Stochastic Approximation with Unbounded Markovian Noise: A General-Purpose Theorem

Shaan Ul Haque
Georgia Institute of Technology

Siva Theja Maguluri
Georgia Institute of Technology

Abstract

Motivated by engineering applications such as resource allocation in networks and inventory systems, we consider average-reward Reinforcement Learning with unbounded state space and reward function. Recent work Murthy et al. (2024) studied this problem in the actor-critic framework and established finite sample bounds assuming access to a critic with certain error guarantees. We complement their work by studying Temporal Difference (TD) learning with linear function approximation and establishing finite-time bounds with the optimal sample complexity. These results are obtained using the following general-purpose theorem for non-linear Stochastic Approximation (SA).

Suppose that one constructs a Lyapunov function for a non-linear SA with certain drift condition. Then, our theorem establishes finite-time bounds when this SA is driven by unbounded Markovian noise under suitable conditions. It serves as a black box tool to generalize sample guarantees on SA from i.i.d. or martingale difference case to potentially unbounded Markovian noise. The generality and the mild assumptions of the setup enables broad applicability of our theorem. We illustrate its power by studying two more systems: (i) We improve upon the finite-time bounds of Q-learning in Chen et al. (2024) by tightening the error bounds and also allowing for a larger class of behavior policies. (ii) We establish the first ever finite-time bounds for distributed stochastic optimization of high-dimensional smooth strongly convex function using cyclic block coordinate descent.

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

1 INTRODUCTION

Reinforcement Learning (RL) is an important paradigm in machine learning that provides a powerful framework for learning optimal decision-making strategies in uncertain environments (Sutton, 2018; Szepesvári, 2022). Since its inception, it has been employed in a variety of practical problems such as health care (Dann et al., 2019), robotics (Kober et al., 2013), autonomous vehicles (Aradi, 2020), and stochastic networks (Liu et al., 2019). This remarkable success has led to an extensive study of its convergence behavior both asymptotically (Bertsekas, 1996; Tsitsiklis, 1994; Sutton, 1988) and in finite-time (Beck and Srikant, 2012; Bhandari et al., 2018; Srikant and Ying, 2019; Qu and Wierman, 2020; Chandak et al., 2022; Chen et al., 2024; Zhang and Xie, 2024).

The underlying problem structure in RL is typically modeled by a Markov Decision Process (MDP) (Puterman, 2014) whose transition dynamics are unknown. Several real-world problems such as inventory management systems or queueing models of resource allocation in stochastic networks involve infinite state spaces, and moreover rewards or costs usually go to infinity with the state. Despite these challenges, RL algorithms have shown promising empirical results in these extreme settings (Liu et al., 2019; Cuartas and Aguilar, 2023; Wei et al., 2024; Bharti et al., 2020). In contrast, there is little analytical understanding of their performance in the unbounded setting. In particular, their finite time/sample performance is not well understood. Most of the literature focusing on the finite-time analysis of RL algorithms either assumes finite state space for the underlying MDP (Chen et al., 2024; Khodadadian et al., 2023; Qiu et al., 2021; Chen et al., 2022) or bounded rewards (Wu et al., 2020; Yang et al., 2018; Wang et al., 2017). Furthermore, these assumptions are crucial to their analysis, and thus, their results cannot be easily extended.

One of the widely adopted approaches to find the optimal policy is the actor-critic (AC) framework (Barto et al., 1983). In this method, the actor improves the

current policy by updating it in a direction that maximizes the expected long-term rewards, while the critic evaluates the performance of the policy based on the data samples from the MDP. A recent prior work that analytically studied infinite state MDPs in this context is Murthy et al. (2024), where the authors focus on the actor phase and established finite-time convergence bounds of policy optimization algorithms assuming that the critic evaluates a given policy with certain error guarantees. In this paper, we complement their work by providing finite sample guarantees of such a critic. In particular, we analyze Temporal Difference (TD) learning, a popular algorithm for policy evaluation in critic, and establish finite-time bounds on the mean square error.

The main contributions of the paper are as follows.

Performance of TD-Learning in Unbounded State Space and Rewards: In the policy evaluation problem, the infiniteness of the state space manifests itself through unbounded feature vectors and rewards in the algorithm. We analyze average-reward TD(λ) with linear function approximation (LFA) under asynchronous updates, a popular algorithm for policy evaluation in RL. We establish the first known finite-time convergence bounds for this setting, and show an optimal $\mathcal{O}(1/k)$ convergence rate under appropriate choice of step sizes. Due to the challenges in the average-reward setting, to the best of our knowledge, even the asymptotic convergence has not been formally established in the literature. By a careful projection of the iterates to an appropriate subspace, we also establish its almost-sure (a.s.) convergence.

Finite-Time Convergence Guarantees for SA with Unbounded Markov noise: TD learning is based on using SA to solve the underlying Bellman equation of the MDP. The aforementioned results on TD learning is obtained by studying a general class of non-linear SA corrupted by unbounded Markovian noise, and establishing the following general-purpose result.

Informal Theorem. *Consider a nonlinear SA, and suppose that a Lyapunov function satisfying certain drift condition is constructed in the setting when the noise is i.i.d. or martingale difference. Then, we establish finite sample bounds when the same SA is driven by Markovian noise with unbounded state space under appropriate assumptions.*

In other words, we decouple the challenge of handling Markovian noise from the issue of analyzing the SA itself. Our result complements the existing literature by enabling one to generalize any SA result to the case of unbounded Markovian noise. Therefore, we believe that this powerful result is of independent interest due to its applicability in a wide variety of settings.

Methodological Contribution: The key technique that enables us to establish these results is the use of the solution of the Poisson equation to analyze Markov noise. Recent works control Markov noise by exploiting the geometric mixing properties of Markov chains (Bhandari et al., 2018; Srikant and Ying, 2019; Qu and Wierman, 2020; Mou et al., 2021; Xu and Liang, 2021; Khodadadian et al., 2022; Chen et al., 2024). However, it is unclear if this approach enables one to analyze unbounded Markovian noise. We instead adopt the use of Poisson equation, which has been used to study asymptotic convergence and statistics of SA (Harold et al., 1997; Benveniste et al., 2012; Borkar et al., 2024; Lauand and Meyn, 2024; Allmeier and Gast, 2024). Although this approach has also been recently used to study linear SA under bounded Markovian noise in Kaledin et al. (2020); Haque et al. (2023); Agrawal et al. (2024), we use it to obtain finite sample bounds for nonlinear SA under unbounded noise settings. Compared to the mixing-time approach, this approach is not only more elegant but also has the added advantage of giving tighter bounds (in terms of log factors) and allows for a larger class of Markov chains (such as periodic chains). The next two contributions focus on exploiting these improvements.

Performance of Q-learning Algorithm: As an illustrative application in the case of finite-state Markovian noise, we consider Q-learning in the discounted setting. Using our black box, we immediately obtain finite-sample bounds for Q-learning using the Lyapunov function constructed in Chen et al. (2024). Our result improves Chen et al. (2024) by (i) shaving off additional $\mathcal{O}(\log(k))$, and $\mathcal{O}\left(\log\left(1/1-\gamma\right)\right)$ factors in the convergence bounds and (ii) allowing for a larger class of behavior policies, including those that may not have geometric mixing or lead to periodic behavior.

Performance of Stochastic Cyclic Block Coordinate Descent: The general setup of our theorem also allows us to consider settings beyond RL. We study the stochastic optimization of a high-dimensional smooth strongly convex function, where one is only allowed to update a subset of components at each time. This is commonly done using a variant of stochastic gradient descent called cyclic block coordinate descent (CBCD). While other versions of block coordinate descent were studied in the literature (Nesterov, 2012; Diakonikolas and Orecchia, 2018; Lan, 2020), finite sample bounds of CBCD in the stochastic setting were not known. We provide a new perspective to handle the cyclic nature of updates by viewing each block as the states of a periodic Markov chain. This outlook in conjunction with our black box immediately gives optimal $\mathcal{O}\left(1/k\right)$ convergence rate.

We have provided a detailed comparison with the related literature in Appendix A.

2 PROBLEM SETTING AND MAIN RESULT

Consider a non-linear operator $\bar{F} : \mathbb{R}^d \to \mathbb{R}^d$. Our objective is to find the solution x^* to the following equation:

$$\bar{F}(x) = \mathbb{E}_{Y\sim\mu}[F(x, Y)] = 0, \tag{1}$$

where Y represents random noise sampled from a Markov chain with a unique stationary distribution μ, and F is a general non-linear operator. The state space of the Markov chain is denoted by $\mathcal{Y}$.

Suppose $\bar{F}(\cdot)$ is known, then Eq. (1) can be solved using the simple fixed-point iteration $x_{k+1} = \bar{F}(x_k)$. The convergence of this iteration is guaranteed if one can construct a potential function—also known as Lyapunov function in stochastic approximation theory—that strictly decreases over time. However, when the distribution μ is unknown, and thus $\bar{F}(x)$ is unknown, we consider solving Eq. (1) using the stochastic approximation iteration proposed as follows.

Let $\{Y_k\}_{k\geq 0}$ be a Markov process with stationary distribution μ. Then, the algorithm iteratively updates the estimate x_k by:

$$x_{k+1} = \Pi_{\mathcal{X}}\left(x_k + \alpha_k(F(x_k, Y_k) + M_k)\right), \tag{2}$$

where $\{\alpha_k\}_{k\geq 0}$ is a sequence of step-sizes, $\{M_k\}_{k\geq 0}$ is a random process representing the additive external noise, and $\Pi_{\mathcal{X}}(\cdot)$ is ℓ_2-norm projection of the iterates to set $\mathcal{X}$. The projection on the set $\mathcal{X}$ is included for generality, where $\mathcal{X}$ can be either a compact set or the entire space $\mathbb{R}^d$, depending on the context. We emphasize the importance of projection operator $\Pi_{\mathcal{X}}$ to get meaningful mean square bounds here. In a recent study Borkar et al. (2024), the authors constructed an SA with unbounded noise that operates without any projection step. It is shown that such an algorithm will converge to the stationary point a.s., however, the mean square error diverges (Proposition 4, Section 3.3, (Borkar et al., 2024)). Thus, projecting the iterates to a bounded set is not a proof artifact, but rather a technical necessity.

We begin by outlining the set of assumptions for Algorithm 2. These assumptions are motivated by practical applications of SA algorithms, such as those in RL and optimization algorithms, which will be studied in Sections 3 and 4. Let $\|\cdot\|_c$ be an arbitrary norm in $\mathbb{R}^d$.

Assumption 2.1. There exist functions $A_1(y), B_1(y) : \mathcal{Y} \to [0, \infty)$ such that for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ the operator F satisfies the following:

$$\|F(x, y)\|_c \leq A_1(y)\|x - x^*\|_c + B_1(y).$$

Remark. Prior works such as Srikant and Ying (2019); Mou et al. (2021); Chen et al. (2024) assumed that the state space of the Markov chain is bounded, thus could replace the functions $A_1(y)$ and $B_1(y)$ by their upper bounds. However, in contrast, we consider the case of unbounded state space where these functions can possibly be unbounded as well.

Next, we state the assumptions about the Markov process. Let $P : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ be the transition kernel, and let us denote the one-step expectation of any measurable function G conditioned on $z \in \mathcal{Y}$ as $\mathbb{E}_z[G(Y_1)] = \int_{\mathcal{Y}} G(y)P(z, dy)$. Note that if $\mathcal{Y}$ is countable, then $\mathbb{E}_z[G(Y_1)] = \sum_{j\in\mathcal{Y}} G(j)P(j|z)$.

Assumption 2.2. We assume the following properties on the Markov process:

(a) The Markov process has a unique stationary distribution denoted by μ. Moreover, $\mathbb{E}_{Y\sim\mu}[F(x, Y)]$ exists for all $x \in \mathbb{R}^d$ and is denoted by $\bar{F}(x)$.

(b) There exists a function $V_x(z)$ for all $z \in \mathcal{Y}$ and $x \in \mathbb{R}^d$ which satisfies the Poisson equation:

$$V_x(z) = F(x, z) + \mathbb{E}_z[V_x(Y_1)] - \bar{F}(x). \tag{3}$$

(c) There exist functions $A_2(y), B_2(y) : \mathcal{Y} \to [0, \infty)$ such that for all $x_1, x_2 \in \mathcal{X}$ and $y \in \mathcal{Y}$, the solution V_x satisfies the following:

$$\begin{aligned} \|V_{x_2}(y) - V_{x_1}(y)\|_c &\leq A_2(y)\|x_2 - x_1\|_c, \\ \|V_{x^*}(y)\|_c &\leq B_2(y). \end{aligned} \tag{4}$$

(d) Let $\{Y_k\}_{k\geq 0}$ be a sample path starting with an arbitrary initial state $Y_0 = y_0$, then for all $k \geq 0$ we have the following:

(1) Starting from any initial state y_0, second moment of the functions $A_1(\cdot), A_2(\cdot), B_1(\cdot)$, and $B_2(\cdot)$ are finite and given as follows:

$$\begin{aligned} \max\{\mathbb{E}_{y_0}[A_1^2(Y_k)], 1\} &= \hat{A}_1^2(y_0), \\ \max\{\mathbb{E}_{y_0}[A_2^2(Y_k)], 1\} &= \hat{A}_2^2(y_0), \\ \mathbb{E}_{y_0}[B_1^2(Y_k)] = \hat{B}_1^2(y_0),\ \mathbb{E}_{y_0}[B_2^2(Y_k)] &= \hat{B}_2^2(y_0), \end{aligned}$$

where $\mathbb{E}_{y_0}[\cdot] = \mathbb{E}[\cdot|Y_0 = y_0]$.

(2) If the state space is bounded then we denote the upper bound on these functions as follows:

$$\begin{aligned} \max\{|\max_{y\in\mathcal{Y}} A_1(y)|, 1\} &= A_1, \\ \max\{|\max_{y\in\mathcal{Y}} A_2(y)|, 1\} &= A_2, \\ |\max_{y\in\mathcal{Y}} B_1(y)| = B_1,\ |\max_{y\in\mathcal{Y}} B_2(y)| &= B_2. \end{aligned}$$

Remark. This set of assumptions is inspired by the asymptotic analysis of SA in Benveniste et al. (2012). Implicit in them is the fact that the Markov process exhibits a certain degree of stability. It is important to note that these assumptions are always satisfied for bounded state space Markov chains under fairly general conditions. Additionally, they also hold in many scenarios of practical interest when the state space $\mathcal{Y}$ is unbounded, as will be demonstrated in Section 3.

Remark. There is a parallel line of research that studies the a.s. convergence of SA under even more general setting where the Poisson's equation may not have a solution Yu (2012, 2017, 2018); Liu et al. (2025). This line of work relies on the ergodic theorem for Markov chains to get a handle on the noise, and focuses on asymptotic convergence. For a comprehensive discussion, see detailed literature survey in Appendix A.

Let $\{\mathcal{F}_k\}$ be a set of increasing families of σ-fields, where $\mathcal{F}_k = \sigma\{x_0, Y_0, M_0, \ldots, x_{k-1}, Y_{k-1}, M_{k-1}, Y_k\}$.

Assumption 2.3. Let $A_3, B_3 \geq 0$. Then, process $\{M_k\}_{k\geq 0}$ satisfies the following conditions: (a) $\mathbb{E}[M_k|\mathcal{F}_k] = 0$ for all $k \geq 0$, (b) $\|M_k\|_c \leq A_3\|x_k - x^*\|_c + B_3$.

Remark. Assumption 2.3 implies that $\{M_k\}_{k\geq 0}$ forms a martingale difference sequence with respect to the filtration $\mathcal{F}_k$, and its growth is at most linear with respect to the iterate x_k.

Let $\|\cdot\|_s$ be a norm in $\mathbb{R}^d$. To study the convergence behavior of Eq. 1, we assume the existence of a smooth Lyapunov function with respect to $\|\cdot\|_s$ that has negative drift with respect to the iterates x_k. More concretely, the Lyapunov function satisfies the following assumption.

Assumption 2.4. Given a Lyapunov function $\Phi(x)$, there exists constants $\eta, L_s, l, u > 0$, such that we have

$$\langle \nabla\Phi(x - x^*), \bar{F}(x)\rangle \leq -\eta\Phi(x - x^*), \tag{5}$$

$$\Phi(y) \leq \Phi(x) + \langle \nabla\Phi(x), y - x\rangle + \frac{L_s}{2}\|x - y\|_s^2, \tag{6}$$

$$l\Phi(x) \leq \|x\|_c^2 \leq u\Phi(x), \tag{7}$$

$$\Phi(\Pi_{\mathcal{X}}(x) - x^*) \leq \Phi(x - x^*), \tag{8}$$

where Eq. (5) is the negative drift condition, Eq. (6) is smoothness with respect to $\|\cdot\|_s$, Eq. (7) is equivalence relation with the norm $\|\cdot\|_c$, and Eq. (8) is nonexpansivity of the Lyapunov function $\Phi(x)$. Note that we allow $\|\cdot\|_c$ and $\|\cdot\|_s$ to be two different norms for generality. Often, one can fine-tune the s-norm to get tighter bounds.

Assumption 2.5. Finally, we assume that the stepsize sequence is of the following form:

$$\alpha_k = \frac{\alpha}{(k + K)^\xi},$$

where $\alpha > 0$, $K \geq 2$, and $\xi \in [0, 1]$.

2.1 Unbounded State Space

We will now present finite bounds for the two most popular choices of step size, which are common in practice. Let $\mathcal{X}$ denote an ℓ_2-ball of sufficiently large radius chosen such that $x^* \in \mathcal{X}$. Then, the resulting mean-squared error is as follows.

Theorem 2.1. *Suppose that we run the Markov chain with initial state y_0. When the state space $\mathcal{Y}$ is unbounded and the set $\mathcal{X}$ is an ℓ_2-ball, then under the Assumptions 2.1-2.5, $\{x_k\}_{k\geq 0}$ in the iterations (1) satisfy the following:*

(a) When $\alpha_k \equiv \alpha \leq 1$, then for all $k \geq 0$:

$$\mathbb{E}[\|x_{k+1} - x^*\|_c^2] \leq \varphi_0 \exp(-\eta\alpha k) + 3\varphi_1\hat{C}(y_0)\alpha + \frac{6\varphi_1\hat{C}(y_0)\alpha}{\eta}.$$

(b) When $\xi = 1$, $\alpha > \frac{1}{\eta}$ and $K \geq \max\{\alpha, 2\}$, then for all $k \geq 0$:

$$\mathbb{E}[\|x_{k+1} - x^*\|_c^2] \leq \varphi_0\left(\frac{K}{k + K}\right)^{\eta\alpha} + \frac{\varphi_1\hat{C}(y_0)\alpha}{k + K} + \frac{4(6 + 4\eta)\varphi_1\hat{C}(y_0)e\alpha^2}{(\eta\alpha - 1)(k + K)}.$$

The rate of convergence under other choices of stepsize and the constants $\{\varphi_i\}_i$ and $\hat{C}(y_0)$ are defined in Appendix C.

Remark. In part (a), the error never converges to 0 due to noise variance, however, the expected error of the iterates converges to a ball around x^* at an exponential rate. In part (b), using a decreasing step size with appropriately chosen α, we obtain the $\mathcal{O}(1/k)$ convergence rate, which leads to the sample complexity of $\mathcal{O}(1/\epsilon^2)$ to achieve $\mathbb{E}[\|x_{k+1} - x^*\|_c] \leq \epsilon$.

Remark. Note that $\Pi_{\mathcal{X}}(x) = \arg\min_{x'\in\mathcal{X}} \|x - x'\|_2$. Thus, from a computational standpoint, this only involves rescaling the iterates, as the projection operator $\Pi_{\mathcal{X}}(x)$ reduces to $\frac{\|x\|_2}{\text{radius}(\mathcal{X})}x$ if $\|x\|_2 \geq \text{radius}(\mathcal{X})$ and is x otherwise.

We introduce the projection onto the ball $\mathcal{X}$ for analytical tractability [1]. The interplay of the unbounded state space $\mathcal{Y}$ and the iterate space $\mathbb{R}^d$ makes the analysis significantly challenging. Prior works assume that the set $\mathcal{Y}$ is bounded and thus do not need projection. In contexts like queueing systems, truncating

[1] One way to bypass projection is if one can show by other means that the iterates remain bounded, such as in discounted bounded rewards settings (Chapter 1, (Bertsekas et al., 2011))

the state space would change the stationary distribution, thereby altering the optimal policy. However, projecting the iterates is a more realistic solution in such cases, as it does not change the solution provided that the projecting set is taken to be large enough. Nonetheless, it is worth noting that even after projection, handling the noise is substantially challenging, and no previous work handles this.

2.2 Bounded State Space

Now we state the sample complexity when $\mathcal{Y}$ is bounded and $\mathcal{X} \equiv \mathbb{R}^d$ which implies no projection is required.

Theorem 2.2. *When the state space $\mathcal{Y}$ is bounded and the set $\mathcal{X} \equiv \mathbb{R}^d$, then under the Assumptions 2.1-2.5, $\{x_k\}_{k\geq 0}$ in the iterations (1) satisfy the following:*

(a) When $\alpha_k \equiv \alpha \leq \min\left\{1, \frac{\eta}{A(5+2\eta)\varrho_1}\right\}$, then for all $k \geq 0$:

$$\mathbb{E}[\|x_{k+1} - x^*\|_c^2] \leq \varrho_0 \exp\left(\frac{-\eta\alpha k}{2}\right) + 18B\varrho_1\alpha + \frac{40B\varrho_1\alpha}{\eta}.$$

(b) When $\xi = 1$, $\alpha > \frac{2}{\eta}$ and $K \geq \max\{A\alpha(5\alpha + 8)\varrho_1, 2\}$, then for all $k \geq 0$:

$$\mathbb{E}[\|x_{k+1} - x^*\|_c^2] \leq \varrho_0 \left(\frac{K}{k+K}\right)^{\frac{\eta\alpha}{2}} + \frac{2B\varrho_1\alpha}{k+K} + \frac{8B(5+4\eta)\varrho_1 e\alpha^2}{\left(\frac{\eta\alpha}{2} - 1\right)(k+K)}.$$

Please refer to Appendix C for error bounds under other choices of step-sizes and the constants $\{\varrho_i\}_i$, A, and B.

Remark. Compared to the existing literature with finite state Markov chains, such as Chen et al. (2024), we do not need any mixing property of the Markov chain to establish the bounds. We instead assume that the solution to the Poisson equation exists, which is true in finite state Markov chains even when there is no mixing (such as under periodic behavior). This also has an additional benefit of eliminating the poly-logarithmic factors from the bounds.

3 APPLICATION IN REINFORCEMENT LEARNING: POLICY EVALUATION

In this section, we consider the infinite-horizon average-reward MDP which is specified by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, P)$. Here, $\mathcal{S}$ is the state space which may be countably infinite, $\mathcal{A}$ is the finite action space, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, and $P : \mathcal{S}\times\mathcal{S}\times\mathcal{A} \to [0,1]$ is the transition kernel. At each time step $k = 0, 1, 2, \ldots$, the agent in state $S_k \in \mathcal{S}$ selects an action $A_k \in \mathcal{A}$ sampled from a policy $\pi(\cdot|S_k)$, receives a reward $\mathcal{R}(S_k, A_k)$, and transitions to the next state S_{k+1} sampled from $P(\cdot|S_k, A_k)$.

Consider the problem of evaluating the performance of a policy π from data generated by applying π to the MDP. Let us denote $P_\pi(s'|s) = \sum_{a\in S} P(s'|s,a)\pi(a|s)$ as the transition probabilities, and $\mathcal{R}_\pi(s) = \sum_{a\in\mathcal{A}} \mathcal{R}(s,a)\pi(a|s)$ as the average reward for each state. For clarity, we will drop the subscript π from the notation wherever it is evident. We assume that the policy π has the following property:

Assumption 3.1. The Markov chain generated by policy π is irreducible, aperiodic and has a unique stationary distribution given by μ. Furthermore, the rewards have a finite fourth moment under μ.

$$\mathbb{E}_\mu[\mathcal{R}^4(S_k, A_k)] = \hat{r}^2 < \infty,$$

where $\mathbb{E}_\mu[\cdot]$ denotes the stationary expectation. Let $\bar{r} = \sum_{s\in\mathcal{S}} \mu(s)\mathcal{R}_\pi(s)$. Then, Assumption 3.1 is sufficient for the existence of a differential value function $V^* : \mathcal{S} \to \mathbb{R}$ that satisfies the Bellman/Poisson equation $\mathcal{B}_\pi V^* = V^*$ (Derman and Veinott, 1967). Here $\mathcal{B}$ is defined as:

$$\mathcal{B}_\pi V(s) = \mathcal{R}_\pi(s) + \sum_{s'\in\mathcal{S}} P(s'|s)V(s') - \bar{r},\ \forall s \in \mathcal{S}.$$

It should be noted that the average-reward Bellman equation has a unique solution only up to the additive constant. Thus, if there exists a $c \in \mathbb{R}$ such that $V'(s) = V^*(s) + c$, for all $s \in \mathcal{S}$, then V' is also a solution to the Bellman equation.

Since the state space is infinite, directly estimating V^* from data samples is intractable. Hence, we will use linear function approximation (LFA) to approximate V^*. Denote $\psi(s) = (\psi_1(s), \psi_2(s), \ldots, \psi_d(s))^T \in \mathbb{R}^d$ as the feature vector for state s. Consider an arbitrary indexing of the state space $\mathcal{S} = \{s_1, s_2, s_3, \ldots\}$ such that $P(s_{i+1}|s_i) > 0$ for all $i \geq 1$. For notational ease, we will concatenate the feature vectors $\psi(s_i)$ into a feature matrix Ψ which has d columns and infinite rows such that the i-th row of Ψ corresponds to s_i. Let Λ be a diagonal matrix (infinite dimensional) with diagonal entries as $\mu(s)$. Then, we have the following assumption on Ψ.

Assumption 3.2. (a) The columns of matrix Ψ are linearly independent. More explicitly,

$$\sum_{j=1}^{d} a_j\psi_j(s) = 0,\ \forall s \in \mathcal{S} \implies a_j = 0,\ 1 \leq j \leq d.$$

(b) The columns ψ_j of Ψ satisfy:

$$\|\psi_j\|_\Lambda^2 := \sum_{i=1}^{\infty} \mu(s_i)\psi_j^2(s_i) \leq \hat{\psi}^2 < \infty,\ 1 \leq j \leq d.$$

Remark. The assumption of linear independence holds without loss of generality. If any columns of the matrix are linearly dependent, they can be removed without affecting the approximation accuracy. Part (b) states that the μ-weighted ℓ_2 norm of columns of Ψ are bounded. Note that due to Assumption 3.1, $\mu(s) > 0$ for all $s \in \mathcal{S}$, hence $\|\cdot\|_\Lambda$ is a valid norm.

