4\theta_5^*}\right), \\
\theta_1^* &= \gamma c_1\left[\left(\theta_1^* - \frac{\theta_2^*\theta_3^*}{2\theta_5^*}\right) + 2c_0\left(\theta_4^* - \frac{(\theta_3^*)^2}{4\theta_5^*}\right)\right],
\end{aligned}$$

$$
\begin{aligned}
\theta_2^* &= r_1 + \gamma c_2\left[\left(\theta_1^* - \frac{\theta_2^*\theta_3^*}{2\theta_5^*}\right) + 2c_0\left(\theta_4^* - \frac{(\theta_3^*)^2}{4\theta_5^*}\right)\right],\\
\theta_3^* &= r_2 + 2\gamma c_1 c_2\left(\theta_4^* - \frac{(\theta_3^*)^2}{4\theta_5^*}\right),\\
\theta_4^* &= \gamma c_1^2\left(\theta_4^* - \frac{(\theta_3^*)^2}{4\theta_5^*}\right),\\
\theta_5^* &= r_3 + \gamma c_2^2\left(\theta_4^* - \frac{(\theta_3^*)^2}{4\theta_5^*}\right).
\end{aligned}
$$

Given the values of r_i and c_i, $i = 0, 1, 2$, we may then solve for θ^* from the above system of equations.

G.2.2. Value Function of Linear Policies For an arbitrary linear policy $\pi_\omega(s) = \omega_0 + \omega_1 s$, its value function $V_\pi(s)$ can be calculated as follows. Assume that $V_{\pi_\omega}(s) := v_0 + v_1 s + v_2 s^2$. By Bellman's equation,

$$
\begin{aligned}
V_{\pi_\omega}(S_0) &= r(S_0, A_0) + \gamma\,\mathbb{E}[V_{\pi_\omega}(S_1)|S_0]\\
&= r(S_0, \pi_\omega(S_0)) + \gamma\mathbb{E}[v_0 + v_1 S_1 + v_2 S_1^2|S_0]\\
&= r_0 + r_1(\omega_0 + \omega_1 S_0) + r_2 S_0(\omega_0 + \omega_1 S_0) + r_3(\omega_0 + \omega_1 S_0)^2 + \gamma\mathbb{E}[v_0 + v_1 S_1 + v_2 S_1^2|S_0]. \quad \text{(G.1)}
\end{aligned}
$$

Note that under policy π_ω,

$$
S_1 = c_0 + c_1 S_0 + c_2 A_0 + \eta_0 = (c_0 + c_2\omega_0) + (c_1 + c_2\omega_1)S_0 + \eta_0.
$$

It is easy to show that the RHS of (G.1) is a quadratic function in S_0. Since $V_{\pi_\omega}(S_0)$ is also quadratic in S_0, matching the coefficients yields

$$
\begin{aligned}
v_0 &= r_0 + r_1\omega_0 + r_3\omega_0^2 + \gamma[v_0 + v_1(c_0 + c_2\omega_0) + v_2((\beta_0 + c_2\omega_0)^2 + \sigma_\eta^2)],\\
v_1 &= r_1\omega_1 + r_2\omega_0 + r_3 2\omega_0\omega_1 + \gamma[v_1(c_1 + c_2\omega_1) + 2v_2(\beta_0 + c_2\omega_0)(c_1 + c_2\omega_1)],\\
v_2 &= r_2\omega_1 + r_3\omega_1^2 + \gamma v_2(c_1 + c_2\omega_1)^2.
\end{aligned}
$$

This is a system of linear equations in (v_0, v_1, v_2) and the solution can be easily calculated.