Stochastic Gradient Descent with Adaptive Data

Ethan Che, Jing Dong
Columbia University, eche25@gsb.columbia.edu, jing.dong@gsb.columbia.edu
Xin T. Tong
National University of Singapore, xin.t.tong@nus.edu.sg

Stochastic gradient descent (SGD) is a powerful optimization technique that is particularly useful in online learning scenarios. Its convergence analysis is relatively well understood under the assumption that the data samples are independent and identically distributed (iid). However, applying SGD to policy optimization problems in operations research involves a distinct challenge: the policy changes the environment and thereby affects the data used to update the policy. The adaptively generated data stream involves samples that are non-stationary, no longer independent from each other, and affected by previous decisions. The influence of previous decisions on the data generated introduces bias in the gradient estimate, which presents a potential source of instability for online learning not present in the iid case. In this paper, we introduce simple criteria for the adaptively generated data stream to guarantee the convergence of SGD. We show that the convergence speed of SGD with adaptive data is largely similar to the classical iid setting, as long as the mixing time of the policy-induced dynamics is factored in. Our Lyapunov-function analysis allows one to translate existing stability analysis of stochastic systems studied in operations research into convergence rates for SGD, and we demonstrate this for queueing and inventory management problems. We also showcase how our result can be applied to study the sample complexity of an actor-critic policy gradient algorithm.

Key words: Stochastic gradient descent, adaptive data, gradient estimation

1. Introduction

We consider the following stochastic optimization problem

$$\min_{\theta\in\Theta} \ell(\theta) = \mathbb{E}_{\mu_\theta}[\mathcal{L}(\theta, z)]. \tag{1}$$

In (1), θ is an m-dimensional policy parameter that parametrizes a Markov chain with transition kernel P_θ. The probability measure μ_θ is the unique invariant distribution of P_θ, and $z \sim \mu_\theta$ is a random outcome. Here, the set Θ could either be a convex constraint set or $\mathbb{R}^m$ in the unconstrained setting.

For example, when designing service systems, we are interested in finding the optimal pricing and/or capacity sizing policies to strike a balance between revenue and service quality (Kim and Randhawa 2018, Chen et al. 2023a). Here, the policy parameter θ may consist of the price, which affects the demand, and the capacity, which affects the service speed. Jointly, they control a Markov chain P_θ that describes the queueing dynamics, and the service quality is often measured by the

steady-state average waiting time. Similarly, in inventory management, we are interested in finding the optimal inventory ordering policy (e.g., base-stock level) to minimize the long-run average holding and backlog costs (Huh et al. 2009, Zhang et al. 2020a). Here, the replenishment policy, indexed by θ, controls the dynamics of the inventory level, whose stationary distribution in turn determines the long-run average costs. The stochastic optimization problem (1) also arises in machine-learning applications, such as policy gradient-based reinforcement learning (Sutton and Barto 2018, Agarwal et al. 2021), strategic classification with adaptive best response (Mendler-Dünner et al. 2020, Li and Wai 2022), and adaptive experimental design with temporal carryovers (Glynn et al. 2020, Hu and Wager 2022).

In many problems of practical interest, direct access to the distribution μ_θ may not be available. Instead, at each time point t, we can apply a new candidate policy θ_t on the concurrent system state z_{t-1} and obtain a new data point following

$$z_t \sim P_{\theta_t}(\,\cdot\,|z_{t-1}). \tag{2}$$

We will refer to this as the *adaptive* data setting. Instead of minimizing (1) using the true gradient $\nabla\ell(\theta)$, which relies on the unknown distribution μ_θ, one only has access to a gradient estimator $g(\theta_t, z_t)$ based on the adaptive data stream, which satisfies

$$\mathbb{E}_{z\sim\mu_\theta}[g(\theta, z)] = \nabla\ell(\theta). \tag{3}$$

This resembles the problem setup for reinforcement learning (RL) (Sutton and Barto 2018), and includes policy-gradient algorithms as a special case. However, the RL literature generally focuses on a specific class of gradient estimators derived from the REINFORCE estimator (Williams 1992) or the Q-function. In contrast, the adaptive data setting we consider is applicable for any gradient estimator satisfying (3) and thereby covers a much larger range of gradient estimation strategies (Mohamed et al. 2020), including infinitesimal perturbation analysis (IPA) (Heidelberger et al. 1988, Glasserman 1992) and general likelihood-ratio gradient estimation (Glynn 1990). These gradient estimation strategies are outside the scope of the existing RL convergence analysis, but are of particular relevance to operations research applications, such as queueing and inventory management (see, e.g., Chen et al. (2023a), Huh et al. (2009)).

The core challenge of applying stochastic gradient descent (SGD) to the adaptive data setting is that z_t is not only determined by the current action θ_t but also depends on previous actions through z_{t-1}. As a result, using z_t to form a stochastic gradient estimator $g(\theta_t, z_t)$ leads to *biased* estimation of the true gradient $\nabla\ell(\theta_t)$, as $\mathbb{E}[g(\theta_t, z_t)|z_{t-1}] \neq \nabla\ell(\theta)$ in general. If the effects of previous actions persist in the system for a long time, the bias can be significant. SGD with biased gradients may

not be able to converge to the desired minimum (or even a small enough neighborhood of the minimum).

Consider, for example, the optimal pricing problem in a GI/GI/1 queue, where the arrival rate is determined by the price charged according to a demand function $\lambda(p)$. Let T_{t+1} denote the baseline interarrival time between the t-th and $(t+1)$-th arrivals, S_t denote the service time, and W_t denote the waiting time of customer t, i.e., the t-th arrival. For a fixed price p, by Lindley's recursion, $\{W_t : t \geq 0\}$ is a Markov chain satisfying

$$W_{t+1} = \left(W_t + S_t - \frac{T_{t+1}}{\lambda(p)} \right)^+.$$

Our goal is to choose the optimal price to maximize the revenue minus the long-run average cost of waiting, or equivalently,

$$\min_p \ \ell(p) = -\left(p\lambda(p) - h\lambda(p)\mathbb{E}_{\pi_p}[W] \right),$$

where π_p denotes the steady-state distribution of the Markov chain $\{W_t : t \geq 0\}$ with arrival rate $\lambda(p)$, and h denotes the holding/waiting cost per unit time per customer. If we are to update the price after each arrival and let p_t denote the price charged for the t-th arrival, then the waiting times satisfy

$$\tilde{W}_{t+1} = \left(\tilde{W}_t + S_t - \frac{T_{t+1}}{\lambda(p_{t+1})} \right)^+,$$

which is a non-stationary Markov process due to varying p_t's.

Note that the waiting time $\tilde{W}_t$ is not only affected by p_t, but also by the previous prices charged, i.e., p_s for $s < t$, as they affect the current congestion level in the queue. As a result, $\mathbb{E}[\tilde{W}_t] \neq \mathbb{E}_{\pi_{p_t}}[W_t]$ due to transient behavior of the Markov chain. In this case, using the current waiting time to estimate the steady-state waiting time leads to a biased estimation of $\nabla\ell(p_t)$, which may derail the convergence of the standard SGD updates to the optimal p^*. For example, if one sets very low prices $p_s \ll p_t$ for $s < t$, the waiting time $\tilde{W}_t$ could be much larger than $\mathbb{E}_{\pi_{p_t}}[W_t]$ due to the congestion induced by high demands in previous periods (as result of the low prices p_s's). Since $\tilde{W}_t$ will be an overestimate of the stationary waiting time $\mathbb{E}_{\pi_{p_t}}[W_t]$, using this estimate to update the price could cause the new price p_{t+1} to be too high. The oscillations due to delayed feedback can lead to instability, which is a well-known issue in control theory.

Optimization algorithms that use data streams to sequentially update the solution are often referred to as *online* algorithms. The convergence behavior of online algorithms using independent and identically distributed (iid) data stream has been well studied in the literature (Robbins and Monro 1951, Shapiro et al. 2021, Moulines and Bach 2011). More recent results have shown that the iid requirement can be relaxed. Duchi et al. (2012), Agarwal and Duchi (2012), Sun et al. (2018)

study the performance of online learning algorithms on a dependent data stream. In particular, they assume z_k's are generated from a fixed suitably ergodic Markov chain. But these results cannot be applied to our problem directly, since they would require that the invariant distribution does not depend on the policy θ. In contrast, in the service system design example discussed above, the price changes the transition dynamics of the queue, which changes the steady-state waiting time distribution.

One indirect way to apply the existing results with iid data to the adaptive data setting is to only update θ periodically, where the period (or batch size) $B \in \mathbb{N}^+$ is set to be long enough such that the data distribution is close to stationarity towards the end of the period. In other words, for all $t \in \{0, ..., B-1\}$ we would maintain the same policy $\theta_0 = ... = \theta_{B-1}$ in order for z_{B-1} to resemble a draw from the stationary distribution μ_{θ_0}. This would allow for a good estimate of $\nabla_\theta \mathbb{E}_{\mu_{\theta_0}}[\mathcal{L}(\theta_0, z)]$, which would then be used to update the policy at time B. One may even specify a schedule of batch sizes B_k to reliably control the non-stationarity induced by the actions (see, e.g., Chen et al. (2023a), Huh et al. (2009), Hu and Wager (2022)). However, reducing the frequency at which the policy is updated curtails the adaptivity of the algorithm, exposing a potential trade-off between adaptivity and bias from non-stationarity. Balancing this tradeoff would require carefully choosing the batch size B. This may require detailed knowledge about the ergodicity property, e.g., the exact rate of convergence of the underlying Markov chain, which is unavailable or hard to obtain in many applications. Overall, it is a priori unclear how the length of the period/batch size affects the performance of the learning algorithm in the adaptive setting.

The papers Mendler-Dünner et al. (2020) and Drusvyatskiy and Xiao (2023) study stochastic optimization when one can draw independent samples from the updated data distribution μ_θ. This is similar to the periodic updating design discussed above, because, in most practical settings, one can only generate independent samples from μ_θ by applying the same policy for a long enough time, i.e., running the Markov chain under P_θ until it reaches stationarity.

In this paper, we study the convergence of SGD where only one sample or a minibatch of samples, i.e., B_k's being a fixed $O(1)$ constant, is used at each iteration. Under certain ergodicity and continuity conditions on the Markov transition kernels, we show that SGD with adaptive data can achieve $O((\log T)^2/\sqrt{T})$ convergence to a stationary point in the nonconvex case. In the convex case, for projected SGD where the projection set is convex (can be $\mathbb{R}^d$, which corresponds to the case without projection), we show that it can achieve $O((\log T)^4/\sqrt{T})$ convergence to the optimal when ℓ is convex, and $O((\log T)^2/T)$ convergence when ℓ is strongly convex. These rates are similar to the iid case (Shapiro et al. 2021). We also show how the mixing time of the underlying Markov chains is incorporated into the convergence rate analysis. It is important to note that knowledge of the mixing time is not necessary for the implementation of the SGD algorithm. Overall, our

results show that under the conditions we specify, non-stationarity induced by the policy updates does not impose fundamental limitations on adaptivity.

Our finite-time convergence analysis for SGD with adaptive data can be applied to study a wide range of problems of practical relevance. In particular, we demonstrate how the analysis can be applied to study online learning algorithms for service and inventory systems. As mentioned before, our setting also covers policy-gradient approaches in RL, and we show how our analysis applies to an actor-critic policy gradient algorithm. These examples demonstrate that the assumptions we impose are easy to verify and widely applicable.

When applying SGD with adaptive data to service and inventory management problems, we demonstrate how to construct gradient estimators using the sample path derivative, which can be updated recursively. In particular, we augment the Markov chain to include both the original Markov chain and a derivative process. This sample-path derivative construction utilizes developments in the simulation literature on infinitesimal perturbation analysis (IPA) (Heidelberger et al. 1988, Glasserman 1992). However, in our applications of IPA, we need to establish stronger ergodicity results than the standard convergence results in the literature.

The rest of the paper is organized as follows. We conclude this section with a review of the literature to highlight our contribution. In Section 2, we present our main results – the finite-time performance bounds for SGD with adaptive data in various settings, i.e., nonconvex, convex with/without projection, and strongly convex with/without projection. In Sections 3 – 5, we demonstrate how to apply the main results to study various online learning problems. We also discuss how to apply the sample path derivative to construct gradient estimators in Sections 3 and 4. Lastly, we conduct numerical experiments to demonstrate the performance of the SGD algorithm with adaptive data in various applications in Section 6 and conclude in Section 7. All the proofs are provided in the appendices.

Literature Review

In this paper, we study the convergence rate of the SGD updates with adaptive data:

$$\theta_{t+1} = \theta_t - \eta_t g(\theta_t, z_t)$$

where $z_t \sim P_{\theta_t}(\cdot|z_{t-1})$, $g(\theta_t, z_t)$ is a gradient estimator, and we assume $\nabla l(\theta) = \mathbb{E}_{\mu_\theta}[g(\theta, z)]$. As discussed above, motivated by different applications, most previous literature on SGD either assumes iid data or dependent but non-adaptive data (see, e.g., (Moulines and Bach 2011, Ghadimi and Lan 2013, Agarwal and Duchi 2012, Chen et al. 2023b)). Extending stochastic gradient/ stochastic approximation algorithms to adaptive data is an active area of research (Benveniste et al. 2012). When the data is drawn adaptively from a policy-dependent Markov chain, some simple gradient

estimators can suffer from estimation biases even when evaluated according to the corresponding stationary distribution. Doucet and Tadic (2017) studies the asymptotic behavior of SGD with a biased gradient estimator. Karimi et al. (2019), Li and Wai (2022), Roy et al. (2022) further establishes non-asymptotic convergence results of SGD with adaptive data. They demonstrate how to apply the results to analyze the regularized online expectation maximization algorithm, policy gradient method, and strategic classification. Huh and Rusmevichientong (2014) study stochastic online optimization with biased gradients under dependent data. They introduce the notion of sequential convexity and propose an adaptive algorithm with regret guarantees, demonstrating its effectiveness in applications such as inventory control, capacity allocation, and lifetime buy problems with censored demand. Recently, Li et al. (2025) demonstrate how to apply the framework in (Benveniste et al. 2012, Karimi et al. 2019) to optimize the design of GI/GI/1 queues and base-stock inventory replenishment. Similar to (Karimi et al. 2019, Li and Wai 2022, Roy et al. 2022), our work also establishes non-asymptotic performance bounds. However, we require a different set of assumptions that are easier to verify and can be applied to many online learning problems in operations research. We next discuss some key differences.

First, most of the previous works require the stochastic gradient estimator, $g(\theta, z)$, to be Lipschitz continuous in θ (see, e.g., Assumptions 1 and 2 in Li and Wai (2022) and Assumption 2.4 in Roy et al. (2022)). However, this assumption does not hold in many applications, e.g., when $g(\theta, z)$ involves indicators. To handle this challenge, our results only require the population gradient $\nabla l(\theta)$ to be Lipschitz. This is a much weaker assumption since ∇l is the weighted average of $g(\theta, z)$'s, and its smoothness is in general satisfied (see, e.g., Proposition 2.1 of Roy et al. (2022)). Moreover, our analysis leverages the recent developments in the perturbation theory for Markov chains (Rudolf and Schweizer 2018) to explain why ∇l is Lipschitz in general.

Second, many existing works impose assumptions that require verifying certain continuity properties of the solution to a Poisson equation associated with P_θ and $g(\theta, \cdot)$ (see, e.g., assumptions A5 and A6 in (Karimi et al. 2019)). While these assumptions require only that a solution with the desired properties exists, rather than explicitly solving the Poisson equation, verifying such properties often demands substantial additional analytical effort (see, e.g., Appendix D in (Karimi et al. 2019) and the development in (Li et al. 2025)). In contrast, our results provide a set of assumptions that are more directly verifiable. In particular, the assumptions we impose are rather standard Markov chain mixing properties, which are fairly well known for many operations research applications.

Third, our work also provides a more complete picture of the convergence of SGD with adaptive data (nonconvex, convex with/without projection, and strongly convex with/without projection), while previous literature often analyzes convergence under only one or two settings. In addition,

with non-convex loss, the estimate (Corollary 1) in Li and Wai (2022) has a nonvanishing bias, while Roy et al. (2022) gives only $\tilde{O}(T^{-2/5})$ convergence rate, which is slower than the standard $\tilde{O}(T^{-1/2})$ rate. Our analysis combines Markov chain perturbation theory with IPA to achieve a faster rate of convergence.

From the application perspective, our work is related to the online learning literature in operations research, especially in queueing and inventory systems. This has been an active area of research. For example, Chen et al. (2023a) study learning pricing and capacity sizing in single server queues, Krishnasamy et al. (2021), Zhong et al. (2022) study learning the scheduling policy in multiclass queues. See Walton and Xu (2021) for a review of recent developments in learning in stochastic networks. Huh et al. (2009), Zhang et al. (2020a) study learning the replenishment policy in lost sales inventory systems. Tang et al. (2023) study learning dual-index policies in dual sourcing systems. Cheung et al. (2022) study learning how to allocate limited resources to heterogeneous customers. Our work complements these works by taking a closer look at how many samples need to be collected before a policy update, i.e., how often the policy can be updated. We provide easy-to-verify conditions under which updating the policy after each new sample or a constant batch of samples leads to a near-optimal rate of convergence.

Our work is also related to the literature on policy gradient in reinforcement learning, especially recent developments on sample complexity analysis of policy gradient algorithms (Wang et al. 2019, Zhang et al. 2020b, Xiong et al. 2021, Yuan et al. 2022, Xu et al. 2020). The convergence of policy gradient with exact gradient information –thus eliminating approximation errors – has been studied in several different settings (Mei et al. 2020, Agarwal et al. 2021, Xiao 2022, Bhandari and Russo 2024). However, when implementing policy gradient methods, a key challenge is how to estimate the gradient in a sample-efficient manner. For example, if we are to estimate the gradient by sampling the trajectories of the MDP under the current policy π^θ, what would be the horizon for the trajectory (i.e., where to truncate since we cannot sample an infinite-horizon trajectory) and how many trajectories do we need to sample? Note that in addition to the standard stochastic noise, we also need to handle the bias due to truncation (i.e., the underlying Markov chain has not reached stationarity yet). In this paper, we demonstrate how our finite-time convergence analysis for SGD with adaptive data can be applied to study an actor-critic policy gradient algorithm, where we use temporal difference learning (TD) to estimate the state-action value function/Q-function under policy π^θ. Our algorithm only requires one TD update in each iteration, which is substantially less than what is required in Wang et al. (2019), Yuan et al. (2022), Xiong et al. (2021), Xu et al. (2020). To achieve this, we study the Markov chain (s_t, a_t, Q_t) induced by the TD sampler with a constant learning rate, where s_t and a_t are the state and action visited and Q_t is the state-action value function (Q-function) at time t. We show that even though Q_t does not

converge to the desired Q-function, i.e., the Q-function under policy π^θ, Q_t evaluated under the corresponding stationary measure is equal to the desired Q-function.

The SGD-based methods/analyses have also been applied to conduct finite-time analysis of TD algorithms (Bhandari et al. 2018, Dalal et al. 2018, Srikant and Ying 2019, Qu and Wierman 2020). There, because the policy is fixed, i.e., the Markov chain dynamics is fixed, the data is Markovian but not adaptive, which is similar to the setting studied (Duchi et al. 2012, Agarwal and Duchi 2012, Sun et al. 2018).

Notations

Let (Ω, d) denote a metric space for the data stream z. We denote $\|\cdot\|$ as the L^2 norm. For a transition kernel P, we write $Pf(x) = \int P(x, dx')f(x')$, and we write the distribution of the Markov chain starting from x after n steps of transition as $\delta_x P^n$. For a nonnegative sequence $\{a_t : t \geq 0\}$ and a nonnegative function $f(t)$, we say $a_t = O(f(t))$ if there is a constant C such that $a_t \leq Cf(t)$ for any $t \geq 0$. We also use the notation $a_t = \tilde{O}(f(t))$ when we ignore the logrighmic factors, e.g., when $a_t = O(f(t)(\log f(t))^k)$, we can write $a_t = \tilde{O}(f(t))$. We say $a_t = \Omega(f(t))$ if there is a constant C such that $a_t \geq Cf(t)$ for any $t \geq 0$. Let μ and ν be two probability measures defined on a common measurable space Ω, and $d : \Omega \times \Omega \to [0, \infty)$ be a measurable cost function. We define the Wasserstein distance W_d as

$$W_d(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \mathbb{E}_{(x,y)\sim\gamma}[d(x, y)],$$

where $\Gamma(\mu, \nu)$ denotes the set of all couplings of μ and ν.

2. Main Results

To solve the stochastic optimization problem (1), we consider the SGD update:

$$\theta_{t+1} = \theta_t - \eta_t \hat{g}_t(\theta_t, z_t). \tag{4}$$

We assume the data sequence comes from a Markov chain controlled by θ_t. In particular, we assume there is a transition kernel P_{θ_t} on Ω such that

$$z_t \sim P_{\theta_t}(\cdot|z_{t-1}).$$

For each given θ, P_θ generates an ergodic Markov chain with invariant measure μ_θ. In addition, we assume there is a stochastic gradient estimator $g(\theta, z_t)$ satisfying

$$\nabla \ell(\theta) = \mathbb{E}_{\mu_\theta}[g(\theta, z)].$$

We will discuss how to find such a gradient estimator using the sample path derivative in Sections 3 and 4 for some examples. Note that $g(\theta, z_t)$ is unbiased only under the corresponding invariant measure. Due to the transience and non-stationarity of the underlying Markov chain, $\mathbb{E}[g(\theta_t, z_t)|\theta_t, z_{t-1}] \neq \nabla \ell(\theta_t)$ in general. Meanwhile, to accommodate more general settings, we assume $\hat{g}_t(\theta, z_t)$ in (4) is only an approximation of $g(\theta, z_t)$ with diminishing errors.

We make the following assumptions.

ASSUMPTION 1. *P_θ's have a common Lyapunov function $V \geq 1$ where*

$$P_\theta V(z) \leq \rho V(z) + K,$$

for some $\rho \in (0,1)$ and $K \in (1, \infty)$. In addition, P_θ's are Wasserstein contractive with respect to the metric d on Ω, i.e., for any $x, y \in \Omega$,

$$W_d(\delta_x P_\theta^n, \delta_y P_\theta^n) \leq K \rho^n d(x, y).$$

ASSUMPTION 2. *There exists $L > 0$ such that*

$$\|\nabla \ell(\theta) - \nabla \ell(\theta')\| \leq L \|\theta - \theta'\|,$$

$$\|g(\theta, z) - g(\theta, z')\| \leq L d(z, z'),$$

and

$$W_d(\delta_x P_\theta, \delta_x P_{\theta'}) \leq L \|\theta - \theta'\| V(x).$$

ASSUMPTION 3. *There exists $M \in (0, \infty)$, such that*

$$\|g(\theta, z)\| \leq M V(z), \|\hat{g}_t(\theta, z)\| \leq M V(z), \text{ and } \|\nabla l(\theta)\| \leq M.$$

ASSUMPTION 4. *There exits $C > 0$, such that for any $t \geq 0$,*

$$P_\theta^t V(z) \leq C(V(z)^4 + 1).$$

Note that a sufficient condition for Assumption 4 is that $V(z)^4$ is also a Lyapunov function.

ASSUMPTION 5. *There exists a stochastic sequence e_t with $\mathbb{E}[e_t^2] < \infty$, such that*

$$\|g(\theta, z_t) - \hat{g}_t(\theta, z_t)\| \leq e_t.$$

Assumption 1 requires that the underlying Markov chains are suitably ergodic. This assumption ensures that the stationary distribution μ_θ is well defined and the Markov chain converges to stationarity sufficiently fast. In particular, the existence of the Lyapunov function implies that the dynamics return to the "center" of the state space regularly and the length of the excursions from

the center can be properly controlled. Note that by Harris' ergodic theorem, the existence of the Lyapunov function together with a uniform "minorization" condition localized to the interior of a level set implies geometric ergodicity and thus Wasserstein contraction under an appropriate metric (Roberts and Rosenthal 1997). This Markov chain convergence framework has been well-studied in the literature (Meyn and Tweedie 2012). For many existing models, e.g., queueing models, inventory models, etc, Assumption 1 has already been verified.

The first condition in Assumption 2 requires $\nabla\ell$ to be Lipschitz continuous, which is satisfied in many applications. This condition is weaker and more broadly applicable than requiring $g(\theta, z)$ to be Lipschitz continuous in θ, a common assumption in the literature (see e.g., Proposition 2.1 of Roy et al. (2022) and Assumption A13 of Karimi et al. (2019)). For example, $g(\theta, z) = 1\{z \le \theta\}$ is not Lipschitz in θ. We also require $g(\theta, z)$ to be Lipschitz continuous in z, but have some flexibility in choosing the metric d. In particular, by choosing an appropriate d, this condition is easily satisfied in many applications. For example, if we consider the total variation distance, $d(z, z') = 2 \cdot 1\{z \neq z'\}$, then this condition will be a consequence of Assumption 3. More generally, we allow considerable flexibility in the choice of the metric d, and hence the corresponding Wasserstein distance W_d, rather than working exclusively with total variation distance. This flexibility proves useful in applications, as we demonstrate in Sections 3 – 5, where we apply problem-specific metric d. Additionally, we require $\delta_x P_\theta$ to be Lipschitz continuous in θ. Note that this condition is for the one-step transition kernel, which is easy to verify in practice. By Markov chain perturbation theory, the Lipschitz continuity of the one-step transition kernel together with the ergodicity condition, i.e, Assumption 1, implies the Lipschitz continuity of the corresponding stationary measure (Rudolf and Schweizer 2018).

Assumption 3 imposes some boundedness conditions, which are weaker than those commonly adopted in the literature on gradient-based algorithms with Markovian or adaptive data streams. In particular, we allow the magnitude of $g(\theta, z)$ to grow with a measurable function $V(z)$ and require only that $|g(\theta, z)| \le MV(z)$ for some constant $M > 0$, rather than imposing a uniform bound independent of z (see, e.g., Assumption A6 in Karimi et al. (2019) and Assumption 5 in (Sun et al. 2018)).

For the ease of exposition, we also introduce the concept of mixing time.

Definition 1. The Markov chain P_θ has a mixing time $\tau < \infty$ if for any $z \in \Omega$,

$$W_d(\delta_z P_\theta^\tau, \mu_\theta(\cdot)) \le \frac{1}{4} V(z).$$

Note that we incorporate the Lyapunov function V in the definition of the mixing time. This formulation reflects the fact that we are dealing with Markov chains on general (possibly unbounded)

state spaces. In such settings, uniform convergence rates are often unattainable. The Lyapunov function V serves as a tool to capture the "size" of the initial state z, enabling us to express convergence rates that adapt to the location of the initial condition. Under Assumption 1 and with a suitably chosen Lyapunov function, for example, one satisfying $d(x,y) \le V(x) + V(y)$, we have $\tau = O(1/|\log \rho|)$ (Cui et al. 2025).

Nonconvex case: We have the following convergence result for a general loss function l.

THEOREM 1. *Suppose Assumptions 1 – 5 hold and P_θ has a mixing time τ. The iterates according to* (4) *satisfy*

$$\min_{0\le t<T} \mathbb{E}\|\nabla \ell(\theta_t)\|^2 = O\left(\frac{1}{\sum_{t=0}^{T-1}\eta_t}\left(\tau \log T + \tau \log T \sum_{t=0}^{T-1}\eta_t^2 + \frac{1}{\sqrt{T}}\sum_{t=0}^{T-1}\eta_t + \sum_{t=0}^{T-1}\eta_t \mathbb{E}e_t\right)\right),$$

where O hides a polynomial of M and L. If we fix $\eta_t = \eta_0 t^{-1/2}$ for some $\eta_0 > 0$, and assume $\mathbb{E}e_t = O(1/\sqrt{t})$, we can further simplify the bound to

$$\min_{0\le t<T} \mathbb{E}\|\nabla \ell(\theta_t)\|^2 = O\left(\tau(\log T)^2/\sqrt{T}\right).$$

Theorem 1 implies that the SGD updates can achieve an $O(\tau(\log T)^2/\sqrt{T})$ convergence to a stationary point, which is similar to the $O(\log(T)/\sqrt{T})$ convergence in Theorem 2 of (Karimi et al. 2019), for which they also establish a matching lower bound. Here, we explicitly characterize how the mixing time affects the convergence rate. The convergence results in (Karimi et al. 2019) have an extra asymptotic bias term, which does not arise in our setting since our result works under a different set of assumptions and a more refined gradient estimator.

REMARK 1. Assumption 5, along with the requirement that $\mathbb{E}[e_t] = O(1/\sqrt{t})$, is introduced to handle scenarios where it is not possible to construct a gradient estimator $g(\theta, z)$ satisfying $\nabla \ell(\theta) = \mathbb{E}_{\mu_\theta}[g(\theta, z)]$, and one must instead rely on an "approximate" estimator $\hat{g}_t$. In the examples presented in Sections 3 and 4, we are able to construct an exact estimator $g(\theta, z)$ using the sample path derivative. In general, designing effective gradient estimators is a nontrivial task. The sample path derivative approach, which is also known as the IPA method in the simulation literature (Heidelberger et al. 1988, Glasserman 1992), is a versatile and analytically grounded approach that facilitates the construction of gradient estimators in many operations research applications. However, when such estimators are not available, one must resort to approximate methods that introduce additional estimation error. To ensure convergence in these settings, it is necessary to improve the estimator's accuracy over time. For example, when using finite-difference approximations, the perturbation size must be carefully reduced as a function of t to control the bias.

Convex case: For general (constrained) convex optimization, we consider a projected SGD update:

$$\theta_{t+1} = \mathcal{P}_\Theta(\theta_t - \eta_t \hat{g}_t(\theta_t, z_t)), \tag{5}$$

where Θ is a properly defined convex set, and $\mathcal{P}_\Theta$ is the associated projection. The set Θ can be the constrained set for constrained optimization, or $\mathbb{R}^m$ in the unconstrained setting. The projection allows us to relax Assumptions 1 – 5 such that they only need to hold for $\theta \in \Theta$.