Assumption 3.3. For any initial state $s_0 \in \mathcal{S}$ and $m \geq 0$, there exist functions $f_1, f_2, f_3 : \mathcal{S} \to [0, \infty)$ and a constant $\rho \in (0, 1)$ satisfying the following:

$$\begin{aligned}
|\mathbb{E}_{s_0}[\mathcal{R}(S_k, A_k)] - \bar{r}| &\leq \rho^k f_1(s_0),\\
\|\mathbb{E}_{s_0}[\psi(S_k)] - \mathbb{E}_\mu[\psi(S_k)]\|_2 &\leq \rho^k f_1(s_0),\\
\|\mathbb{E}_{s_0}[\psi(S_k)\psi(S_{k+m})^T] - \mathbb{E}_\mu[\psi(S_k)\psi(S_{k+m})^T]\|_2 &\\
\leq \rho^k f_1(s_0),&\\
\|\mathbb{E}_{s_0}[\psi(S_k)\mathcal{R}(S_{k+m}, A_{k+m})] - \mathbb{E}_\mu[\psi(S_k)\mathcal{R}_\pi(S_{k+m})]\|_2 &\\
\leq \rho^k f_1(s_0),&
\end{aligned}$$

where for all $k \geq 0$, $f_1(\cdot)$ satisfies: $\mathbb{E}_{s_0}[f_1^4(S_k)] \leq f_2(s_0)$. Furthermore, for all $k \geq 0$, we have

$$\mathbb{E}_{s_0}[\|\psi(S_k)\|_2^4] \leq f_3(s_0);\ \ \mathbb{E}_{s_0}[\mathcal{R}^4(S_k, A_k)] \leq f_3(s_0).$$

Remark. Note that these assumptions are always true for finite state space. For infinite-state space, they quantify the stability of the Markov chain. Overall, while these assumptions are technical, they are fairly mild in many practical applications of interest. For example, in stable queueing systems with downward drift, the stationary distribution is light-tailed and decreases geometrically fast with queue length which forms the state space of the Markov chain. Moreover, one can show rapid convergence of $P^m(\cdot|s_0) \to \mu$ for these Markov chains (Stamoulis and Tsitsiklis, 1990; Meyn and Tweedie, 1994; Lund and Tweedie, 1996). Thus, for rewards and feature vectors with polynomial growth with the queue length, these assumptions are readily satisfied.

Remark. Although, these assumptions imply certain level of mixing in the Markov chain, it remains unclear if one can use the techniques in the existing literature for this setting. This is because these works use the finiteness of the noise to control the growth rate of the iterate by picking small enough step-size. However, this is not possible in the case of unbounded space since the noise can be arbitrarily large.

3.1 Average-Reward TD(λ)

We now define a modified version of the Bellman operator which is essential for understanding TD(λ). For any $\lambda \in [0, 1)$, define $\mathcal{B}_\pi^{(\lambda)}(V) = (1-\lambda)\left(\sum_{m=0}^{\infty} \lambda^m \mathcal{B}_\pi^m(V)\right)$, where $\mathcal{B}_\pi^m(\cdot)$ is the m-step Bellman operator. It is easy to verify that $\mathcal{B}_\pi^{(\lambda)}$ and $\mathcal{B}_\pi$ have the same set of fixed points. Since we are restricting our search for the value function in the subspace spanned by Ψ, we instead solve the corresponding projected Bellman equation:

$$\Psi\theta = \Pi_{\Lambda,\Psi}\left(\mathcal{B}_\pi^{(\lambda)}\Psi\theta\right), \tag{9}$$

where $\theta \in \mathbb{R}^d$ is the parameter variable and $\Pi_{D,\Psi} = \Psi(\Psi^T\Lambda\Psi)^{-1}\Psi^T\Lambda$ ($(\Psi^T\Lambda\Psi)^{-1}$ is a $d \times d$ matrix which well defined due to Assumption 3.2) is the projection operator onto the column space of Ψ with respect to $\|\cdot\|_\Lambda$. Let E_Ψ be a subspace defined as follows:

$$\begin{aligned}
E_\Psi &= \text{span}\{\theta | \psi(s_i)^T\theta = 1, \forall i\}\\
&= \begin{cases} \{c\theta_e | c \in \mathbb{R}\}, & \text{if } \exists \theta_e \in \mathbb{R}^d \text{ and } \psi(s_i)^T\theta_e = 1, \forall i\\ \{0\}, & \text{otherwise} \end{cases}
\end{aligned}$$

Observe that if $E_\Psi \neq \{0\}$, then Eq. 9 has infinitely many solutions of the form $\{\theta^* + c\theta_e | c \in \mathbb{R}\}$, where θ^* is the solution in the orthogonal complement of E_Ψ denoted by $E_\Psi^\perp$. As shown in Part (b) of Proposition E.1, this solution is unique in the subspace $E_\Psi^\perp$. Let $\Pi_{2,E_\Psi^\perp}(\theta) = \theta - \langle\theta, \theta_e\rangle/(\|\theta_e\|_2^2)\theta_e$, which is the projection operator onto the subspace $E_\Psi^\perp$[2] and $\mathcal{X}$ be an ℓ_2-ball in $\mathbb{R}^{d+1}$ such that $[\bar{r}, \theta^{*T}]^T \in \mathcal{X}$. Then, we have Algorithm 1 to estimate $\bar{r}$ and θ^*.

Algorithm 1: Average-reward TD(λ) with LFA

Input : $\lambda \in [0, 1)$, $c_\alpha > 0$, basis functions $\{\psi_i\}_{i=1}^d$, step-size sequence $\{\alpha_k\}_{k\geq 0}$.
Initialize $z_{-1} = 0$, $\bar{r}_0 \in \mathbb{R}$ and $\theta_0 \in E_\Psi^\perp$ arbitrarily.
for $k = 0, 1, \ldots$ **do**
  Observe $(S_k, \mathcal{R}(S_k, A_k), S_{k+1})$
  $\delta_k = \mathcal{R}(S_k, A_k) - \bar{r}_k + \psi(S_{k+1})^T\theta_k - \psi(S_k)^T\theta_k$
  $z_k = \lambda z_{k-1} + \psi(S_k)$
  $\tilde{r}_{k+1} = \bar{r}_k + c_\alpha\alpha_k(\mathcal{R}(S_k, A_k) - \bar{r}_k)$
  $\tilde{\theta}_{k+1} = \theta_k + \alpha_k\delta_k\Pi_{2,E_\Psi^\perp} z_k$
  $[\bar{r}_{k+1}, \theta_{k+1}^T] = \Pi_{\mathcal{X}}([\tilde{r}_{k+1}, \tilde{\theta}_{k+1}^T])$
end

The algorithm is essentially equivalent to TD(λ) for average-reward setting, however one difference lies in the update of the parameter θ_k. Specifically, we project the eligibility trace vector z_k onto $E_\Psi^\perp$ which restricts the iterates $\{\theta_k\}_{k\geq 0}$ to the subspace $E_\Psi^\perp$. This ensures that the convergent point of the algorithm unique. Finally, we use projection onto $\mathcal{X}$ to control the growth of the concatenated vector $[\bar{r}_{k+1}, \theta_{k+1}^T]$.

[2]This should not be confused with $\Pi_{\mathcal{X}}(\cdot)$ which is projection onto the compact set $\mathcal{X}$.

3.1.1 Properties of TD(λ) algorithm

To transform Algorithm 1 in the form of iteration (2), we construct a process $Y_k = (S_k, A_k, S_{k+1}, z_k)$ taking values in the space $\mathcal{Y} := \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \mathbb{R}^d$. It is easy to verify that Y_k forms a Markov process in the continuous unbounded state space[3]. Let us define $x_k := [\bar{r}_k, \theta_k^T]^T$, then the iterations can be compactly written as:

$$x_{k+1} = \Pi_{\mathcal{X}} \left(x_k + \alpha_k F(x_k, Y_k) \right), \tag{10}$$

where $F(x_k, Y_k) = T(Y_k)x_k + b(Y_k)$ and

$$T(Y_k) = \begin{bmatrix} -c_\alpha & 0 \\ -\Pi_{2,E_\Psi^\perp} z_k & \Pi_{2,E_\Psi^\perp} z_k \left(\psi(S_{k+1})^T - \psi(S_k)^T \right) \end{bmatrix};$$

$$b(Y_k) = \begin{bmatrix} c_\alpha \mathcal{R}(S_k, A_k) \\ \Pi_{2,E_\Psi^\perp} \mathcal{R}(S_k, A_k) z_k \end{bmatrix}.$$

To consider the stationary behavior of Y_k, let $\{\tilde{S}_k, A_k\}_{k \geq 0}$ denote the stationary process. Then, $\tilde{z}_k := \sum_{\ell=-\infty}^{k} \lambda^{k-\ell} \psi(\tilde{S}_\ell)$ and $\tilde{Y}_k = (\tilde{S}_k, A_k, \tilde{S}_{k+1}, \tilde{z}_k)$ are the stationary analogs of z_k and Y_k, respectively. Let the stationary expectation of the matrices $T(\tilde{Y}_k)$ and $b(\tilde{Y}_k)$ be denoted by $\bar{T}$ and $\bar{b}$. Then, we have the following lemma for $\bar{T}$ and $\bar{b}$ whose proof is given in Appendix E.1.

Lemma 3.1. *Under Assumption 3.1 and 3.2, the stationary expectations $\bar{T}$ and $\bar{b}$ are finite and given by*

$$\bar{T} = \begin{bmatrix} -c_\alpha & 0 \\ \frac{1}{(\lambda-1)} \Pi_{2,E_\Psi^\perp} \Psi^T \mu & \Pi_{2,E_\Psi^\perp} (\Psi^T \Lambda P^{(\lambda)} \Psi - \Psi^T \Lambda \Psi) \end{bmatrix}$$

$$\bar{b} = \begin{bmatrix} c_\alpha \bar{r} \\ \Pi_{2,E_\Psi^\perp} \Psi^T \Lambda \mathcal{R}^{(\lambda)} \end{bmatrix},$$

where $P^{(\lambda)} = (1-\lambda) \sum_{m=0}^{\infty} \lambda^m P^{m+1}$ and $\mathcal{R}^{(\lambda)} = (1-\lambda) \sum_{m=0}^{\infty} \lambda^m \sum_{l=0}^{m} P^l \mathcal{R}_\pi$.

Using Lemma 3.1 and Assumptions 3.1-3.3 one can verify with some algebra that TD(λ) satisfies all the Assumptions 2.1-2.2. We present the proof of this in Appendix E.2.

3.1.2 Finite Sample Bounds for TD(λ)

We pick $\Phi(x_k - x^*) = (\bar{r}_k - \bar{r})^2 + \|\theta_k - \theta^*\|_2^2$ as our Lyapunov function. A key insight in Zhang et al. (2021) that established the negative drift was the observation that for any function in the set $\{V | \sum_{s \in \mathcal{S}} V(s) = 0, \sum_{s \in \mathcal{S}} V^2(s) = 1\}$, there exists a $\Delta > 0$ such that

$$V^T \Lambda V - V^T \Lambda P^{(\lambda)} V \geq \Delta.$$

However, in general, such an inequality is not true for infinite state space, as explained in Appendix E.5.

To establish a similar drift condition in our setting, we leverage the fact that the matrix Ψ has a finite number of linearly independent columns. This effectively restricts the value function to a finite-dimensional subspace, allowing us to prove the following lemma. For proof, please refer to Appendix E.3.

Lemma 3.2. *Under Assumption 3.1 and 3.2, we have*

$$\Delta := \min_{\substack{\theta \in E_\Psi^\perp \\ \|\theta\|_2 = 1}} \theta^T \Pi_{2,E_\Psi^\perp} \left(\Psi^T \Lambda \Psi - \Psi^T \Lambda P^{(\lambda)} \Psi \right) \theta > 0.$$

Furthermore, when $c_\alpha \geq \Delta + \sqrt{\frac{d^2 \hat{\psi}^4}{\Delta^2 (1-\lambda)^4} - \frac{d\hat{\psi}^2}{(1-\lambda)^2}}$, we have $-x^T \bar{T} x \geq \frac{\Delta}{2} \|x\|_2^2$ for all $x \in \mathbb{R} \times E_\Psi^\perp$.

Since $\Phi(x)$ is ℓ_2-norm squared and has a negative drift, Assumption 2.4 is also verified. With all the assumptions now satisfied, we can apply Theorem 2.1 to get the following sample complexity for TD(λ).

Theorem 3.1. *Consider the iterates $\{\theta_k, \bar{r}_k\}_{k \geq 0}$ generated by Algorithm 1 under Assumption 3.1-3.3 and c_α be chosen as in Lemma 3.2. When $\xi = 1$, $\alpha > \frac{1}{\Delta}$ and $K \geq \max\{\alpha, 2\}$, then for all $k \geq 0$:*

$$\mathbb{E}[(\bar{r}_k - \bar{r})^2 + \|\theta_k - \theta^*\|_2^2] \leq \varphi_{V,0} \left(\frac{K}{k+K} \right)^{\frac{\Delta\alpha}{2}} + \frac{\hat{C}_V(s_0, s_1)\alpha}{k+K} + \frac{4(6 + 2\Delta)\hat{C}_V(s_0, s_1) e \alpha^2}{\left(\frac{\Delta\alpha}{2} - 1\right)(k+K)}.$$

Refer to Appendix E.4 for rate of convergence for constant step-size and the constants $\varphi_{V,0}$ and $\hat{C}_V(s_0, s_1)$.

Remark. It is evident from the above bound that to find a pair (r, θ) such that $\mathbb{E}[|r - \bar{r}|] \leq \epsilon$ and $\mathbb{E}[\|\theta - \theta^*\|] \leq \epsilon$, one needs at most $\mathcal{O}\left(1/\epsilon^2\right)$ number of samples.

Remarkably, one can show the a.s. convergence of Algorithm 1 even without using $\Pi_{\mathcal{X}}(\cdot)$ in the final step. Recall that due to the projection of the iterates on $E_\Psi^\perp$, the fixed point θ^* is unique. Thus, with an additional assumption on Markov chain, we can apply the general result on SA from Benveniste et al. (2012) (Theorem 17, Page 239) to show a.s. convergence.

Theorem 3.2. *Suppose that in addition to Assumptions 3.1-3.3, we have $\mathbb{E}_{s_0}[f_1^q(S_k)] \leq f_1^q(s_0)$ for all $q > 0$. Then, the Algorithm 1 a.s. converges to $(\bar{r}, \theta^*)$.*

Remark. Previous works showed a.s. convergence when $E_\Psi = \{0\}$ since the limit point is unique in this case. However, by utilizing the uniqueness of solution in $E_\Psi^\perp$ and the ease of projection in such a space, we eliminate any such assumptions on Ψ. As highlighted before, one does not need the final projection onto the bounded set for asymptotic convergence.

[3] We would like to highlight the distinction between infinite and unbounded state space here. In the analysis of TD(λ) for finite MDPs, the process Y_k also has infinite state space, but it is *bounded*.

3.2 Q-learning for Discounted-Reward Setting in Finite State Space

As discussed in Section 2.2, our bounds are applicable for finite state Markov chains which do not mix and hence removing the requirement on the behavior policy to induce an aperiodic chain. Due to this flexibility, often one can design more effective behavior policies from a wider class of distributions to balance the trade-off between exploration and exploitation. Due to space constraints, we present the details in Appendix B.

4 APPLICATION IN OPTIMIZATION: CYCLIC BLOCK COORDINATE DESCENT

Consider an optimization problem $\min_{x\in\mathbb{R}^d} f(x)$ where the objective function $f(x)$ is μ-strongly convex and L-smooth. Denote x^* as the unique minimizer of $f(x)$. We assume that any vector x can be partitioned into p blocks as follows:

$$x = (x(1), x(2), \ldots, x(p)),$$

where $x(i) \in \mathbb{R}^{d_i}$ with $d_i \geq 1$ for all $1 \leq i \leq p$ and satisfying $\sum_{i=1}^p d_i = d$. Furthermore, $\nabla_i f(x)$ denotes the partial derivatives with respect to the i-th block. Suppose that we have access to the partial gradients only through a noisy oracle which for any $x \in \mathbb{R}^d$ and block i returns $\nabla_i f(x) + w$. Here w represents the noise with appropriate dimension which satisfies the following assumption.

Assumption 4.1. Let $\mathcal{F}_k$ be the σ-field generated by $\{x_i, w_i\}_{0\leq i\leq k-1} \cup \{x_k\}$. Then, there exists constants $C_1, C_2 \geq 0$ such that for all $k \geq 0$: (a) $\mathbb{E}[w_k|\mathcal{F}_k] = 0$, (b) $\|w_k\|_2 \leq C_1\|x_k - x^*\|_2 + C_2$.

Assumption 4.1 is a standard assumption in optimization and basically implies that w_k is a martingale difference sequence with respect to $\mathcal{F}_k$ and grows linearly with the iterates. Then, we have the following stochastic iterative algorithm to estimate x^*:

Algorithm 2: Stochastic Cyclic Block Coordinate Descent (SCBCD)

Initialize $x_0 \in \mathbb{R}^d$, and step-size $\{\alpha_k\}_{k\geq 0}$.
for $k = 0, 1, \ldots$ **do**
  Set $i(k) = k \bmod p + 1$
  $x_{k+1}(j) =$
  $\begin{cases} x_k(j) + \alpha_k(-\nabla_j f(x_k) + w_k), & \text{if } j = i(k) \\ x_k(j), & \text{otherwise} \end{cases}$
end

Without loss of generality, we assume that at $k = 0$ we update the first block. At each time step k, we cyclically update a block, where the block index $i(k)$ is determined through the modulo function. The oracle provides a noise gradient $-\nabla_j f(x_k)+w_k$ and the block corresponding to $i(k)$ gets updated while the rest of the blocks remain unchanged.

4.1 Properties of SCBCD

To fit Algorithm 2 in the framework of (2), we will set up some notation. Define the matrices $U_i \in \mathbb{R}^{d\times d_i}$, $1 \leq i \leq p$ which satisfy,

$$(U_1, U_2, \ldots, U_p) = I_d.$$

Note that $x(i) = U_i^T x$ for any vector $x \in \mathbb{R}^d$ and similarly the partial derivatives with respect to the i-th block can be written as $\nabla_i f(x_k) = U_i^T \nabla f(x_k)$. Thus, we rewrite the update equation as follows:

$$\begin{aligned} x_{k+1} &= x_k + \alpha_k(-U_{i(k)}\nabla_{i(k)} f(x_k) + U_{i(k)} w_k) \\ &= x_k + \alpha_k(F(x_k, i(k)) + M_k) \end{aligned} \tag{11}$$

where $F(x_k, i(k)) = -U_{i(k)}\nabla_{i(k)} f(x_k)$ and $M_k = U_{i(k)} w_k$. Observe that $\mathcal{M}_U = \{i(k)\}_{k\geq 0}$ can be viewed as a periodic Markov chain defined on the state space $\mathcal{S} = \{1, 2, \ldots, p\}$ with transition probabilities given as $P(i \bmod p + 1|i) = 1, \quad \forall i \in \mathcal{S}$. Furthermore, it is easy to verify that $\mu(i) = \frac{1}{p}$, $\forall i \in \mathcal{S}$ is the unique stationary distribution for this Markov chain. This implies $\mathbb{E}_{i\sim\mu}[U_i \nabla_i f(x)] = \frac{1}{p}\nabla f(x)$. Thus, solving for $\nabla f(x) = 0$ is equivalent to finding the root of $\mathbb{E}_{i\sim\mu}[U_i \nabla_i f(x)] = 0$. It is now easy to verify from Eq. (11) that all the Assumptions 2.1-2.2 are satisfied. For completeness, we provide the proof for the verification in Appendix F.1.

4.2 Finite Sample Bounds for SCBCD

We choose $\Phi(x - x^*) = \frac{1}{2}\|x - x^*\|_2^2$ as our Lyapunov function. This immediately implies the properties of $\Phi(x - x^*)$ in Assumption 2.4. In addition, $\eta = \frac{\mu}{p}$ and $L_s = 1$ by smoothness and strong convexity of $f(x)$. We apply Theorem 2.2 to SCBCD to obtain the following finite-time sample complexity.

Theorem 4.1. *Consider the iterates $\{x_k\}_{k\geq 0}$ generated by Algorithm 2 under Assumption 4.1.*

(a) When $\alpha_k \equiv \alpha \leq \min\left\{1, \frac{\mu}{A_G(5p+2\mu)\varrho_{G,1}}\right\}$, then for all $k \geq 0$:

$$\mathbb{E}[\|x_{k+1} - x^*\|_2^2] \leq \varrho_{G,0} \exp\left(\frac{-\mu\alpha k}{2p}\right) + 18B_G\varrho_{G,1}\alpha + \frac{40pB_G\varrho_{G,1}\alpha}{\mu}.$$

(b) When $\xi = 1$, $\alpha > \frac{2p}{\mu}$ *and* $K \geq \max\{A_G\alpha(5\alpha + 8)\varrho_{G,1}, 2\}$*, then for all* $k \geq 0$*:*

$$\mathbb{E}[\|x_{k+1} - x^*\|_2^2] \leq \varrho_{G,0}\left(\frac{K}{k+K}\right)^{\frac{\mu\alpha}{p}} + \frac{2B_G\varrho_{G,1}\alpha}{k+K} + \frac{16B_G\left(5p+4\mu\right)\varrho_{G,1}e\alpha^2}{(\mu\alpha - 2p)(k+K)}.$$

For the constants $\{\varrho_{G,i}\}_i$ A_G, and B_G refer to Appendix F.2.

Remark. In the noisy case, we obtain the $\mathcal{O}\left(1/k\right)$ rate of convergence similar to the randomized BCD in Lan (2020). Moreover, setting $B_G = 0$ in the noiseless case, one obtains a geometric rate of convergence with a sample complexity of $\mathcal{O}\left((p^2L^3/\mu^2)\log\left(1/\epsilon\right)\right)$. In the most general setting, our bound is optimal with respect to p as shown in Sun and Ye (2021). However, we remark that the dependence on the condition number $\frac{L}{\mu}$ is sub-optimal due to universal framework of our theorem. Nevertheless, one can improve upon the constants by using $f(x) - f(x^*)$ as the Lyapunov function and refining our analysis with the additional structure.

Some interesting future directions are extending our analysis to non-smooth functions, analyzing SCBCD with block dependent step-size, and reducing p dependence in specialized cases.

Acknowledgment. This work was partially supported by NSF Grant EPCN-2144316, CMMI-2140534, and CMMI-2112533. We also thank Zaiwei Chen for pointing out that one can model the cyclic update of the iterates in CBCD as a periodic Markov chain and use our theorem to obtain finite time bounds for stochastic CBCD.

References

Agrawal, S., Maguluri, S. T., et al. (2024). Markov chain variance estimation: A stochastic approximation approach. *arXiv preprint arXiv:2409.05733*.

Allmeier, S. and Gast, N. (2024). Computing the bias of constant-step stochastic approximation with markovian noise.

Aradi, S. (2020). Survey of deep reinforcement learning for motion planning of autonomous vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 23(2):740–759.

Barto, A. G., Sutton, R. S., and Anderson, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE transactions on systems, man, and cybernetics*, 5:834–846.

Beck, A. (2017). *First-order methods in optimization*. SIAM.

Beck, A. and Tetruashvili, L. (2013). On the convergence of block coordinate descent type methods. *SIAM journal on Optimization*, 23(4):2037–2060.

Beck, C. L. and Srikant, R. (2012). Error bounds for constant step-size q-learning. *Systems & control letters*, 61(12):1203–1208.

Benveniste, A., Métivier, M., and Priouret, P. (2012). *Adaptive algorithms and stochastic approximations*, volume 22. Springer Science & Business Media.

Bertsekas, D. (1996). Neuro-dynamic programming. *Athena Scientific*.

Bertsekas, D. P. et al. (2011). Dynamic programming and optimal control 3rd edition, volume ii. *Belmont, MA: Athena Scientific*, 1.

Bhandari, J., Russo, D., and Singal, R. (2018). A finite time analysis of temporal difference learning with linear function approximation. In *Conference on learning theory*, pages 1691–1692. PMLR.

Bharti, S., Kurian, D. S., and Pillai, V. M. (2020). Reinforcement learning for inventory management. In *Innovative Product Design and Intelligent Manufacturing Systems: Select Proceedings of ICIPDIMS 2019*, pages 877–885. Springer.

Borkar, V. (1991). Topics in controlled markov chains. pitman research notes in mathematics series# 240. *Pitman Research Notes in Mathematics Series*, 240.

Borkar, V., Chen, S., Devraj, A., Kontoyiannis, I., and Meyn, S. (2024). The ode method for asymptotic statistics in stochastic approximation and reinforcement learning.

Borkar, V. S. (2008). *Stochastic approximation: a dynamical systems viewpoint*, volume 9. Springer.

Chandak, S., Borkar, V. S., and Dodhia, P. (2022). Concentration of contractive stochastic approximation and reinforcement learning. *Stochastic Systems*, 12(4):411–430.

Chang, C.-C. and Lin, C.-J. (2011). Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3).

Chen, Z., Khodadadian, S., and Maguluri, S. T. (2022). Finite-sample analysis of off-policy natural actor–critic with linear function approximation. *IEEE Control Systems Letters*, 6:2611–2616.

Chen, Z., Maguluri, S. T., Shakkottai, S., and Shanmugam, K. (2020). Finite-sample analysis of contractive stochastic approximation using smooth convex envelopes. *Advances in Neural Information Processing Systems*, 33:8223–8234.

Chen, Z., Maguluri, S. T., Shakkottai, S., and Shanmugam, K. (2024). A lyapunov theory for finite-sample guarantees of markovian stochastic approximation. *Operations Research*, 72(4):1352–1367.

Chou, H.-Y., Lin, P.-Y., and Lin, C.-J. (2020). Dual coordinate-descent methods for linear one-class svm and svdd. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, pages 181–189. SIAM.

Cuartas, C. and Aguilar, J. (2023). Hybrid algorithm based on reinforcement learning for smart inventory management. *Journal of intelligent manufacturing*, 34(1):123–149.

Dann, C., Li, L., Wei, W., and Brunskill, E. (2019). Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, pages 1507–1516. PMLR.

Derman, C. and Veinott, A. F. (1967). A solution to a countable system of equations arising in markovian decision processes. *The Annals of Mathematical Statistics*, 38(2):582–584.

Diakonikolas, J. and Orecchia, L. (2018). Alternating randomized block coordinate descent. In *International Conference on Machine Learning*, pages 1224–1232. PMLR.

Falin, G. and Falin, A. (1999). Heavy traffic analysis of m/g/1 type queueing systems with markov-modulated arrivals. *Top*, 7(2):279–291.

Fercoq, O. and Richtárik, P. (2015). Accelerated, parallel, and proximal coordinate descent. *SIAM Journal on Optimization*, 25(4):1997–2023.

Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *Journal of computational and graphical statistics*, 7(3):397–416.

Glynn, P. W. and Iglehart, D. L. (1989). Importance sampling for stochastic simulations. *Management science*, 35(11):1367–1392.

Grosof, I., Hong, Y., Harchol-Balter, M., and Scheller-Wolf, A. (2023). The reset and marc techniques, with application to multiserver-job analysis.

Gurbuzbalaban, M., Ozdaglar, A., Parrilo, P. A., and Vanli, N. (2017). When cyclic coordinate descent outperforms randomized coordinate descent. *Advances in Neural Information Processing Systems*, 30.

Haque, S. U., Khodadadian, S., and Maguluri, S. T. (2023). Tight finite time bounds of two-time-scale linear stochastic approximation with markovian noise. *arXiv preprint arXiv:2401.00364*.

Harold, J., Kushner, G., and Yin, G. (1997). Stochastic approximation and recursive algorithm and applications. *Application of Mathematics*, 35(10).

Joachims, T. (1998). Making large-scale svm learning practical. Technical report, Technical report.

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

Khodadadian, S., Doan, T. T., Romberg, J., and Maguluri, S. T. (2023). Finite-sample analysis of two-time-scale natural actor–critic algorithm. *IEEE Transactions on Automatic Control*, 68(6):3273–3284.

Khodadadian, S., Sharma, P., Joshi, G., and Maguluri, S. T. (2022). Federated reinforcement learning: Linear speedup under markovian sampling. In *International Conference on Machine Learning*, pages 10997–11057. PMLR.

Kober, J., Bagnell, J. A., and Peters, J. (2013). Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274.

Konda, V. and Tsitsiklis, J. (1999). Actor-critic algorithms. *Advances in neural information processing systems*, 12.

Kumar, H., Koppel, A., and Ribeiro, A. (2023). On the sample complexity of actor-critic method for reinforcement learning with function approximation.

Kushner, H. J. and Yin, G. (1997). Stochastic approximation algorithms and applications. In *Applied Mathematics*.

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

Lauand, C. K. and Meyn, S. (2024). Revisiting step-size assumptions in stochastic approximation.

Li, G., Ee, Cai, C., and Wei, Y. (2021). Is q-learning minimax optimal? a tight sample complexity analysis. *Oper. Res.*, 72:222–236.

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. (2020). Sample complexity of asynchronous q-learning: Sharper analysis and variance reduction. *Advances in neural information processing systems*, 33:7031–7043.

Liu, B., Xie, Q., and Modiano, E. (2019). Reinforcement learning for optimal control of queueing systems. In *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 663–670.

Liu, S. D., Chen, S., and Zhang, S. (2025). The ode method for stochastic approximation and reinforcement learning with markovian noise.

Lund, R. B. and Tweedie, R. L. (1996). Geometric convergence rates for stochastically ordered

markov chains. *Mathematics of Operations Research*, 21(1):182–194.

Meyn, S. P. and Tweedie, R. L. (1994). Computable Bounds for Geometric Convergence Rates of Markov Chains. *The Annals of Applied Probability*, 4(4):981 – 1011.

Meyn, S. P. and Tweedie, R. L. (2012). *Markov chains and stochastic stability*. Springer Science & Business Media.

Mou, W., Pananjady, A., Wainwright, M. J., and Bartlett, P. L. (2021). Optimal and instance-dependent guarantees for markovian linear stochastic approximation. *arXiv preprint arXiv:2112.12770*.

Murthy, Y., Grosof, I., Maguluri, S. T., and Srikant, R. (2024). Performance of npg in countable state-space average-cost rl. *arXiv preprint arXiv:2405.20467*.

Nesterov, Y. (2012). Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362.

Nesterov, Y. and Stich, S. U. (2017). Efficiency of the accelerated coordinate descent method on structured optimization problems. *SIAM Journal on Optimization*, 27(1):110–123.

Nutini, J., Schmidt, M., Laradji, I., Friedlander, M., and Koepke, H. (2015). Coordinate descent converges faster with the gauss-southwell rule than random selection. In *International Conference on Machine Learning*, pages 1632–1641. PMLR.

Pananjady, A. and Wainwright, M. J. (2020). Instance-dependent ℓ_∞-bounds for policy evaluation in tabular reinforcement learning. *IEEE Transactions on Information Theory*, 67(1):566–585.

Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Qiu, S., Yang, Z., Ye, J., and Wang, Z. (2021). On finite-time convergence of actor-critic algorithm. *IEEE Journal on Selected Areas in Information Theory*, 2(2):652–664.

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

Richtárik, P. and Takáč, M. (2016). Parallel coordinate descent methods for big data optimization. *Mathematical Programming*, 156:433–484.

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

Sardy, S., Bruce, A. G., and Tseng, P. (2000). Block coordinate relaxation methods for nonparametric wavelet denoising. *Journal of computational and graphical statistics*, 9(2):361–379.

Shah, D., Xie, Q., and Xu, Z. (2020). Stable reinforcement learning with unbounded state space. In *Learning for Dynamics and Control*, pages 581–581. PMLR.

Song, C. and Diakonikolas, J. (2023). Cyclic coordinate dual averaging with extrapolation. *SIAM Journal on Optimization*, 33(4):2935–2961.

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

Stamoulis, G. D. and Tsitsiklis, J. N. (1990). On the settling time of the congested gi/g/1 queue. *Advances in Applied Probability*, 22(4):929–956.

Sun, R. and Ye, Y. (2021). Worst-case complexity of cyclic coordinate descent: O(n2) gap with randomized version. *Math. Program.*, 185(1–2):487–520.

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

Sutton, R. S. (2018). Reinforcement learning: An introduction. *A Bradford Book*.

Szepesvári, C. (2022). *Algorithms for reinforcement learning*. Springer Nature.

Tsitsiklis, J. N. (1994). Asynchronous stochastic approximation and q-learning. *Machine learning*, 16:185–202.

Tsitsiklis, J. N. and Van Roy, B. (1997). An analysis of temporal-difference learning with function approximation. *IEEE transactions on automatic control*, 42(5):674–690.

Tsitsiklis, J. N. and Van Roy, B. (1999). Average cost temporal-difference learning. *Automatica*, 35(11):1799–1808.

Wang, L., Cai, Q., Yang, Z., and Wang, Z. (2019). Neural policy gradient methods: Global optimality and rates of convergence.

Wang, Y., Chen, W., Liu, Y., Ma, Z.-M., and Liu, T.-Y. (2017). Finite sample analysis of the gtd policy evaluation algorithms in markov setting. *Advances in Neural Information Processing Systems*, 30.

Watkins, C. J. and Dayan, P. (1992). Q-learning. *Machine learning*, 8:279–292.

Wei, H., Liu, X., Wang, W., and Ying, L. (2024). Sample efficient reinforcement learning in mixed systems through augmented samples and its applications to

queueing networks. *Advances in Neural Information Processing Systems*, 36.

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

Xu, T. and Liang, Y. (2021). Sample complexity bounds for two timescale value-based reinforcement learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 811–819. PMLR.

Yang, Z., Zhang, K., Hong, M., and Başar, T. (2018). A finite sample analysis of the actor-critic algorithm. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 2759–2764.

Yu, H. (2012). Least squares temporal difference methods: An analysis under general conditions. *SIAM Journal on Control and Optimization*, 50(6):3310–3343.

Yu, H. (2017). On convergence of emphatic temporal-difference learning.

Yu, H. (2018). On convergence of some gradient-based temporal-differences algorithms for off-policy learning.

Zhang, S., Zhang, Z., and Maguluri, S. T. (2021). Finite sample analysis of average-reward td learning and q-learning. *Advances in Neural Information Processing Systems*, 34:1230–1242.

Zhang, Y. and Xie, Q. (2024). Constant stepsize q-learning: Distributional convergence, bias and extrapolation. *arXiv preprint arXiv:2401.13884*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

A DETAILED LITERATURE SURVEY

Reinforcement Learning: RL has been extensively studied in the literature, starting with asymptotic convergence which was established in Tsitsiklis (1994); Tsitsiklis and Van Roy (1997); Bertsekas (1996). Off late, there has been a growing interest in obtaining finite sample complexity of these algorithms as established in Beck and Srikant (2012); Bhandari et al. (2018); Srikant and Ying (2019); Qu and Wierman (2020); Li et al. (2020); Pananjady and Wainwright (2020); Chen et al. (2024); Zhang and Xie (2024). AC algorithms were initially proposed and studied in Barto et al. (1983); Konda and Tsitsiklis (1999) with their finite time performance analyzed in Kumar et al. (2023); Qiu et al. (2021); Wang et al. (2019); Chen et al. (2022). However, these finite time studies are primarily focused on finite state space settings. Some notable exceptions include recent works such as Shah et al. (2020); Murthy et al. (2024). In Shah et al. (2020), the authors focus on designing RL policies that ensure stable behavior in queueing systems without emphasizing optimality. On the other hand, Murthy et al. (2024) establishes finite time bounds for policy optimization using natural policy gradient in infinite state settings under an oracular critic having guaranteed error margins. We focus on constructing such a critic based on TD learning and characterizing its performance.

TD Learning: TD Learning is one of the most common algorithms for the critic phase, i.e., policy evaluation, and has been extensively studied both in discounted and average reward settings. The asymptotic behavior of TD learning in these regimes was characterized in Tsitsiklis and Van Roy (1997, 1999). Finite-sample complexity of TD has been established in Bhandari et al. (2018); Srikant and Ying (2019); Pananjady and Wainwright (2020) in the discounted reward setting and in Zhang et al. (2021) in the average-reward setting. However, most of the prior work on finite-sample guarantees considers only finite state MDPs. Motivated by applications in engineering systems, we study infinite state MDPs in the average reward setting, and establish finite sample guarantees.

Asymptotic Analysis of Stochastic Approximation: SA was first proposed by Robbins and Monro (1951) as a family of iterative algorithms to find the roots of an operator and has been extensively studied since then. Asymptotic convergence of SA was studied in Borkar (2008); Benveniste et al. (2012); Kushner and Yin (1997). More recent work including Borkar et al. (2024); Lauand and Meyn (2024); Allmeier and Gast (2024) studies SA with unbounded Markovian noise using the Poisson's equation, as we do in this paper. However, their focus is on establishing a central limit theorem, i.e., an asymptotic result of the form $(x_k - x^*)/\sqrt{\alpha_k} \xrightarrow{d} \mathcal{N}(0, \Sigma)$, for appropriate choice of Σ. In contrast, the focus of our work is on establishing a finite-time bound. Another line of work, inspired by off-policy algorithms in RL studies the asymptotic convergence of SA under more general setup where the solution to Poisson's equation may not exist. This approach, as seen in works by Yu (2012, 2017, 2018); Liu et al. (2025) only assumes the ergodic theorem for Markov chains, which is arguably the most general framework for controlling the noise in the algorithm (Kushner and Yin, 1997). Note that even when one has a finite-state MDP, off-policy RL algorithms such as Least Squares TD, Emphatic TD, and Gradient TD(λ) lead to SA with unbounded Markovian noise due to the product of importance sampling ratios. In some cases, such noise can even have infinite variance (Glynn and Iglehart, 1989). While we make a more restrictive assumption on the existence and moments of the solution of the Poisson's equation, we obtain finite-time mean-square bounds.

Finite Time Bounds for Stochastic Approximation: Finite time analysis has gained significant attraction in recent works such as Chen et al. (2020); Srikant and Ying (2019); Chen et al. (2024); Mou et al. (2021). In particular, these works demonstrate that finite-time bounds on general SA algorithms immediately imply performance guarantees of a large class of RL algorithms including V-Trace, Q-learning, n-step TD, etc. We contribute to this line of work by providing a general-purpose theorem for SA with unbounded Markovian noise which furnishes finite-sample bounds on various RL algorithms for infinite state MDPs, and we illustrate such use in the context of TD learning.

Poisson Equation for Markov Chains: Recent works on finite sample bounds of SA such as Bhandari et al. (2018); Srikant and Ying (2019); Mou et al. (2021); Qu and Wierman (2020); Xu and Liang (2021); Chen et al. (2024) have exploited geometric mixing of the underlying Markov chain. It is unclear if this approach generalizes to the case of unbounded setting. In this paper, we adopt the use of Poisson equation to analyze Markov noise which has been extensively used for this purpose in classical work on asymptotic convergence of SA, such as Benveniste et al. (2012); Harold et al. (1997), and also in other domains such as queueing theory in Grosof et al. (2023); Falin and Falin (1999). More recently, while this approach has recently been used to study linear SA in Chandak et al. (2022); Kaledin et al. (2020); Agrawal et al. (2024); Haque et al. (2023), their analysis is restricted to finite state space.

Block Coordinate Descent: Block coordinate descent (BCD) methods have been widely explored due to their effectiveness in large-scale distributed optimization (Fercoq and Richtárik, 2015; Richtárik and Takáč, 2016) for machine learning, such as in L1-regularized least squares (LASSO) (Fu, 1998; Sardy et al., 2000) and support vector machines (SVMs) (Joachims, 1998; Chang and Lin, 2011; Chou et al., 2020). While a substantial number of studies have investigated the Randomized and Greedy variants of BCD (Nesterov, 2012; Nutini et al., 2015; Nesterov and Stich, 2017; Diakonikolas and Orecchia, 2018; Lan, 2020), the literature on CBCD is not as rich. Some of the works that have analyzed it in deterministic settings include Beck and Tetruashvili (2013); Gurbuzbalaban et al. (2017); Song and Diakonikolas (2023) but to the best of our knowledge, no prior work has explored the stochastic version of CBCD.

B Q-LEARNING FOR DISCOUNTED-REWARD RL IN FINITE-STATE MDPs

In this section, we will consider the control problem in discounted-reward RL. In this setting, the goal is to maximize the expected cumulative discounted-reward. More formally, let $\gamma \in (0,1)$ be the discount factor and π be a policy, define the state value function $V_\pi : \mathcal{S} \to \mathbb{R}$ as:

$$V_\pi(s) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k \mathcal{R}(S_k, A_k) | S_0 = s \right]$$

Then the objective of the control problem is to directly find an optimal policy π^* such that $V_{\pi^*}(s) \geq V_\pi(s)$, $\forall s \in \mathcal{S}$ and any policy π. It can be shown that, under mild conditions, such a policy always exists (Puterman, 2014).

Q-learning (Watkins and Dayan, 1992) is one of the most popular algorithms for finding the optimal policy by running the following iteration

$$Q_{k+1}(s,a) = Q_k(s,a) + \alpha_k \mathbb{1}\{S_k = s, A_k = a\} \times \left(\mathcal{R}(s,a) + \gamma \max_{a' \in \mathcal{A}} Q_k(S_{k+1}, a') - Q_k(s,a) \right) \tag{12}$$

where $\{(S_k, A_k)\}_{k \geq 0}$ is a sample trajectory collected using a suitable behavior policy π_b and $\mathbb{1}\{\cdot\}$ is the indicator function. It can be shown that the Algorithm (12) converges to Q^* which is the unique fixed point of the Bellman optimality operator $\mathcal{B}(Q)$ defined by

$$\mathcal{B}(Q) = \mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) \max_{a' \in \mathcal{A}} Q(s', a')$$

Since Q^* and the optimal policy π^* satisfy the following relation: $\pi^*(\cdot|s) \in \arg\max_{a \in \mathcal{A}} Q^*(s,a)$ (Bertsekas, 1996), estimation of Q^* directly related to finding the optimal policy.

We make the following standard assumption on the Markov chain generated by π_b.

Assumption B.1. The behavior policy π_b satisfies $\pi_b(a|s) > 0$ for all (s,a) and the Markov chain $\mathcal{M}_S^{\pi_b} = \{S_k\}$ induced by π_b is irreducible.

Remark. The condition that $\pi(a|s) > 0$ for all (s,a) and the irreducibility of the induced Markov chain $\mathcal{M}_S^{\pi_b}$ is a standard assumption which ensures that all state action pairs are visited infinitely often Bertsekas (1996). Moreover, since the MDP is finite, Assumption B.1 implies that there exists a unique stationary distribution, which we denote as $\mu_b \in \Delta^{|\mathcal{S}|}$.

Remark. Recent works on the finite-time analysis of Q-learning often leverage the geometric mixing of Markov chain to handle Markovian noise (Li et al., 2020; Qu and Wierman, 2020; Chen et al., 2024; Zhang and Xie, 2024). To ensure geometric mixing, these works commonly assume that $\mathcal{M}_S^{\pi_b}$ is also aperiodic, which is crucial to achieving this property. However, in our case, we do not require the aperiodicity assumption, since we utilize the solution to the Poisson equation, which exists under Assumption B.1. This flexibility is significant; often, one can design more effective behavior policies from a wider class of distributions to balance the trade-off between exploration and exploitation.

B.1 Properties of the Q-learning Algorithm

To apply Theorem 2.2 to Q-learning we first rearrange the iteration (12) in the form of (1) and verify the assumptions.

$$\begin{aligned} Q_{k+1}(s,a) &= Q_k(s,a) + \alpha_k \mathbb{1}\{S_k = s, A_k = a\} \times \left(\mathcal{R}(s,a) + \gamma \max_{a' \in \mathcal{A}} Q_k(S_{k+1}, a') - Q_k(s,a) \right) \\ &= Q_k(s,a) + \alpha_k \big(F(Q_k, (S_k, A_k))(s,a) + M_k(Q_k)(s,a) \big) \end{aligned} \tag{13}$$

where

$$F(Q,(S,A))(s,a) = \mathbb{1}\{S = s, A = a\} \times \left(\mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) \max_{a' \in \mathcal{A}} Q(s',a') - Q(s,a) \right),$$

and

$$M_k(Q)(s,a) = \gamma \mathbb{1}\{S_k = s, A_k = a\} \times \left(\max_{a' \in \mathcal{A}} Q(S_{k+1}, a') - \sum_{s' \in \mathcal{S}} P(s'|S_k, A_k) \max_{a' \in \mathcal{A}} Q(s',a') \right).$$

Furthermore, denote $Y_k = (S_k, A_k)$. It is easy to verify that the process $\mathcal{M}_Q = \{Y_k\}$ is a Markov chain whose state space $\mathcal{Y} := \mathcal{S} \times \mathcal{A}$ is finite. Then, Q-learning algorithm can be written as

$$Q_{k+1} = Q_k + \alpha_k (F(Q_k, Y_k) + M_k(Q_k))$$

Note that by Assumption B.1, the Markov chain $\mathcal{M}_Q$ is irreducible and therefore it has a unique stationary distribution given by $\mu_Q(s,a) = \mu_b(s)\pi_b(a|s)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Let y_0 be some arbitrary state in $\mathcal{Y}$. For any $y \in \mathcal{Y}/\{y_0\}$, define $\tau_{y_0}^y$ as the expected hitting time of state y_0 starting from state y. Let τ_{y_0} denote $\max_{y \in \mathcal{Y}} \tau_{y_0}^y$, which is a well-defined quantity in finite state space (Meyn and Tweedie, 2012). Furthermore, let $\Lambda \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}$ be diagonal matrix with $\{\mu_Q(s,a)\}$ as diagonal entries. Then, we have the following proposition whose proof can be found in Appendix G.1.

Proposition B.1. *Under Assumption B.1, the Q-learning algorithm satisfies the following:*

(a) For any $Q, Q_1, Q_2 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ and $y \in \mathcal{Y}$, the operator $F(Q,y)$ has the following properties:

- *(1) The operator $F(Q,y)$ satisfies: $\|F(Q,y)\|_\infty \le 2\|Q - Q^*\|_\infty + \|Q^*\|_\infty$ and $\|F(Q_1,y) - F(Q_2,y)\|_\infty \le 2\|Q - Q^*\|_\infty$.*
- *(2) Define $\bar{F}(Q) = \mathbb{E}_{Y \sim \mu_Q}[F(Q,Y)]$. Then, $\bar{F}(Q) = \Lambda(\mathcal{B}(Q) - Q)$, where $\mathcal{B}(Q)$ is the Bellman optimality operator.*
- *(3) The solution to Bellman equation, i.e., Q^* is also the unique root of equation $\bar{F}(Q) = 0$.*

(b) There exists a solution to the Poisson equation (3) for the Markov chain $\mathcal{M}_Q$ which satisfies Assumption 2.2 with $A_2 = 4\tau_{y_0}$ and $B_2 = 0$.

(c) The noise sequence $M_k(Q_k)$ is a martingale difference sequence and satisfies Assumption 2.3 with constants $A_3 = 2$ and $B_3 = 2\|Q^*\|_\infty$.

Finally, we highlight the construction of a suitable Lyapunov function to study the convergence properties of the Q-learning algorithm.

B.2 Finite Sample Bounds for Q-Learning

The authors in Chen et al. (2020) showed that the Generalized Moreau Envelope can serve as a Lyapunov function for any operator which has the contraction property under a non-smooth norm. Specifically, consider the function $\Phi(x) = \min_{u\in\mathbb{R}^d}\left\{\frac{1}{2}\|u\|_c^2 + \frac{1}{2\omega}\|x-u\|_p^2\right\}$ where $\omega > 0$ and $p \geq 2$. The function $\Phi(\cdot)$ is known to be a smooth approximation of the function $\frac{1}{2}\|x\|_c^2$, with the smoothness parameter $\frac{p-1}{\omega}$. Further details on the properties of $\Phi(\cdot)$ can be found in Beck (2017).

For Q-learning $\|\cdot\|_c = \|\cdot\|_\infty$, which by the properties of ℓ_p norms implies that $l_{cs} = 1$ and $u_{cs} = (|\mathcal{S}||\mathcal{A}|)^{1/p}$. To verify Assumptions 2.4 in the context of Q-learning, we will need the following lemma whose proof can be found in Chen et al. (2020).

Lemma B.1. *Assign* $p = 2\log(|\mathcal{S}||\mathcal{A}|)$ *and* $\omega = \left(\frac{1}{2} + \frac{1}{2(1-(1-\gamma)\Lambda_{min})}\right)^2 - 1$, *where* $\Lambda_{min} = \min_{(s,a)}\{\mu_b(s)\pi_b(a|s)\} > 0$ *due to Assumption B.1. Then, the function* $\Phi(x)$ *satisfies the following properties:*

(a) For all $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ *we have* $\langle\nabla\Phi(Q-Q^*), \bar{F}(Q)\rangle \leq -(1-\gamma)\Lambda_{min}\Phi(Q-Q^*)$.

(b) $\Phi(x)$ *is convex, and* $\frac{p-1}{\omega}$*-smooth with respect to* $\|\cdot\|_p$. *That is* $\Phi(y) \leq \Phi(s) + \langle\nabla\Phi(x), y-x\rangle + \frac{p-1}{2\omega}\|x-y\|_p^2$ *for all* $x, y \in \mathbb{R}^d$.

(c) Let $l = 2(1+\omega/\sqrt{e})$ *and* $u = 2(1+\omega)$. *Then, we have* $l\Phi(x) \leq \|x\|_c^2 \leq u\Phi(x)$.

With all the assumptions satisfied, we can apply Theorem 2.2 to Q-learning. The exact characterization of the constants can be found in Appendix G.2.

Theorem B.1. *Consider the iterates* $\{Q_k\}_{k\geq 0}$ *generated by iteration* (12) *under Assumption B.1.*

(a) When $\alpha_k \equiv \alpha \leq \min\left\{1, \frac{\eta_Q}{A_Q(5+2\eta_Q)\varrho_{Q,1}}\right\}$, *then for all* $k \geq 0$*:*

$$\mathbb{E}[\|Q_{k+1} - Q^*\|_\infty^2] \leq \varrho_{Q,0}\exp\left(\frac{-\eta_Q\alpha k}{2}\right) + \frac{58B_Q\varrho_{Q,1}\alpha}{\eta_Q}.$$

(b) When $\xi = 1$, $\alpha > \frac{2}{\eta_Q}$ *and* $K \geq \max\{A_Q\alpha(5\alpha+8)\varrho_{Q,1}, 2\}$, *then for all* $k \geq 0$*:*

$$\mathbb{E}[\|Q_{k+1} - Q^*\|_\infty^2] \leq \varrho_{Q,0}\left(\frac{K}{k+K}\right)^{\frac{\eta_Q\alpha}{2}} + \frac{2B_Q\varrho_{Q,1}\alpha}{k+K} + \frac{72B_Q\varrho_{Q,1}e\alpha^2}{\left(\frac{\eta_Q\alpha}{2}-1\right)(k+K)}.$$

Remark. Note that for the case of constant step size, the above sample complexity immediately implies $\mathcal{O}\left(\frac{\log(1/\epsilon)}{\epsilon^2}\right)\mathcal{O}\left(\frac{1}{(1-\gamma)^5}\right)\mathcal{O}\left(\Lambda_{min}^{-3}\right)$, Corollary G.3. Compared to Chen et al. (2024), we do not have any poly-logarithmic factors and our bounds hold for a broader class of Markov chains (periodic or non-mixing). However, we note that our bounds are sub-optimal with respect to Λ_{min} and $1/(1-\gamma)$ as compared to Li et al. (2021). It is possible to improve the bounds by exploiting the specific nature of updates in Q-Learning instead of using our general-purpose theorem. Nevertheless, we include Q-learning as an illustrative application of our theorem and show that one can also get reasonable bounds for finite-state MDPs.

C PROOF OF MAIN THEOREMS 2.1 AND 2.2

Before presenting the proof of the main theorems, we will set up some notations for ease of exposition.

Common Notation: Since we are working with finite-dimensional space $\mathbb{R}^d$, there exists positive constants such that $l_{cs}\|x\|_c \leq \|x\|_s \leq u_{cs}\|x\|_c$ and $l_{2s}\|x\|_2 \leq \|x\|_s \leq u_{2s}\|x\|_2$. Denote $\|\cdot\|_{s^*}$ as the dual norm of $\|\cdot\|_s$ and $\kappa := \left(\frac{\xi}{\alpha} + \eta\right)$.