Let θ^* denote a minimizer of $\ell(\theta)$. Define the weighted average of the iterates as

$$\bar{\theta}_T = \frac{1}{\sum_{t=0}^{T-1} \eta_t} \sum_{t=0}^{T-1} \eta_t \theta_t.$$

THEOREM 2. *Suppose Assumptions 1 – 5 hold and P_θ has a mixing time τ under the restriction that $\theta \in \Theta$.*

1. If ℓ is convex and $\mathbb{E}[e_t] = O(1/\sqrt{t})$, by setting the step size as $\eta_t = \eta_0/\sqrt{t}$, the weighted average of iterates according to (5) *satisfy*

$$\mathbb{E}\ell(\bar{\theta}_T) - \ell(\theta^*) = O\left(\tau^2 (\log T)^4 / \sqrt{T}\right),$$

where O hides a polynomial of M and L.

2. If ℓ is strongly convex with a convexity constant c and $\mathbb{E}[e_t^2] = O(1/t)$, by setting the step size as $\eta_t = 2\eta_0/(ct)$ for $t \geq 1$ with $\eta_0 > 2$, the iterates according to (5) *satisfy*

$$\mathbb{E}\|\theta_T - \theta^*\|^2 = O\left(\tau (\log T)^2 / T\right),$$

where O hides a polynomial of M and L.

Theorem 2 indicates that when ℓ is strongly convex, we achieve $O(\tau(\log T)^2/T)$ convergence rate, which is similar to the $O(1/T)$ convergence rate established in Theorem 1 of (Li and Wai 2022). In Appendix A, we provide a lower bound on the convergence rate of the SGD update (4) when ℓ is strongly convex, showing that $\mathbb{E}[\|\theta_T - \theta^*\|^2] = \Omega(\tau/T)$.

In the convex (but not strongly convex) case, the dependence on τ is likely suboptimal and primarily reflects limitations in our proof techniques. In particular, since we allow the parameter space Θ to be unbounded, we have to develop some bounds for $\mathbb{E}[\|\theta_t - \theta^*\|]$, which depends on $\tau(\log T)^2$ (see Lemma 10 in Appendix B). If Θ is bounded, we would directly have $\|\theta_t - \theta^*\| \leq C$ for some constant $C < \infty$, which would yield an improved convergence rate of $O\left(\tau(\log T)^2/\sqrt{T}\right)$.

In what follows, we will demonstrate how to apply Theorems 1 and 2 to various applications. In particular, we will show that the assumptions required in the theorems are satisfied and easy to verify in many online learning problems in operations research. We will also discuss an extension to an actor-critic policy gradient algorithm.

3. Inventory Control with Stock-Out Damping

We consider the problem of selecting a base-stock level in a single-product multi-period inventory system with endogenous demand. Motivated by recent empirical findings, demand is temporarily reduced whenever a stock-out occurs (Anderson et al. 2006). We model demand as a Markovian autoregressive process subject to a dampening effect when demand exceeds the base-stock level:

$$D_{t+1} = (\alpha \min\{D_t + u_{t+1}, \theta\} + (1-\alpha)m + \epsilon_{t+1})^+, \tag{6}$$

where $\epsilon_t \overset{iid}{\sim} N(0,\sigma^2)$, $u_t \overset{iid}{\sim} N(0,\sigma^2)$, u_t is independent of ϵ_t, and $(\cdot)^+ = \max\{\cdot, 0\}$. The goal is to select the base-stock level $\theta \in \Theta := [0, \bar{\theta}]$, for any finite upper bound $\bar{\theta} > 0$, to minimize the underage and overage costs under the stationary distribution induced by θ:

$$\min_{\theta\in\Theta} \ell(\theta) := \mathbb{E}_{\mu_\theta}\left[h(\theta - D_t)^+ + b(D_t - \theta)^+\right]. \tag{7}$$

Note that the base-stock level endogenously affects the demand dynamics, and we write $D_t(\theta)$ to mark the dependence explicitly when necessary. To obtain an appropriate gradient estimator, we consider taking a pathwise derivative of $D_t(\theta)$:

$$\begin{aligned} L_{t+1}(\theta) := \frac{dD_{t+1}(\theta)}{d\theta} =& \frac{\partial}{\partial D_t}(\alpha \min\{D_t(\theta) + u_{t+1}, \theta\} + (1-\alpha)m + \epsilon_{t+1})^+ \frac{dD_t(\theta)}{d\theta} \\ &+ \frac{\partial}{\partial \theta}(\alpha \min\{D_t(\theta) + u_{t+1}, \theta\} + (1-\alpha)m + \epsilon_{t+1})^+ \\ =& \alpha 1\{D_{t+1}(\theta) > 0\} 1\{D_t(\theta) + u_{t+1} \le \theta\} L_t(\theta) \\ &+ \alpha 1\{D_{t+1}(\theta) > 0\} 1\{\theta < D_t(\theta) + u_{t+1}\}. \end{aligned}$$

This gives us a recursive way to update L_t, i.e.,

$$L_{t+1}(\theta) = 1\{D_{t+1}(\theta) > 0\}\alpha\left(1\{D_t(\theta) + u_{t+1} > \theta\} + 1\{D_t(\theta) + u_{t+1} \le \theta\} L_t(\theta)\right).$$

Consider the augmented Markov chain $Z_t(\theta) = (D_t(\theta), L_t(\theta))$. Our next result shows that $Z_t(\theta)$ is well defined with a proper stationary distribution.

Lemma 1. *For any $\theta \in \Theta$, the pathwise derivative $L_t(\theta)$ exists and the Markov chain $Z_t(\theta) = (D_t(\theta), L_t(\theta))$ converges in distribution to $Z_\infty(\theta) = (D_\infty(\theta), L_\infty(\theta))$ as $t \to \infty$ with*

$$\nabla \ell(\theta) = \mathbb{E}\left[\left(h1\{D_\infty(\theta) < \theta\} - b1\{D_\infty(\theta) \ge \theta\}\right)(1 - L_\infty(\theta))\right].$$

Based on Lemma 1, we have the following gradient estimator:

$$g(\theta, Z_t) = \left(h1\{D_t < \theta\} - b1\{D_t \ge \theta\}\right)(1 - L_t). \tag{8}$$

The SGD update using the gradient estimator defined in (8) with step-size η_t then takes the form

$$\theta_{t+1} = \mathcal{P}_\Theta\left(\theta_t - \eta_t g(\theta_t, Z_t)\right).$$

THEOREM 3. *Suppose the objective function* (7) *is convex and let* θ^* *be a minimizer. Then, using the gradient estimator* $g(\theta, Z_t)$ *defined in* (8) *with step-size* $\eta_t = \eta_0 t^{-1/2}$*, the average SGD iterate* $\bar{\theta}_T$ *satisfies*

$$\mathbb{E}[\ell(\bar{\theta}_T) - \ell(\theta^*)] = O\left(\tau^2 (\log T)^4 / \sqrt{T}\right),$$

where τ *is an upper bound of the mixing time of the augmented Markov chain* $Z_t(\theta)$ *for* $\theta \in \Theta$*. Moreover, if the objective function is strongly convex with a convexity constant* c*, by setting the step size as* $\eta_t = 2\eta_0/(ct)$ *for* $t \geq 1$ *with* $\eta_0 > 2$*, we have*

$$\mathbb{E}[\|\theta_T - \theta^*\|^2] = O\left(\tau (\log T)^2 / T\right).$$

Note that the existing approaches, i.e., those in (Li and Wai 2022, Roy et al. 2022), require Lipschitzness of the gradient estimator g in both θ and z under the Euclidean distance, which does not hold in this example. In contrast, our framework only requires Lipschitzness of $\nabla \ell(\theta)$ in θ. While the gradient estimator is highly non-smooth, the averaged gradient is Lipschitz. In fact, the Lipschitzness of $\nabla \ell(\theta)$ is a direct consequence of the Lipschitzness of the transition kernel P_θ, as we show in the proof of Theorem 3. At a high level, our framework reveals that randomness in the transition dynamics implicitly smooths the gradient estimator, enabling a greater range of gradient estimators while maintaining convergence guarantees.

4. Pricing and Capacity Sizing in Single-Server Queue

We consider the problem of pricing and capacity sizing in a single-server queue, as studied in Chen et al. (2023a). Consider a $GI/GI/1$ queue, i.e., a single-server queue with generally distributed interarrival times and service times. The interarrival times and service times are scaled by the arrival rate and service rate, respectively. The arrival rate is determined by the price charged according to a known demand function $\lambda(p)$ for feasible prices $p \in [\underline{p}, \bar{p}]$. The service provider also selects the service rate $\mu \in [\underline{\mu}, \bar{\mu}]$, which incurs a service cost $c(\mu)$ per unit of time. Let T_{t+1} denote the baseline interarrival time between the t-th and $(t+1)$-th arrivals, S_t denote the baseline service time, and W_t denote the waiting time of customer t. For given μ and p, the system dynamics follow

$$W_{t+1} = \left(W_t + \frac{S_t}{\mu} - \frac{T_{t+1}}{\lambda(p)}\right)^+. \tag{9}$$

Let $\Theta = [\underline{\mu}, \bar{\mu}] \times [\underline{p}, \bar{p}]$ denote the feasible set of service rates and prices. The goal of the service provider is to select $(\mu, p) \in \Theta$, i.e., a fixed service rate μ and price p, to maximize expected net profit, which is the long-run average revenue minus the service cost and holding cost:

$$\max_{(\mu,p)\in\Theta} V(\mu, p) := p\lambda(p) - c(\mu) - h_0 \mathbb{E}[Q_\infty(\mu, p)], \tag{10}$$

where $Q_\infty(\mu,p)$ is the stationary queue length (number of people waiting in the system) under (μ,p), and h_0 is the per-unit-time per-customer holding cost (cost of waiting). By Little's law, the maximization problem in (10) is equivalent to the following minimization problem involving the stationary waiting time $W_\infty(\mu,p)$ under (μ,p):

$$\min_{(\mu,p)\in\Theta} \ell(\mu,p) := h_0\lambda(p)\left(\mathbb{E}\left[W_\infty(\mu,p)\right]+\frac{1}{\mu}\right)+c(\mu)-p\lambda(p). \tag{11}$$

To obtain an appropriate gradient estimator, we consider taking a pathwise derivative of $W_t(\mu,p)$. Define

$$\begin{aligned}
L_{\mu,t+1}(\mu,p) &:= \frac{\partial W_{t+1}(\mu,p)}{\partial\mu} = \left(\frac{\partial W_t(\mu,p)}{\partial\mu}-\frac{S_t}{\mu^2}\right)1\{W_{t+1}(\mu,p)>0\},\\
L_{p,t+1}(\mu,p) &:= \frac{\partial W_{t+1}(\mu,p)}{\partial p} = \left(\frac{\partial W_t(\mu,p)}{\partial p}+\frac{T_t}{\lambda(p)^2}\lambda'(p)\right)1\{W_{t+1}(\mu,p)>0\}.
\end{aligned}$$

For the augmented Markov chain $(W_t, L_{\mu,t}, L_{p,t})$, by verifying the conditions in Glasserman (1992), we can show that $L_{\mu,t}$ and $L_{p,t}$ are well-defined, the Markov chain $(W_t, L_{\mu,t}, L_{p,t})$ converges to a unique stationary distribution, and

$$\mathbb{E}[L_{\mu,\infty}(\mu,p)] = \frac{\partial\mathbb{E}[W_\infty(\mu,p)]}{\partial\mu}, \quad \mathbb{E}[L_{p,\infty}(\mu,p)] = \frac{\partial\mathbb{E}[W_\infty(\mu,p)]}{\partial p}.$$

Indeed, this has been studied in Chen et al. (2023a). For the single-server queue, we can achieve further simplification of the derivative processes by considering a simpler augmented Markov chain $Z_t=(W_t,X_t)$, where X_t denotes the server's busy time seen by the t-th arrival. In particular, the dynamics of Z_t are as follows:

$$\begin{aligned}
W_{t+1} &= \left(W_t+\frac{S_t}{\mu}-\frac{T_t}{\lambda(p)}\right)^+\\
X_{t+1} &= \left(X_t+\frac{T_t}{\lambda(p)}\right)1\{W_{t+1}>0\}.
\end{aligned}$$

Lemma 5 in Chen et al. (2023a) shows that

$$\begin{aligned}
\frac{\partial}{\partial p}\ell(\mu,p) &= -\lambda(p)-p\lambda'(p)+h_0\lambda'(p)\left(\mathbb{E}[W_\infty(\mu,p)]+\mathbb{E}[X_\infty(\mu,p)]+\frac{1}{\mu}\right)\\
\frac{\partial}{\partial\mu}\ell(\mu,p) &= c'(\mu)-h_0\frac{\lambda(p)}{\mu}\left(\mathbb{E}[W_\infty(\mu,p)]+\mathbb{E}[X_\infty(\mu,p)]+\frac{1}{\mu}\right).
\end{aligned} \tag{12}$$

Based on (12), we define

$$g(\mu,p,Z_t) = (g_p(\mu,p,Z_t), g_\mu(\mu,p,Z_t)),$$

where

$$\begin{aligned}
g_p(\mu,p,Z_t) &= -\lambda(p)-p\lambda'(p)+h_0\lambda'(p)\left(W_t+X_t+\frac{1}{\mu}\right),\\
g_\mu(\mu,p,Z_t) &= c'(\mu)-h_0\frac{\lambda(p)}{\mu}\left(W_t+X_t+\frac{1}{\mu}\right).
\end{aligned} \tag{13}$$

Then, the SGD update for $\theta_t = (\mu_t, p_t)$ with the gradient estimator (13) and step-size η_t takes the form

$$\theta_{t+1} = \mathcal{P}_\Theta\left(\theta_t - \eta_t g(\theta_t, Z_t)\right).$$

To prove the convergence of the SGD update, we impose some regularity conditions on the functions λ and c and the distributions of the baseline interarrival time and service time, T and S.

ASSUMPTION 6. *For the demand and cost functions we have:*

(i) The constraint set $\Theta = [\underline{p}, \overline{p}] \times [\underline{\mu}, \overline{\mu}]$ is such that $\lambda(\underline{p}) < \underline{\mu}$.

(ii) $\lambda(p) \in C^2$ on $[\underline{p}, \overline{p}]$ and non-increasing in p.

(iii) $c(\mu) \in C^2$ on $[\underline{\mu}, \overline{\mu}]$ and non-decreasing in μ.

ASSUMPTION 7. *The baseline service time S and inter-arrival time T are iid respectively and*

(i) There exists $\alpha^* > 0$ such that $\mathbb{E}[e^{4\alpha^* S}] < \infty$ and $\mathbb{E}[e^{4\alpha^* T}] < \infty$.

(ii) There exist $0 < \alpha_2 < \alpha_1 < \min\{\alpha^*/\underline{\mu}, \alpha^*/\lambda(\underline{p})\}$ such that

$$\mathbb{E}\left[e^{4\alpha_1 \frac{T}{\underline{\mu}}}\right] \mathbb{E}\left[e^{-4(\alpha_1 - \alpha_2)\frac{S}{\lambda(\underline{p})}}\right] < 1.$$

(iii) The density functions of T and S, which are denoted as f_T and f_S respectively, are continuously differentiable. In addition, there exist $c, D_1, D_2 > 0$ and $k \in \mathbb{N}_+$ such that for all $x \geq \frac{c}{\min\{\underline{\mu}, \underline{\lambda}\}}$, $\left|\frac{d}{dx}\log f_T(x)\right| \leq D_1 + D_2|x|^k$ and $\left|\frac{d}{dx}\log f_S(x)\right| \leq D_1 + D_2|x|^k$. Lastly, $f_S(\mu x) \leq C(f_S(\underline{\mu} x) + f_S(\bar{\mu} x))$ for $\mu \in [\underline{\mu}, \bar{\mu}]$ and $f_T(\lambda x) \leq C(f_T(\underline{\lambda} x) + f_T(\bar{\lambda} x))$ for $\lambda \in [\underline{\lambda}, \bar{\lambda}]$.

Assumption 7 requires S and T to be sufficiently light-tailed and their density functions to be smooth enough, which enables us to verify that the transition kernel is Lipschitz continuous in the arrival and service rates. Commonly used service and interarrival time distributions, such as exponential, Erlang, Weibull(λ, k) with $k \geq 1$, etc., satisfy Assumption 7. For example, the Weibull$(1, k)$, $k \geq 1$, satisfies that for $x \geq 1$:

$$\frac{d}{dx}\log f(x) = \frac{d}{dx}\left[k \log x - x^k\right] \leq k + k|x|^{k-1}.$$

In addition, for $\mu \in [\underline{\mu}, \bar{\mu}]$,

$$f(\mu x) \leq \left(\frac{\bar{\mu}}{\underline{\mu}}\right)^{k-1} (f(\underline{\mu} x) + f(\bar{\mu} x)).$$

Under Assumptions 6 and 7, we can verify the conditions of Theorem 2 (Assumptions 1 – 5 restricted to $\theta \in \Theta$), which leads to the following theorem.

THEOREM 4. *Suppose Assumptions 6 and 7 hold. Suppose the objective (11) is convex and let (μ^*, p^*) be a minimizer. Then, using the gradient estimator $g(\mu, p, z_t)$ defined in (13) with step-size $\eta_t = \eta_0 t^{-1/2}$, the average SGD iterate $\bar{\theta}_T = (\bar{\mu}_T, \bar{p}_T)$ satisfies*

$$\mathbb{E}[\ell(\bar{\mu}_T, \bar{p}_T)] - \ell(\mu^*, p^*) = O\left(\tau^2 (\log T)^4 / \sqrt{T}\right),$$

where τ is an upper bound of the mixing time of the augmented Markov chain $Z_t(\theta)$ for $\theta \in \Theta$. Moreover, if the objective function is strongly convex with a convexity constant c, by setting the step size as $\eta_t = 2\eta_0/(ct)$ for $t \geq 1$ with $\eta_0 > 2$, we have

$$\mathbb{E}\|\theta_T - \theta^*\|^2 = O\left(\tau(\log T)^2/T\right).$$

The online learning algorithm proposed by Chen et al. (2023a) achieves a regret that is logarithmic in the number of customers assuming ℓ is strongly convex, and Theorem 4 yields a similar regret. However, their algorithm requires calibrating the number of customers seen before making a gradient update. The algorithm we consider does not require such calibration, i.e., we can update the parameter after each arrival. For this specific example, our convergence result is similar to the one developed in (Li et al. 2025). However, the development in (Li et al. 2025) requires verifying the Lipschitz continuity of the corresponding Poisson equation solution, which is more involved.

5. Application to policy gradient in reinforcement learning

We consider the classic Markov decision process (MDP) with a finite state space $\mathcal{S}$, a finite action space $\mathcal{A}$, a collection of transition probabilities $\{P(\cdot|s,a)\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$, and an initial distribution $\rho(\cdot)$. The policy is parameterized by θ, where $\pi^\theta(a|s)$ denotes the probability of taking action a when in state s. We focus on the infinite-horizon discounted cost formulation with the discount factor $\gamma \in (0,1)$:

$$\min_\theta \ell(\theta) := \mathbb{E}_\rho^\theta \left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t)\right], \tag{14}$$

where $c(s,a)$ is the expected instantaneous cost incurred by taking action a at state s.

When applying policy gradient to solve (14), the gradient can be expressed as (Sutton and Barto 2018)

$$\nabla\ell(\theta) = \sum_{s,a} \nu^\theta(s,a) Q^\theta(s,a) \nabla_\theta \log \pi^\theta(a|s), \tag{15}$$

where Q^θ is the state-action value function under policy π^θ,

$$\nu^\theta(s,a) = \frac{1}{1-\gamma} \sum_{t=0}^{\infty} \mathbb{E}_\rho^\theta[\gamma^t 1\{s_t = s, a_t = a\}]$$

is the state-action occupancy measure, which can also be viewed as the stationary distribution of the Markov chain with transition kernel

$$\mathcal{P}^\theta((s,a),(s',a')) := (1-\gamma)\rho(s')\pi^\theta(a'|s') + \gamma P(s'|s,a)\pi^\theta(a'|s').$$

In general, $\ell(\theta)$ is a non-convex function of θ. Recent works have shown that under certain regularity conditions, any stationary point of the policy gradient loss function is globally optimal (Agarwal

et al. 2021, Bhandari and Russo 2024, Wang et al. 2019). For instance, for finite state and action MDPs with natural parameterization, Bhandari and Russo (2024) show that $\ell(\theta)$ has no suboptimal stationary point. In addition, (Agarwal et al. 2020) show that $\nabla\ell(\theta)$ is Lipschitz continuous.

From the gradient formula (15), we can take $Q^\theta(s,a)\nabla_\theta \log \pi^\theta(a|s)$ as a gradient estimator. However, if (s,a)'s are not sampled from $\nu^\theta(s,a)$, the gradient estimator is biased. In addition, $Q^\theta(s,a)$ also needs to be estimated. One way to overcome the challenge is to simulate the Markov chain under policy π^θ for a long enough time so that we get an accurate enough estimate Q^θ and the distribution of (s_t,a_t) is close enough to $\nu^\theta(s,a)$. This idea has been employed in the literature (see, e.g., Wang et al. (2019), Xu et al. (2020), Xiong et al. (2021)). In this section, we are interested in understanding how adaptive the policy gradient algorithm can be while still achieving fast convergence to a stationary point.

We consider an actor-critic scheme where we update the state-action value function using the following temporal-difference (TD) update:

$$\begin{aligned} &Q_{t+1}(s_t,\hat{a}_t) = Q_t(s_t,\hat{a}_t) + \alpha[c(s_t,\hat{a}_t) - Q_t(s_t,\hat{a}_t) + \gamma Q_t(s'_{t+1},a'_{t+1})],\\ &Q_{t+1}(s,a) = Q_t(s,a) \text{ for } (s,a)\neq(s_t,\hat{a}_t), \end{aligned} \tag{16}$$

where $\hat{a}_t$ is sampled uniformly at random from $\mathcal{A}$, and (s'_{t+1},a'_{t+1}) is a random state generated from $\mathcal{P}^{\theta_t}((s_t,\hat{a}_t),\cdot)$, independent of (s_{t+1},a_{t+1}), which is generated from $\mathcal{P}^{\theta_t}((s_t,a_t),\cdot)$. Denote $Z_t=(s_t,a_t,Q_t)$.

We first establish the ergodicity property of the Markov chain $Z_t=(s_t,a_t,Q_t)$ under a fixed policy π^θ. In particular, given Z_t, Z_{t+1} is generated as follows: Sample $\hat{a}_t$ uniformly at random from $\mathcal{A}$. Sample (s_{t+1},a_{t+1}) from $\mathcal{P}^{\theta_t}((s_t,a_t),\cdot)$, and independently, sample (s'_{t+1},a'_{t+1}) from $\mathcal{P}^{\theta_t}((s_t,\hat{a}_t),\cdot)$. Update Q_{t+1} according to (16).

For a Markov chain s_t with a finite state space $\mathcal{S}$ and a unique stationary distribution ν. Let $\kappa_s=\inf\{t\geq 0: s_t=s\}$. We define the hitting time (Levin and Peres 2017) as

$$t_{hit}=\max_{s,s'\in\mathcal{S}}\mathbb{E}_s[\kappa_{s'}].$$

In addition, recall that under the total variation distance $\|\cdot\|_{\text{TV}}$, the mixing time of s_t, t_{mix}, satisfies

$$\max_{s\in\mathcal{S}}\|P_s^{t_{mix}}-\nu\|_{\text{TV}}\leq\frac{1}{4}.$$

Proposition 1. *For a fixed value of θ (i.e., under a fixed policy π^θ), suppose s_t is a finite-state Markov chain with a mixing time t_{mix} under the total variation distance and a finite hitting time t_{hit}. In addition, suppose $Q_t(s,a)\leq M$ for some $M<\infty$. Then, the triad $Z_t=(s_t,a_t,Q_t)$ induced by the TD sampler is a Markov chain with an ϵ-mixing time, $\bar{\eta}_\epsilon$, satisfying*

$$\begin{aligned}\bar{\eta}_\epsilon \leq& \frac{|\log(\epsilon/(8M+4))|}{|\log(1/4)|}t_{mix}\\ &+\max\left\{\frac{|\log(\epsilon/(4M))|}{|\log(1-\alpha+\alpha\gamma)|},12|\log(\epsilon/(8M+4))|\right\}(1+t_{hit})|\mathcal{A}|\sum_{k=1}^{|\mathcal{S}||\mathcal{A}|-1}\frac{1}{k}\end{aligned}$$

under the metric

$$\tilde{d}(z,\tilde{z}) = 1\{s \neq \tilde{s}, a \neq \tilde{a}\} + \|Q - \tilde{Q}\|_\infty.$$

PROPOSITION 2. *Under the assumptions of Proposition 1, for α in the TD sampler takes value outside a finite set of values, the invariant distribution of the triad $Z_t = (s_t, a_t, Q_t)$, μ^θ, satisfies*

$$\mathbb{E}_{\mu^\theta}[1\{s_t = s, a_t = a\}] = \nu^\theta(s,a), \quad \mathbb{E}_{\mu^\theta}[Q_t(s_t,a_t)|s_t = s, a_t = a] = Q^\theta(a,s).$$

We note from Propostion 2 that Q_{t+1} does not converge to Q^θ even if (s_t, a_t)'s are generated under π^θ. Instead, (s_t, a_t, Q_t) will be a Markov chain with the stationary measure μ^θ satisfying

$$\mathbb{E}_{\mu^\theta}[Q_t(s_t,a_t)|s_t = s, a_t = a] = Q^\theta(s,a).$$

REMARK 2. In the TD update (16), we require sampling two independent samples, (s_{t+1}, a_{t+1}) and (s'_{t+1}, a'_{t+1}), drawn from $\mathcal{P}^{\theta_t}((s_t, \hat{a}_t), \cdot)$ and $\mathcal{P}^{\theta_t}((s_t, a_t), \cdot)$, respectively. It is important to note that without this double sampling, the standard constant step-size TD update will lead to a bias of order α (Huo et al. 2023), i.e., $\mathbb{E}_{\mu^\theta}[Q_t(s_t,a_t)|s_t = s, a_t = a] = Q^\theta(s,a) + C\alpha + O(\alpha^2)$. In addition, the update in (16) requires sampling $\hat{a}_t$ uniformly at random from $\mathcal{A}$ rather than following the current policy π^{θ_t}. This is to ensure that all state-action pairs, (s,a)'s, are visited sufficiently often during the Q-value updates. By doing so, we avoid imposing some strong sufficient exploration condition on π^{θ_t}.

Based on Propositions 1 and 2, Z_t generated under a fixed policy π^θ is a properly defined ergodic Markov chain with stationary distribution μ^θ, where μ^θ satisfies

$$\mathbb{E}_{\mu^\theta}[Q_t(s_t,a_t)\nabla_\theta \log \pi^\theta(a_t|s_t)] = \sum_{s\in\mathcal{S}, a\in\mathcal{A}} \nu^\theta(s,a) Q^\theta(s,a) \nabla_\theta \log \pi^\theta(a|s).$$

This indicates that we only need "one transition" for each policy update. Algorithm 1 summarizes our actor-critic scheme.

We next establish the convergence of Algorithm 1. We first introduce some assumptions about the MDP.

ASSUMPTION 8. *The instantaneous costs are bounded, i.e., $|c(s,a)| \leq \tilde{M}$, almost surely. Q_0 is initialized such that $\|Q_0\|_\infty \leq \tilde{M}/(1-\gamma)$.*

ASSUMPTION 9. *The initial distribution $\rho(s) > 0$ for all $s \in \mathcal{S}$.*

ASSUMPTION 10. *For all $\theta \in \Theta$ and all $a \in \mathcal{A}$ and $s \in \mathcal{S}$, $\nabla_\theta \log \pi^\theta(a|s)$ is bounded and Lipschitz continuous in θ, i.e. there exists $R, L \in (0,\infty)$ such that for any $\theta, \theta' \in \Theta$,*

$$||\nabla_\theta \log \pi^\theta(a|s)|| \leq R, \quad ||\nabla_\theta \log \pi^\theta(a|s) - \nabla_{\theta'} \log \pi^{\theta'}(a|s)|| \leq L||\theta - \theta'||.$$

Algorithm 1: Actor-critic based policy gradient

1 Initialize s_0, a_0, Q_0, θ_0. Set $t=0$. ;

2 Sample $\hat{a}_t$ uniformly at random from $\mathcal{A}$. Sample (s_{t+1}, a_{t+1}) and (s'_{t+1}, a'_{t+1}) as two independent samples from $\mathcal{P}^{\theta_t}((s_t, a_t), \cdot)$ and $\mathcal{P}^{\theta_t}((s_t, \hat{a}_t), \cdot)$ respectively. Set

$$Q_{t+1}(s_t, \hat{a}_t) = Q_t(s_t, \hat{a}_t) + \alpha[c(s_t, \hat{a}_t) - Q_t(s_t, \hat{a}_t) + \gamma Q_t(s'_{t+1}, a'_{t+1})]$$

$$Q_{t+1}(s, a) = Q_t(s, a) \text{ for } (s, a) \neq (s_t, \hat{a}_t),$$

and $\theta_{t+1} = \theta_t - \eta_t Q_{t+1}(s_{t+1}, a_{t+1}) \nabla_\theta \log \pi^{\theta_t}(a_{t+1}|s_{t+1})$;

3 Set $t = t+1$. If $t < T$, go back to Step 2; otherwise, output $\bar{\theta}_T$.

Assumption 8 is a mild assumption since the state and action spaces are finite, and this assumption is standard in the literature (see, e.g., Assumption 1 in Mei et al. (2020)). Assumption 9 ensures sufficient exploration of the state space and is also a standard assumption for the convergence analysis of vanilla policy gradient (see, e.g., Assumption 2 in Mei et al. (2020)). Assumption 10 is a regularity condition on the score function that prevents the policy gradient estimator from having an infinite variance and is a standard assumption in the analysis of policy gradient under estimated gradients (see, e.g., Assumption 3.1 in Zhang et al. (2020b)). This is naturally satisfied under the softmax parameterization $\pi(a|s) \propto \exp(\theta_{s,a})$ with Θ being a bounded subset of $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$.

THEOREM 5. *Suppose Assumptions 8-10 hold. Consider step size $\eta_t = \eta_0/\sqrt{t}$ and let θ_t denote the policy parameters under Algorithm 1. Then*

$$\min_{0 \leq t < T} \mathbb{E}\|\nabla \ell(\theta_t)\|^2 = O\left(\tau (\log T)^2 / \sqrt{T}\right).$$

where τ is the mixing time of $Z_t = (s_t, a_t, Q_t)$.

Theorem 5 indicates that Algorithm 1 is able to attain an ϵ-stationary point with $O(\epsilon^{-2}\log(1/\epsilon))$ samples. For a similar $\tilde{O}(1/\sqrt{T})$ convergence, Wang et al. (2019) requires running $O(T^8)$ steps in the TD update to estimate the Q-function in each iteration and the ability to sample from (s, a) from $v^\theta(s, a)$ directly. In particular, for a policy π^θ, they approximate the gradient using

$$\frac{1}{B}\sum_{k=1}^{B} \hat{Q}^\theta(\tilde{s}_k, \tilde{a}_k),$$

where $\hat{Q}^\theta$ is estimated by running TD under π^θ for $O(T^8)$ steps and $(\tilde{s}_k, \tilde{a}_k)$, $k = 1, \ldots, B$ are B iid samples drawn from v^θ. When using the REINFORCE gradient estimator, the method in Yuan et al. (2022) requires running the Markov chain under policy π^θ for $O(\log(1/T))$ steps in each iteration, and the overall sample complexity is $\tilde{O}(\epsilon^{-4})$. Meanwhile, we would also like to

acknowledge that Wang et al. (2019), Yuan et al. (2022) study more complicated settings than the tabular setting we study here. For example, Wang et al. (2019) considers using neural networks to approximate the policies and the state-action value functions. Yuan et al. (2022) considers different and more general regularity conditions than what we assume in this section. Our sample complexity results are comparable to some of the best-known sample complexity results for policy gradient algorithms. In particular, Xiong et al. (2021) establishes an $\tilde{O}(\epsilon^{-2})$ complexity for an Adam-type policy gradient algorithm, which requires sampling a long enough trajectory under a policy π to estimate the corresponding Q-function. Xu et al. (2020) establishes an $\tilde{O}(\epsilon^{-2})$ complexity for a mini-batch actor-critic policy gradient algorithm. Unlike these works, we do not assume that the Markov chain on the states induced by the policy is uniformly ergodic across policies, which is a strong assumption that is difficult to verify. Rather, we use the regenerations induced by the discount factor γ to control the mixing rate, which allows our analysis to be applied to a wider range of problems.

The bound in Thereom 5 also depends polynomially on $(1-\gamma)^{-1}$. Although it is not the focus of this work, we make explicit the dependence on the discount factor in the following corollary.

COROLLARY 1. *Suppose Assumptions 8-10 hold. Consider step size $\eta_t = (1-\gamma)^5/\sqrt{t}$ and let θ_t denote the policy parameters under Algorithm 1. Then,*

$$\min_{0\le t<T} \mathbb{E}\|\nabla \ell(\theta_t)\|^2 = O\left((1-\gamma)^{-10}(\log T)^2/\sqrt{T}\right).$$

As discussed above, our work differentiates itself from many existing results (e.g., Xiong et al. (2021), Xu et al. (2020)) by not requiring the Markov chain s_t to be uniformly ergodic across policies. Our assumptions align with those in Zhang et al. (2020b), which establishes a convergence rate of $O\left((1-\gamma)^{-7}/\sqrt{T}\right)$ for a policy gradient algorithm. However, the algorithm in Zhang et al. (2020b) requires sampling a path of length $O((1-\gamma)^{-1})$ at each iteration to obtain an unbiased Q-function estimate under the current policy. Although our bound exhibits a worse dependence on $(1-\gamma)^{-1}$, our method uses only $O(1)$ samples per iteration while still achieving a convergence rate of $\tilde{O}(1/\sqrt{T})$, even with biased Q-function estimates. Moreover, our work aims to provide a unified framework for SGD with adaptive data. To streamline the analysis, we treat most constants in assumptions as $O(1)$. This simplification enhances conciseness, but may make the resulting bounds less sharp for specific problems where these constants are not $O(1)$.

6. Numerical Experiments

Motivated by our theoretical results, we proceed to empirically study the performance of the SGD algorithm with adaptive data. Specifically, for the policy optimization examples considered

in Sections 3 - 5, we examine how the adaptivity of the algorithm affects performance by varying the batch size: the number of data points collected before updating the policy parameters.

We observe broadly that even in the fully adaptive setting where policy parameters are updated after every data point, i.e., the batch size is 1, SGD can achieve an equivalent rate of convergence to larger-batch variants, which is consistent with our theoretical results. This holds even for highly non-stationary environments where the policy parameters change the environment quite a bit and convergence to stationarity is slow. Nevertheless, a carefully chosen larger batch size can provide small improvements in convergence speed in some cases (likely by improving the constant term), especially if the step sizes are tuned appropriately. Above all, our results show that under the ergodicity and smoothness conditions characterized in our theoretical analysis, in non-stationary, adaptive environments, SGD is robust to the level of adaptivity and the convergence speed is largely similar to the iid setting.

For the following numerical results, we compare the performance of batch sizes $B \in \{1, 10, 100\}$. We index the number of data points by $t \in \mathbb{N}_+$. For each example, we test out a step-size schedule $\eta_t \propto t^{-1/2}$ with iterate averaging $\bar{\theta}_t$ and average the performance across 100 independent runs of the algorithm. For the queuing and inventory examples, we also look at the loss of θ_t under the step-size schedule $\eta_t \propto t^{-1}$. We plot the performance as a function of the number of data samples rather than SGD iterations. In this setting, running the algorithm with a larger batch size will have identical θ_t across multiple data points until an update of the policy is made. To facilitate comparison, in the figures below, we also plot dashed lines corresponding to the theoretical convergence rates Ct^{-1} or $Ct^{-1/2}$, where C is a constant selected to match the empirical trends observed in the SGD algorithms.

6.1. Inventory Control with Stock-Out Damping

We empirically test the performance of the adaptive SGD algorithm for the inventory control problem discussed in Section 3. Recall that in this problem, demand is endogenously affected by the base-stock level θ. We consider a setting where the newsvendor loss is parameterized by an overage cost of $h = 1$, an underage cost of $b = 10$, and the demand process has a demand drift term $m = 5$ and a noise level $\sigma = 1.0$. We consider two values of the $AR(1)$ parameter, $\alpha = 0.8$ and $\alpha = 0.9$, to compare performance across relatively low and high levels of non-stationarity/mixing rate. Note that α also calibrates the degree to which θ affects the dynamics, with $\alpha = 0$ involving zero damping of the demand from stockouts.

We compare the SGD algorithm with step sizes scaled by the batch size B across different batch sizes for the two settings of α in Figure 1. We compare two versions of the SGD algorithm, one with step-size schedule $\eta_t \propto t^{-1/2}$ and with iterate averaging; the other with $\eta_t \propto t^{-1}$ and without

iterate averaging. We find that for both $\alpha = 0.8$ and $\alpha = 0.9$, the fully adaptive SGD algorithm ($B = 1$) achieves an equivalent rate of convergence as larger batch variants, with the large batch sizes performing better (as the step sizes are scaled with the batch size). Interestingly, both the $t^{-1/2}$ and t^{-1} step-sizes empirically achieve a convergence rate of $1/t$ (indicated by the dashed line). In addition, while we observe that convergence of SGD is slower for the larger value of $\alpha = 0.9$, because the underlying Markov chain mixes more slowly in this case and the base-stock policy has a greater impact on the dynamics, the overall rate of convergence resembles the rate observed for $\alpha = 0.8$. This provides numerical evidence that when the conditions of Theorem 1 are satisfied, the performance of SGD in the adaptive environment resembles that in the iid setting.

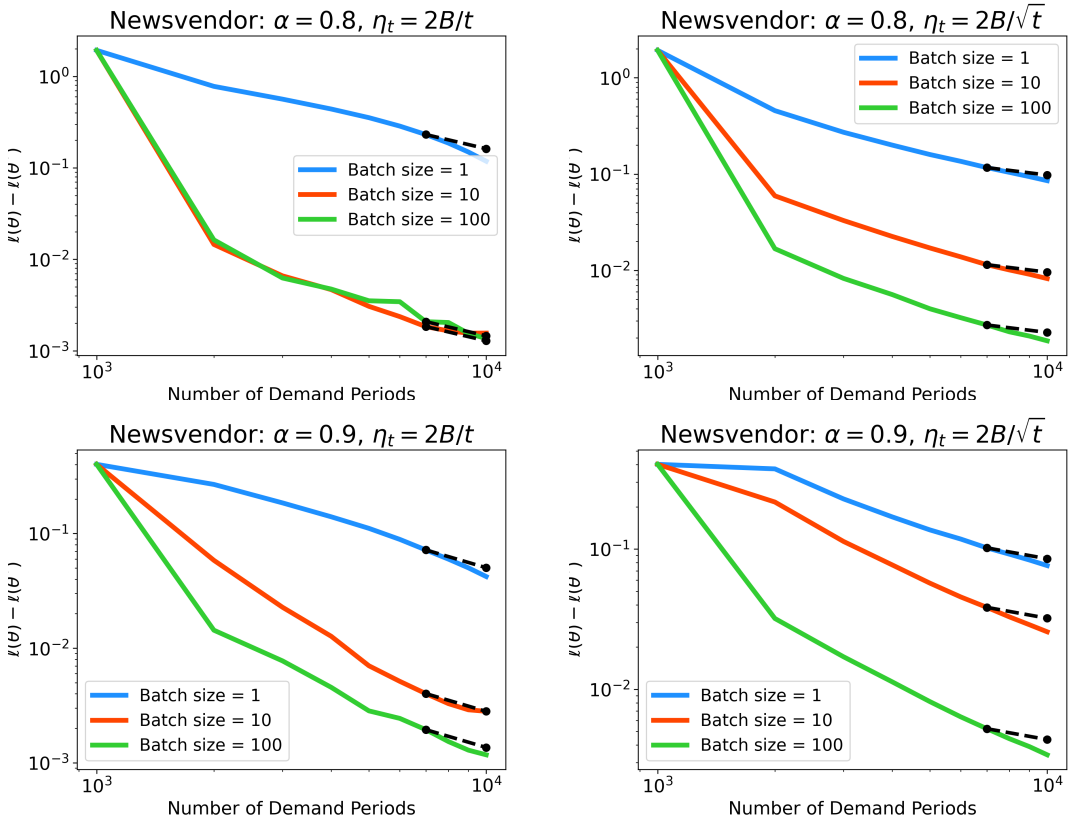

Figure 1 Inventory control with stock-out damping (overage cost $b = 10$, underage cost $h = 1$, and noise level $\sigma = 1.0$). The dashed line indicates the reference t^{-1} convergence rate in all plots. (Top) Newsvendor loss gap for the SGD iterates with $AR(1)$ parameter $\alpha = 0.8$. (Top Left) Step-size schedule $\eta_t = 2Bt^{-1}$ across batch sizes $B \in \{1, 10, 100\}$. (Top Right) Step size schedule $\eta_t = 2Bt^{-1/2}$ with iterate averaging. (Bottom) Newsvendor loss gap for the SGD iterates with $AR(1)$ parameter $\alpha = 0.9$. (Bottom Left) Step-size schedule $\eta_t = 2Bt^{-1}$. (Bottom Right) Step-size schedule $\eta_t = 2Bt^{-1/2}$ with iterate averaging.

6.2. Pricing and Capacity Sizing in Single-Server Queue

We consider the pricing and capacity sizing problem studied in Section 4. Following the numerical example described in Chen et al. (2023a), we consider an $M/M/1$ queue where arrivals to the queue follow a Poisson process with rate $\lambda(p) = n\lambda_0(p)$ with $n > 0$ and

$$\lambda_0(p) = \frac{\exp(a-p)}{1+\exp(a-p)}$$

for some $a > 0$. The service time is exponentially distributed and the service rate μ entails a quadratic cost $c(\mu) = c_0\mu^2$ with $c_0 > 0$. There is a holding cost $h_0 > 0$. In this simple example, the pricing and capacity sizing problem (10) can be written in closed form as

$$\max_{p,\mu\in\Theta} \left\{ np\lambda_0(p) - c_0\mu^2 - h_0 \frac{\lambda(p)/\mu}{1-\lambda(p)/\mu} \right\}. \tag{17}$$

We set $n = 10$, $a = 4.1$, $c_0 = 0.1$, $h_0 = 1$, and the step-size parameter $\eta_0 = 1$.

The joint pricing and capacity sizing problem (17) is known to be strongly convex. As in the inventory example, we evaluate two versions of the SGD algorithm, one with step-size schedule $\eta_t \propto t^{-1/2}$ and with iterate averaging; the other with step-size schedule $\eta_t \propto t^{-1}$ and without iterate averaging. Figure 2 displays the performance for both versions of the algorithm. As predicted by our theoretical results, across all batch sizes, step-size schedule $\eta_t = 1/\sqrt{t}$ with iterate averaging achieves an $O(t^{-1/2})$ convergence. Whereas step-size schedule $\eta_t = 1/t$ without iterate averaging achieves an $O(t^{-1})$ convergence.

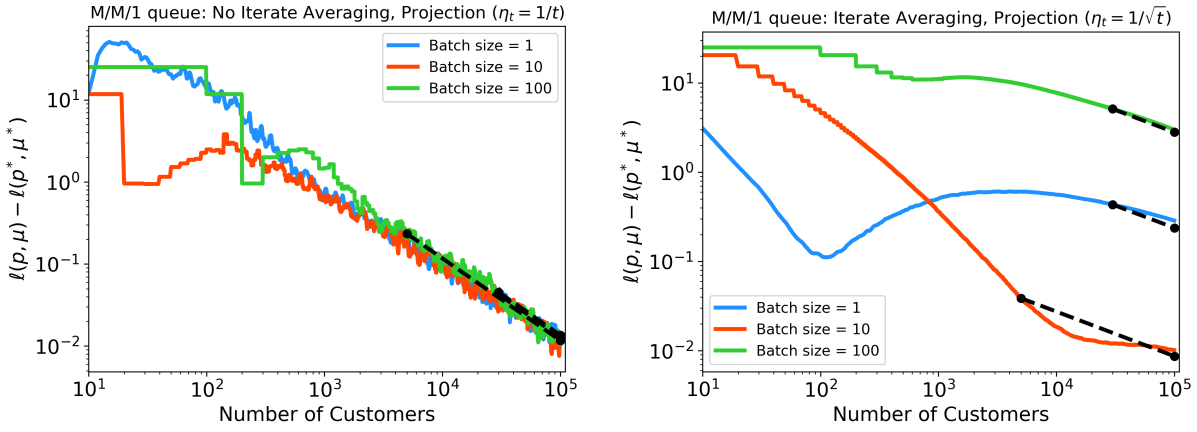

Figure 2 Pricing and capacity sizing in the single server queue. (Left) Last iterate loss gap for the SGD iterates with $\eta_t = 1/t$ across batch sizes $B \in \{1, 10, 100\}$. The dashed line indicates the reference t^{-1} convergence rate. (Right) Loss gap for the average iterate with $\eta_t = 1/\sqrt{t}$. The dashed line indicates the reference $t^{-1/2}$ convergence rate.

We also see in our experiments that instabilities can appear under adaptive feedback if the assumptions for Theorem 2 are not satisfied. For example, Figure 3 compares the empirical convergence of SGD with step-size schedule $\eta_t = 1/\sqrt{t}$, without versus with averaging, and without projection. Note that the two scenarios are outside the scope of our theoretical results because we do not project the parameters to the set in which the queue will be uniformly stable. We observe in the left panel of Figure 3 that without projection and iterate averaging, the adaptive algorithm can be highly unstable. For example, when $B = 1$, the cost/loss oscillates between 10 and 10^4. In the right panel, with iterate averaging, the algorithm converges at rate $t^{-1/2}$. The instabilities in the left panel are a result of the bias incurred by the adaptive feedback. Intuitively, instabilities emerge because the algorithm lowers the service capacity without fully anticipating the increase in congestion that this will incur, since it takes time for customers to arrive. Waiting for more customers to arrive before updating parameters allows the algorithm to gauge better congestion, which leads to better performance for larger batch variants. Averaging or smaller step sizes, i.e., t^{-1}, can help smooth things in this case and lead to more stable learning. Overall, this demonstrates that features of the algorithm, such as projection, iterate averaging, and small and properly decreasing step sizes, which are usually benign for the empirical performance of SGD with iid data, can be crucial for stable convergence in the adaptive setting.

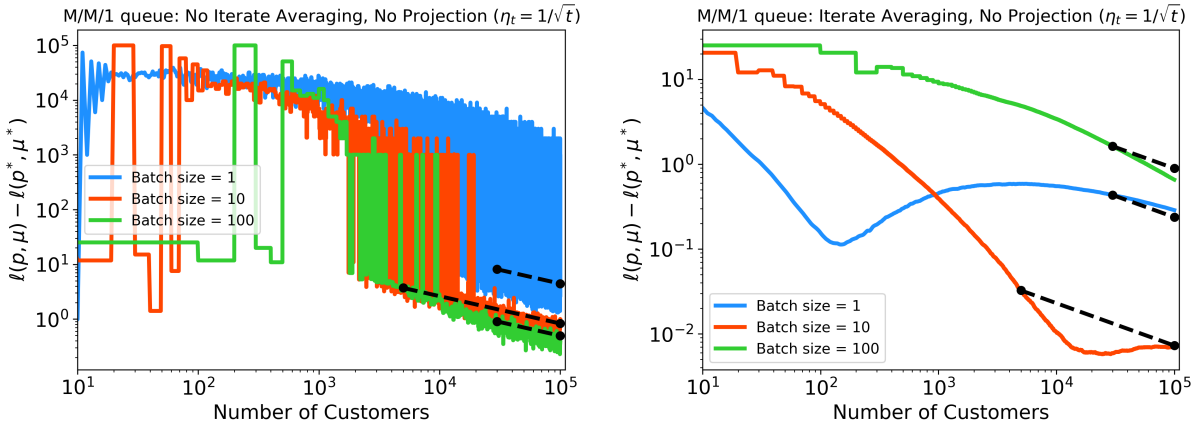

Figure 3 Pricing and capacity sizing in the single server queue. (Left) Loss gap for the SGD iterates without projection and without iterate averaging with $\eta_t = 1/\sqrt{t}$ across batch sizes $B \in \{1, 10, 100\}$. (Right) Loss gap for the SGD iterates without projection but with iterate averaging and $\eta_t = 1/\sqrt{t}$. The dashed line indicates the reference $t^{-1/2}$ convergence rate.

6.3. Policy Gradient in Reinforcement Learning

We evaluate the performance of the actor-critic algorithm (Algorithm 1) in simple tabular RL examples. We consider a direct softmax parameterization of the policy over all states and actions, i.e. $\theta \in \mathbb{R}^{|S|\times|A|}$ and

$$\pi^\theta(a|s) = \frac{\exp(\theta_{s,a})}{\sum_{a\in A}\exp(\theta_{s,a})}.$$

Since the loss function $\ell(\theta) = \mathbb{E}_\rho^\theta[\sum_{t=0}^\infty \gamma^t c(s_t, a_t)]$ is not strongly convex in θ, we consider step size schedule with $\eta_t \propto t^{-1/2}$ and iterate averaging. We set the discount factor $\gamma = 0.8$, the TD update parameter $\alpha = 0.5$, and the step-size schedule $\eta_t = 2B/t^{-1/2}$, which scales linearly with the batch size. We consider randomly generated MDPs with $|\mathcal{S}| = 5$ states and $|\mathcal{A}| = 5$ actions, and larger instances with $|\mathcal{S}| = 10$ states and $|\mathcal{A}| = 10$ actions. For each instance type, we average performance across 100 randomly generated MDPs by rescaling the cost in each MDP as

$$\frac{\ell(\bar{\theta}_t) - \ell(\theta^*)}{\ell(\theta^*)}. \tag{18}$$

Figure 4 displays the averaged scaled optimality gap (18) for different batch sizes as a function of the total number of samples drawn from the MDP. Note that because each TD update uses 2 samples from the MDP, we effectively use $2B$ samples per iteration, which is reflected in the x-axis. We observe that the fully adaptive actor-critic algorithm, which updates the policy after only a single TD update to the Q-function, is able to achieve an almost identical convergence rate with variants that perform multiple TD updates before updating the policy. This convergence rate is $O(t^{-1/2})$ as indicated by the dashed line.

We also compare the performance of our online actor-critic algorithm (Algorithm 1) with the random-horizon policy gradient method proposed in (Zhang et al. 2020b, Algorithm 3). This comparison evaluates whether it is more effective to update the policy frequently using biased gradients (our approach) or to accumulate enough samples to compute an unbiased policy gradient, as in Zhang et al. (2020b). For the method of Zhang et al. (2020b), each iteration k begins by sampling a trajectory under the current policy π^{θ_k} with a random horizon $T_k \sim$Geometric$(1-\gamma)$. Let (s_{T_k}, a_{T_k}) denote the state-action pair at the end of this trajectory. An unbiased estimate of the Q-function at (s_{T_k}, a_{T_k}) is then computed via another trajectory, starting from (s_{T_k}, a_{T_k}), with a random horizon $\hat{T}_k \sim$Geometric$(1-\gamma^{1/2})$, using the discounted sum

$$\hat{Q}_k(s_{T_k}, a_{T_k}) \equiv \frac{1}{1-\gamma}\sum_{l=0}^{\hat{T}_k}\gamma^{l/2}c(s_l, a_l).$$

The policy is then updated using this estimated Q-function:

$$\theta_{k+1} = \theta_k - \eta_k \hat{Q}_k(s_{T_k}, a_{T_k})\nabla_{\theta_k}\log\pi^{\theta_k}(a_{T_k}|s_{T_k}).$$

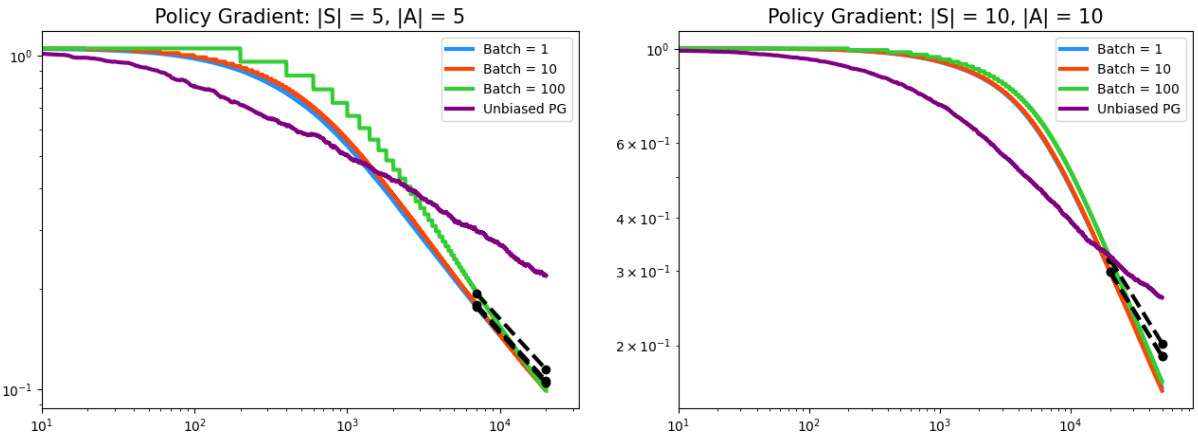

Figure 4 Policy gradient for tabular MDPs. The dashed line indicates the reference $t^{-1/2}$ convergence rate. (Left) Scaled loss gap of the averaged SGD iterates for batch sizes $B \in 1, 10, 100$ and the random-horizon policy gradient algorithm from Zhang et al. (2020b) (unbiased PG) across 100 randomly generated MDPs with $|\mathcal{S}| = |\mathcal{A}| = 5$. The step-size schedule for batched SGD is $\eta_t = 2B/\sqrt{t}$. For the unbiased PG, we report the best performance using the step-size schedule $\eta_t = \eta_0/\sqrt{t}$ with $\eta_0 \in 0.1, 1, 10, 100$. (Right) Scaled loss gap for the averaged SGD iterates with batch sizes $B \in \{1, 10, 100\}$ and the unbiased PG for 100 randomly generated MDPs with $|\mathcal{S}| = |\mathcal{A}| = 10$.

Note that our algorithm differs in three key ways: First, we update the policy using gradients from all visited state-action pairs, rather than only the final pair of a random-horizon trajectory. Second, we use a biased, online TD estimate of the Q-function instead of an unbiased Monte Carlo estimate. Finally, our method uses only $O(1)$ samples per gradient step, since each TD update draws two samples from the transition kernel, while the method in Zhang et al. (2020b) requires an expected $O((1-\gamma)^{-1})$ samples per update to maintain unbiasedness.

Figure 4 also shows the performance of the random-horizon policy gradient algorithm from Zhang et al. (2020b), labeled as "Unbiased PG," plotted as a function of the total number of samples drawn from the MDP. Note that the total sample count varies across iterations because the gradient estimator in this algorithm uses trajectories of random lengths. Similar to our work, Zhang et al. (2020b) establish a convergence rate of $O(1/\sqrt{T})$ under a step-size schedule of the form $\eta_t = O(1/\sqrt{t})$. Accordingly, we implement their algorithm using $\eta_t = \eta_0/\sqrt{t}$ and report the best performance across $\eta_0 \in 0.1, 1, 10, 100$. We observe that while the unbiased gradient estimator achieves faster initial progress, the performance improvement slows down after approximately 10^3 or 10^4 samples and is then outperformed by our actor-critic algorithm, even with batch size $B = 1$. This highlights that unbiased gradient estimates are not strictly necessary for convergence, and that frequent updates with biased estimates can be more effective than less frequent, unbiased updates.

7. Conclusion

In this work, we study SGD with adaptive data, which arises in many online learning problems in operations research. We provide easy-to-verify conditions under which the fully adaptive SGD update achieves a similar convergence speed as the classical stationary setting. Our results provide guidance and assurance as to how to choose the appropriate batch size in online learning problems where stationary (long-run average) performance is involved. For example, we demonstrate how to apply the results to study online learning algorithms for some service and inventory management problems.

When applying stochastic gradient descent to optimize policy design in practice, two key considerations arise. First, how to parameterize the policy. In Sections 3 and 4, we focus on relatively restrictive classes of stationary policies, e.g., a base-stock policy in Section 3, and static service rate and price in Section 4. Richer classes of stationary policies can be considered, e.g., $\pi^\theta(a|s)$ where θ parameterizes a neural network. This added flexibility may come at the cost of more complex, nonconvex optimization challenges in practice. Second, we must consider how to construct a gradient estimator. In this work, we advocate the use of pathwise gradient estimators. Similar techniques have recently gained traction in large-scale reinforcement learning problems arising in operations research (Alvo et al. 2023, Che et al. 2024). Developing more efficient and accurate gradient estimators remains an important direction for future research.

We conclude the paper with some remarks about the limitations of this work that suggest promising directions for future research. First, the conditions we required in Theorems 1 and 2 are only sufficient conditions. It would be an interesting future research direction to see both theoretically and empirically whether similar convergence speeds hold under more general conditions. In particular, while we think the ergodicity and smoothness conditions are important for the convergence of the algorithm, it would be interesting to see if these conditions only need to hold locally, i.e., around the optimal solution. Second, the upper bounds we established in Theorems 1 and 2 are unlikely to be tight, especially regarding the logarithmic terms. We leave it as a future research direction to establish tight bounds. Lastly, from our numerical experiments, even though the batch size does not affect the convergence rate of the algorithm, it can improve the convergence speed through the constant term. It would be valuable to develop theoretical results that can guide the choice of the optimal batch size.

Acknowledgments

We thank the anonymous Associate Editor and reviewers for their constructive and insightful feedback. Jing Dong gratefully acknowledges support from the National Science Foundation through Grant 1944209. Xin Tong gratefully acknowledges support from the Singapore Ministry of Education grant A-8002956-00-00.

References

Alekh Agarwal and John C Duchi. The generalization ability of online algorithms for dependent data. *IEEE Transactions on Information Theory*, 59(1):573–587, 2012.

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. In *Conference on Learning Theory*, pages 64–66. PMLR, 2020.

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

Matias Alvo, Daniel Russo, and Yash Kanoria. Neural inventory control in networks via hindsight differentiable policy optimization. *arXiv preprint arXiv:2306.11246*, 2023.

Eric T Anderson, Gavan J Fitzsimons, and Duncan Simester. Measuring and mitigating the costs of stockouts. *Management science*, 52(11):1751–1763, 2006.

Peter H Baxendale. Renewal theory and computable convergence rates for geometrically ergodic markov chains. 2005.

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

Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *Operations Research*, 2024.

Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Conference on learning theory*, pages 1691–1692. PMLR, 2018.

Ethan Che, Jing Dong, and Hongseok Namkoong. Differentiable discrete event simulation for queuing network control. *arXiv preprint arXiv:2409.03740*, 2024.

Xinyun Chen, Yunan Liu, and Guiyu Hong. An online learning approach to dynamic pricing and capacity sizing in service systems. *Operations Research*, 2023a.

Zaiwei Chen, Siva T Maguluri, Sanjay Shakkottai, and Karthikeyan Shanmugam. A lyapunov theory for finite-sample guarantees of markovian stochastic approximation. *Operations Research*, 2023b.

Wang Chi Cheung, Will Ma, David Simchi-Levi, and Xinshang Wang. Inventory balancing with online learning. *Management Science*, 68(3):1776–1807, 2022.

Tiangang Cui, Jing Dong, Ajay Jasra, and Xin T Tong. Convergence speed and approximation accuracy of numerical mcmc. *Advances in Applied Probability*, 57(1):101–133, 2025.

Gal Dalal, Balázs Szörényi, Gugan Thoppe, and Shie Mannor. Finite sample analyses for td (0) with function approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

A Doucet and V Tadic. Asymptotic bias of stochastic gradient search. *Annals of Applied Probability*, 27(6), 2017.

Dmitriy Drusvyatskiy and Lin Xiao. Stochastic optimization with decision-dependent distributions. *Mathematics of Operations Research*, 48(2):954–998, 2023.

John C Duchi, Alekh Agarwal, Mikael Johansson, and Michael I Jordan. Ergodic mirror descent. *SIAM Journal on Optimization*, 22(4):1549–1578, 2012.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Paul Glasserman. Stationary waiting time derivatives. *Queueing systems*, 12:369–389, 1992.