Notation for Theorem 2.1: Let $\max_{x\in\mathcal{X}}\|x\|_c = M/2$, when $\mathcal{X}$ is an ℓ_2 ball such that $x^* \in \mathcal{X}$. Then, we define the following constants for Theorem 2.1.

$$\hat{A}(y_0) = \hat{A}_1^2(y_0) + \hat{A}_2^2(y_0) + A_3^2;\ \hat{B}(y_0) = \hat{B}_1^2(y_0) + \hat{B}_2^2(y_0) + B_3^2;\ \hat{C}(y_0) = \hat{A}(y_0)M^2 + \hat{B}(y_0);$$

$$\varphi_1 = \frac{uL_s u_{2s} u_{cs}^2}{l_{2s}};\ \varphi_0 = \frac{u}{l}\|x_0 - x^*\|_c^2 + 2\varphi_1 \hat{C}(y_0).$$

Notation for Theorem 2.2: When the state space $\mathcal{Y}$ is bounded, the we define the following constants for Theorem 2.2.

$$A = (A_1 + A_3 + 1)^2;\ B = \left(B_1 + B_3 + \frac{B_2}{A_2}\right)^2;$$

$$\varrho_1 = uL_s u_{cs}^2 A_2;\ \varrho_0 = \frac{2u(1 + 2A\varrho_1)}{l}\|x_0 - x^*\|_c^2 + 4B\varrho_1.$$

Theorem C.1. *Suppose that we run the Markov chain with initial state y_0. When the state space $\mathcal{Y}$ is unbounded and the set $\mathcal{X}$ is an ℓ_2-ball of radius $R/2$, then under the Assumptions 2.1-2.5, $\{x_k\}_{k\geq 0}$ in the iterations (1) satisfy the following:*

(a) When $\alpha_k \equiv \alpha \leq 1$, then for all $k \geq 0$:

$$\mathbb{E}[\|x_{k+1} - x^*\|_c^2] \leq \varphi_0 \exp(-\eta\alpha k) + 3\varphi_1 \hat{C}(y_0)\alpha + \frac{6\varphi_1 \hat{C}(y_0)\alpha}{\eta}.$$

(b) When $\xi = 1$, $\alpha > \frac{1}{\eta}$ and $K \geq \max\{\alpha, 2\}$, then for all $k \geq 0$:

$$E[\|x_{k+1} - x^*\|_c^2] \leq \varphi_0 \left(\frac{K}{k+K}\right)^{\eta\alpha} + \frac{\varphi_1 \hat{C}(y_0)\alpha}{k+K} + \frac{4(6+4\eta)\varphi_1 \hat{C}(y_0) e\alpha^2}{\left(\frac{\eta\alpha}{2} - 1\right)(k+K)}.$$

(c) When $\xi < 1$, $\alpha > 0$ and $K \geq \max\{\alpha^{1/\xi}, 2\}$, then for all $k \geq 0$:

$$\mathbb{E}[\|x_{k+1} - x^*\|_c^2] \leq \varphi_0 \exp\left(\frac{-\eta\alpha}{(1-\xi)}\left[(k+K)^{1-\xi} - K^{1-\xi}\right]\right) + \frac{\varphi_1 \hat{C}(y_0)\alpha}{(k+K)^\xi} + \frac{2(6+4\kappa)\varphi_1 \hat{C}(y_0)\alpha}{\eta(k+K)^\xi}.$$

Theorem C.2. *When the state space $\mathcal{Y}$ is compact and the set $\mathcal{X} \equiv \mathbb{R}^d$, then under the Assumptions 2.1-2.5, $\{x_k\}_{k\geq 0}$ in the iterations (1) satisfy the following:*

(a) When $\alpha_k \equiv \alpha \leq \min\left\{1, \frac{\eta}{A(5+2\eta)\varrho_1}\right\}$, then for all $k \geq 0$:

$$\mathbb{E}[\|x_{k+1} - x^*\|_c^2] \leq \varrho_0 \exp\left(\frac{-\eta\alpha k}{2}\right) + 18B\varrho_1\alpha + \frac{40B\varrho_1\alpha}{\eta}.$$

(b) When $\xi = 1$, $\alpha > \frac{2}{\eta}$ and $K \geq \max\{A\alpha(5\alpha+8)\varrho_1, 2\}$, then for all $k \geq 0$:

$$E[\|x_{k+1} - x^*\|_c^2] \leq \varrho_0 \left(\frac{K}{k+K}\right)^{\frac{\eta\alpha}{2}} + \frac{2B\varrho_1\alpha}{k+K} + \frac{8B(5+4\eta)\varrho_1 e\alpha^2}{\left(\frac{\eta\alpha}{2} - 1\right)(k+K)}.$$

(c) When $\xi < 1$, $\alpha > 0$ and $K \geq \max\left\{\left(\frac{2A\alpha(5+2\kappa)\varrho_1}{\eta}\right)^{1/\xi}, 2\right\}$, then for all $k \geq 0$:

$$\mathbb{E}[\|x_{k+1} - x^*\|_c^2] \leq \varrho_0 \exp\left(\frac{-\eta\alpha}{2(1-\xi)}\left[(k+K)^{1-\xi} - K^{1-\xi}\right]\right) + \frac{2B\varrho_1\alpha}{(k+K)^\xi} + \frac{8B(5+2\kappa)\varrho_1\alpha}{\eta(k+K)^\xi}.$$

Before starting the proof of the theorems, we have the following lemma which decomposes the Lyapunov function at time $k+1$ using its properties in Assumption 2.4 and the recursion (2), thereby establishing a one-step recursive relation. Define the following terms:

$$\begin{aligned}
T_{1,1} &= \alpha_k \langle \nabla\Phi(x_{k+1} - x^*) - \nabla\Phi(x_k - x^*), V_{x_k}(Y_{k+1}))\rangle,\\
T_{1,2} &= \alpha_k \langle \nabla\Phi(x_{k+1} - x^*), V_{x_{k+1}}(Y_{k+1}) - V_{x_k}(Y_{k+1}))\rangle,\\
T_2 &= \frac{L_s\alpha_k^2}{2}\|F(x_k, Y_k) + M_k\|_s^2,\\
d_k &= \langle \nabla\Phi(x_k - x^*), V_{x_k}(Y_k)\rangle.
\end{aligned}$$

Lemma C.1. *Under the Assumptions 2.1-2.5, we have the following one-step recursive relation:*

$$\mathbb{E}[\Phi(x_{k+1}-x^*)] \leq (1-\eta\alpha_k)\mathbb{E}[\Phi(x_k-x^*)] + \alpha_k(\mathbb{E}[d_k]-\mathbb{E}[d_{k+1}]) + \mathbb{E}[T_{1,1}] + \mathbb{E}[T_{1,2}] + \mathbb{E}[T_2]. \tag{14}$$

Proof. Using the property (8) of the Lyapunov function and the iteration (1), we have

$$\begin{aligned}
\Phi(x_{k+1}-x^*) =& \Phi(\Pi_{\mathcal{X}}(x_k+\alpha_k(F(x_k,Y_k)+M_k))-x^*)\\
\leq& \Phi(x_k+\alpha_k(F(x_k,Y_k)+M_k)-x^*)\\
\leq& \Phi(x_k-x^*) + \langle\nabla\Phi(x_k-x^*), \alpha_k(F(x_k,Y_k)+M_k)\rangle + \frac{L_s}{2}\|\alpha_k(F(x_k,Y_k)+M_k)\|_s^2\\
=& \Phi(x_k-x^*) + \alpha_k\langle\nabla\Phi(x_k-x^*), \bar{F}(x_k)\rangle + \underbrace{\alpha_k\langle\nabla\Phi(x_k-x^*), F(x_k,Y_k)-\bar{F}(x)+M_k\rangle}_{T_1}\\
&+ \underbrace{\frac{L_s\alpha_k^2}{2}\|F(x_k,Y_k)+M_k\|_s^2}_{T_2}.
\end{aligned} \tag{15}$$

We begin by re-organizing T_1 with the help of solution to the Poisson's equation as follows:

$$\begin{aligned}
T_1 &= \alpha_k\langle\nabla\Phi(x_k-x^*), V_{x_k}(Y_k) - \mathbb{E}_{Y_k}[V_{x_k}(Y_{k+1})] + M_k\rangle\\
&= \alpha_k\langle\nabla\Phi(x_k-x^*), V_{x_k}(Y_{k+1}) - \mathbb{E}_{Y_k}[V_{x_k}(Y_{k+1})] + M_k\rangle + \alpha_k\langle\nabla\Phi(x_k-x^*), V_{x_k}(Y_k) - V_{x_k}(Y_{k+1})\rangle.
\end{aligned}$$

Observe that the first term is a martingale difference sequence with respect to the σ-field $\mathcal{F}_k$. We rewrite the second term as follows:

$$\begin{aligned}
\alpha_k\langle\nabla\Phi(x_k-x^*), V_{x_k}(Y_k) - V_{x_k}(Y_{k+1})\rangle = \alpha_k(d_k-d_{k+1}) &+ \underbrace{\alpha_k\langle\nabla\Phi(x_{k+1}-x^*)-\nabla\Phi(x_k-x^*), V_{x_k}(Y_{k+1}))\rangle}_{T_{1,1}}\\
&+ \underbrace{\alpha_k\langle\nabla\Phi(x_{k+1}-x^*), V_{x_{k+1}}(Y_{k+1}) - V_{x_k}(Y_{k+1}))\rangle}_{T_{1,2}}.
\end{aligned}$$

Taking expectation conditioned on $\mathcal{F}_k$ on both sides of Eq. (15), we get

$$\begin{aligned}
\mathbb{E}[\Phi(x_{k+1}-x^*)|\mathcal{F}_k] \leq \Phi(x_k-x^*) &+ \alpha_k\langle\nabla\Phi(x_k-x^*), \bar{F}(x_k)\rangle + \alpha_k\mathbb{E}[(d_k-d_{k+1})|\mathcal{F}_k]\\
&+ \mathbb{E}[T_{1,1}|\mathcal{F}_k] + \mathbb{E}[T_{1,2}|\mathcal{F}_k] + \mathbb{E}[T_2|\mathcal{F}_k].
\end{aligned}$$

Using Tower property and Eq. (5), we have

$$\mathbb{E}[\Phi(x_{k+1}-x^*)] \leq (1-\eta\alpha_k)\mathbb{E}[\Phi(x_k-x^*)] + \alpha_k(\mathbb{E}[d_k]-\mathbb{E}[d_{k+1}]) + \mathbb{E}[T_{1,1}] + \mathbb{E}[T_{1,2}] + \mathbb{E}[T_2].$$

□

Now we can proceed by bounding each of the terms in accordance with the specific settings.

Proof for Theorem 2.1. Using Eq. (16) in Lemma D.1 and Eq. (18) in Lemma (D.2) for $\mathbb{E}[T_{1,1}]$ and $\mathbb{E}[T_{1,2}]$ respectively, we get

$$\mathbb{E}[T_1] \leq \frac{4\alpha_k^2\varphi_1}{u}.$$

We use Eq. (21) in Lemma D.3 to get a bound on $\mathbb{E}[T_2]$,

$$\mathbb{E}[T_2] \leq \frac{2\alpha_k^2\varphi_1}{u}$$

Furthermore, to upper bound the second term in Eq. (14), we use Eq. (26) in Lemma D.5. Combining all the bounds, we get

$$\mathbb{E}[\Phi(x_{k+1}-x^*)] \leq (1-\eta\alpha_k)\mathbb{E}[\Phi(x_k-x^*)] + (1-\eta\alpha_k)\,\alpha_{k-1}\mathbb{E}[d_k] - \alpha_k\mathbb{E}[d_{k+1}] + \frac{\alpha_k^2(6+2\kappa)\varphi_1\hat{C}(y_0)}{u}$$

$$\leq \mathbb{E}[\Phi(x_0 - x^*)] \prod_{n=0}^{k} (1 - \eta\alpha_n) + \alpha_{-1}\mathbb{E}[d_0] \prod_{n=0}^{k} (1 - \eta\alpha_n) - \alpha_k \mathbb{E}[d_{k+1}]$$
$$+ \frac{(6+2\kappa)\varphi_1 \hat{C}(y_0)}{u} \sum_{n=0}^{k} \alpha_n^2 \prod_{\ell=n+1}^{k} (1 - \eta\alpha_\ell)$$

Since $K \geq 2$, α_{-1} is well-defined and is bounded above as $\alpha_{-1} \leq 2\alpha_0 \leq 2$. Furthermore, using Eq. (23) in Lemma D.4 for second and third term, we have

$$\mathbb{E}[\Phi(x_{k+1} - x^*)] \leq \left(\mathbb{E}[\Phi(x_0 - x^*)] + \frac{2\varphi_1}{u}\right) \prod_{n=0}^{k} (1 - \eta\alpha_n) + \frac{\alpha_k \varphi_1}{u} + \frac{(6+2\kappa)\varphi_1 \hat{C}(y_0)}{u} \sum_{n=0}^{k} \alpha_n^2 \prod_{\ell=n+1}^{k} (1 - \eta\alpha_\ell)$$

Using Eq. (7) in Assumption 2.4, we get

$$\mathbb{E}[\|x_{k+1} - x^*\|_c^2] \leq \left(\frac{u}{l}\|x_0 - x^*\|_c^2 + 2\varphi_1\right) \prod_{n=0}^{k} (1 - \eta\alpha_n) + \alpha_k \varphi_1 + (6+2\kappa)\varphi_1 \hat{C}(y_0) \sum_{n=0}^{k} \alpha_n^2 \prod_{\ell=n+1}^{k} (1 - \eta\alpha_\ell)$$

The finite time bounds for all the choices of step sizes can be obtained using the above bound by a straightforward application of Corollary 2.1.1 and Corollary 2.1.2 in Chen et al. (2020). □

Proof for Theorem 2.2. Using Eq. (17) Lemma D.1 and Eq. in (19) in Lemma D.2 for $\mathbb{E}[T_{1,1}]$ and $\mathbb{E}[T_{1,2}]$ respectively, we get

$$\mathbb{E}[T_1] \leq \frac{4\alpha_k^2 \varrho_1}{u} \left(uA\mathbb{E}[\Phi(x_k - x^*)] + B\right).$$

For $\mathbb{E}[T_2]$, we use (21) in Lemma D.3 to get,

$$\mathbb{E}[T_2] \leq \frac{\alpha_k^2 \varrho_1}{u} \left(uA\mathbb{E}[\Phi(x_k - x^*)] + B\right)$$

Furthermore, to upper bound the second term in Eq. (14), we use Eq. (26) in Lemma D.5. Using all the bounds, we get

$$\mathbb{E}[\Phi(x_{k+1} - x^*)] \leq (1 - \eta\alpha_k)\mathbb{E}[\Phi(x_k - x^*)] + \left(1 - \frac{\eta\alpha_k}{2}\right)\alpha_{k-1}\mathbb{E}[d_k] - \alpha_k \mathbb{E}[d_{k+1}]$$
$$+ \frac{\alpha_k^2 (5+2\kappa)\varrho_1}{u} \left(uA\mathbb{E}[\Phi(x_k - x^*)] + B\right).$$

Assume that α_k is small enough such that we have

$$\frac{\eta\alpha_k}{2} \geq A(5+2\kappa)\varrho_1 \alpha_k^2.$$

Using the above condition, we get

$$\mathbb{E}[\Phi(x_{k+1} - x^*)] \leq \left(1 - \frac{\eta\alpha_k}{2}\right)\mathbb{E}[\Phi(x_k - x^*)] + \left(1 - \frac{\eta\alpha_k}{2}\right)\alpha_{k-1}\mathbb{E}[d_k] - \alpha_k \mathbb{E}[d_{k+1}] + \alpha_k^2 \frac{B(5+2\kappa)\varrho_1}{u}.$$

Recursively writing the above inequality, we get

$$\mathbb{E}[\Phi(x_{k+1} - x^*)] \leq \mathbb{E}[\Phi(x_0 - x^*)] \prod_{n=0}^{k} \left(1 - \frac{\eta\alpha_n}{2}\right) + \alpha_{-1}\mathbb{E}[d_0] \prod_{n=0}^{k} \left(1 - \frac{\eta\alpha_n}{2}\right) - \alpha_k \mathbb{E}[d_{k+1}]$$
$$+ \frac{B(5+2\kappa)\varrho_1}{u} \sum_{n=0}^{k} \alpha_n^2 \prod_{\ell=n+1}^{k} \left(1 - \frac{\eta\alpha_\ell}{2}\right).$$

Again, since $K \geq 2$, α_{-1} is well-defined and is bounded above as $\alpha_{-1} \leq 2\alpha_0 \leq 2$. Furthermore, using Eq. (24) in Lemma D.4 for the second and the third term, we have

$$\mathbb{E}[\Phi(x_{k+1}-x^*)] \leq \left(\mathbb{E}[\Phi(x_0-x^*)] + \frac{2\varrho_1}{u}(uA\mathbb{E}[\Phi(x_0-x^*)]+B)\right)\prod_{n=0}^{k}\left(1-\frac{\eta\alpha_n}{2}\right) + \frac{\alpha_k\varrho_1}{u}(uA\mathbb{E}[\Phi(x_{k+1}-x^*)]+B) + \frac{B(5+2\kappa)\varrho_1}{u}\sum_{n=0}^{k}\alpha_n^2\prod_{\ell=n+1}^{k}\left(1-\frac{\eta\alpha_\ell}{2}\right)$$

Note that $5+2\kappa > \eta$. Thus, $\alpha_k \leq \frac{\eta}{2A(5+2\kappa)\varrho_1}$ implies that $\alpha_k A\varrho_1 \leq 0.5,\ \forall k \geq 0$. Thus, we have

$$\mathbb{E}[\Phi(x_{k+1}-x^*)] \leq \left(2\left(1+2A\varrho_1\right)\mathbb{E}[\Phi(x_0-x^*)] + \frac{4B\varrho_1}{u}\right)\prod_{n=0}^{k}\left(1-\frac{\eta\alpha_n}{2}\right) + \frac{2\alpha_k B\varrho_1}{u} + +\frac{2B(5+2\kappa)\varrho_1}{u}\sum_{n=0}^{k}\alpha_n^2\prod_{\ell=n+1}^{k}\left(1-\frac{\eta\alpha_\ell}{2}\right)$$

Using Eq. (7) in Assumption 2.4, we get

$$\mathbb{E}[\|x_{k+1}-x^*\|_c^2] \leq \left(\frac{2u(1+2A\varrho_1)}{l}\|x_0-x^*\|_c^2 + 4B\varrho_1\right)\prod_{n=0}^{k}\left(1-\frac{\eta\alpha_n}{2}\right) + 2\alpha_k B\varrho_1 + 2B(5+2\kappa)\varrho_1\sum_{n=0}^{k}\alpha_n^2\prod_{\ell=n+1}^{k}\left(1-\frac{\eta\alpha_\ell}{2}\right)$$

Again, the above bound immediately implies finite time bounds for various choices of step sizes by applying Corollary 2.1.1 and Corollary 2.1.2 in Chen et al. (2020).

□

D PROOF OF THE MAIN LEMMAS USED IN THEOREMS 2.1 AND 2.2

Lemma D.1. *Under the Assumptions 2.1-2.5, we have the following:*

(a) When the set $\mathcal{X}$ is an ℓ_2-ball of sufficiently large such that $x^* \in \mathcal{X}$, then

$$\mathbb{E}[T_{1,1}] \leq \frac{2\alpha_k^2\varphi_1\hat{C}(y_0)}{u}. \tag{16}$$

(b) When $\mathcal{Y}$ is bounded and $\mathcal{X} \equiv \mathbb{R}^d$, then

$$\mathbb{E}[T_{1,1}] \leq \frac{2\alpha_k^2\varrho_1}{u}\left(uA\mathbb{E}[\Phi(x_k-x^*)]+B\right). \tag{17}$$

Proof. To handle $\mathbb{E}[T_{1,1}]$, we use Holder's inequality and the smoothness of $\Phi(\cdot)$ to get

$$\begin{aligned}\mathbb{E}[T_{1,1}] &\leq \alpha_k\mathbb{E}[\|\nabla\Phi(x_{k+1}-x^*)-\nabla\Phi(x_k-x^*)\|_{s^*}\|V_{x_k}(Y_{k+1})\|_s]\\ &\leq \alpha_k L_s\mathbb{E}[\|x_{k+1}-x_k\|_s\|V_{x_k}(Y_{k+1})\|_s]\\ &\leq \alpha_k L_s u_{cs}^2\mathbb{E}[\|x_{k+1}-x_k\|_c\|V_{x_k}(Y_{k+1})\|_c]\\ &\leq \alpha_k L_s u_{cs}^2\mathbb{E}[\|x_{k+1}-x_k\|_c(A_2(Y_{k+1})\|x_k-x^*\|_c+B_2(Y_{k+1}))]. \qquad \text{(Eq. (4) in Assumption 2.2)}\end{aligned}$$

(a) Using Eq. (27) in Lemma D.6 and $\|x_k-x^*\|_c \leq M$, we get

$$\mathbb{E}[T_{1,1}] \leq \alpha_k^2\frac{L_s u_{2s}u_{cs}^2}{l_{2s}}\mathbb{E}[(A_1(Y_k)M+B_1(Y_k)+A_3M+B_3)(A_2(Y_{k+1})M+B_2(Y_{k+1}))]$$

$$\leq \alpha_k^2 \frac{\varphi_1}{2u}\mathbb{E}[(A_1(Y_k)M + B_1(Y_k) + A_3M + B_3)^2 + (A_2(Y_{k+1})M + B_2(Y_{k+1}))^2] \qquad (ab \leq \tfrac{a^2+b^2}{2})$$
$$\leq \alpha_k^2 \frac{2\varphi_1}{u}\left(\mathbb{E}[A_1^2(Y_k)M^2 + B_1^2(Y_k) + A_3^2M^2 + B_3^2 + A_2^2(Y_{k+1})M^2 + B_2^2(Y_{k+1})\right). \qquad (\left(\textstyle\sum_{i=1}^n a_i\right)^2 \leq n\left(\textstyle\sum_{i=1}^n a_i^2\right))$$

Finally, using part (d) in Assumption 2.2, we get

$$\mathbb{E}[T_{1,1}] \leq \frac{2\varphi_1\alpha_k^2}{u}\hat{C}(y_0).$$

(b) From part (d) in Assumption 2.2, we get

$$\begin{aligned}
\mathbb{E}[T_{1,1}] &\leq \alpha_k^2 L_s u_{cs}^2 \mathbb{E}[((A_1+A_3)\|x_k - x^*\|_c + B_1 + B_3)(A_2\|x_k - x^*\|_c + B_2)] \\
&\leq \alpha_k^2 L_s u_{cs}^2 A_2 \mathbb{E}\left[((A_1+A_3)\|x_k - x^*\|_c + B_1 + B_3)\left(\|x_k - x^*\|_c + \frac{B_2}{A_2}\right)\right] \\
&\leq \alpha_k^2 \frac{\varrho_1}{u}\mathbb{E}\left[\left((A_1+A_3+1)\|x_k - x^*\|_c + B_1 + B_3 + \frac{B_2}{A_2}\right)^2\right] \qquad (A_1, A_3 \geq 0, A_1 + A_3 + 1 \geq 1) \\
&\leq \frac{2\alpha_k^2\varrho_1}{u}\mathbb{E}\left[(A\|x_k - x^*\|_c^2 + B\right] \qquad ((a_1+a_2)^2 \leq 2\left(a_1^2 + a_2^2\right)) \\
&\leq \frac{2\alpha_k^2\varrho_1}{u}\left(uA\mathbb{E}[\Phi(x_k - x^*)] + B\right). \qquad \text{(Eq. (7) in Assumptions (2.4))}
\end{aligned}$$

□

Lemma D.2. *Under the Assumptions 2.1-2.5, we have the following:*

(a) When the set $\mathcal{X}$ is an ℓ_2-ball of sufficiently large such that $x^* \in \mathcal{X}$, then

$$\mathbb{E}[T_{1,2}] \leq \frac{2\alpha_k^2\varphi_1\hat{C}(y_0)}{u}. \tag{18}$$

(b) When $\mathcal{Y}$ is bounded and $\mathcal{X} \equiv \mathbb{R}^d$, then

$$\mathbb{E}[T_{1,2}] \leq \frac{2\alpha_k^2\varrho_1}{u}\left(uA\mathbb{E}[\Phi(x_k - x^*)] + B\right). \tag{19}$$

Proof. Denote s^* as the dual norm of s. To handle $T_{1,2}$ we use Holder's inequality to get

$$\begin{aligned}
\mathbb{E}[T_{1,2}] &\leq \alpha_k\mathbb{E}\left[\|\nabla\Phi(x_{k+1} - x^*)\|_{s^*}\|V_{x_{k+1}}(Y_{k+1}) - V_{x_k}(Y_{k+1}))\|_s\right] \\
&\leq \alpha_k\mathbb{E}\left[A_2(Y_{k+1})\|\nabla\Phi(x_{k+1} - x^*)\|_{s^*}\|x_{k+1} - x_k\|_s\right]. \qquad \text{(Eq. (4) in Assumption 2.2)}
\end{aligned}$$

Since $\Phi(\cdot)$ is convex-differentiable and achieves its minima at 0, $\nabla\Phi(0) = 0$. Along with smoothness of the function, we get

$$\begin{aligned}
&\|\nabla\Phi(x_{k+1} - x^*) - \nabla\Phi(0)\|_{s^*} \leq L_s\|x_{k+1} - x^*\|_s \\
&\implies \|\nabla\Phi(x_{k+1} - x^*)\|_{s^*} \leq L_s\|x_{k+1} - x^*\|_s
\end{aligned} \tag{20}$$

$$\begin{aligned}
\mathbb{E}[T_{1,2}] &\leq \alpha_k L_s\mathbb{E}\left[A_2(Y_{k+1})\|x_{k+1} - x^*\|_s\|x_{k+1} - x_k\|_s\right] \\
&\leq \alpha_k L_s u_{cs}\mathbb{E}\left[A_2(Y_{k+1})\|x_{k+1} - x^*\|_s\|x_{k+1} - x_k\|_c\right]
\end{aligned}$$