Peter W Glynn. Likelihood ratio gradient estimation for stochastic systems. *Communications of the ACM*, 33(10):75–84, 1990.

Peter W Glynn, Ramesh Johari, and Mohammad Rasouli. Adaptive experimental design with temporal interference: A maximum likelihood approach. *Advances in Neural Information Processing Systems*, 33:15054–15064, 2020.

Philip Heidelberger, Xi-Ren Cao, Michael A Zazanis, and Rajan Suri. Convergence properties of infinitesimal perturbation analysis estimates. *Management Science*, 34(11):1281–1302, 1988.

Yuchen Hu and Stefan Wager. Switchback experiments under geometric mixing. *arXiv preprint arXiv:2209.00197*, 2022.

Woonghee Tim Huh and Paat Rusmevichientong. Online sequential optimization with biased gradients: theory and applications to censored demand. *INFORMS Journal on Computing*, 26(1):150–159, 2014.

Woonghee Tim Huh, Ganesh Janakiraman, John A Muckstadt, and Paat Rusmevichientong. An adaptive algorithm for finding the optimal base-stock policy in lost sales inventory systems with censored demand. *Mathematics of Operations Research*, 34(2):397–416, 2009.

Dongyan Huo, Yudong Chen, and Qiaomin Xie. Bias and extrapolation in markovian linear stochastic approximation with constant stepsizes. In *Abstract Proceedings of the 2023 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pages 81–82, 2023.

Belhal Karimi, Blazej Miasojedow, Eric Moulines, and Hoi-To Wai. Non-asymptotic analysis of biased stochastic approximation scheme. In *Conference on Learning Theory*, pages 1944–1974. PMLR, 2019.

Jeunghyun Kim and Ramandeep S Randhawa. The value of dynamic pricing in large queueing systems. *Operations Research*, 66(2):409–425, 2018.

Subhashini Krishnasamy, Rajat Sen, Ramesh Johari, and Sanjay Shakkottai. Learning unknown service rates in queues: A multiarmed bandit approach. *Operations research*, 69(1):315–330, 2021.

David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Qiang Li and Hoi-To Wai. State dependent performative prediction with stochastic approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 3164–3186. PMLR, 2022.

Xiang Li, Jiadong Liang, Xinyun Chen, and Zhihua Zhang. Convergence and inference of stream sgd, with applications to queueing systems and inventory control. *arXiv preprint arXiv:2309.09545*, 2025.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International conference on machine learning*, pages 6820–6829. PMLR, 2020.

Celestine Mendler-Dünner, Juan Perdomo, Tijana Zrnic, and Moritz Hardt. Stochastic optimization for performative prediction. *Advances in Neural Information Processing Systems*, 33:4929–4939, 2020.

Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

Shakir Mohamed, Mihaela Rosca, Michael Figurnov, and Andriy Mnih. Monte carlo gradient estimation in machine learning. *Journal of Machine Learning Research*, 21(132):1–62, 2020.

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

Guannan Qu and Adam Wierman. Finite-time analysis of asynchronous stochastic approximation and q-learning. In *Conference on Learning Theory*, pages 3185–3205. PMLR, 2020.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Gareth Roberts and Jeffrey Rosenthal. Geometric ergodicity and hybrid markov chains. 1997.

Abhishek Roy, Krishnakumar Balasubramanian, and Saeed Ghadimi. Constrained stochastic nonconvex optimization with state-dependent markov data. *Advances in Neural Information Processing Systems*, 35:23256–23270, 2022.

Daniel Rudolf and Nikolaus Schweizer. Perturbation theory for markov chains via wasserstein distance. 2018.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. *Lectures on stochastic programming: modeling and theory*. SIAM, 2021.

Rayadurgam Srikant and Lei Ying. Finite-time error bounds for linear stochastic approximation andtd learning. In *Conference on Learning Theory*, pages 2803–2830. PMLR, 2019.

Tao Sun, Yuejiao Sun, and Wotao Yin. On markov chain gradient descent. *arXiv preprint arXiv:1809.04216*, 2018.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Jingwen Tang, Boxiao Chen, and Cong Shi. Online learning for dual-index policies in dual-sourcing systems. *Manufacturing & Service Operations Management*, 2023.

Neil Walton and Kuang Xu. Learning and information in stochastic networks and queues. In *Tutorials in Operations Research: Emerging Optimization Methods and Modeling Techniques with Applications*, pages 161–198. INFORMS, 2021.

Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*, 2019.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256, 1992.

Lin Xiao. On the convergence rates of policy gradient methods. *Journal of Machine Learning Research*, 23 (282):1–36, 2022.

Huaqing Xiong, Tengyu Xu, Yingbin Liang, and Wei Zhang. Non-asymptotic convergence of adam-type reinforcement learning algorithms under markovian sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10460–10468, 2021.

Tengyu Xu, Zhe Wang, and Yingbin Liang. Improving sample complexity bounds for (natural) actor-critic algorithms. *Advances in Neural Information Processing Systems*, 33:4358–4369, 2020.

Rui Yuan, Robert M Gower, and Alessandro Lazaric. A general sample complexity analysis of vanilla policy gradient. In *International Conference on Artificial Intelligence and Statistics*, pages 3332–3380. PMLR, 2022.

Huanan Zhang, Xiuli Chao, and Cong Shi. Closing the gap: A learning algorithm for lost-sales inventory systems with lead times. *Management Science*, 66(5):1962–1980, 2020a.

Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Basar. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 58(6):3586–3612, 2020b.

Yueyang Zhong, John R Birge, and Amy Ward. Learning the scheduling policy in time-varying multiclass many server queues with abandonment. *Available at SSRN*, 2022.

A. Lower bound on the Convergence Rate of SGD

In this section, we construct an example with a strongly convex loss function ℓ and show that the convergence rate of the SGD update is $\Omega(\tau/T)$ under the assumption that the step size η_t is non-increasing in t for t large enough.

LEMMA 2. *Suppose $\ell(\theta) = (y - \theta x)^2$ where $y = \theta^* x$. Assume x is uniform on $\{-1, 1\}$. Consider the SGD update $\theta_{t+1} = \theta_t - \eta_t(\nabla \ell(\theta_t) + \epsilon_t)$. We further assume that $\{\epsilon_t\}_{t\geq 0}$ follows a Markov chain with transition probability*

$$P_{i,i} = 1 - \frac{1}{\tau}, P_{i,j} = \frac{1}{\tau} \text{ for } i, j \in \{-1, 1\} \text{ and } i \neq j.$$

Suppose η_t is non-increasing in t for t large enough. Then for any τ, there exists $N_\tau < \infty$, such that for any $T > N_\tau$,

$$\mathbb{E}[(\theta_T - \theta^*)^2] \geq \frac{\tau}{23T}.$$

Proof. Plug in the specific form of $\ell(\theta)$ and let $\Delta_t = \theta_t - \theta^*$, we have

$$\Delta_{t+1} = (1 - 2\eta_t)\Delta_t - \eta_t \epsilon_t.$$

This further implies that

$$\Delta_{t+1} = \Delta_1 \prod_{s=1}^{t}(1 - 2\eta_s) - \sum_{k=1}^{t} \eta_k \epsilon_k \prod_{s=k+1}^{t} (1 - 2\eta_s),$$

where we define $\prod_{s=t+1}^{t}(1 - 2\eta_s) = 1$. Let $w_k(t) = \eta_k \prod_{s=k+1}^{t}(1 - 2\eta_s)$. Since $\mathbb{E}_\pi[\epsilon_t] = 0$ and $\mathbb{E}_\pi[\epsilon_i \epsilon_j] = (1 - 2/\tau)^{|i-j|}$,

$$\begin{aligned}
\mathbb{E}[\Delta_{t+1}^2] &= \Delta_1^2 \prod_{s=1}^{t}(1 - 2\eta_s)^2 + \sum_{k=1}^{t} w_k(t)^2 + 2 \sum_{1 \leq i < j \leq t} w_i(t) w_j(t) \left(1 - \frac{2}{\tau}\right)^{j-i} \\
&\geq \Delta_1^2 \prod_{s=1}^{t}(1 - 2\eta_s)^2 + \sum_{k=1}^{t} w_k(t)^2 + 2 \sum_{i=1}^{t-\tau/4} \sum_{d=1}^{\tau/4} w_i(t) w_{i+d}(t), \qquad (19)
\end{aligned}$$

because $(1 - 2/\tau)^k \geq 1 - 2k/\tau \geq 1/2$ for $k \leq \tau/4$. We next look at the three terms in (19) one by one.

For the first term, to have convergence to 0, we need $\sum_{s=1}^{t} \eta_s \to \infty$ as $t \to \infty$.

For the second term, since $w_t(t) = \eta_t$, to have convergence to 0, we need $\eta_t \to 0$ as $t \to \infty$. In addition, note that

$$2w_k = \prod_{s=k+1}^{t} (1 - 2\eta_s) - \prod_{s=k}^{t}(1 - 2\eta_s).$$

By Cauchy–Schwarz inequality, we have

$$\sum_{k=1}^{t} w_k(t)^2 \geq \frac{1}{t} \left(\sum_{k=1}^{t} w_k(t) \right)^2 = \frac{1}{t} \left(1 - \prod_{s=1}^{t}(1 - 2\eta_s) \right).$$

Since $\prod_{s=1}^{t}(1-2\eta_s) \to 0$ as $t \to \infty$, we have there exists $n_0 > 0$, such that for $t > n_0$, $\prod_{s=1}^{t}(1-2\eta_s) < 1/2$. Then for $t > n_0$,

$$\sum_{k=1}^{t} w_k(t)^2 \geq \frac{1}{2t}.$$

For the third term, since $\eta_t \to 0$ as $t \to \infty$, we have for a fixed τ, there exists $n'_\tau < \infty$, such that for any $t > n'_\tau$ and $d \leq \tau/4$, $\prod_{s=t+1}^{t+d}(1-2\eta_s) > 1/2$. Then, for $t > n'_\tau$,

$$\begin{aligned} w_i(t)w_{i+d}(t) &= \eta_i\eta_{i+d} \prod_{s=i+1}^{t}(1-2\eta_s) \prod_{s=i+d+1}^{t}(1-2\eta_s) \\ &\geq \left(\prod_{s=i+1}^{i+d}(1-2\eta_s)\right)\eta_{i+d}^2 \prod_{s=i+d+1}^{t}(1-2\eta_s)^2 \geq \frac{1}{2}w_{i+d}(t)^2. \end{aligned}$$

This further implies that there exists $N_\tau < \infty$ such that for all $t > N_\tau$,

$$\begin{aligned} 2\sum_{i=1}^{t-\tau/4}\sum_{d=1}^{\tau/4} w_i(t)w_{i+d}(t) &\geq \sum_{i=n'_\tau}^{t-\tau/4}\sum_{d=1}^{\tau/4} w_{i+d}(t)^2 \\ &\geq \frac{\tau}{4}\sum_{j=n'_\tau+\tau/4}^{t-\tau/4} w_j(t)^2 \\ &\geq \frac{\tau}{4}\frac{2}{t}\left(\sum_{j=n'_\tau+\tau/4}^{t-\tau/4} w_j(t)\right)^2 \\ &= \frac{\tau}{2t}\left(\prod_{s=t-\tau/4+1}^{t}(1-2\eta_s) - \prod_{s=n'_\tau+\tau/4}^{t}(1-2\eta_s)\right)^2 \\ &= \frac{\tau}{2t}\left(\prod_{s=t-\tau/4+1}^{t}(1-2\eta_s)\right)^2\left(1 - \prod_{s=n'_\tau+\tau/4}^{t-\tau/4}(1-2\eta_s)\right)^2 \\ &\geq \frac{\tau}{2t}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^2 = \frac{\tau}{32t}. \end{aligned}$$

□

B. Proofs of Theorems 1 and 2

Define $\eta_{t:t+n} := \sum_{s=t}^{t+n-1}\eta_s$, $\eta^2_{t:t+n} := \sum_{s=t}^{t+n-1}\eta_s^2$, $(V\eta)_{t:t+n} = \sum_{s=t}^{t+n-1}V(z_s)\eta_s$, $(V\eta)^2_{t:t+n} = \sum_{s=t}^{t+n-1}V(z_s)^2\eta_s^2$ and $P_{\theta_{m:n}} = P_{\theta_m}P_{\theta_1}\dots P_{\theta_{n-1}}$. We also define the ϵ-mixing time τ_ϵ as a time for which under Assumption 1,

$$W_d(\mu_\theta, \delta_z P_\theta^{\tau_\epsilon}) \leq \epsilon V(z). \tag{20}$$

We first prove some auxiliary lemmas. These lemmas hold for both the unprojected and projected updates, i.e., (4) and (5) respectively. For concision, we only list the assumptions for the unprojected case. For the projected case, these assumptions are only required to hold for $\theta \in \Theta$.

LEMMA 3. *If P_θ has a mixing time τ, then for $\tau_\epsilon = \lceil |\log \epsilon|/|\log(1/4)| \rceil \tau$, we have*

$$W_d(\delta_z P_\theta^{\tau_\epsilon}, \mu_\theta(\cdot)) \le \epsilon V(z),$$

i.e., τ_ϵ is an ϵ-mixing time.

Proof. For any function f with $|f(z)| \le V(z)$ for all z, we define $f_0(z) := f(z) - \int_{z'} \mu_\theta(dz') f(z')$. Note that $\int_z f_0(z)\mu_\theta(dz) = 0$. We also define $f_1(z) := \mathbb{E}_z f_0(z_\tau)$. Note that

$$\int_z f_1(z)\mu_\theta(dz) = \int_z \mu_\theta(dz) \int_{z'} \delta_z P_\theta^\tau(dz') f_0(z') = \int_{z'} \mu_\theta(dz') f_0(z') = 0,$$

and

$$|f_1(z)| = \left| \int_{z'} f(z') \left(\delta_z P_\theta^\tau(dz') - \mu_\theta(dz')\right) \right| \le \frac{1}{4} V(z).$$

We can repeat the above procedure and define a sequence of f_k's. For $f_k(z) := \mathbb{E}_z[f_0(z_{k\tau})]$ for $k \ge 2$, suppose $|f_{k-1}(z)| \le \frac{1}{4^{k-1}} V(z)$. Then, we have

$$\int_z \mu(dz) f_k(z) = \int_z \mu_\theta(dz) \int_{z'} \delta_z P_\theta^{k\tau}(dz') f_0(z') = \int_{z'} \mu_\theta(dz') f_0(z') = 0,$$

and

$$|f_k(s)| = \frac{1}{4^{k-1}} \left| \int_{z'} 4^{k-1} f_{k-1}(z') \left(\delta_z P_\theta^\tau(dz') - \mu_\theta(dz')\right) \right| \le \frac{1}{4^k} V(z).$$

Let $n = \lceil |\log \epsilon|/|\log(1/4)| \rceil$. Then,

$$f_n(z) = \mathbb{E}_z f(z_{\tau_\epsilon}) - \int_{z'} \mu(dz') f(z') \text{ and } |f_n(z)| \le \frac{1}{4^n} V(z) \le \epsilon V(z).$$

Since f is any function with $|f(z)| \le V(z)$, we have $W_d(\delta_z P_\theta^{\tau_\epsilon}, \mu(\cdot)) \le \epsilon V(z)$. □

LEMMA 4. *Under Assumptions 3, the iterates satisfy*

$$\|\theta_{t+n} - \theta_t\| \le M(V\eta)_{t:t+n}.$$

Proof. We first note that for the unprojected cases, $\|\theta_{t+1} - \theta_t\| = \|\eta_t \hat{g}_t(\theta_t, z_t)\|$. For the projected case, since $\mathcal{C}$ is convex, $\|\theta_{t+1} - \theta_t\| \le \|\eta_t \hat{g}_t(\theta_t, z_t)\|$. Thus,

$$\|\theta_{t+1} - \theta_t\| \le \|\eta_t \hat{g}_t(\theta_t, z_t)\| \le MV(z_t)\eta_t.$$

Then,

$$\|\theta_{t+n} - \theta_t\| \le \sum_{k=t}^{t+n-1} \|\theta_{k+1} - \theta_k\| \le M(V\eta)_{t:t+n}.$$

□

Before proceeding, we refine Assumption 2 slightly by specifying that

$$W_d(\delta_x P_\theta, \delta_x P_{\theta'}) \leq L_d \|\theta - \theta'\| V(x),$$

instead of the original bound $W_d(\delta_x P_\theta, \delta_x P_{\theta'}) \leq L\|\theta - \theta'\|V(x)$. We introduce this adjustment because, in the reinforcement learning example, the constant L_d is significantly smaller than L.

LEMMA 5. *Under Assumptions 1 and 2, the following holds almost surely,*

$$W_d(\delta_x P_{\theta_{0:n}}, \delta_x P_{\theta_0}^n) \leq \frac{L_d K^2 M V(x)}{(1-\rho)^2} (V\eta)_{0:n}.$$

Proof. First, let γ^* denote the optimal coupling between μ and ν under d. Then,

$$\begin{aligned} W_d(\mu P_{\theta_0}^k, \nu P_{\theta_0}^k) &\leq \int_{x,y} d(\delta_x P_{\theta_0}^k, \delta_y P_{\theta_0}^k) \gamma^*(dx, dy) \\ &\leq K\rho^k \int_{x,y} d(x,y)\gamma^*(dx,dy) = K\rho^k W_d(\mu, \nu). \end{aligned}$$

Next,

$$W_d(\mu P_\theta, \mu P_{\theta'}) \leq \int_x W_d(\delta_x P_\theta, \delta_x P_{\theta'}) \mu(dx) \leq L_d \|\theta - \theta'\| \int_x V(x)\mu(dx).$$

We also note that

$$\begin{aligned} \int_{x'} V(x') \delta_x P_{\theta_{0:m}}(dx') &\leq \rho P_{\theta_{0:m-1}} V(x) + K \\ &\leq \rho^m V(x) + \rho^{m-1} K + \cdots + \rho K + K \\ &\leq \frac{KV(x)}{1-\rho}. \end{aligned}$$

Lastly,

$$\begin{aligned} W_d(\delta_x P_{\theta_{0:n}}, \delta_x P_{\theta_0}^n) &\leq \sum_{m=2}^n W_d(\delta_x P_{\theta_{0:m}} P_{\theta_0}^{n-m}, \delta_x P_{\theta_{0:m-1}} P_{\theta_0}^{n-m+1}) \\ &\leq \sum_{m=2}^n K\rho^{n-m} W_d(\delta_x P_{\theta_{0:m-1}} P_{\theta_{m-1}}, \delta_x P_{\theta_{0:m-1}} P_{\theta_0}) \\ &\leq \sum_{m=2}^n K\rho^{n-m} L_d \|\theta_{m-1} - \theta_0\| \int_{x'} V(x') \delta_x P_{\theta_{0:m-1}}(dx') \\ &\leq \sum_{m=2}^n K\rho^{n-m} L_d M (V\eta)_{0:m-1} \frac{KV(x)}{1-\rho} \leq \frac{L_d K^2 M V(x)}{(1-\rho)^2} (V\eta)_{0:n}. \end{aligned}$$

□

Similar to Lemma 5, we also have

LEMMA 6. *Under Assumptions 1 and 2,*

$$W_d(\delta_x P_{\theta_{0:n}}, \delta_x P_{\theta_n}^n) \leq \frac{L_d K^2 M V(x)}{(1-\rho)^2} (V\eta)_{0:n}.$$

Proof. Define $P_{\theta_{m:m}} = I$ and $P^0_{\theta_n} = I$.

$$\begin{aligned}
W_d(\delta_x P_{\theta_{0:n}}, \delta_x P^n_{\theta_n}) &\le \sum_{m=1}^{n} W_d(\delta_x P_{\theta_{0:m}} P^{n-m}_{\theta_n}, \delta_x P_{\theta_{0:m-1}} P^{n-m+1}_{\theta_n}) \\
&\le \sum_{m=1}^{n} K\rho^{n-m} W_d(\delta_x P_{\theta_{0:m-1}} P_{\theta_{m-1}}, \delta_x P_{\theta_{0:m-1}} P_{\theta_n}) \\
&\le \sum_{m=1}^{n} K\rho^{n-m} L_d \|\theta_{m-1} - \theta_n\| \int_{x'} V(x') \delta_x P_{\theta_{0:m-1}}(dx') \\
&\le \sum_{m=1}^{n} K\rho^{n-m} L_d M (V\eta)_{m-1:n} \frac{KV(x)}{1-\rho} \le \frac{L_d K^2 M V(x)}{(1-\rho)^2} (V\eta)_{0:n}.
\end{aligned}$$

□

Lemma 7. *Under Assumptions 1 and 2, we have*

$$\|\nabla \ell(\theta_t) - \mathbb{E}_t g(\theta_t, z_{t+\tau_\epsilon})\| \le L\epsilon V(z_{t-1}) + \frac{LL_d K^2 M}{(1-\rho)^2} \eta_{t:t+\tau_\epsilon} (V(z_{t-1}) + K)^2.$$

Proof. Note that

$$\begin{aligned}
&\|\nabla \ell(\theta_t) - \mathbb{E}_t g(\theta_t, z_{t+\tau_\epsilon})\| \\
\le& \left\| \int_z g(\theta_t, z) \left(\mu_{\theta_t}(dz) - \delta_{z_{t-1}} P^{\tau_\epsilon}_{\theta_t}(dz) \right) \right\| + \left\| \mathbb{E}_t \int_z g(\theta_t, z) \left(\delta_{z_{t-1}} P_{\theta_{t:t+\tau_\epsilon}}(dz) - \delta_{z_{t-1}} P^{\tau_\epsilon}_{\theta_t}(dz) \right) \right\| \\
\le& L\epsilon V(z_{t-1}) + \frac{LL_d K^2 M V(z_{t-1})}{(1-\rho)^2} \mathbb{E}_t[(V\eta)_{t:t+\tau_\epsilon}] \\
\le& L\epsilon V(z_{t-1}) + \frac{LL_d K^2 M V(z_{t-1})}{(1-\rho)^2} (V(z_{t-1}) + K) \eta_{t:t+\tau_\epsilon},
\end{aligned}$$

by Lemma 5, Kantorovich-Rubenstein duality, and the fact that V is a Lyapunov function. □

Lemma 8. *Under Assumptions 1 and 2, we have*

$$\|\mathbb{E}_t[\nabla \ell(\theta_{t+\tau_\epsilon}) - g(\theta_{t+\tau_\epsilon}, z_{t+\tau_\epsilon})]\| \le L\epsilon V(z_{t-1}) + \frac{LL_d K^2 M}{(1-\rho)^2} \eta_{t:t+\tau_\epsilon} (V(z_{t-1}) + K)^2.$$

Proof. Note that

$$\begin{aligned}
&\|\mathbb{E}_t[\nabla \ell(\theta_{t+\tau_\epsilon}) - g(\theta_{t+\tau_\epsilon}, z_{t+\tau_\epsilon})]\| \\
\le& \left\| \int_z g(\theta_{t+\tau_\epsilon}, z) \left(\mu_{\theta_{t+\tau_\epsilon}}(dz) - \delta_{z_{t-1}} P^{\tau_\epsilon}_{\theta_{t+\tau_\epsilon}}(dz) \right) \right\| \\
&+ \left\| E_t \int_z g(\theta_{t+\tau_\epsilon}, z) \left(\delta_{z_{t-1}} P_{\theta_{t:t+\tau_\epsilon}}(dz) - \delta_{z_{t-1}} P^{\tau_\epsilon}_{\theta_{t+\tau_\epsilon}}(dz) \right) \right\| \\
\le& L\epsilon V(z_{t-1}) + \frac{LL_d K^2 M V(z_{t-1})}{(1-\rho)^2} \mathbb{E}_t[(V\eta)_{t:t+\tau_\epsilon}] \\
\le& L\epsilon V(z_{t-1}) + \frac{LL_d K^2 M V(z_{t-1})}{(1-\rho)^2} (V(z_{t-1}) + K) \eta_{t:t+\tau_\epsilon},
\end{aligned}$$

by Lemma 6, Kantorovich-Rubenstein duality, and the fact that V is a Lyapunov function. □

LEMMA 9. *Under Assumptions 2 and 3, the iterates satisfy the following*

$$\|\mathbb{E}_t[g(\theta_t, z_{t+\tau_\epsilon}) - g(\theta_{t+\tau_\epsilon}, z_{t+\tau_\epsilon})]\| \leq \left(2L\epsilon + 2\frac{LL_dK^2M}{(1-\rho)^2}\eta_{t:t+\tau_\epsilon} + ML\eta_{t:t+\tau_\epsilon}\right)(V(z_{t-1}) + K)^2.$$

Proof. Note that

$$\begin{aligned}
&\|\mathbb{E}_t[g(\theta_t, z_{t+\tau_\epsilon}) - g(\theta_{t+\tau_\epsilon}, z_{t+\tau_\epsilon})]\| \\
\leq&\|\mathbb{E}_t g(\theta_t, z_{t+\tau_\epsilon}) - \nabla l(\theta_t)\| + \|\mathbb{E}_t[g(\theta_{t+\tau_\epsilon}, z_{t+\tau_\epsilon}) - \nabla l(\theta_{t+\tau_\epsilon})]\| + \|\mathbb{E}_t \nabla l(\theta_{t+\tau_\epsilon}) - \nabla l(\theta_t)\| \\
\leq&2L\epsilon V(z_{t-1}) + 2\frac{LL_dK^2M}{(1-\rho)^2}\eta_{t:t+\tau_\epsilon}(V(z_{t-1}) + K)^2 + ML(V(z_{t-1}) + K)\eta_{t:t+\tau_\epsilon}
\end{aligned}$$

by Lemmas 4, 7, and 8. □

We are now ready to prove the main theorems.

Proof of Theorem 1. Since $\theta_{t+1} = \theta_t - \eta_t \hat{g}_t(\theta_t, z_t)$ and $\nabla\ell(\theta)$ is L-smooth,

$$\ell(\theta_{t+1}) \leq \ell(\theta_t) - \eta_t\langle\nabla\ell(\theta_t), \hat{g}_t(\theta_t, z_t)\rangle + \frac{L}{2}\eta_t^2\|\hat{g}_t(\theta_t, z_t)\|^2.$$

Since $\|\hat{g}_t(\theta_t, z_t)\| \leq MV(z_t)$, the above inequality can be rearranged as

$$\eta_t\langle\nabla\ell(\theta_t), \nabla\ell(\theta_t)\rangle \leq \ell(\theta_t) - \ell(\theta_{t+1}) + \frac{L}{2}\eta_t^2M^2V(z_t)^2 + \eta_t\langle\nabla\ell(\theta_t) - \hat{g}_t(\theta_t, z_t), \nabla\ell(\theta_t)\rangle.$$

Taking the sum over t, we have

$$\sum_{t=0}^{T-1}\eta_t\|\nabla\ell(\theta_t)\|^2 \leq \ell(\theta_0) + \frac{1}{2}LM^2(V\eta)_{0:T}^2 + \sum_{t=0}^{T-1}\eta_t\langle\nabla\ell(\theta_t) - \hat{g}_t(\theta_t, z_t), \nabla\ell(\theta_t)\rangle.$$

We next establish an appropriate bound for $\langle\nabla\ell(\theta_t) - \hat{g}_t(\theta_t, z_t), \nabla\ell(\theta_t)\rangle$. Consider the decomposition

$$\begin{aligned}
&\langle\nabla\ell(\theta_t) - \hat{g}_t(\theta_t, z_t), \nabla\ell(\theta_t)\rangle \\
=&\underbrace{\langle g(\theta_t, z_t) - \hat{g}_t(\theta_t, z_t), \nabla\ell(\theta_t)\rangle}_{(a)} + \underbrace{\langle g(\theta_{t+\tau_\epsilon}, z_{t+\tau_\epsilon}) - g(\theta_t, z_t), \nabla\ell(\theta_t)\rangle}_{(b)} \\
&+\underbrace{\langle g(\theta_t, z_{t+\tau_\epsilon}) - g(\theta_{t+\tau_\epsilon}, z_{t+\tau_\epsilon}), \nabla\ell(\theta_t)\rangle}_{(c)} + \underbrace{\langle\nabla\ell(\theta_t) - g(\theta_t, z_{t+\tau_\epsilon}), \nabla\ell(\theta_t)\rangle}_{(d)},
\end{aligned}$$

where $z_{t+\tau_\epsilon} \sim P_{\theta_{t:t+\tau_\epsilon}}$.