(a) Using Eq. (27) in Lemma D.6 and the fact that $\|x - x^*\|_c \leq M,\ \forall\ x \in \mathcal{X}$, we get

$$\mathbb{E}[T_{1,2}] \leq \alpha_k^2 L_s\frac{u_{2s}u_{cs}^2}{l_{2s}}\mathbb{E}[A_2(Y_{k+1})M(A_1(Y_k)M + B_1(Y_k) + A_3M + B_3)] \qquad (\|x_{k+1} - x^*\|_s \leq u_{cs}M)$$

$$\leq \alpha_k^2 \frac{\varphi_1}{2u}\mathbb{E}[A_2^2(Y_{k+1})M^2 + (A_1(Y_k)M + B_1(Y_k) + A_3M + B_3)^2] \qquad (ab \leq \tfrac{a^2+b^2}{2})$$
$$\leq \frac{2\alpha_k^2\varphi_1}{u}\left(\mathbb{E}[A_1^2(Y_k)M^2 + B_1^2(Y_k) + A_3^2M^2 + B_3^2 + A_2^2(Y_{k+1})M^2\right). \qquad ((\textstyle\sum_{i=1}^n a_i)^2 \leq n\left(\sum_{i=1}^n a_i^2\right))$$

Finally, using part (d) in Assumption 2.2, we get

$$\mathbb{E}[T_{1,2}] \leq \frac{2\alpha_k^2\varphi_1}{u}\left(\left(\hat{A}_1^2(y_0) + \hat{A}_2^2(y_0) + A_3^2\right)M^2 + \hat{B}_1^2(y_0) + B_3^2\right)$$
$$= \frac{2\alpha_k^2\varphi_1\hat{C}(y_0)}{u}.$$

(b) Since in this case $\mathcal{X} \equiv \mathbb{R}^d$, we use Eq. (1), to get

$$\|x_{k+1} - x^*\|_s \leq \|x_k - x^*\|_s + \|x_{k+1} - x_k\|_s$$
$$\leq \|x_k - x^*\|_s + \alpha_k u_{cs}((A_1 + A_3)\|x_k - x^*\|_c + B_1 + B_3) \qquad \text{(Eq. (28) in Lemma D.6)}$$
$$\leq u_{cs}\left((A_1 + A_3 + 1)\|x_k - x^*\|_c + B_1 + B_3\right). \qquad \text{(Assuming } \alpha_k \leq 1)$$

Furthermore, from part (d) and Eq. (28) in Lemma D.6, we get

$$\mathbb{E}[T_{1,2}] \leq \alpha_k^2 u_{cs}^2 L_s A_2 \mathbb{E}\left[\left((A_1 + A_3 + 1)\|x_k - x^*\|_c + B_1 + B_3\right)\right)\left((A_1 + A_3)\|x_k - x^*\|_c + B_1 + B_3\right)]$$
$$\leq \frac{\varrho_1\alpha_k^2}{u}\mathbb{E}\left[(A_1 + A_3 + 1)\|x_k - x^*\|_c + B_1 + B_3)^2\right]$$
$$\leq \frac{2\varrho_1\alpha_k^2}{u}\mathbb{E}\left[(A_1 + A_3 + 1)^2\|x_k - x^*\|_c^2 + (B_1 + B_3)^2\right] \qquad ((a_1 + a_2)^2 \leq 2\left(a_1^2 + a_2^2\right))$$
$$\leq \frac{2\varrho_1\alpha_k^2}{u}\left(uA\mathbb{E}[\Phi(x_k - x^*)] + B\right). \qquad \text{(Eq. (7) in Assumptions (2.4))}$$

□

Lemma D.3. *Under the Assumptions 2.1-2.5, we have the following:*

(a) When the set $\mathcal{X}$ is an ℓ_2-ball of sufficiently large such that $x^* \in \mathcal{X}$, then

$$\mathbb{E}[T_2] \leq \frac{2\alpha_k^2\varphi_1\hat{C}(y_0)}{u}. \tag{21}$$

(b) When $\mathcal{Y}$ is bounded and $\mathcal{X} \equiv \mathbb{R}^d$, then

$$\mathbb{E}[T_2] \leq \frac{\alpha_k^2\varrho_1}{u}\left(uA\mathbb{E}[\Phi(x_k - x^*)] + B\right). \tag{22}$$

Proof. (a)

$$E[T_2] \leq \frac{\alpha_k^2 L_s u_{cs}^2}{2}\mathbb{E}[(\|F(x_k, Y_k)\|_c + \|M_k\|_c)^2]$$
$$\leq \frac{\alpha_k^2 L_s u_{cs}^2}{2}\mathbb{E}[(A_1(Y)\|x_k - x^*\|_c + B_1(Y_k) + A_3\|x_k - x^*\|_c + B_3)^2] \qquad \text{(Assumptions 2.1 and 2.3)}$$
$$\leq \frac{\alpha_k^2 L_s u_{cs}^2}{2}\mathbb{E}[(A_1(Y)M + B_1(Y_k) + A_3M + B_3)^2] \qquad (\|x - x^*\|_c \leq M, \quad \forall x \in \mathcal{X})$$
$$\leq 2\alpha_k^2 L_s u_{cs}^2\mathbb{E}[A_1^2(Y)M^2 + B_1^2(Y_k) + A_3^2M^2 + B_3^2] \qquad ((\textstyle\sum_{i=1}^n a_i)^2 \leq n\left(\sum_{i=1}^n a_i^2\right))$$
$$\leq 2\alpha_k^2 L_s u_{cs}^2\left((\hat{A}_1^2(y_0) + A_3^2)M^2 + \hat{B}_1^2(y_0) + B_3^2\right) \qquad \text{(Part (d) in Assumption 2.2)}$$
$$\leq \frac{2\alpha_k^2\varphi_1\hat{C}(y_0)}{u}. \qquad (\tfrac{u_{2s}}{l_{2s}} \geq 1 \text{ and } \hat{A}_2^2(y_0), \hat{B}_2^2(y_0) \geq 0)$$

(b)

$$\begin{aligned}
\mathbb{E}[T_2] &\leq \frac{\alpha_k^2 L_s u_{cs}^2}{2}\mathbb{E}[(\|F(x_k,Y_k)\|_c + \|M_k\|_c)^2] \\
&\leq \frac{\alpha_k^2 L_s u_{cs}^2}{2}\mathbb{E}[(A_1(Y)\|x_k - x^*\|_c + B_1(Y_k) + A_3\|x_k - x^*\|_c + B_3)^2] && \text{(Assumptions 2.1 and 2.3)} \\
&\leq \frac{\alpha_k^2 L_s u_{cs}^2}{2}\mathbb{E}[(A_1\|x_k - x^*\|_c + B_1 + A_3\|x_k - x^*\|_c + B_3)^2] && \text{(Part (d) in Assumption 2.2)} \\
&\leq \alpha_k^2 L_s u_{cs}^2 \mathbb{E}[(A_1 + A_3)^2\|x_k - x^*\|_c^2 + (B_1 + B_3)^2] \\
&\leq \frac{\alpha_k^2 \varrho_1}{u}\left(uA\mathbb{E}[\Phi(x_k - x^*)] + B\right). && (A_2 \geq 1 \text{ and Eq. (7) in Assumptions (2.4)})
\end{aligned}$$

□

Lemma D.4. *Under the Assumptions 2.1-2.5, we have the following:*

(a) When the set $\mathcal{X}$ is an ℓ_2-ball of sufficiently large such that $x^* \in \mathcal{X}$, then

$$\mathbb{E}[|d_k|] \leq \frac{\varphi_1 \hat{C}(y_0)}{u}. \tag{23}$$

(b) When $\mathcal{Y}$ is bounded and $\mathcal{X} \equiv \mathbb{R}^d$, then

$$\mathbb{E}[|d_k|] \leq \frac{\varrho_1}{u}(uA\mathbb{E}[\Phi(x_k - x^*)] + B). \tag{24}$$

Proof. Using Holder's inequality, we have

$$\mathbb{E}[|d_k|] \leq \mathbb{E}[\|\nabla\Phi(x_k - x^*)\|_{s^*}\|V_{x_k}(Y_k)\|_s].$$

Using the same argument as in Eq. (20), we get

$$\begin{aligned}
\mathbb{E}[|d_k|] &\leq L_s\mathbb{E}[\|x_k - x^*\|_s\|V_{x_k}(Y_k)\|_s] \\
&\leq L_s u_{cs}^2\mathbb{E}[\|x_k - x^*\|_c\|V_{x_k}(Y_k)\|_c] \\
&\leq L_s u_{cs}^2\mathbb{E}[\|x_k - x^*\|_c(A_2(Y_k)\|x_k - x^*\|_c + B_2(Y_k))]. && \text{(Using Eq. (4) in Assumption 2.2)}
\end{aligned}$$

(a) Since $\|x_k - x^*\|_c \leq M$, we get

$$\begin{aligned}
\mathbb{E}[|d_k|] &\leq L_s u_{cs}^2\mathbb{E}[M(A_2(Y_k)M + B_2(Y_k))] \\
&\leq \frac{L_s u_{cs}^2}{2}\mathbb{E}[M^2 + (A_2(Y_k)M + B_2(Y_k))^2] \\
&\leq L_s u_{cs}^2\mathbb{E}[M^2 + A_2^2(Y_k)M^2 + B_2^2(Y_k)] && ((a_1+a_2)^2 \leq 2\left(a_1^2 + a_2^2\right)) \\
&\leq L_s u_{cs}^2\hat{C}(y_0) && \text{(Part (d) in Assumption 2.2)} \\
&\leq \frac{\varphi_1\hat{C}(y_0)}{u}. && (\tfrac{u_{2s}}{l_{2s}} \geq 1)
\end{aligned}$$

(b) For this part, we have

$$\begin{aligned}
\mathbb{E}[|d_k|] &\leq \frac{L_s u_{cs}^2 A_2}{2}\left(\mathbb{E}\left[\|x_k - x^*\|_c^2 + \left(\|x_k - x^*\|_c + \frac{B_2}{A_2}\right)^2\right]\right) && (ab \leq \tfrac{a^2+b^2}{2}) \\
&\leq L_s u_{cs}^2 A_2\left(\mathbb{E}[\|x_k - x^*\|_c^2] + \left(\frac{B_2}{A_2}\right)^2\right) \\
&\leq \frac{\varrho_1}{u}(uA\mathbb{E}[\Phi(x_k - x^*)] + B). && \text{(Eq. (7) in Assumptions (2.4))}
\end{aligned}$$

□

Lemma D.5. *Under the Assumptions 2.1-2.5, we have the following:*

(a) When the set $\mathcal{X}$ is an ℓ_2-ball of sufficiently large such that $x^* \in \mathcal{X}$, then

$$\alpha_k(\mathbb{E}[d_k] - \mathbb{E}[d_{k+1}]) \leq (1 - \eta\alpha_k)\,\alpha_{k-1}\mathbb{E}[d_k] - \alpha_k\mathbb{E}[d_{k+1}] + \frac{2\alpha_k^2\kappa\varphi_1\hat{C}(y_0)}{u}. \tag{25}$$

(b) When $\mathcal{Y}$ is bounded and $\mathcal{X} \equiv \mathbb{R}^d$, then

$$\alpha_k(\mathbb{E}[d_k] - \mathbb{E}[d_{k+1}]) \leq \left(1 - \frac{\eta\alpha_k}{2}\right)\alpha_{k-1}\mathbb{E}[d_k] - \alpha_k\mathbb{E}[d_{k+1}] + \frac{2\alpha_k^2\kappa\varrho_1}{u}\left(uA\mathbb{E}[\Phi(x_k - x^*)] + B\right). \tag{26}$$

Proof. (a) Re-writing the expression, we get

$$\begin{aligned}
\alpha_k(\mathbb{E}[d_k] - \mathbb{E}[d_{k+1}]) &= (1 - \eta\alpha_k)\,\alpha_{k-1}\mathbb{E}[d_k] - \alpha_k\mathbb{E}[d_{k+1}] + \alpha_k\mathbb{E}[d_k] - (1 - \eta\alpha_k)\,\alpha_{k-1}\mathbb{E}[d_k] \\
&= (1 - \eta\alpha_k)\,\alpha_{k-1}\mathbb{E}[d_k] - \alpha_k\mathbb{E}[d_{k+1}] + (\alpha_k - \alpha_{k-1} + \eta\alpha_k\alpha_{k-1})\,\mathbb{E}[d_k] \\
&\leq (1 - \eta\alpha_k)\,\alpha_{k-1}\mathbb{E}[d_k] - \alpha_k\mathbb{E}[d_{k+1}] + 2\alpha_k^2\left(\frac{\xi}{\alpha} + \eta\right)\mathbb{E}[d_k]. \qquad \text{(Lemma H.1)}
\end{aligned}$$

Using Eq. (23) in Lemma D.4, we get

$$\alpha_k(\mathbb{E}[d_k] - \mathbb{E}[d_{k+1}]) \leq (1 - \eta\alpha_k)\,\alpha_{k-1}\mathbb{E}[d_k] - \alpha_k\mathbb{E}[d_{k+1}] + \frac{2\alpha_k^2\kappa\varphi_1\hat{C}(y_0)}{u}.$$

(b) For this part, we re-write the expression as follows:

$$\begin{aligned}
\alpha_k(\mathbb{E}[d_k] - \mathbb{E}[d_{k+1}]) &= \left(1 - \frac{\eta\alpha_k}{2}\right)\alpha_{k-1}\mathbb{E}[d_k] - \alpha_k\mathbb{E}[d_{k+1}] + \alpha_k\mathbb{E}[d_k] - \left(1 - \frac{\eta\alpha_k}{2}\right)\alpha_{k-1}\mathbb{E}[d_k] \\
&= \left(1 - \frac{\eta\alpha_k}{2}\right)\alpha_{k-1}\mathbb{E}[d_k] - \alpha_k\mathbb{E}[d_{k+1}] + \left(\alpha_k - \alpha_{k-1} + \frac{\eta\alpha_k\alpha_{k-1}}{2}\right)\mathbb{E}[d_k] \\
&\leq \left(1 - \frac{\eta\alpha_k}{2}\right)\alpha_{k-1}\mathbb{E}[d_k] - \alpha_k\mathbb{E}[d_{k+1}] + \alpha_k^2\left(\frac{2\xi}{\alpha} + \eta\right)\mathbb{E}[d_k]. \qquad \text{(Lemma H.1)}
\end{aligned}$$

Using Eq. (24) in Lemma D.4 to get

$$\begin{aligned}
\alpha_k(\mathbb{E}[d_k] - \mathbb{E}[d_{k+1}]) &\leq \left(1 - \frac{\eta\alpha_k}{2}\right)\alpha_{k-1}\mathbb{E}[d_k] - \alpha_k\mathbb{E}[d_{k+1}] \\
&\quad + \alpha_k^2\left(\frac{2\xi}{\alpha} + \eta\right)\frac{\varrho_1}{u}\left(uA\mathbb{E}[\Phi(x_k - x^*)] + B\right) \\
&\leq \left(1 - \frac{\eta\alpha_k}{2}\right)\alpha_{k-1}\mathbb{E}[d_k] - \alpha_k\mathbb{E}[d_{k+1}] + \frac{2\alpha_k^2\kappa\varrho_1}{u}\left(uA\mathbb{E}[\Phi(x_k - x^*)] + B\right).
\end{aligned}$$

□

Lemma D.6. *Under the Assumptions 2.1 and 2.3, $\forall\ k \geq 0$, we have*

(a) When the set $\mathcal{X}$ is an ℓ_2-ball of sufficiently large such that $x^* \in \mathcal{X}$, then

$$\|x_{k+1} - x_k\|_c \leq \alpha_k\frac{u_{2c}}{l_{2c}}(A_1(Y_k)M + B_1(Y_k) + A_3M + B_3) \tag{27}$$

(b) When $\mathcal{Y}$ is bounded and $\mathcal{X} \equiv \mathbb{R}^d$, then

$$\|x_{k+1} - x_k\|_c \leq \alpha_k((A_1 + A_3)\|x_k - x^*\|_c + B_1 + B_3) \tag{28}$$

Proof. (a) Using the iteration (2) and the fact that $x_k \in \mathcal{X}$, we have

$$\begin{aligned}
\|x_{k+1} - x_k\|_c &\leq u_{2c}\|x_{k+1} - x_k\|_c \\
&= u_{2c}\|\Pi_{\mathcal{X}}\left(x_k + \alpha_k(F(x_k, Y_k) + M_k)\right) - \Pi_{\mathcal{X}}(x_k)\|_2 \\
&\leq \alpha_k u_{2c}\|F(x_k, Y_k) + M_k\|_2 && \text{(Non-expansive projection)} \\
&\leq \alpha_k \frac{u_{2c}}{l_{2c}}\|F(x_k, Y_k) + M_k\|_c \\
&\leq \alpha_k \frac{u_{2c}}{l_{2c}}(\|F(x_k, Y_k)\|_c + \|M_k\|_c) \\
&\leq \alpha_k \frac{u_{2c}}{l_{2c}}((A_1(Y_k) + A_3)\|x_k - x^*\|_c + B_1(Y_k) + B_3) && \text{(Assumptions (2.1) and (2.3))} \\
&\leq \alpha_k \frac{u_{2s}}{l_{2s}}((A_1(Y_k) + A_3)M + B_1(Y_k) + B_3). && (\|x_k - x^*\|_c \leq M)
\end{aligned}$$

(b) Using the iteration (2) and the fact that $\mathcal{X} \equiv \mathbb{R}^d$, we have

$$\begin{aligned}
\|x_{k+1} - x_k\|_c &= \alpha_k\|F(x_k, Y_k) + M_k\|_c \\
&\leq \alpha_k(\|F(x_k, Y_k)\|_c + \|M_k\|_c) \\
&\leq \alpha_k(\|F(x_k, Y_k)\|_c + \|M_k\|_c) \\
&\leq \alpha_k((A_1(Y_k) + A_3)\|x_k - x^*\|_c + B_1(Y_k) + B_3) && \text{(Assumptions (2.1) and (2.3))} \\
&\leq \alpha_k((A_1 + A_3)\|x_k - x^*\|_c + B_1 + B_3). && \text{(Part (d) in Assumption 2.2)}
\end{aligned}$$

□

E PROOF OF THE TECHNICAL RESULTS IN SECTION 3

Before beginning the proofs of the lemmas and propositions in this section, we need Lemma 7 from Tsitsiklis and Van Roy (1997) in order to prove Lemma 3.1. We state it here for completeness, but we omit the proof as it is essentially repeating the same arguments with the contraction factor being 1.

Lemma E.1. *Under Assumptions 3.1 and 3.2, the following relations hold in the steady state of the Markov process Y_k.*

(a) $\mathbb{E}_\mu[\psi(\tilde{S}_k)\psi(\tilde{S}_{k+m})^T] = \Psi^T \Lambda P^m \Psi$, *for all* $m \geq 0$.

(b) $\|E_\mu[\psi(\tilde{S}_k)\psi(\tilde{S}_{k+m})^T]\|_2 = \psi' < \infty$, *for all* $m \geq 0$.

(c) $E_\mu[\tilde{z}_k \psi(\tilde{S}_k)^T] = \Psi^T \Lambda \left(\sum_{m=0}^{\infty} \lambda^m P^m\right) \Psi$.

(d) $E_\mu[\tilde{z}_k \psi(\tilde{S}_{k+1})^T] = \Psi^T \Lambda \left(\sum_{m=0}^{\infty} \lambda^m P^{m+1}\right) \Psi$.

(e) $E_\mu[\tilde{z}_k \mathcal{R}(\tilde{S}_k, A_k)] = \Psi^T \Lambda \left(\sum_{m=0}^{\infty} \lambda^m P^m\right) \mathcal{R}_\pi$.

E.1 Proof of Lemma 3.1

Proof. Using Lemma E.1, we have

$$\begin{aligned}
\mathbb{E}_\mu[T(\tilde{Y}_k)] &= \begin{bmatrix} -c_\alpha & 0 \\ -\Pi_{2, E_\Psi^\perp}\mathbb{E}_\mu[\tilde{z}_k] & \Pi_{2, E_\Psi^\perp}\mathbb{E}_\mu\left[\tilde{z}_k\left(\psi(\tilde{S}_{k+1})^T\theta_k - \psi(\tilde{S}_k)^T\right)\right] \end{bmatrix} \\
&= \begin{bmatrix} -c_\alpha & 0 \\ -\frac{1}{(1-\lambda)}\Pi_{2, E_\Psi^\perp}\Psi^T\mu & \Pi_{2, E_\Psi^\perp}\left(\sum_{m=0}^{\infty}\lambda^m \Psi^T \Lambda P^{m+1}\Psi - \Psi^T\Lambda P^m \Psi\right) \end{bmatrix}.
\end{aligned}$$

Note that for any $\lambda \in [0, 1)$, we can rewrite $\lambda^m = (1-\lambda)\sum_{l=m}^{\infty}\lambda^l$. Then, it follows that for any $j \geq 0$, we have

$$\sum_{m=0}^{\infty}\lambda^m P^{m+j} = (1-\lambda)\sum_{m=0}^{\infty} P^{m+j}\sum_{l=m}^{\infty}\lambda^l$$

$$= (1-\lambda)\sum_{l=0}^{\infty}\lambda^l\sum_{m=j}^{l+j}P^m.$$

Using the above relation for $j=0$ and $j=1$, we get

$$\begin{aligned}\sum_{m=0}^{\infty}\lambda^m\Psi^T\Lambda P^{m+1}\Psi-\Psi^T\Lambda P^m\Psi &= (1-\lambda)\sum_{l=0}^{\infty}\lambda^l\Psi^T\Lambda\left(\sum_{m=1}^{l+1}P^m-\sum_{m=0}^{l}P^m\right)\Psi\\ &= (1-\lambda)\sum_{l=0}^{\infty}\lambda^l\left(\Psi^T\Lambda P^{l+1}\Psi-\Psi^T\Lambda\Psi\right)\\ &= \Psi^T\Lambda P^{(\lambda)}\Psi-\Psi^T\Lambda\Psi.\end{aligned}$$

Thus, we have

$$\mathbb{E}_\mu[T(\tilde{Y}_k)] = \begin{bmatrix}-c_\alpha & 0\\ \frac{1}{(\lambda-1)}\Pi_{2,E_\Psi^\perp}\Psi^T\mu & \Pi_{2,E_\Psi^\perp}\left(\Psi^T\Lambda P^{(\lambda)}\Psi-\Psi^T\Lambda\Psi\right)\end{bmatrix} = \bar{T}.$$

Similarly, using Lemma E.1 the steady-state expectation of $b(Y_k)$ is given by

$$\begin{aligned}\mathbb{E}_\mu[b(\tilde{Y}_k)] &= \begin{bmatrix}c_\alpha\mathbb{E}_\mu[\mathcal{R}(\tilde{S}_k,A_k)]\\ \Pi_{2,E_\Psi^\perp}\mathbb{E}_\mu[\mathcal{R}(\tilde{S}_k,A_k)\tilde{z}_k]\end{bmatrix}\\ &= \begin{bmatrix}c_\alpha\bar{r}\\ \Pi_{2,E_\Psi^\perp}\Psi^T\Lambda\left((1-\lambda)\sum_{l=0}^{\infty}\lambda^l\sum_{m=0}^{l}P^m\mathcal{R}_\pi\right)\end{bmatrix}\\ &= \begin{bmatrix}c_\alpha\bar{r}\\ \Pi_{2,E_\Psi^\perp}\Psi^T\Lambda\mathcal{R}^{(\lambda)}\end{bmatrix} = \bar{b}.\end{aligned}$$

□

E.2 Properties of TD(λ)

Next, we state the following lemma which will be crucial for proving desired properties of TD(λ). Define $Y_k=(S_k,A_k,S_{k+1},z_k)$ and $g(Y_k)=\frac{f_1(S_{k+1})}{(1-\lambda)(1-\rho)}+\frac{\|z_k\|_2\rho^{-1}(f_1(S_k)+f_1(S_{k+1}))}{(1-\lambda\rho)}+\frac{\|z_k\|_2\left(1+\hat{r}+\hat{\psi}\sqrt{d}\right)}{(1-\lambda)}+\frac{\hat{\psi}+\psi'+\hat{\psi}\hat{r}\sqrt{d}}{(1-\lambda)^2}$. Let $Y_0=y_0=(s_0,a_0,s_1,z_0)$.

Lemma E.2. *Assume that the eligibility trace vector z_k was initialized from z_0. Then, the following relations hold, for all $y_0\in\mathcal{Y}$:*

(a) $\mathbb{E}_\mu[\tilde{z}_k]=\frac{1}{(1-\lambda)}\Psi^T\mu$. *Furthermore,* $\sum_{k=0}^{\infty}\|\mathbb{E}_{y_0}[z_k]-\mathbb{E}_\mu[\tilde{z}_k]\|_2\le g(y_0)$.

(b) $\sum_{k=0}^{\infty}\|\mathbb{E}_{y_0}[z_k\psi(S_k)^T]-\mathbb{E}_\mu[\tilde{z}_k\psi(\tilde{S}_k)^T]\|_2\le g(y_0)$.

(c) $\sum_{k=0}^{\infty}\|\mathbb{E}_{y_0}[z_k\psi(S_{k+1})^T]-\mathbb{E}_\mu[\tilde{z}_k\psi(\tilde{S}_{k+1})^T]\|_2\le g(y_0)$.

(d) $\|\Psi^T\Lambda P^m\mathcal{R}_\pi\|_2\le\hat{\psi}\hat{r}\sqrt{d}$. *Furthermore,* $\sum_{k=0}^{\infty}\|\mathbb{E}_{y_0}[z_k\mathcal{R}(S_k,A_k)]-\mathbb{E}_\mu[\tilde{z}_k\mathcal{R}(\tilde{S}_k,A_k)]\|_2\le g(y_0)$.

(e) $\mathbb{E}_{y_0}[\|z_k\|_2^4]\le\frac{\|z_0\|_2^4+f_3(s_1)}{(1-\lambda)^4}$.

(f) $\mathbb{E}_{y_0}[g^2(Y_k)]\le\frac{4\sqrt{f_2(s_1)}}{(1-\lambda)^2(1-\rho)^2}+\frac{16\rho^{-2}\sqrt{\|z_0\|_2^4+f_3(s_1)}\sqrt{f_2(s_0)+f_2(s_1)}}{(1-\lambda\rho)^2(1-\lambda)^2}+\frac{4\left(1+\hat{r}+\hat{\psi}\sqrt{d}\right)^2\sqrt{\|z_0\|_2^4+f_3(s_1)}}{(1-\lambda)^4}+\frac{4(\hat{\psi}+\psi'+\hat{\psi}\hat{r}\sqrt{d})^2}{(1-\lambda)^4}$.