For (a), we have

$$|\mathbb{E}_t\langle g(\theta_t, z_t) - \hat{g}_t(\theta_t, z_t), \nabla\ell(\theta_t)\rangle| \leq \|\nabla\ell(\theta_t)\|\mathbb{E}_t\|g(\theta_t, z_t) - \hat{g}_t(\theta_t, z_t)\| \leq M\mathbb{E}_t e_t.$$

For (c), by Lemma 9, we have

$$\begin{aligned}
&|\mathbb{E}_t\langle g(\theta_t, z_{t+\tau_\epsilon}) - g(\theta_{t+\tau_\epsilon}, z_{t+\tau_\epsilon}), \nabla\ell(\theta_t)\rangle| \\
\leq&\|\mathbb{E}_t[g(\theta_t, z_{t+\tau_\epsilon}) - g(\theta_{t+\tau_\epsilon}, z_{t+\tau_\epsilon})]\|\|\nabla\ell(\theta_t)\| \\
\leq&M\left(2L\epsilon + 2\frac{LL_dK^2M}{(1-\rho)^2}\eta_{t:t+\tau_\epsilon} + ML\eta_{t:t+\tau_\epsilon}\right)(V(z_{t-1}) + K)^2 \\
\leq&M\left(2L\epsilon + 2\frac{LL_dK^2M}{(1-\rho)^2}\tau_\epsilon\eta_t + ML\tau_\epsilon\eta_t\right)(V(z_{t-1}) + K)^2.
\end{aligned}$$

For (d), following Lemma 7, we have

$$\begin{aligned}
&|\mathbb{E}_t[\langle \nabla \ell(\theta_t) - g(\theta_t, z_{t+\tau_\epsilon}), \nabla \ell(\theta_t)\rangle]| \\
\leq &\|\mathbb{E}_t[\nabla \ell(\theta_t) - g(\theta_t, z_{t+\tau_\epsilon})]\| \|\nabla \ell(\theta_t)\| \\
\leq &M\left(L\epsilon + \frac{LL_dK^2M}{(1-\rho)^2}\eta_{t:t+\tau_\epsilon}\right)(V(z_{t-1}) + K)^2 \\
\leq &M\left(L\epsilon + \frac{LL_dK^2M}{(1-\rho)^2}\tau_\epsilon\eta_t\right)(V(z_{t-1}) + K)^2.
\end{aligned}$$

Lastly, for (b), taking the sum over t, we have

$$\begin{aligned}
&\left|\mathbb{E}\sum_{t=0}^{T-1}\eta_t\langle g(\theta_{t+\tau_\epsilon}, z_{t+\tau_\epsilon}) - g(\theta_t, z_t), \nabla \ell(\theta_t)\rangle\right| \\
\leq &\left|\mathbb{E}\sum_{t=\tau_\epsilon}^{T-1}\langle g(\theta_t, z_t), \eta_{t-\tau_\epsilon}\nabla \ell(\theta_{t-\tau_\epsilon}) - \eta_t \nabla \ell(\theta_t)\rangle\right| + \left|\mathbb{E}\sum_{t=0}^{\tau_\epsilon-1}\langle g(\theta_t, z_t), \eta_t \nabla \ell(\theta_t)\rangle\right| \\
&+ \left|\mathbb{E}\sum_{t=T}^{T+\tau_\epsilon-1}\langle g(\theta_t, z_t), \eta_{t-\tau_\epsilon}\nabla \ell(\theta_{t-\tau_\epsilon})\rangle\right| \\
\leq &\left|\mathbb{E}\sum_{t=\tau_\epsilon}^{T-1}\langle g(\theta_t, z_t), \eta_t\nabla \ell(\theta_{t-\tau_\epsilon}) - \eta_t \nabla \ell(\theta_t)\rangle\right| + \left|\mathbb{E}\sum_{t=\tau_\epsilon}^{T-1}\langle g(\theta_t, z_t), (\eta_{t-\tau_\epsilon} - \eta_t)\nabla \ell(\theta_{t-\tau_\epsilon})\rangle\right| \\
&+ \left|\mathbb{E}\sum_{t=0}^{\tau_\epsilon-1}\langle g(\theta_t, z_t), \eta_t \nabla \ell(\theta_t)\rangle\right| + \left|\mathbb{E}\sum_{t=T}^{T+\tau_\epsilon-1}\langle g(\theta_t, z_t), \eta_{t-\tau_\epsilon}\nabla \ell(\theta_{t-\tau_\epsilon})\rangle\right| \\
\leq &M^2L\mathbb{E}\sum_{t=\tau_\epsilon}^{T-1}\eta_t V(z_t)(V\eta)_{t-\tau_\epsilon:t} + M^2(V(z_0) + K)\eta_{0:\tau_\epsilon} \\
&+ M^2\mathbb{E}(V\eta)_{0:\tau_\epsilon} + M^2(V(z_0) + K)\eta_{T-\tau_\epsilon:T} \\
\leq &M^2L\tau_\epsilon\sum_{t=0}^{T-1}\mathbb{E}(V(z_t) + K)^2\eta_t^2 + 3M^2\tau_\epsilon(V(z_0) + K).
\end{aligned}$$

Putting the bounds for (a) – (d) together, we have

$$\begin{aligned}
&\left|\mathbb{E}\sum_{t=0}^{T-1}\eta_t\langle \nabla \ell(\theta_t) - \hat{g}_t(\theta_t, z_t), \nabla \ell(\theta_t)\rangle\right| \\
\leq &M\sum_{t=0}^{T-1}\eta_t\mathbb{E}e_t + 3M^2\tau_\epsilon(V(z_0) + K) + 3ML\epsilon\sum_{t=0}^{T-1}\eta_t\mathbb{E}(V(z_t) + K)^2 \\
&+ \left(3\frac{M^2LL_dK^2}{(1-\rho)^2}\tau_\epsilon + 2M^2L\tau_\epsilon\right)\sum_{t=0}^{T-1}\eta_t^2\mathbb{E}(V(z_t) + K)^2.
\end{aligned}$$

Thus,

$$\begin{aligned}
\sum_{t=0}^{T-1}\eta_t\mathbb{E}\|\nabla \ell(\theta_t)\|^2 \leq &\ell(\theta_0) + \frac{1}{2}LM^2(V\eta)_{0:T} + \sum_{t=0}^{T-1}\eta_t\mathbb{E}\langle \nabla \ell(\theta_t) - \hat{g}_t(\theta_t, z_t), \nabla \ell(\theta_t)\rangle \\
\leq &\ell(\theta_0) + 3M^2\tau_\epsilon(V(z_0) + K) + M\sum_{t=0}^{T-1}\eta_t\mathbb{E}e_t + 3ML\epsilon\sum_{t=0}^{T-1}\eta_t\mathbb{E}(V(z_t) + K)^2
\end{aligned}$$

$$+\left(3\frac{M^2LL_dK^2}{(1-\rho)^2}\tau_\epsilon+2M^2L\tau_\epsilon\right)\sum_{t=0}^{T-1}\eta_t^2\mathbb{E}(V(z_t)+K)^2. \tag{21}$$

For $\epsilon=1/\sqrt{T}$, $\tau_\epsilon=O(\tau\log T)$. Let $\tilde{\eta}_t=\eta_t/\eta_{0:T}$. Then, we have

$$\sum_{t=0}^{T-1}\tilde{\eta}_t\mathbb{E}\|\nabla\ell(\theta_t)\|^2=O\left(\frac{1}{\eta_{0:T}}\left(\tau\log T+\tau\log T\eta_{0:T}^2+\frac{1}{\sqrt{T}}\eta_{0:T}+\sum_{t=0}^{T-1}\eta_t\mathbb{E}e_t\right)\right).$$

This implies that

$$\min_{0\le t<T}\mathbb{E}\|\nabla\ell(\theta_t)\|^2=O\left(\frac{1}{\eta_{0:T}}\left(\tau\log T+\tau\log T\eta_{0:T}^2+\frac{1}{\sqrt{T}}\eta_{0:T}+\sum_{t=0}^{T-1}\eta_t\mathbb{E}e_t\right)\right).$$

□

Before we prove Case 1 of Theorem 2, we provide a bound for $\mathbb{E}\|\theta_t-\theta^*\|^2$ first. Note that when Θ is bounded, $\|\theta_t-\theta^*\|^2$ is bounded. We will show in Lemma 10 that even when Θ is not bounded, we can still establish appropriate bounds for $\mathbb{E}\|\theta_t-\theta^*\|^2$.

LEMMA 10. *Assume ℓ is convex and $\mathbb{E}e_t=O(1/\sqrt{t})$. Let $\eta_t=\eta_0/\sqrt{t}$. There exists a constant $C\in(0,\infty)$, such that for any fixed $T>0$, we have*

$$\mathbb{E}\|\theta_T-\theta^*\|^2\le C^2\left(\|\theta_0-\theta^*\|+\tau(\log T)^2+(V(z_0)+K)^2\right)^2,$$

for $t\le T$.

Proof. Let $\epsilon=1/\sqrt{T}$. Then $\tau_\epsilon=O(\tau\log T)$.

By the convexity of Θ and the fact that $\theta^*\in\Theta$, we have

$$\begin{aligned}\|\theta_{t+1}-\theta^*\|^2&\le\|\theta_t-\theta^*-\eta_t\hat{g}_t(\theta_t,z_t)\|^2\\&\le\|\theta_t-\theta^*\|^2+2\eta_t\langle\hat{g}_t(\theta_t,z_t),\theta^*-\theta_t\rangle+\eta_t^2M^2V(z_t)^2\\&=\|\theta_t-\theta^*\|^2+2\eta_t\langle\nabla\ell(\theta_t),\theta^*-\theta_t\rangle+\eta_t^2M^2V(z_t)^2+2\eta_t\langle\hat{g}_t(\theta_t,z_t)-\nabla\ell(\theta_t),\theta^*-\theta_t\rangle\\&\le\|\theta_t-\theta^*\|^2+\eta_t^2M^2V(z_t)^2+2\eta_t\langle\hat{g}_t(\theta_t,z_t)-\nabla\ell(\theta_t),\theta^*-\theta_t\rangle,\end{aligned}$$

since ℓ is convex. By induction, we have

$$\|\theta_t-\theta^*\|^2\le\|\theta_0-\theta^*\|^2+M^2(V\eta)_{0:t}^2+2\sum_{s=0}^{t-1}\eta_s\langle\hat{g}_s(\theta_s,z_s)-\nabla\ell(\theta_s),\theta^*-\theta_s\rangle.$$

Next, note that

$$\begin{aligned}&\langle\nabla\ell(\theta_s)-\hat{g}_s(\theta_s,z_s),\theta_s-\theta^*\rangle\\=&\underbrace{\langle g(\theta_s,z_s)-\hat{g}_s(\theta_s,z_s),\theta_s-\theta^*\rangle}_{(a)}+\underbrace{\langle g(\theta_{s+\tau_\epsilon},z_{s+\tau_\epsilon})-g(\theta_s,z_s),\theta_s-\theta^*\rangle}_{(b)}\\&+\underbrace{\langle g(\theta_s,z_{s+\tau_\epsilon})-g(\theta_{s+\tau_\epsilon},z_{s+\tau_\epsilon}),\theta_s-\theta^*\rangle}_{(c)}+\underbrace{\langle\nabla\ell(\theta_s)-g(\theta_s,z_{s+\tau_\epsilon}),\theta_s-\theta^*\rangle}_{(d)}.\end{aligned}$$

We shall bound each of the terms in the decomposition. For (a), we have

$$\begin{aligned}\eta_s|\mathbb{E}_s\langle g(\theta_s,z_s)-\hat{g}_s(\theta_s,z_s),\theta_s-\theta^*\rangle| \leq& \eta_s\|\mathbb{E}_s[g(\theta_s,z_s)-\hat{g}_s(\theta_s,z_s)]\|\|\theta_s-\theta^*\|\\ \leq& \eta_s\mathbb{E}_s e_s\|\theta_s-\theta^*\|.\end{aligned}$$

For (c), by Lemma 9, we have

$$\begin{aligned}&\eta_s|\mathbb{E}_s\langle g(\theta_s,z_{s+\tau_\epsilon})-g(\theta_{s+\tau_\epsilon},z_{s+\tau_\epsilon}),\theta_s-\theta^*\rangle|\\ \leq&\eta_s\|\mathbb{E}_s[g(\theta_s,z_{s+\tau_\epsilon})-g(\theta_{s+\tau_\epsilon},z_{s+\tau_\epsilon})]\|\|\theta_s-\theta^*\|\\ \leq&\eta_s\left(2L\epsilon+2\frac{LL_dK^2M}{(1-\rho)^2}\eta_{s:s+\tau_\epsilon}+ML\eta_{s:s+\tau_\epsilon}\right)(V(z_{s-1})+K)^2\|\theta_s-\theta^*\|.\end{aligned}$$

For (d), by Lemma 7, we have

$$\begin{aligned}&\eta_s\left|\mathbb{E}_s\langle\nabla\ell(\theta_s)-g(\theta_s,z_{s+\tau_\epsilon}),\theta_s-\theta^*\rangle\right|\\ &\leq\eta_s\left(L\epsilon+\frac{LL_dK^2M}{(1-\rho)^2}\eta_{s:s+\tau_\epsilon}\right)(V(z_{s-1})+K)^2\|\theta_s-\theta^*\|.\end{aligned}$$

Lastly, for (b), taking the sum over s, we have

$$\begin{aligned}&\left|\mathbb{E}\sum_{s=0}^{t-1}\eta_s\langle g(\theta_{s+\tau_\epsilon},z_{s+\tau_\epsilon})-g(\theta_s,z_s),\theta_s-\theta^*\rangle\right|\\ \leq&\left|\mathbb{E}\sum_{s=\tau_\epsilon}^{t-1}\langle g(\theta_s,z_s),\eta_s(\theta_{s-\tau_\epsilon}-\theta_s)\rangle\right|+\left|\mathbb{E}\sum_{s=\tau_\epsilon}^{t-1}\langle g(\theta_s,z_s),(\eta_{s-\tau_\epsilon}-\eta_s)(\theta_{s-\tau_\epsilon}-\theta^*)\rangle\right|\\ &+\left|\mathbb{E}\sum_{s=0}^{\tau_\epsilon-1}\eta_s\langle g(\theta_s,z_s),\theta_s-\theta^*\rangle\right|+\left|\mathbb{E}\sum_{s=t}^{t+\tau_\epsilon-1}\eta_{s-\tau_\epsilon}\langle g(\theta_s,z_s),\theta_{s-\tau_\epsilon}-\theta^*\rangle\right|\\ \leq&\mathbb{E}\sum_{s=\tau_\epsilon}^{t-1}M^2V(z_s)\eta_s(V\eta)_{s-\tau_\epsilon:s}+M\mathbb{E}\sum_{s=0}^{t-\tau_\epsilon-1}\eta_s^3\tau_\epsilon\|\theta_s-\theta^*\|V(z_{s+\tau_\epsilon})\\ &+M\mathbb{E}\sum_{s=0}^{\tau_\epsilon-1}\eta_sV(z_s)\|\theta_s-\theta^*\|+M\mathbb{E}\sum_{s=t-\tau_\epsilon}^{t-1}\eta_sV(z_{s+\tau_\epsilon})\|\theta_s-\theta^*\|.\end{aligned}$$

Putting the bounds for (a) to (d) together, we have

$$\begin{aligned}&\left|\mathbb{E}\sum_{s=0}^{t-1}\eta_s\langle\nabla\ell(\theta_s)-\hat{g}_s(\theta_s,z_s),\theta_s-\theta^*)\rangle\right|\\ \leq&\sum_{s=0}^{t-1}\eta_s\mathbb{E}e_s\mathbb{E}\|\theta_s-\theta^*\|+M^2\tau_\epsilon\sum_{s=\tau_\epsilon}^{t-1}\eta_s^2\mathbb{E}(V(z_s)+K)^2\\ &+M\tau_\epsilon\sum_{s=1}^{t-1}\eta_s^3\mathbb{E}(V(z_s)+K)^2\|\theta_s-\theta^*\|+M\sum_{s=0}^{\tau_\epsilon-1}\eta_s\mathbb{E}V(z_s)\|\theta_s-\theta^*\|\\ &+M\sum_{s=t-\tau_\epsilon}^{t-1}\eta_s\mathbb{E}(V(z_s)+K)\|\theta_s-\theta^*\|+3L\epsilon\sum_{s=0}^{t-1}\eta_s\mathbb{E}(V(z_s)+K)^2\|\theta_s-\theta^*\|\\ &+\left(3\frac{LL_dK^2M}{(1-\rho)^2}\tau_\epsilon+ML\tau_\epsilon\right)\sum_{s=0}^{t-1}\eta_s^2\mathbb{E}(V(z_s)+K)^2\|\theta_s-\theta^*\|.\end{aligned}$$

Next, we prove the result by induction. Suppose for $s \le t$,

$$\mathbb{E}\|\theta_s - \theta^*\|^2 \le C^2 \left(\|\theta_0 - \theta^*\| + \tau_\epsilon \log t + \epsilon\sqrt{s} + (V(z_0)+K)^2\right)^2.$$

Then, since $\mathbb{E}\left[(V(z_{s-1})+K)^2\|\theta_{s-1}-\theta^*\|\right] \le \sqrt{\mathbb{E}(V(z_s)+K)^4}\sqrt{\mathbb{E}\|\theta_s-\theta^*\|^2}$,

$$\begin{aligned}
&\left|\mathbb{E}\sum_{s=0}^{t-1}\eta_s\langle\nabla\ell(\theta_s) - \hat{g}_s(\theta_s, z_s), \theta_s - \theta^*\rangle\right| \\
\le&\sum_{s=0}^{t-1}\eta_s\mathbb{E}e_s C(\|\theta_0-\theta^*\| + \tau_\epsilon\log t + \epsilon\sqrt{s} + (V(z_0)+K)^2) + M^2\tau_\epsilon\sum_{s=\tau_\epsilon}^{t-1}\eta_s^2\mathbb{E}(V(z_s)+K)^2 \\
&+ M\tau_\epsilon\sum_{s=1}^{t-1}\eta_s^3\sqrt{\mathbb{E}(V(z_s)+K)^4}C(\|\theta_0-\theta^*\| + \tau_\epsilon\log t + \epsilon\sqrt{s} + (V(z_0)+K)^2) \\
&+ M\sum_{s=0}^{\tau_\epsilon-1}\eta_s\sqrt{\mathbb{E}V(z_s)^2}C(\|\theta_0-\theta^*\| + \tau_\epsilon\log t + \epsilon\sqrt{s} + (V(z_0)+K)^2) \\
&+ M\sum_{s=t-\tau_\epsilon}^{t-1}\eta_s\sqrt{\mathbb{E}(V(z_s)+K)^2}C(\|\theta_0-\theta^*\| + \tau_\epsilon\log t + \epsilon\sqrt{s} + (V(z_0)+K)^2) \\
&+ 3L\epsilon\sum_{s=0}^{t-1}\eta_s\sqrt{\mathbb{E}(V(z_s)+K)^4}C(\|\theta_0-\theta^*\| + \tau_\epsilon\log t + \epsilon\sqrt{s} + (V(z_0)+K)^2) \\
&+ \left(3\frac{L^2K^2M}{(1-\rho)^2}\tau_\epsilon + ML\tau_\epsilon\right)\sum_{s=0}^{t-1}\eta_s^2\sqrt{\mathbb{E}(V(z_s)+K)^4}C(\|\theta_0-\theta^*\| + \tau_\epsilon\log t + \epsilon\sqrt{s} + (V(z_0)+K)^2) \\
\le&C'(\|\theta_0-\theta^*\| + \tau_\epsilon\log t + (V(z_0)+K)^2)C(\tau_\epsilon\log t + \epsilon\sqrt{t} + (V(z_0)+K)^2),
\end{aligned}$$

where C' is a constant that does not depend on C.

Thus,

$$\begin{aligned}
\mathbb{E}\|\theta_t-\theta^*\|^2 \le& \|\theta_0-\theta^*\|^2 + M^2\log t(V(z_0)+K)^2 \\
&+ C'C(\|\theta_0-\theta^*\| + \tau_\epsilon\log t + (V(z_0)+K)^2)(\tau_\epsilon\log t + \epsilon\sqrt{t} + (V(z_0)+K)^2) \\
&\le C^2(\|\theta_0-\theta^*\| + \tau_\epsilon\log t + \epsilon\sqrt{t} + (V(z_0)+K)^2)^2.
\end{aligned}$$

□

Proof of Case 1 of Theorem 2. By the convexity of Θ and the fact that $\theta^* \in \Theta$, we have

$$\begin{aligned}
\|\theta_{t+1}-\theta^*\|^2 &\le \|\theta_t - \theta^* - \eta_t\hat{g}_t(\theta_t, z_t)\|^2 \\
&= \|\theta_t-\theta^*\|^2 + 2\eta_t\langle\hat{g}_t(\theta_t,z_t), \theta^*-\theta_t\rangle + \eta_t^2\|\hat{g}_t(\theta_t,z_t)\|^2 \\
&\le \|\theta_t-\theta^*\|^2 + 2\eta_t\langle\hat{g}_t(\theta_t,z_t), \theta^*-\theta_t\rangle + \eta_t^2M^2V(z_t)^2.
\end{aligned}$$

Therefore,

$$\langle\hat{g}_t(\theta_t,z_t), \theta^*-\theta_t\rangle \ge \frac{1}{2\eta_t}(\|\theta_{t+1}-\theta^*\|^2 - \|\theta_t-\theta^*\|^2) - \frac{1}{2}\eta_t M^2 V(z_t)^2.$$

Next, by the convexity of ℓ,

$$\begin{aligned}\ell(\theta^*) - \ell(\theta_t) &\geq \langle \nabla \ell(\theta_t), \theta^* - \theta_t \rangle \\ &= \langle \hat{g}_t(\theta_t, z_t), \theta^* - \theta_t \rangle + \langle \nabla \ell(\theta_t) - \hat{g}_t(\theta_t, z_t), \theta^* - \theta_t \rangle \\ &\geq \frac{1}{2\eta_t}(\|\theta_{t+1} - \theta^*\|^2 - \|\theta_t - \theta^*\|^2) - \frac{1}{2}\eta_t V(z_t)^2 M^2 + \langle \nabla \ell(\theta_t) - \hat{g}_t(\theta_t, z_t), \theta^* - \theta_t \rangle.\end{aligned}$$

This leads to

$$\begin{aligned}\frac{1}{\eta_{0:T}} \sum_{t=0}^{T-1} \eta_t (\ell(\theta_t) - \ell(\theta^*)) \leq & \frac{1}{2\eta_{0:T}} \|\theta_0 - \theta^*\|^2 + \frac{1}{2} M^2 \frac{(V\eta)_{0:T}^2}{\eta_{0:T}} \\ & + \frac{1}{\eta_{0:T}} \sum_{t=0}^{T-1} \eta_t \langle \nabla \ell(\theta_t) - \hat{g}_t(\theta_t, z_t), \theta_t - \theta^* \rangle.\end{aligned}$$

From the proof of Lemma 10, for $\epsilon = 1/\sqrt{T}$, $\tau_\epsilon = O(\tau \log T)$, we have

$$\begin{aligned}&\left| \mathbb{E} \sum_{t=0}^{T-1} \eta_t \langle \nabla \ell(\theta_t) - \hat{g}_t(\theta_t, z_t), \theta_t - \theta^* \rangle \right| \\ \leq & C'(\|\theta_0 - \theta^*\| + \tau_\epsilon \log T + (V(z_0) + K)^2) C(\tau_\epsilon \log T + \epsilon\sqrt{T} + (V(z_0) + K)^2).\end{aligned}$$

Then,

$$\begin{aligned}&\frac{1}{\eta_{0:T}} \sum_{t=0}^{T-1} \eta_t (\ell(\theta_t) - \ell(\theta^*)) \\ \leq & \frac{\|\theta_0 - \theta^*\|^2}{\eta_{0:T}} + \frac{M^2}{2\eta_{0:T}} \sum_{t=0}^{T-1} \eta_t^2 \mathbb{E} V(z_t)^2 \\ & + \frac{C'C}{\eta_{0:T}} (\|\theta_0 - \theta^*\| + \tau_\epsilon \log T + (V(z_0) + K)^2)(\tau_\epsilon \log T + \epsilon\sqrt{T} + (V(z_0) + K)^2) \\ = & O\left(\frac{1}{\sqrt{T}} + \frac{\log T}{\sqrt{T}} + \frac{\tau (\log T)^2}{\sqrt{T}} + \frac{\tau^2 (\log T)^4}{\sqrt{T}} \right).\end{aligned}$$

For $\bar{\theta}_T = \frac{1}{\eta_{0:T}} \sum_{t=0}^{T-1} \eta_t \theta_t$, by the convexity of ℓ, we have

$$\mathbb{E}\ell(\bar{\theta}_T) - \ell(\theta^*) \leq \frac{1}{\eta_{0:T}} \sum_{t=0}^{T-1} \eta_t (\ell(\theta_t) - \ell(\theta^*)) = O\left(\tau^2 (\log T)^4 / \sqrt{T} \right).$$

□

We next present an auxiliary result about our choice of step size in the strongly convex case.

LEMMA 11. *If $\eta_t = 2\eta_0/(ct)$ and $\eta_0 > 2$, we have $\exp(-\frac{1}{2} c \eta_{t+1:T}) \eta_t$ is increasing in t.*

Proof. Through induction, we only need to show that $\exp\left(-\frac{1}{2} c \eta_{t+1}\right) \eta_t \leq \eta_{t+1}$, which is equivalent to showing

$$\exp(-\frac{1}{2} c \eta_{t+1}) \leq \frac{t}{t+1},$$

which is true with $\eta_{t+1} \geq \frac{4}{c(t+1)}$. □

Proof of Case 2 in Theorem 2. First, note that

$$\begin{aligned}\|\theta_{t+1}-\theta^*\|^2 &= \|\theta_t-\eta_t\hat{g}(\theta_t,z_t)-\theta^*\|^2\\ &\le \|\theta_t-\theta^*\|^2-2\langle\theta_t-\theta^*,\hat{g}(\theta_t,z_t)\rangle\eta_t+M^2\eta_t^2V(z_t)^2.\end{aligned}$$

Consider the following decomposition of $\hat{g}(\theta_t,z_t)$:

$$\begin{aligned}\hat{g}(\theta_t,z_t) =& \underbrace{[\hat{g}(\theta_t,z_t)-g(\theta_t,z_t)]}_{(a)}+\underbrace{[g(\theta_t,z_t)-g(\theta_{t+\tau_\epsilon},z_{t+\tau_\epsilon})]}_{(b)}+\underbrace{[g(\theta_{t+\tau_\epsilon},z_{t+\tau_\epsilon})-g(\theta_t,z_{t+\tau_\epsilon})]}_{(c)}\\ &+\underbrace{[g(\theta_t,z_{t+\tau_\epsilon})-\nabla\ell(\theta_t)]}_{(d)}+\underbrace{\nabla\ell(\theta_t)}_{(e)}.\end{aligned}$$

We next bound the inner product of each part in (22) with $\theta^*-\theta_t$, except for part (b). Part (b) will be treated separately later.

For (e), since ℓ is c-strongly convex,

$$-\langle\nabla\ell(\theta_t),\theta_t-\theta^*\rangle\le-c\|\theta_t-\theta^*\|^2.$$

For (a), under Assumption 5,

$$\begin{aligned}-\langle\hat{g}(\theta_t,z_t)-g(\theta_t,z_t),\theta_t-\theta^*\rangle &\le \frac{1}{c}\|\hat{g}(\theta_t,z_t)-g(\theta_t,z_t)\|^2+\frac{1}{4}c\|\theta_t-\theta^*\|^2\\ &\le \frac{1}{c}e_t^2+\frac{1}{4}c\|\theta_t-\theta^*\|^2.\end{aligned}$$

For (c), following Lemma 9, we have

$$\begin{aligned}&-\mathbb{E}_t\langle g(\theta_{t+\tau_\epsilon},z_{t+\tau_\epsilon})-g(\theta_t,z_{t+\tau_\epsilon}),\theta_t-\theta^*\rangle\\ \le&\frac{1}{c}\|\mathbb{E}_t[g(\theta_{t+\tau_\epsilon},z_{t+\tau_\epsilon})-g(\theta_t,z_{t+\tau_\epsilon})]\|^2+\frac{1}{4}c\|\theta_t-\theta^*\|^2\\ \le&\frac{1}{c}\left(2L\epsilon+2\frac{LL_dK^2M}{(1-\rho)^2}\eta_{t:t+\tau_\epsilon}+ML\eta_{t:t+\tau_\epsilon}\right)^2(V(z_{t-1})+K)^4+\frac{1}{4}c\|\theta_t-\theta^*\|^2.\end{aligned}$$

For (d), following Lemma 7, we have

$$\begin{aligned}&-\mathbb{E}_t\langle g(\theta_t,z_{t+\tau_\epsilon})-\nabla\ell(\theta_t),\theta_t-\theta^*\rangle\\ \le&\frac{1}{c}\|\mathbb{E}_tg(\theta_t,z_{t+\tau_\epsilon})-\nabla\ell(\theta_t)\|^2+\frac{1}{4}c\|\theta_t-\theta^*\|^2\\ \le&\frac{1}{c}\left(L\epsilon+\frac{LL_dK^2M}{(1-\rho)^2}\eta_{t:t+\tau_\epsilon}\right)^2(V(z_{t-1})+K)^4+\frac{1}{4}c\|\theta_t-\theta^*\|^2\\ \le&\frac{2}{c}L^2\epsilon^2(V(z_{t-1})+K)^4+\frac{2}{c}\frac{L^2L_d^2K^4M^2}{(1-\rho)^4}\tau_\epsilon^2\eta_t^2(V(z_{t-1})+K)^4+\frac{1}{4}c\|\theta_t-\theta^*\|^2.\end{aligned}$$

Putting together the bounds for parts (a), (c), (d), and (e), we have

$$\begin{aligned}\mathbb{E}_t\|\theta_{t+1}-\theta^*\|^2\le&\left(1-\frac{1}{2}c\eta_t\right)\|\theta_t-\theta^*\|^2+M^2\eta_t^2V(z_t)^2+\frac{2}{c}\eta_te_t^2\\ &+\frac{2}{c}\left(18L^2\epsilon^2\eta_t+2L^2M^2\tau_\epsilon^2\eta_t^3+18\frac{L^2L_d^2K^4M^2}{(1-\rho)^4}\tau_\epsilon^2\eta_t^3\right)(V(z_t)+K)^4\\ &-2\eta_t\mathbb{E}_t\langle g(\theta_t,z_t)-g(\theta_{t+\tau_\epsilon},z_{t+\tau_\epsilon}),\theta_t-\theta^*\rangle.\end{aligned}$$

Define $\eta_{T:T} = 0$. Then,

$$\begin{aligned}\mathbb{E}\|\theta_T - \theta^*\|^2 \leq & \exp\left(-\frac{1}{2}c\eta_{0:T}\right)\|\theta_0 - \theta^*\|^2 + \sum_{t=0}^{T-1}\exp\left(-\frac{1}{2}c\eta_{(t+1):T}\right)M^2\eta_t^2 V(z_t)^2 \\ & + \frac{2}{c}\sum_{t=0}^{T-1}\exp\left(-\frac{1}{2}c\eta_{(t+1):T}\right)(18L^2\epsilon^2\eta_t\mathbb{E}(V(z_t)+K)^4 + \eta_t\mathbb{E}e_t^2) \\ & + \frac{2}{c}\sum_{t=0}^{T-1}\exp\left(-\frac{1}{2}c\eta_{(t+1):T}\right)\left(2L^2M^2\tau_\epsilon^2\eta_t^3 + 18\frac{L^2L_d^2K^4M^2}{(1-\rho)^4}\tau_\epsilon^2\eta_t^3\right)\mathbb{E}(V(z_t)+K)^4 \\ & \underbrace{-2\sum_{t=0}^{T-1}\exp\left(-\frac{1}{2}c\eta_{(t+1):T}\right)\eta_t\mathbb{E}\langle g(\theta_t,z_t) - g(\theta_{t+\tau_\epsilon},z_{t+\tau_\epsilon}),\theta_t-\theta^*\rangle}_{(f)}.\end{aligned}$$

Lastly, we develop a proper bound for (f). We first rearrange the summation as

$$\begin{aligned}& \underbrace{-2\sum_{t=\tau_\epsilon}^{T-1}\left\langle \exp\left(-\frac{1}{2}c\eta_{(t+1):T}\right)\eta_t(\theta_t-\theta^*) - \exp\left(-\frac{1}{2}c\eta_{(t+1-\tau_\epsilon):T}\right)\eta_{t-\tau_\epsilon}(\theta_{t-\tau_\epsilon}-\theta^*), g(\theta_t,z_t)\right\rangle}_{(f1)} \\ & \underbrace{-2\sum_{t=0}^{\tau_\epsilon-1}\exp\left(-\frac{1}{2}c\eta_{(t+1):T}\right)\eta_t\langle g(\theta_t,z_t),\theta_t-\theta^*\rangle}_{(f2)} \\ & \underbrace{+2\sum_{t=T-\tau_\epsilon}^{T-1}\exp\left(-\frac{1}{2}c\eta_{(t+1):T}\right)\eta_t\langle g(\theta_{t+\tau_\epsilon},z_{t+\tau_\epsilon}),\theta_t-\theta^*\rangle}_{(f3)}.\end{aligned}$$

For (f1), we have

$$\begin{aligned}& \left|-\left\langle \exp\left(-\frac{1}{2}c\eta_{(t+1):T}\right)\eta_t(\theta_t-\theta^*) - \exp\left(-\frac{1}{2}c\eta_{(t+1-\tau_\epsilon):T}\right)\eta_{t-\tau_\epsilon}(\theta_{t-\tau_\epsilon}-\theta^*), g(\theta_t,z_t)\right\rangle\right| \\ \leq & \left|\exp\left(-\frac{1}{2}c\eta_{(t+1):T}\right)\eta_t\langle\theta_t-\theta_{t-\tau_\epsilon}, g(\theta_t,z_t)\rangle\right| \\ & + \left|\left\langle\left(\exp\left(-\frac{1}{2}c\eta_{(t+1):T}\right)\eta_t - \exp\left(-\frac{1}{2}c\eta_{(t+1-\tau_\epsilon):T}\right)\eta_{t-\tau_\epsilon}\right)(\theta_{t-\tau_\epsilon}-\theta^*), g(\theta_t,z_t)\right\rangle\right| \\ \leq & M^2\exp\left(-\frac{1}{2}c\eta_{(t+1):T}\right)\eta_t V(z_t)(V\eta)_{t-\tau_\epsilon:t} \\ & + \left(\exp\left(-\frac{1}{2}c\eta_{(t+1):T}\right)\eta_t - \exp\left(-\frac{1}{2}c\eta_{(t+1-\tau_\epsilon):T}\right)\eta_{t-\tau_\epsilon}\right)\|\theta_{t-\tau_\epsilon}-\theta^*\|MV(z_t).\end{aligned}$$

Since,

$$\eta_{t-\tau_\epsilon:t} \leq \frac{\eta_{1:\tau_\epsilon}}{\eta_{\tau_\epsilon}}\eta_t \leq \tau_\epsilon\log\tau_\epsilon\eta_t,$$

then,

$$\begin{aligned}& M^2\exp\left(-\frac{1}{2}c\eta_{(t+1):T}\right)\mathbb{E}\eta_t V(z_t)(V\eta)_{t-\tau_\epsilon:t} \\ \leq & M^2\exp\left(-\frac{1}{2}c\eta_{(t+1):T}\right)C(V(z_0)^2+K)\tau_\epsilon\log\tau_\epsilon\eta_t^2.\end{aligned}$$

Next, note that

$$
\begin{aligned}
&\sum_{t=\tau_\epsilon}^{T-1}\left(\exp\left(-\frac{1}{2}c\eta_{(t+1):T}\right)\eta_t-\exp\left(-\frac{1}{2}c\eta_{(t+1-\tau_\epsilon):T}\right)\eta_{t-\tau_\epsilon}\right)\mathbb{E}\|\theta_{t-\tau_\epsilon}-\theta^*\|MV(z_t)\\
\leq& M\sum_{t=\tau_\epsilon}^{T-1}\left(\exp\left(-\frac{1}{2}c\eta_{(t+1):T}\right)\eta_t-\exp\left(-\frac{1}{2}c\eta_{(t+1-\tau_\epsilon):T}\right)\eta_{t-\tau_\epsilon}\right)\sqrt{\mathbb{E}\|\theta_{t-\tau_\epsilon}-\theta^*\|^2}\sqrt{\mathbb{E}V(z_t)^2}\\
\leq& MC(V(z_0)^2+1)\left\{\sum_{t=\tau_\epsilon}^{T-\tau_\epsilon-1}\left(\sqrt{\mathbb{E}\|\theta_{t-\tau_\epsilon}-\theta^*\|^2}-\sqrt{\mathbb{E}\|\theta_t-\theta^*\|^2}\right)\exp\left(-\frac{1}{2}c\eta_{(t+1):T}\right)\eta_t\right.\\
&\left.-\sum_{t=1}^{\tau_\epsilon-1}\exp\left(-\frac{1}{2}c\eta_{(t+1):T}\right)\eta_t\sqrt{\mathbb{E}\|\theta_t-\theta^*\|^2}+\sum_{t=T-\tau_\epsilon}^{T-1}\exp\left(-\frac{1}{2}c\eta_{(t+1):T}\right)\eta_t\sqrt{\mathbb{E}\|\theta_{t-\tau_\epsilon}-\theta^*\|^2}\right\}\\
\leq& MC(V(z_0)^2+1)\left\{\sum_{t=\tau_\epsilon}^{T-\tau_\epsilon-1}\eta_T\sqrt{\mathbb{E}\|\theta_{t-\tau_\epsilon}-\theta_t\|^2}-\sum_{t=1}^{\tau_\epsilon-1}\exp\left(-\frac{1}{2}c\eta_{(t+1):T}\right)\eta_t\sqrt{\mathbb{E}\|\theta_t-\theta^*\|^2}\right.\\
&\left.+\sum_{t=T-\tau_\epsilon}^{T-1}\eta_T\sqrt{\mathbb{E}\|\theta_{t-\tau_\epsilon}-\theta^*\|^2}\right\}\text{ since }\exp\left(-\tfrac{1}{2}c\eta_{t+1:T}\right)\eta_t\leq\eta_T\\
\leq& MC(V(z_0)^2+1)\left\{\eta_T\sum_{t=\tau_\epsilon}^{T-\tau_\epsilon-1}\sqrt{\mathbb{E}(M(V\eta)_{t-\tau_\epsilon:t})^2}-\sum_{t=1}^{\tau_\epsilon-1}\exp\left(-\frac{1}{2}c\eta_{(t+1):T}\right)\eta_t\sqrt{\mathbb{E}\|\theta_t-\theta^*\|^2}\right.\\
&\left.+\eta_T\sum_{t=T-\tau_\epsilon}^{T-1}\sqrt{\mathbb{E}\|\theta_{t-\tau_\epsilon}-\theta^*\|^2}\right\}\\
=& MC(V(z_0)^2+1)\left\{MC\tau_\epsilon\eta_T\eta_{0:T-2\tau_\epsilon}(V(z_0)^2+1)^{1/2}-\sum_{t=1}^{\tau_\epsilon-1}\exp\left(-\frac{1}{2}c\eta_{(t+1):T}\right)\eta_t\sqrt{\mathbb{E}\|\theta_t-\theta^*\|^2}\right.\\
&\left.+\eta_T\sum_{t=T-\tau_\epsilon}^{T-1}\sqrt{\mathbb{E}\|\theta_{t-\tau_\epsilon}-\theta^*\|^2}\right\}.
\end{aligned}
$$

For (f2), we have

$$
\begin{aligned}
&\left|-\sum_{t=0}^{\tau_\epsilon-1}\exp\left(-\frac{1}{2}c\eta_{(t+1):T}\right)\eta_t\mathbb{E}\langle g(\theta_t,z_t),\theta_t-\theta^*\rangle\right|\\
\leq& M\sum_{t=0}^{\tau_\epsilon-1}\exp\left(-\frac{1}{2}c\eta_{(t+1):T}\right)\eta_t\mathbb{E}V(z_t)\|\theta_t-\theta^*\|\\
\leq& M\sum_{t=0}^{\tau_\epsilon-1}\exp\left(-\frac{1}{2}c\eta_{(t+1):T}\right)\eta_t\sqrt{\mathbb{E}V(z_t)^2}\sqrt{\mathbb{E}\|\theta_t-\theta^*\|^2}\\
\leq& MC(V(z_0)^2+1)\sum_{t=1}^{\tau_\epsilon-1}\exp\left(-\frac{1}{2}c\eta_{(t+1):T}\right)\eta_t\sqrt{\mathbb{E}\|\theta_t-\theta^*\|^2}.
\end{aligned}
$$

Similarly, for (f3), we have

$$
\begin{aligned}
&\left|\sum_{t=T-\tau_\epsilon}^{T-1}\exp\left(-\frac{1}{2}c\eta_{(t+1):T}\right)\eta_t\mathbb{E}\langle g(\theta_{t+\tau_\epsilon},z_{t+\tau_\epsilon}),\theta_t-\theta^*\rangle\right|\\
\leq& MC(V(z_0)^2+1)\eta_T\sum_{t=T-\tau_\epsilon}^{T-1}\sqrt{\mathbb{E}\|\theta_t-\theta^*\|^2}.
\end{aligned}
$$

Putting the bounds of (f1) – (f3) together, we have

$$\begin{aligned}
&\mathbb{E}\|\theta_T-\theta^*\|^2\\
\le&\exp\left(-\frac{1}{2}c\eta_{0:T}\right)\|\theta_0-\theta^*\|^2\\
&+\sum_{t=0}^{T-1}\exp\left(-\frac{1}{2}c\eta_{(t+1):T}\right)\left(M^2\eta_t^2\mathbb{E}V(z_t)^2+M^2C(V(z_0)^2+K)\tau_\epsilon\log\tau_\epsilon\eta_t^2\right)\\
&+\frac{2}{c}\sum_{t=0}^{T-1}\exp\left(-\frac{1}{2}c\eta_{(t+1):T}\right)(18L^2\epsilon^2\eta_t\mathbb{E}(V(z_t)+K)^4+\eta_t\mathbb{E}e_t^2)\\
&+\frac{2}{c}\sum_{t=0}^{T-1}\exp\left(-\frac{1}{2}c\eta_{(t+1):T}\right)\left(2L^2M^2\tau_\epsilon^2\eta_t^3+18\frac{L^2L_d^2K^4M^2}{(1-\rho)^4}\tau_\epsilon^2\eta_t^3\right)\mathbb{E}(V(z_t)+K)^4\\
&+2M^2C^2(V(z_0)^2+1)^2\tau_\epsilon\eta_T\eta_{0:T-2\tau_\epsilon}+2M\eta_T\sum_{t=T-2\tau_\epsilon}^{T-1}\sqrt{\mathbb{E}\|\theta_t-\theta^*\|^2}.
\end{aligned}$$

Since $\eta_t=2\eta_0/(ct)$ for some $\eta_0>0$, $\exp\left(-\frac{1}{2}c\eta_{(t+1):T}\right)\le\left(\frac{t}{T}\right)^2$. In addition, since $\mathbb{E}e_t^2=O(1/t)$, we have

$$\begin{aligned}
\mathbb{E}\|\theta_T-\theta^*\|^2\le&\frac{1}{T^2}\|\theta_0-\theta^*\|^2\\
&+\sum_{t=0}^{T-1}\frac{t^2}{T^2}\frac{4\eta_0^2}{c^2t^2}\left(M^2\mathbb{E}V(z_t)^2+M^2C(V(z_0)^2+K)\tau_\epsilon\log\tau_\epsilon\right)\\
&+\frac{2}{c}\sum_{t=0}^{T-1}\frac{t^2}{T^2}\frac{2\eta_0}{ct}(18L^2\epsilon^2\mathbb{E}(V(z_t)+K)^4+\mathbb{E}e_t^2)\\
&+\frac{2}{c}\sum_{t=0}^{T-1}\frac{t^2}{T^2}\frac{8\eta_0^3}{c^3t^3}\left(2L^2M^2\tau_\epsilon^2+18\frac{L^2L_d^2K^4M^2}{(1-\rho)^4}\tau_\epsilon^2\right)\mathbb{E}(V(z_t)+K)^4\\
&+8M^2C^2(V(z_0)^2+1)^2\tau_\epsilon\frac{\eta_0^2}{c^2T}\log T+2M\eta_T\sum_{t=T-2\tau_\epsilon}^{T-1}\sqrt{\mathbb{E}\|\theta_t-\theta^*\|^2}\\
\le&\frac{1}{T^2}\|\theta_0-\theta^*\|^2+C_1\tau_\epsilon\frac{\log T}{T}+C_2\epsilon^2+C_3\tau_\epsilon\frac{\log T}{T}+C_4\frac{1}{T}\sum_{t=T-2\tau_\epsilon}^{T-1}\sqrt{\mathbb{E}\|\theta_t-\theta^*\|^2},\qquad(22)
\end{aligned}$$

where C_1,C_2,C_3,C_4 are some suitably defined constants that do not depend on T.

Next, we prove by induction. Suppose for $t<T$,

$$\mathbb{E}\|\theta_t-\theta^*\|^2\le C\left(\tau_\epsilon\frac{\log t}{t}+\epsilon^2\right).$$

Then, by (22), we have

$$\mathbb{E}\|\theta_T-\theta^*\|^2\le C\left(\tau_\epsilon\frac{\log T}{T}+\epsilon^2\right).$$

Lastly, set $\epsilon=\frac{1}{\sqrt{T}}$. Then, $\tau_\epsilon=O(\tau\log T)$ and

$$\mathbb{E}\|\theta_T-\theta^*\|^2=O\left(\tau\frac{(\log T)^2}{T}\right).$$

□

C. Proofs of the results in Section 3

C.1. Proof of Lemma 1

Our proof is based on the development in Glasserman (1992), who establish sufficient conditions for the derivative process to be well-defined, converge to a unique stationary distribution, and be an unbiased estimator of $\frac{d}{d\theta}\mathbb{E}[D_\infty(\theta)]$ at stationarity.

Let

$$\phi(D_t, u_{t+1}, \epsilon_{t+1}; \theta) = (\alpha \min\{D_t + u_{t+1}, \theta\} + (1-\alpha)m + \epsilon_{t+1})^+.$$

Then, $D_{t+1} = \phi(D_t, u_{t+1}, \epsilon_{t+1}; \theta)$, where u_t's and ϵ_t's are iid respectively, and

$$L_{t+1} = 1\{D_{t+1} > 0\}\alpha \left(1\{\theta < D_t + u_{t+1}\} + 1\{D_t + u_{t+1} \le \theta\} L_t\right).$$

Let $G_\phi \subseteq \mathbb{R}^4$ be defined as the set where ϕ is continuously differentiable. It follows that G_ϕ is the complement $\mathbb{R}^4 \backslash C_\phi$ where C_ϕ is

$$C_\phi = \{a, b, c, \theta : \alpha \min\{a+b, \theta\} + (1-\alpha)m + c = 0\} \cup \{a, b, c, \theta : a + b = \theta\}.$$

Note that for any θ and t, $\mathbb{P}((D_t, u_{t+1}, \epsilon_{t+1}, \theta) \in G_\phi) = 1$ since u_{t+1} and ϵ_{t+1} have densities on $\mathbb{R}$. By Lemma 2.1 of Glasserman (1992), it follows that $D_t(\theta)$ is differentiable with probability 1 for any $\theta \in \Theta$ and $t \ge 0$, and so the derivative process L_t exists.

Moreover, note that ϕ is Lipschitz in a, b, c, θ with Lipschitz constant 1 and for all t,

$$\begin{aligned}\mathbb{E}[D_t] &= \mathbb{E}\left[(\alpha \min\{D_{t-1} + u_{t+1}, \theta\} + (1-\alpha)m + \epsilon_{t+1})^+\right] \\ &\le \mathbb{E}\left[(\alpha\theta + (1-\alpha)m + \epsilon_{t+1})^+\right] < \infty.\end{aligned}$$

By Lemma 2.3 of Glasserman (1992), we have that $\mathbb{E}[D_t(\theta)]$ is differentiable in θ for all t and $\frac{d}{d\theta}\mathbb{E}[D_t(\theta)] = \mathbb{E}[\frac{d}{d\theta}D_t(\theta)] := \mathbb{E}[L_t(\theta)]$.

Finally, it remains to show that $\frac{d}{d\theta}\mathbb{E}[D_\infty(\theta)] = \mathbb{E}[L_\infty(\theta)]$. We first verify the conditions in Theorem 4.1 of Glasserman (1992), which guarantees that L_t has a stationary distribution. First, we have that (ϵ_t, η_t) is a stationary sequence as they are iid. Second, for any $\theta \in \Theta$, the stationary distribution of $D_t(\theta)$ exists (see the proof of Theorem 3) and we can consider a stationary process $\widetilde{D}_t$ where $\widetilde{D}_0$ is drawn from the stationary distribution of $D_t(\theta)$. Note that using the same sequence of (ϵ_t, u_t)'s, the processes D_t and $\widetilde{D}_t$ will couple whenever both are equal to zero, which will happen almost surely, since at every t, the probability of hitting zero is at least $\bar{\Phi}\left(\hat{\theta}/\sigma\right)$, where $\hat{\theta} = \alpha\bar{\theta} + (1-\alpha)m$ and $\bar{\Phi}(x)$ is the tail cumulative distribution function of the standard Normal distribution. Finally, we have that for all $\theta \in \Theta$

$$\mathbb{P}\left(\frac{\partial}{\partial d}\phi(\widetilde{D}_0, u_1, \epsilon_1; \theta) = 0\right) > 0,$$

This is because the derivative $\frac{\partial}{\partial d}\phi(\widetilde{D}_0, u_1, \epsilon_1; \theta)$ will be zero whenever $\alpha \min\{D_0 + u_1, \theta\} + (1-\alpha)m + \epsilon_1 = 0$, which happens with probability at least $\bar{\Phi}\left(\hat{\theta}/\sigma\right)$. By Theorem 4.1 of Glasserman (1992), for any $\theta \in \Theta$, $(D_t(\theta), L_t(\theta))$ converges in distribution to $(D_\infty(\theta), L_\infty(\theta))$. Since $L_\infty(\theta)$ is bounded, it follows that $\mathbb{E}[L_t(\theta)] \to \mathbb{E}[L_\infty(\theta)]$ as $t \to \infty$.

Next, note that ϕ is Lipschitz in a, b, c, θ with Lipschitz constant 1. $\mathbb{E}[D_t(\theta)] < \infty$ and $\mathbb{E}[L_t(\theta)] < \infty$ for any t. Moreover, $\bar{L}_t(\theta) := \frac{1}{n}\sum_{s=1}^{t} L_s(\theta) \le 1$ almost surely, which implies that for $c > 1$

$$\sup_t \int_0^{\bar{\theta}} |\bar{L}_t(\theta)| 1\{\bar{L}_t(\theta) > c\} d\theta = 0.$$

Then, by Theorem 5.1 of Glasserman (1992), we have

$$\mathbb{E}[L_\infty(\theta)] = \frac{d}{d\theta}\mathbb{E}[D_\infty(\theta)].$$

□

C.2. Proof of Theorem 3

We prove Theorem 3 by verifying that the conditions in Thereom 2 hold for the augmented Markov chain $Z_t = (D_t, L_t)$. We use the following for the choice of the Lyapunov function and the metric over $z = (d, l) \in \Omega$:

$$V(d, l) = 1 + 1\{d > 0 \text{ or } l > 0\};$$
$$d(z, z') = 2 \cdot 1\{z \neq z'\}.$$

Since under $d(z, z') = 2 \cdot 1\{z \neq z'\}$, induces the total variation distance, we write $\|\mu - \nu\|_{\mathrm{TV}}$ instead of $W_d(\mu, \nu)$ in this section to make the metric explicit. We first show that V and V^4 are both valid Lyapunov functions, and Z_t satisfies Wasserstein contraction with respect to the metric $d(z, z')$. In particular, we shall verify Assumptions 1 and 4 holds for Z_t.

Lyapunov function. The key observation is that for any $\theta \in \Theta$, $C = \{d = 0, l = 0\}$ is a recurrent aperiodic atom of the Markov chain. This is because it is possible to reach zero demand with probability at least $\bar{\Phi}\left(\hat{\theta}/\sigma\right)$. Then,

$$\begin{aligned}
&\mathbb{E}[V(D_1, L_1] - V(d_0, l_0) \\
=&\mathbb{E}[1 + 1\{D_1 > 0 \text{ or } L_1 > 0\}] - (1 + 1\{d_0 > 0 \text{ or } l_0 > 0\}) \\
\le&\mathbb{E}[1 + 1 - 1\{D_1 = 0\}] - (1 + 1\{d_0 > 0 \text{ or } l_0 > 0\}) \\
\le&-\bar{\Phi}\left(\hat{\theta}/\sigma\right) + 1 - 1\{d_0 > 0 \text{ or } l_0 > 0\}) \\
\le&-\bar{\Phi}\left(\hat{\theta}/\sigma\right)(1 + 1\{d_0 > 0 \text{ or } l_0 > 0\}) + 1 \\
=&-\bar{\Phi}\left(\hat{\theta}/\sigma\right) V(d_0, l_0) + 1.
\end{aligned}$$

Thus, $V(d,l)$ is a Lyapunov function.

Since $V(d,x)^4 = 1 + 15 \cdot 1\{d > 0, x < 1\}$, $V(d,x)^4$ is also a Lyapunov function. We have thus verified Assumption 4.

Wasserstein ergodicity. Let $\nu(\cdot) = \bar{\Phi}\left(\hat{\theta}/\sigma\right)\delta_{(0,0)}(\cdot)$ where $\delta_{(0,0)}$ is the delta measure on $(0,0)$. By construction, $\nu(0,0) > 0$. In addition, we have for any $z \in \Omega$,

$$\mathbb{P}_z(D_t = 0, L_t = 0) \geq \bar{\Phi}\left(\hat{\theta}/\sigma\right).$$

Thus, Z_t satisfies the Doeblin's condition. By Theorem 16.2.4 of Meyn and Tweedie (2012), we have

$$\sup_{z\in\Omega} \|\delta_z P_\theta^n - \pi^\theta\|_{\mathrm{TV}} \leq \left(1 - \bar{\Phi}\left(\hat{\theta}/\sigma\right)\right)^n.$$

We next verify Assumption 2, i.e., the Lipschitz continuity properties.

Lipschitzness of the transition kernel. We would like to show that P_θ is Lipschitz in θ with respect to the total variation distance. To simplify the notation, we write $Z_t = Z_t(\theta)$ and $Z_t' = Z_t(\theta')$. Since

$$\begin{aligned} D_1 &= (\alpha \min\{d_0 + u_1, \theta\} + (1-\alpha)m + \epsilon_1)^+, \\ D_1' &= (\alpha \min\{d_0 + u_1, \theta'\} + (1-\alpha)m + \epsilon_1)^+, \end{aligned}$$

and

$$\begin{aligned} L_1 &= 1\{D_1 > 0\}\alpha\left(1\{d_0 + u_1 > \theta\} + 1\{d_0 + u_1 \leq \theta\} l_0\right), \\ L_1' &= 1\{D_1' > 0\}\alpha\left(1\{d_0 + u_1 > \theta\} + 1\{d_0 + u_1 \leq \theta'\} l_0\right). \end{aligned}$$

we have

$$\begin{aligned} &\sup_{|f|\leq 1} \left|\mathbb{E}\left[f(D_1, L_1) - f(D_1', L_1')\right]\right| \\ (a) =& \sup_{|h|\leq 1} \left|\mathbb{E}h(\epsilon_1 + \alpha\theta, u_1 - \theta) - \mathbb{E}h(\epsilon_1 + \alpha\theta', u_1 - \theta')\right| \\ =& d_{\mathrm{TV}}\left((\epsilon_1 + \alpha\theta, u_1 - \theta), (\epsilon_1 + \alpha\theta', u_1 - \theta')\right) \\ (b) \leq& d_{\mathrm{TV}}\left(\epsilon_1 + \alpha\theta, \epsilon_1 + \alpha\theta'\right) + d_{\mathrm{TV}}\left(u_1 - \theta, \eta_1 - \theta'\right) \\ (c) \leq& \frac{2|\theta - \theta'|}{\sigma}. \end{aligned}$$

For (a), we use the fact that any bounded function f of the update can be rewritten as a bounded function h of $\epsilon_1 + \alpha\theta$ and $\eta_1 - \theta$. For (b), we use the independence of u_1 and ϵ_1. The inequality in (c) follows from a standard upper bound on the total variation distance between two normal distributions.

The above bound implies

$$\sup_{z\in\Omega} \|\delta_z P_\theta - \delta_z P_{\theta'}\|_{\mathrm{TV}} \leq \frac{2}{\sigma}|\theta - \theta'|.$$

Lipschitzness of the gradient. We can show that $\nabla\ell(\theta)$ is Lipschitz in θ as a consequence of Theorem 3.1 of Rudolf and Schweizer (2018), which shows that Lipschitzness of the transition kernel, which we have established above, implies Lipschitzness of the stationary distribution under the total variation distance:

$$\|\mu_\theta - \mu_{\theta'}\|_{\mathrm{TV}} \leq \frac{\alpha/\sigma}{\bar{\Phi}\left(\hat{\theta}/\sigma\right)^2}\|\theta - \theta'\|.$$

$$g(\theta, Z_t) = \left(b1\{D_t > \theta\} - h1\{D_t < \theta\}\right)(1 + L_t).$$

Since $\nabla\ell(\theta) = \mathbb{E}_{\mu_\theta}[(b1\{D_t > \theta\} - h1\{D_t < \theta\})(1 + L_t)]$, $\nabla\ell(\theta)$ is Lipschitz.

In addition, we note that since $|g(\theta, z)| \leq 2\max\{h, b\}$,

$$\|g(\theta, z) - g(\theta, z')\| \leq 4\max\{h, b\}1\{z \neq z'\}.$$

Lastly, we verify Assumption 3.

Bounded objective and gradients For the objective function, we have

$$\ell(\theta) \leq \bar{\theta} + m + 5\sigma.$$

For the gradients, we have

$$\nabla\ell(\theta) \leq 2\max\{h, b\}$$

and

$$|g(\theta, z)| \leq 2\max\{h, b\}$$

for any $\theta \in \Theta$. □

D. Proof of Theorem 4

We prove Theorem 4 by verifying that the conditions in Theorem 2 hold for the augmented Markov chain $Z_t = (W_t, X_t)$.

We first verify Assumptions 1 and 4, which requires establishing the Lyapunov drift conditions and Wasserstein contraction of P_θ for $\theta \in \Theta$. Let

$$V(w, x) = e^{\alpha_1 w + \alpha_2 x}$$

for $0 < \alpha_2 < \alpha_1 < \min\{\alpha^*/\underline{\mu}, \alpha^*/\lambda(\underline{p})\}$, where $\alpha^*, \alpha_1, \alpha_2$ are defined in Assumption 7.

LEMMA 12 **(Lyapunov drift condition)**. *There exists $\rho \in (0, 1)$ such that*

$$\sup_{\theta\in\Theta} P_\theta V(w_0, x_0) \leq \rho V(w_0, x_0) + 1$$

and

$$\sup_{\theta\in\Theta} P_\theta V(w_0, x_0)^4 \leq \rho V(w_0, x_0)^4 + 1.$$

Proof. Let W_0 and X_0 be the current waiting time and server's busy time respectively, and $Y = S/\mu - T/\lambda(p)$ be the increment brought by the next customer.

$$\begin{aligned}
\mathbb{E}[e^{\alpha_1 W_1 + \alpha_2 X_1}] &= \mathbb{E}\left[e^{\alpha_1 (W_0+Y)^+ + \alpha_2 \left(X_0 + \frac{T}{\lambda(p)}\right) 1\{(W_0+Y)^+ > 0\}}\right] \\
&\leq \mathbb{E}\left[e^{\left(\alpha_1 (W_0+Y) + \alpha_2 (X_0 + \frac{T}{\lambda(p)})\right) 1\{(W+Y)^+ > 0\}}\right] + \mathbb{E}[1\{(W+Y)^+ = 0\}] \\
&\leq \mathbb{E}\left[e^{\alpha_1 (W_0+Y) + \alpha_2 (X_0 + \frac{T}{\lambda(p)})}\right] + 1 \\
&= e^{\alpha_1 W_0 + \alpha_2 X_0} \mathbb{E}\left[e^{\alpha_1 \frac{S}{\mu} + (\alpha_2 - \alpha_1) \frac{T}{\lambda(p)}}\right] + 1.
\end{aligned}$$

By Assumption 7, there exists $0 < \alpha_2 < \alpha_1 < \alpha^*$ to be small enough so that

$$\mathbb{E}\left[e^{4\alpha_1 \frac{S}{\underline{\mu}}}\right] \mathbb{E}\left[e^{-4(\alpha_1 - \alpha_2) \frac{T}{\lambda(\underline{p})}}\right] < 1,$$

which is possible because $\lambda(\underline{p}) < \underline{\mu}$. By the convexity of $h(x) = x^4$ and Jensen's inequality, we have:

$$\mathbb{E}\left[e^{\alpha_1 \frac{S}{\underline{\mu}}}\right]^4 \mathbb{E}\left[e^{-(\alpha_1 - \alpha_2) \frac{T}{\lambda(\underline{p})}}\right]^4 \leq \mathbb{E}\left[e^{4\alpha_1 \frac{T}{\underline{\mu}}}\right] \mathbb{E}\left[e^{-4(\alpha_1 - \alpha_2) \frac{S}{\lambda(\underline{p})}}\right] < 1.$$

Let $\rho = \mathbb{E}[e^{\alpha_1 \frac{S}{\underline{\mu}}}] \mathbb{E}[e^{-(\alpha_1 - \alpha_2) \frac{T}{\lambda(\underline{p})}}]$. This condition implies that

$$P_\theta V(w_0, x_0) \leq \rho V(w_0, x_0) + 1.$$

Similarly (by repeating the same arguments), we can show that

$$P_\theta V(w_0, x_0)^4 \leq \rho V(w_0, x_0)^4 + 1.$$

□

The above Lyapunov drift condition (Lemma 12) implies for $\beta = \frac{1}{2}(1 - \rho)$ and a set

$$C = \{z \in \mathbb{R}^2 : V(z) \leq 1/\beta\},$$

we have, by Lemma 15.2.8 of Meyn and Tweedie (2012),

$$P_\theta V(z) - V(z) \leq -\beta V(z) + \mathbf{1}_C(z). \tag{23}$$

Consider the metric

$$d_V(z, z') := (V(z) + V(z')) 1\{z \neq z'\}. \tag{24}$$

In what follows, we write $\|\mu - \nu\|_V$ instead of $W_d(\mu, \nu)$ to make the metric explicit.