Proof. (a) From the definition of $\tilde{z}_k$, we have

$$\mathbb{E}_\mu[\tilde{z}_k]=\mathbb{E}_\mu\left[\sum_{m=-\infty}^{k}\lambda^{k-m}\psi(\tilde{S}_k)\right]$$

$$= \sum_{m=-\infty}^{k} \lambda^{k-m} \mathbb{E}_\mu[\psi(\tilde{S}_k)] \qquad \text{(Assumption 3.2 and Dominated Convergence Theorem)}$$
$$= \sum_{m=-\infty}^{k} \lambda^{k-m} \left(\sum_{s\in\mathcal{S}} \mu(s)\psi(s) \right)$$
$$= \frac{1}{1-\lambda} \left(\sum_{s\in\mathcal{S}} \mu(s)\psi(s) \right) = \frac{1}{1-\lambda} \Psi^T \mu.$$

Recall that $z_k = \lambda^k z_0 + \sum_{j=1}^{k} \lambda^{k-j}\psi(S_k)$. Using Assumption 3.3 and the above relation, we have

$$\mathbb{E}_{y_0}[z_k] - \mathbb{E}_\mu[\tilde{z}_k] = \lambda^k z_0 + \mathbb{E}_{y_0}\left[\sum_{j=0}^{k-1} \lambda^j \psi(S_{k-j}) \right] - \sum_{j=0}^{\infty} \lambda^j \sum_{s\in\mathcal{S}} \mu(s)\psi(s)$$
$$= \lambda^k z_0 + \mathbb{E}_{y_0}\left[\sum_{j=0}^{k-1} \lambda^j \left(\psi(S_{k-j}) - \sum_{s\in\mathcal{S}} \mu(s)\psi(s) \right) \right] - \sum_{j=k}^{\infty} \lambda^j \sum_{s\in\mathcal{S}} \mu(s)\psi(s)$$

Taking norm both sides and using triangle inequality, we get

$$\|\mathbb{E}_{y_0}[z_k] - \mathbb{E}_\mu[\tilde{z}_k]\|_2 \le \lambda^k \|z_0\|_2 + \left\| \mathbb{E}_{y_0}\left[\sum_{j=0}^{k-1} \lambda^j (\psi(S_{k-j}) - \sum_{s\in\mathcal{S}} \mu(s)\psi(s) \right] \right\|_2 + \| \sum_{s\in\mathcal{S}} \mu(s)\psi(s)\|_2 \sum_{j=k}^{\infty} \lambda^j$$
$$\le \lambda^k \|z_0\|_2 + \sum_{j=0}^{k-1} \lambda^j \left\| \mathbb{E}_{y_0}\left[(\psi(S_{k-j}) - \mathbb{E}_\mu[\psi(\tilde{S}_k)] \right] \right\|_2 + \frac{\lambda^k \hat{\psi}}{1-\lambda}$$
(Jensen's inequality and Assumption 3.2)
$$\le \sum_{j=0}^{k-1} \lambda^j f_1(s_1)\rho^{k-j-1} + \lambda^k \|z_0\|_2 + \frac{\lambda^k \hat{\psi}}{1-\lambda} \qquad \text{(Assumption 3.3)}$$
$$\le f_1(s_1)\rho^{-1} \sum_{j=0}^{k-1} \lambda^j \rho^{k-j} + \lambda^k \|z_0\|_2 + \frac{\lambda^k \hat{\psi}}{1-\lambda}.$$

Summing over all k, we get

$$\sum_{k=0}^{\infty} \|\mathbb{E}_{y_0}[z_k] - \mathbb{E}_\mu[\tilde{z}_k]\|_2 \le \sum_{k=0}^{\infty} \left(f_1(s_1)\rho^{-1} \sum_{j=0}^{k-1} \lambda^j \rho^{k-j} + \lambda^k \|z_0\|_2 + \frac{\lambda^k \hat{\psi}}{1-\lambda} \right)$$
$$= f_1(s_1)\rho^{-1} \sum_{k=0}^{\infty} \left(\sum_{j=0}^{k-1} \lambda^j \rho^{k-j} \right) + \frac{\|z_0\|_2}{1-\lambda} + \frac{\hat{\psi}}{(1-\lambda)^2}$$
$$= \frac{f_1(s_1)}{(1-\lambda)(1-\rho)} + \frac{\|z_0\|_2}{1-\lambda} + \frac{\hat{\psi}}{(1-\lambda)^2} \qquad \text{(Fubini-Tonelli Theorem)}$$
$$\le g(y_0).$$

(b) Using the formula for z_k and part (c) of Lemma E.1, we have

$$\mathbb{E}_{y_0}[z_k\psi(S_k)^T] - \mathbb{E}_\mu[\tilde{z}_k\psi(\tilde{S}_k)^T] = \lambda^k z_0 \mathbb{E}_{y_0}[\psi(S_k)^T] + \mathbb{E}_{y_0}\left[\sum_{j=0}^{k-1} \lambda^j \psi(S_{k-j})\psi(S_k)^T \right] - \sum_{j=0}^{\infty} \lambda^j \mathbb{E}_\mu[\psi(\tilde{S}_{k-j})\psi(\tilde{S}_k)^T]$$
$$= \lambda^k z_0 \mathbb{E}_{y_0}[\psi(S_k)^T] + \mathbb{E}_{y_0}\left[\sum_{j=0}^{k-1} \lambda^j \left(\psi(S_{k-j})\psi(S_k)^T - \mathbb{E}_\mu[\psi(\tilde{S}_{k-j})\psi(\tilde{S}_k)^T] \right) \right]$$

$$- \sum_{j=k}^{\infty} \lambda^j \mathbb{E}_{\mu}[\psi(\tilde{S}_{k-j})\psi(\tilde{S}_k)^T]$$

Taking norm both sides and using triangle inequality, we get

$$\begin{aligned}\|\mathbb{E}_{y_0}[z_k\psi(S_k)^T] - \mathbb{E}_{\mu}[\tilde{z}_k\psi(\tilde{S}_k)^T]\|_2 \leq \lambda^k\|z_0\|_2\|\mathbb{E}_{y_0}[\psi(S_k)^T]\|_2 \\ + \left\|\mathbb{E}_{y_0}\left[\sum_{j=0}^{k-1}\lambda^j\left(\psi(S_{k-j})\psi(S_k)^T - \mathbb{E}_{\mu}[\psi(\tilde{S}_{k-j})\psi(\tilde{S}_k)^T]\right)\right]\right\|_2 \\ + \sum_{j=k}^{\infty}\lambda^j\|\mathbb{E}_{\mu}[\psi(\tilde{S}_{k-j})\psi(\tilde{S}_k)^T]\|_2\end{aligned}$$

To bound the first term, we use Assumption 3.3 to get

$$\begin{aligned}\|\mathbb{E}_{y_0}[\psi(S_k)^T]\|_2 &\leq \|\mathbb{E}_{y_0}[\psi(S_k)^T] - \mathbb{E}_{\mu}[\psi(\tilde{S}_k)]\|_2 + \|\mathbb{E}_{\mu}[\psi(\tilde{S}_k)]\|_2 \\ &\leq \rho^{k-1}(f_1(s_0) + f_1(s_1)) + \|\mathbb{E}_{\mu}[\psi(\tilde{S}_k)]\|_2 \\ &\leq \rho^{k-1}(f_1(s_0) + f_1(s_1)) + \hat{\psi}\sqrt{d} \qquad \text{(Jensen's inequality and Assumption 3.2)}\end{aligned}$$

With the above bound, we have

$$\begin{aligned}\|\mathbb{E}_{y_0}[z_k\psi(S_k)^T] - \mathbb{E}_{\mu}[\tilde{z}_k\psi(\tilde{S}_k)^T]\|_2 &\leq \sum_{j=0}^{k-1}\lambda^j\left\|\mathbb{E}_{y_0}[\psi(S_{k-j})\psi(S_k)^T] - \mathbb{E}_{\mu}[\psi(\tilde{S}_{k-j})\psi(\tilde{S}_k)^T]\right\|_2 \\ &\quad + \lambda^k\|z_0\|_2\left(\rho^{k-1}(f_1(s_0) + f_1(s_1)) + \hat{\psi}\sqrt{d}\right) + \sum_{j=k}^{\infty}\lambda^j\|\mathbb{E}_{\mu}[\psi(\tilde{S}_{k-j})\psi(\tilde{S}_k)^T]\|_2 \\ &\leq \sum_{j=0}^{k-1}\lambda^j\left\|\mathbb{E}_{y_0}[\psi(S_{k-j})\psi(S_k)^T] - \mathbb{E}_{\mu}[\psi(\tilde{S}_{k-j})\psi(\tilde{S}_k)^T]\right\|_2 \\ &\quad + \lambda^k\|z_0\|_2\left(\rho^{k-1}(f_1(s_0) + f_1(s_1)) + \hat{\psi}\sqrt{d}\right) + \frac{\lambda^k\psi'}{1-\lambda} \qquad \text{(Part (b) of Lemma E.1)} \\ &\leq \sum_{j=0}^{k-1}\lambda^j f_1(s_1)\rho^{k-j-1} + \lambda^k\|z_0\|_2\left(\rho^{k-1}(f_1(s_0) + f_1(s_1)) + \hat{\psi}\sqrt{d}\right) + \frac{\lambda^k\psi'}{1-\lambda} \\ &\qquad \text{(Assumption 3.3)} \\ &\leq f_1(s_1)\rho^{-1}\sum_{j=0}^{k-1}\lambda^j\rho^{k-j} + \lambda^k\|z_0\|_2\left(\rho^{k-1}(f_1(s_0) + f_1(s_1)) + \hat{\psi}\sqrt{d}\right) + \frac{\lambda^k\psi'}{1-\lambda}.\end{aligned}$$

Summing over all k, we get

$$\begin{aligned}\sum_{k=0}^{\infty}\|\mathbb{E}_{y_0}[z_k\psi(S_k)^T] - \mathbb{E}_{\mu}[\tilde{z}_k\psi(\tilde{S}_k)^T]\|_2 &\leq \sum_{k=0}^{\infty}\left(f_1(s_1)\rho^{-1}\sum_{j=0}^{k-1}\lambda^j\rho^{k-j}\right. \\ &\quad \left. + \lambda^k\|z_0\|_2\left(\rho^{k-1}(f_1(s_0) + f_1(s_1)) + \hat{\psi}\sqrt{d}\right) + \frac{\lambda^k\psi'}{1-\lambda}\right) \\ &= f_1(s_1)\rho^{-1}\sum_{k=0}^{\infty}\left(\sum_{j=0}^{k-1}\lambda^j\rho^{k-j}\right) + \frac{\|z_0\|_2\rho^{-1}(f_1(s_0) + f_1(s_1))}{(1-\lambda\rho)} \\ &\quad + \frac{\|z_0\|_2\hat{\psi}\sqrt{d}}{(1-\lambda)} + \frac{\psi'}{(1-\lambda)^2} \\ &= \frac{f_1(s_1)}{(1-\lambda)(1-\rho)} + \frac{\|z_0\|_2\rho^{-1}(f_1(s_0) + f_1(s_1))}{(1-\lambda\rho)}\end{aligned}$$

$$+ \frac{\|z_0\|_2 \hat{\psi}\sqrt{d}}{(1-\lambda)} + \frac{\psi'}{(1-\lambda)^2} \qquad \text{(Fubini-Tonelli Theorem)}$$
$$\leq g(y_0).$$

(c) It is easy to verify that an identical argument as in the previous part can be carried out for $\mathbb{E}_{y_0}[z_k\psi(S_{k+1})^T] - \mathbb{E}_{\mu}[\tilde{z}_k\psi(\tilde{S}_{k+1})^T]$. Thus to avoid repetition, we omit the proof for this part.

(d) We bound j-th element of the vector $\Psi^T\Lambda P^m\mathcal{R}_\pi$ as follows:

$$\begin{aligned}
(\Psi^T\Lambda P^m\mathcal{R}_\pi)^2(j) &= \left(\sum_{s\in\mathcal{S}}\mu(s)\psi_j(s)\sum_{s'\in\mathcal{S}}P^m(s'|s)\mathcal{R}_\pi(s)\right)^2 \\
&\leq \left(\sum_{s\in\mathcal{S}}\mu(s)\psi_j^2(s)\right)\left(\sum_{s\in\mathcal{S}}\mu(s)\left(\sum_{s'\in\mathcal{S}}P^m(s'|s)\mathcal{R}_\pi(s)\right)^2\right) && \text{(Cauchy-Schwartz inequality)} \\
&\leq \left(\sum_{s\in\mathcal{S}}\mu(s)\psi_j^2(s)\right)\left(\sum_{s\in\mathcal{S}}\mu(s)\sum_{s'\in\mathcal{S}}P^m(s'|s)(\mathcal{R}_\pi(s))^2\right) && \text{(Jensen's inequality)} \\
&\leq \hat{\psi}^2\left(\sum_{s\in\mathcal{S}}\mu(s)(\mathcal{R}_\pi(s))^2\right) && \text{(Assumption 3.2 and Fubini-Tonelli Theorem)} \\
&\leq \hat{\psi}^2\hat{r}^2. && \text{(Assumption 3.1)}
\end{aligned}$$

Thus, the norm can be bounded as

$$\|\Psi^T\Lambda P^m\mathcal{R}_\pi\|_2 \leq \hat{\psi}\hat{r}\sqrt{d}.$$

Proceeding in a similar fashion as in part (c), we have

$$\begin{aligned}
\mathbb{E}_{y_0}[z_k\mathcal{R}(S_k,A_k)] - \mathbb{E}_{\mu}[\tilde{z}_k\mathcal{R}(\tilde{S}_k,A_k)] &= \lambda^k z_0\mathbb{E}_{y_0}[\mathcal{R}(S_k,A_k)] + \mathbb{E}_{y_0}\left[\sum_{j=0}^{k-1}\lambda^j\psi(S_{k-j})\mathcal{R}(S_k,A_k)\right] \\
&\quad - \sum_{j=0}^{\infty}\lambda^j\mathbb{E}_{\mu}[\psi(\tilde{S}_{k-j})\mathcal{R}(\tilde{S}_k,A_k)] \\
&= \lambda^k z_0\mathbb{E}_{y_0}[\mathcal{R}(S_k,A_k)] \\
&\quad + \mathbb{E}_{y_0}\left[\sum_{j=0}^{k-1}\lambda^j\left(\psi(S_{k-j})\mathcal{R}(S_k,A_k) - \mathbb{E}_{\mu}[\psi(\tilde{S}_{k-j})\mathcal{R}(\tilde{S}_k,A_k)]\right)\right] \\
&\quad - \sum_{j=k}^{\infty}\lambda^j\mathbb{E}_{\mu}[\psi(\tilde{S}_{k-j})\mathcal{R}(\tilde{S}_k,A_k)]
\end{aligned}$$

Taking norm both sides and using triangle inequality, we get

$$\begin{aligned}
\|\mathbb{E}_{y_0}[z_k\mathcal{R}(S_k,A_k)] - \mathbb{E}_{\mu}[\tilde{z}_k\mathcal{R}(\tilde{S}_k,A_k)]\|_2 &\leq \lambda^k\|z_0\|_2|\mathbb{E}_{y_0}[\mathcal{R}(S_k,A_k)]| \\
&\quad + \left\|\mathbb{E}_{y_0}\left[\sum_{j=0}^{k-1}\lambda^j\left(\psi(S_{k-j})\mathcal{R}(S_k,A_k) - \mathbb{E}_{\mu}[\psi(\tilde{S}_{k-j})\mathcal{R}(\tilde{S}_k,A_k)]\right)\right]\right\|_2 \\
&\quad + \sum_{j=k}^{\infty}\lambda^j\|\mathbb{E}_{\mu}[\psi(\tilde{S}_{k-j})\mathcal{R}(\tilde{S}_k,A_k)]\|_2
\end{aligned}$$

To bound the first term, we use Assumption 3.3 to get

$$|\mathbb{E}_{y_0}[\mathcal{R}(S_k,A_k)]| \leq |\mathbb{E}_{y_0}[\mathcal{R}(S_k,A_k)] - \mathbb{E}_{\mu}[\mathcal{R}_\pi(\tilde{S}_k)]| + |\mathbb{E}_{\mu}[\mathcal{R}_\pi(\tilde{S}_k)]|$$

$$\begin{aligned}
&\leq \rho^{k-1}(f_1(s_0)+f_1(s_1)) + |\mathbb{E}_\mu[\mathcal{R}_\pi(\tilde{S}_k)]| \\
&\leq \rho^{k-1}(f_1(s_0)+f_1(s_1)) + \hat{r} \qquad \text{(Jensen's inequality and Assumption 3.1)}
\end{aligned}$$

With the above bound, we have

$$\begin{aligned}
\|\mathbb{E}_{y_0}[z_k\mathcal{R}(S_k,A_k)] - \mathbb{E}_\mu[\tilde{z}_k\mathcal{R}(\tilde{S}_k,A_k)]\|_2 &\leq \lambda^k\|z_0\|_2(\rho^{k-1}(f_1(s_0)+f_1(s_1))+\hat{r}) \\
&\quad + \left\|\mathbb{E}_{y_0}\left[\sum_{j=0}^{k-1}\lambda^j\left(\psi(S_{k-j})\mathcal{R}(S_k,A_k) - \mathbb{E}_\mu[\psi(\tilde{S}_{k-j})\mathcal{R}(\tilde{S}_k,A_k)]\right)\right]\right\|_2 \\
&\quad + \sum_{j=k}^{\infty}\lambda^j\|\mathbb{E}_\mu[\psi(\tilde{S}_{k-j})\mathcal{R}(\tilde{S}_k,A_k)]\|_2 \\
&\leq \lambda^k\|z_0\|_2(\rho^{k-1}(f_1(s_0)+f_1(s_1))+\hat{r}) \\
&\quad + \sum_{j=0}^{k}\lambda^j\left\|\mathbb{E}_{y_0}[\psi(S_{k-j})\mathcal{R}(S_k,A_k)] - \mathbb{E}_\mu[\psi(\tilde{S}_{k-j})\mathcal{R}(\tilde{S}_k,A_k)]\right\|_2 \\
&\quad + \frac{\lambda^k\hat{\psi}\hat{r}\sqrt{d}}{1-\lambda} \\
&\leq \sum_{j=0}^{k-1}\lambda^j f_1(s_1)\rho^{k-j-1} + \lambda^k\|z_0\|_2(\rho^{k-1}(f_1(s_0)+f_1(s_1))+\hat{r}) \\
&\quad + \frac{\lambda^k\hat{\psi}\hat{r}\sqrt{d}}{1-\lambda} \qquad \text{(Assumption 3.3)} \\
&\leq f_1(s_1)\rho^{-1}\sum_{j=0}^{k-1}\lambda^j\rho^{k-j} + \lambda^k\|z_0\|_2(\rho^{k-1}(f_1(s_0)+f_1(s_1))+\hat{r}) \\
&\quad + \frac{\lambda^k\hat{\psi}\hat{r}\sqrt{d}}{1-\lambda}.
\end{aligned}$$

Similar to part (d), summing over all k, we get

$$\begin{aligned}
\sum_{k=0}^{\infty}\|\mathbb{E}_{y_0}[z_k\mathcal{R}(S_k,A_k)] - \mathbb{E}_\mu[\tilde{z}_k\mathcal{R}(\tilde{S}_k,A_k)]\|_2 &\leq \frac{f_1(s_1)}{(1-\lambda)(1-\rho)} + \frac{\|z_0\|_2\rho^{-1}(f_1(s_0)+f_1(s_1))}{(1-\lambda\rho)} \\
&\quad + \frac{\|z_0\|_2\hat{r}}{(1-\lambda)} + \frac{\hat{\psi}\hat{r}\sqrt{d}}{(1-\lambda)^2} \\
&\leq g(y_0).
\end{aligned}$$

(e) Using triangle inequality on the formula for z_k, we have

$$\begin{aligned}
\|z_k\|_2 &\leq \lambda^k\|z_0\|_2 + \sum_{j=1}^{k}\lambda^{k-j}\|\psi(S_k)\|_2 \\
&= \frac{1-\lambda^{k+1}}{(1-\lambda)}\left(\frac{(1-\lambda)\lambda^k}{1-\lambda^{k+1}}\|z_0\|_2 + \sum_{j=1}^{k}\frac{(1-\lambda)\lambda^{k-j}}{1-\lambda^{k+1}}\|\psi(S_j)\|_2\right)
\end{aligned}$$

By taking fourth power both sides, we get

$$\|z_k\|_2^4 \leq \frac{(1-\lambda^{k+1})^4}{(1-\lambda)^4}\left(\frac{(1-\lambda)\lambda^k}{1-\lambda^{k+1}}\|z_0\|_2 + \sum_{j=1}^{k}\frac{(1-\lambda)\lambda^{k-j}}{1-\lambda^{k+1}}\|\psi(S_j)\|_2\right)^4.$$

Since the weights $\frac{(1-\lambda)\lambda^{k-j}}{1-\lambda^{k+1}}$ form a probability distribution, we can apply Jensen's inequality to get

$$\begin{aligned}\|z_k\|_2^4 &\leq \frac{(1-\lambda^{k+1})^4}{(1-\lambda)^4}\left(\frac{(1-\lambda)\lambda^k}{1-\lambda^{k+1}}\|z_0\|_2^4 + \sum_{j=1}^{k}\frac{(1-\lambda)\lambda^{k-j}}{1-\lambda^{k+1}}\|\psi(S_j)\|_2^4\right)\\ &\leq \frac{1}{(1-\lambda)^3}\left(\lambda^k\|z_0\|_2^4 + \sum_{j=1}^{k}\lambda^{k-j}\|\psi(S_j)\|_2^4\right).\end{aligned}$$

Taking expectation both sides conditioned on the initial state, we have

$$\begin{aligned}\mathbb{E}_{y_0}[\|z_k\|_2^4] &\leq \frac{1}{(1-\lambda)^3}\left(\lambda^k\|z_0\|_2^4 + \sum_{j=1}^{k}\lambda^{k-j}\mathbb{E}_{y_0}[\|\psi(S_j)\|_2^4]\right)\\ &\leq \frac{1}{(1-\lambda)^3}(\|z_0\|_2^4 + f_3(s_1))\left(\sum_{j=0}^{k}\lambda^{k-j}\right) && \text{(Assumption 3.3)}\\ &\leq \frac{\|z_0\|_2^4 + f_3(s_1)}{(1-\lambda)^4}.\end{aligned}$$

(f) Recall $Y_k = (S_k, A_k, S_{k+1}, z_k) \in \mathcal{Y}$. Then, we have

$$\begin{aligned}g^2(Y_k) \leq{}& \frac{4f_1^2(S_{k+1})}{(1-\lambda)^2(1-\rho)^2} + \frac{4\|z_k\|_2^2\rho^{-2}(f_1(S_k)+f_1(S_{k+1}))^2}{(1-\lambda\rho)^2}\\ &+ \frac{4\|z_k\|_2^2\left(1+\hat{r}+\hat{\psi}\sqrt{d}\right)^2}{(1-\lambda)^2} + \frac{4(\hat{\psi}+\psi'+\hat{\psi}\hat{r}\sqrt{d})^2}{(1-\lambda)^4}. && \text{(Using } \left(\textstyle\sum_{i=1}^n a_i\right)^2 \leq n\left(\sum_{i=1}^n a_i^2\right)\text{)}\end{aligned}$$

Taking expectation both sides, conditioned on initial state y_0,

$$\begin{aligned}\mathbb{E}_{y_0}[g^2(Y_k)] \leq{}& \frac{4\mathbb{E}_{y_0}[f_1^2(S_{k+1})]}{(1-\lambda)^2(1-\rho)^2} + \frac{4\rho^{-2}\mathbb{E}_{y_0}[\|z_k\|_2^2(f_1(S_k)+f_1(S_{k+1}))^2]}{(1-\lambda\rho)^2} + \frac{4\left(1+\hat{r}+\hat{\psi}\sqrt{d}\right)^2\mathbb{E}_{y_0}[\|z_k\|_2^2]}{(1-\lambda)^2}\\ &+ \frac{4(\hat{\psi}+\psi'+\hat{\psi}\hat{r}\sqrt{d})^2}{(1-\lambda)^4}.\end{aligned}$$

Using Assumption 3.3, we can bound the first term as

$$\frac{4\mathbb{E}_{y_0}[f_1^2(S_{k+1})]}{(1-\lambda)^2(1-\rho)^2} \leq \frac{4\sqrt{f_2(s_1)}}{(1-\lambda)^2(1-\rho)^2}. \qquad \text{(Jensen's inequality)}$$

For second term, we use Cauchy-Schwartz inequality for expectations, to get

$$\begin{aligned}\frac{4\rho^{-2}\mathbb{E}_{y_0}[\|z_k\|_2^2(f_1(S_k)+f_1(S_{k+1}))^2]}{(1-\lambda\rho)^2} &\leq \frac{4\rho^{-2}\sqrt{\mathbb{E}_{y_0}[\|z_k\|_2^4]}\sqrt{\mathbb{E}_{y_0}[(f_1(S_k)+f_1(S_{k+1}))^4]}}{(1-\lambda\rho)^2}\\ &\leq \frac{8\rho^{-2}\sqrt{\mathbb{E}_{y_0}[\|z_k\|_2^4]}\sqrt{\mathbb{E}_{y_0}[f_1^4(S_k)+f_1^4(S_{k+1})]}}{(1-\lambda\rho)^2} && ((a+b)^2 \leq 2a^2+2b^2)\\ &\leq \frac{16\rho^{-2}\sqrt{\|z_0\|_2^4+f_3(s_1)}\sqrt{f_2(s_0)+f_2(s_1)}}{(1-\lambda\rho)^2(1-\lambda)^2} && \text{(Assumption 3.3)}\end{aligned}$$

For the third term, we use part (e) and Jensen's inequality to get

$$\frac{4\left(1+\hat{r}+\hat{\psi}\sqrt{d}\right)^2\mathbb{E}_{y_0}[\|z_k\|_2^2]}{(1-\lambda)^2} \leq \frac{4\left(1+\hat{r}+\hat{\psi}\sqrt{d}\right)^2\sqrt{\|z_0\|_2^4+f_3(s_1)}}{(1-\lambda)^4}. \qquad \text{(Jensen's inequality)}$$

The claim follows by combining all the bounds.