LEMMA 13 **(Wasserstein contraction)**. *There exists $K \in (0, \infty)$ and $\rho \in (0, 1)$ such that for all $t \in \mathbb{N}$,*

$$\sup_{\theta\in\Theta}\sup_{z,z'\in\Omega}\frac{\|P_\theta^t(z,\cdot)-P_\theta^t(z',\cdot)\|_V}{d_V(z,z')}\leq K\rho^t$$

Proof. Let $\mathbf{0}:=(0,0)$, which is a recurrent atom for the Markov chain $Z_t=(W_t,X_t)$. Let τ_0 denote the first return time to $\mathbf{0}$, i.e.,

$$\tau_0=\inf\{t\geq 1: Z_t=\mathbf{0}\},$$

and $\pi^{\theta,0}$ denote the probability of $\mathbf{0}$ under the stationary measure of Z_t. We first note the following bound on the generating function of the distance to the stationary distribution, which was developed in Proposition 4.2 of (Baxendale 2005), for any $r>1$,

$$\begin{aligned}\sum_{t=1}^{\infty}r^t\|\delta_z P_\theta^t-\mu_\theta\|_V \leq&\mathbb{E}_z\left[\sum_{t=0}^{\tau_0}V(Z_t)r^t\right]+\mathbb{E}_{\mathbf{0}}\left[\sum_{t=0}^{\tau_0}V(Z_t)r^t\right]\frac{\mathbb{E}_z[r^{\tau_0}]-1}{r-1}\\&+\mathbb{E}_z[r^{\tau_0}]\mathbb{E}_0\left[\sum_{t=0}^{\tau_0}V(Z_t)r^t\right]\left|\sum_{t=1}^{\infty}(P_\theta^t(\mathbf{0},\mathbf{0})-\pi^{\theta,0})r^t\right|\\&+\frac{\mathbb{E}_{\mathbf{0}}\left[\sum_{t=0}^{\tau_0}V(Z_t)r^t\right]+r\mathbb{E}_{\mathbf{0}}\left[\sum_{t=0}^{\tau_0}V(Z_t)\right]}{r-1}.\end{aligned}$$

Since $V(z)\geq 1$, we have $\mathbb{E}_z[r^{\tau_0}]\leq\mathbb{E}_z\left[\sum_{t=0}^{\tau_0}V(Z_t)r^t\right]$. Theorem 15.2.5 of Meyn and Tweedie (2012) shows that the geometric drift condition (23) implies a bound on the generating function of $V(Z_t)$. In particular, for any $r\in(1,(1-\beta)^{-1})$, let $\varepsilon=r^{-1}-(1-\beta)>0$. Then,

$$\begin{aligned}\mathbb{E}_z\left[\sum_{t=0}^{\tau_0}V(Z_t)r^t\right]&\leq(r\varepsilon)^{-1}V(z)+\varepsilon^{-1}\mathbb{E}_z\left[\sum_{t=0}^{\tau_0}\mathbf{1}_C(Z_t)r^t\right]\\&\leq(r\varepsilon)^{-1}V(z)+\frac{r\varepsilon^{-1}}{r-1}\sup_{z\in C}\mathbb{E}_z\left[r^{\tau_0}\right].\end{aligned}$$

In what follows, we first develop a bound for $\sup_{\theta\in\Theta}\sup_{z\in C}\mathbb{E}_z\left[r^{\tau_0}\right]$. We write $\tau_0(\theta)$ and $Y(\theta)=S/\mu-T/\lambda(p)$ to mark dependence of the distribution of τ_0 and $Y(\theta)$ on θ explicitly. Conditional on that $Z_0\in C$, the maximum workload is $W_0=\frac{1}{\alpha_1}\log\frac{1}{\beta}$. Meanwhile, by coupling, for any fixed starting workload w_0 and (μ,p), $\tau_0(\mu,p)\leq_{st}\tau_0(\underline{\mu},\underline{p})$. Above all, we can bound $\tau_0(\theta)$ by the first return time to zero of a random walk starting from $\frac{1}{\alpha_1}\log\frac{1}{\beta}$ and with increments distributed as $Y(\underline{\mu},\underline{p})$. For this random walk, let $\tilde{\tau}_0$ denote its first return time to 0. Then, there exists $\kappa>1$ such that $\mathbb{E}_z\left[\kappa^{\tilde{\tau}_0}\right]<\infty$ for all $z\in C$. Let $\bar{r}=\min\{\kappa,(1-\beta)^{-1}-\varepsilon\}$ for any fixed $\varepsilon>0$. Let $\bar{K}=\mathbb{E}_{\frac{1}{\alpha_1}\log\frac{1}{\beta}}\left[\bar{r}^{\tau_0(\underline{\mu},\underline{p})}\right]$. Then,

$$\sup_{\theta\in\Theta}\sup_{z\in C}\mathbb{E}_z\left[r^{\tau_0}\right]\leq\bar{K}\text{ and }\sup_{\theta\in\Theta}\sup_{z\in C}\mathbb{E}_z\left[\sum_{t=0}^{\tau_0}V(Z_t)\bar{r}^t\right]\leq\bar{K}V(z).\tag{25}$$

We next bound $\left|\sum_{n=1}^{\infty}(P_\theta^n(\mathbf{0},\mathbf{0})-\pi^{\theta,0})r^n\right|$ following the construction in Theorem 3.2 of Baxendale (2005). The sequence $P_\theta^n(\mathbf{0},\mathbf{0})$ can be seen as a renewal process with increment distribution

$b_n := \mathbb{P}_{\mathbf{0}}(\tau_0 = n)$. The key conditions required for bounding $\sum_{n=1}^{\infty} |P_\theta^n(\mathbf{0},\mathbf{0}) - \pi^{\theta,0}| r^n$ uniformly across $\theta \in \Theta$ are some bounds for b_n. First, note that

$$\sum_{n=0}^{\infty} b_n \bar{r}^n = \mathbb{E}_0[\bar{r}^{\tau_0}] \leq \bar{K}.$$

In addition,

$$b_1 = \mathbb{P}_0(\tau_0 = 1) = \mathbb{P}\left(\frac{S}{\mu} - \frac{T}{\lambda(p)} \leq 0\right) \geq \mathbb{P}\left(\frac{S}{\underline{\mu}} - \frac{T}{\lambda(\underline{p})} \leq 0\right) =: \beta.$$

By Theorem 3.2 of Baxendale (2005), the radius of convergence of $\sum_{t=1}^{\infty} |P_\theta^t(\mathbf{0}) - \pi^{\theta,0}| r^t$ is at least $R_1 > 1$, where

$$R_1 := \left\{ r \in (1, \bar{r}) : \frac{r-1}{r(\log \bar{r}/r)^2} = \frac{e^2 \beta}{8 \frac{\bar{K}-1}{\bar{r}-1}} \right\}.$$

Then, for any $r \leq R_1$, there exists K_1 that only depends on $r, \beta, \bar{r}, \bar{K}$ such that

$$\sup_{\theta \in \Theta} \left| \sum_{t=1}^{\infty} |P_\theta^t(\mathbf{0},\mathbf{0}) - \pi^{\theta,0}| r^t \right| \leq K_1. \tag{26}$$

Putting the two bounds (25) and (26) together, taking any $\hat{r} < R_1$, there exists K_2 depending only on $\kappa, \beta, \bar{K}, K_1, \bar{r}, \epsilon$ such that

$$\sup_{\theta \in \Theta} \sum_n \hat{r}^t \|\delta_z P_\theta^t - \mu_\theta\|_V \leq K_2 V(z).$$

This implies that there exists $\rho < \frac{1}{\hat{r}} < 1$ and a constant K such that

$$\sup_{\theta \in \Theta} \|\delta_z P_\theta^t - \mu_\theta\|_V \leq K \rho^t V(z),$$

which further implies that

$$\sup_{\theta \in \Theta} \sup_{z \in \Omega} \frac{\|\delta_z P_\theta^t - \mu_\theta\|_V}{V(z)} \leq K \rho^t.$$

By Lemma 3.2 of Rudolf and Schweizer (2018), we have

$$\sup_{\theta \in \Theta} \sup_{z, z' \in \Omega} \frac{\|\delta_z P_\theta^n - \delta_{z'} P_\theta^n\|_V}{d_V(z, z')} \leq K \rho^t.$$

□

We next verify Assumption 2, i.e., the Lipschitz continuity of the one-step transition kernel and the gradients.

LEMMA 14 **(Lipschitz continuity of the one-step transition kernel)**. *There exists a constant* $\Gamma \in (0, \infty)$ *such that for all* $z \in \mathbb{R}_+^2$

$$d(\delta_z P_\theta^n, \delta_z P_{\theta'}^n) \leq \Gamma \|\theta - \theta'\| V(z).$$

Proof. Let C be a generic constant whose value may change from line to line. By Assumption 7, S and T have $\mathcal{C}^1$ densities. The density of S/μ is $\mu f_S(\mu x)$ and the density of $T/\lambda(p)$ is $\lambda(p) f_T(\lambda(p)x)$. A key property we require is that these densities are sufficiently smooth in μ and $\lambda(p)$ respectively, i.e., Assumption 7 (iii).

Consider a fixed starting state $z_0 = (w_0, x_0)$. Let ϕ_θ be the joint probability density function of $(w_0 + \frac{S}{\mu} - \frac{T}{\lambda(p)}, x_0 + \frac{T}{\lambda(p)})$ for $\theta = (\mu, p)$. By the law of the unconscious statistician, we have

$$\mathbb{E}_{\delta_{z_0} P_\theta}[f(W_1, X_1)] = \int_{\mathbb{R}} \int_{\mathbb{R}} [f(w,x) 1\{w > 0\} + f(0,0) 1\{w \le 0\}] \, \phi_\theta(w,x) dw dx.$$

Then, for any measurable f with $|f| \le V$, we have

$$\begin{aligned}
&|\mathbb{E}_{\delta_{z_0} P_\theta}[f(W_1, X_1)] - \mathbb{E}_{\delta_{z_0} P_{\theta'}}[f(W_1, X_1)]| \\
=&\left| \int_{\mathbb{R}} \int_{\mathbb{R}} f(w,x) 1\{w > 0\} \phi_\theta(w,x) dw dx - \int_{\mathbb{R}} \int_{\mathbb{R}} f(w,x) 1\{w > 0\} \phi_{\theta'}(w,x) dw dx \right| \\
&+ \left| \int_{\mathbb{R}} \int_{\mathbb{R}} f(0,0) 1\{w \le 0\} \phi_\theta(w,x) dw dx - \int_{\mathbb{R}} \int_{\mathbb{R}} f(0,0) 1\{w \le 0\} \phi_{\theta'}(w,x) dw dx \right| \\
\le& \int_{\mathbb{R}} \int_{\mathbb{R}} V(w,x) |\phi_\theta(w,x) - \phi_{\theta'}(w,x)| dw dx \qquad (27) \\
&+ \left| \mathbb{P}\left(w_0 + \frac{S}{\mu} - \frac{T}{\lambda(p)} \le 0 \right) - \mathbb{P}\left(w_0 + \frac{S}{\mu'} - \frac{T}{\lambda(p')} \le 0 \right) \right|. \qquad (28)
\end{aligned}$$

We first bound (27). Note that

$$\begin{aligned}
&\int_{\mathbb{R}} \int_{\mathbb{R}} V(w,x) |\phi_\theta(w,x) - \phi_{\theta'}(w,x)| dw dx \\
=& \int_{\mathbb{R}} \int_{\mathbb{R}} e^{\alpha_1 w + \alpha_2 x} |\phi_\theta(w,x) - \phi_{\theta'}(w,x)| dw dx \\
=& e^{\alpha_1 w_0 + \alpha_2 x_0} \int_{\mathbb{R}} \int_{\mathbb{R}} e^{\alpha_1 w + \alpha_2 x} |\psi_\theta(w,x) - \psi_{\theta'}(w,x)| dw dx,
\end{aligned}$$

where for the last equation, we take a change of variables, and ψ_θ is the joint density of $(\frac{S}{\mu} - \frac{T}{\lambda(p)}, \frac{T}{\lambda(p)})$, i.e.,

$$\psi_\theta(w,x) = \mu f_S(\mu(w+x)) \cdot \lambda(p) f_T(\lambda(p) x)$$

by independence of f_S and f_T. To be more concise, in what follows we denote $\lambda = \lambda(p)$ and $\lambda' = \lambda(p')$.

Next,

$$\begin{aligned}
&\int_0^\infty \int_{\mathbb{R}} e^{\alpha_1 w + \alpha_2 x} |\psi_\theta(w,x) - \psi_{\theta'}(w,x)| dw dx \\
=& \int_0^\infty \int_{-x}^\infty e^{\alpha_1 w + \alpha_2 x} |\mu f_S(\mu(w+x)) \lambda f_T(\lambda x) - \mu' f_S(\mu'(w+x)) \lambda' f_T(\lambda' x)| dw dx \\
\le& \int_0^\infty \int_{-x}^\infty e^{\alpha_1 w + \alpha_2 x} |\mu f_S(\mu(w+x)) - \mu' f_S(\mu'(w+x))| \cdot \lambda f_T(x) dw dx \qquad (29) \\
&+ \int_0^\infty \int_{-x}^\infty e^{\alpha_1 w + \alpha_2 x} |\lambda f_T(\lambda x) - \lambda' f_T(\lambda' x)| \cdot \mu' f_S(\mu'(w+x)) dw dx. \qquad (30)
\end{aligned}$$

Since f_S and f_T are $\mathcal{C}^1$, by the mean-value theorem applied pointwise, we have

$$|\mu f_S(\mu x) - \mu' f_S(\mu' x)| \le L_S(x)|\mu - \mu'|$$
$$|\lambda f_T(\lambda x) - \lambda' f_T(\lambda' x)| \le L_T(x)|\lambda - \lambda'|,$$

where

$$L_S(x) = \sup_{\mu \in [\underline{\mu}, \bar{\mu}]} \left| \frac{d}{d\mu} \mu f_S(\mu x) \right|$$
$$L_T(x) = \sup_{\lambda \in [\lambda(\overline{p}), \lambda(\underline{p})]} \left| \frac{d}{d\lambda} \lambda f_T(\lambda x) \right|.$$

Note that

$$\begin{aligned}
L_S(x) &= \sup_{\mu \in [\underline{\mu}, \bar{\mu}]} |f_S(\mu x) + \mu x f_S'(\mu x)| \\
&\le K 1\{0 \le x \le c\} + \sup_{\mu \in [\underline{\mu}, \bar{\mu}]} \left[\left(1 + x \left| \frac{d}{dx} \log f_S(\mu x) \right| \right) f_S(\mu x) 1\{x \ge c\} \right] \\
&\le K 1\{0 \le x \le c\} + \sup_{\mu \in [\underline{\mu}, \bar{\mu}]} \left[(1 + D_1(\mu x) + D_2(\mu x)^{k+1}) f_S(\mu x) \right] \\
&\le K 1\{0 \le x \le c\} + (1 + D_1(\bar{\mu} x) + D_2(\bar{\mu} x)^{k+1}) \sup_{\mu \in [\underline{\mu}, \bar{\mu}]} f_S(\mu x).
\end{aligned}$$

In the first inequality, we use the Weierstrass theorem to obtain an upper bound for f_S and f_S' over $[0, \bar{\mu} c]$ with the assumption that $f_S \in \mathcal{C}^1$. For the second inequality, we use Assumption 7 (iii) to bound the log-derivative. Then, for (29), we have

$$\begin{aligned}
&\int_0^\infty \int_{-x}^\infty e^{\alpha_1 w + \alpha_2 x} |\mu f_S(\mu(w+x)) - \mu' f_S(\mu'(w+x))| \cdot \lambda f_T(\lambda x) dw dx \\
=&\int_0^\infty \int_0^\infty e^{\alpha_1 (w-x) + \alpha_2 x} |\mu f_S(\mu w) - \mu' f_S(\mu' w))| \cdot \lambda f_T(\lambda x) dw dx \\
\le& C\lambda |\mu - \mu'| \cdot \int_0^\infty \int_0^\infty e^{\alpha_1 w + (\alpha_2 - \alpha_1) x} L_S(w) \lambda f_T(\lambda x) dw dx \\
\le& C\lambda |\mu - \mu'| \cdot \int_0^\infty \int_0^\infty e^{\alpha_1 w} L_S(w) \lambda f_T(\lambda x) dw dx.
\end{aligned}$$

since $\alpha_2 < \alpha_1$. For the above integral to be finite, it is sufficient to show that $L(w)$ is exponentially integrable:

$$\begin{aligned}
&\int_0^\infty e^{\alpha_1 w} L_S(w) dw \\
\le& \int_0^\infty e^{\alpha_1 w} K 1\{0 \le w \le c\} + e^{\alpha_1 w} (1 + D_1(\bar{\mu} w) + D_2(\bar{\mu} w)^{k+1}) \sup_{\mu \in [\underline{\mu}, \bar{\mu}]} f_S(\mu w) dw \\
\le& K c e^{\alpha_1 c} + \int_0^\infty e^{\alpha_1 w} (1 + D_1(\bar{\mu} w) + D_2(\bar{\mu} w)^{k+1}) \sup_{\mu \in [\underline{\mu}, \bar{\mu}]} f_S(\mu w) dw < \infty.
\end{aligned}$$

Thus, there exists $C' \in (0, \infty)$, such that

$$\int_0^\infty \int_{-x}^\infty e^{\alpha_1 w + \alpha_2 x} |\mu f_S(\mu(w+x)) - \mu' f_S(\mu'(w+x))| \cdot \lambda f_T(\lambda x) dw dx \leq C' |\mu - \mu'|.$$

Following similar arguments, we can bound (30) as

$$\begin{aligned}
&\int_0^\infty \int_{-x}^\infty e^{\alpha_1 w + \alpha_2 x} |\lambda f_T(\lambda x) - \lambda' f_T(\lambda' x)| \cdot \mu' f_S(\mu'(w+x)) dw dx \\
\leq & C|\lambda - \lambda'| \int_0^\infty \int_0^\infty e^{\alpha_1 w + (\alpha_2 - \alpha_1) x} L_T(x) \cdot \mu' f_S(\mu' w) dw dx \\
\leq & C|\lambda - \lambda'| \int_0^\infty e^{\alpha_1 w} \cdot \mu' f_S(\mu' w) \int_0^\infty L_T(x) dx dw \\
\leq & C|\lambda - \lambda'| \left(\int_0^\infty e^{\alpha_1 w} \mu' f_S(\mu' w) dw \right) \\
& \cdot \left[Kc + \int_0^\infty (1 + D_1(\bar{\lambda} x) + D_2(\bar{\lambda} x)^{k+1}) \sup_{\lambda \in [\underline{\lambda}, \bar{\lambda}]} f_T(\lambda x) dx \right] \\
\leq & C' |\lambda - \lambda'|,
\end{aligned}$$

where $C' \in (0, \infty)$ since $\lambda f_T(\lambda x)$ and $\mu' f_S(\mu' w)$ have finite moment-generating functions.

We next bound (28). Let h_θ denote the density of $\frac{S}{\mu} - \frac{T}{\lambda}$. h_θ is Lipschitz since it is a convolution of Lipschitz densities:

$$h_\theta(x) = \int_0^\infty \mu f_S(\mu(x+t)) \lambda f_T(\lambda t) dt.$$

Then,

$$\begin{aligned}
&\left| \mathbb{P}\left(\frac{S}{\mu} - \frac{T}{\lambda} \leq -w_0 \right) - \mathbb{P}\left(\frac{S}{\mu'} - \frac{T}{\lambda'} \leq -w_0 \right) \right| \\
\leq & \int_{-\infty}^{-w_0} |h_\theta(x) - h_\theta(x)| dx \\
= & \int_{-\infty}^{-w_0} \left| \int_0^\infty \mu f_S(\mu(x+t)) \cdot \lambda f_T(\lambda t) dt - \int_0^\infty \mu' f_S(\mu'(x+t)) \cdot \lambda' f_T(\lambda' t) dt \right| dx \\
\leq & \int_{-\infty}^{-w_0} \int_0^\infty |\mu f_S(\mu(x+t)) - \mu' f_S(\mu'(x+t))| \lambda f_T(\lambda t) dt dx \\
& + \int_{-\infty}^{-w_0} \int_0^\infty |f_T(\lambda t) - f_T(\lambda' t)| \mu' f_S(\mu'(x+t)) dt dx \\
\leq & |\mu - \mu'| \int_{-\infty}^{-w_0} \int_0^\infty L_S(x+t) \lambda f_T(\lambda t) dt dx \\
& + |\lambda - \lambda'| \int_{-\infty}^{-w_0} \int_0^\infty L_T(t) \mu' f_S(\mu'(x+t)) dt dx \\
\leq & C(|\mu - \mu'| + |\lambda - \lambda'|).
\end{aligned}$$

using the previous bounds for $L_S(x)$ and $L_T(t)$. $\square$

LEMMA 15 **(Smoothness of the gradients)**. *There exists a constant $L \in (0, \infty)$ such that*

$$\|\nabla \ell(\mu, p) - \nabla \ell(\mu, p)\| \leq L\left(|\mu - \mu'| + |p - p'|\right),$$
$$\|g(\mu, p, z) - g(\mu, p, z')\| \leq L d_V(z, z').$$

Proof. Recall the characterization of ∇l and g in (12) and (13). For the first bound, we first note that under Assumption 6, c'' and λ'' are uniformly bounded. Second, the Lipschitz continuity of $\mathbb{E}[W_\infty(\mu, p)]$ has been established in Lemma 4 of Chen et al. (2023a). We next establish the Lipschitz continuity of $\mathbb{E}[X_\infty(\mu, p)]$. For this, we leverage the result from Rudolf and Schweizer (2018), which shows that the Lipschitzness of the one-step transition kernel and Wasserstein ergodicity imply the Lipschitzness of the stationary distribution. In particular, by Theorem 3.1 of Rudolf and Schweizer (2018),

$$d_V(Z_\infty(\theta), Z_\infty(\theta')) \leq \frac{K\Gamma}{(1-\rho)^2}\|\theta - \theta'\|,$$

where Γ is the Lipschitz constant for the one-step transition kernel:

$$\Gamma = \sup_z \sup_{\theta, \theta'} \frac{d_V(\delta_z P_\theta, \delta P_{\theta'})}{V(z)\|\theta - \theta'\|},$$

which was established in Lemma 14, ρ is the Lyapunov drift in Lemma 12 and the Wasserstein contraction rate in Lemma 13, and K is the constant term in the Wasserstein contraction in Lemma 13. Since $X \leq A e^{\alpha_1 W + \alpha_2 X}$ for some constant $A \in (0, \infty)$ large enough, we have

$$\begin{aligned}|\mathbb{E}[X_\infty(\theta) - X_\infty(\theta')]| &\leq A d_V(Z_\infty(\theta), Z_\infty(\theta')) \\ &\leq A\frac{K\Gamma}{(1-\rho)^2}\|\theta - \theta'\|.\end{aligned}$$

For the second bound, since g is linear in z and there exists L large enough so that

$$\begin{aligned}\|z - z'\|_1 &\leq (|w| + |x| + |w'| + |x'|)1\{z \neq z'\} \\ &\leq L\left(e^{\alpha_1|w| + \alpha_2|x|} + e^{\alpha_1|w'| + \alpha_2|x'|}\right)1\{z \neq z'\},\end{aligned}$$

we have $\|g(\mu, p, z) - g(\mu, p, z')\| \leq L d(z, z')$. □

We next verify Assumption 3, i.e., bounds of the gradients.

LEMMA 16 **(Bounds of the gradients)**. *There exists $M \in (0, \infty)$, such that $\|g(\mu, p, z)\| \leq MV(z)$ and $\|\nabla \ell(\mu, p)\| \leq M$.*

Proof. Since $e^x \leq 1 + x$, under Assumption 6, there exists $M > 0$, such that

$$\begin{aligned}g_p(\mu, p, z) &= -\lambda(p) - p\lambda'(p) + h_0\lambda'(p)\left(w + x + \frac{1}{\mu}\right) \\ &\leq M\exp(\alpha_1 w + \alpha_2 x),\end{aligned}$$

and

$$g_\mu(\mu,p,Z_t) = c'(\mu) - h_0\frac{\lambda'(p)}{\mu}\left(W_t + X_t + \frac{1}{\mu}\right)$$
$$\leq M\exp(\alpha_1 w + \alpha_2 x).$$

Next, since $\mathbb{E}[W_\infty(\mu,p)] < \infty$ and $\mathbb{E}[X_\infty(\mu,p)] < \infty$, under Assumption 6, there exists $M > 0$, such that

$$\frac{\partial}{\partial p}\ell(\mu,p) = -\lambda(p) - p\lambda'(p) + h_0\lambda'(p)\left(\mathbb{E}[W_\infty(\mu,p)] + \mathbb{E}[X_\infty(\mu,p)] + \frac{1}{\mu}\right) \leq M,$$

and

$$\frac{\partial}{\partial \mu}\ell(\mu,p) = c'(\mu) - h_0\frac{\lambda'(p)}{\mu}\left(\mathbb{E}[W_\infty(\mu,p)] + \mathbb{E}[X_\infty(\mu,p)] + \frac{1}{\mu}\right) \leq M.$$

□

Lastly, since we can sample directly from g, no further approximation, $\hat{g}_t$, is needed, i.e., Assumption 5 holds trivially. This concludes the proof of Theorem 3.

E. Proofs of the results in Section 5

E.1. Proofs of Propositions 1 and 2

We first present and prove some auxiliary lemmas.

Let τ_{cov} be the first time at which s_t has visited all the states. Then, we define the cover time (Levin and Peres 2017) as

$$t_{cov} = \max_{s\in\mathcal{S}} \mathbb{E}_x[\tau_{cov}].$$

Note that from Theorem 11.2 in (Levin and Peres 2017)

$$t_{cov} \leq t_{hit}\sum_{k=1}^{|\mathcal{S}|-1}\frac{1}{k}.$$

Similarly, we define $\hat{t}_{hit}$ and $\hat{t}_{cov}$ as the hitting time and cover time of the finite-state Markov chain $(s_t, \hat{a}_t)$, respectively.

LEMMA 17. *Suppose the Markov chain s_t has a finite hitting time t_{hit}. Then, the cover time of the Markov chain $(s_t, \hat{a}_t)$ satisfies*

$$\hat{t}_{cov} \leq (1 + t_{hit})|\mathcal{A}|\sum_{k=1}^{|\mathcal{S}||\mathcal{A}|-1}\frac{1}{k}.$$

Proof. Define $\hat{\kappa}_{s,\hat{a}} = \min\{t \geq 0 : s_t = s, \hat{a}_t = \hat{a}\}$. Consider two arbitrary states $(s,\hat{a})$ and $(s',\hat{a}')$ and we are interested in bounding $\mathbb{E}_{s,\hat{a}}[\hat{\kappa}_{s',\hat{a}'}]$. Let $\zeta_0 = 0$, and $\zeta_k, k \geq 1$ be the sequence of stopping time defined as $\zeta_k = \inf\{t > \zeta_{k-1} : s_t = s'\}$. We also write $\Delta\zeta_k = \zeta_{k+1} - \zeta_k$. Since $\hat{a}_t$ are sampled uniformly at random from $\mathcal{A}$, independent of s_t, we have

$$\hat{\kappa}_{s',\hat{a}'} \overset{d}{=} \sum_{k=1}^{N}\Delta\zeta_k,$$

where N is a Geometric random variable with probability of success $1/|\mathcal{A}|$ and is independent of $\Delta\zeta_k$. In addition, note that $\mathbb{E}_s[\Delta\zeta_1] \le t_{hit}$ and $\mathbb{E}_{s'}[\Delta\zeta_k] \le 1 + t_{hit}$ for $k \ge 2$. Then,

$$\mathbb{E}_{s,\hat{a}}[\hat{\kappa}_{s',\hat{a}'}] \le \mathbb{E}[N](1 + t_{hit}) = (1 + h_{hit})|\mathcal{A}|.$$

Thus, $\hat{t}_{hit} \le (1 + t_{hit})|\mathcal{A}|$. Then, the bound for the cover time follows from Theorem 11.2 in (Levin and Peres 2017). □

LEMMA 18. *For a sequence of stopping times $\{\tau_k\}_{k\ge 0}$ with $N(t) = \sup_k\{\tau_k \le t\}$, if there exists $t_0 > 0$, such that $\mathbb{P}_t(N(t + t_0) > N(t)) \le \frac{1}{2}$, then*

$$\mathbb{P}_{\tau_0}(\tau_K - \tau_0 > 12Kt_0) \le \exp(-K).$$

Proof. Without loss of generality, we assume $\tau_0 = 0$. Let $A_m = 1\{N(mt_0) > N((m-1)t_0)\}$. Then,

$$\mathbb{E}_{(m-1)t_0} A_m = \mathbb{P}_{(m-1)t_0}(N(mt_0) > N((m-1)t_0)) \ge 1/2.$$

Next note that $N(12Kt_0) \ge \sum_{m=1}^{12K} A_m$ and $|A_m - \mathbb{E}_{(m-1)t_0} A_m| < 1$. Then,

$$\begin{aligned}
\mathbb{P}_{\tau_0}(\tau_K - \tau_0 > 12Kt_0) =& \mathbb{P}(N(12Kt_0) < K) \\
\le& \mathbb{P}\left(\sum_{m=1}^{12K} A_m < K\right) \\
\le& \mathbb{P}\left(\sum_{m=1}^{12K} (A_m - \mathbb{E}_{(m-1)t_0} A_m) < K - 12K\frac{1}{2}\right) \\
\le& \exp\left(-\frac{(-5K)^2}{2 \times 12K}\right) \text{ by Azuma's inequality} \\
\le& \exp(-K).
\end{aligned}$$

□

LEMMA 19. *Given two matrices A and B, if $det(A + \alpha B) = 0$ for all α, then there exist nonzero W_0 and W_1 such that that $AW_0 = 0$ and $AW_1 + BW_0 = 0$.*

Proof. Let $A = U\Lambda V^T$ denote the singular value decomposition of A. Suppose the first r entries of Λ are zero. Let $C = U^T BV$. We write

$$\Lambda = \begin{bmatrix} 0 & 0 \\ 0 & \Lambda_4 \end{bmatrix}, \quad C = \begin{bmatrix} C_1 & C_2 \\ C_3 & C_4 \end{bmatrix},$$

where Λ_4 is invertible. Then, there exists α_0 such that for all $\alpha \le \alpha_0$, $[\Lambda_4 + \alpha C_4]^{-1}$ exists and its L_2 norm is smaller than a constant. Moreover,

$$0 = \det(A + \alpha B) = \det(\Lambda + \alpha C) = \det(\Lambda_4 + \alpha C_4)\det(\alpha C_1 - \alpha^2 C_2[\Lambda_4 + \alpha C_4]^{-1} C_3).$$

This implies that

$$\det(C_1 - \alpha C_2[\Lambda_4 + \alpha C_4]^{-1} C_3) = 0,$$

which further implies $\det(C_1) = 0$. Let w_0 be a nonzero r-dimensional vector such that $C_1 w_0 = 0$. Next let $w_1 = -\Lambda_4^{-1} C_3 w_0$ which is a $n-r$ dimensional vector. Next, extend w_0, w_1 to an n-dimension vector $\bar{w}_0 = [w_0^T, 0, \ldots, 0]^T, \bar{w}_1 = [0, \ldots, 0, w_1^T]^T$. Let $W_i = V\bar{w}_i, i = 0$ and 1, then

$$AW_0 = U\Lambda V^T V\bar{w}_0 = 0, \quad AW_1 = U\Lambda V^T V\bar{w}_1 = U[0, (\Lambda_4 w_1)^T]^T,$$

$$BW_0 = UCV^T V\bar{w}_0 = U[0, (C_3 w_0)^T]^T,$$

which leads to our conclusion. □

Proof of Proposition 1. Consider running two coupled Z_t and $\tilde{Z}_t$ under π^θ, where $\tilde{Z}_0 \sim \mu^\theta$, and the transition is coupled so that

$$(\hat{a}_t, a_t, s_{t+1}, s'_{t+1}, a'_{t+1}) = (\tilde{\hat{a}}_t, \tilde{a}_t, \tilde{s}_{t+1}, \tilde{s}'_{t+1}, \tilde{a}'_{t+1})$$

if $s_t = \tilde{s}_t$. First, there exists a stopping time $\tau_0 := \min\{t \geq 0 : s_t = \tilde{s}_t\}$, since s_t is ergodic under π^θ. Note that under the coupling, $(s_t, \hat{a}_t, a_t) = (\tilde{s}_t, \tilde{\hat{a}}_t, \tilde{a}_t)$ for $t \geq \tau_0$. In addition,

$$\mathbb{P}\left(\tau_0 \geq \frac{|\log(\epsilon/(8M+4))|}{|\log(1/4)|} t_{mix}\right) \leq \frac{\epsilon}{4(2M+1)}.$$

Let $\Delta_t := Q_t - \tilde{Q}_t$. Then, for $t \geq \tau_0$,

$$\Delta_{t+1}(s_t, \hat{a}_t) = (1-\alpha)\Delta_t(s_t, \hat{a}_t) + \gamma\alpha\Delta_t(s'_{t+1}, a'_{t+1}).$$

Consider a sequence of covering times, $\{\tau_k\}_{k\geq 1}$, where τ_k is the time it takes $(s_t, \hat{a}_t)$ to visit all the state-action pairs at least k times:

$$\tau_k = \min\left\{t > \tau_{k-1} \text{ s.t. for any } (s,a) \in \mathcal{S}\times\mathcal{A} \text{ there is a } u \in (\tau_{k-1}, t] \text{ s.t. } s_u = s, \hat{a}_u = a\right\}.$$

We next use induction to show that

$$\|\Delta_t\|_\infty \leq (1 - (1-\gamma)\alpha)^{k-1} \|\Delta_{\tau_0}\|_\infty, \quad \forall \tau_k < t \leq \tau_{k+1}.$$

The claim holds trivially when $k = 0$. Suppose it is true for τ_k, then for $\tau_k < t < \tau_{k+1}$, we first note that if $(s,a) \neq (s_t, \hat{a}_t)$, $\Delta_{t+1}(s,a) = \Delta_t(s,a)$. If $(s,a) = (s_t, \hat{a}_t)$,

$$\Delta_{t+1}(s,a) \leq (1 - \alpha + \alpha\gamma)\|\Delta_t\|_\infty.$$

This indicates that $\|\Delta_{t+1}\|_\infty \leq \|\Delta_t\|_\infty$. Thus,

$$\Delta_{t+1}(s,a) \leq (1-\alpha+\alpha\gamma)\|\Delta_t\|_\infty \leq (1-\alpha+\alpha\gamma)\|\Delta_{\tau_k}\|_\infty.$$

This further indicates that when all state-action pairs are visited at least once after τ_k, We have

$$\|\Delta_{\tau_{k+1}}\|_\infty \le (1-\alpha+\alpha\gamma)\|\Delta_{\tau_k}\|_\infty.$$

Next, let $N(t) = \sup_k\{\tau_k \le t\}$. Recall that $\hat{t}_{cov}$ is the cover time of the Markov chain $(s_t, \hat{a}_t)$. Since $\mathbb{E}_t[\tau_{N(t)+1} - t] \le \hat{t}_{cov}$,

$$\hat{t}_{cov} \ge \sum_{k>2\hat{t}_{cov}} k\mathbb{P}_t(\tau_{N(t)+1} - t = k) \ge 2\hat{t}_{cov}\mathbb{P}_t(\tau_{N(t)+1} - t > 2\hat{t}_{cov}),$$

which implies that

$$\mathbb{P}_t(\tau_{N(t)+1} - t > 2\hat{t}_{cov}) \le \frac{1}{2} \quad \text{and} \quad \mathbb{P}_t(N(t+2\hat{t}_{cov}) > N(t)) \ge \frac{1}{2}.$$

Then, by Lemma 18, $\mathbb{P}((\tau_K - \tau_0) \ge 12K\hat{t}_{cov}) \le e^{-K}$. Since $\tilde{d}(Z, \tilde{Z}) \le 1 + 2M$ and $\|\Delta_{\tau_0}\|_\infty \le 2M$, for any

$$\bar{\eta}_\epsilon \ge \frac{|\log(\epsilon/(8M+4))|}{|\log(1/4)|} t_{mix} + \max\left\{\frac{|\log(\epsilon/(4M))|}{|\log(1-\alpha+\alpha\gamma)|}, 12|\log(\epsilon/(8M+4))|\right\}\hat{t}_{cov},$$

we have

$$\begin{aligned} d(Z_{\bar{\eta}_\epsilon}, \tilde{Z}_{\bar{\eta}_\epsilon}) \le & (2M+1)\mathbb{P}\left(\tau_0 > \frac{|\log(\epsilon/(8M+4))|}{|\log(1/4)|} t_{mix}\right) \\ & + (2M+1)\mathbb{P}(\tau_K - \tau_0 \ge 12|\log(\epsilon/(8M+4))|\hat{t}_{cov}) \\ & + 2M(1-(1-\gamma)\alpha)^{\frac{|\log(\epsilon/(4M))|}{|\log(1-\alpha+\alpha\gamma)|}} \\ \le & (2M+1)\frac{\epsilon}{4(2M+1)} + (2M+1)\frac{\epsilon}{4(2M+1)} + \frac{\epsilon}{4M}2M \le \epsilon. \end{aligned}$$

Lastly, by Lemma 17, $\hat{t}_{cov} \le (1+t_{hit})|\mathcal{A}|\sum_{k=1}^{|\mathcal{S}||\mathcal{A}|-1} 1/k$. $\square$

Proof of Proposition 2. The first claim is due to the fact that $x_t = (s_t, a_t)$ is a Markov chain with invariant measure ν^θ.

For the second claim, to simplify the notation, we denote x, y as two state-action pairs. Let

$$W(y,x) := \frac{1}{\nu^\theta(x)}\sum_Q \mu^\theta(x,Q)Q(y), \tag{31}$$

which can be seen as an $|\mathcal{S}||\mathcal{A}|$-dimensional vector.

For fixed x and y, we consider a test function $G(Q,s,a) = Q(y)1\{(s,a) = x\}$. Under the invariant measure, we should have

$$\begin{aligned} \mathbb{E}_{z_t\sim\mu^\theta}[G(Q_{t+1}, s_{t+1}, a_{t+1})] &= \mathbb{E}_{z_t\sim\mu^\theta}[G(Q_t, s_t, a_t)] \\ &= \sum_Q \mu^\theta(x,Q)Q(y) = \nu^\theta(x)W(y,x). \end{aligned} \tag{32}$$

For clarity with notation, let $y = (s_y, a_y)$ and let $\tilde{x} = (\tilde{s}, \tilde{a})$. Let $\hat{a}$ denote the random action sampled for the TD learning step (sampled uniformly at random from $\mathcal{A}$. We can expand the one-step expectation as

$$\begin{aligned}
&\mathbb{E}_{z_t \sim \mu^\theta}[G(Q_{t+1}, s_{t+1}, a_{t+1})] \\
=&\sum_{\tilde{x}, Q} \mu^\theta(x_t = \tilde{x}, Q_t = Q)\Big[(1\{\tilde{s} \neq s_y\} + 1\{\tilde{s} = s_y\}\mathbb{P}(\hat{a} \neq a_y))\mathcal{P}^\theta(\tilde{x}, x)Q(y) \\
&+ 1\{\tilde{s} = s_y\}\mathbb{P}(\hat{a} = a_y)\mathcal{P}^\theta(y, x)((1-\alpha)Q(y) + \alpha r(y) + \alpha\gamma\mathbb{E}^\theta[Q(s'_{t+1}, a'_{t+1})])\Big] \\
=&\sum_{\tilde{x}, Q} \mu^\theta(\tilde{x}, Q)\mathcal{P}^\theta(\tilde{x}, x)Q(y) \\
&- \alpha\frac{1}{|A|}\sum_{\tilde{x}:\tilde{s}=s_y}\sum_{Q} \mu^\theta(\tilde{x}, Q)\mathcal{P}^\theta(\tilde{x}, x)\Big[Q(y) - r(y) - \gamma\sum_{x'}\mathcal{P}^\theta(y, x')Q(x')\Big] \\
=&\sum_{\tilde{x}} \nu^\theta(\tilde{x})W(y, \tilde{x})\mathcal{P}^\theta(\tilde{x}, x) \\
&- \alpha\frac{1}{|A|}\sum_{\tilde{x}:\tilde{s}=s_y} \nu^\theta(\tilde{x})\mathcal{P}^\theta(\tilde{x}, x)\Big[W(y, \tilde{x}) - r(y) - \gamma\sum_{x'}\mathcal{P}^\theta(y, x')W(x', \tilde{x})\Big].
\end{aligned} \tag{33}$$

Putting (32) and (33) together, we have

$$\begin{aligned}
&\nu^\theta(x)W(y, x) \\
=&\sum_{\tilde{x}} \nu^\theta(\tilde{x})W(y, \tilde{x})\mathcal{P}^\theta(\tilde{x}, x) \\
&- \alpha\frac{1}{|A|}\sum_{\tilde{x}:\tilde{s}=s_y} \nu^\theta(\tilde{x})\mathcal{P}^\theta(\tilde{x}, x)\Big[W(y, \tilde{x}) - r(y) - \gamma\sum_{x'}\mathcal{P}^\theta(y, x')W(x', \tilde{x})\Big].
\end{aligned} \tag{34}$$

By enumerating all possible x, y combinations, we have a system of $|\mathcal{S} \times \mathcal{A}|^2$ linear equations. Then, it suffices to show this linear equation system has a unique solution W and verify that $W(y, x) = Q^\theta(y)$ is a solution.

To show the solution is unique, we note that the system of linear equations can be written as

$$AW - \alpha BW = \alpha b,$$

where A, B are $|\mathcal{S} \times \mathcal{A}|^2 \times |\mathcal{S} \times \mathcal{A}|^2$ matrices, and b is an $|\mathcal{S} \times \mathcal{A}|^2$-dimensional vector. For such a system to have multiple solutions, $f(\alpha) := \det(A - \alpha B)$ has to be zero. Note that since $A - \alpha B$ is linear in α, $f(\alpha)$ is a polynomial function of α. This implies that $f(\alpha)$ either has finitely many roots or $f(\alpha) \equiv 0$. If $f(\alpha) \equiv 0$, then there exist nonzero W_0 and W_1 such that $AW_0 = 0$ and $AW_1 + BW_0 = 0$ (see Lemma 19). $AW_0 = 0$ implies that for all x, y,

$$\nu^\theta(x)W_0(y, x) - \sum_{\tilde{x}} \nu^\theta(\tilde{x})W_0(y, \tilde{x})\mathcal{P}^\theta(\tilde{x}, x) = 0,$$

which indicates that $W_0(y,x) = W_0(y,\tilde{x})$ for all $\tilde{x}$. We write $W_0(y) := W_0(y,x)$. We plug this into the second equation $AW_1 + BW_0 = 0$, and obtain for all x, y

$$\begin{aligned}
&\nu^\theta(x) W_1(y,x) - \sum_{\tilde{x}} \nu^\theta(\tilde{x}) W_1(y,\tilde{x}) \mathcal{P}^\theta(\tilde{x},x) \\
&+ \frac{1}{|A|} \sum_{\tilde{x}:\tilde{s}=s_y} \nu^\theta(\tilde{x}) \mathcal{P}^\theta(\tilde{x},x) \Big[W_0(y) - \gamma \sum_{x'} \mathcal{P}^\theta(y,x') W_0(x') \Big] = 0.
\end{aligned}$$

For each fixed y, we sum the equations above for all x and obtain

$$\frac{1}{|A|} \nu_s^\theta(s_y) [W_0(y) - \gamma \sum_{x'} \mathcal{P}^\theta(y,x') W_0(x')] = 0.$$

This gives

$$W_0(y) = \gamma \sum_{x'} \mathcal{P}^\theta(y,x') W_0(x').$$

Let $y = \arg\max_x W_0(x)$, then we have $W_0(y) \equiv 0$, which is a contradiction.

We next verify that $W(y,x) = Q^\theta(y)$ is a solution. Since

$$\nu^\theta(x) = \sum_{\tilde{x}} \nu^\theta(\tilde{x}) \mathcal{P}^\theta(\tilde{x},x) \text{ and } Q^\theta(y) = r(y) + \gamma \sum_{x'} \mathcal{P}^\theta(y,x') Q^\theta(x'),$$

we have

$$\begin{aligned}
\nu^\theta(x) Q^\theta(y) = &\sum_{\tilde{x}} \nu^\theta(\tilde{x}) Q^\theta(y) \mathcal{P}^\theta(\tilde{x},x) \\
&- \alpha \frac{1}{|A|} \sum_{\tilde{x}:\tilde{s}=s_y} \nu^\theta(\tilde{x}) \mathcal{P}^\theta(\tilde{x},x) \Big[Q^\theta(y) - r(y) - \gamma \sum_{x'} \mathcal{P}^\theta(y,x') Q^\theta(x') \Big].
\end{aligned} \tag{35}$$

Lastly, note that

$$\begin{aligned}
\mathbb{E}_{\mu^\theta}[Q(s_t,a_t) \nabla_\theta \pi^\theta(a_t|s_t)] &= \sum_{s,a} \nu^\theta(s,a) \mathbb{E}_{\mu^\theta}[Q(s_t,a_t)|s_t = s, a_t = a] \nabla_\theta \pi^\theta(a|s) \\
&= \sum_{s,a} \nu^\theta(s,a) Q^\theta(s,a) \nabla_\theta \pi^\theta(a|s).
\end{aligned}$$

□

E.2. Proof of Theorem 5

We proceed to verify that the assumptions required in Theorem 1 hold for the actor-critic update in Algorithm 1.

First, it is worth noting that under Assumption 8, if Q_0 is initialized to be $\|Q_0\|_\infty \leq \frac{\tilde{M}}{1-\gamma}$, then $\|Q_t\|_\infty \leq \frac{\tilde{M}}{1-\gamma}$ almost surely for all t, under any policy π^θ.

Lyapunov Function. For the Markov chain $Z_t = (s_t, a_t, Q_t)$. we denote $\widetilde{\mathcal{P}}_\theta$ as its transition kernal. Since there are finitely many states and actions and $\|Q_t\|_\infty \le \frac{\tilde{M}}{1-\gamma}$ almost surely, we can construct a Lyapunov function for which the drift inequality holds trivially:

$$V(z) = 1 + \|Q\|_\infty.$$

The corresponding drift inequality is

$$\widetilde{\mathcal{P}}_\theta V(z) \le 1 + \frac{\tilde{M}}{1-\gamma},$$

where $\rho = 0$ and $K = 1 + \frac{\tilde{M}}{1-\gamma}$. We have thus verified the first part of Assumption 1. Similarly, Assumption 4 also holds trivially.

Wasserstein Ergodicity. For $z = (s, a, Q)$, we consider the metric

$$\tilde{d}(z, \tilde{z}) = 1\{(s, a) \neq (\tilde{s}, \tilde{a})\} + \|Q - \tilde{Q}\|_\infty. \tag{36}$$

We next show that for any $\theta \in \Theta$ and any z and $\tilde{z}$,

$$W_{\tilde{d}}(\delta_z \widetilde{\mathcal{P}}_\theta^t, \delta_{\tilde{z}} \widetilde{\mathcal{P}}_\theta^t) \le (8M + 4) r^t$$

where

$$r = \exp\left\{ -\left(\frac{1}{\log(1/\gamma)} + \left(\frac{1}{\log(1/(1-\alpha+\alpha\gamma))} + 12 \right) \frac{(1-\gamma)\min_s \rho(s) + 1}{(1-\gamma)\min_s \rho(s)} |\mathcal{A}|(1 + \log(|\mathcal{S}||\mathcal{A}|)) \right)^{-1} \right\}, \tag{37}$$

which verifies the second part of Assumption 1. Note that (37) implies that

$$\frac{1}{1-r} = O\left(\frac{1}{(1-\gamma)^2} \right).$$

We first bound the mixing time and hitting time of s_t, which we denote as t_{mix} and t_{hit}. Recall that s_t is a finite-state Markov chain with the transition kernel

$$\bar{\mathcal{P}}_\theta(s_{t+1}|s_t) := \gamma \sum_{a_t \in \mathcal{A}} P(s_{t+1}|s_t, a_t) \pi^\theta(a_t|s_t) + (1-\gamma)\rho(s_{t+1}).$$

Note that s_t satisfies a strong version of Doeblin's condition: for any states $s_0, s \in \mathcal{S}$,

$$\bar{\mathcal{P}}_\theta(s|s_0) \ge (1-\gamma)\rho(s).$$

Let $\bar{\nu}_\theta$ denote the stationary distribution of s_t. Since $1 - (1-\gamma)\sum_s \rho(s) = 1 - (1-\gamma) = \gamma$, by Theorem 16.2.4 in Meyn and Tweedie (2012), we have

$$||\delta_s \bar{\mathcal{P}}_\theta^n - \bar{\nu}_\theta||_{\text{TV}} \le \gamma^n,$$

which is uniform across all policy parameters θ. As a result, the mixing time t_{mix} of s_t satifies $t_{\text{mix}} \leq \log(1/4)/\log(1/\gamma)$.

For the hitting time, we have for any $s_0, s \in \mathcal{S}$,

$$\mathbb{E}_{s_0}[\kappa_s] \leq \frac{1}{(1-\gamma)\rho(s)}.$$

Thus, $t_{\text{hit}} \leq 1/((1-\gamma)\min_s \rho(s))$.

Given the bounds for t_{mix} and t_{hit}, by Proposition 1, we have that the ϵ-mixing time of Z_t satisfies

$$\begin{aligned}
\bar{\eta}_\epsilon \leq & \frac{|\log(\epsilon/(8M+4))|}{|\log(1/4)|}\frac{\log(1/4)}{\log(1/\gamma)} + \max\left\{\frac{|\log(\epsilon/(4M))|}{|\log(1-\alpha+\alpha\gamma)|}, 12|\log(\epsilon/(8M+4))|\right\} \\
& \times \left(1 + \frac{1}{(1-\gamma)\min_s \rho(s)}\right)|\mathcal{A}| \sum_{k=1}^{|\mathcal{S}||\mathcal{A}|-1} \frac{1}{k} \\
\leq & \frac{|\log(\epsilon/(8M+4))|}{|\log(1/\gamma)|} + \max\left\{\frac{|\log(\epsilon/(4M))|}{|\log(1-\alpha+\alpha\gamma)|}, 12|\log(\epsilon/(8M+4))|\right\} \\
& \times \frac{(1-\gamma)\min_s \rho(s)+1}{(1-\gamma)\min_s \rho(s)}|\mathcal{A}|(1+\log(|\mathcal{S}||\mathcal{A}|)) \\
\leq & \left(\log\frac{1}{\epsilon} + \log(8M+4)\right) \\
& \times \left(\frac{1}{\log(1/\gamma)} + \left(\frac{1}{\log(1/(1-\alpha+\alpha\gamma))} + 12\right)\frac{(1-\gamma)\min_s \rho(s)+1}{(1-\gamma)\min_s \rho(s)}|\mathcal{A}|(1+\log(|\mathcal{S}||\mathcal{A}|))\right) \\
= & \left(\log\frac{1}{\epsilon} + \log(8M+4)\right)\frac{1}{\log(1/r)}.
\end{aligned}$$

Thus, given t, we can achieve $W_{\tilde{d}}(Z_t, \tilde{Z}_t) \leq \epsilon_t \tilde{d}(z, \tilde{z})$, where $\epsilon_t = (8M+4)r^t$, since $\bar{\eta}_{\epsilon_t} \leq t$.

Lipschitz Gradients and Transition Kernel. To verify Assumption 2, we first show that the transition kernel is Lipschitz according to the metric (36):

$$W_{\tilde{d}}(\delta_z \widetilde{\mathcal{P}}_\theta, \delta_z \widetilde{\mathcal{P}}_{\tilde{\theta}}) \leq (1+\gamma\|Q\|_\infty) R|\mathcal{A}|\|\theta - \tilde{\theta}\|.$$

This implies that $L_d \leq (1+\|Q\|_\infty) R|\mathcal{A}|$, where the bound is independent of γ. Under Assumption 10, the action probabilities are Lipschitz:

$$\begin{aligned}
& |\pi^\theta(a_{t+1}|s_{t+1}) - \pi^{\tilde{\theta}}(a_{t+1}|s_{t+1})| \\
\leq & \left(\sup_{\theta' \in \Theta} \|\nabla_\theta \pi^{\theta'}(a_{t+1}|s_{t+1})\|\right)\|\theta - \tilde{\theta}\| \\
\leq & \left(\sup_{\theta' \in \Theta} \|\pi^{\theta'}(a_{t+1}|s_{t+1})\nabla_\theta \log \pi^{\theta'}(a_{t+1}|s_{t+1})\|\right)\|\theta - \tilde{\theta}\| \\
\leq & R\|\theta - \tilde{\theta}\|.
\end{aligned}$$

Next, note that

$$W_{\tilde{d}}(\delta_z \widetilde{\mathcal{P}}_\theta, \delta_{z'} \widetilde{\mathcal{P}}_\theta) = \|\delta_z \mathcal{P}_\theta - \delta_{z'} \mathcal{P}_\theta\|_{\text{TV}} + \mathbb{E}[\|Q_1 - \tilde{Q}_1\|_\infty],$$

where Q_1 and $\tilde{Q}_1$ are properly coupled. We can bound the total variation term using the Lipschitzness of the policy, i.e.,

$$\begin{aligned}\|\delta_z \mathcal{P}_\theta - \delta_z \mathcal{P}_{\tilde{\theta}}\|_{\mathrm{TV}} =& \frac{1}{2} \sum_{(s_1,a_1)\in\mathcal{S}\times\mathcal{A}} |P(s_1|s_0,a_0)\pi^\theta(a_1|s_1) - P(s_1|s_0,a_0)\pi^{\tilde{\theta}}(a_1|s_1)| \\ \leq& \frac{1}{2}|\mathcal{A}|R\|\theta-\tilde{\theta}\|.\end{aligned}$$

Next, we bound the difference for the Q-function. Note that starting from the same Q-function Q_0, a_0', which is sampled uniformly at random from $\mathcal{A}$, can be coupled and are thus identical under π^θ and $\pi^{\tilde{\theta}}$. In addition, s_1' which is sampled according to $P(\cdot|s_0,a_0')$ can also be coupled and are thus identical under π^θ and $\pi^{\tilde{\theta}}$. Then, the only thing that differs between π^θ and $\pi^{\tilde{\theta}}$ is the next action a_1' in the Q-update, i.e., $a_1' \sim \pi^\theta(\cdot|s_1')$ versus $\tilde{a}_1' \sim \pi^{\tilde{\theta}}(\cdot|s_1')$. Under the coupling described above, we have

$$\begin{aligned}\mathbb{E}[\|Q_1 - \tilde{Q}_1\|_\infty] =& \alpha\gamma\mathbb{E}[|Q_0(s_1',a_1') - Q_0(s_1',\tilde{a}_1')|] \\ \leq& \alpha\gamma\|Q_0\|_\infty \frac{1}{2}|\mathcal{A}|\cdot R\|\theta-\tilde{\theta}\|,\end{aligned}$$

since Q_0 is bounded function of (s,a).

Finally, we show that $\nabla\ell(\theta)$ is also Lipschitz. Recall that μ^θ denotes the stationary distribution of Z_t under π^θ. By Assumption 1, which we have already verified, and the Lipschitzness of $\delta_z\widetilde{P}_\theta$ verified above, we can apply Theorem 3.1 of Rudolf and Schweizer (2018), which gives us

$$W_{\tilde{d}}(\mu^\theta, \mu^{\tilde{\theta}}) \leq \frac{R|\mathcal{A}|(8M+4)}{(1-r)(1-\gamma)}\|\theta-\tilde{\theta}\|,$$

where r is the geometric rate of Wasserstein ergodicity in (37). Since $Q_t(s_t,a_t)\nabla_\theta \log\pi^\theta(a_t|s_t)$ is Lipschitz with respect to the metric $\tilde{d}(z,\tilde{z})$, the above bound for $W_{\tilde{d}}(\mu^\theta,\mu^{\tilde{\theta}})$ implies that $\nabla\ell(\theta)$ is Lipschitz with Lipschitz constant

$$L = \frac{R|\mathcal{A}|(8M+4)}{(1-r)(1-\gamma)} = O\left(\frac{1}{(1-\gamma)^3}\right).$$

Note that this matches the dependence on γ for the Lipschitz constant of $\nabla\ell(\theta)$ in Zhang et al. (2020b).

Bounded Gradients. For $g(\theta,z) = Q(s_t,a_t)\nabla_\theta\log\pi^\theta(a_t|s_t)$, we have

$$\|g(\theta,z)\| \leq R(1+\|Q\|_\infty).$$

In addition,

$$\nabla\ell(\theta) \leq \frac{\tilde{M}}{1-\gamma}.$$

We have thus verified Assumption 3. □

E.3. Proof of Corollary 1

To characterize the dependence on the discount factor γ, we revisit the constant term in bound characterize in Theorem (1), with $\eta_t = \eta_0 t^{-1/2}$. In particular, from the proof of Theorem (1), i.e., the bound in (21), for $\epsilon = 1/\sqrt{T}$, we have

$$\min_{0\le t<T} \mathbb{E}\|\nabla \ell(\theta_t)\|^2 \le \frac{C}{\sqrt{T}} \left\{ \eta_0^{-1}\ell(\theta_0) + \eta_0^{-1} M^2 \tau_\epsilon (V(z_0)+K) + ML \frac{1}{\sqrt{T}} \sum_{t=0}^{T-1} \frac{1}{\sqrt{t}} \mathbb{E}(V(z_t)+K)^2 \right.$$
$$\left. + \left(\frac{M^2 L L_d K^2}{(1-\rho)^2} \tau_\epsilon + M^2 L \tau_\epsilon \right) \eta_0 \sum_{t=0}^{T-1} \frac{1}{t} \mathbb{E}(V(z_t)+K)^2 \right\},$$

where $C < \infty$ is a constant that does not depend on $(1-\gamma)^{-1}$. Note that in bound above, M is the bound of the gradient norm in Assumption 3, L is the bound of the Lipschitz constant in Assumption 2, K and ρ are the ergodicity constants in Assumption 1, and τ_ϵ is the ϵ mixing time. From the analysis in Appendix E.2, we have $L_d = O(1)$,

$$M = O\left(\frac{1}{1-\gamma}\right), \quad K = O\left(\frac{1}{1-\gamma}\right), \quad (V(z)+K) = O\left(\frac{1}{1-\gamma}\right),$$
$$L = O\left(\frac{1}{(1-\gamma)^3}\right), \quad \frac{1}{1-\rho} = \frac{1}{1-r} = O\left(\frac{1}{(1-\gamma)^2}\right), \quad \tau = O\left(\frac{1}{(1-\gamma)^2}\right).$$

Thus,

$$\min_{0\le t<T} \mathbb{E}\|\nabla \ell(\theta_t)\|^2 = O\left(\frac{(\log T)^2}{\sqrt{T}} \left(\eta_0^{-1}(1-\gamma)^{-5} + (1-\gamma)^{-6} + \eta_0 (1-\gamma)^{-15} \right) \right).$$

Next, setting $\eta_0 = (1-\gamma)^5$, we have

$$\min_{0\le t<T} \mathbb{E}\|\nabla \ell(\theta_t)\|^2 = O\left(\frac{(\log T)^2}{\sqrt{T}} \frac{1}{(1-\gamma)^{10}} \right).$$