□

Define $\hat{T}(s_0,s_1) = c_\alpha^2 + \frac{\sqrt{f_3(s_0)+f_3(s_1)}}{(1-\lambda)^2} + \frac{4(f_3(s_0)+f_3(s_1))}{(1-\lambda)^2}$, $\hat{b}(s_0,s_1) = c_\alpha^2\sqrt{f_3(s_0)+f_3(s_1)} + \frac{f_3(s_0)+f_3(s_1)}{(1-\lambda)^2}$ and $\hat{g}(s_0,s_1) = \mathbb{E}_{y_0}[g^2(s_0,a_0,s_1,\psi(s_0))]$. Furthermore, for ease of notation, we will denote I as the identity matrix (infinite dimensional).

Proposition E.1. *The TD(λ) algorithm satisfies the following:*

(a) The operator $F(x_k,Y_k)$ defined in Eq. (10) has the following properties:

(1) $\|F(x_k,Y_k)\|_2 \le \|T(Y_k)\|_2\|x-x^*\|_2 + \|b(Y_k)\|_2 + \|T(Y_k)\|_2\|x^*\|_2$, where $\mathbb{E}_{y_0}[\|T(Y_k)\|_2^2] \le \hat{T}(s_0,s_1)$ and $\mathbb{E}_{y_0}[(\|b(Y_k)\|_2^2] \le \hat{b}(s_0,s_1)$.

(2) Define $\bar{F}(x) = \mathbb{E}_{Y\sim\mu}[F(x,Y)]$. Then, under Assumptions 3.1 and 3.2, $\bar{F}(x)$ exists and is given by $\bar{F}(x) = \bar{T}x + \bar{b}$.

(3) There exists a unique $\theta^* \in E_\Psi^\perp$ such that $x^* = (\bar{r}, \theta^{*T})^T$ solves $\bar{T}x + \bar{b} = 0$. Furthermore, it is also one of the solutions to the Projected-Bellman equation $\mathcal{B}_\pi^{(\lambda)}(\Psi\theta) = \Psi\theta$.

(b) There exists a solution to the Poisson equation (3) for the Markov chain $\mathcal{M}_Y$ which satisfies Assumption 2.2 with $\hat{A}_2^2(y_0) = 9\hat{g}(s_0,s_1)$ and $\hat{B}_2^2(y_0) = 2(9\|x^*\|_2^2+1)\hat{g}(s_0,s_1) + \frac{8c_\alpha^2\sqrt{f_2(s_0)+f_2(s_1)}}{(1-\rho)^2}$.

Proof. (a) (1) Since $T(Y_k)$ is partitioned in a block form, we use Lemma H.3 and the non-expansivity of the projection operator $\Pi_{2,E_\Psi^\perp}$ to get

$$\begin{aligned}
\|T(Y_k)\|_2^2 &\le c_\alpha^2 + \|\Pi_{2,E_\Psi^\perp} z_k\|_2^2 + \|\Pi_{2,E_\Psi^\perp} z_k(\psi(S_{k+1})^T - \psi(S_k)^T)\|_2^2 \\
&\le c_\alpha^2 + \|z_k\|_2^2 + \|z_k(\psi(S_{k+1})^T - \psi(S_k)^T)\|_2^2 \\
&\le c_\alpha^2 + \|z_k\|_2^2 + (\|z_k\|_2(\|\psi(S_{k+1})\|_2 + \|\psi(S_k)\|_2))^2 \\
&\le c_\alpha^2 + \|z_k\|_2^2 + 2(\|z_k\|_2^2(\|\psi(S_{k+1})\|_2^2 + \|\psi(S_k)\|_2^2)) \qquad ((\text{Using } (a+b)^2 \le 2(a^2+b^2))
\end{aligned}$$

Note that $z_{-1} = 0$ implies $z_0 = \psi(s_0)$. Taking expectation both sides, we get

$$\begin{aligned}
\mathbb{E}_{y_0}[\|T(Y_k)\|_2^2] &\le c_\alpha^2 + \mathbb{E}_{y_0}[\|z_k\|_2^2] + 2\mathbb{E}_{y_0}[\|z_k\|_2^2\|\psi(S_{k+1})\|^2] + 2\mathbb{E}_{y_0}[\|z_k\|_2^2\|\psi(S_k)\|_2^2] \\
&\le c_\alpha^2 + \mathbb{E}_{y_0}[\|z_k\|_2^2] + 2\sqrt{\mathbb{E}_{y_0}[\|z_k\|_2^4]}\sqrt{\mathbb{E}_{y_0}[\|\psi(S_{k+1})\|_2^4]} + 2\sqrt{\mathbb{E}_{y_0}[\|z_k\|_2^4]}\sqrt{\mathbb{E}_{y_0}[\|\psi(S_k)\|_2^2]} \\
&\qquad \text{(Cauchy-Schwartz for expectation)} \\
&\le c_\alpha^2 + \frac{\sqrt{\|\psi(s_0)\|_2^4 + f_3(s_1)}}{(1-\lambda)^2} + \frac{4\sqrt{\|\psi(s_0)\|_2^4 + f_3(s_1)}\sqrt{f_3(s_0)+f_3(s_1)}}{(1-\lambda)^2} \\
&\qquad \text{(Part (e) Lemma E.2 and Assumption 3.3)} \\
&\le c_\alpha^2 + \frac{\sqrt{f_3(s_0)+f_3(s_1)}}{(1-\lambda)^2} + \frac{4(f_3(s_0)+f_3(s_1))}{(1-\lambda)^2} = \hat{T}(s_0,s_1).
\end{aligned}$$

Next, we bound $b(Y_k)$

$$\|b(Y_k)\|_2^2 = c_\alpha^2\mathcal{R}^2(S_k,A_k) + \mathcal{R}^2(S_k,A_k)\|\Pi_{2,E_\Psi^\perp} z_k\|_2^2$$

Again using the non-expansivity of the projection operator $\Pi_{2,E_\Psi^\perp}$ and taking expectation, we get

$$\begin{aligned}
\mathbb{E}_{y_0}[\|b(Y_k)\|_2^2] &= c_\alpha^2\mathbb{E}_{y_0}[\mathcal{R}^2(S_k,A_k)] + \mathbb{E}_{y_0}[\mathcal{R}^2(S_k,A_k)\|z_k\|_2^2] \\
&\le c_\alpha^2\sqrt{\mathbb{E}_{y_0}[\mathcal{R}^4(S_k,A_k)]} + \sqrt{\mathbb{E}_{y_0}[\mathcal{R}^4(S_k,A_k)]}\sqrt{\mathbb{E}_{y_0}[\|z_k\|_4^2]} \\
&\qquad \text{(Cauchy-Schwartz for expectation)} \\
&\le c_\alpha^2\sqrt{f_3(s_0)+f_3(s_1)} + \frac{\sqrt{f_3(s_0)+f_3(s_1)}\sqrt{\|\psi(s_0)\|_2^4 + f_3(s_1)}}{(1-\lambda)^2}. \\
&\qquad \text{(Part (e) Lemma E.2 and Assumption 3.3)} \\
&\le c_\alpha^2\sqrt{f_3(s_0)+f_3(s_1)} + \frac{f_3(s_0)+f_3(s_1)}{(1-\lambda)^2} = \hat{b}(s_0,s_1). \qquad (29)
\end{aligned}$$

Combining both the bounds, we have

$$\|F(x_k, Y_k)\|_2 \leq \|T(Y_k)\|_2\|x - x^*\|_2 + \|b(Y_k)\|_2 + \|T(Y_k)\|_2\|x^*\|_2,$$

where $\mathbb{E}_{y_0}[\|T(Y_k)\|_2^2] \leq \hat{T}(s_0, s_1)$ and $\mathbb{E}_{y_0}[\|b(Y_k)\|_2^2] \leq \hat{b}(s_0, s_1)$.

(2) From Lemma 3.1, the stationary expectations of $T(Y_k)$ and $b(Y_k)$ are finite. Thus,

$$\mathbb{E}_\mu[F(\tilde{Y}_k, x)] = \bar{T}x + \bar{b}.$$

(3)
- $\nexists\ \theta \in \mathbb{R}^d$ such that $\psi(s)^T\theta = 1,\ \forall s \in \mathcal{S}$: In this case, $E_\Psi^\perp \equiv \mathbb{R}^d$. Lemma 3.2 implies that all the eigenvalues of $\Psi^T\Lambda(P^{(\lambda)} - I)\Psi$ are strictly negative, immediately suggesting that $\Psi^T\Lambda(P^{(\lambda)} - I)\Psi$ is invertible. Thus, there exists a unique solution θ^*

$$-\frac{\bar{r}}{(1-\lambda)}\Psi^T\mu + \Psi^T\Lambda(P^{(\lambda)} - I)\Psi\theta^* + \Psi^T\Lambda\mathcal{R}^{(\lambda)} = 0.$$

- $\exists\ \theta_e \in \mathbb{R}^d$ such that $\psi(s)^T\theta_e = 1,\ \forall s \in \mathcal{S}$: Note that due to linear independence of columns of Ψ and the irreducibility of P, θ_e is the unique left and right eigenvector of $\Psi^T\Lambda(P^{(\lambda)} - I)\Psi$ corresponding to eigenvalue 0. This implies that all the other generalized eigenvectors of $\Psi^T\Lambda(P^{(\lambda)} - I)\Psi$ are perpendicular to θ_e and hence, they span $E_\Psi^\perp$. Furthermore, note that

$$\theta_e^T\left(\Psi^T\Lambda\mathcal{R}^{(\lambda)} - \frac{\bar{r}}{(1-\lambda)}\Psi^T\mu\right) = \mu^T\mathcal{R}^{(\lambda)} - \frac{\bar{r}}{(1-\lambda)} = 0.$$

Thus, the vector $\Psi^T\Lambda\mathcal{R}^{(\lambda)} - \frac{\bar{r}}{(1-\lambda)}\Psi^T\mu$ is perpendicular to θ_e and therefore lies in $E_\Psi^\perp$. By the properties of generalized eigenvectors, it is easy to verify that there exists unique $\theta^* \in E_\Psi^\perp$ which satisfies

$$-\frac{\bar{r}}{(1-\lambda)}\Psi^T\mu + \Psi^T\Lambda(P^{(\lambda)} - I)\Psi\theta^* + \Psi^T\Lambda\mathcal{R}^{(\lambda)} = 0.$$

Note that $\Pi_{2,E_\Psi^\perp}\Psi^T\Lambda(P^{(\lambda)} - I)\Psi\theta = \Psi^T\Lambda(P^{(\lambda)} - I)\Psi\theta$, for all $\theta \in E_\Psi^\perp$. Consider the expression $\bar{T}x^* + \bar{b}$. Expanding the matrix $\bar{T}$, we get

$$\bar{T}x^* + \bar{b} = \begin{bmatrix} -\bar{r} + \bar{r} \\ -\frac{\bar{r}}{(1-\lambda)}\Pi_{2,E_\Psi^\perp}\Psi^T\mu + \Pi_{2,E_\Psi^\perp}\Psi^T\Lambda(P^{(\lambda)} - I)\Psi\theta^* + \Pi_{2,E_\Psi^\perp}\Psi^T\Lambda\mathcal{R}^{(\lambda)} \end{bmatrix} = 0$$

Thus, x^* is the unique solution to $\bar{T}x^* + \bar{b} = 0$ in the subspace $E_\Psi^\perp$. Furthermore, rearranging the terms in $\bar{T}x^* + \bar{b} = 0$, we get

$$\Psi^T\Lambda\Psi\theta^* = -\frac{\bar{r}}{(1-\lambda)}\Psi^T\mu + \Psi^T\Lambda P^{(\lambda)}\Psi\theta^* + \Psi^T\Lambda\mathcal{R}^{(\lambda)}.$$

Multiplying both sides by $\Psi(\Psi^T\Lambda\Psi)^{-1}$ ($\Psi^T\Lambda\Psi$ is a $d \times d$ invertible matrix), we get

$$\Psi\theta^* = \Psi(\Psi^T\Lambda\Psi)^{-1}\left(\Psi^T\Lambda\mathcal{R}^{(\lambda)} + \Psi^T\Lambda P^{(\lambda)}\Psi\theta^* - \frac{\bar{r}}{(1-\lambda)}\Psi^T\mu\right)$$
$$\Psi\theta^* = \Pi_{\Lambda,\Psi}\mathcal{B}_\pi^{(\lambda)}(\Psi\theta^*).$$

Thus, θ^* is also one of the solutions for the Projected-Bellman equation for TD(λ).

(b) We will use similar arguments as in Lemma 1 of Chapter 2, Part 2 from Benveniste et al. (2012) to show the existence of a solution to the Poisson equation. Define $V_x(y)$ for all $y = (s, a, s', z) \in \mathcal{Y}$ as follows:

$$V_x(y) = \left(\sum_{k=0}^{\infty}(\mathbb{E}_y[T(Y_k)] - \bar{T})\right)x + \sum_{k=0}^{\infty}(\mathbb{E}_y[b(Y_k)] - \bar{b})$$

Then, using Lemma E.2 we can bound each infinite summation as follows:

$$\begin{aligned}
\left\|\sum_{k=0}^{\infty}(\mathbb{E}_y[T(Y_k)]-\bar{T})\right\|_2 &\leq \sum_{k=0}^{\infty}\left\|\mathbb{E}_y[T(Y_k)]-\bar{T}\right\|_2 \\
&\leq \sum_{k=0}^{\infty}\Bigg(\|\Pi_{2,E_\Psi^\perp}\left(\mathbb{E}_y[z_k]-\mathbb{E}_\mu[\tilde{z}_k]\right)\|_2+\left\|\Pi_{2,E_\Psi^\perp}\left(\mathbb{E}_y[z_k\psi(S_k)^T]-\mathbb{E}_\mu[\tilde{z}_k\psi(\tilde{S}_k)^T\right)\right]\right\|_2 \\
&\quad+\left\|\Pi_{2,E_\Psi^\perp}\left(\mathbb{E}_y[z_k\psi(S_{k+1})^T]-\mathbb{E}_\mu[\tilde{z}_k\psi(\tilde{S}_{k+1})^T]\right)\right\|_2\Bigg) \\
&\qquad\qquad\text{(Lemma H.3 and triangle inequality)} \\
&\leq 3g(y). \qquad\qquad\text{(Lemma E.2)}
\end{aligned}$$

Next, for the second summation, we have

$$\begin{aligned}
\left\|\sum_{k=0}^{\infty}(\mathbb{E}_y[b(Y_k)]-\bar{b})\right\|_2 &\leq \sum_{k=0}^{\infty}\left\|\mathbb{E}_y[b(Y_k)]-\bar{b}\right\|_2 \\
&\leq \sum_{k=0}^{\infty}\Bigg(c_\alpha|\mathbb{E}_y[\mathcal{R}(S_k,A_k)]-\mathbb{E}_\mu[\mathcal{R}(\tilde{S}_k,A_k)]| \\
&\quad+\left\|\Pi_{2,E_\Psi^\perp}\left(\mathbb{E}_y[z_k\mathcal{R}(S_k,A_k)]-\mathbb{E}_\mu[\tilde{z}_k\mathcal{R}(\tilde{S}_k,A_k)]\right)\right\|_2\Bigg) \\
&\leq \sum_{k=0}^{\infty}c_\alpha\rho^k(f_1(s)+f_1(s'))+g(y) \qquad\text{(Assumption 3.3 and Lemma E.2)} \\
&= \frac{c_\alpha(f_1(s)+f_1(s'))}{1-\rho}+g(y).
\end{aligned}$$

Thus, both the series are convergent. Furthermore, note that Assumption 3.3 implies that $\mathbb{E}_y[f_1(S_k)] \leq \sqrt[4]{f_2(s)+f_2(s')}$, for all $y\in\mathcal{Y}$ and $k\geq 0$. Thus, following dominated convergence theorem,

$$\begin{aligned}
\mathbb{E}_y[V_x(Y_1)] &= \mathbb{E}_y\left[\left(\sum_{k=1}^{\infty}(\mathbb{E}_{Y_1}[T(Y_k)]-\bar{T})\right)x+\sum_{k=1}^{\infty}(\mathbb{E}_{Y_1}[b(Y_k)]-\bar{b})\right] \\
&= \left(\sum_{k=1}^{\infty}(\mathbb{E}_y[\mathbb{E}_{Y_1}[T(Y_k)]-\bar{T}])\right)x+\sum_{k=1}^{\infty}(\mathbb{E}_y[\mathbb{E}_{Y_1}[b(Y_k)]-\bar{b}]) \\
&= \left(\sum_{k=1}^{\infty}(\mathbb{E}_y[T(Y_k)]-\bar{T}])\right)x+\sum_{k=1}^{\infty}(\mathbb{E}_y[b(Y_k)]-\bar{b}]) \\
&= V_y(x)-((T(y)-\bar{T})x+b(y)-\bar{b}).
\end{aligned}$$

The claim follows. Now to show bounded expectations, we use part (f) of Lemma E.2 and the fact that $z_0=\psi(s_0)$, to get

$$\begin{aligned}
\hat{A}_2^2(y_0) &= \mathbb{E}_{Y_0=(s_0,a_0,s_1,z_0)}\left[\left\|\sum_{k=0}^{\infty}(\mathbb{E}_y[T(Y_k)]-\bar{T})\right\|_2^2\right]\leq 9\hat{g}(s_0,s_1), \\
\hat{B}_2^2(y_0) &= \mathbb{E}_{Y_0=(s_0,a_0,s_1,z_0)}\left[\|V_{x^*}(Y_k)\|_2^2\right]\leq \mathbb{E}_{Y_0=(s_0,a_0,s_1,z_0)}\left[\left(3g(Y_k)x^*+\frac{c_\alpha(f_1(S_k)+f_1(S_{k+1}))}{1-\rho}+g(Y_k)\right)^2\right], \\
&\leq 2(9\|x^*\|_2^2+1)\hat{g}(s_0,s_1)+\frac{8c_\alpha^2\sqrt{f_2(s_0)+f_2(s_1)}}{(1-\rho)^2}.
\end{aligned}$$

□

E.3 Proof of Lemma 3.2

Proof. The proof largely involves similar arguments as in Lemma 2 in Zhang et al. (2021) but we will also need Lemma E.3 to adapt the infinite state space. Since $P^{(\lambda)}$ is an irreducible and aperiodic Markov kernel, for any non-zero $\theta \in E_\Psi^\perp$, by Lemma E.4 we have

$$\theta^T(\Psi^T\Lambda\Psi - \Psi^T\Lambda P^{(\lambda)}\Psi)\theta > 0.$$

Consider the set $\{\theta \in E_\Psi^\perp | \|\theta\|_2 = 1\}$. Note that this set is compact and closed, thus by the extreme value theorem, we have

$$\Delta := \min_{\theta \in E_\Psi^\perp, \|\theta\|_2 = 1} \theta^T(\Psi^T\Lambda\Psi - \Psi^T\Lambda P^{(\lambda)}\Psi)\theta > 0.$$

By Lemma 3.1, in steady state $\mathbb{E}_\mu[T(Y_k)]$ is given by

$$\bar{T} = \begin{bmatrix} -c_\alpha & 0 \\ -\frac{1}{(1-\lambda)}\Pi_{2,E_\Psi^\perp}\Psi^T\mu & \Pi_{2,E_\Psi^\perp}(\Psi^T\Lambda\Psi - \Psi^T\Lambda P^{(\lambda)}\Psi) \end{bmatrix}.$$

Thus, the minimization problem $\min_{x \in \mathbb{R}\times E_\Psi^\perp, \|x\|_2=1} -x^T\bar{T}x$ can be written as

$$\min_{\theta \in E_\Psi^\perp, r\in\mathbb{R}, r^2+\|\theta\|_2^2=1} c_\alpha r^2 + \frac{r}{1-\lambda}\theta^T\Pi_{2,E_\Psi^\perp}\Psi^T\mu + \theta^T\Pi_{2,E_\Psi^\perp}(\Psi^T\Lambda\Psi - \Psi^T\Lambda P^{(\lambda)}\Psi)\theta.$$

Since $\theta \in E_\Psi^\perp$, $\theta^T\Pi_{2,E_\Psi^\perp} = (\Pi_{2,E_\Psi^\perp}\theta)^T = \theta$, we have

$$\min_{\theta \in E_\Psi^\perp, r\in\mathbb{R}, r^2+\|\theta\|_2^2=1} c_\alpha r^2 + \frac{r}{1-\lambda}\theta^T\Psi^T\mu + \theta^T(\Psi^T\Lambda\Psi - \Psi^T\Lambda P^{(\lambda)}\Psi)\theta.$$

First, we bound the second term.

$$\begin{aligned}\left|\frac{r}{1-\lambda}\theta^T\Psi^T\mu\right| &= \frac{|r|}{1-\lambda}\left|\theta^T\Psi^T\mu\right| \\ &\le \frac{|r|}{1-\lambda}\|\theta\|_2\|\Psi^T\mu\|_2. \qquad \text{(Cauchy-Schwartz Inequality)}\end{aligned}$$

Since $\|\Psi^T\mu\|_2^2 = \sum_{i=1}^d \left(\sum_{s\in\mathcal{S}}\mu(s)\psi_i(s)\right)^2$ which by Jensen's inequality can be bounded as

$$\begin{aligned}\|\Psi^T\mu\|_2^2 &\le \sum_{i=1}^d \sum_{s\in\mathcal{S}} \mu(s)\psi_i^2(s) \\ &\le d\hat{\psi}^2.\end{aligned}$$

Thus,

$$\left|\frac{r}{1-\lambda}\theta^T\Psi^T\mu\right| \le \frac{\hat{\psi}|r|\|\theta\|_2\sqrt{d}}{1-\lambda}, \quad \forall r \in \mathbb{R},\ \theta \in E_\Psi^\perp.$$

and

$$\theta^T(\Psi^T\Lambda\Psi - \Psi^T\Lambda P^{(\lambda)}\Psi)\theta \ge \Delta\|\theta\|_2^2, \quad \forall\theta \in E_\Psi^\perp.$$

Combining all the bounds, we get

$$\begin{aligned}&\min_{\theta \in E_\Psi^\perp, r\in\mathbb{R}, r^2+\|\theta\|_2^2=1} c_\alpha r + \frac{r}{1-\lambda}\theta^T\Psi^T\mu + \theta^T(\Psi^T\Lambda\Psi - \Psi^T\Lambda P^{(\lambda)}\Psi)\theta \\ &\ge \min_{\theta \in E_\Psi^\perp, r\in\mathbb{R}, r^2+\|\theta\|_2^2=1} c_\alpha r^2 - \frac{\hat{\psi}|r|\|\theta\|_2\sqrt{d}}{1-\lambda} + \Delta\|\theta\|_2^2\end{aligned}$$

$$
\begin{aligned}
&= \min_{r\in[-1,1]} c_\alpha |r|^2 - \frac{\sqrt{d}\hat{\psi}|r|\sqrt{1-|r|^2}}{1-\lambda} + \Delta(1-r^2) \\
&= \min_{z\in[0,1]} c_\alpha z - \frac{\sqrt{d}\hat{\psi}\sqrt{z(1-z)}}{1-\lambda} + \Delta(1-z) \\
&= \Delta + \min_{z\in[0,1]} (c_\alpha - \Delta)z - \frac{\sqrt{d}\hat{\psi}\sqrt{z(1-z)}}{1-\lambda}.
\end{aligned}
$$

When $c_\alpha \geq \Delta + \sqrt{\frac{d^2\hat{\psi}^4}{\Delta^2(1-\lambda)^4} - \frac{d\hat{\psi}^2}{(1-\lambda)^2}}$, we have

$$
\begin{aligned}
\min_{z\in[0,1]} (c_\alpha - \Delta)z - \frac{\sqrt{d}\hat{\psi}\sqrt{z(1-z)}}{1-\lambda} &= \frac{1}{2}\left((c_\alpha - \Delta) - \sqrt{(c_\alpha-\Delta)^2 + \frac{d^2\hat{\psi}^2}{(1-\lambda)^2}}\right) \\
&\geq \frac{1}{2}\left(\sqrt{\frac{d^2\hat{\psi}^4}{\Delta^2(1-\lambda)^4} - \frac{d\hat{\psi}^2}{(1-\lambda)^2}} - \frac{d\hat{\psi}^2}{\Delta(1-\lambda)^2}\right). \\
&\geq -\frac{\Delta}{2}
\end{aligned}
$$

where for the last inequality we used the following fact

$$
\sqrt{\frac{x^2}{\Delta^2} - x} - \frac{x}{\Delta} \geq -\Delta \quad \forall x. \qquad (x = \Delta^2 \text{ is the minimizer})
$$

Therefore, it follows that

$$
\min_{x\in\mathbb{R}\times E_\Psi^\perp, \|x\|_2=1} -x^T T x \geq \frac{\Delta}{2}.
$$

□

E.4 Proof of Theorem 3.1

Theorem E.1. *Consider the iterates $\{\theta_k, \bar{r}_k\}_{k\geq 0}$ generated by Algorithm 1 under Assumption 3.1-3.3 and $c_\alpha \geq \Delta + \sqrt{\frac{d^2\hat{\psi}^4}{\Delta^2(1-\lambda)^4} - \frac{d\hat{\psi}^2}{(1-\lambda)^2}}$.*

(a) When $\alpha_k \equiv \alpha \leq 1$, then for all $k \geq 0$:

$$
\mathbb{E}[(\bar{r}_k - \bar{r})^2 + \|\theta_k - \theta^*\|_2^2] \leq \varphi_{V,0}\exp\left(-\frac{\Delta\alpha k}{2}\right) + 3\hat{C}_V(s_0,s_1)\alpha + \frac{12\hat{C}_V(s_0,s_1)\alpha}{\Delta}.
$$

(b) When $\xi = 1$, $\alpha > \frac{1}{\Delta}$ and $K \geq \max\{\alpha, 2\}$, then for all $k \geq 0$:

$$
\mathbb{E}[(\bar{r}_k - \bar{r})^2 + \|\theta_k - \theta^*\|_2^2] \leq \varphi_{V,0}\left(\frac{K}{k+K}\right)^{\frac{\Delta\alpha}{2}} + \frac{\hat{C}_V(s_0,s_1)\alpha}{k+K} + \frac{4(6+2\Delta)\hat{C}_V(s_0,s_1)e\alpha^2}{\left(\frac{\Delta\alpha}{2}-1\right)(k+K)}.
$$

Proof. Since there is no martingale noise in the algorithm $A_3 = B_3 = 0$. From Proposition E.1, we have $\hat{A}_1^2(y_0) = \hat{T}(s_0,s_1)$ and $\hat{A}_2^2(y_0) = 9\hat{g}(s_0,s_1)$ which gives

$$
\hat{A}(y_0) = \hat{A}_1^2(y_0) + \hat{A}_2^2(y_0) + A_3^2 = \hat{T}(s_0,s_1) + 9\hat{g}(s_0,s_1).
$$

Next, $\mathbb{E}_{y_0}[(\|b(Y_k)\|_2^2] \leq \hat{b}(s_0,s_1)$, we have

$$
\mathbb{E}_{y_0}[(\|b(Y_k)\|_2 + \|T(Y_k)\|_2\|x^*\|_2)^2] \leq 2\hat{b}(s_0,s_1) + 2\hat{T}(s_0,s_1)\|x^*\|_2^2.
$$

Combining above with $B_2^2(y_0)$, we get

$$\begin{aligned}\hat{B}(y_0)^2 &= \hat{B}_1^2(y_0) + \hat{B}_2^2(y_0) + B_3^2 \\ &= 2\hat{b}(s_0,s_1) + 2\hat{g}(s_0,s_1) + 2\|x^*\|_2^2\left(9\hat{g}(s_0,s_1) + \hat{T}(s_0,s_1)\right) + \frac{8c_\alpha^2\sqrt{f_2(s_0)+f_2(s_1)}}{(1-\rho)^2}.\end{aligned}$$

Let $\max_{x\in\mathcal{X}}\|x\|_2 \le M/2$. Then, $\hat{C}_V(s_0,s_1) = \hat{C}(y_0) = \hat{A}(y_0)M^2 + \hat{B}(y_0)$. Since $\|\cdot\|_c = \|\cdot\|_s = \|\cdot\|_2$, we have

$$\varphi_1 = \frac{uL_su_{2s}u_{cs}^2}{l_{2s}} = 1;\ \varphi_{V,0} = (\bar{r}_0 - r^*)^2 + \|\theta_0 - \theta^*\|_c^2 + 2\hat{C}_V(s_0,s_1).$$

□

E.5 Challenges in the infinite state space

Consider a birth-death chain. Let the state space be given as $\mathcal{S} = \{s_i\}_{i\ge 0}$ with the transition kernel $P(s_{i+1}|s_i) = p$ and $P(s_{i-1}|s_i) = 1-p$, where $p < 1/2$. Furthermore, $P(s_0|s_0) = 1-p$. It is well known that for $p < 1/2$, this chain is positive recurrent and the stationary distribution is given by $\mu(s_i) = \frac{(1-2p)p^i}{(1-p)^{i+1}}$. For simplicity, we will consider the setting when $\lambda = 0$, which implies $P^{(0)} = P$. Consider a sequence of functions $\{V_j\}_{j\ge 1}$ in the set $\{V|\sum_{s\in\mathcal{S}}V(s) = 0, \sum_{s\in\mathcal{S}}V^2(s) = 1\}$ that satisfy the following:

$$V_j(s_i) = \begin{cases} \frac{1}{\sqrt{2}}, & i = j \\ -\frac{1}{\sqrt{2}}, & i = j+1 \\ 0, & \text{otherwise.}\end{cases}$$

Then, we have the following:

$$\begin{aligned}V_j^T\Lambda V_j - V_j^T\Lambda P V_j &= \frac{1}{2}\mathbb{E}_\mu[(V_j(S_{k+1}) - V_j(S_k))^2] \\ &= \mu(s_{j-1})p\left(\frac{1}{\sqrt{2}}\right)^2 + \mu(s_j)p\left(\sqrt{2}\right)^2 + \mu(s_{j+1})(1-p)\left(\sqrt{2}\right)^2 + \mu(s_{j+2})(1-p)\left(\frac{1}{\sqrt{2}}\right)^2 \\ &= \frac{(1-2p)p^j}{(1-p)^j}\left(\frac{1}{2} + \frac{4(1-2p)p}{(1-p)} + \frac{(1-2p)p^2}{2(1-p)^2}\right).\end{aligned}$$

Note that $p < 1/2$, therefore $\lim_{j\to\infty} V_j^T\Lambda V_j - V_j^T\Lambda P V_j \to 0$. Thus,

$$\inf_V V_j^T\Lambda V_j - V_j^T\Lambda P V_j = 0.$$

Observe that in the above example the vector has infinite dimension, thus the state at which the function value has a variation for the first time can drift to infinity. However, by using linear function approximation with a finite number of columns, we are essentially restricting the function in a finite-dimensional setting. More concretely, in Lemma E.3 we establish that for any function given by a linear combination of columns of Ψ, the point of variation in the value of the function cannot drift to infinity.

E.6 Auxiliary Lemmas for TD(λ)

Lemma E.3. *Let $\mathcal{S} = \{s_1, s_2, s_3, \dots\}$ be an indexing of the state space such that $\psi(s_i)^T$ is the i-th row in Ψ. Then, under the Assumption 3.2, there exists a finite N such that for all $\theta \in E_\Psi^\perp$ the following relation holds*

$$\sum_{j=1}^d \theta(j)\psi_j(s_N) \ne \sum_{j=1}^d \theta(j)\psi_j(s_i),\ 1 \le i \le N-1. \tag{30}$$

Proof. From Lemma H.2, there exists N_1 such that the span of the first N_1 rows of Ψ is $\mathbb{R}^d$. Therefore, there exists a set of d vectors that are linearly independent. Denote these row vectors by $\{\psi(s_{i_1}), \psi(s_{i_2}), \dots, \psi(s_{i_d})\}$ and construct a matrix $\hat{\Psi}_d$ by concatenating these row vectors. Let $e \in \mathbb{R}^d$ be the vector of all ones. Now, we have two cases: (i) E_Ψ is non-empty, or (ii) E_Ψ is empty. We will consider these two cases separately.

- E_Ψ is non-empty: Since $\hat{\Psi}_d$ is full rank and θ_e is unique, $\hat{\Psi}_d\theta \neq e$ for all $\theta \in E_\Psi^\perp$. The claim follows immediately.
- E_Ψ is empty: In this case $E_\Psi^\perp \equiv \mathbb{R}^d$. Let θ'_e be the vector for which we have $\Psi_d\theta'_e = e$. Then, from Assumption 3.2, there exists a finite N_2 such that $\sum_{j=1}^d \theta'_e(j)\psi_j(s_{N_2}) \neq 1$. Furthermore, $\Psi_d\theta \neq e$ for any $\theta \neq \theta'_e$. Thus, for all $\theta \in \mathbb{R}^d$, $\sum_{j=1}^d \theta(j)\psi_j(s_N) \neq \sum_{j=1}^d \theta(j)\psi_j(s_i)$, where $N = \max\{N_1, N_2\}$.

□

Lemma E.4. *Let P be the transition kernel for an irreducible and aperiodic Markov chain. Then, for any $\theta \in E_\Psi^\perp$, the following is true:*

$$\theta^T(\Psi^T\Lambda\Psi - \Psi^T\Lambda P\Psi)\theta > 0. \tag{31}$$

Proof. Let $\mathcal{V}(s_i) = \psi(s_i)^T\theta$ be a non-constant function of the states, where $\psi(s_i)^T$ is the i-th row of Ψ. Note that due to Lemma E.3, there exists a finite N where $\mathcal{V}(s_N) \neq \mathcal{V}(s_{N-1})$. Since the Markov chain is irreducible and $\mathcal{V}(\cdot)$ is a non-constant function of time, we have

$$\begin{aligned}
0 &< \frac{1}{2}\sum_{i=1}^\infty \mu(s_i)\sum_{s\in\mathcal{S}} P(s|s_i)(\mathcal{V}(s_i) - \mathcal{V}(s))^2 && (P(s_{i+1}|s_i) > 0 \text{ by the construction of } \Psi)\\
&= \sum_{i=1}^\infty \mu(s_i)\left(\mathcal{V}^2(s_i) - \mathcal{V}(s_i)\sum_{s\in\mathcal{S}} P(s|s_i)\mathcal{V}(s)\right)\\
&= \theta^T(\Psi^T\Lambda\Psi - \Psi^T\Lambda P\Psi)\theta.
\end{aligned}$$

□

F PROOF OF THE TECHNICAL RESULTS IN SECTION 4

F.1 Properties of SCBCD

Proposition F.1. *The SCBCD algorithm has the following properties:*

(a) The operator $F(x,i)$, satisfies: $\|F(x,i)\|_2 \leq L\|x - x^*\|_2,\ \forall i \in \mathcal{S}$.

(b) There exists a solution to the Poisson equation (3) for the Markov chain $\mathcal{M}_U$ which satisfies Assumption 2.2 with $A_2 = \max\{L, 1\}$ and $B_2 = 0$.

(c) The noise sequence M_k is a martingale difference sequence and satisfies: $\|M_k\|_2 \leq C_1\|x - x^*\|_2 + C_2$.

Proof. (a) Using L-smoothness of the $f(x)$ and the fact that $\nabla f(x^*) = 0$, we get

$$\begin{aligned}
\| - U_i\nabla_i f(x)\|_2 &= \| - U_iU_i^T\nabla f(x) - U_iU_i^T\nabla f(x^*)\|_2\\
&\leq L\|U_iU_i^T\|_2\|x - x^*\|_2\\
&= L\|x - x^*\|_2. && (\|U_iU_i^T\|_2 = 1)
\end{aligned}$$

(b) Note that $\mathbb{E}_i[G(Y)] = G(i \bmod p + 1)$ for any $G : \mathcal{S} \to \mathbb{R}^d$. Thus, we can write Poisson equation as

$$V_x(i) = -U_iU_i^T\nabla f(x) + V_x(i \bmod p + 1) + \frac{1}{p}\nabla f(x).$$

Set $V_x(1) = 0$, then we have

$$V_x(i) = -\sum_{k=0}^{p-i}\left(U_{k+1}U_{k+1}^T\nabla f(x) - \frac{1}{p}\nabla f(x)\right),\ \forall i \in \mathcal{S}/\{1\}$$

Note that $U_k U_k^T$ is a also diagonal matrix. Hence, $\sum_{k=0}^{p-i}\left(U_{k+1}U_{k+1}^T - \frac{1}{p}I_d\right)$ is a diagonal matrix with entries $\frac{i-1}{p}$ in the first i places and $-\frac{1}{p}$ in the remaining places. This implies that $\left\|\sum_{k=0}^{p-i}\left(U_{k+1}U_{k+1}^T - \frac{1}{p}I_d\right)\right\|_2 \leq \frac{i-1}{p} \leq 1$. Thus, for all $i \in \mathcal{S}$, we have

$$\begin{aligned}
\|V_x(i) - V_y(i)\|_2 &= \left\|\sum_{k=0}^{p-i} U_{k+1}U_{k+1}^T\nabla f(x) - \frac{1}{p}\nabla f(x) - U_{k+1}U_{k+1}^T\nabla f(y) + \frac{1}{p}\nabla f(y)\right\|_2 \\
&\leq \left\|\sum_{k=0}^{p-i}\left(U_{k+1}U_{k+1}^T - \frac{1}{p}I_d\right)\right\|_2 \|\nabla f(x) - \nabla f(y)\|_2 \\
&\leq \|\nabla f(x) - \nabla f(y)\|_2 \\
&\leq L\|x-y\|_2 && \text{(Smoothness of } f(x)\text{)}
\end{aligned}$$

Additionally, $\nabla f(x^*) = 0$ implies $V_{x^*}(i) = 0,\ \forall i \in \mathcal{S}$.

(c) Using Assumption 4.1, we have

$$\begin{aligned}
\mathbb{E}[M_k|\mathcal{F}_k] &= \mathbb{E}[U_{i(k)}w_k|\mathcal{F}_k] \\
&= U_{i(k)}\mathbb{E}[w_k|\mathcal{F}_k] && (U_{i(k)} \text{ is deterministic}) \\
&= 0.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\|M_k\|_2 &\leq \|U_{i(k)}w_k\|_2 \\
&\leq \|w_k\|_2 && (\|U_i\|_2 = 1) \\
&\leq C_1\|x_k - x^*\| + C_2.
\end{aligned}$$

□

F.2 Proof of Theorem 4.1

From Proposition F.1, we have

$$A_G = (L + C_1 + 1)^2;\ \ B_G = C_2^2;$$
$$\varrho_{G,0} = 2(1 + 2(L + C_1 + 1)^2\max\{L,1\})\|x_0 - x^*\|_c^2 + 4C_2^2\max\{L,1\};\ \ \varrho_{G,1} = \max\{L,1\}.$$

G PROOF OF TECHNICAL RESULTS FOR Q-LEARNING IN SECTION B

G.1 Proof of Proposition B.1

Proof. (a) (1) Recall that $Q^*(s,a) = \mathcal{R}(s,a) + \gamma\sum_{s'\in\mathcal{S}}P(s'|s,a)\max_{a'\in\mathcal{A}}Q^*(s',a')$. Thus, we have

$$\begin{aligned}
\|F(Q,y)\|_\infty &\leq \left\|\gamma\sum_{s'\in\mathcal{S}}P(s'|s,a)\left(\max_{a'\in\mathcal{A}}Q(s',a') - \max_{a'\in\mathcal{A}}Q^*(s',a')\right) - Q(s,a) + Q^*(s,a)\right\|_\infty + \|Q^*(s,a)\|_\infty \\
&\leq 2\|Q - Q^*\|_\infty + \|Q^*\|_\infty.
\end{aligned}$$

Similarly, for any Q_1 and Q_2 and $y \in \mathcal{Y}$ we have

$$\begin{aligned}
\|F(Q_1,y) - F(Q_2,y)\|_\infty &\leq \left\|\gamma\sum_{s'\in\mathcal{S}}P(s'|s,a)\left(\max_{a'\in\mathcal{A}}Q_1(s',a') - \max_{a'\in\mathcal{A}}Q_2(s',a')\right) - Q_1(s,a) + Q_2(s,a)\right\|_\infty \\
&\leq 2\|Q_1 - Q_2\|_\infty.
\end{aligned}$$

(2) Using the Markov property, we have for any $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ and (s,a):

$$\begin{aligned}
&\mathbb{E}_{S_k\sim\mu_b}[F(Q,(S_k,A_k))(s,a)] \\
&\qquad = \mathbb{E}_{S_k\sim\mu_b}\left[\mathbb{1}\{S_k=s,A_k=a\}\left(\mathcal{R}(s,a)+\gamma\sum_{s'\in\mathcal{S}}P(s'|s,a)\max_{a'\in\mathcal{A}}Q(s',a')-Q(s,a)\right)\right] \\
&\qquad = \mu_b(s)\pi_b(a|s)(\mathcal{B}(Q)(s,a)-Q(s,a)).
\end{aligned}$$

Thus, $\bar{F}(Q) = \Lambda(\mathcal{B}(Q)-Q)$.

(3) Since Q^* is the solution to the Bellman equation, we have

$$\begin{aligned}
\bar{F}(Q^*) &= \Lambda(\mathcal{B}(Q^*)-Q^*) \\
&= 0.
\end{aligned}$$

Hence, Q^* is a solution to the equation $\bar{F}(Q^*)=0$. The uniqueness of the solution is immediate from the fact that $\mu_b(s)\pi_b(a|s)>0$ and Q^* is the unique solution to $\mathcal{B}(Q)-Q=0$.

(b) Fix a state $y_0=(s_0,a_0)\in\mathcal{Y}$ and define $\tau=\min\{n>0: Y_n=y_0\}$ and $\mathbb{E}_y[\cdot]=\mathbb{E}[\cdot|Y_0=y]$, then for all $y\in\mathcal{Y}$

$$V_Q(y) = \mathbb{E}_y\left[\sum_{n=0}^{\tau-1}\left(F(Q,Y_n)-\bar{F}(Q)\right)\right]$$

is a solution to the Poisson equation (Lemma 4.2 and Theorem 4.2 of Section VI.4, pp. 85-91, of Borkar (1991)). Thus, we have

$$\begin{aligned}
\|V_{Q_1}(y)-V_{Q_2}(y)\|_\infty &= \left\|\mathbb{E}_y\left[\sum_{n=0}^{\tau-1}\left(F(Q_1,Y_n)-F(Q_2,Y_n)-(\bar{F}(Q_1)-\bar{F}(Q_2))\right)\right]\right\|_\infty \\
&\le \mathbb{E}_y\left[\sum_{n=0}^{\tau-1}\left(\|F(Q_1,Y_n)-F(Q_2,Y_n)\|_\infty+\|\bar{F}(Q_1)-\bar{F}(Q_2)\|_\infty\right)\right] \\
&\le \mathbb{E}_y\left[\sum_{n=0}^{\tau-1}\left(2\|Q_1-Q_2\|_\infty+\|\Lambda(\mathcal{B}(Q_1)-(\mathcal{B}(Q_2))\|_\infty+\|\Lambda(Q_1-Q_2)\|_\infty\right)\right] \\
&\qquad\qquad \text{(Using property 1 and 2 from part (a))} \\
&\le 4\|Q_1-Q_2\|_\infty\mathbb{E}_y[\tau] \\
&\le 4\tau_y\|Q_1-Q_2\|_\infty.
\end{aligned}$$

Furthermore, since Q^* solves the Bellman equation, for all $y\in\mathcal{Y}$ we have

$$V_{Q^*}(y)=0.$$

(c) Define $\mathcal{F}_k=\{Q_0,Y_0,\ldots,Q_{k-1},Y_{k-1},Q_k,Y_k\}$. Then due to the Markov property, we have

$$\mathbb{E}[M_k(Q_k)|\mathcal{F}_k]=0.$$

Furthermore,

$$\begin{aligned}
\|M_k(Q_k)\|_\infty &\le \gamma\max_{s,a}\left(\left|\max_{a'\in\mathcal{A}}Q_k(S_{k+1},a')-\sum_{s'\in\mathcal{S}}P(s'|S_k,A_k)\max_{a'\in\mathcal{A}}Q_k(s',a')\right|\right) \\
&\le 2\|Q_k\|_\infty \\
&\le 2\left(\|Q_k-Q^*\|_\infty+\|Q^*\|_\infty\right).
\end{aligned}$$

□

G.2 Proof of Theorem B.1

Proof. To identify the constants A_Q, B_Q and η_Q, we use Lemma B.1 to get

$$A_Q = (A_1 + A_3 + 1)^2 = 25; \quad B_Q = \left(B_1 + B_3 + \frac{B_2}{A_2}\right)^2 = 9\|Q^*\|_\infty^2; \quad \eta_Q = (1-\gamma)\Lambda_{min}.$$

Since $\eta_Q \leq 1$, we have $\frac{B_Q}{\eta_Q} \geq B_Q$. Furthermore, from Lemma B.1 we get $\varrho_{Q,1}$ as follows:

$$\begin{aligned}
\varrho_{Q,1} = uL_s u_{cs}^2 A_2 &= \frac{2(1+\omega)}{\omega}(p-1)\left(|\mathcal{S}||\mathcal{A}|\right)^{2/p} 4\tau_{y_0} \\
&\leq \frac{(1+\omega)}{\omega} 16e\tau_{y_0} \log\left(|\mathcal{S}||\mathcal{A}|\right) \\
&\leq \frac{32e\tau_{y_0}\log\left(|\mathcal{S}||\mathcal{A}|\right)}{(1-\gamma)\Lambda_{min}}.
\end{aligned}$$

$$\begin{aligned}
\varrho_{Q,0} &= \frac{2(1+\omega)(1+50\varrho_{Q,1})}{(1+\omega/\sqrt{e})}\|Q_0 - Q^*\|_c^2 + 36\|Q^*\|_\infty^2 \varrho_{Q,1} \\
&\leq 4(1+50\varrho_{Q,1})\|Q_0 - Q^*\|_c^2 + 36\|Q^*\|_\infty^2 \varrho_{Q,1}.
\end{aligned}$$

□

G.3 Sample complexity for Q-Learning

To find an estimate Q, such that $\mathbb{E}[\|Q - Q^*\|_\infty] \leq \epsilon$, we need

$$\begin{aligned}
\frac{58 B_Q \varrho_{Q,1}\alpha}{\eta_Q} &\leq \frac{\epsilon^2}{2} \\
\implies \alpha &\leq \mathcal{O}\left(\frac{\eta_Q}{\|Q^*\|_\infty^2 \varrho_{Q,1}}\right).
\end{aligned}$$

Using this bound on α, we have

$$\begin{aligned}
\varrho_{Q,0}\exp\left(\frac{-\eta_Q \alpha k}{2}\right) &\leq \frac{\epsilon^2}{2} \\
\implies k &\leq \mathcal{O}\left(\frac{1}{\alpha\eta_Q}\log(1/\epsilon)\right).
\end{aligned}$$

Since $\varrho_{Q,1} \leq \frac{32e\tau_{y_0}\log(|\mathcal{S}||\mathcal{A}|)}{(1-\gamma)\Lambda_{min}}$ and $\|Q^*\|_\infty^2 \leq \mathcal{O}\left(\frac{1}{(1-\gamma)^2}\right)$, we have

$$k \leq \mathcal{O}\left(\frac{\log(1/\epsilon)}{\epsilon^2}\right)\mathcal{O}\left(\frac{1}{(1-\gamma)^5}\right)\mathcal{O}\left(\Lambda_{min}^{-3}\right).$$

H AUXILIARY LEMMAS

Lemma H.1. *The step-size sequence in Assumption 2.5 has following properties:*

$$\alpha_k \leq \alpha_{k-1}; \quad \alpha_{k-1} \leq 2\alpha_k; \quad \alpha_{k-1} - \alpha_k \leq \frac{2\xi}{\alpha}\alpha_k^2 \tag{32}$$

Proof. The step-size is non-decreasing by construction. Now consider the ratio

$$\frac{\alpha_{k-1}}{\alpha_k} = \left(\frac{K+k}{K+k-1}\right)^\xi$$

$$= \left(1 + \frac{1}{K+k-1}\right)^{\xi}.$$

Since $k \geq 1$ we have $\frac{1}{K+k-1} \leq \frac{1}{K} \leq 1$. Putting together with the expression above, we get

$$\frac{\alpha_{k-1}}{\alpha_k} \leq 2^{\xi} \leq 2. \qquad (\xi \leq 1)$$

For the final part, consider the function $f(x) = \frac{1}{(k+x)^{\xi}}$ for $x \in [0,1]$ and $k \geq 1$. Using Taylor's series expansion, there exists $z \in [x,1]$ such that we have

$$\begin{aligned} f(1) &= f(x) + (1-x)f'(z) \\ &= f(x) - \frac{(1-x)\xi}{(k+z)^{1+\xi}} \\ f(x) - f(1) &= \frac{(1-x)\xi}{(k+z)^{1+\xi}} \\ &\leq \frac{(1-x)\xi}{k^{1+\xi}} && (z \geq 0) \\ &\leq \frac{(1-x)\xi}{k^{2\xi}}. && (\xi \leq 1) \end{aligned}$$

Substituting $x = 0$, we have $\forall\ k \geq 1$

$$\begin{aligned} &\frac{1}{k^{\xi}} - \frac{1}{(k+1)^{\xi}} \leq \frac{\xi}{k^{2\xi}} \\ \implies &\alpha_{k-1} - \alpha_k \leq \frac{\alpha\xi}{(k+K-1)^{2\xi}} \leq \frac{2\xi}{\alpha}\alpha_k^2. && (\alpha_{k-1} \leq 2\alpha_k) \end{aligned}$$

□

Lemma H.2. *Consider the following infinite set of equations:*

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + ... + a_{1d}x_d &= 0 \\ a_{21}x_1 + a_{22}x_2 + ... + a_{2d}x_d &= 0 \\ &\vdots \end{aligned} \qquad (33)$$

Assume that the set of equations has a unique solution, i.e., there exists a unique x^* which satisfies $\sum_{j=1}^{d} a_{ij}x_j^* = 0$ for all $i \geq 1$. Define $A(i)$ as a row vector, $A(i) = [a_{i1}, a_{i2}, ..., a_{id}]$. Then, there exists a finite N such that the span of $\{A(i)\}_{i \leq N} = \mathbb{R}^d$.

Proof. Denote $A_i \in \mathbb{R}^{i \times d}$ as the concatenation of the row vectors $\{A(j)\}_{j \leq i}$ into a matrix. Let $r_i = rank(A_i)$. Then, r_i is a non-decreasing sequence that is bounded above by d. This follows from the observation that the span of a set of vectors is non-decreasing as new vectors are added to the set. Thus, by Monotone convergence theorem, $\lim_{i \to \infty} r_i$ must exist. Let us denote the limit by r.

Now, to show that $r = d$, we will use the method of contradiction. Assume that $r < d$. Since $r_i \in \mathbb{Z}$, there exists a finite number N such that $r_i = r,\ \forall i \geq N$. However, this implies that the null space of A_i is nonempty for all i, further implying that the set of equations (33) has more than one solution. Hence, we have a contradiction and r must be equal to d. Since the column rank and the row rank of a finite-dimensional matrix are equal, this implies $dim(\{A(i)\}_{i \leq N}) = rank(A_N) = d$. The claim follows. □

Lemma H.3. *Let P be a square matrix with dimension $d_1 + d_2$ which partitioned as follows*

$$P = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

where $A \in \mathbb{R}^{d_1 \times d_1}$, $B \in \mathbb{R}^{d_1 \times d_2}$, $C \in \mathbb{R}^{d_2 \times d_1}$, and $D \in \mathbb{R}^{d_2 \times d_2}$. Then, $\|P\|_2 \leq \|A\|_2 + \|B\|_2 + \|C\|_2 + \|D\|_2$.

Proof. Using the definition of matrix norm, we have

$$\|P\|_2 = \max_{\|x\|_2=1} \|Px\|_2$$

Let $x = \begin{bmatrix} y \\ z \end{bmatrix}$, where $y \in \mathbb{R}^{d_1}$ and $z \in \mathbb{R}^{d_2}$. Then, Px can be written as

$$Px = \begin{bmatrix} Ay + Bz \\ Cy + Dz \end{bmatrix} = \begin{bmatrix} Ay \\ 0 \end{bmatrix} + \begin{bmatrix} Bz \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ Cy \end{bmatrix} + \begin{bmatrix} 0 \\ Dz \end{bmatrix}$$

Then, by triangle inequality, we have

$$\|Px\|_2 \leq (\|A\|_2 + \|C\|_2)\|y\|_2 + (\|B\|_2 + \|D\|_2)\|z\|_2$$

Note that since $\max\{\|y\|_2, \|z\|_2\} \leq \|x\|_2 \leq 1$, we have

$$\max_{\|x\|_2=1} \|Px\|_2 \leq \|A\|_2 + \|B\|_2 + \|C\|_2 + \|D\|_2.$$

□